Median Partitioning: A Novel Method for the Selection of Representative Subsets from Large Compound Pools

Jeffrey W. Godden,[†,‡] Ling Xue,[†,‡] Douglas B. Kitchen,[†] Florence L. Stahura,[†,‡] E. James Schermerhorn,[†] and Jürgen Bajorath*[,†,‡,§]

Department of Computer-Aided Drug Discovery, Albany Molecular Research, Inc. (AMRI), 21 Corporate Circle, Albany, New York 12212-5098, AMRI Bothell Research Center (AMRI-BRC), 18804 North Creek Parkway, Bothell, Washington 98011, and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received March 16, 2002

A method termed Median Partitioning (MP) has been developed to select diverse sets of molecules from large compound pools. Unlike many other methods for subset selection, the MP approach does not depend on pairwise comparison of molecules and can therefore be applied to very large compound collections. The only time limiting step is the calculation of molecular descriptors for database compounds. MP employs arrays of property descriptors with little correlation to divide large compound pools into partitions from which representative molecules can be selected. In each of n subsequent steps, a population of molecules is divided into subpopulations above and below the median value of a property descriptor until a desired number of 2^n partitions are obtained. For descriptor evaluation and selection, an entropy formulation was embedded in a genetic algorithm. MP has been applied here to generate a subset of the Available Chemicals Directory, and the results have been compared with cell-based partitioning.

INTRODUCTION

Selection of subsets from large compound pools such as combinatorial libraries, inventories, or collections from vendor catalogs is an important topic in molecular diversity analysis, for example, when developing compound acquisition strategies.[1,2] Major efforts in diversity analysis include subset selection and diversity design.[3] By definition, subset selection starts from given compound data sets and is in essence a deductive approach, whereas the design of diverse libraries is more inductive in nature. Various methods have been introduced to facilitate the selection of representative or diverse subsets from compound collections. Prominent among those are clustering techniques,[4,5] especially hierarchical clustering,[6] stochastic methods combining different diversity functions and search algorithms,[7] and dissimilarity-based methods,[3,8] which include, among others, different versions of the popular MaxMin algorithm.[9,10] Like molecular fingerprint-based approaches in diversity selection,[11,12] these techniques essentially rely on pairwise comparisons of property distances between compounds. In principle, diversity functions that rely on molecular comparisons display quadratic dependence on the number of compounds in the data set. In consequence, the underlying combinatorial problem substantially increases with the size of both databases and subsets and becomes computationally infeasible if the data sets are very large.

Different types of dissimilarity-based methods with modulated complexity have been developed.[3] For example, the complexity of maximum dissimilarity selection methods is on the order of $O(kn)$ to $O(k^2n)$, with k being the size of the subset and n the size of the original collection. More efficient techniques for diversity analysis such as the centroid-based diversity sorting algorithm[13] have been introduced where complexity only scales with the size of the original data set and for which further improvements in calculation speed have recently been proposed.[14] In addition, other algorithms have been designed that rely on probability sampling rather than complete enumeration of pairwise distances[15] and thereby largely circumvent the combinatorial problem.

Cell-based methods to partition compound data sets do not depend on distance or nearest neighbor calculations and thus represent a different approach for compound classification and selection.[16−18] In this case, positions of molecules in low-dimensional property spaces are calculated, and the cells into which compounds fall are identified. Cells are subdivisions of chemical space obtained by application of binning schemes.[19] Similar to the situation in cluster analysis,[4] representative compounds can then be selected from each computed cell. Since partitioning does not require calculation of pairwise property distances, the complexity of the methods is lower than in the case of clustering or maximum dissimilarity methods, on the order of $O(n)$, similar to centroid-based diversity sorting, as mentioned above. It follows that cell-based methods should, in principle, be amenable to the analysis of much larger compound pools than methods depending on pairwise comparisons. However, cell-based approaches generally require a dimension reduction of chemical descriptor space,[17,18] which can be accomplished, for example, by principal component analysis (PCA).[20] At least in this case, however, increasing size of the original compound pool is again becoming an issue, due

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com.
† Albany Molecular Research, Inc. (AMRI).
‡ AMRI Bothell Research Center (AMRI-BRC).
§ University of Washington.

10.1021/ci0203693 CCC: $22.00
Published on Web 06/12/2002

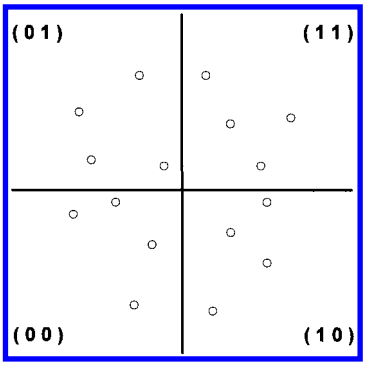

Figure 1. Median partitioning. The schematic representation illustrates the basic idea of the MP approach. As an example, a two-dimensional chemical space is shown. The axes represent the medians of two descriptors that divide the compound set into equal or at least nearly equal subpopulations. Each partition is characterized by a unique binary code.

to the increasing complexity of eigenvalue and eigenvector calculations when computing principal components.[20] Not all partitioning methods are cell-based. For example, recursive partitioning,[21,22] which is mostly applied for hit or lead identification, generates subsets along decision trees.

We were interested in developing an efficient and conceptually straightforward method to facilitate the selection of diverse subsets. Specifically, we aimed to design an approach that does not depend on pairwise comparison of compounds and that can be applied to very large pools of, ultimately, millions of molecules. To achieve this end, we have developed and implemented median partitioning, a statistical approach, whereby a population of molecules is divided in subsequent steps into subpopulations above and below the median values of property descriptors. In these calculations, n descriptors produce a total of 2^n distinct partitions having unique binary signatures. Herein we report the design of the MP method and results of initial applications.

METHODS

Median Partitioning Concept. In statistics, the median is defined as the value within a distribution that divides the population into two equal subpopulations (above and below the median value).[23] If a property descriptor is calculated for a database of molecules, its median value can be determined, and, following a binary classification scheme, molecules having a descriptor value above the median are assigned a "1" and those having a value below the median a "0". As illustrated in Figure 1, the process can be repeated for different descriptors in subsequent steps, whereby n descriptors yield 2^n unique partitions. Each partition is then characterized by a binary signature pattern consisting of n bits. This concept can be applied to divide a compound set of a given size into a desired number of partitions. For example, the use of 10 descriptors in subsequent partitioning steps will create 2^{10} or 1024 partitions into which the source compounds fall. To obtain a representative subset of the original data set, given this classification scheme, compounds can be selected from each populated partition. Thus, from very large compound pools, subsets of desired size can be selected by adjusting the number of (descriptors applied and) partitions created. While the basic idea of MP is quite simple, descriptor-based MP analysis has a number of specific requirements, as discussed in the following. The algorithms required to facilitate MP analysis were implemented using Perl.

Relevant Descriptors. Given the MP concept, property descriptors must be selected that are suitable to calculate reasonable median values. What are important requirements? First of all, descriptors should produce "broad" value distributions, as illustrated in Figure 2. In other words, they should capture a significant amount of information so that median values become meaningful population characteristics. Thus, following our implementation of the Shannon (SE) entropy concept[24] for descriptor analysis,[25,26] we have chosen as a selection criterion for MP descriptors detectable and, if possible, significant information content.[26]

Shannon entropy is defined as

$$SE = -\sum p_i \log_2 p_i$$

In this formulation, p is the sample probability of a data point to fall as a count c within a specific data range i, and p is obtained as

$$p_i = c_i / \sum c_i$$

The logarithm to the base two is a scale factor which makes it possible to consider SE as a metric of information content. It can be rationalized as a binary detector of counts (i.e., does the count appear in a given data interval?). Histograms provide a convenient way to establish the bit framework for data representation (here descriptor value distributions). The major advantage of this concept is that the information content of descriptors having very different distributions and value ranges can be compared. Since SE values calculated from histograms are bin number-dependent, descriptor variability may vary from zero for a single valued descriptor to a maximum of the logarithm to the base two of the number of chosen histogram bins. Therefore, it is useful to establish a bin-independent SE value, called scaled SE, which can be directly compared, regardless of the number of histogram bins. A scaled SE value is obtained by dividing an observed SE value by the maximum possible SE value for the number of bins used:

$$sSE = SE/\log_2 (bins)$$

Based on the analysis of value distributions of many molecular descriptors in large compound collections, we have previously established[26] generally applicable threshold values for low (<0.30), medium ($0.30-0.60$), and high scaled SE (>0.6). From an original pool of 143 1D and 2D molecular property descriptors,[26] we excluded here all descriptors having single values (and thus no information content) in the compound collections under investigation, yielding a total of 111 descriptors for further study. Among these descriptors, scaled SE values ranged from 0.02 to 0.90. In addition, selected descriptors should display as little correlation as possible, as explained in the following.

Descriptor Correlation Analysis. Why should descriptors selected for MP analysis be uncorrelated? The reason is illustrated in Figure 3. When using correlated descriptors for MP, the data distributions are skewed along the diagonal of correlation. As a consequence, the four partitions shown in Figure 3 no longer contain the ideal amount, $^1/_4$ of the total

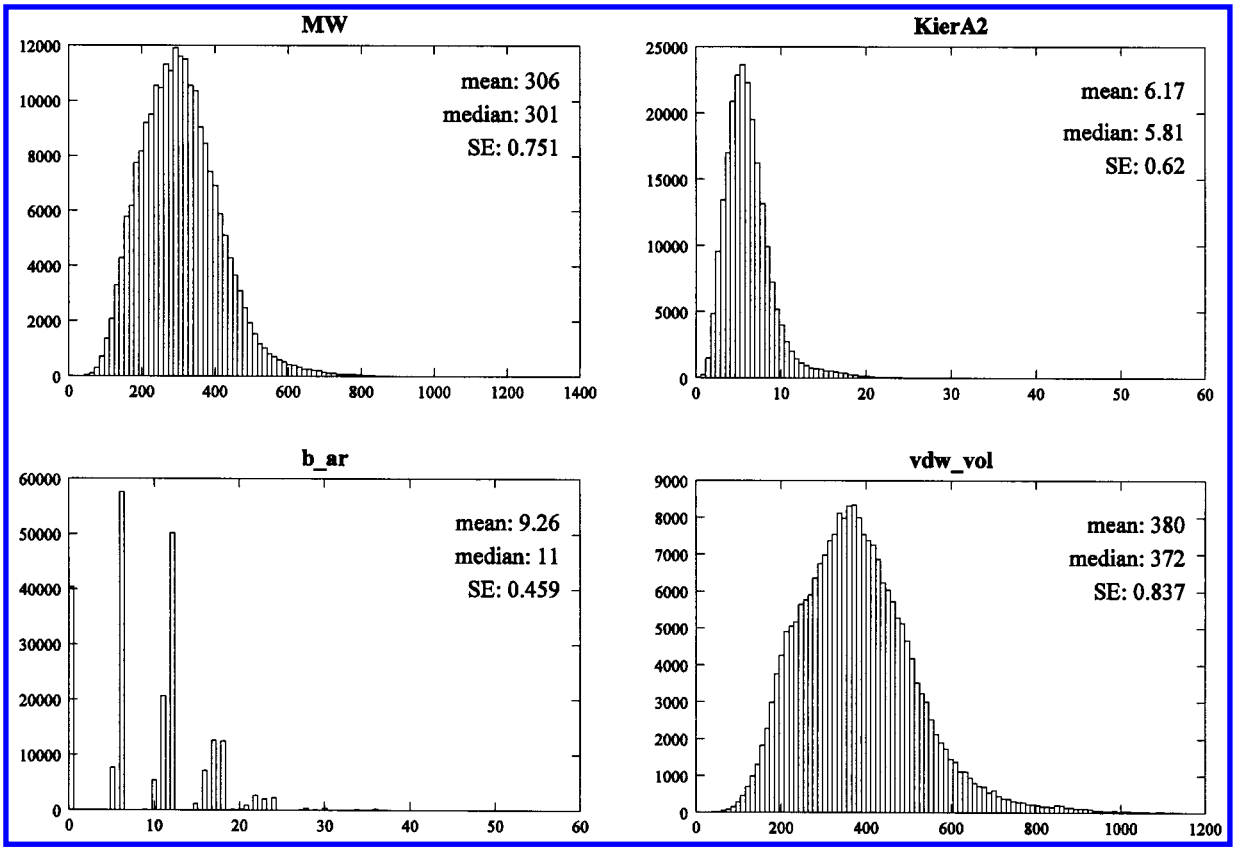

Figure 2. Descriptor distributions. Shown are distributions of four molecular descriptors ("mw", molecular weight; "b_ar", number of aromatic bonds; "KierA2", a Kier & Hall index;[37] "vdw_vol"; van der Waals volume) calculated for a total of 229 529 ACD compounds. These distributions provide examples of information-rich descriptors that are generally favored for MP analysis. Descriptor distributions are monitored in histograms consistently having 100 bins, and mean, median, and scaled SE values[26] are reported for each descriptor.

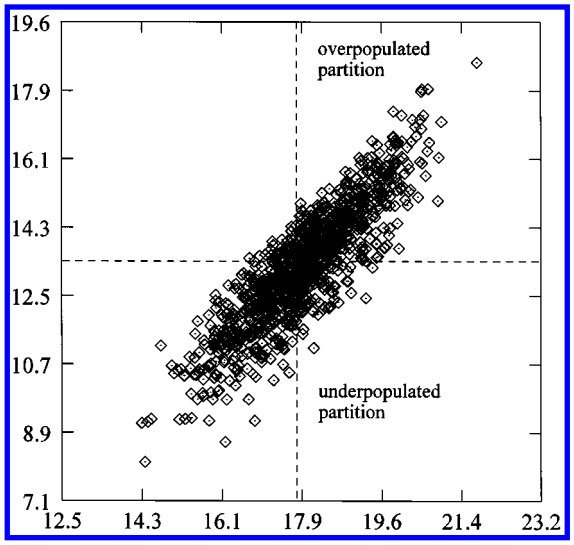

Figure 3. Descriptor correlation. The diagram shows the value distributions of two strongly correlated topological descriptors and illustrates the effects of descriptor correlation discussed in the text.

population of compounds, but rather fewer or more molecules. Simply put, the use of (even only marginally) correlated descriptors for MP will produce both empty and overpopulated partitions. To identify information-rich descriptors with little correlation, all n-by-n descriptor correlation coefficients were calculated for a set of 111 molecular property descriptors and retained for further study. As to be expected,[18] the analysis revealed that it was not possible to identify combinations of completely uncorrelated chemical descriptors, at least within the descriptor pool used for this study. Thus, as further described below, we attempted to optimize descriptor combinations and minimize correlation effects as much as possible.

Databases and Descriptor Calculations. Descriptor values were calculated with the Molecular Operating Environment (MOE)[27] for the Available Chemicals Directory (ACD)[28] and also for an in-house virtual library of ~2.5 million compounds. Scaled SE values were calculated from histograms as described previously.[25,26] To remove "exotic" compounds that would distort the descriptor values distributions, median absolute deviations[23] were calculated, defined as $\mathrm{Mad} = |x-M|/D$, where "x" stands for each descriptor value in a population, "M" is the median value of the population, and "D" is the median of $|x-M|$. Mad values essentially correspond to standard deviations but do not depend on the presence of normal data distributions. Compounds were omitted from the source database, if their Mad values were greater than nine for at least 10 of the selected descriptors. This stringent protocol was applied to remove only those compounds whose presence would skew distributions to a degree that the compound would be separated from all others. Examples of some of such "outlier" compounds are shown in Figure 4.

Descriptor Analysis. To identify preferred combinations of descriptors, a genetic algorithm (GA)[29] was implemented, as summarized in Figure 5. The chromosomes consisted of 111 bits each of which represented one of the descriptors and, if set on, added this descriptor to the calculation. The scoring function $S = \langle SE \rangle / \langle CC \rangle$ ("CC" means correlation

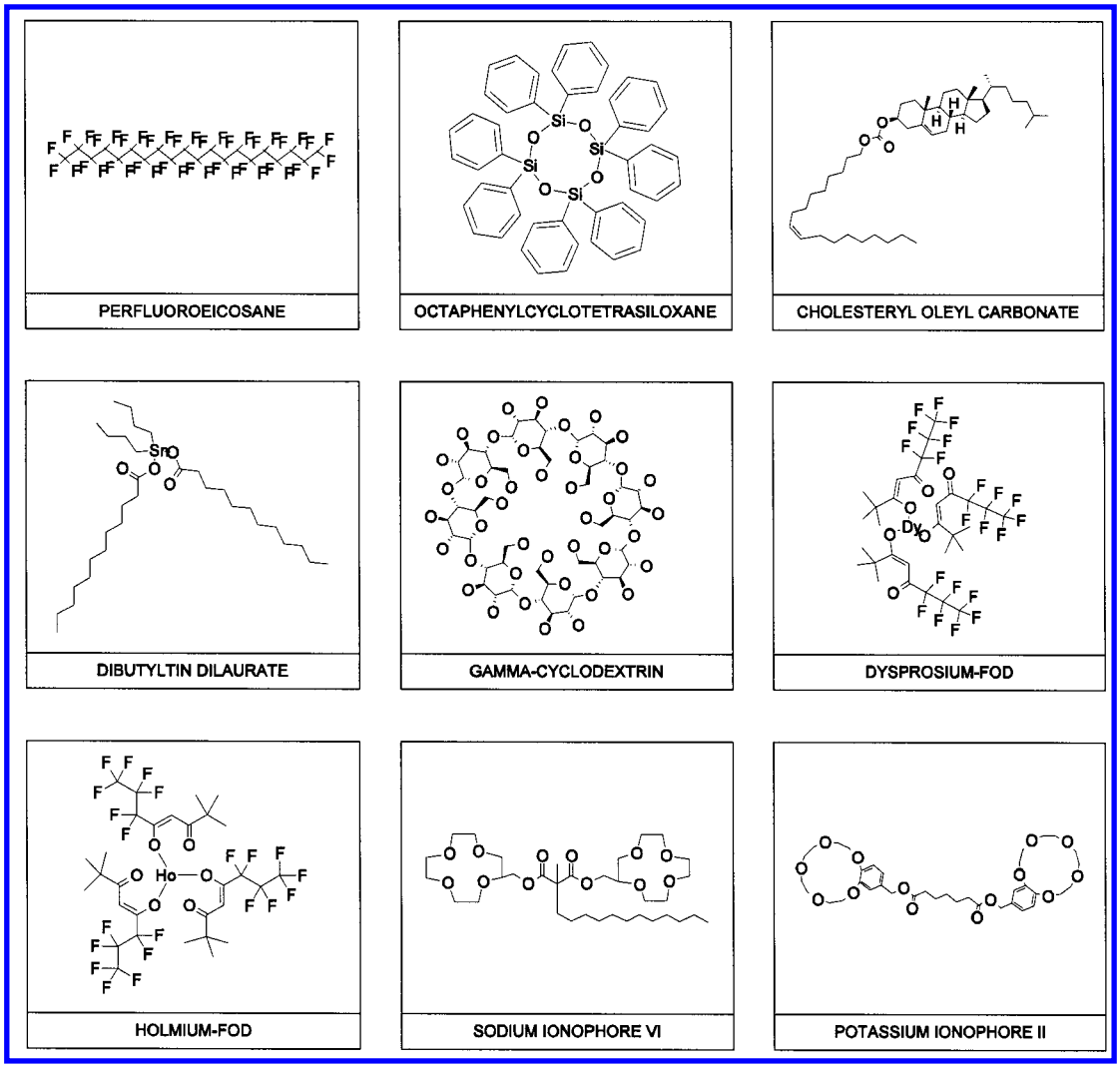

Figure 4. Excluded compounds. Examples of compounds are shown that were omitted from MP analysis based on calculation of median absolute deviations.

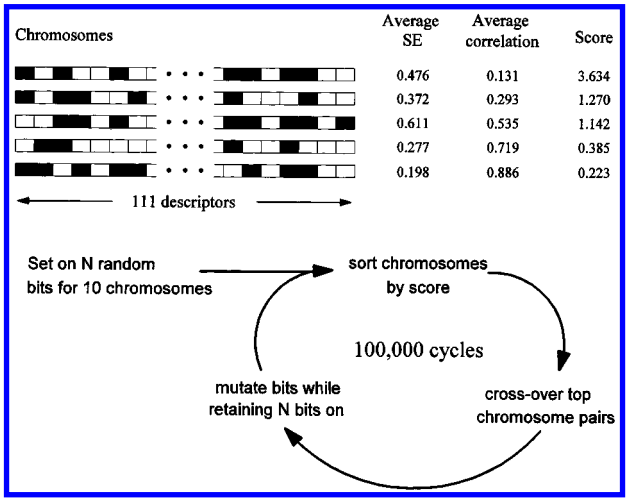

Figure 5. Genetic algorithm. The figure describes the GA technique used for descriptor set optimization. At the top, model chromosomes are shown encoding specific descriptor combinations and their scores. The diagram at the bottom summarizes the GA calculations including the crossover and mutation operations.

coefficient) was used to maximize average scaled SE values of descriptor combinations and minimize their average correlation coefficient. At each cycle, the crossover operation was applied to the top two chromosome pairs, the resulting chromosomes were mutated at a rate of 25%, and the calculations proceeded for 100 000 GA cycles.

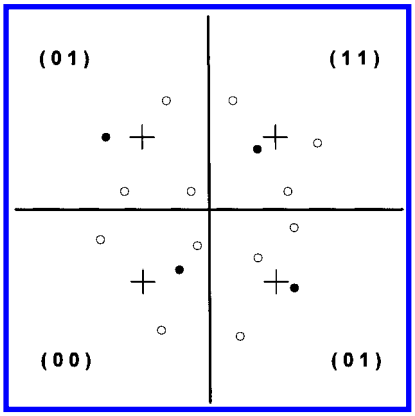

Figure 6. Compound selection. This schematic, an extension of Figure 1, illustrates how compounds are chosen from central positions in partitions. Crosses represent the calculated quartile positions, and compounds nearest to the quartiles (filled circles) are selected to represent each partition.

Compound Selection. To produce a representative subset, molecules were selected from partitions based on closest scaled Euclidian distance[23] from the quartile, as illustrated

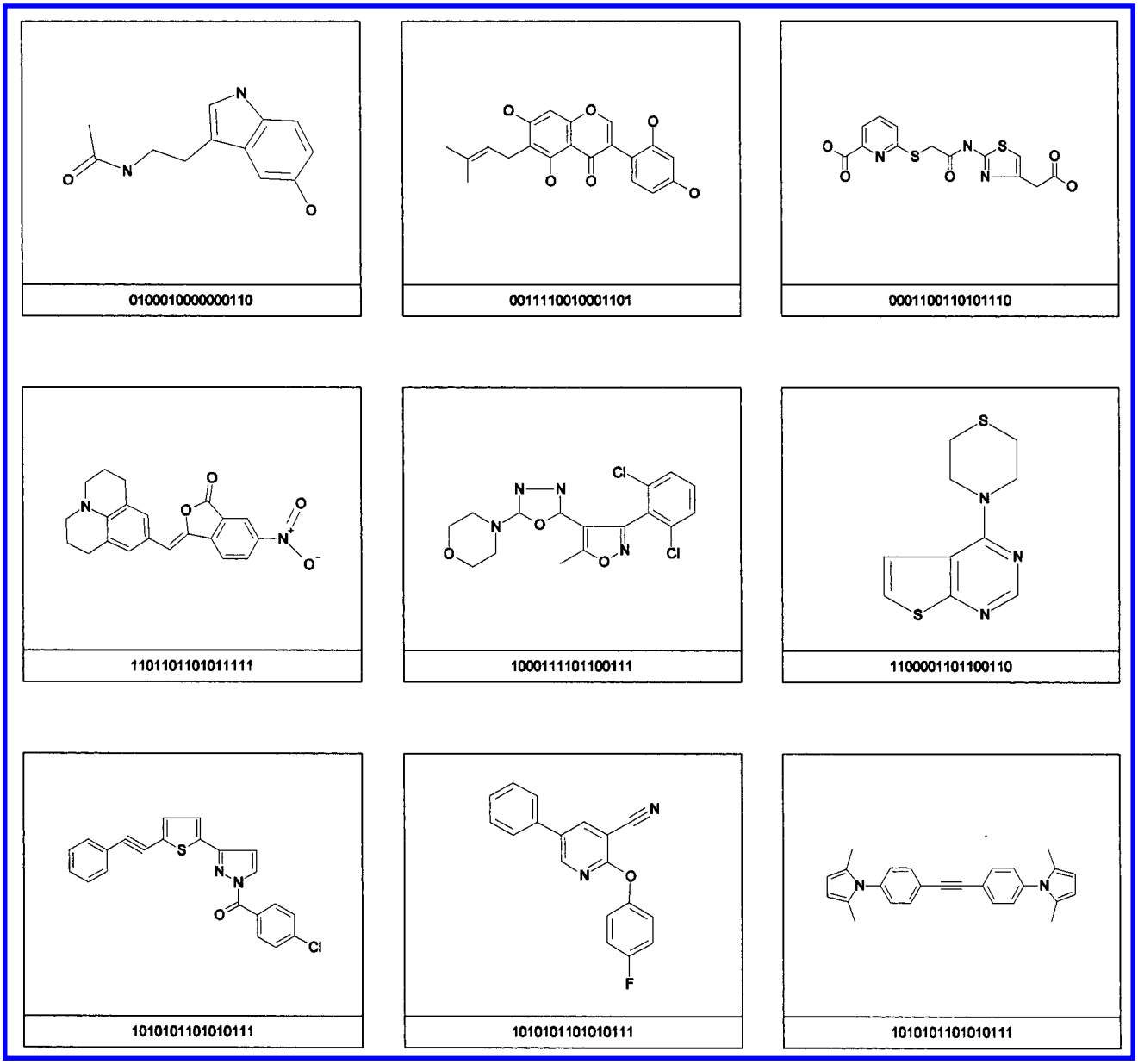

Figure 7. Partitioned compounds. Shown are examples of compounds from ACD-MP partitions and their binary signatures including singletons (top), compounds from multiply populated partitions (middle), and molecules from the most populated partition (bottom).

in Figure 6. The quartile is defined as the median value of every descriptor coordinate within each partition. Euclidian distances were scaled by dividing the distance by the range of each descriptor value. The procedure essentially selects compounds from the center of each partition, thus avoiding boundary effects. In addition to quartile selections from each multiply populated partition, all singletons (i.e., partitions containing only one compound) were included in the subset.

Cell-Based Partitioning. The diversity of the MP-selected ACD compound subset (ACD-MP) was compared to a subset of similar size produced by PCA-based partitioning[28] (ACD-PCA) using the same descriptor set. This subset was obtained by PCA, calculated with MOE, including the first five principal components and applying eight bins per PC axis. This effectively created a five-dimensional space from the contributions of 16 descriptors and adjusted the total number of cells for partitioning. In these calculations, the first five nonscaled principal components accounted for greater than 90% of the ACD variance with respect to the chosen descriptors. ACD-PCA was obtained by selecting all singletons and a randomly chosen compound from each multiply populated cell.

Diversity Calculations. The degree of diversity in ACD-MP and ACD-PCA was evaluated using Euclidian distance calculations as well as atom pair descriptor-based diversity[30] and molecular scaffold distribution[31] analysis. Molecular scaffolds were isolated from compounds following a hierarchical molecular description, whereby substituents were systematically removed from these molecules until ring-containing core structures were obtained.[31] Furthermore, these two subsets were graphically compared in three-dimensional PCA space.

RESULTS AND DISCUSSION

Descriptor Selection. For MP analysis of the ACD, we arbitrarily focused on the selection of 16 descriptors, yielding a total of 2^{16} or 65 536 possible partitions. Thus, values of all 111 descriptors in our original pool were calculated for

Table 1. Descriptors Selected for ACD-MP[a]

descriptor	scaled SE	definition
Fcharge	0.17	sum of formal charges
PEOE_RPC-	0.84	relative negative partial charge[38]
PEOE_VSA_FNEG	0.86	fractional negative vdw surface area[32,38]
PEOE_VSA_POL	0.48	total polar vdw surface area[32,38]
a_aro	0.48	number of aromatic atoms
a_don	0.28	number of H-bond donor atoms
a_nP	0.02	number of phosphorus atoms
a_nS	0.17	number of sulfur atoms
b_rotR	0.84	fraction of rotatable bonds
b_triple	0.06	number of triple bonds
density	0.56	mass density
logP(o/w)	0.49	log octanol/water partition coefficient
vsa_acc	0.47	vdw acceptor surface area
vsa_acid	0.13	vdw acidic surface area
vsa_don	0.21	vdw donor surface area
weinerPol	0.61	Weiner polarity number[39]

[a] Average scaled SE: 0.42. Average absolute value of correlation coefficient: 0.14. "vdw" stands for van der Waals.

all ACD compounds, and different combinations were analyzed for their information content and degree of correlation. Ideally, descriptors selected for MP should not only be rich in chemical information but also uncorrelated, as explained in the methods section. Among the 2D property descriptors evaluated here, not a single combination of 10 or more uncorrelated descriptors could be identified by GA calculations. The net effect of descriptor correlation on MP is that partitions remain empty or become variably populated. While there is no stringent a priori requirement in MP subset selection to avoid empty or overpopulated partitions, we attempted to minimize the effects of descriptor correlation. The most favorable (i.e., information-rich and least correlated) descriptor combination identified in our GA calculations is reported in Table 1. The selected descriptors included various charge terms and approximate van der Waals surface area descriptors[32] as well as atom or bond counts and some bulk properties. This descriptor combination had an average SE value of 0.42 and an average absolute value of the pairwise correlation coefficient of 0.14.

ACD Subset Calculation. Initially, salts and noncovalent complexes were removed from the ACD, yielding a total of 231 187 compounds. This compound set was subjected to Mad calculations using the 111 descriptors to remove unusual compounds, as described in the methods section. These calculations further reduced the number of compounds to 225 929 molecules that served as the pool for MP analysis. Of the 65 536 theoretically possible partitions, MP produced a total of 8103 populated partitions, thus yielding an occupancy rate of 12.4%. This illustrates the cumulative effects of descriptor correlations, even if they are relatively small. The obtained ACD partitions were variably populated and included 1191 singletons. The largest partition contained a total of 1918 molecules. Examples of compounds in differently populated partitions are shown in Figure 7. Filtering of ACD-MP revealed that 16% of the selected compounds had undesired reactive groups,[33] 79% had between one and seven desired pharmacophore groups,[34] and 87% followed Lipinski's rules.[35] These relatively favorable characteristics were in part due to the fact that several thousand unusual compounds were removed from the ACD by Mad analysis prior to partitioning.

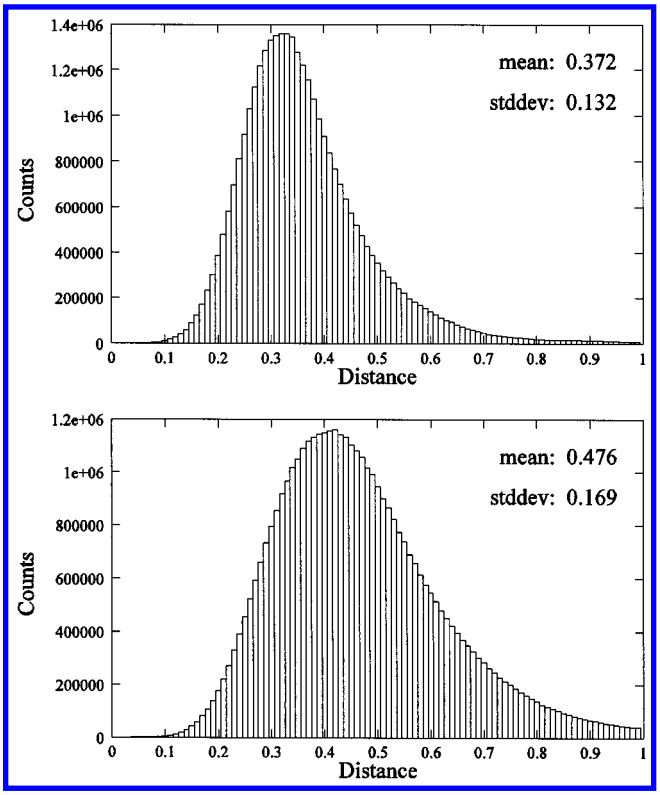

Figure 8. Diversity analysis: Euclidian distances. The histograms display the distributions of Euclidian distances in ACD-MP (top) and ACP-PCA (bottom), which contains more compounds. Mean values and their standard deviations (stddev) are reported.

Diversity Assessment. It is well established that the evaluation of molecular diversity depends on the definition of chemical spaces and diversity metrics[36] and that it is difficult, if at all possible, to "objectively" compare diversity distributions calculated using different methods. Nevertheless, as a reference, we have also generated a subset of the ACD by PCA-based partitioning starting from the same descriptor set. ACD-PCA consisted of 9256 compounds, as compared to the 8103 molecules in ACD-MP. In principle, PCA partitioning should be capable of creating a greater degree of diversity than MP because PCA generates its own reference frame from preselected molecular descriptors and the PCA axes can have a varying number of bins. By contrast, MP is based on a binary classification scheme relative to each descriptor without the ability to linearly combine different descriptor contributions. What are the differences and how significant are they? Figure 8 shows a comparison of the Euclidian distance distributions in ACD-MP and ACD-PCA. Both subsets display overall very similar Gaussian-like distributions. The mean Euclidian distance is slightly larger for ACD-PCA (0.48) than for ACD-MP (0.37). Figure 9 shows the results of atom pair descriptor-based diversity analysis. The average pairwise similarity for ACD-MP is 19% and average nearest neighbor similarity 67%, as compared to 17% and 66% for ACD-PCA, respectively. Thus, at the level of these calculations, the differences are insignificant. When assessed with our previously reported algorithm,[31] ACD compounds used for partitioning contained a total number of 43 979 unique scaffolds including 27 170 singletons. Here differences appear to be more significant because ACD-PCA captures ~12% of these scaffolds, whereas MP selects only ~9%. However, these differences

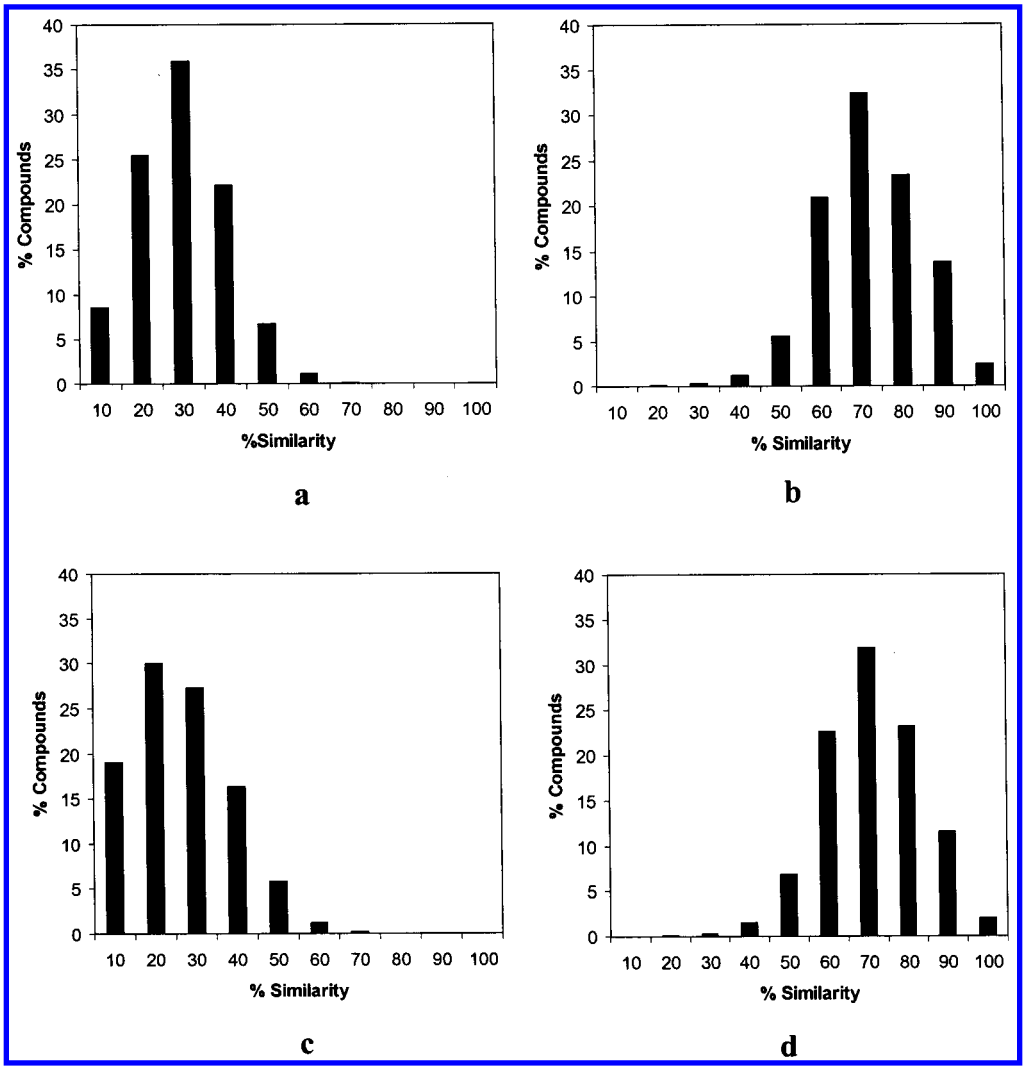

Figure 9. Diversity analysis: Atom pair descriptors. The results of atom pair descriptor calculations are summarized. Panels (a) and (c) report the distributions of average atom pair descriptor similarity for ACD-MP and ACD-PCA, respectively. Panels (b) and (d) show the distributions of nearest neighbor atom pair similarity for ACD-MP and ACD-PCA, respectively.

resulted from the fact that PCA-based partitioning selected ~1000 singletons that are not found in ACD-MP. In addition, ACD-PCA was found to contain 536 nonring compounds, whereas ACD-MP contained only 177. These observations suggest that PCA-based partitioning has a greater tendency than MP to create singletons and select compounds at the "edges" of diversity space, as we would expect. For example, whereas 87% of ACD-MP compounds follow the Lipinski rules, as stated above, only 67% of ACD-PCA compounds do so. Finally, we have also compared ACD-PCA and ACD-MP in PCA space with ACD serving as a reference frame. Here our original ACD compound pool was subjected to PCA using the 16-descriptor set, and the first three resulting principal components were selected as the coordinate system for visualization of the subsets. The comparison is shown in Figure 10. As can be seen, the graphical distributions of ACD-PCA and ACD-MP are very similar, and any qualitative difference in the achieved level of diversity between these subsets is not obvious. Thus, in summary, we do not detect what we would consider to be significant differences when comparing the overall diversity of ACD-PCA and ACD-MP.

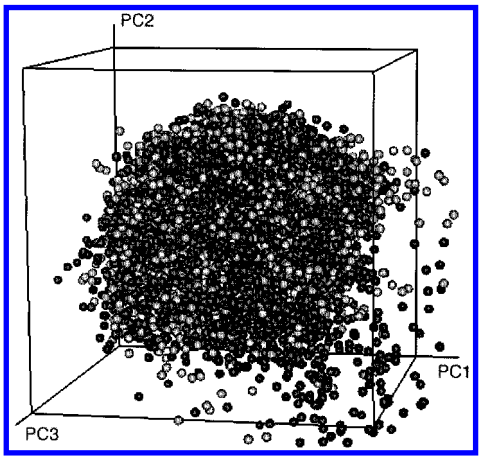

Figure 10. Comparison in PC space. Shown is a comparison of ACD-MP (light gray) and ACD-PCA (dark gray) compounds in reference space defined by the first three principal components calculated for the ACD. The figure was generated with MOE.[27]

Library Design. As another application of the MP method, we also selected a screening library for acquisition and internal use from a pool of approximately 2.5 million compounds collected from catalogs of various chemistry vendors. In this case, our target library size was about 100 000 compounds, and we selected a total of 19 descriptors for partitioning. This descriptor set had an average absolute

value of the correlation coefficient of 0.13. In these calculations, a partition occupancy rate of 21% was achieved and a total of 110 039 compounds was selected. In this more medicinal chemistry-oriented library, only 2% of the compounds contained undesired reactive groups, 92% had between one and seven desired pharmacophore groups, and 83% were within the Lipinski rule-of-5. Selection of this library from a large source revealed the computational efficiency and potential of MP. Excluding initial calculations of descriptor values for the compound pools, which had already been completed for other purposes,[26] median value statistics, partitioning, and code assignments only required approximately 2 h on a 600 MHz PC processor.

CONCLUSIONS

The primary purpose of developing the MP approach has been to facilitate the selection of subsets that does not depend on pairwise comparisons of molecules and is thus applicable to very large compound pools. What are the approaches most similar to MP? Among currently available partitioning methods, recursive partitioning,[21,22] another nonparametric statistical method, also divides compound databases in descriptor or property space. In contrast to MP, however, recursive partitioning divides a data set into statistically distinct subsets along decision trees until a minimal subset containing objects with desired properties is obtained.[22] In chemoinformatics, recursive partitioning is typically based on the analysis of learning sets to correlate descriptor values with compound properties, most often biological activity, and is thus conceptually distinct from MP. The major attraction of MP is its ability to efficiently generate subsets of targeted size from very large compound pools. Although we would expect MP to be a less efficient "diversity selector" than, for example PCA-based partitioning, differences revealed at the level of test calculations carried out thus far do not appear to be very significant. In fact, results obtained by MP and PCA are overall comparable. PCA has the advantage that preselected descriptor sets will be decorrelated when calculating principal components. This, however, makes the complexity of these calculations dependent on the size of the original data set, and the decorrelation per se has no immediate chemical meaning. On the other hand, quartile selection renders MP much less prone to boundary effects than PCA classification. MP calculations can employ many types of molecular descriptors but benefit from minimizing descriptor correlation effects. However, the occupancy rates of partitions are easily monitored, and different numbers of compounds can be selected from variably populated partitions to mirror the composition of source data sets. When combined with molecular scaffold analysis, for example, compound selection can be biased toward certain classes of compounds, if so desired. In addition, the MP scheme provides a convenient basis for the extension of subsets, for example, by screening additional compound sets for molecules to populate empty partitions. This is thought to be particularly relevant for increasing the size of compound libraries selected from external sources.

ACKNOWLEDGMENT

The authors thank Harold Meckler of AMRI for his help in the assembly of database compounds from external sources.

REFERENCES AND NOTES

(1) Shemetulskis, N. E.; Dunbar, J. B. Jr.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput-Aided Mol. Des.* **1995**, *9*, 407−416.
(2) Rhodes, N.; Willett, P.; Dunbar, J. B., Jr.; Humblet, C. Bit-string methods for selective compound acquisition. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 210−214.
(3) Willett, P. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. *J. Comput. Biol.* **1999**, *6*, 447−457.
(4) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(5) 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.
(6) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.
(7) Agrafiotis, D. K. Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.
(8) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model.* **1997**, *15*, 372−285.
(9) Higgs, R. E.; Bemis, K. G.; Watson, I. A.; Wikel, J. H. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 861−870.
(10) Clark, R. D. OptiSim: An extended dissimilarity selection method for finding diverse representative subsets. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181−1188.
(11) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.
(12) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. A dual-fingerprint based metric for the design of focused compound libraries and analogues. *J. Mol. Model.* **2001**, *7*, 125−131.
(13) 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.
(14) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivaschenko, A. A. New diversity calculations algorithms used for compound selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249−258.
(15) Agrafiotis, D. K. A constant time algorithm for estimating the diversity of large chemical libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 159−167.
(16) 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.
(17) Pearlman, R. S.; Smith, K. M. Novel software tools for chemical diversity. *Perspect. Drug Discov. Design* **1998**, *9*, 339−353.
(18) Xue, L.; Bajorath, J. Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 801−809.
(19) Bayley, M. J.; Willett, P. Binning schemes for partition-based compound selection. *J. Mol. Graph. Model.* **1999**, *17*, 10−18.
(20) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.
(21) Friedman, J. A. Recursive partitioning decision rules for nonparametric classification. *IEEE Trans. Comput.* **1977**, *26*, 404−408.
(22) Rusinko, A. III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(23) Meier, P. C.; Zünd, R. E. *Statistical methods in analytical chemistry;* John Wiley & Sons: New York, 2000.
(24) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, 1963.
(25) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796−800.
(26) Godden, J. W.; Bajorath, J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87−93.
(27) MOE (Molecular Operating Environment), version 2001.01, Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
(28) Available Chemicals Directory, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.
(29) Forrest, S. Genetic algorithms − Principles of natural selection applied to computation. *Science* **1993**, *261*, 872−878.

(30) Carhart, R. E.; Smith, D. H.; Vankataraghavan, R. Atom pairs as molecular features in structure−activity studies: Definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(31) Xue, L.; Bajorath, J. Distribution of molecular scaffolds and R-groups isolated from large compound databases. *J. Mol. Model.* **1999**, *5*, 97−102.
(32) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model*. **2000**, 18, 464−477.
(33) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic pooling of compounds for high-throughput screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.
(34) Muegge, I.; Heald, S. L.; Britelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(35) Lipinski, C. A.; Lombardo, F.; Dominy, B. W., Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **1997**, *23*, 3−25.
(36) Martin, Y. C. Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **2001**, *3*, 1−20.
(37) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. *Rev. Comput. Chem.* **1991**, *2*, 367−422.
(38) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(39) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

CI0203693