Similarity Measures for Rational Set Selection and Analysis of Combinatorial Libraries: The Diverse Property-Derived (DPD) Approach

Richard A. Lewis,* Jonathan S. Mason,† and Iain M. McLay

Computer-Aided Drug Design, Dagenham Research Centre, Rhone-Poulenc Rorer, Dagenham, Essex, RM10 7XS

Received November 29, 1996⊗

The generation of new chemical leads for biological targets is a very challenging task for researchers in the pharmaceutical industry. The design of representative screening sets and combinatorial libraries is central to achieving this objective. In this paper, we describe a novel molecular descriptor, the Diverse Property-Derived (DPD) code, that contains information about key molecular and physicochemical properties of a molecule. The utility of this descriptor is explored through its application for the selection of a maximally diverse representative screening set, through the selection of secondary screening sets to obtain more information concerning the structure−activity relationships (SAR) of a particular target receptor, and through the profiling of combinatorial libraries. The usefulness of physicochemical/molecular property descriptors, such as the DPD code, is discussed critically.

INTRODUCTION

The generation of new chemical leads for biological targets is a very challenging task. There are several strategies for finding new leads, including quasi-random biological screening which has played an important role in drug discovery for many years. It would be preferable, in our view, to incorporate as much explicit design as possible into the lead generation process, preferably such that a better understanding of structure−activity relationships can be gained. The design of focused and diverse screening sets using *a priori* hypotheses will give such an insight. In this paper, we describe a novel molecular descriptor, the Diverse Property-Derived (DPD) code, that is designed to contain information about key molecular and physicochemical properties of a molecule. We will discuss its application to the selection of a representative screening set, the selection of secondary screening sets to obtain more information concerning the SAR of a particular target receptor, and the profiling of combinatorial libraries. The usefulness of molecular and physicochemical descriptors, such as the DPD code, is discussed critically.

The original goals of our studies into molecular similarity were to provide a rational framework for selecting representative sets of compounds for biological screening and to provide a mechanism for selecting further compounds to follow up initial leads. Corporate databases of chemical compounds contain a wealth of information and provide a very rich source of compounds for screening. At the genesis of the project in 1991,[1] the large size of our corporate databases (RPR > 150 000, RP Agrochemicals > 350 000 compounds) precluded the systematic screening of all compounds. Even today, the brute-force method of high-throughput screening is still relatively expensive to perform in terms of designing and validating assays and providing protein, and it may not be possible to screen all possible compounds for every screen. An alternative strategy is to abstract a small subset of a large database so that the subset represents as many as possible of the key features of that database in an economical and nonredundant fashion. This requires the *a priori* formulation of hypotheses, by the team selecting the sets, concerning what constitutes a key feature, and how that feature should be measured. The same arguments apply to the design of combinatorial libraries. It is our experience that the needs of different medicinal chemistry projects will emphasize different features; the procedure described in this paper attempts to provide a framework that avoids such biases.

Other workers in this field have classified molecules on the basis of the functional groups the molecules contain.[2] We have also performed this type of structural (chemical family) classification;[3] however, we feel that this type of analysis is unsatisfying from the perspective of understanding ligand-receptor interactions. A receptor or enzyme does not recognize particular atoms or groups, it interacts with the properties in space (electrostatic- and orbital-based) projected by a certain geometric arrangement of these atoms. Functional group classification ignores the possibility of bioisosterism and gives little idea of how similar two groups are in terms of their receptor binding properties. A similarity metric based on molecular properties was developed in an attempt to provide a general measure of molecular similarity based around this view of ligand-receptor interactions and to answer some of the objections to functional group classification. Molecular properties do not describe specific geometries of interaction, and they smear conformational space into a single lump: in this sense, molecular properties are not the absolute answer to molecular similarity, but they are interesting and useful descriptors nonetheless. As part of this work, several novel molecular descriptors were developed to represent the electronic and steric properties of a molecule that were statistically uncorrelated to other descriptors. Forty-nine molecular descriptors (including standard physicochemical descriptors) were considered; a subset of descriptors was selected based on a statistical analysis and on our biases about ligand-receptor interactions. The validation of the method of subset selection was provided

† Current address: Collegeville Research Center, Rhone-Poulenc Rorer, 500 Arcola Road, Collegeville, PA 19426.

⊗ Abstract published in *Advance ACS Abstracts,* April 15, 1997.

S0095-2338(96)00471-4 CCC: $14.00

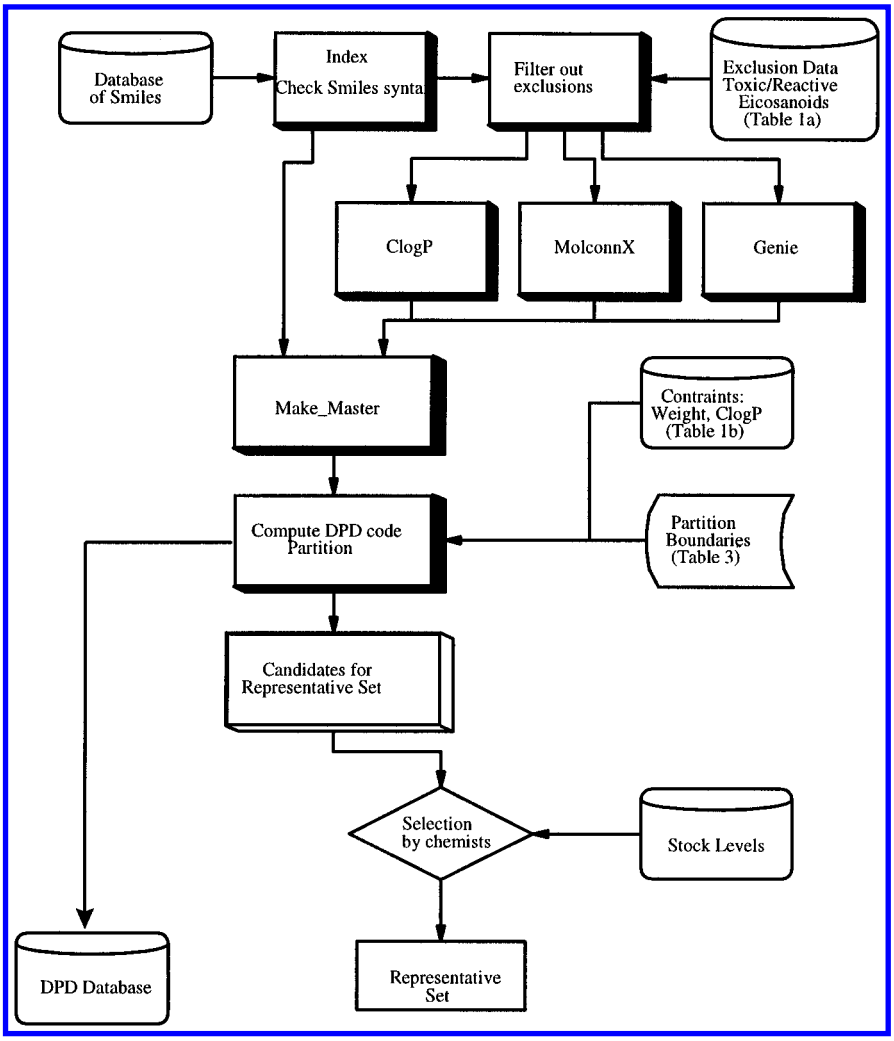


Figure 1. A flowchart describing the logic of how databases of diverse-property set data are built up, and how the representative sets are selected.

by examining several independent databases: two publicly available databases, Pharma Projects (PP; 6000 compounds in 1991), and the Standard Drug File (SDF; 25 000 compounds in 1991) were used. The RPR database itself was broken into three significant selections that were treated independently, derived historically from the different research sites of the company: RPR_1 (UK, >60 000 compounds), RPR_2 (France; >70 000 compounds) and RPR_3 (U.S.A,, >18 000 compounds). As our goal was a small set "representative" of the diversity of a database, rather than of the actual proportion of compounds, a partitioning approach was followed; a complementary approach involving clustering of "fingerprints" of substructural information is the subject of another paper.[3]

A general screening set selected from the RPR_1 database, containing about 350 compounds, was prepared and assayed in several screens; the final set of the desired size of ~1000 compounds contained compounds from all three segments of the database. The molecules in this set were chosen to provide the widest possible diversity of compounds for test, given our model of molecular similarity, covering the whole range of suitable structures that occur in the database. This collection has provided a number of hits in biological assays; one of these hits will be discussed critically to try to understand what role this type of similarity measure can play in future diversity studies.

This paper describes the construction of a similarity measure based on molecular descriptors and properties and issues concerning the selection of rational sets and design of combinatorial libraries and gives brief details of the analysis and utility programs that have been developed during the project. Future applications and developments are also discussed.

1. METHODS

The structure of a compound can be represented by a SMILES code, which contains chemically coherent information about the atom types and bonding patterns in the molecule.[4,5] Many molecular descriptors and properties can be estimated from this atom and bond information, for instance, hydrophobicity (octanol/water partition coefficient[6]), flexibility, charge distribution, and molecular volume. Much more information can be calculated using the 3D structures generated from these codes; this topic is discussed in another paper in this series.[7] Software has been written to process automatically large numbers of these SMILES codes (derived from structural databases) into a master file. A flowchart outlining the logic of the method is given in Figure 1. Each line of a master file contains 49 molecular descriptors, and 46 functional group counts for each compound. A statistical and literature analysis of the descriptors was performed, and a subset of six key descriptors was chosen. To generate a

general screening set from a corporate database, the database was partitioned into six-dimensional boxes, and representatives from each box were chosen for inclusion in the general DPD screening set. This was done by choosing a compound from a list of eight randomly chosen candidates from each partition from each of the three segments of the database (RPR_1, RPR_2, and RPR_3); the actual selection was made or validated by an experienced medicinal chemist. In the case where none of the eight compounds were suitable, further choices were provided. This strategy of partitioning generates a bin number, which is unique to a box. Using the bin number, all compounds (from the same or a different database) in a particular partition can be located and used to follow up initial leads. We now describe each part of the process in more detail.

1.1. Generation of Molecular Descriptors. The molecular descriptors were generated from three separate commercial software packages: atom and group counts were produced using GENIE,[8] electronic and flexibility indices were calculated by MOLCONN-X,[9] and *C*log*P* and CMR using Daylight v353.[10] These programs are described in more detail below. All other programs were written and developed in house. The data are combined into a single master file by the MAKE_MASTER program, and the DPD code for each molecule is computed by the PARTITION program. A brief listing of the descriptors computed is given in Chart 1. Further information is available in ref 9. These programs are controlled by a single batch control program, MAKE_DPD, to make the production of DPD sets as straightforward, error-free, and efficient as possible.

All initial studies were performed under VAX VMS, using DCL scripts to control job submission to the VAX cluster. As this platform is no longer supported by Daylight, all current work is now performed on an SGI cluster. The implications of this change will be discussed.

1.1.1. MOLCONN-X. The MOLCONN-X software, from Kier and Hall,[9] computes a wide range of topological indices for a molecular structure. These indices have been widely and successfully used in the development of QSAR and QSPR equations,[11] demonstrating their power as molecular descriptors. In addition to the κ molecular shape indices, χ molecular connectivity indices, topological state indices, and atomic electrotopological state indices, other composite indices considered to be of potential use were also calculated. These included the flexibility index (The product of the κ α1 and α2 shape indices divided by the number of vertices.) and the molecular normalized electrotopological indices used in this work. We devised the normalized molecular electrotopological indices to combine information about the electronic interactions and the topological environment of each atom into a single overall molecular value. Inspection of the results for a diverse sampling of compounds and functional groups indicated that the values calculated gave a sensible classification of "polarity" (from a medicinal chemistry perception) and were superior to an index calculated from atomic partial charges. The standard valence state electronegativity index (including perturbations from neighboring atoms in the molecule) is computed for each atom in a molecule.[12] The normalized index is the sum of the squares of the atomic indices, divided by the number of atoms. We also devised a descriptor that we have called the aromatic density descriptor; this was developed as other potentially useful indices we had calculated were too correlated with other indices of interest (see below). Aromatic density is simply the number of aromatic rings in a molecule divided by the molar volume (computed by Schroeder's method[13]). We believe that this descriptor represents, in a very approximate way, the ability of the molecule to form aromatic interactions with the receptor, and some aspects of the shape of a molecule.

Chart 1. DPD Master File Format

	ClogP Error
Variable	CMR
Registry Name	CMR Error
DPD code	#basic N
Quality Flag	#aniline NH
#aromatic rings	#aromatic NH
#h-acceptors	#amide NH
#h-donors	#acid
#rotatable bonds	#alcohol
molecular volume	#thioalcohol
#heavy atoms	#urea
formula weight	#thiourea
flexibility index	#amide
Electrotopological index (with squares)	#thioamide
Normalized Electrotopological index	#ketone
Electrotopological index (with modulos)	#thioketone
Normalized Electrotopological index	#aldehyde
Kappa 0	#thioaldehyde
Kappa 1	#ester
Kappa 2	#ether
Kappa 3	#thioether
Kappa 4	#nitro
Kappa 5	#aromatic N:
Kappa 6	#sulfoxide
Shannon index	#sulfone
Total topological index	#sulfonamide
#paths	#nitrile
Sum intrinsic I values	#aromatic F
Sum delta I values	#N-oxide
Total electrotopological state index	#N-(O,N)H
#bonds	#N-(O,N):
#elements	#hydrazineH
Idw Bonchev-Trinajstic information indices	#hydrazine:
Average Idw	#tetrazole
Idc Bonchev-Trinajstic information indices	#sulfonic acid
Average Idc	#phosphor acid
Terminal group	#carbamic acid
Terminal group3	#acid sulfonamide
Terminal methyl	#acidic amide
Terminal methyl3	#acidic groups
Wiener number	#toxic groups
Wiener p	#reactive halides
Platt f	#reactive epoxides
total Wiener	#epoxide
KnotP	#sulfonates
KnotVP	#anhydride
#N atoms	#NCS,NCO,CO_3H
#O atoms	#organo-P
#S atoms	#C+, N+
Andrews' binding constant	#cations
ClogP	

1.1.2. *C*log*P*. The BDRIVE module of the Daylight v3.54 software running under VAX VMS was originally used for the calculation of *C*log*P* (calculated log P from the CLOGP3 algorithm) and CMR (Calculated Molar Refractivity) values. The current implementation now uses v441 from Daylight running under SGI Irix5.2. This switch is not without its effects. The values of *C*log*P* computed by the two versions were compared for a set of 10 000 compounds. The results were that the *C*log*P* values for 609 compounds changed by 5−10%, 2255 compounds changed by more than 10%, of which 1077 were different by 0.2−0.5 log units and 1140 by more than 0.5 log units. This is consistent with the reparameterization of fragments that occurs with each new release,[14] but it does imply that to make meaningful comparisons between databases, *C*log*P* values of all databases should be derived using the same version of the program. The mean and spread of the values did not seem to be affected.

1.1.3. GENIE. A GENIE routine (Daylight software running under VAX/VMS) was written to derive various fragmental properties from the SMILES codes, using interpreted SMARTS substructure queries. These included the following: number of hydrogen-bond acceptor groups; number of hydrogen-bond donor groups; number of aromatic rings; number of flexible bonds; and molecular volume. In addition, frequency counts of 46 functional groups were computed. Although these are not used in the work described in this paper, the extra information was generated for future use in alternative measures of molecular similarity. The switch to SGI IRIX again has caused difficulties, as the GENIE program has not been implemented on this platform. A basic GCL parser[15] was therefore written in C, using the Daylight toolkits, to replicate the functions of GENIE; to date, our program can interpret standard SMARTS and compound SMARTS statements, additions, conditional statements, simple multiplications, and print formatting statements. As this was sufficient to our needs, other functions were not implemented. The parser has been useful in several other roles, including filtering for toxicity, and lately in selection of reagents for combinatorial libraries.

1.1.4. Filtering of the Data. Molecules in the database can be subjected to several optional layers of filtering, based on chemical formula, molecular weight, and charge. We made the explicit assumption that we would exclude from our analysis all compounds that were chemically unsuitable for general screening. The reasoning was that the inclusion of these compounds could add noise to the analysis, as we were only interested in compounds that were reasonable for further study by medicinal chemists. We have applied this principle both in the selection of rational sets and in profiling to aid the design of combinatorial libraries. As for the GENIE program, the FILTER program is written as a series of substructural queries, interpreted by our GCL parser. The FILTER program can be used to flag molecules that are reactive and may bind nonspecifically to proteinaceous material (e.g., acid halides), cytotoxic or that exhibit a wide range of potent biological activities and so are not suitable for general screening (e.g., prostanoids). The filtering criteria are given in Table 1a. Flagged molecules were removed from the DPD analysis.

1.2. Utility Programs. The MAKE_MASTER program combines all the sources of data into one file; the PARTITION program analyses each line of the master file, computes the DPD code, and inserts it back into the master file. Further filtering mechanisms were incorporated into these programs: trivial compounds with low molecular weight or small numbers of atoms and very large molecules can be flagged, as can compounds for which some of the molecular descriptors cannot be accurately computed. This normally indicated a problem in the original SMILES code, that can then be corrected. One error that occurs more than others is that the number of marked ring bonds exceeds 10; as our version of MOLCONN-X could not interpret the %N syntax, an error is produced. It should be stated that, where possible, molecular descriptors and DPD codes are produced for all molecules in the database. The filters were applied for the statistical analysis of the molecular descriptors and, in our view, should also be applied to the selection of rational sets for general screening; they are also useful indicators in the profiling and design of combinatorial libraries, particularly those intended for general screening.

Table 1. The Chemical Filtering Criteria Used in the Form of Substructural Queries That Remove Unwanted Structures from the Database and To Create the PDP Databases and the Partition Sets

a. In the Form of the Substructural Queries

toxic or reactive substructures	very biologically active substructures
reactive epoxides, thiepoxides, aziridines	prostaglandins
acid halides, acid anhydrides	prostacyclins
reactive carbon halides	thromboxanes
sulfonyl or phosphoryl halides, sulfonates	
NCS, NCO, COOOH	
silyl halides, silanes, silates	
acyclic aminals, cyanohydrins	
unstabilized enols and enolates	

b. To Create the DPD Databases and the Partition Sets

filter	allowed values	used in Make_Master	used in partitioning
formula weight	$150 < X < 565$	√	x
no. of heavy atoms	$10 < X < 50$	√	x
*C*log*P* error	$X < 59$	x	√
CMR error	$X < 41$	x	√
no. of N, O, S atoms	$X \geq 1$	x	√

The MAKE_DPD program takes as input a file of SMILES codes and generates a master file, and, if required, a partitioning of the data. There are several options to select the levels of filtering that are used on the data, depending on the application. The program was designed to take advantage of the implicit parallel computing facilities in a VAX cluster, when the tasks can be run in parallel on different machines. A similar facility was written for our SGI cluster, but the queuing and job scheduling facilities are not as well implemented.

The DATA_SEARCH program is a data-mining utility that can take as input a list of registry numbers, a DPD code, or a single compound ID and create files containing DPD data or lists of similar compounds contained in the corporate database in various formats. It can also perform several other functions that are beyond the scope of this paper. The output can be used to create local databases in any system that can import SDF files.[16]

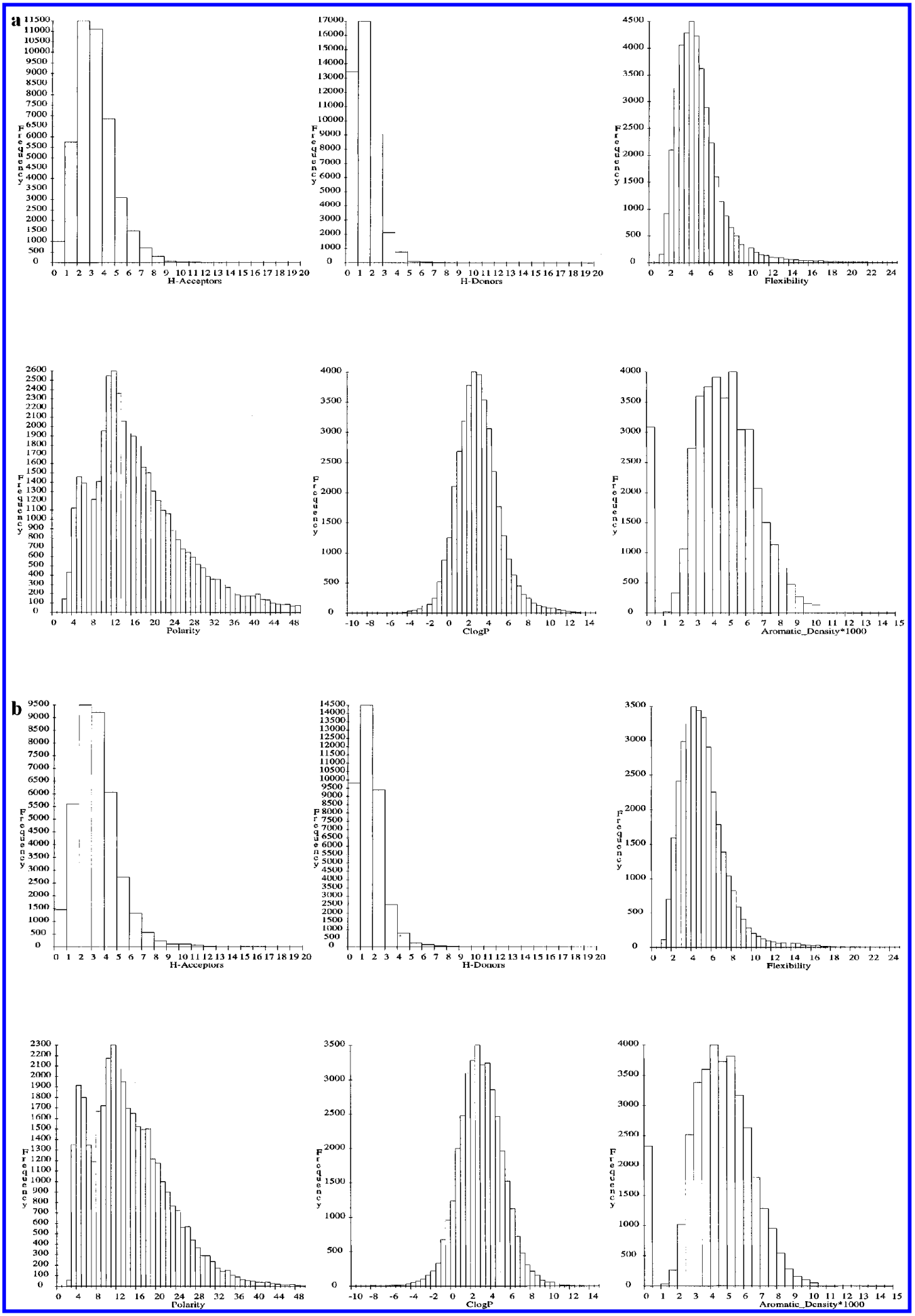
1.3. Statistical Analysis of the Molecular Descriptors. The filtered data in the RPR_1 master file were used to determine which molecular descriptors should be used to classify the database; this, in SAR terms, was our training set. The filtering criteria are given in Tables 1a,b. We

decided to select a set of descriptors that could measure hydrophobicity, polarity, flexibility, shape, hydrogen-bonding properties, and aromatic interactions, reflecting our biases about ligand-receptor interactions. We chose to work in chemical space rather than looking at biological activity space, because our goal was to derive an assay-independent metric. We looked at the literature which gave the original derivation and use of each descriptor where possible and eliminated most of them because their primary function was not to describe hydrophobicity, polarity, flexibility, shape, hydrogen-bonding, or aromaticity. For example, the Shannon index is useful for describing molecular symmetry,[17] but we felt that it would not be useful as a generalized descriptor of shape. Preliminary statistical analyses showed that many potentially relevant descriptors were highly correlated, this led us to develop the aromatic density index, which combined potentially relevant information into a reasonably noncorrelated descriptor. As with the functional group counts, we still calculate all the descriptors, as they may be useful in other similarity metrics, based on different assumptions. A final statistical analysis of the 17 molecular descriptors (#rotatable bonds, #h-acceptors, #h-donors, molecular volume, flexibility index, electrotopological index (with squares), normalized electrotopological index, electrotopological index (with modulos), normalized electrotopological index, #paths, $\kappa 2$, total topological index, sum intrinsic *I* values, total electrotopological state index, Andrews' binding constant, *C*log*P*, CMR) remaining after this literature analysis was performed using RS/1[18] for each compound in an database of 42 700 molecules derived from the RPR_1 collection (the remainder having been removed by the various filters); six descriptors stood out clearly as being only weakly correlated. The other descriptors (for example, CMR) showed much higher correlations. The final correlation table is shown below (Table 2a). The largest magnitude correlation between any pair of descriptors was 0.5, between *C*log*P* and the flexibility index. This is perhaps understandable, if we assume that rotatable bonds will be mostly composed of saturated groups. As an aside, it has been pointed out that the *C*log*P* values for flexible compounds are themselves likely to be overestimates, as the extra groups are still treated additively, not allowing for the possibility of internal collapse to bury hydrophobic surface. There is also a correlation of −0.4 between aromaticity and flexibility, which again is intuitive. The magnitudes of the other correlation coefficients were of the order of 0.25, indicating that there is not much pairwise correlation between the descriptors. This is important to the partitioning strategy. Pairwise uncorrelated descriptors pass the first test for orthogonality, so the data are more likely to be distributed evenly across the descriptor space (the descriptors could be related by multiple collinearities, which would invalidate our remarks about orthogonality; we did not investigate this possibility). We made the assumption that our descriptors were orthogonal and hence that partitioning would be able to divide space evenly and so produce a representative sampling.

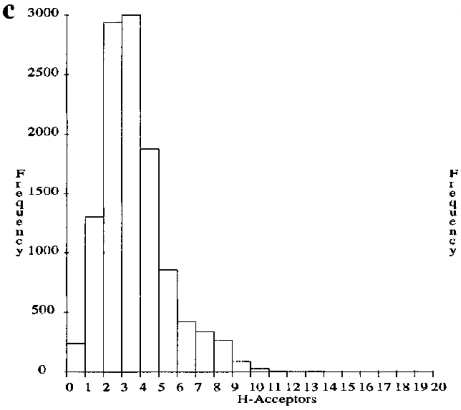
Correlation matrices are not the only method for selecting key descriptors;[19] we decided against techniques such as principal component analysis because we wanted to maintain the essential simplicity of the descriptors. Another method would have been to look for the spanning set of descriptors: if the set of the six chosen descriptors can explain (using principal component analysis) or predict the variance (using partial least squares) in the other descriptors, then they are a spanning set, and the information contained in the other descriptors can be said to be largely redundant. Our six descriptors are probably not a spanning set, as our motivation in choosing them was not only purely statistical but also chemical. Correlation matrices were obtained from the other databases examined (RPR_2, RPR_3, PP, SDF; Table 2b−e) using the same protocol. As may be seen from the tables, the patterns of correlation values are broadly similar. As the databases are independent sources of data and have been treated separately, we believe that these correlations are a general phenomenon and provide a justification for the unsophisticated statistical approach followed. It should be noted that the SDF database has some higher correlations (five values with magnitude > 0.5), the largest in the numbers of h-bond acceptors and donors (0.63), whereas the PP database, which should contain similar compounds (both being databases of available drugs), has lower correlations. We know that the SDF database also contains dentifrices, spermicides, and disinfectants which we speculate might be the cause of these observations, but we have no hard evidence for this position. Frequency histograms of each of the six descriptors plotted for each of the five databases show a remarkably similar profile across the databases, again indicating that the properties of the descriptors are not tied to the origin of the medicinal chemistry compound databases (Figure 2a−f). However, we would be wary of extending these conclusions to other types of nonmedicinal chemistry databases (for example, the CAS or Available Chemicals

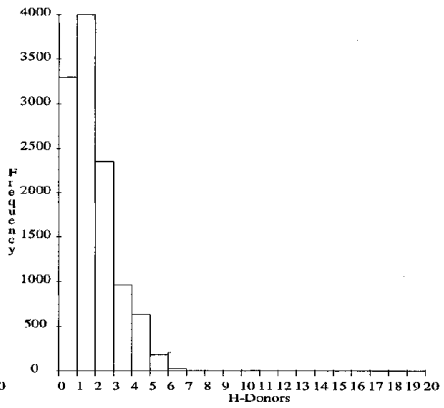
Table 2. The Correlation Matrix for the Molecular Descriptors Derived from the RPR_1 RPR_2, RPR_3, PP, and SDF Database

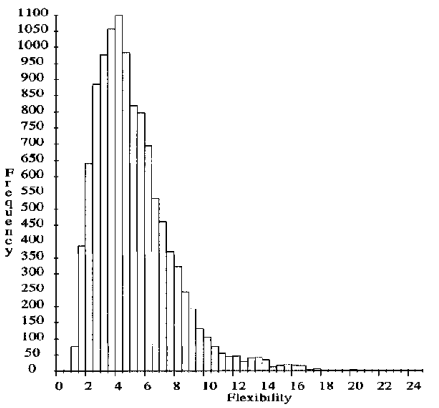
descriptor	H-acc	H-donor	flexib	electro	*C*log*P*	arom. d.
			a. RPR_1 Database			
H-acceptor	1.00	0.28	0.22	0.24	−0.16	0.13
H-donor		1.00	0.23	0.00	−0.19	−0.10
flexibility			1.00	−0.06	0.50	−0.40
electrotopological				1.00	−0.27	−0.10
*C*log*P*					1.00	0.04
aromatic density						1.00
			b. RPR_2 Database			
H-acceptor	1.00	0.20	0.18	0.43	−0.36	0.06
H-donor		1.00	0.28	0.01	−0.19	−0.22
flexibility			1.00	−0.05	0.39	−0.42
electrotopological				1.00	−0.40	−0.11
*C*log*P*					1.00	0.11
aromatic density						1.00
			c. RPR_3 Database			
H-acceptor	1.00	0.49	0.33	0.29	−0.44	0.05
H-donor		1.00	0.24	0.17	−0.49	−0.16
flexibility			1.00	0.14	0.05	−0.45
electrotopological				1.00	−0.38	−0.30
*C*log*P*					1.00	0.17
aromatic density						1.00
			d. PP Database			
H-acceptor	1.00	0.23	0.20	0.35	−0.38	−0.11
H-donor		1.00	0.20	0.17	−0.44	0.08
flexibility			1.00	0.04	0.17	−0.13
electrotopological				1.00	−0.46	0.01
*C*log*P*					1.00	−0.05
aromatic density						1.00
			e. SDF Database			
H-acceptor	1.00	0.63	0.15	0.52	−0.54	−0.01
H-donor		1.00	0.14	0.32	−0.50	−0.10
flexibility			1.00	−0.03	0.25	−0.42
electrotopological				1.00	−0.50	0.18
*C*log*P*					1.00	0.10
aromatic density						1.00

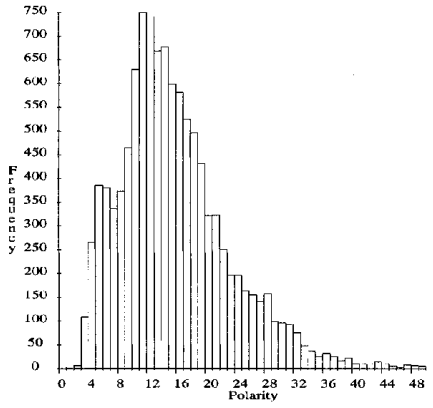

a
Frequency
H-Acceptors
H-Donors
Flexibility
Polarity
ClogP
Aromatic_Density*1000
b
Frequency
H-Acceptors
H-Donors
Flexibility
Polarity
ClogP
Aromatic_Density*1000

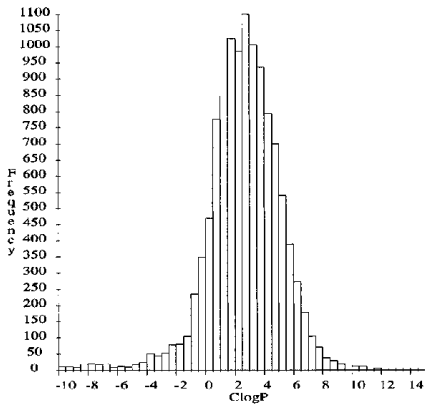
c

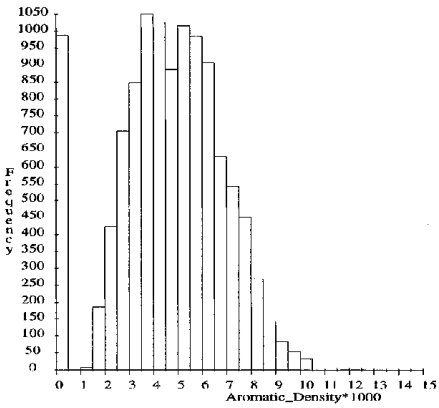

Frequency
H-Acceptors


Frequency
H-Donors

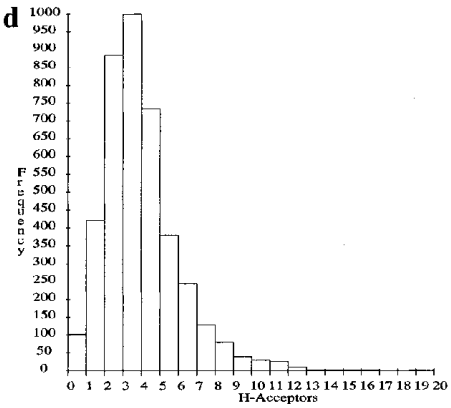

Frequency
Flexibility

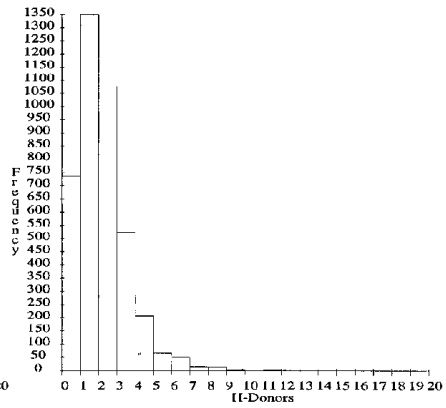

Frequency
Polarity

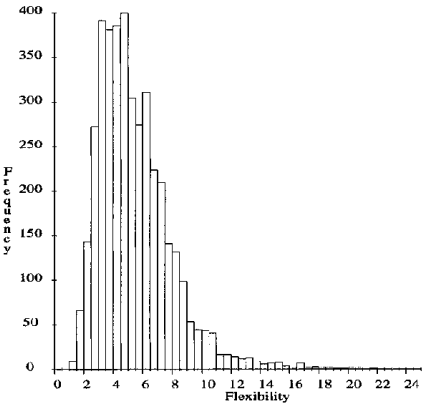

Frequency
ClogP

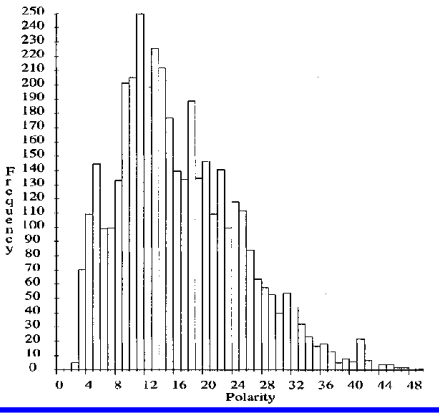

Frequency
Aromatic_Density*1000

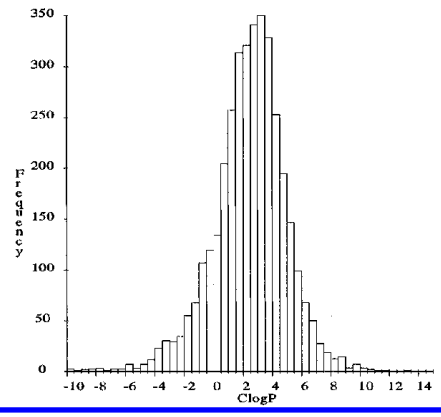
d

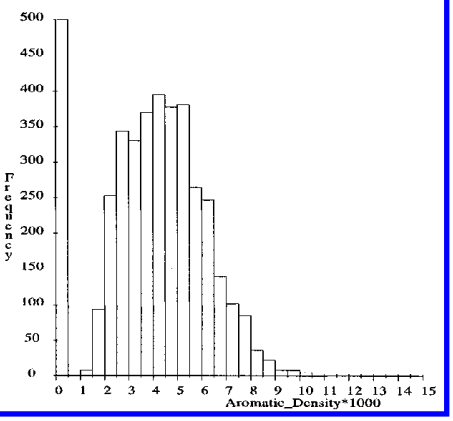

Frequency
H-Acceptors

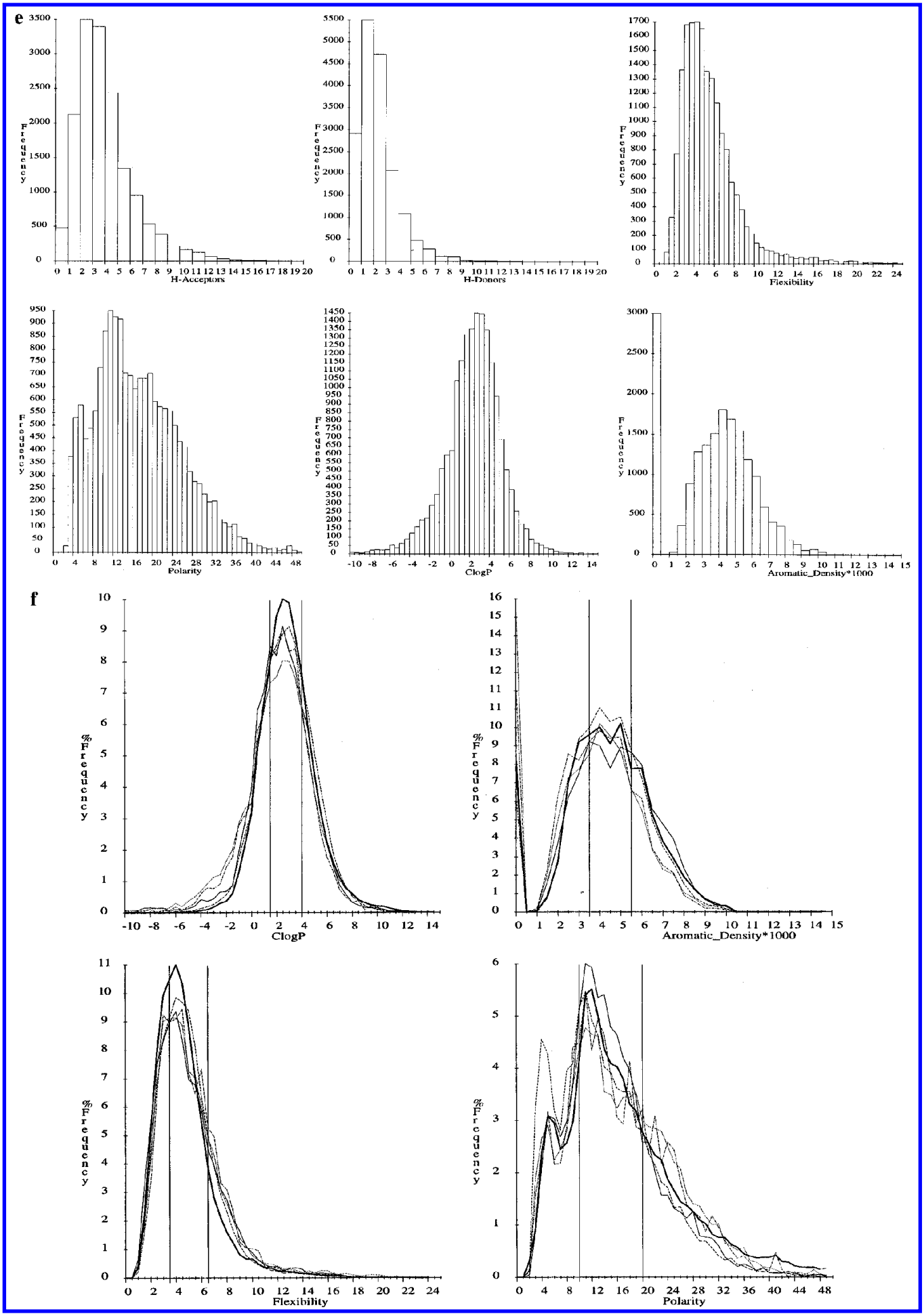

Frequency
H-Donors


Frequency
Flexibility


Frequency
Polarity


Frequency
ClogP


Frequency
Aromatic_Density*1000


e
Frequency
H-Acceptors
H-Donors
Flexibility
Polarity
ClogP
Aromatic_Density*1000
f
%Frequency
ClogP
Aromatic_Density*1000
Flexibility
Polarity

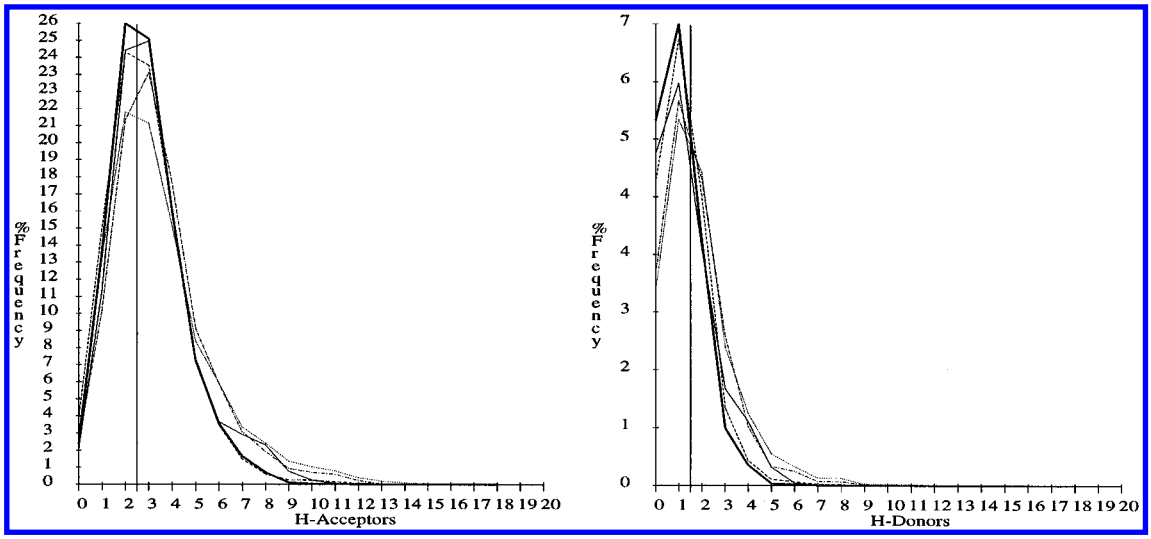


Figure 2. Frequency histograms of h-bond donor, h-bond acceptor, flexibility, polarity, *C*log*P* and aromatic density for each of the databases. The data is first plotted by database: (a) RPR_1; (b) RPR_2; (c) RPR_3; (d) PP; (e) SDF. (f) The data is plotted by descriptor, with the partitions marked by vertical lines, and the curve for the RPR_1 database drawn in the heaviest line, RPR_2 by a dashed, RPR_3 by a solid, PP by dotdashed, and SDF by a dotted line, respectively.

Directory databases) without performing a test analysis first, simply because we have little knowledge of the nature and behavior of the data from these sources. We also note that statistical comparisons of the data (both means tests for normal distributions and χ^2 tests were performed) showed that all the distributions were significantly different at the $p < 0.005$ level. We are therefore committing a type I error (rejecting a hypothesis when statistical tests suggest it should be accepted) in saying that the distributions are similar; we believe that this is justified by observation, and the difficulties created by large population sizes. The large number of data points makes it possible to detect very small differences in two distributions, differences that are not necessarily meaningful.

1.4. Partitioning of the Data to Create the DPD Descriptor. The next stage of our analysis was to combine the six descriptors into a single descriptor, the DPD code, that is a linear combination of the components. Two methods were considered: clustering and partitioning. Techniques for clustering chemical objects have been well reviewed by other researchers.[20] For the purpose of clustering, the object would be a point in the six-dimensional descriptor space, and the Euclidean distance between objects would be the dissimilarity of the objects, assuming equal weighting between the descriptors. A simple composite code would then be the cluster identifier after clustering of the data. The number of clusters can be set arbitrarily. There were two factors that weighed against the use of clustering in this study: the application of a clustering method makes the assumption that the data is in fact amenable to clustering (in other words, most clustering methods will produce a clustering whatever the data); to the authors' knowledge, there are no simple ways of testing if this assumption is justified for a very large dataset. Certainly, cluster significance tests have been proposed,[21,22] but they are quite computationally expensive and not practicable to apply to very large datasets. The second and most important factor is the lack of generality of the descriptor. If the descriptor was defined by the clustering of one database, it is hard to define the descriptors for compounds in a second database without a large number of expensive distance calculations, and some arbitrary definitions of cluster dimensions. Partitioning is best described as a boxing algorithm: each descriptor is divided into ranges; a combination of descriptor ranges makes a partition or box. The composite descriptor is then effectively the coordinate vector of one of the vertices of the box. At an even simpler level, the coordinate vector can be made up of the (integer) names of the lower values of defining ranges. For instance, ethyl benzoate, with property values of number of H-acceptors = 1, number of H-donors = 0, flexibility = 2.81, electrotopology = 14.9, *C*log*P* = 2.64, aromatic density = 4.76 would be assigned to partition 111 222, using the descriptor ranges described below in Table 3. The complete set of partitions is formed by taking all the combinations of all the ranges into which the molecular descriptors have been divided. It is completely portable between different databases provided the same descriptors and ranges are used. We freely acknowledge that there are disadvantages to the partitioning algorithm, in the arbitrary way in which the ranges must be set, and the introduction of edge effects when a partition boundary slices between two very similar compounds; an answer to this issue may come though the application of fuzzy logic. However, we felt that the portability of the descriptor was key to its usefulness. Other workers have also addressed the issue of which method, clustering or partitioning, gives the better performance; there is not yet a consensus on this subject.[23,24]

The initial partitioning study was again performed using only the filtered RPR_1 database. Molecular properties were calculated for 42 700 structures. The structures for which a value of *C*log*P* could be computed at a reasonable error level (no missing fragments, incorrect bonding), and which had a formula weight between 150 and 565 daltons (24 828 compounds) were used in the statistical analysis. The justification for the molecular weight limits were based on the extremes of molecular weight found in small molecule drugs (metronidazole and pristinamycin). Heavier and lighter bioactive compounds can be found, but these are the exception rather than the rule. The filtering rules are given in Table 1a,b.

Frequency histograms for each descriptor were plotted and used to select divisions for the partitions. The frequency

Table 3. Divisions created for Each Molecular Descriptor[a]

descriptor	class	range of *X*
no. of H-bond acceptors	1	$X < 2.5$
	2	$X > 2.5$
no. of H-bond donors	1	$X < 1.5$
	2	$X > 1.5$
flexibility index	1	$X < 3.5$
	2	$3.5 < X < 6.5$
	3	$X > 6.5$
normalized sum of the squares of the electrotopological indices	1	$X < 10.0$
	2	$10.0 < X < 20.0$
	3	$X > 20.0$
*C*log*P*	0	contains N+
	1	$X < 1.5$
	2	$1.5 < X < 4.0$
	3	$X > 4.0$
aromatic density	0	nonaromatic
	1	$X < 3.5$
	2	$3.5 < X < 5.5$
	3	$X > 5.5$

[a] The number of partitions is equal to the product of the number of divisions.

histograms show a reasonable approximation to normal distributions, allowing for the integer nature of the number of hydrogen-bond acceptors or donors. The dividing values for the descriptors were set to split the histograms from RPR_1 into regions with equal areas under the curve. A histogram was produced for each property (Figure 2a−f). There were two exceptions to this binomial pattern, and extra divisions were set up to allow for them: compounds that contain no aromatic rings which were assigned to aromatic density partition 0 (this has fully aliphatic compounds such as steroids and long alkyl chain acids and bases) and compounds containing a quarternary nitrogen, which were assigned to *C*log*P* partition 0. The reason for the latter assignment is that the *C*log*P* values for compounds containing quaternary nitrogens are not at all reliable (missing fragments). The divisions used (Table 3) give a total of 432 partitions for all combinations of property values, or 576 partitions if the *C*log*P* class 0 (containing only quaternary N+ compounds) is included. The number of partitions is simply the product of the number of ranges for all the descriptors and is arbitrary. Pragmatic screening considerations dictated our choice of 432 partitions, but any number of partitions could be chosen. An increase in the number of ranges increases the resolution of the descriptor.

Selection of General Screening Sets Using the DPD Partitioning. A database of compounds can be classified using the partitioning descriptors and ranges so that the compounds fall into one of 432 or 576 partitions. The partitioning of a database of 25 000 compounds (after filtering) will put, on average, 50 (0.2%) compounds into each partition (The average is, of course, dependent on the size of the original database. The graphs of the distribution of compounds against percentage partition occupancy are given in Figure 3a−e.). The DPD general screening set is made up by selecting one member from a random selection of eight molecules taken from each partition (with the proviso that stocks of that compound are available!): if none of the eight were suitable, a further tranche was taken. The assistance of experienced medicinal chemists at this stage is invaluable. Not all partitions are filled, as the structures that would fall in these partitions may not be chemically or pharmaceutically feasible, and for partitions with few representatives, there may not be a compound with adequate stock levels available. The number of unfilled boxes across the full set of 576 partitions is given in Table 4. There are 40 boxes (7%) that are unfilled across all five databases: 16 of these are from the *C*log*P* class 0 (which probably reflects poor sampling), 9 have flexibility class 3 and *C*log*P* class 1, 11 have flexibility class 1 and *C*log*P* class 3, and the remaining 4 have extreme values of the aromatic density descriptor. Because of the correlations between descriptors and basic chemical intuition, one would not expect many compounds to have a high flexibility index and a low value of *C*log*P*. The converse is not true (polyarenes, for example), but these classes of compounds may not be well represented in medicinal chemistry databases. It would also be difficult to imagine a small organic molecule (<565 daltons) with high hydrogen-bonding power that was also strongly hydrophobic. However, the issue of the missing partitions cannot be dismissed so lightly. An arbitrary set of compounds should be spread evenly over the whole of property space as described by an orthogonal set of axes (descriptors), assuming an adequate population size. If there are correlations between descriptors (as there are in our case), the distribution of points will be skewed toward the axis of correlation. A thought-experiment can be constructed consisting of a set of data plotted in XY-space and lying between (0,0) and (10,10), partitioned at intervals of 1 unit. If the X- and Y-values are uncorrelated, then we would expect that all partitions would be equally occupied with mean and standard deviation related to the density of points and the number of partitions. If the values are positively correlated, we would expect a higher occupancy for partitions near to the line X = Y and much less near the regions around (0,10) and (10,0). Interestingly, the empty partitions (which are 6D hypercubes) may well be demonstrating that there are higher-order correlations between the descriptors. The fact that there are more positive correlations than negative ones may also mean that the assumption about orthogonality is less secure.

The question then arises of whether the partitions could be theoretically adjusted to minimize the number of empty partitions. There may be a way of relating the correlation with the probability of occupancy, but that is beyond the scope of this paper. The number of void partitions may be reduced by having an irregular partitioning (small intervals in regions of high data population, large elsewhere); that is what we tried to achieve qualitatively with our equal-area approach. The final DPD set was initially created by selecting one compound from each partition (with an available compound) from each of the three segments of the database (RPR_1, RPR_2, RPR_3); because of the historical differences in the types of compounds in these three segments, it was believed that this would give an added dimension of diversity in the final set. So by selecting three representatives where possible from each partition, the goal of a diverse "representative" set of about 1000 compounds was achieved.

It should be clear from the discussion above, that a DPD set will not contain a fixed number of compounds. Some partitions will be empty, and some will contain only compounds that are out of stock. Conversely, partitions that contain many compounds ought to be represented by more than one compound in the DPD set.

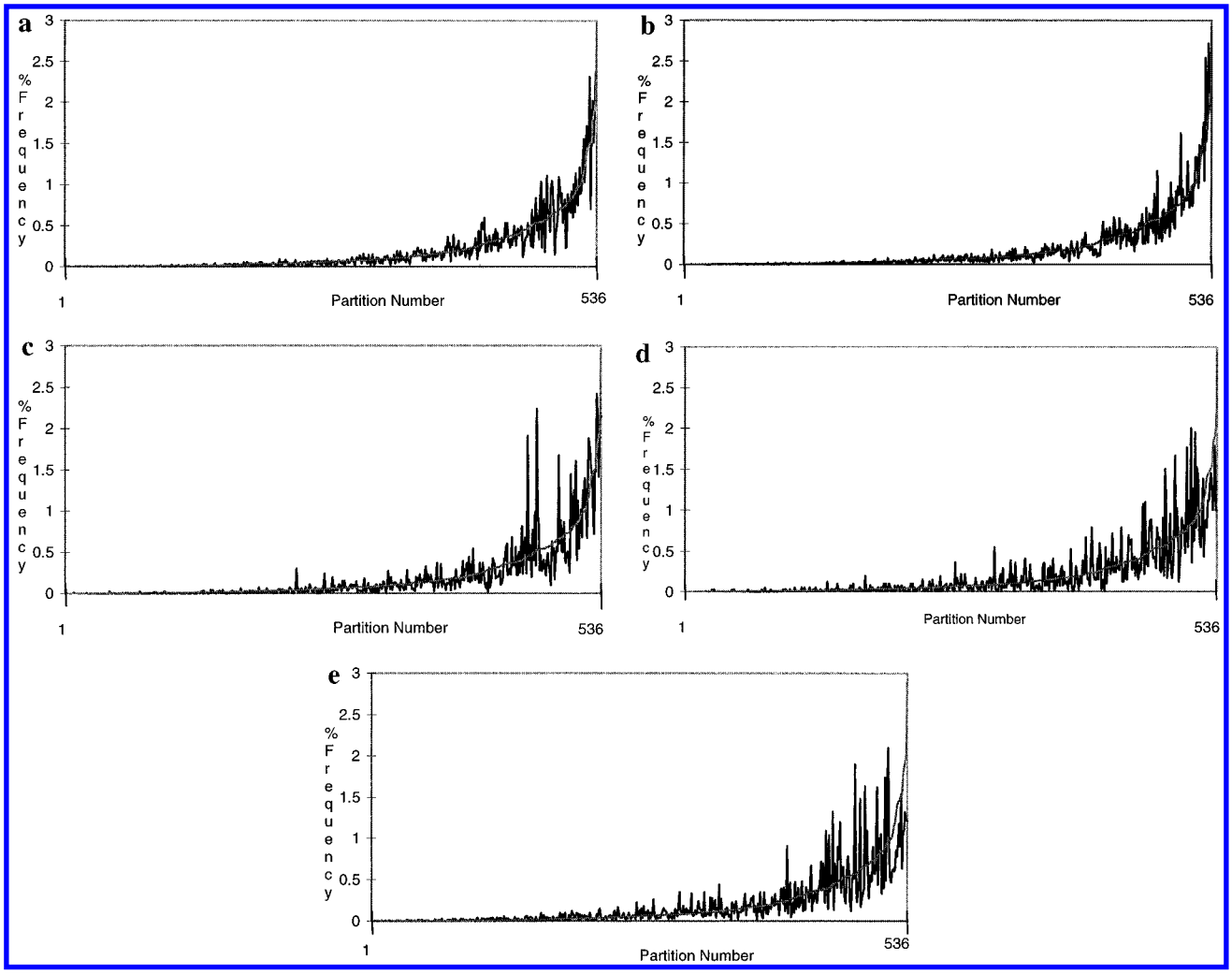


Figure 3. The profiles of each database against the combined databases. The Y-axis gives the percentage of compounds found in each bin, the X-axis is an arbitrary ordering of the bins such that the profile of the combined databases is a simple rising curve. (a) RPR_1; (b) RPR_2; (c) RPR_3; (d) PP; (e) SDF.

Table 4. Running Times (in CPU h:min) and Number of Compounds Present in the Database at Each Stage in the Production of a DPD Set of 576 Partitions

database	filter	compds after filter	*C*log*P*	GENIE	MCX	Master	compds	partition	final compds	empty boxes
RPR_1	6:35	61457	23:58	38:29	4:57	0:02	57178	0:27	45122	46
RPR_2	6:47	57301	13:00	23:38	7:53	0:02	53551	1:35	38978	52
RPR_3	1:52	16392	3:35	6:20	1:31	0:02	15335	0:03	11748	131
PP	2:53	5474	1:26	10:41	5:41	0:01	4851	0:07	4195	182
SDF	3:55	25501	6:23	10:53	5:04	0:01	21665	0:38	18685	60

The filtered RPR_1 database was used to test the usefulness of the partitioning paradigm. In this study, the *C*log*P* class 0 was not included. It was found that 404 out of 432 partitions contained compounds suitable for screening. Of the 404 partitions for which a structure was initially identified, a further 61 partitions could not be represented in the final set, as either all the compounds were rejected by the chemists, or no in-stock compounds could be found, giving a new filled partition number of 343.

1.6. Profiling of Combinatorial Libraries Using the DPD Code. The DPD method of partitioning can also be used to profile a proposed combinatorial library. The normalized distribution of compounds across the DPD partitions is reasonably similar for the five compound databases examined (see Results section). It is relatively simple to construct a database of SMILES codes for the proposed library (we have developed an in-house C program to do this), to compute the molecular descriptors, the DPD codes, and then to partition this virtual database using the methods described above. A reasonable match of the DPD profile of the proposed library against the general DPD profile can then be made one of the design criteria that the library seeks to meet if it is intended for general screening; for focused and biased libraries, the DPD profile can provide warning or confirmation of deviation from the profile of the reference libraries. We would hesitate to recommend that the DPD profile be the sole criterion for design, as we believe that pharmacophoric descriptors are more important. We do think that a library whose compounds fall mainly in partitions that are very sparsely filled by the compound databases in this study should be looked at very carefully before synthesis because it has an unusual profile. It is not necessary for the library to have a very close similarity in terms of its DPD partitioning compared to the reference

databases in this study to be acceptable, providing it falls within the well populated partitions. Combinatorial libraries can be considerably smaller that the reference databases, raising serious objections about inadequate sample size [576 members would be required to even have a chance of occupying all the filled partitions in our study]. The largest common substructure present in all library members may have a biasing effect, as we have observed in analogue series within the reference databases. The profile should be used to highlight possible problems in the design, rather than force conformity.

1.7. Finding Related Compounds Using the DPD Code. Another beneficial effect of the use of partitioning to set up the DPD code is that it is very simple and quick to find all compounds in a database that match a query code. For all but the largest of databases, a straightforward linear search will suffice, but the database could be keyed on the DPD code to make search times faster. As we often search our databases on several different keys, file inversion is not worthwhile. A database of 150K structures can be searched in a few seconds on an SGI Indigo2 R4400. By contrast, a cluster-based code would require the use of a more expensive nearest-neighbor algorithm. The partition code can potentially miss related compounds due to edge effects; in our searching program, we allow the option of widening the search criteria to include neighboring ranges. This allows the user to make decisions about what factors might be most important in designing a follow-up set of compounds after an initial screening hit. If *C*log*P*, for instance, was decided not to be a crucial factor, then the search could be set to include all the *C*log*P* ranges; similarly the search can be broadened to include less and/or more flexible compounds.

2. RESULTS

The MAKE_DPD program was used to partition structures from the RPR_1, RPR_2, and RPR_3 databases. Two external databases of compounds that show biological activity, the Pharma Projects (PP) database and the Standard Drug File (SDF), were also classified as a reference set.[25,26] The standard sets of default filters and partition values were applied. The original work for this study was performed on the local VAX cluster. The running times given (Table 4) should therefore be taken only as a guide, as the machines in the cluster have different specifications. The SGI IRIX version of MAKE_DPD benefits from the faster chip sets of the SGI computers and runs 30−50 times faster. The numbers of compounds that were left after each stage of filtering are also given. The observation that there are relatively large reductions in database size after each level of filtering indicates that some care must be exercised in choosing a random screening set of compounds from a raw database, especially when the database contains compounds intended for agrochemical and pharmaceutical screens. Although the current implementation has been designed for use in a pharmaceutical context, it is simple to change the control files to apply a different set of filters more appropriate to other applications.

2.1. Comparison of the Descriptor Distributions and Correlations from Different Databases. A statistical analysis was performed on each database, to check correlations of descriptors, the distributions of descriptor values, and the profile of partitions. The profile of the partitions examines the relative numbers of compounds in each partition for all the databases. It has been used to check whether the method gives similar classifications on independent databases. Despite their very different origins, the distributions of properties in all databases are broadly similar; the histograms organised by databases are given in Figure 2a−e, and comparative plots organized by category are given in Figure 2f. The observation of the trend leads to the belief that the DPD method is generally applicable.

Table 5. Correlation Matrix for the Partition Profiles of the Various Databases

	RPR_1	RPR_2	RPR_3	PP	SDF
RPR_1	1	0.84	0.78	0.68	0.65
RPR_2		1	0.83	0.77	0.72
RPR_3			1	0.72	0.70
PP				1	0.92
SDF					1

2.2. Comparison of the Partitioning of Different Databases. The profiles of different databases were compared using the percentage normalized occupancy of each partition (100*number of compounds in partition/total number of compounds). To make comparison more simple to visualize, a reference database, made by combining the results of the RPR_1, RPR_2, and RPR_3 databases is used as a reference in each plot and the data sorted into ascending order of occupancy in the reference database (Figure 3a−e). It is clear that the profiles are broadly similar, again confirming the idea that the DPD method can be applied to any diverse chemical database. The local discrepancies in the overall trend can often be traced back to individual project families within a database. For example, in column 475 of the RPR_3 plot (Figure 3c), corresponding to bin 222 213, the RPR_3 database has a much higher occupancy than the reference. This corresponds to a family of compounds made for an antihypertension project, which make up 75% of the members of the bin. In column 477 of the sorted plot, corresponding to bin 213 233, families of compounds made for a leukotriene D4 project (43%) and a leukotriene B4 project (34%), account for the difference. The PP database, which should not contain large analogue families, again shows that the trend of the profile is similar (Figure 3d); such a database is of course only representative of biological targets already exploited and will tend to contain series of related "me-too" compounds. The extra variation may be due to the relatively small size of the PP database, which would tend to highlight the absence or presence of compounds in certain partitions. However, a very similar pattern of variation is seen in the SDF database (Figure 3e), which would tend to reinforce the arguments concerning the presence of analogue series within the data.

Although the graphical profiles are useful for looking at general trends and identifying specific differences, it is not easy to quantify the overall similarity of two databases in this way. A correlation matrix of the number of compounds in each partition is given in the upper half of Table 5. The smallest coefficient is 0.65, and the average value is 0.76, which is encouraging.

2.3. Screening Results. The DPD rational set has been used in biological screens at RPR since the beginning of 1992. It has been routinely used for screens at the Dagenham Research Centre where a number of weak leads have been identified in enzyme-based assays and in whole-cell assays.

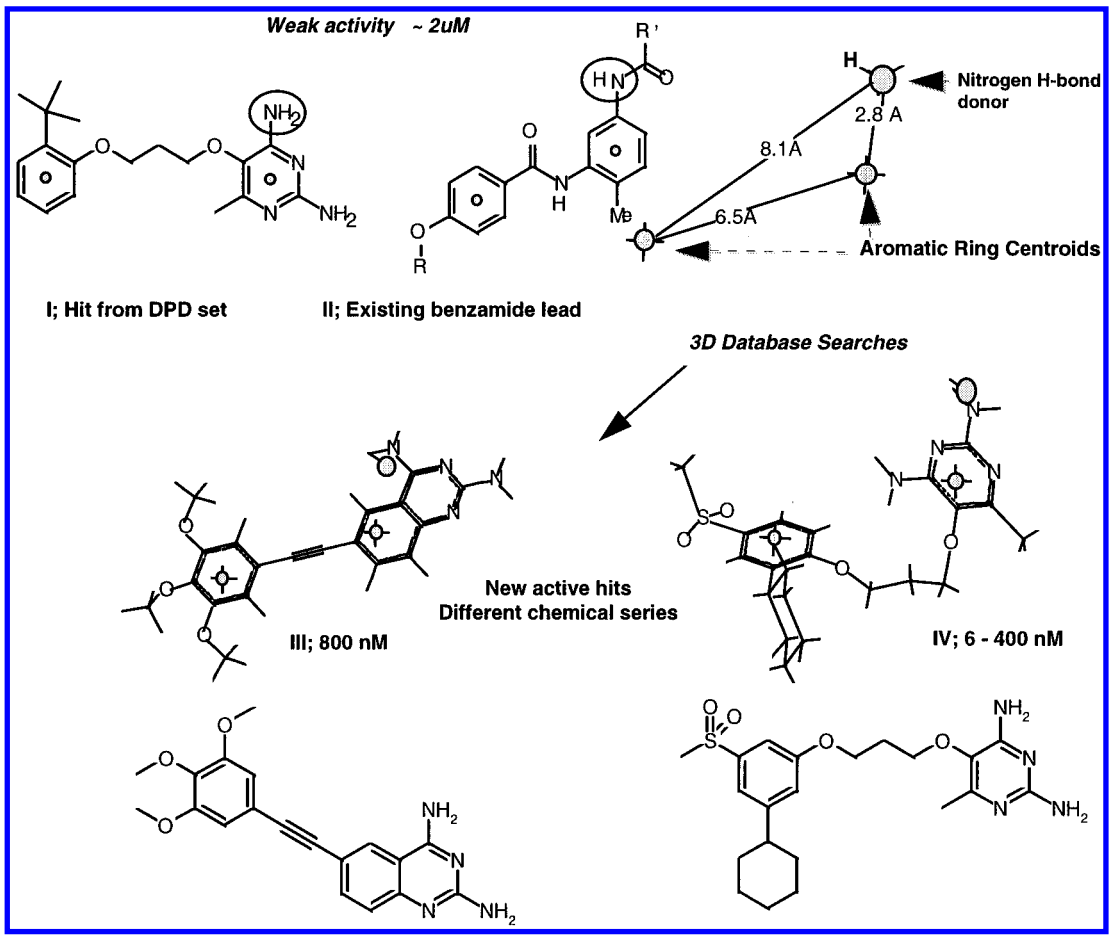


Figure 4. Lead generation for low-density lipoprotein receptor upregulation using hits from a DPD set combined with 3D database searching.

The activities are in the range 1−50 μM. Only the Low Density Lipoprotein (LDL)[27] series will be discussed.

The goal from screening was to identify compounds that reduce blood LDL concentrations. The process of LDL production is thought to be controlled in the cell nucleus, and as the assay was cell-based, transport of the compounds to the potential site of action was an issue. This motivated us to screen the DPD representative screening set, as it was selected on a physicochemical basis. Screening gave one hit, a diaminopyrimidine compound; multiple hits would indicate that some of the descriptors used to partition the database (excluding flexibility) were perhaps not appropriate for the assay. A follow-up screening set of compounds with the same DPD code, to test the hypothesis that the DPD code was relevant to activity, gave further hits which were analogues of the diaminopyrimidine (**1**; illustrated in Figure 4). Low activity (IC_{50} 1.7 μM) compounds based on a dibenzamide structure (**2**) had been identified,[1] but optimization to more potent compounds was elusive. In parallel to this work, a three-point 3D pharmacophore model had been derived, and 3D searches of the corporate databases using this query in ChemDBS-3D had produced screening sets that yielded two new lead series including one of compounds related to the DPD hit. The 3D database query was then further refined using the common features of the different active compounds. From the diaminopyrimidine series, compounds already existing in our corporate registry, were identified with activities of ca. 6 nM (compounds **3** and **4** in Figure 4). This illustrates the point that DPD codes only give initial hits that have physicochemical properties suitable for binding. The advances in activity came only through a 3D pharmacophore model. The final stages of optimization would involve fine-tuning and the production of close analogues, so that a substructure-based similarity measure would be of most use at this stage. DPD representative sets do not give high quality leads, they give hits. However, the hits can be obtained rapidly, and because the sets are representative of the diversity of molecular properties present, rather than of chemical series, can be particularly relevant for new screens (receptor, enzyme, protein−protein or whole cell). The information used in the design of the DPD code can also be used to guide further screening and modeling studies.

3. DISCUSSION

The initial goal of this study was to develop a methodology for rationally selecting subsets of corporate databases, to enhance the efficiency of lead generation over random screening. Any screening strategy should find hit compounds quickly, and furthermore should enable those hit compounds to be turned into a lead series. Design of a general screening set will allow a greater diversity of the whole database to be sampled more rapidly and, given the underlying design assumptions, will allow rational approaches to be adopted for turning any hits into a lead series. In particular, similar compounds to the hit can be rapidly assembled to give a secondary screening set that help establish the validity of the first hit. This is not to say that a general screening set will always give better results than screening at random,[23] but the results can be analyzed in a clear frame of reference, that is, the initial assumptions that went into the design of the set. In the same way, the design of combinatorial

libraries can be enhanced by requiring that the library be (dis)similar to a reference set of compounds. For these reasons, we have studied several methods for computing molecular similarity.

There have been many studies on molecular similarity, as recent reviews bear out.[28] Our line of reasoning has been driven both by theoretical and pragmatic considerations. A commonly used similarity metric is based on the functional groups of paths contained within a molecule. Although this method is very powerful when splitting a database up into chemical families, a receptor or enzyme does not recognize particular functional groups or paths *per se*; rather it interacts with the properties in space projected by these atoms. A more satisfying metric might be one based on matches of molecular properties, for instance, a Carbo index.[29] We ruled this out on purely practical grounds, as the index, even using Gaussian methods,[30] is slow to compute and is dependent on conformation. In addition, there is the complication of how many maxima in the similarity matches between just two molecules should be carried through. Although we are convinced of the usefulness of these method in some contexts, they did not seem to be appropriate for our requirements here. Another method would be to use pharmacophoric similarity: that is the subject of another paper in this series,[7] and at the time this study was started was not a practical proposition for large databases of conformationally flexible molecules. Given the perceived limitations in these other methods, we decided to base our similarity descriptors on overall molecular properties; other groups have also taken a parallel approach but with very different criteria for selecting descriptors.[31] The COUSIN program, developed independently by workers at Upjohn, has grown along very similar lines but with the emphasis on using an experimental design paradigm for selection of the diverse set.[32]

A ligand with high affinity for a receptor site will have a very high degree of complementarity, that is, the molecule will fit snugly into the site and will match the spatial hydrogen-bonding and electrostatic profile of the site. In contrast, one would expect that an initial screening hit may only have micromolar affinity. This implies a qualitative match to the site, that is, the lead is of the right overall polarity/hydrophobicity to get to the site and contains a fragment that can fit into the target site. The issue of transport to the site is important in whole-cell assays. Molecular properties are a reasonable compromise; they represent the ensemble average of the individual conformations of a molecule, they contain some notions of the properties important in ligand-receptor interactions, and they are very quick to compute. The hydrophobicity and polarity measures are properties of the whole molecule and will reflect the general environment of the receptor site. The hydrogen-bond and steric descriptors will probe the specific nature of the site. Of course, they are not without their limitations, in particular the gross averaging of conformation space and the absence of dipole descriptors (to avoid difficulties with coordinate frames of reference). However, we feel that for performing inter- and intralibrary comparisons to obtain rational screening sets or to aid the of design combinatorial libraries, the use of molecular properties as a similarity metric is justifiable.

It has been argued that any molecular diversity measure should be validated by testing to see how effective it is at separating active from inactive molecules in a biological assay, and studies to this effect have been performed.[33] We do not believe that this prescription will give a proper validation of a similarity metric for two reasons: first, the difficulty in setting an appropriate activity cut-off and, second, the absence of any truly representative data sets. If one were to conduct a retrospective analysis of a typical medicinal chemistry study and plot a graph of the biological activity of the molecules in the study against their 2D substructural similarity to a reference (highly active) molecule, one would expect to see a graph similar to that shown in Figure 5. The substructural similarity metric is the one most keenly perceived by organic chemists. The trend of increasing activity with 2D similarity is often observed in medicinal chemistry studies, but it does not mean anything in terms of ligand-receptor interactions. The trend is mainly the understandable tendency of chemists to reinforce success and to make analogues of lead compounds. The better the activity, the smaller the changes as fine-tuning occurs, which is why the 2D measures discriminate well. If the activity cut-off is set at 10^{-8} M, one would expect 2D similarity measures to perform best at separating actives from inactives. Our metric is targeted at much lower levels of activity (10^{-5}–10^{-6} M). A similarity metric for the most interesting range (10^{-6}–10^{-8}) is covered in another paper in this series.[7] Perhaps it would be more appropriate to determine for a particular similarity measure the cut-off level that gave most discrimination. The other difficulty is the absence of data sets that contain consistent biological data against a range of targets for a large number of compounds (not just analogues). Provision of such a data set is a major challenge.

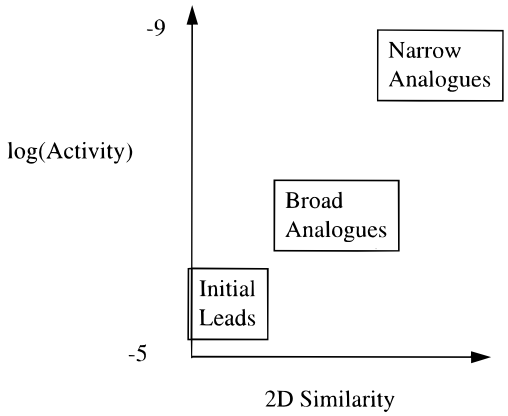


Figure 5. A schematic graph that plots biological activity against 2D similarity to a reference molecule. The correlation of activity with 2D similarity to the most active molecule is often observed in medicinal chemistry projects and is reflected in corporate databases.

The data subjected to statistical analysis were prefiltered to remove undesirable structures; about 25% of the RPR_1 database was rejected. This will obviously bias the statistics. However, we feel that this procedure is justified, as we were trying to look at profiles of databases of potential pharmaceutical drug molecules (leaving aside anticancer and antifungal drugs, where some degree of eukaryotic cytotoxicity/reactivity is necessary for action, and the eicosanoids that would be handled separately). Our target is the formation of a reversible, specific complex between a small molecule ligand and a protein receptor. Filtering is justified for this frame of reference.

3.1. Clustering and Partitioning. Partitioning is not the only algorithm that can be used to divide up a database into classes. The data can also be clustered into families, on the basis of a similarity function. The similarity function measures the distance between two objects, and each cluster is made up of objects that are close to each other. The advantage of clustering is that it does not suffer from edge effects at partition boundaries, whereas partitioning is more portable between independent data sets. We have experience in clustering databases using both hierarchical and parametric methods. The disadvantages of clustering are that the programs take a very long time to run for large data sets, and the criteria for family inclusion/exclusion are somewhat arbitrary. In addition, a similarity function that encompassed all six dimensions would need to have relative weights for the dimensions. An appropriate similarity metric could be obtained through a principal components analysis, but at the possible cost of losing the key dimensions we postulated from our understanding of ligand-receptor interactions. Partitioning seemed to be more appropriate for large chemical databases; this assumption is supported by the similar partition profiles derived from several independent databases. Other workers have favored clustering.[34]

3.2. The Rational Selection of General Screening Sets. The first application of the DPD method was in the rational selection of small general screening sets. It can be argued with justification that a set consisting of only 1% of the complete database cannot be truly representative. On the other hand, it is very difficult, if not impossible, to say what fraction would give a proper sampling. The number of partitions that we set up were dictated by the then needs of our screening unit and has no theoretical basis. The DPD method has been designed so that any number of partitions could be created, with the proviso that consistent partitioning values need to be used if the DPD code is to remain transferable between databases. When analyzing the results of assaying a screening set, the underlying assumptions behind the partitioning strategy should be taken into consideration. In an ideal case, each partition would have a distinct and unique molecular property profile, and only one profile would match the profile of the site. This means that only one lead should be found when a DPD set is scanned. In real life, the target profile may match more than one partition, because a particular descriptor is not relevant to binding; this should be detectable. In addition, the DPD descriptors are not truly orthogonal, so there will be correlation effects. If several hits are found, all apparently unrelated in their DPD code, the assumptions on which the set was selected are probably not valid for the assay.

3.3. Design of Combinatorial Libraries. The DPD method can be used to generate a profile for a proposed combinatorial library. The profile of the library can be compared to arbitrary reference database (for example, the SDF database or a corporate registry) using the bin occupancy measures. In addition, the *C*log*P* and molecular weight descriptors contained within the DPD databases can be plotted as separate frequency histograms and compared to distributions from reference databases. We do not of course advocate that a design for a combinatorial library be rejected because the profiles do not match the reference standards exactly. The DPD profile is however an indication that the design may need to be closely examined to see what is causing the observed differences, although designs focused or biased to a particular target may well have, or indeed need, a different occupancy. Conversely, if one were so minded, the DPD method can easily be inverted to favor libraries that are the complement of a given reference database. The comparison of libraries has been automated into a single C-shell script and the profiling results can be obtained in a few minutes and can be used by medicinal chemists routinely. In our hands, the DPD profile is used at an early stage in the design, to identify combinatorial products with more extreme physicochemical properties. These products can be examined in the training phase to give an interpolative picture of how the other compounds in the initial design will behave. For focussed or optimization libraries, where the reference library may consist of only a few ligands, the DPD method is not particularly useful, other factors like pharmacophore descriptors should be given more weight.

3.4. Validity of Partitioning for Selecting Secondary Screening Sets. The DPD method will quickly find compounds that are related physicochemically, but which may have little substructure similarity. The lead compound is examined using the same procedure as before to determine its DPD code. Once the code has been determined, then other compounds in the same class (and which therefore have similar molecular properties) can be sent for secondary screening, to try to find other potential lead compounds from different chemical families. For a database of 40 000 compounds, we have found on average about 50 follow-up compounds. These molecules should have a much higher probability of showing activity, if the assumptions concerning the importance of the DPD code are valid for the assay concerned, with the added bonus of possibly identifying a new chemical series of leads. This last point is important as a chemically diverse series of leads will assist the production of a pharmacophore model and will provide a choice of synthetic targets to follow up.

However, there is a general weakness inherent in all classification procedures, regarding objects that lie on the periphery of a family. The classification rules will force an object into one family when it may go just as well into another one. In partitioning, objects that lie close to the dividing boundaries may be misclassified. This has implications for follow-up screening of molecules in the same partition as a lead. To avoid missing compounds that just fall outside a partition, the surrounding partitions should also be tested. However, for a 6-dimensional system, with perhaps 50 compounds per partition, this could provide a further potential 36 000 compounds to be screened, which defeats the object of the exercise. For properties such as flexibility it would be reasonable to search routinely for related compounds which differ only in this parameter, particularly for less flexible compounds when the hit is a flexible compound. For broader searches, a nearest-neighbor list could be constructed using the absolute values of the descriptors. We have decided against using this approach, as it requires us to construct a weighting scheme between the descriptors (for the Cartesian calculation of distance), and the scheme may not always be appropriate for every situation. We prefer to allow the medicinal chemist to make explicit assumptions as to the SAR of the system when selecting which descriptor ranges to broaden. However, we do not dismiss the other approach, and each case should be decided on its merits. The DPD method groups together compounds with similar properties, so even though the

representative compound may have only low affinity, related compounds from several different databases are already identified and can be rapidly evaluated.

4. CONCLUSIONS

The generation of new chemical leads for biological targets is a very challenging task, and the design of screening sets and combinatorial libraries have provided useful insights. In this paper, we have described a novel molecular similarity descriptor, the DPD code, that contains information about key molecular and physicochemical properties of the molecule, the application of this descriptor to the selection of a representative screening set, the selection of secondary screening sets to obtain more information concerning the SAR of a particular target receptor, and the profiling of combinatorial libraries. The general applicability of the method has been validated by comparing results from four independent compound libraries. General screening sets derived using the DPD method are in use within Rhone-Poulenc Rorer, and have provided useful hits. The DPD method for measuring molecular similarity offers new capabilities for comparing and profiling libraries and compounds and thus for lead generation and exploitation.

ACKNOWLEDGMENT

We would like to thank our colleagues within the RPR CADD groups for useful discussions and the referees who, through their diligence, greatly improved this work.

REFERENCES AND NOTES

(1) Preliminary communications of this work have appeared in the following: (a) Mason, J. S.; McLay, I. M.; Lewis, R. A. Applications of Computer-Aided Drug Design Techniques to Lead Generation. In *New Perspectives in Drug Design*; Dean, P. M., Jolles, G., Newton, C. G., Eds.; Academic Press: London, 1995; Chapter 12, pp 225−253. (b) 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. (c) Lewis, R. A.; McLay, I. M.; Mason, J. S. Diverse Property-derived Sets: A novel method for selecting representative screening sets using molecular and physicochemical properties. *Chemical Design Automation News* **1995**, *10*, 37−38.

(2) 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.

(3) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering. Manuscript in Preparation.

(4) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(5) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(6) Hansch, C.; Leo, A. In *Substituent Constants for Correlation Analysis in Chemistry and Biology*; Wiley Interscience: New York, pp 18−43.

(7) 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.

(8) *Daylight Software Manual*: *Theory*; Daylight Chemical Information Systems: Santa Fe, NM 87501 (Daylight daymodels software).

(9) Hall, L. H.; Kier, L. B. In *Reviews in Computational Chemistry*; Boyd, D. B., Lipowitz, K. B., Eds.; VCH: 1991; Vol. 2, pp 367−422. Hall, L. H. Molconn-X; Hall Associates Computing: 2 Davis Street, Quincy, MA 02170.

(10) Daylight Chemical Information Systems; Santa Fe, NM 87501.

(11) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd.: UK, 1986.

(12) Hall, L. H.; Mohney, B. K.; Kier, L. B. The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.

(13) The formula for computing molar volume by Schroeder's method is molar volume = 7.0*(*n*C + *n*N + *n*O) + 31.5**n*Br + 24.5**n*Cl + 10.5**n*F + 38.5**n*I + 21.0**n*S + 7.0**n*Double + 14.0**n*Triple + 10.5**n*Aromatic + 7.0**n*AH1 + 14.0**n*AH2 + 21.0**n*AH3 + 28.0**n*AH4 − 7.0*(1 − *n*Atoms + *n*Bonds) where the symbols means number of carbon, nitrogen, oxygen, bromine, fluorine, iodine, or sulfur atoms, the number of double or triple aromatic bonds, the number of atoms bonded to one, two, three, or four hydrogens, the number of atoms and the number of bonds in the molecule, respectively.

(14) Leo, A. Personal communication.

(15) This program is available on request from RAL; it requires Smiles and Smarts toolkit licenses from Daylight Chemical Information Systems.

(16) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(17) Kier, L. B. Use of molecular negentropy to encode structure governing biological activity. *J. Pharm. Sci.* **1980**, 69, 807−810.

(18) RS/1. BBN Software Products Corporation; Cambridge, MA 02238.

(19) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining Structural Similarity of Chemicals using Graph-theoretic Indices. *Discrete Appl. Math.* **1988**, 19, 17.

(20) Downs, G. M.; Willett, P. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1094−1102.

(21) MacFarlane, J. W.; Gans, D. J. Cluster Significance Analysis. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: 1995; pp 295−308.

(22) Lawson R. G.; Jurs P. C. New Index for Clustering Tendency and Its Application to Chemical Problems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 36−41.

(23) 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.

(24) Lajiness, M. Evaluation of the Performance of Dissimilarity Selection Methodology. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo C., Vittoria A., Eds., Escom: 1991; pp 201−204.

(25) Pharma Projects; Pharma Projects Ltd.: 18-20 Hill Rise, Richmond, Surrey, TW10 6UA UK.

(26) Standard Drug File (now known as the World Drug Index). Derwent Information Ltd.: 14 Great Queen Street, London WC2B 5DF UK.

(27) Ashton, M. J.; Brown, T. J.; Fenton, G.; Halley, F.; Harper, M. F.; Lockey, P. M.; Porter, B.; Roach, A. G.; Stuttle, K. A. J.; Vicker, N. New Low-Density Lipoprotein Receptor Upregulators Acting via a Novel Mechanism. *J. Med. Chem.* **1996**, *39*, 3343−3356.

(28) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie Academic and Professional: Glasgow, 1995.

(29) Carbo, R.; Leyda, L.; Arnau, M. An electron density measure of the similarity between two compounds. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.

(30) Good, A. C.; Hodgkin, E. E.; Richards, W. G. The Utilisation of Gaussian Functions for the Rapid Evaluation of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 112−116.

(31) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(32) Lajiness, M. Applications of Molecular Similarity/Dissimilarity in Drug Research. In *Structure-Property Correlations in Drug Research*; van de Waterbeemd, H., Ed.; R. G. Landes: 1996; Chapter 5.

(33) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical similarity using physicochemical property descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.

(34) Shemetulskis, N. E.; Dunbar, J. B., Jr.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemcial database clustering and analysis. *J. Comput-Aided Mol. Design* **1995**, *9*, 407−416.

CI960471Y