PERSPECTIVE

Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening

Jürgen Bajorath[†]

New Chemical Entities, Inc., 18804 North Creek Parkway, Suite 100, Bothell, Washington 98011, and
Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received October 7, 2000

INTRODUCTION

Compound classification and virtual screening methods are capable of exploring and exploiting molecular similarity beyond chemistry, in accordance with the similar property principle.[1] They can be used to analyze and predict biologically active compounds and correlate structural features and chemical properties of molecules with specific activities. This explains why such approaches are highly attractive tools in pharmaceutical research,[2] although a number of the underlying scientific concepts have originally been developed for different purposes. Since it is increasingly recognized that simply synthesizing and screening more and more compounds does not necessarily provide a sufficiently large number of high-quality leads and, ultimately, clinical candidates, much effort is spent in developing and implementing computational concepts that help to identify and refine leads. Typical applications include the identification of compounds with desired activity by database searching, derivation of predictive models of activity for database mining, selection of representative subsets from large compound libraries, or analysis of druglike properties.

The aim of this contribution is to review and comment on some major developments in compound classification and molecular similarity research, reflect their diversity, and highlight some of the questions that remain unanswered. In a single contribution, it is difficult, if not impossible, to provide a complete account of, and give full credit to, all methods and developments relevant to compound classification and virtual screening. Therefore, some areas have been, rather subjectively, more emphasized than others or even omitted. For example, the discussion of virtual screening approaches is limited to those that focus on the small molecular level, as opposed to target structure-based design or docking methods, which have been reviewed elsewhere.[3−5] Fingerprint-based approaches to virtual screening are a major focal point of this report. Nevertheless, recent advances in other areas of three-dimensional (3D) similarity and database searching and statistical approaches are also discussed, albeit to a lesser extent. To illustrate the discussion with practical examples, some data produced in our laboratory are reported that address the performance of selected two-dimensional (2D) descriptors in compound classification and similarity searching. For example, a "simulated" virtual screening experiment is described that indicates the level of "success" that may be expected in similarity searching using relatively simple metrics. In addition, a few comparisons of 2D and 3D fingerprints in detecting remote or "difficult" similarity relationships are presented.

MOLECULAR DESCRIPTORS AND CHEMICAL SPACES

A prerequisite for most approaches to compound classification and library design or analysis is the definition of theoretical "chemical space". Similar to quatitative structure−activity relationship (QSAR) investigations, this typically requires the use of descriptors that capture a broad range of molecular characteristics.[6,7] Such molecular descriptors may have very different complexity but can often be classified according to their "dimensionality", referring to the molecular representations from which they are calculated.[7] According to this scheme, one-dimensional (1D) descriptors include bulk properties and physiochemical parameters (e.g., log P(o/w), molecular weight); 2D descriptors, for example, defined structural fragments or connectivity indices; and 3D descriptors solvent-accessible surface area, molecular volumes, or spatial pharmacophores. Hundreds of molecular descriptors are available in the literature, and a critical task is how to best select descriptors beyond chemical intuition. For example, are there sets of descriptors that are generally superior to others or, alternatively, is descriptor performance largely dependent on the particular problem under investigation? A number of studies, as further discussed below, are beginning to address these and related questions. In addition to more conventional molecular descriptors complex pharmacophore ensembles and fingerprint-type representations of molecular structure, conformations, and properties have been developed specifically for drug discovery applications.

The definition of chemical space using molecular descriptors is conceptually simple, as illustrated in Figure 1. Each of n selected descriptors adds a dimension to n-dimensional chemical space, and each molecule under investigation is assigned coordinates in this space based on calculation of its descriptor values. Thus, in simple terms, decreasing distance in chemical space should correlate with increasing similarity between molecules. For many applications, it is desirable to reduce the dimensionality of chemical space to take only the most important descriptors and their contribu-

[†] To whom correspondence should be addressed at NCE, Inc. Telephone: (425) 424-7297. Fax: (425) 424-7299. E-mail: jbajorath@nce-mail.com.

10.1021/ci0001482 CCC: $20.00 © 2001 American Chemical Society
Published on Web 01/30/2001

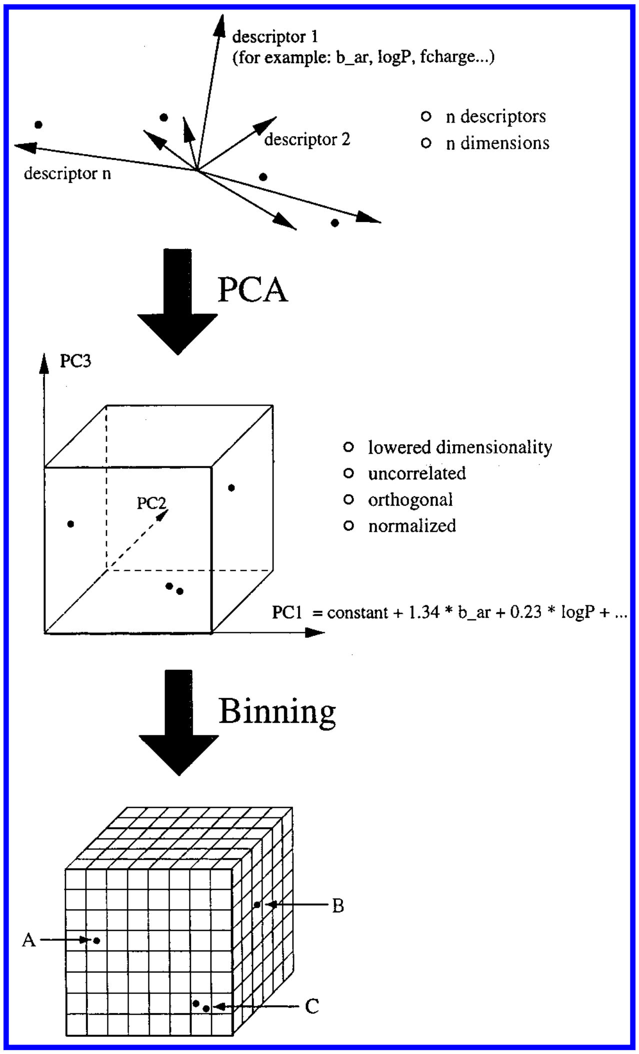


Figure 1. Definition and reduction of chemical space. The figure schematically illustrates how an *n*-dimensional chemical space is defined by use of *n* molecular descriptors. The positions of four compounds, as determined by their descriptor values, are shown as black dots. Principal component analysis (PCA) provides a way to reduce the dimensionality of the space and remove descriptor correlations. Three-dimensional chemical space defined by the first three principal components (PC1−3) is shown, each of which is a linear combination of the original descriptors. Application of binning schemes to the axes of this low-dimensional space produces cells, thereby providing a basis for compound classification. Three areas are labeled (A, B, C). "A" points to a cell that is occupied by a single compound (singleton). "B" indicates the position of a compound located between two cells. Such boundary effects represent potential problems in cell-based partitioning. "C" indicates a cell populated with two "similar" compounds.

tions into account and reduce redundancies and correlations between descriptors (for example, those accounting for "aromatic character" and "chemical saturation"). As also illustrated in Figure 1, one way to do so is to subject descriptor combinations and sets of molecules to principal component analysis (PCA),[8] which produces an uncorrelated and lower dimensional space on the basis of calculated principal components. Each principal component is a linear combination of the original descriptors, and the coefficients indicate the relative importance of these descriptors accounting for the variance of the data. Other approaches are also available to generate low-dimensional chemical space. Among these, the introduction of so-called "BCUT" descriptors in the context of the "receptor-relevant subspace" concept[9,10] has been one of the most interesting methods for applications in QSAR, compound classification, and library design. These complex descriptors are derived from combinations of preselected 2D and 3D descriptors, thought to be particularly relevant for the treatment of receptor−ligand interactions (e.g., accounting for partial charge or hydrogen bond potential), by eliminating features that are correlated or statistically insignificant.

BASIC CLASSIFICATION CONCEPTS

The majority of compound classification approaches are based on clustering[11] or partitioning[12] methods. Clustering of compounds in chemical space, however defined, typically involves the calculation of intermolecular distances, and compounds that are "close" to each other are combined into clusters. In partitioning, on the other hand, chemical space is subdivided into sections, based on ranges of descriptor values, and compounds that fall into the same section are combined. For compound partitioning, it is critical how chemical space is divided into cells, and this process depends on the way descriptor value ranges are binned.[13] As shown in Figure 1, binning produces "cells" in chemical space, and the analysis of how these subspaces are populated with compounds is a common theme of cell-based partitioning methods.[10,12] Such approaches benefit from the ability to generate low-dimensional chemistry space.

Cluster algorithms can be divided into two major categories, hierarchical and nonhierarchical methods. The distinguishing factor is whether initially obtained clusters are combined or divided to obtain the final result (hierarchical) or clustered in a single step, without any relation or hierarchy between clusters (nonhierarchical). Methods that have become particularly popular for clustering of chemical structures include Ward's clustering,[14] a hierarchical method, and Jarvis−Patrick clustering,[15] a nonhierarchical method.

A major goal of many compound classification studies has been, and continues to be, to select representative subsets of large libraries, for example, those that mirror their overall diversity. Another attractive application is the selection of active compounds or the separation of active and inactive molecules. In the latter cases, the calculations attempt to produce clusters or cells that are enriched with molecules having desired activity or that contain only molecules with a specific activity, while minimizing the number of classes that mix compounds with different activities and the number of singletons (i.e., clusters or cells containing only one compound). Since the choice of calculation parameters and descriptors influences the number, size, and composition of clusters or cells, many investigations aim to identify combinations of algorithms and calculation conditions that optimally separate compounds in benchmark databases.

SIMILARITY SEARCHING

On the small molecular level, virtual screening calculations usually start from one or more molecules having some desired activity and aim to identify "similar" molecules with, for example, increased potency or better synthetic accessibility. In its simplest form, similarity searching is based on the detection of 2D substructures or fragments that are shared by molecules.[16] Two-dimensional similarity searches that go beyond the detection of common substructures mostly rely on capturing molecular descriptors in the form of binary bit string representations or "fingerprints". Such fingerprints consist of varying numbers of bits, dependent on the type and number of descriptors and the values they capture. In simple representations, each bit accounts for the presence or absence of one specific molecular feature, but designs may be increasingly complex. For example, a fingerprint representation may capture all possible connectivity pathways through a molecule up to specified length, as in the Daylight fingerprint,[17] one of the pioneering developments in this area. Fingerprints that detect the presence or absence of defined structural fragments in molecules can be easily generated using the binary bit string format. Prominent among these fragment-based tools are MDL or MACCS structural keys,[18,19] the most widely used version of which consists of 166 structural fragments. For comparison of fingerprint representations, a variety of metrics and similarity coefficients have been adapted.[20] The most widely used metric to quantify fingerprint overlap (and thus "molecular similarity") is the Tanimoto coefficient,[20] defined as $Tc = bc/(b1 + b2 - bc)$, with b1 being the number of bits set on in the first molecule, b2 the number of bits set on in the second molecule, and bc the number of bits that both molecules have in common.

In addition to 2D similarity methods, different concepts have been developed to search for similar molecules in three dimensions. In principle, 3D structures of molecules could be directly used to search databases for similar molecules, on the basis of however defined similarity scores or functions. However, most contemporary methods either start from specific geometric patterns of functional groups in molecules that are known or thought to be critical for a specific activity or, alternatively, attempt to survey and compare possible spatial arrangements of such groups in different molecules. This spectrum of approaches ranges from single 3D search queries,[20-22] steric field descriptors,[6] and pharmacophores[22-24] to 3D- or 4D-QSAR models[25,26] and pharmacophore fingerprints.[27-30] In its simplest form, a 3D query may consist of a single spatial pattern in a molecule, for example, a pair of specified atoms and its interatomic distance, calculated through space.[21] In its most complex form, it may consist of millions of potential pharmacophore arrangements the presence or absence of which can be recorded in a fingerprint.[29,30] For pharmacophore searching, the introduction of four-point pharmacophores has advanced the conventional three-point pharmacophore approach. When four-point pharmacophore arrangements are calculated, the number of potential pharmacophores increases dramatically. For example, for six features (e.g., hydrogen bond donor or acceptor, aromatic ring, hydrophobic moiety, negatively charged group, etc.) and 10 distance ranges, a three-point method produces potentially 33 000 pharmacophores, whereas 13 million four-point pharmacophores would be obtained.[24] Some encouraging results have been reported that support the development of four-point pharmacophores. For example, on the basis of four-point pharmacophores, it was possible to differentiate between different types of protease inhibitors and fibrinogen receptor antagonists, which was difficult to do using three-point techniques.[24,27]

PERFORMANCE OF CLASSIFICATION METHODS

The performance of clustering, partitioning, and statistical approaches to identify active compounds, distinguish between activity classes, or select representative subsets has been investigated in a number of studies over the past few years. An important point is that evaluation of classification concepts is much influenced by the choice of molecular descriptors and their overall performances. Thus, it is not possible to completely separate these issues. In addition, the majority of studies are difficult to compare directly because selected test cases, methods, and descriptors differ, at least to some extent, and it often remains difficult to draw firm conclusions concerning the relative performance of different classification methods. Nevertheless, considerable progress has been made in this area.

Early studies by Willett and colleagues demonstrated that Jarvis−Patrick clustering was the most effective nonhierarchical method to cluster molecules according to their chemical and biological characteristics.[31] Jarvis−Patrick clustering was used in conjunction with Daylight fingerprints to analyze the diversity of compound libraries,[32] and the method has been adapted to handle large data sets.[33] Furthermore, a fuzzy clustering modification has been introduced to address some of the shortcomings of the Jarvis−Patrick method,[34] for example, to better control the size of obtained clusters. The availability of the Daylight clustering tools[35] has significantly aided in many of these studies.

Brown and Martin have analyzed the performance of different clustering methods and various 2D and 3D descriptors for classification of active and inactive compounds.[36,37] They found that hierarchical clustering methods and 2D descriptors were in general preferred and that the combination of Ward's clustering[14] and MACCS structural keys[18] performed best. Subsequent studies support superior performance of hierarchical over nonhierarchical clustering techniques for similar applications.[38,39] On the other hand, statistical methods including linear discriminant analysis[40] and recursive partitioning[41] have recently been shown to perform better than hierarchical clustering in compound classification.[42]

Cell-based partitioning of compounds in lower dimensional chemical spaces has become a focal point in compound classification. Popular methods include the BCUT and PCA metrics mentioned before as well as the diverse property-derived (DPD) approach.[43] Similar to the BCUT metric, DPD defines chemical space by use of six pairwise uncorrelated descriptors (hydrogen bond donors and acceptors, molecular flexibility index, aromatic density, clogP, electrotopological index), which were selected on the basis of statistical analysis and intuition. BCUT descriptors have successfully been used in QSAR, cluster analysis, and diversity design.[44-46] In combination with partial least squares discriminant analysis BCUT descriptors accurately classified inhibitors targeting the cofactor binding sites in five different tyrosine and serine/threonine kinases.[47] Compound partitioning based on PCA correctly classified compounds belonging to seven biological activity classes.[48-50] High and prediction accuracy was achieved by systematic evaluation of combinations of large numbers of single (1D and 2D) molecular descriptors, including structural keys.[50] Table 1 lists descriptor combinations that yield high prediction accuracy in PCA-based partitioning. These data demonstrate that relatively few and simple (2D) molecular descriptors, if carefully selected, can be sufficient to effectively classify compounds according to

Table 1. Effective Combinations of 2D Descriptors for Compound Classification Based on Principal Component Analysis[a]

preferred descriptor combinations	N_p	N_m	N_s	*A* (%)
Study C				
a_nI, b_triple, f_conh2, 25, 47, 53, 61, 62, 65, 81, 87, 122, 124, 158, 159	438	0	17	96.3
38, 59, 62, 69, 80, 96, 112, 127, 139, 145, 156	436	7	12	95.8
a_aro, 47, 62, 71, 112, 134	435	3	17	95.6
Study B				
a_aro, b_aro, b_double, b_triple, a_nP, f_so2nh2, LP	415	13	27	91.2
a_aro, f_c=o, LP, vsa_pol	414	0	41	91.0
a_aro, a_nO, LP	404	17	36	88.8
Study A				
b_aro, SS, HB-a	341	80	34	74.9
b_1rotR, chi1, PEOE_PC+, SS, Kier1, Kier2, Kier3	327	76	52	71.9
b_1rotR, b_ar, SS, HB-a	341	76	38	74.9

[a] Compounds belonging to seven activity classes (benzodiazepines, carbonic anhydrase inhibitors, cyclooxygenase-2 inhibitors, HIV protease inhibitors, H3 antagonists, serotonin receptor antagonists, tyrosine kinase inhibitors) were partitioned. Data are in part taken from subsequent studies (A−C) carried out in our laboratory.[48−50] "N_p", "N_m", and "N_s" are the number of compounds in pure cells/classes (i.e., only consisting of compounds belonging to the same activity class), mixed classes (consisting of compounds having different activities), and singletons, respectively. "*A*" is the overall prediction accuracy, defined as N_p divided by the total number of compounds in the test database. The number of systematically studied single descriptors was increased from 17 to 208 in the course of these studies. In A and B, different sets of structural keys were combined and treated as complex descriptors (abbreviated SS). In C, 166 MACCS keys were treated as single descriptors. The analysis in B and C was facilitated by implementation of a genetic algorithm. The following abbreviations are used for descriptors: "a_nI", "a_nP", and "a_nO", number of iodine, phosphorus, and oxygen atoms in a molecule, respectively; "b_double", number of double bonds; "b_triple", number of triple bonds; "a_aro" and "b_aro", number of aromatic atoms and bonds, respectively; "f_conh2", "f_so2nh2", and "f_c=o", number of amide, sulfonamide, and carbonyl groups in a molecule, respectively; "HB-a", number of hydrogen bond acceptors; "LP", number of the lone pair electrons on oxygen and nitrogen atoms (capturing many hydrogen bond acceptors); "chi1", "Kier1", "Kier2", and "Kier3", different connectivity and shape indices;[92] "b_1rotR", fraction of single nonring bonds in a molecule. Numbers in C provide MACCS key numbers for single structural keys.

biological activity. This observation also relates to the BCUT metric, as it relies on contributions of only a few uncorrelated descriptors. Furthermore, the data in Table 1 illustrate that different descriptor combinations can have similarly high predictive power. Since these descriptors were identified by PCA, descriptors within each combination are largely uncorrelated. Thus, in some cases, key features in molecules that determine their structure−activity characteristics may be captured in similar ways by different types of descriptors, for example, those representing structural fragments or molecular properties.[51]

EVALUATION OF MOLECULAR DESCRIPTORS

Since so many different molecular descriptors are available, longstanding questions have been how to identify widely applicable descriptors, if they exist, or how to select the "best" descriptors for specific applications. Moreover, there is considerable debate in the literature whether 2D or 3D descriptors are a priori superior. The predictive power of structural key-type descriptors initially observed by Brown and Martin has been further established for a variety of classification techniques.[42,50,52] Structural key-type descriptors are likely to implicitly account for a number of physicochemical parameters, in addition to atom types and connectivity, which may explain their high performance in discriminating molecular characteristics. In compound partitioning based on PCA, the performance of selected structural keys could be further enhanced by the addition of very few numerical 2D descriptors.[50] Other studies have also emphasized the value of 2D descriptors. Two-dimensional UNITY fingerprints[53] performed better than their 3D counterparts and molecular field descriptors in selecting representative subsets of biologically active molecules.[54] These 2D fingerprints were also superior to 3D pharmacophore triplets in hierarchical clustering of biological compound classes.[55] On the other hand, Patterson et al. found that molecular field descriptors performed as well as some 2D fingerprints and better than others in the design of diverse compound libraries.[56] Similarly, linear discriminant analysis in combination with BCUT descriptors (that combine 2D and 3D information) produced better results in classifying kinase inhibitors than clustering with Daylight fingerprints.[47] In addition, 3D pharmacophore descriptors were found to be highly selective (four-point pharmacophores performed better than three-point pharmacophores).[47] In fact, combining information provided by 2D and 3D descriptors, for example, 2D fingerprints and 3D fields,[57] may lead to the best performance in molecular similarity analysis[57] or compound classification.[55]

Given the fact that molecules are active in three dimensions, 2D descriptors can be surprisingly powerful, as revealed by several comparisons. This suggests that molecular features critical for activity can often be deduced from 2D representations or molecular graphs and do not require explicitly taking conformational parameters into account. However, results vary dependent on the particular application. Rather than deciding a priori for one or the other dimensionality, a key question appears to be how to best select and combine (2D and/or 3D) descriptors that capture highly complementary information, a principle that is already inherent in the BCUT and PCA concepts.

A major difficulty in evaluating molecular property descriptors is that many of these descriptors have very different units and value ranges. Thus, their distributions and variability in compound databases are difficult to compare. Only recently, an entropy-based approach has been introduced to compare the intrinsic and extrinsic variability of different descriptors, regardless of their units and value ranges.[58] The method is based on the Shannon entropy concept,[59] originally applied in digital communication theory, and calculates descriptor entropy values from histogram representations. Shannon entropy is defined as SE $= -\sum p_i$

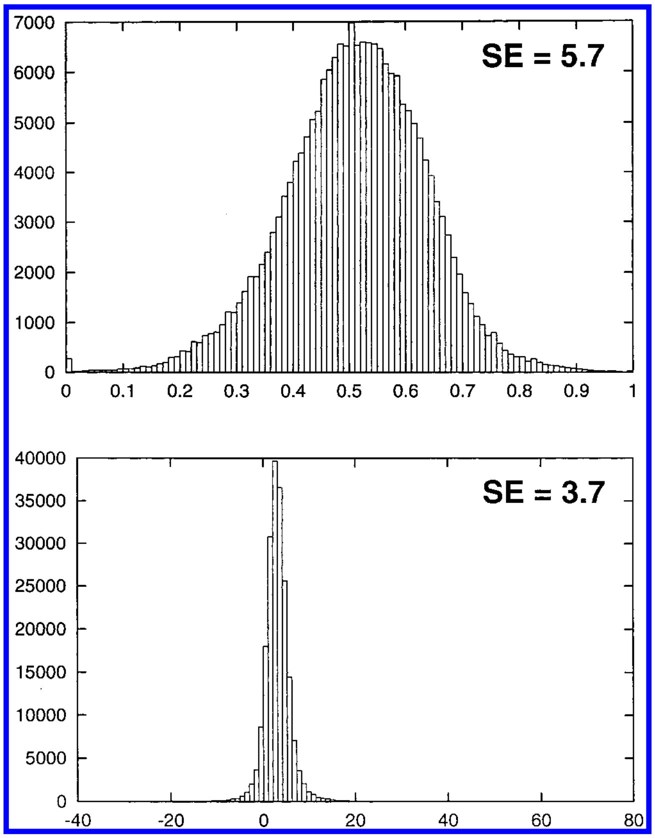


Figure 2. Descriptor distributions for the calculation of Shannon entropy. For two different descriptors, the distribution of values in a large compound database is calculated using uniform data representation. This means that the horizontal axis of each histogram is divided into exactly 100 bins covering the entire data range for each descriptor. The vertical axis reports the number of occurrences. The top histogram reports the distribution of a negative partial charge descriptor (accounting for a fraction of van der Waals surface area in a molecule), and the histogram at the bottom shows the distribution of log *P*(o/w), the logarithm of the octanol/water partition coefficient. For both distributions, Shannon entropy (SE) values were calculated as described in the text. The broader descriptor distribution indicates greater variability, which is quantified by calculation of SE values.

$\log_2 p_i$, with p being the probability of a data point c to adopt a value falling within a data interval i. The probability p is calculated as $p_i = c_i/\sum c_i$. Probabilities and SE values can be calculated and compared for any sets of data divided into evenly spaced intervals, and, therefore, SE values for different types of data can be directly compared as long as the data representation is uniform. This is the case when data sets are represented as histograms where any data range is divided into the same number of intervals or bins. Figure 2 shows some representative descriptor distributions recorded in such histograms. By combining descriptor entropy analysis and binary QSAR[60] calculations, it has been possible to systematically distinguish between compounds from synthetic and natural sources.[61]

DIVERSE FINGERPRINTS FOR VIRTUAL SCREENING

Fingerprints reported in the literature may have very different design and complexity. Representative four-point pharmacophore fingerprints monitor the presence or absence of between nine and 24 million potential pharmacophore arrangements[27,30,62] and consist of corresponding numbers of bits. Thus, these bit string representations are extremely large. The other end of the fingerprint spectrum is defined by short 2D structural fragment-based representations such as MACCS[18] (166 bits) and "mini-fingerprints" (MFPs)[50,63] that consist of only approximately 60 bit positions and combine a number of structural keys and a few other 2D descriptors (e.g., accounting for hydrogen bonding or aromatic character), selected on the basis of high prediction accuracy in compound classification.[50] Despite remarkable differences in size, these types of fingerprints have in common that each bit position can be associated with the presence or absence of a specific feature, be it a potential pharmacophore, a structural key, or a value range of numerical descriptors. This is different, for example, in Daylight fingerprints[17] that consist of up to 2048 bits and are representative of a hashed or folded design. This means that, due to chosen fingerprint size limitations, different chemical features (in this case, 2D connectivity pathways) are mapped to the same bit segments. Consequently, bits in these segments can no longer be associated with a specific feature of property.

The similarity or overlap of binary fingerprints, however designed, is mostly determined by calculation of Tc (or Tc-like) values, as discussed above. What Tc threshold values should be applied to consider molecules similar on the basis of fingerprint comparison? With the introduction of the "neighborhood behavior" concept, a Tc similarity cutoff value of 0.85 was suggested[56] because at this level, 80% of molecules identified as "similar" to an active compound would also be expected to be active.[56] However, systematic evaluation of structural key-type fingerprints designed to recognize molecules with similar activity has revealed an optimum performance at lower Tc threshold values of approximately 0.7.[63] It is clear that both the stringency of similarity criteria and specifics of fingerprint design determine the results of virtual screening calculations.

Although earlier 3D pharmacophore queries were able to take flexibility into account,[64,65] pharmacophore fingerprints present a significant advance for virtual screening. They no longer rely on accurate prediction of active conformations or 3D alignments of molecules, which is difficult in many cases, if not impossible. Moreover, these fingerprints monitor possible spatial arrangements of functional groups at an unprecedented level of resolution, reflected by the large number of potential pharmacophores. Although computationally expensive, the exploration of millions of possible spatial arrangements of functional groups has become readily feasible, due to the availability of powerful computers with large amounts of disk space and memory and efficient 2D/3D structure conversion[66] and conformational search[30] techniques. The underlying idea is that similarity (or overlap) of potential pharmacophores indicates that molecules are related in terms of their binding characteristics or activity. However, whether this idea is generally valid or limited to certain molecules and receptor−ligand interactions remains to be determined. In the absence of detailed knowledge about these interactions, a pharmacophore model only represents a hypothesis, regardless of how and at what resolution it is explored.

What level of resolution is required for effective identification of compounds with similar activity? Support for the application of high-resolution 3D methods comes from a number of examples where pharmacophore fingerprinting was able to detect similarities that could not be established

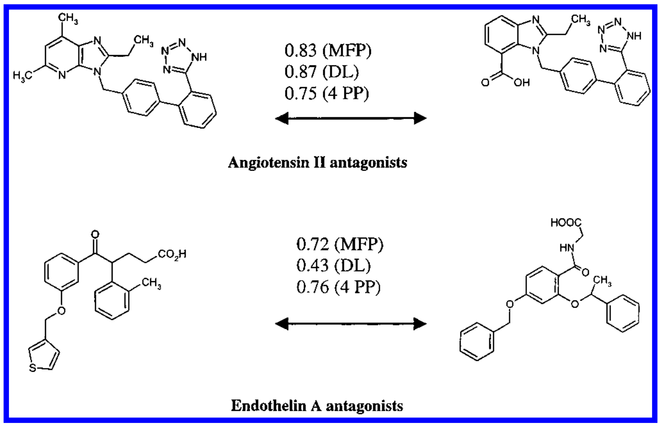


Figure 3. Comparison of angiotensin II and endothelin A antagonists using different fingerprints.The structures and Tc values reported for the Daylight (DL) fingerprint, and Tc-like values for a four-point pharmacophore (4 PP) fingerprint[27] are taken from ref 27. For comparison, Tc values were calculated for a mini-fingerprint (MFP) consisting of only 62 bit positions.[50] The examples are of increasing difficulty. The similarity of the two angiotensin II antagonists can be well detected using the different fingerprints, whereas the similarity of the endothelin A antagonists is difficult to detect using the Daylight fingerprint.

with, for example, the Daylight fingerprint. These examples include endothelin A antagonists,[27] a$_1$-adrenergic receptor ligands,[30] or fibrinogen receptor antagonists that mimic a natural tripeptide ligand motif.[62] However, as illustrated in Figures 3 and 4, 2D fingerprints other than the ones used for initial comparisons may recognize at least some of these relationships, and it is therefore difficult to conclude that 2D methods would generally fail to do so. Furthermore, representations that distinguish fine chemical details may be overly sensitive to differences occurring in analogues or series of molecules having similar activity and may thus not detect certain similarity relationships.[63] On the other hand, test calculations summarized in Figures 5−8 suggest that relatively simple 2D fingerprints are capable of producing meaningful results in virtual screening. Therefore, it may be desirable to balance the level of chemical resolution in fingerprint design and emphasize the detection of features that are critical for specific activities, in analogy to pharmacophore-based approaches. Similar to the situation in molecular descriptor analysis, it thus remains difficult to conclude which of the currently available fingerprint-based approaches to virtual screening may consistently perform best, if any. Just as combining 2D and 3D molecular descriptors can provide highly complementary information, fingerprint-based virtual screening may be most effective if different methods are used in conjunction.

THREE-DIMENSIONAL METHODS FOR VIRTUAL SCREENING

As mentioned earlier, similarity searching in three dimensions is not limited to pharmacophore fingerprints. After pioneering work by Gund,[67] numerous methods for 3D pharmacophore and database searching have been introduced[22,68] and new computational approaches and systems for various types of 3D searches continue to be developed.[68,69] Recent progress has been made in this area that is relevant for the identification and differentiation of active compounds by virtual screening. These developments include extensions of the 3D-QSAR concept and applications of specific shape descriptors. The 4D-QSAR method, introduced by Hopfinger and colleagues,[26] adds molecular flexibility and dynamic parameters to 3D-QSAR analysis by calculation of grid cell occupancy descriptors on the basis of conformational sampling of test molecules.[26,70] Thus, conceptually similar to pharmacophore fingerprints, conformational flexibility of candidate molecules is now taken into account, however, separate from the analysis of 3D molecular alignments. Therefore, possible alignments of compounds in training sets can be efficiently evaluated, and grid cell occupancy descriptors suggest spatial positions in molecules where certain functional groups should be placed or avoided in order to optimize activity. The suitability of the 4D-QSAR approach to generate effective models for virtual screening has been demonstrated, for example, by successful analysis and prediction of glycogen phosphorylase b inhibitors.[70,71]

The idea that similarity in shape correlates with similar biological activity of molecules is intuitive and related to the similar property principle. At the least, active molecules must be capable of fitting the shape of binding sites in their targets and shape complementarity may often be a major determinant of binding. This notion has led to the development of a variety of algorithms to calculate and compare the shape of compounds from different molecular representations.[68] For example, on the basis of single rule-based “topomer” conformations of molecules, corresponding 3D shape descriptors were originally introduced as steric fields.[72] Topomer shape descriptors capture all atoms of a molecule, in contrast to many pharmacophore-based methods, but calculate differences in topomeric shape on the basis of molecular fragments. Recently, these descriptors have been applied to screen analogue libraries and very large virtual libraries for molecules with biological activity similar to selected queries.[73,74] In these studies, topomer shape similarity searching has produced encouraging results. For example, a number of angiotensin II antagonists[73] and molecules belonging to a variety of activity classes taken from the literature[74] were correctly identified.

STATISTICAL MODELS OF ACTIVITY

Elegant statistical methods, recursive partitioning[41,75] and binary QSAR,[60] have been introduced to derive predictive models of biological activity on the basis of data sets consisting of active and inactive compounds. Recursive partitioning derives structure−activity relationships by partitioning compounds with the aid of decision trees and applies the deduced rules to predict the activity of other compounds. Binary QSAR uses learning sets to correlate molecular features with a binary formulation of activity (i.e., active or inactive). Thus, it establishes structure−activity relationships that can then be applied to select compounds from databases that have a high probability of being active. The methods are not limited in their choice of molecular descriptors. An advantage of recursive partitioning is that it can make use of many more descriptors than most related methods. For both approaches, some successful predictions have been reported.[75,76] One of the most attractive features of these statistical methods is that they can be applied to process very large compound data sets. Thus, they are particularly well-

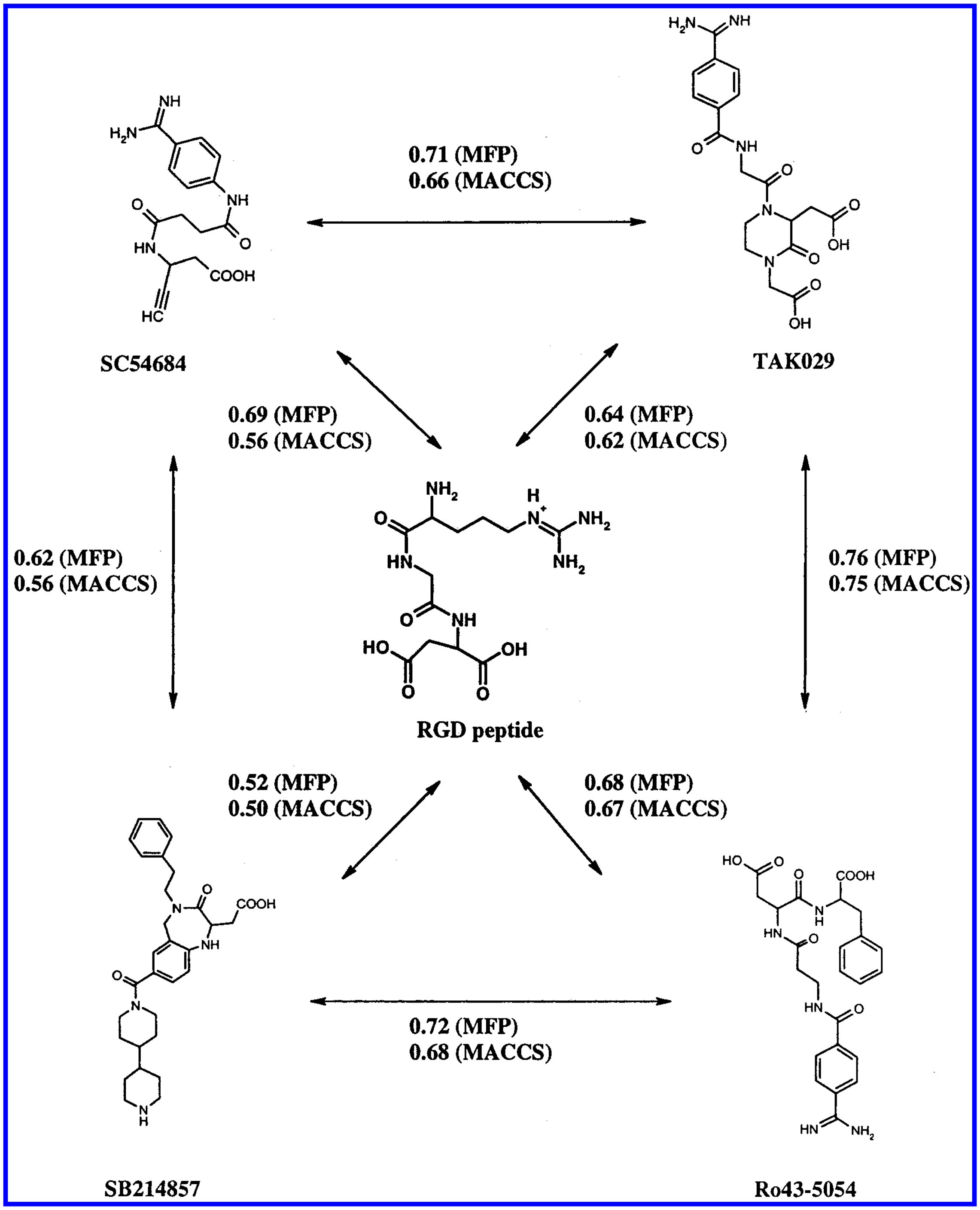


Figure 4. Comparison of fibrinogen receptor antagonists and the RGD motif using 2D fingerprints. Shown in the center is an Arg-Gly-Asp (RGD) peptide motif that occurs in fibrinogen and is recognized by its receptor. This peptide motif is conformationally flexible and should be difficult to mimic by small molecules in the absence of detailed knowledge of its bound conformation. In this regard, it is an excellent test case for pharmacophore fingerprinting,[62] and four-point pharmacophore analysis has successfully revealed similarity of the RGD motif to a number of fibrinogen receptor antagonists belonging to different structural classes.[62] Some of these antagonists are shown here. It should be noted that the peptide motif itself is a natural ligand and therefore an agonist. Thus, similarity evaluated in these studies relates more to binding than function. Here Tc values are reported that were calculated for pairwise comparison of fibrinogen receptor antagonists and the RGD peptide using structural key-based fingerprints (MACCS and the MFP used for the comparison shown in Figure 3). As can be seen, these relatively simple fingerprints detect similarity between antagonists and also between antagonists and the RGD peptide, albeit to a lesser extent. Therefore, even for this challenging test case, it is difficult to conclude that recognition of molecular similarity is strictly limited to 3D pharmacophore analysis.

suited to analyze and exploit data provided by high-throughput screening experiments, even if the results are preliminary (e.g., if a number of poorly characterized hits have been obtained). Without doubt, these approaches significantly extend the spectrum of virtual screening methods.

DRUGLIKE PROPERTIES

Another area of intense research related to compound classification as discussed herein is the analysis of molecular features beyond primary activity that render molecules

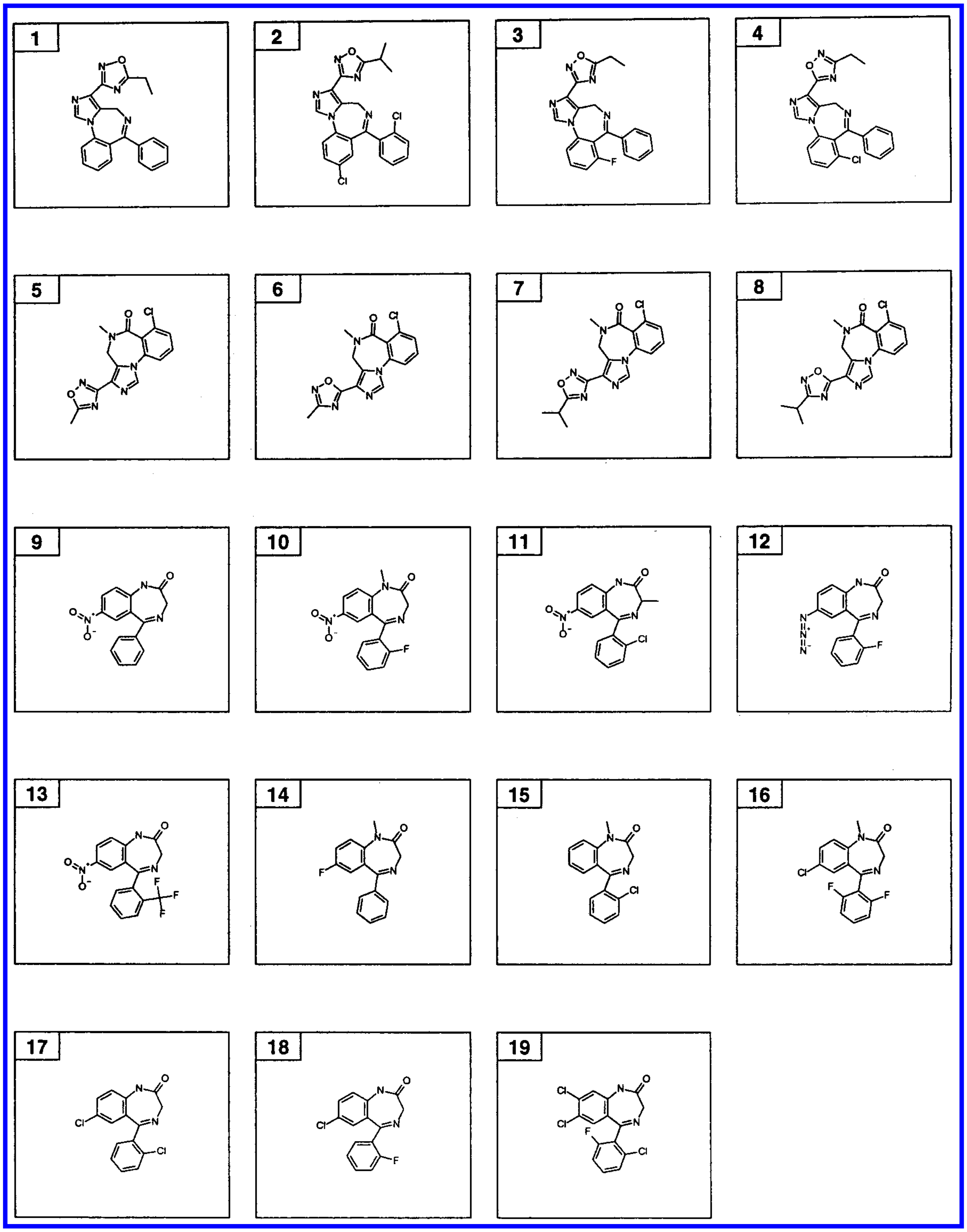


Figure 5. Virtual screening calculations. A number of structurally diverse benzodiazepine receptor ligands[88] were added to 2000 randomly selected ACD[93] compounds. (See also Figure 7 for results of test calculations).

"druglike". Traditionally, the introduction or improvements of such features have been focal points of medicinal chemistry efforts. For example, it is generally known that the polar character of molecules inversely correlates with their ability to diffuse through membranes or that the lipophilic character correlates with (nonspecific) binding to plasma proteins. In recent years, various computational models have been developed and applied to systematically explore druglike properties.[77] These concepts include rule-based approaches,[78] predictive models of membrane permeability,[79] absorption,[80] ADME characteristics,[81] and property profile[52] or neural network-based[82,83] methods to systematically distinguish between druglike and nondruglike molecules. The analysis and prediction of druglike properties is one of the areas of compound design and analysis that is significantly impacted by neural network methods.[84] Insights provided by these and other approaches are beginning to influence library design strategies,[81,85] and it is expected that ADME parameters and druglike characteristics will play an increasing role in compound design and selection.[77,86]

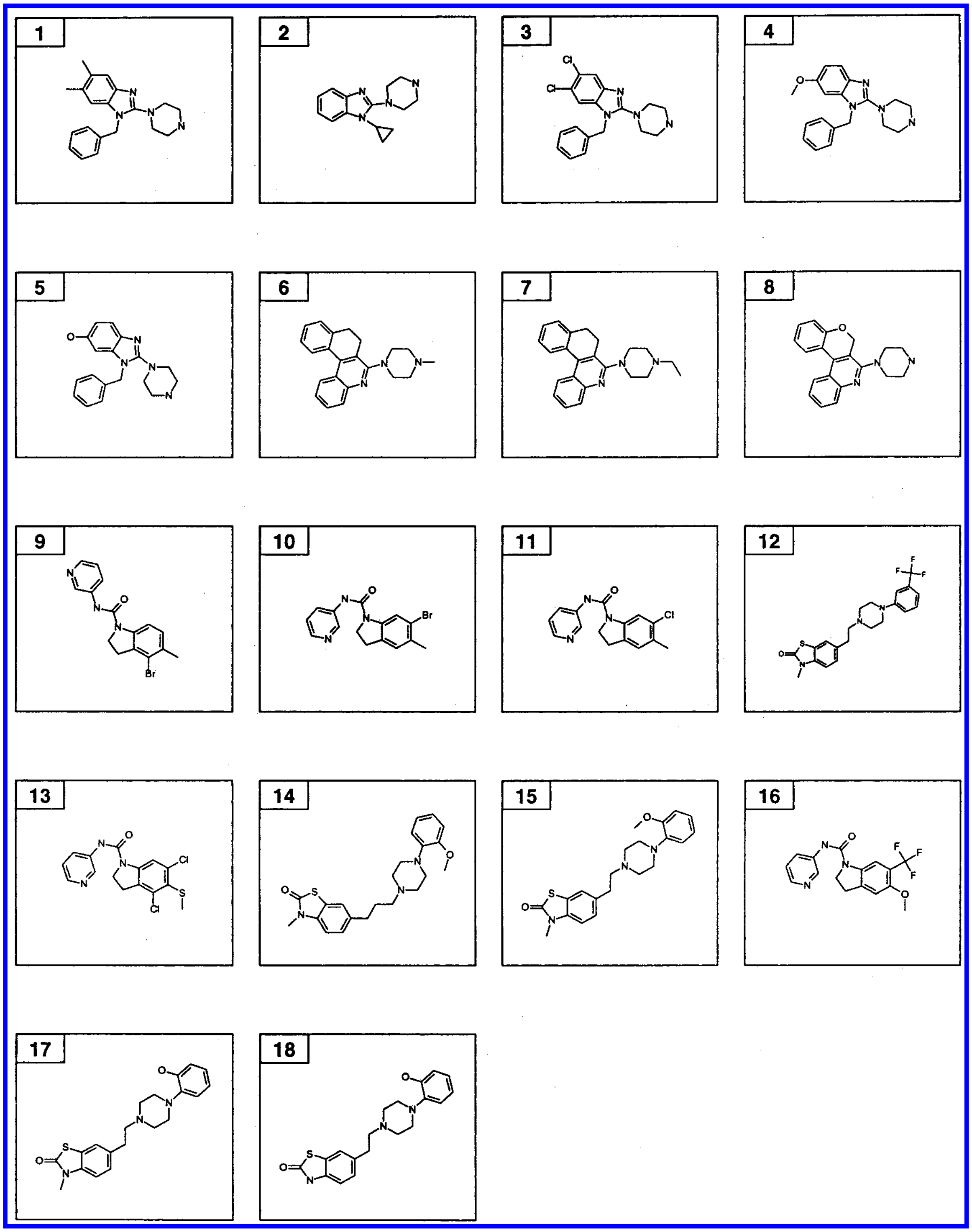


Figure 6. Virtual screening calculations. A number of structurally diverse serotonin receptor antagonists[89−91] were added to 2000 randomly selected ACD[93] compounds. (See also Figure 8 for results of test calculations).

CURRENT TRENDS AND CHALLENGES

Some general trends have become evident in compound classification and virtual screening. Cell-based partitioning techniques are increasingly popular and widely applied. Furthermore, fingerprint representations continue to be developed for the detection of molecular similarity. It is promising to note that diverse fingerprints, albeit strikingly different in their complexity, produce meaningful and sometimes comparable results in cases reported so far. However, it should be noted that although compound selection and virtual screening methods are intensely applied in pharmaceutical settings, it is fair to assume that the vast majority of successful applications, and also failures, are not reported. Consequently, the literature is currently dominated by reports describing computational methods relevant for drug discovery, rather than applications or case studies. The field would certainly benefit from the availability of more “real life” examples of how different methods are actually applied to drug discovery problems. In addition, an organized forum for comparative evaluation of virtual screening

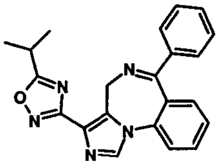
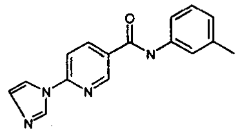
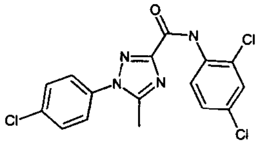
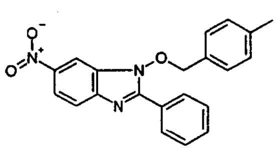
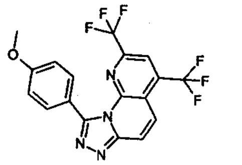
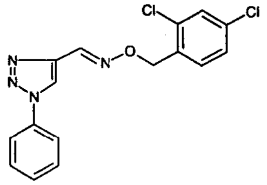
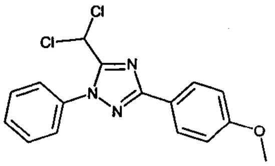
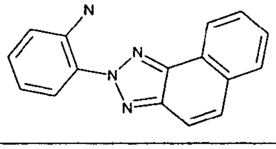
 **Query molecule: Benzodiazepine receptor ligand**			**Database: 19 benzodiazepine receptor ligands added to 2,000 randomly selected ACD compounds.**
	MFP		**MACCS**
	Tc cut-off value: 0.85		
Correctly identified compounds	**4** **(1, 2, 3, 4)**		**3** **(1, 2, 4)**
False positives	**0**		**0**
	Tc cut-off value: 0.70		
Correctly identified compounds	**8** **(1 to 8)**		**8** **(1 to 8)**
False positives	**7**		**0**
			
			
			
			

Figure 7. Results obtained for the benzodiazepine antagonist search (compare to Figure 5). In test calculations, the resulting databases were searched for molecules similar to the query compound. In each case, two fingerprints (MACCS and MFP) were used and two Tc cutoff values applied (0.7 and 0.85). Numbers in parentheses refer to correctly identified compounds according to Figure 5, and the structures of false positive ACD compounds are shown. In each calculation, the number of correctly identified compounds exceeded the number of false positives. At the lower Tc threshold value of 0.7, eight or nine (of 18 or 19) compounds were correctly identified and zero to seven false positives were detected. At the more stringent Tc value of 0.85, between two and four compounds were correctly identified and false positives were completely eliminated.

methods on contributed or specifically assembled data sets, similar perhaps to the protein structure prediction initiative,[87] may be helpful in understanding the opportunities and limitations of different approaches in greater detail.

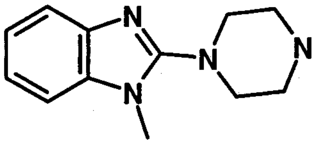
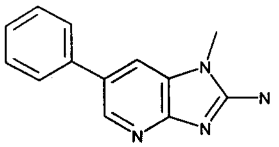
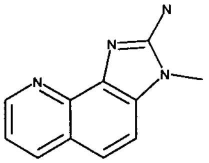
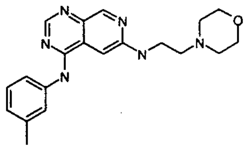
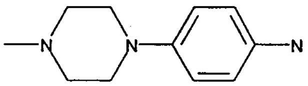
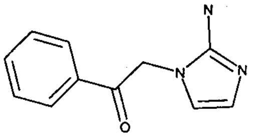
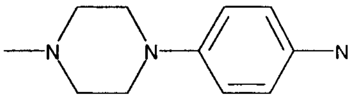
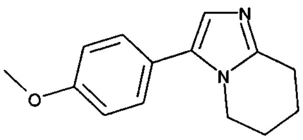
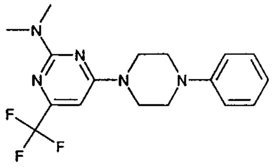
 Query molecule: Serotonin receptor ligand			Database: 18 Serotonin receptor ligands added to 2,000 randomly selected ACD compounds
	MFP		MACCS
	Tc cut-off value: 0.85		
Correctly identified compounds	2 (1, 2)		2 (1, 2)
False positives	0		0
	Tc cut-off value: 0.70		
Correctly identified compounds	7 (1 to 7)		8 (1 to 8)
False positives	5		3
			
			
			

Figure 8. Results obtained for the serotonin antagonist search (compare to Figure 6). In test calculations, the resulting databases were searched for molecules similar to the query compound. In each case, two fingerprints (MACCS and MFP) were used and two Tc cutoff values applied (0.7 and 0.85). Numbers in parentheses refer to correctly identified compounds according to Figure 6, and the structures of false positive ACD compounds are shown. In each calculation, the number of correctly identified compounds exceeded the number of false positives. At the lower Tc threshold value of 0.7, eight or nine (of 18 or 19) compounds were correctly identified and zero to seven false positives were detected. At the more stringent Tc value of 0.85, between two and four compounds were correctly identified and false positives were completely eliminated.

CONCLUSIONS

Over the past few years, much progress has been made in the development of methods for classification, selection, and design of compounds and prediction of biological activity. For example, cell-based partitioning methods and multiple-point pharmacophore-based approaches have significantly-advanced. In addition, a number of studies have begun to evaluate the performance of different molecular descriptors that are central to many approaches in this area. From these studies, we can conclude that increasingly complex representations of molecules are not always better. In fact, relatively simple 2D methods perform remarkably well in a number of applications. As one may perhaps expect, the relative performance of methods and molecular descriptors appears to be much influenced by the specific nature of the problem under investigation. The diversity of methods that have become available for virtual screening and database analysis is one of the most exciting aspects of research in this area and provides opportunities for further development. For example, the complementary nature of 2D and 3D methods and the combination of small molecular- and target structure-based approaches are as of yet relatively little explored avenues in virtual screening.

ACKNOWLEDGMENT

I would like to thank Florence Stahura, Jeffrey Godden, and Ling Xue for help in preparing the figures and reviewing the manuscript.

REFERENCES AND NOTES

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

(2) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening−An overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(3) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078−1082.

(4) Antel, J. Integration of combinatorial chemistry and structure-based drug design. *Curr. Opin. Drug Discoery Dev.* **1999**, *2*, 224−233.

(5) Gane, P. J.; Dean, P. M. Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* **2000**, *10*, 401−404.

(6) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.

(7) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363−372.

(8) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.

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

(10) Pearlman, R. S.; Smith, K. M. Metric validation and the receptor-relevant subspace concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(11) 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.

(12) Mason, J. S.; Pickett, S. D. Partition-based selection. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 85−114.

(13) Bayley, M. J.; Willett, P. Binning schemes for partition-based compound selection. *J. Mol. Graphics Modell.* **1999**, *17*, 10−18.

(14) Ward, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236−244.

(15) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbors. *IEEE Trans. Comput.* **1973**, *C-22*, 1025−1034.

(16) Barnard, J. M. Substructure searching methods. Old and new. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532−538.

(17) James, C. A.; Weininger, D. *Daylight theory manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1995.

(18) *MACCS keys*; MDL Information Systems Inc.: San Leandro, CA.

(19) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

(20) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(21) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128−136.

(22) Martin, Y. C. 3D database searching in drug design. *J. Med. Chem.* **1992**, *35*, 2145−2154.

(23) 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.

(24) Mason, J. S.; Cheney, D. L. Ligand−receptor 3-D similarity studies using multiple 4-point pharmacophores. *Pac. Symp. Biocomput.* **1999**, *4*, 456−467.

(25) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(26) Hopfinger, A. J.; Wang, S.; Tobarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(27) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview over the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(28) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(29) Mason, J. S.; Cheney, D. L. Library design and virtual screening using multiple point pharmacophore fingerprints. *Pac. Symp. Biocomput.* **2000**, *5*, 576−587.

(30) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: Description and application to α_1-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(31) Willett, P.; Wintermann, V.; Bawden, D. Implementation of nonhierarchic cluster analysis methods in chemical information systems: Selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.

(32) 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.

(33) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational screening set design and compound selection: Cascaded clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.

(34) Doman, T. N.; Cibulskis, J. M.; McCray, P. D.; Spangler, D. P. Algorithm5: A technique for fuzzy similarity clustering of chemical inventories. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1195−1204.

(35) Weininger, D.; Delany, J. Clustering package user's guide. *Daylight theory manual*; Daylight Chemical Information Systems, Inc.: Irvine, CA, 1995.

(36) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(37) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(38) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.

(39) Wild, D. J.; Blankley, C. J. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 155−162.

(40) Wesolowsky, G. *Multivariate Regression and Analysis of Variance*; John Wiley & Sons: Toronto, 1976.

(41) Chen, X.; Rusinko, A., III, Young, S. S. Recursive partitioning analysis of a large structure−activity data set using three-dimensional descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.

(42) Dixon, S. L.; Villar, H. O. Investigation of classification methods for the prediction of activity in diverse chemical libraries. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 533−545.

(43) Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity measures for rational set selection and analysis of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599−614.

(44) Stanton, D. T. Evaluation and use of BCUT descriptors in QSAR and QSPR analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.

(45) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of nearest-neighbor and cluster analysis in pharmaceutical lead discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.

(46) Schnur, D. Design and diversity analysis of large compound libraries using cell-based methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.

(47) Pirard, B.; Pickett, S. D. Classification of kinase inhibitors using BCUT descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.

(48) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 699−704.

(49) 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.

(50) Xue, L.; Godden, J.; Bajorath, J. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1227−1234.

(51) Benigni, R.; Gallo, G.; Giorgi, F.; Giuliani, A. On the equivalence between different descriptions of molecules: Value for computational approaches. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 575−578.

(52) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

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

(54) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(55) Matter, H.; Pötter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.

(56) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(57) Schuffenhauer, A.; Gillet, V. J.; Willett, P. Similarity searching in files of three-dimensional chemical structures: Analysis of the BIOSTER database using two-dimensional fingerprints and molecular field descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 295−307.

(58) 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.

(59) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, 1963.

(60) Labute, P. Binary QSAR: A new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* **1999**, *4*, 444−455.

(61) Stahura, F. L.; Godden, J. W.; Xue, L. Bajorath, J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.

(62) Pickett, S. D.; McLay, I. M.; Clark, D. E. Enhancing the hit-to-lead properties of lead optimization libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 263−272.

(63) Xue, L.; Godden, J.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.

(64) Hurst, T. Flexible 3D searching: The directed tweak technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.

(65) Clark, D. E.; Jones, G.; Willett, P. Kenny, P. W.; Glen, R. C. Pharmacophoric pattern matching in files of three-dimensional structures: Comparison of conformational-searching algorithms for flexible searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197−206.

(66) Pearlman, R. S. Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Auto. News* **1987**, *2*, 1−7.

(67) Gund, P. Three-dimensional pharmacophore pattern searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117−143.

(68) Good. A. C.; Mason, J. S. Three-dimensional structure database searches. *Rev. Comput. Chem.* **1996**, *7*, 67−117.

(69) Wang, T.; Zhou, J. 3DFS: A new 3D flexible searching system for use in drug design. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 71−77.

(70) Venkatarangan, P.; Hopfinger, A. J. Prediction of ligand−receptor binding free energy by 4D-QSAR analysis: Application to a set of glucose analogue inhibitors of glycogen phosphorylase. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1141−1150.

(71) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual high throughput screen by 4D-QSAR analysis: Application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1151−1160.

(72) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Ferguson, A. M. Bioisosterism as a molecular diversity descriptor: Steric fields of single topomeric conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.

(73) Cramer, R. D.; Poss, M. A.; Hermsmeier, M. A.; Caulfield, T. J.; Kowala, M. C.; Valentine, M. T. Prospective identification of biologically active structures by topomer similarity searching. *J. Med. Chem.* **1999**, *42*, 3919−3933.

(74) Andrews, K. M.; Cramer, R. D. Toward general methods of targeted library design: Topomer shape similarity searching with diverse structures as queries. *J. Med. Chem.* **2000**, *43*, 1723−1740.

(75) 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.

(76) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure−activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.

(77) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5*, 49−58.

(78) 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. Delivery Rev.* **1997**, *23*, 3−25.

(79) Norinder U.; Osterberg, T.; Artursson, P. Theoretical calculation and prediction of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. *Pharm. Res.* **1997**, *14*, 1786−1791.

(80) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.

(81) Darvas, F.; Dorman, G.; Papp, A. Diversity measures for enhancing ADME admissibility of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 314−322.

(82) Ajay, Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(83) Sadowski, J.; Kubinyi, H. A scoring scheme to distinguish between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(84) Sadowski, J. Optimization of chemical libraries by neural network methods. *Curr. Opin. Chem. Biol.* **2000**, *4*, 280−282.

(85) Ajay, Bemis, G. W., Murcko, M. A. Designing libraries with CNS activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.

(86) Tropsha, A. Recent trends in computer-aided drug discovery. *Curr. Opin. Drug Discovery Dev.* **2000**, *3*, 310−313.

(87) Lattman, E. E., Ed. *Third Meeting on the Critical Assessment of Techniques for Protein Structure Prediction. Proteins: Structure, Function, and Genetics*; 1999; Vol. 37, Supplement S3.

(88) Hadjipavlou-Litina, D.; Hansch, C. Quantitative structure−activity relationships of the benzodiazepines. A review and reevaluation. *Chem. Rev.* **1994**, *94*, 1483−1505.

(89) Bromidge, S. M.; Dabbs, S.; Davies, D. T.; Duckworth, D. M.; Forbes, I. T.; Ham, P.; Jones, G. E.; King, F. D.; Saunders: D. V.; Starr, S.; Thewlis, K. M.; Wyman, P. A.; Blaney, F. E.; Naylor, C. B.; Bailey, F.; Blackburn, T. P.; Holland, V.; Kennett, G. A.; Riley, G. J.; Wood, M. D. Novel and selective 5-HT$_{2c}$/$_{2b}$ receptor antagonists as potential anxiolytic agents: Synthesis, quantitative structure−activity relationships, and molecular modeling of substituted 1-(3-pyridylcarbamolyl)-indolines. *J. Med. Chem.* **1998**, *41*, 1598−1612.

(90) Taverne, T.; Diouf, O.; Depreux, P.; Poupaert, J. H.; Lesieur, D.; Guardiola-Lemaitre, B.; Renard, P.; Rettori, M.-C.; Caignard, D.-H.; Pfeiffer, B. Novel benzothiazolin-2-one and benzoxazin-3-one arylpiperazine derivatives with mixed 5-HT$_{1A}$/D2 affinity as potential atypical antipsychotics. *J. Med. Chem.* **1998**, *41*, 2010−2018.

(91) Morreale, A.; Galvez-Ruano, E.; Iriepa-Canaha, I.; Boyd, D. B. Arylpiperazines with serotonin-3-antagonist activity: A comparative molecular field analysis. *J. Med. Chem.* **1998**, *41*, 2029−2039.

(92) 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.

(93) *Available Chemicals Directory*; MDL Information Systems Inc.: San Leandro, CA.

CI0001482