The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries

Steven L. Dixon* and Ryan T. Koehler†

Telik, Inc., 750 Gateway, South San Francisco, California 94080

Received December 16, 1998

We have carried out a number of sampling experiments in libraries of bioactive compounds to illustrate how size biases introduced by two-dimensional (2D) fragment distance functions may provide misleading information about the diversity of compound subsets. The number of different biological targets covered by a given subset is used as a measure of bioactive diversity, and it is considered to be the relevant property with which 2D diversity should correlate. Since the nature of the size biases depends on the way in which 2D distance is computed, we investigated three different methods of calculating distance. Use of 1-Tanimoto as a dissimilarity measure leads to the spurious conclusion that collections of structurally small compounds are inherently more diverse than other collections which may cover a broader range of sizes and more biological targets. XOR or squared Euclidean distance, by contrast, shows a preference for subsets of structurally larger compounds, but this does not appear to have as many adverse consequences in terms of target coverage. A simple product of 1-Tanimoto and XOR tends to equalize the opposing size effects of the two component distance functions and leads to a relatively unbiased means of comparing structures. Results here suggest that careful consideration should be given to the way in which chemical structures are compared whenever 2D fragment descriptors are used.

Introduction

As the boundaries of combinatorial chemistry and high-throughput biological screening have expanded, so too has the need for fast and meaningful computer-based comparisons of molecular structures. With corporate libraries converging on one million compounds and virtual libraries containing orders of magnitude more, such methodologies have become an absolute necessity. Any means of comparing chemical structures is valid only to the extent that it reflects intuitive notions about similarity that have evolved over decades in the field of medicinal chemistry. These ideas are embodied in the *similar property principle*,[1] which states that compounds with similar structures will tend to exhibit similar physicochemical and biological properties. This concept provides much of the framework on which modern lead optimization is built.

Curiously, though the similar property principle makes no claims regarding dissimilar structures, it is also used as the basis for essentially all work in the field of molecular diversity.[2−4] Basically, the converse of the principle is used to infer that dissimilar-looking structures will exhibit dissimilar properties. Though this may be true to some extent, there is not always a valid, global relationship[5] between structural dissimilarity and differences in measured properties such as biological activity. In general, as compounds become more diverse structurally, we are progressively less certain of how they compare to one another in terms of biological activity.[5] For these reasons, we must be careful in drawing conclusions about diversity based solely on calculated measures of dissimilarity, and we should bear in mind that biological targets provide the ultimate scale on which diversity is usually measured.

Without some a priori knowledge of the structural features that govern activity, or at least knowledge of the best sets of descriptors to use when dealing with specific targets, diversity in any true *bioactive* sense is not something that can be easily manipulated by choice of compounds. There are, however, some basic controllable factors that can have an effect on bioactive properties and certain minimum requirements that should be met in this regard. In particular, when selecting compounds from a library on the basis of dissimilarity, one should be confident that gross structural biases are not being introduced in the process. There is evidence[6,7] to suggest that this sort of thing may be happening when two-dimensional (2D) fragment descriptors are used to measure diversity. Specifically, we are concerned with the way in which widely utilized 2D distance functions introduce biases related to the overall size of compounds and how this may ultimately impact upon the bioactive properties of the subsets selected.

To investigate these effects, we first define appropriate scales on which to measure the properties identified as molecular size, bioactive diversity, and 2D structural diversity. We then carry out three basic types of sampling experiments in libraries of compounds with established pharmacological endpoints. In each experiment, one of the three above properties is varied in a systematic fashion, and the responses in the other two properties are analyzed. This provides a means of isolating the specific effects of 2D size biases and allows a determination to be made as to whether these effects

* To whom correspondence should be addressed. E-mail: sdixon@telik.com.
† E-mail: koehler@telik.com.

10.1021/jm980708c CCC: $18.00
Published on Web 07/07/1999

are detrimental in a true bioactive sense. The sampling experiments demonstrate that Tanimoto coefficient[8] and squared Euclidean distance are prone to measuring diversity in a way that favors selection of compounds which differ markedly in average size from the overall library in which they reside. Results also reveal situations wherein size biases introduced by 2D fragment descriptors lead to conclusions about diversity which appear to be at odds with the range of pharmacological properties of the compounds being considered.

Assessing diversity is now a critical component in the lead discovery process, and success on this front is largely determined by how intelligently we utilize the myriad of tools at our disposal. Whether designing new, innovative techniques in the field of molecular diversity or simply applying principles from well-established methods, it is important to avoid known pitfalls, even if little attention has been paid to them in the past. It is the intent of this paper to raise awareness about some of these pitfalls and provide biologically relevant evidence of why they should be avoided.

Overview of 2D Fragment Descriptors

As a result of their availability through various database packages, 2D fragment descriptors[2,6,9,10] have increased in popularity right along with the movement toward high-throughput biological and chemical methodologies. Their growing use has also been fueled by numerous investigations[5,9−13] which have shown them to be rich in structural information that is relevant to biological activity.

Fragment-based descriptors have been around for decades,[14−16] but only in the past few years have they been routinely cast into the high-dimensional bit string representations that are the focus of this paper. For conceptual purposes, these types of descriptors may be divided into two categories:[2,9] structural keys and hashed fingerprints.

Structural keys are based on a predefined *fragment dictionary*, which is a set of fragments that that are relevant for some purpose, usually efficient database searching, and which occur with varying frequency in 2D depictions of chemical structure. A given compound is represented by a string of ones and zeros that encode the presence, absence, and sometimes the frequency of appearance of each predefined fragment. In the case of frequency, a set of n bits is allocated for a fragment to encode up to n occurrences in the compound.

Hashed fingerprints, by contrast, do not rely upon a fixed fragment dictionary but rather on an exhaustive substructure enumeration procedure that identifies all possible fragments within some restricted domain, e.g., paths containing 1−7 bonds. Each unique fragment in a molecule is assigned a numerical value that is a function of its structure, and this number is used as input to a hashing algorithm that sets a collection of bits along a string of some fixed length. There are usually many more unique fragments in a chemical library than there are bits in the fingerprint string, so it is possible for two different fragments to set some of the same bits. However, the length of the string can be adjusted so that these *collisions* do not occur frequently enough to obscure a significant fraction of the distinguishing structural information.

There are of course advantages and disadvantages to using either type of fragment representation. Because they are derived from a fixed fragment dictionary, structural keys are often criticized for lacking the generality that is inherent in hashed fingerprints. Despite these purported limitations, structural keys have been shown to perform quite well in comparison to hashed fingerprints in studies involving biological activity data.[9,10]

Bit String Comparison Methods

Tanimoto coefficient[8] (TC) is probably the most commonly used parameter for measuring similarity between bit string representations. For compounds i and j, it is defined as

$$\mathrm{TC}_{ij} = N_{ij}/(N_i + N_j - N_{ij}) \quad (1)$$

where N_i is the number of bits set by i, N_j is the number of bits set by j, and N_{ij} is the number of bits set by both i and j. An equivalent formulation based on bit-wise logical operators is

$$\mathrm{TC}_{ij} = \sum_k(\mathrm{bit}_{ik}\ \mathrm{AND}\ \mathrm{bit}_{jk})/\sum_k(\mathrm{bit}_{ik}\ \mathrm{OR}\ \mathrm{bit}_{jk}) \quad (2)$$

The logical AND returns a value of 1 if the kth bit is set in both structures and a value of 0 otherwise; the logical OR returns a value of 1 if the kth bit is set in at least one structure, and 0 if it is set in neither structure. TC increases toward a value of 1 as bit strings differ at fewer and fewer positions. A minimum similarity of 0 is reached when two structures do not set any of the same bits.

For purposes of measuring dissimilarity or distance, 1 − TC is the natural choice. Straightforward manipulation of eq 1 yields

$$1 - \mathrm{TC}_{ij} = [(N_i + N_j - N_{ij}) - N_{ij}]/(N_i + N_j - N_{ij}) \quad (3)$$

Note that the numerator represents the union of set bits minus the intersection of set bits, which is simply the number of positions at which the two bit strings differ. The logical XOR (exclusive OR) operator returns this quantity in a bit-wise fashion:

$$1 - \mathrm{TC}_{ij} = \sum_k(\mathrm{bit}_{ik}\ \mathrm{XOR}\ \mathrm{bit}_{jk})/\sum_k(\mathrm{bit}_{ik}\ \mathrm{OR}\ \mathrm{bit}_{jk}) \quad (4)$$

The potential for size-related biases becomes evident when one considers the behavior of TC in the case of small molecules. As observed by Flower,[6] small structures return characteristically low average Tanimoto similarities when queried against a typical library of compounds. One contributing factor is the tendency of small structures to turn on fewer bits, which restricts the number of set bits that can be shared with other structures. Proceeding quantitatively, we see that the maximum value for the numerator in eq 1 is the smaller of the two numbers N_i, N_j. The denominator is bounded below by the larger of N_i and N_j, so the maximum possible value of TC_{ij} is $\min(N_i,N_j)/\max(N_i,N_j)$. Thus, the smaller the structure of the query, the smaller will be the upper bound for TC.

Perhaps less obvious is the pronounced sensitivity of TC to changes in the bit strings when either compound is small. Lajiness[7] demonstrated this behavior math-

Table 1. General Types of Ligands in the Two Bioactive Libraries

	RBI (445 compounds)		CMC (964 compounds)	
	number	fraction	number	fraction
receptor agonists/ antagonists	337	0.757	776	0.804
enzyme inhibitors	84	0.189	170	0.176
ion channel blockers	24	0.054	18	0.019

ematically by holding fixed the value of N_{ij} and varying the number of positions at which the bit strings differed. The greatest sensitivity in TC was observed for small values of N_{ij}, which, as noted before, are associated with small compounds.

Because there is a correlation between the size of a compound and the number of bits it sets, certain biases related to size are unavoidable when 2D fragment descriptors are used. However, TC tends to amplify these effects because it involves a ratio of two size-dependent factors. Despite this undesirable behavior, TC has historically been the parameter of choice in applications involving 2D fragment descriptors.

Another common method of comparing bit strings is based on Euclidean distance d:

$$d_{ij}^2 = \sum_k (\mathrm{bit}_{ik} - \mathrm{bit}_{jk})^2 \qquad (5)$$

The square root operation is frequently avoided to speed up calculations or in cases where a rank-ordering of distances is all that is required. Since the bit values are 0 or 1 the squared Euclidean distance is simply the number of positions at which the bit strings differ, i.e., the XOR distance:

$$d_{ij}^2 = \mathrm{XOR}_{ij} \equiv \sum_k (\mathrm{bit}_{ik}\ \mathrm{XOR}\ \mathrm{bit}_{jk}) \qquad (6)$$

This form is also commonly referred to as the city-block or Hamming distance.[1] Molecular size effects are still a concern, but one size-dependent factor is removed compared to 1 − TC.

Other methods[7−9,17] of comparing bit strings are sometimes used, the most notable among these being the cosine coefficient. However, this association measure typically correlates strongly[17] with TC, so it is not considered here. As shown by Lajiness,[7] the properties of 1 − TC and XOR are quite different, and because so much published work has relied on them, we focus on these measures of dissimilarity in our study of size biases and diversity.

Bioactive Libraries

The phenomena we wish to demonstrate can be observed in essentially any chemical library, but the consequences of size-related effects may not be appreciated unless they are referenced to some property of biological relevance. For this reason, we have carried out our analyses on two libraries of compounds with established bioactive properties, Table 1.

The first set of compounds was taken from the RBI LOPAC library,[18] which is a collection of 640 biologically active agents and biochemical probes that are available on high-throughput screening plates. Biological targets for these compounds were confirmed via the RBI catalog[19] and other sources.[20−22] Compounds for which the target was unknown or unclear were discarded, as well as compounds which were reported to hit more than one target. Note that in the case of receptors we made no distinction among different subtypes, simply because so many of the associated ligands were reported either to show affinity for more than one subtype or the subtype information was incomplete. Since our ultimate goal was to establish a scale of bioactive diversity, the consequence of merging related targets is primarily the creation of a lower bound for diversity. Overall, the filtering process retained 445 of the 640 RBI compounds.

Table 2. Breakdown of Biological Targets According to Ligand Frequency and Size

target	number of ligands	fraction of library	average molecular weight
RBI compounds			
dopamine receptor	53	0.119	380.0
adenosine receptor	44	0.099	389.3
serotonin receptor	41	0.092	365.9
adrenergic receptor	40	0.090	331.9
cholinergic receptor	35	0.079	355.4
NMDA receptor	27	0.061	244.6
opioid receptor	20	0.045	455.3
glutamate receptor	18	0.040	239.3
GABA receptor	18	0.040	242.2
histamine receptor	15	0.034	309.4
top 10 targets	311	0.699	344.1
remaining 53 targets	134	0.301	333.6
CMC compounds			
cholinergic receptor	194	0.201	355.6
adrenergic receptor	139	0.144	275.9
histamine receptor	105	0.109	329.2
glucocorticoid receptor	71	0.074	456.7
opioid receptor	69	0.072	333.9
cyclooxygenase	51	0.053	296.8
estrogen receptor	48	0.050	362.7
serotonin receptor	32	0.033	330.3
dopamine receptor	29	0.030	356.7
angiotensin-converting enzyme	22	0.023	451.4
top 10 targets	760	0.788	343.1
remaining 53 targets	204	0.212	455.7

The second set of compounds was taken from the CMC (Comprehensive Medicinal Chemistry) database, version 98.1, which is available through MDL Information Systems, Inc.[23] This library contains 7497 compounds that have some biological or medicinal application. An analogous filtering process was used to arrive at a set of 964 compounds for which a primary biological target could be readily identified. Targets are certainly known for a significant fraction of the other compounds in the database, but in most instances only the therapeutic class (analgesic, antiinflammatory, diuretic, sedative, etc.) is listed, and it would have been an unwieldy task to even attempt to identify the exact mode of action of each compound.

Table 1 summarizes general properties of the ligands in the two filtered libraries. For both collections, the vast majority of compounds are seen to act at receptor sites, and this simply reflects the relative importance of this class of macromolecules as therapeutic targets. We note that the ligand categories in Table 1 and elsewhere in this investigation are based on the initial site of binding and not on any secondary events that occur as a consequence of binding.

Table 2 is a breakdown of biological targets according to frequency of appearance and average molecular

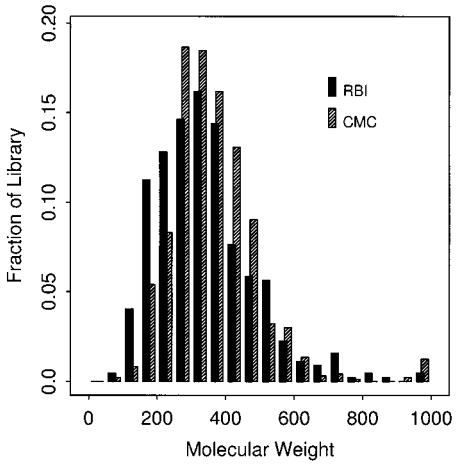

Figure 1. Distribution of molecular weights for the RBI and CMC libraries.

weight of the associated ligands. Molecular weight was selected as a representative measure of size because it is not a function of 3D structure, and it provides a fairly continuous coordinate from which to sample and compare compounds.

Six of the top 10 most frequently appearing targets are common to both libraries, with the primary difference being that the RBI data set is more biased toward targets in the CNS. By sheer coincidence, the libraries each cover a total of 63 targets, 26 of which they have in common. The 53 targets that are not listed appear with frequencies of 1−11 and 1−14 in the RBI and CMC libraries, respectively.

Average molecular weights do not differ significantly between the two sets of compounds when considering only the top 10 targets. However, ligands for the remaining 53 targets differ in average size by more than 120 amu, and as shown in Figure 1, these compounds contribute to a modest overall shift toward higher molecular weights within the CMC set. This fundamental difference is critical for demonstrating size-related effects, and it was one of the primary reasons for selecting these two libraries.

Descriptor Sets

Analyses were carried out using two different sets of fragment descriptors. The ISIS MOLSKEYS[24,25] were chosen as a representative of the structural key class, while the Daylight Fingerprint Toolkit[26] was used to generate hashed fingerprints.

The MOLSKEYS are based on 166 different substructure queries which encode the presence or absence of straight, branched, and cyclic fragments of various sizes, heteroatoms and heterocycles, multiple bonding patterns, and a wide range of other known chemical moieties. While the MOLSKEYS rely on a fixed fragment dictionary, they have been shown to be among the best available sets of 2D descriptors in applications involving biological activity data.[9,10,13]

Daylight hashed fingerprints are generated according to a multi-stage process. First, bits are set to account for each unique atom center in the molecule, with differentiation according to elemental type and the immediate bonding environment. Then, additional bits are set for all unique paths within some specified range of lengths, which, in the present case, was 1−7 bonds. Finally, fragments containing unique cycles and branching are encoded in the fingerprint.

Daylight allows the user to control the size of the fingerprint in two ways. First, one can specify, to the nearest power of two, the length of the bit string onto which the hashed patterns will be mapped. Then one can "fold" the fingerprint in half as many times as is desired, with a logical OR operation being applied to map pairs of bits onto a single bit. The initial length and/or folded length may be adjusted to control the *bit density*, which is the average fraction of bits in the string that are set by the compounds in a particular library. Increasing the bit density, i.e., reducing the size of the fingerprint, saves on computer memory, but it does lead to more frequent collisions among fragments. On the other hand, extremely long fingerprints, though essentially devoid of collisions, are of questionable value because certain bits may never be set, or they may be set so infrequently that they are leveraged on only a handful of compounds in a library. In the present study, fragments were hashed onto a string of 512 bits and no folding was performed. This resulted in bit densities of 0.31 and 0.33 for the RBI and CMC compounds, respectively. These values are very close to the default bit density of 0.30 recommended in the Daylight documentation.[27] By comparison, bit densities for the MOLSKEYS within the two libraries were 0.29 and 0.28.

Effects of Varying Molecular Size

The first set of experiments was designed to see how bioactive diversity and 2D structural diversity vary with average ligand size and to what extent this depends on the library itself. Before proceeding, we must specify exactly how diversity is to be measured, both in terms of biological characteristics and 2D structural characteristics.

Bioactive diversity here is defined as the total number of biological targets covered by a given set compounds. We consider this to be a fairly objective means of expressing the overall range of biological behavior inherent in a set of compounds, and similar approaches have been used by others.[4,5] While this scale of diversity may be somewhat clouded as a result of homology among related biological targets, our procedure of merging different receptor subtypes goes a long way toward eliminating these sorts of ambiguities.

Structural diversity is defined as the average distance observed between all unique pairs of compounds in a given set. Other measures of diversity are certainly possible, such as average nearest neighbor distance, minimum nearest neighbor distance, and so forth. We selected average pairwise distance because it encodes relationships between all pairs of compounds, and it is therefore minimized only when the entire set occupies a single point in descriptor space. Average nearest neighbor distance drops to zero when each compound has a duplicate within the set, and minimum nearest neighbor distance becomes zero when *any* duplicates are present. In either case, the measured diversity is zero, even if the subset, as a whole, spans a wide range of structural classes.

To investigate the effects of varying ligand size, subsets of 50 compounds were selected at random from

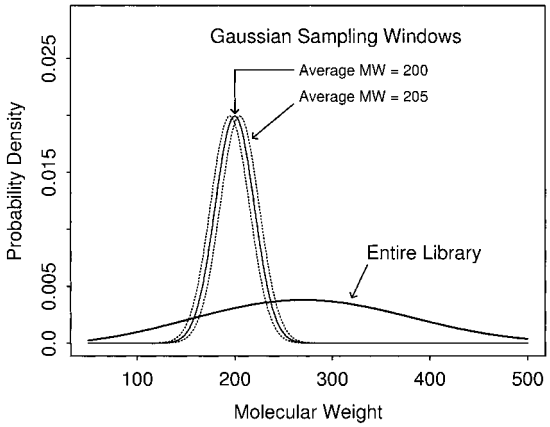

Figure 2. Illustration of the sampling technique used to control average molecular weights and the spread of molecular weights in 50-compound subsets. Reported quantities for each window correspond to the average of 1000 subsets sampled from the associated Gaussian probability distribution.

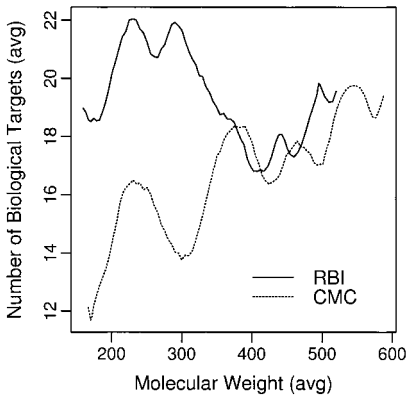

Figure 3. For subsets of 50 compounds, the average number of biological targets covered is monitored as the average molecular weight of the subset is varied.

within narrow, Gaussian-shaped probability windows centered at different points along the molecular weight coordinate, Figure 2. The location of each window determines the average molecular weight of compounds that are sampled, and the standard deviation of the Gaussian controls the overall spread of molecular weights. Gaussians were centered at intervals of 5 amu, and the standard deviation of each window was ±20 amu. Average molecular weights across the series of windows were designed to cover the range of 150−530 for the RBI data set and 150−600 for the CMC compounds. The actual selection of compounds involved generating a random molecular weight from the Gaussian distribution, identifying the compound with the closest molecular weight, then adding it to the current subset if it was not already present. To generate smooth statistics, 1000 subsets were sampled within each probability window, and thus the reported values of molecular weight, biological target counts, and pairwise distances correspond to an average across 1000 trials.

Figure 3 illustrates how the numbers of biological targets covered by the 50-member subsets vary with average ligand size. Target counts tend to be higher for the RBI subsets simply because there are about twice as many targets per ligand in the overall library as compared to the CMC data set. As average ligand size is increased, the two libraries show markedly different behavior, with the number of targets peaking at low

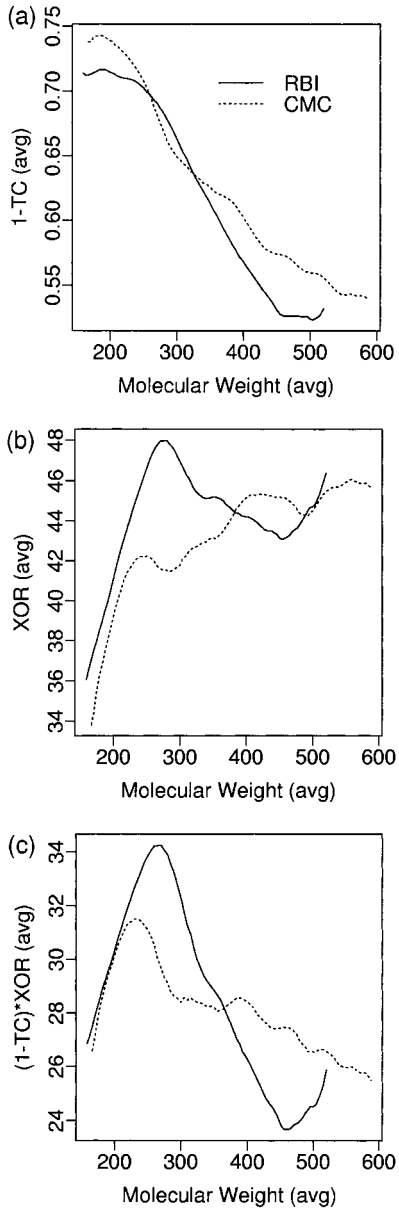

Figure 4. Results for MOLSKEYS. Within subsets of 50 compounds, the average distance between pairs of compounds is monitored as the average molecular weight of the subset is varied.

molecular weights in the RBI collection and at high molecular weights in the CMC collection. This result is consistent with the information in Table 2, and it confirms a strong tendency for subsets of structurally larger compounds to cover more targets in the CMC library.

Figure 4 summarizes the trends in MOLSKEYS distances as average ligand size is varied. When 1 − TC is used, there is a nearly monotonic decrease in average pairwise distance with increasing ligand size in both libraries. The range covered is about 0.2, which corresponds to a swing of 20% in the average Tanimoto similarity. By contrast, the XOR distance shows a peak at about 280 amu in the RBI library and a general tendency to increase with ligand size in the CMC library. Note that the XOR distance does not show as much relative sensitivity to average size as does 1 − TC. The change in the number of dissimilar bits is about 12, which represents only 7.2% of the possible variation

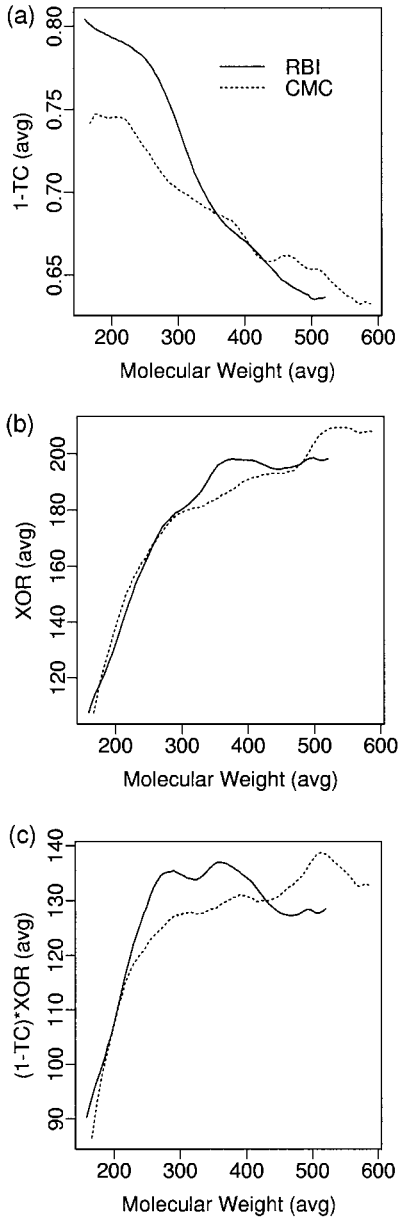

Figure 5. Results for Daylight hashed fingerprints. Within subsets of 50 compounds, the average distance between pairs of compounds is monitored as the average molecular weight of the subset is varied.

in this distance parameter. Because these two methods of measuring distance seem to behave in an opposite sense with regard to molecular size, we have also included results from using the product of the two distances, Figure 4c. This hybrid distance peaks early in both libraries and then tends to drop off with size.

Figure 5 contains the corresponding results for Daylight hashed fingerprints. Here we see the same tendency of 1 − TC to decrease, although the distances are not as sensitive to size, with ranges of only 0.17 and 0.11 for the RBI and CMC libraries, respectively. By contrast, XOR distances from the hashed fingerprints appear to be more sensitive to size than the analogous MOLSKEYS distances. Variations in the number of dissimilar bits for the two libraries correspond to 18% and 20% of the fingerprint length, which is more than twice the relative change seen previously. Curiously, the Daylight XOR distances do not show the same complexity as a function of ligand size that was observed with the MOLSKEYS. Hybrid distances primarily mirror the XOR results since the variations in 1 − TC are comparatively weak for the hashed fingerprints.

The simple fact that average pairwise distances vary with ligand size does not necessarily imply the existence of undesirable features in the fragment descriptors and/or the distance functions. Indeed, as shown in Figure 3, there are clear trends in the libraries with regard to bioactive diversity and ligand size. If a set of fragment descriptors yielded distances which closely tracked the counts of biological targets, then the apparent 2D size biases would not be unwelcome. Unfortunately, no combination of descriptors and distances in Figures 4 and 5 appear to closely mirror the biological target counts in Figure 3.

Effects of Varying Bioactive Diversity

In these experiments, the number of biological targets covered by each subset was varied, and the responses in average ligand size and average pairwise distance were monitored. Combined with the previous set of experiments, these tests help to demonstrate whether the chosen measures of 2D diversity are as sensitive to variations in the bioactive properties of compounds as they are to variations in compound size.

To understand how the sampling was done, first consider the natural distribution of ligands according to biological target. If there are N compounds in a library, and n_i of these are associated with target i, then the natural probability of selecting a ligand of target i is n_i/N. For a collection of M targets, this natural probability distribution can be mapped to the interval (0,1) as follows:

$$\text{Target } 1 \rightarrow (0, n_1/N)$$

$$\text{Target } 2 \rightarrow [n_1/N, n_1/N + n_2/N)$$

$$\ldots$$

$$\text{Target } M \rightarrow [\textstyle\sum_{k=1,M-1} n_k/N, 1)$$

Thus each target is assigned a distinct subinterval, the width of which is proportional to the number of ligands for that target. The ith subinterval can be further subdivided into n_i equal-sized segments, so that each ligand is represented in the distribution. With this mapping scheme, natural sampling proceeds by selecting a random number from the uniform distribution on (0,1) and then choosing the target and ligand that correspond to the subinterval in which the number falls.

To skew sampling away from the natural distribution, the widths of the target subintervals can be manipulated. Any collection of nonnegative numbers $\{p_1, p_2, \ldots, p_M\}$ may be used to partition the interval (0,1) so as to bias sampling in favor of certain targets. In this case, the width of the ith subinterval would be p_i/P, where P is the sum of the numbers $\{p_1, p_2, \ldots, p_M\}$. For the natural distribution, of course, $p_i = n_i$ and $P = N$.

Our goal was to design a sampling procedure that would vary smoothly between low numbers of biological targets and high numbers of biological targets. This was accomplished by defining the partitioning numbers as follows:

$$p_i = (n_i)^t \qquad \text{where } 0 \le t \le 2 \qquad (7)$$

When $t = 0$, the p_i values are all unity and the interval widths are the same for all targets, so there is an equal probability of selecting a ligand from any of the M target classes. This leads to high bioactive diversity in the randomly chosen subsets. When $t = 2$, the disparities in the widths of the intervals are amplified beyond that of the natural distribution ($t = 1$), and there is an even greater tendency to sample ligands from over-represented targets, thus leading to extremely low bioactive diversity.

Sampling experiments were carried out by allowing t to vary over 100 equally spaced values between 0 and 2. At each value of t, 1000 different subsets of 50 compounds were selected, and average values were computed for molecular weight, the number of targets covered, and the three types of pairwise distances.

Since target coverage within the two libraries responds to changes in average ligand size (Figure 3), a second sampling procedure was carried out with the goal of eliminating variations in molecular weight that might naturally accompany the transition from low to high numbers of targets. Accordingly, the average molecular weight across a given subset was monitored and controlled as compounds were added to the subset. If the running average molecular weight dropped below the average observed for the entire library, then only compounds heavier than the running average were accepted. Conversely, only compounds lighter than the running average were accepted whenever this value rose above the library average.

Figure 6 shows how the free and constrained molecular weights change during the course of sampling low to high numbers of biological targets. When no control is exercised over the sizes of the compounds selected, average molecular weights vary by 17 and 98 amu, respectively, in the RBI and CMC libraries. Size effects in the RBI collection are readily controlled by constrained sampling, as the average molecular weight is seen to change by less than 1 amu. A larger drift of about 7 amu is observed for the CMC library, but this is still quite small compared to the unconstrained change of 98 amu. Surprisingly, maintaining control of the molecular weights does not appear to reduce the range of target counts accessible by the basic sampling procedure.

Figure 7 summarizes the corresponding changes that occur in MOLSKEYS pairwise distances as the numbers of biological targets are increased. For ease of comparison, the pairwise distance curves arising from molecular weight variations (Figure 4) are overlaid on this figure. Thus, the horizontal scale for the dotted curves (shown along the top of each plot) is actually molecular weight rather than target counts.

Dissimilarities among compounds in the RBI collection appear to show an overall relationship with bioactive diversity for all three methods of calculating distance and for both sampling procedures (Figure 7a–c). Note, however, that the ranges in pairwise distance observed here are much smaller than they were when molecular weight was systematically varied, so the underlying structure–bioactive diversity relationships in Figure 7a–c are comparatively weak. While there is

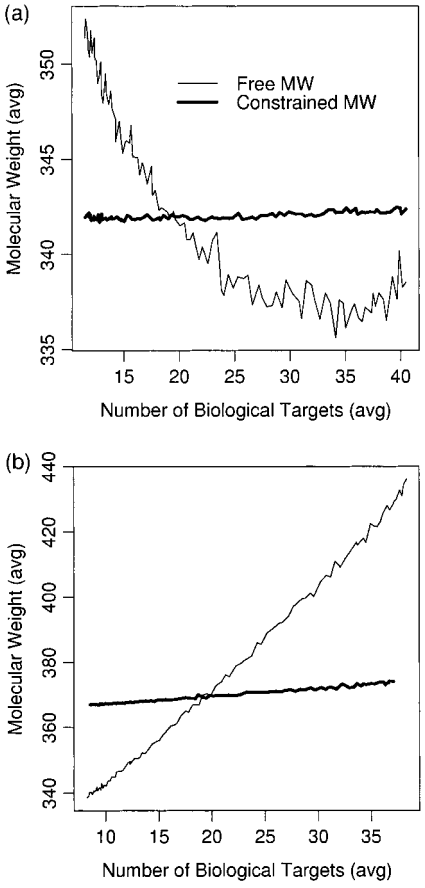

Figure 6. For subsets of 50 compounds, the average molecular weight is monitored as the number of biological targets covered is varied: (a) RBI library; (b) CMC library.

no unequivocal scale on which to compare the effects of varying size and varying bioactive properties, results here suggest that, within the RBI library, all three types of MOLSKEYS distances respond more to changes in molecular weight than to changes in target coverage.

When sampling is done within the CMC library, the 1 − TC results are quite different from the RBI case, and we see a compelling illustration of the biases introduced by this dissimilarity measure, Figure 7d. With unconstrained molecular weight, MOLSKEYS structural diversity is observed to peak and then actually drop off as bioactive diversity increased. The effect is largely removed, however, when molecular weight is constrained, and 1 − TC increases almost monotonically with target counts. These differences are undoubtedly due to the presence or absence of a bias toward larger ligands at the high end of the bioactive diversity scale. When molecular weight is allowed to vary freely, larger compounds are encountered, and 1 − TC measures less structural dissimilarity among these bioactively diverse compounds. The overall effect is small, but it is clearly distinguishable. This phenomenon does not occur with XOR distance (Figure 7e), but it does to a slight degree with the hybrid distance function (Figure 7f), simply because of the factor of 1 − TC.

Figure 8 summarizes results for the same series of tests using Daylight hashed fingerprints. The general behavior here is roughly the same as observed with the MOLSKEYS, with one notable exception: XOR distances in the RBI library decrease almost linearly with increasing bioactive diversity, irrespective of constraints

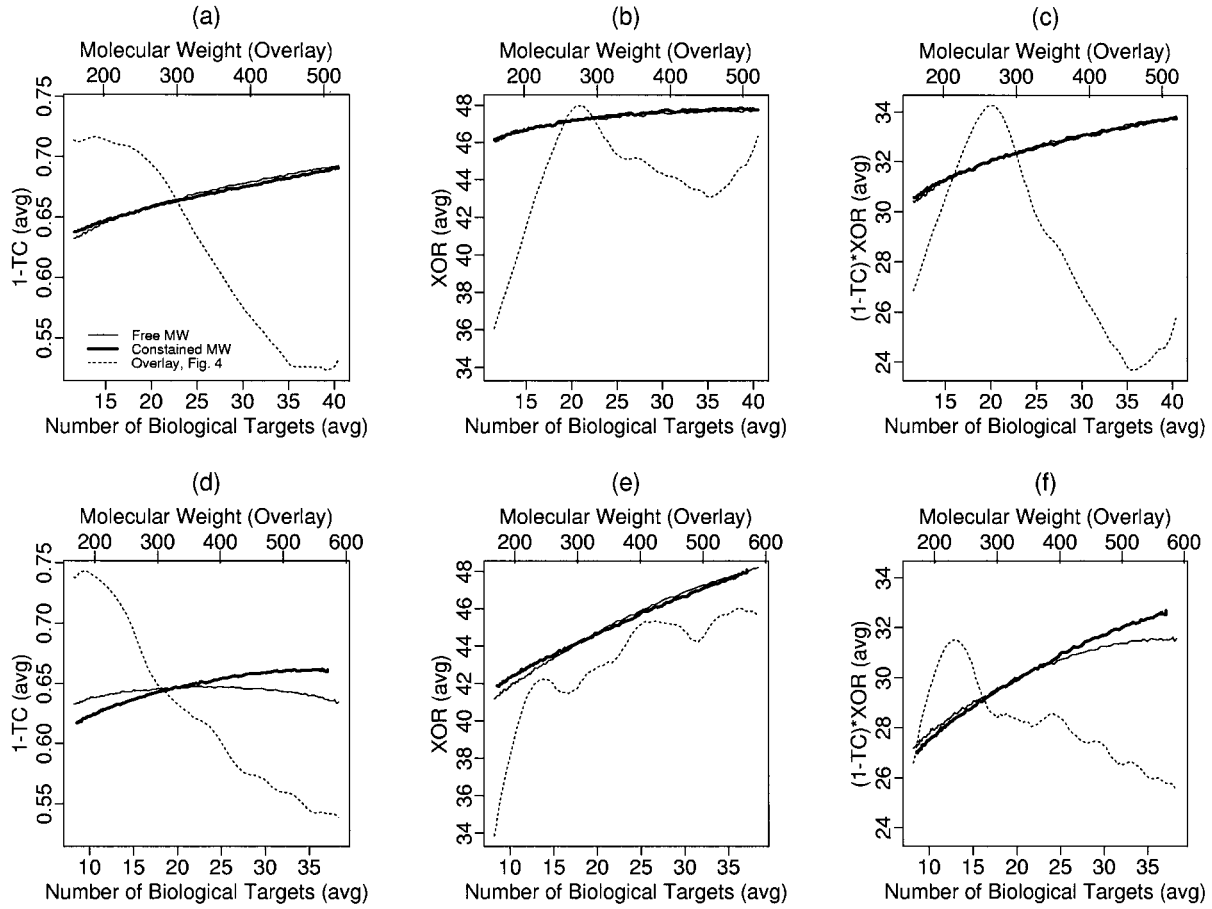

Figure 7. Results for MOLSKEYS. Within subsets of 50 compounds, the average distance between pairs of compounds is monitored as the number of biological targets covered is varied. Dotted curves show the corresponding distances that were obtained when molecular weight was varied (Figure 4): (a−c) RBI library; (d−f) CMC library.

placed on the molecular weight (Figure 8b). That this would occur in the unconstrained case is perhaps not surprising since RBI ligand size, hence the number of potential working bits, decreases somewhat with increasing numbers of targets (Figure 6a). It is surprising, however, that the downward drift in structural diversity persists, even when size changes are controlled. The underlying cause for this behavior is unclear, and the complexity and random characteristics of hashed fingerprints certainly make it difficult to speculate, especially since the distance drift corresponds to only about 1% of the overall fingerprint.

Effects of Varying Structural Diversity

The final set of experiments involved the use of designed structural diversity to probe size effects and their impact on bioactive properties. There are of course any number of ways to select diverse subsets of compounds,[4,7,17,28−31] and it is beyond the scope of this investigation to consider every technique. What is important here is not so much the method itself, but rather the consequences of using any sort of 2D diversity design to select compounds. Biases that are inherent in the descriptors and/or distance functions should have some impact on the results, regardless of which algorithm is utilized.

We have chosen a method which selects compounds that are distributed in an approximately uniform fashion throughout descriptor space. This *spread design* algorithm and its general properties have been detailed elsewhere,[32] and the reader is referred there for additional information. The basic goal is to select a subset of compounds **S**, in which the members are as far away as possible, on average, from their nearest neighbors within the subset. This is achieved by maximizing the following function of distance:

$$f_{\text{spread}} = \sum_{i \in \mathbf{S}} \text{MIN}\{\text{dist}_{ij}:\ j \in \mathbf{S}, j \neq i\} \qquad (8)$$

Here, dist_{ij} is the 2D distance between compounds i and j, and the MIN operation returns the distance between compound i and its nearest neighbor within the subset.

A stochastic procedure is used to maximize f_{spread}, wherein an initial subset of compounds is selected randomly, and then pairwise exchanges between the subset and the remainder of the library are made so as to increase the value of the function. At any step in the algorithm, the two compounds in **S** with the smallest pairwise distance are identified. Of the two, the one which is closer to some other compound in **S** is flagged for ejection. This flagged compound is exchanged for one that is outside of **S** if the exchange will bring about an overall increase in f_{spread}. Pairwise exchanges are continued until no further increase in the function value can be achieved. At this point, a new random subset may be selected and the process repeated, with the highest function value and associated compounds being retained at the end.

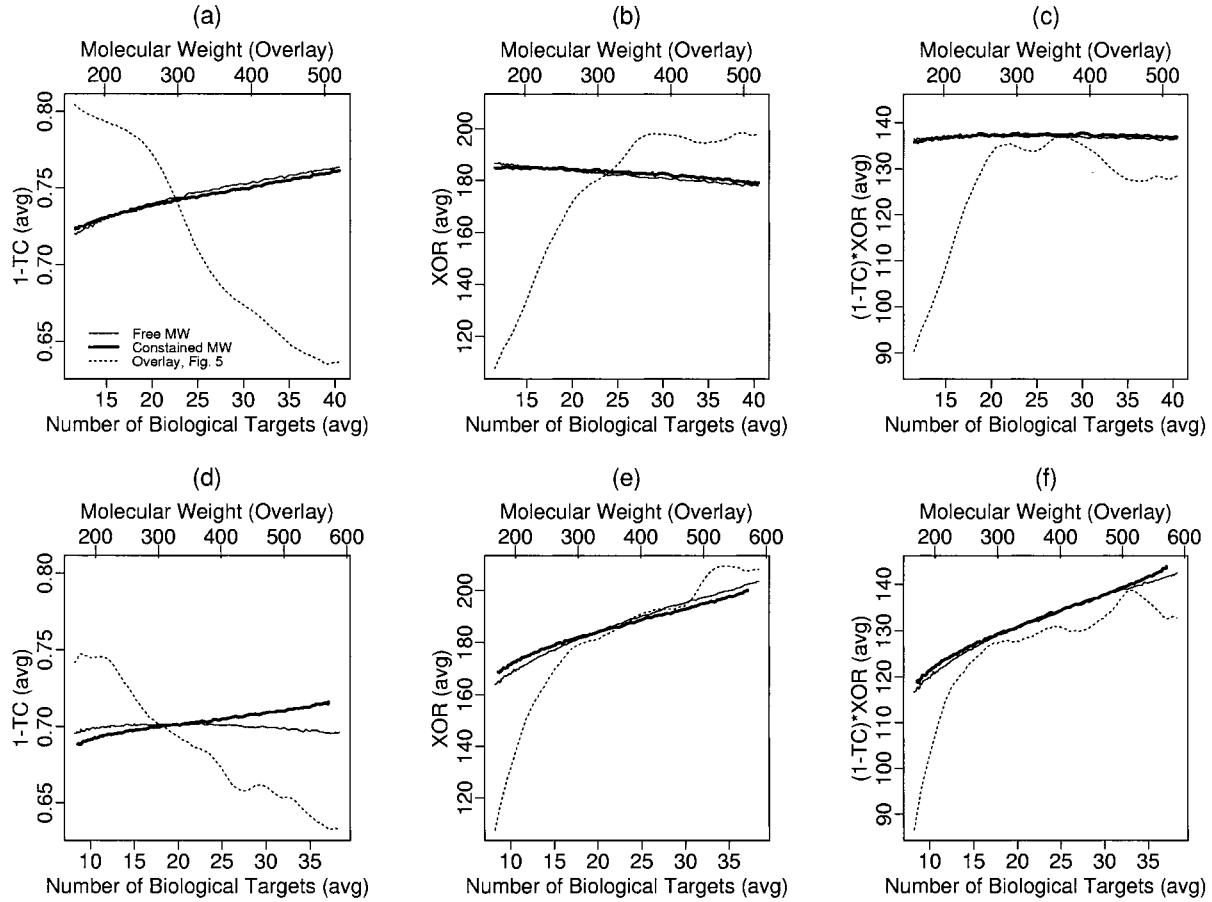

Figure 8. Results for Daylight hashed fingerprints. Within subsets of 50 compounds, the average distance between pairs of compounds is monitored as the number of biological targets covered is varied. Dotted curves show the corresponding distances that were obtained when molecular weight was varied (Figure 5): (a−c) RBI library; (d−f) CMC library.

Note that while we have used average pairwise distance to *measure* diversity throughout the course of this investigation, we are not using it to *design* diversity. This is simply because a function that relies on all pairs of distances typically results in what may be called an edge design.[31,32] Here, the compounds selected tend to lie near the outer boundaries or edges of descriptor space. Such an arrangement may be useful, depending upon the circumstances, but the goal in this investigation is to provide reasonable coverage of the overall descriptor space, and this is better accomplished by maximizing average nearest neighbor distance.

We would like to point out that the premise and characteristics of this algorithm are analogous to those of the well-known maximum dissimilarity spread design.[28,31] The current approach was chosen over others because its stochastic nature affords the possibility of a huge number of different diverse subsets. Maximum dissimilarity algorithms, by contrast, rely on the selection of a single seed compound, followed by a deterministic procedure for adding compounds to the subset.[28,31] Thus, the number of different diverse subsets possible is at most equal to the number of compounds in the overall library.

As an illustrative example of how the stochastic spread algorithm performs, we have included chemical structures for three diverse subsets of 10 compounds selected from the CMC library using the MOLSKEYS and 100 random restarts of the algorithm, Figure 9. The three types of distance functions were used to arrive at the three different subsets in Figure 9. Size effects are readily apparent, as the 1 − TC dissimilarities yield significantly smaller compounds than those of XOR. Average molecular weights for these two subsets are 207 and 378, respectively. Not surprisingly, the hybrid distance function selects intermediate-sized compounds with an average molecular weight of 292.

Table 3 compares average pairwise and nearest neighbor distances among compounds in the three spread-diverse subsets and 100000 randomly selected subsets of 10 compounds. The nearest neighbor of each compound was identified according to the distance function specified in the column heading, so the average nearest neighbor distances across any one row do not necessarily reflect the same pairs of compounds. Boldface entries are the largest values observed in each column, and they of course correspond to cases where the diversity algorithm utilized that particular distance function. What is most interesting here is the fact that the XOR spread design selects compounds which are structurally more diverse than random according to 1 − TC, but the reverse is not true. When compounds are selected purely on the basis of 1 − TC, pairwise and nearest neighbor distances in XOR space are actually shorter than those observed in the case of randomly selected compounds. This is no doubt a direct result of the bias in the sizes of the Tanimoto compounds. These small structures simply do not set enough bits to produce sufficiently large XOR distances.

To examine these sorts of phenomena in a more

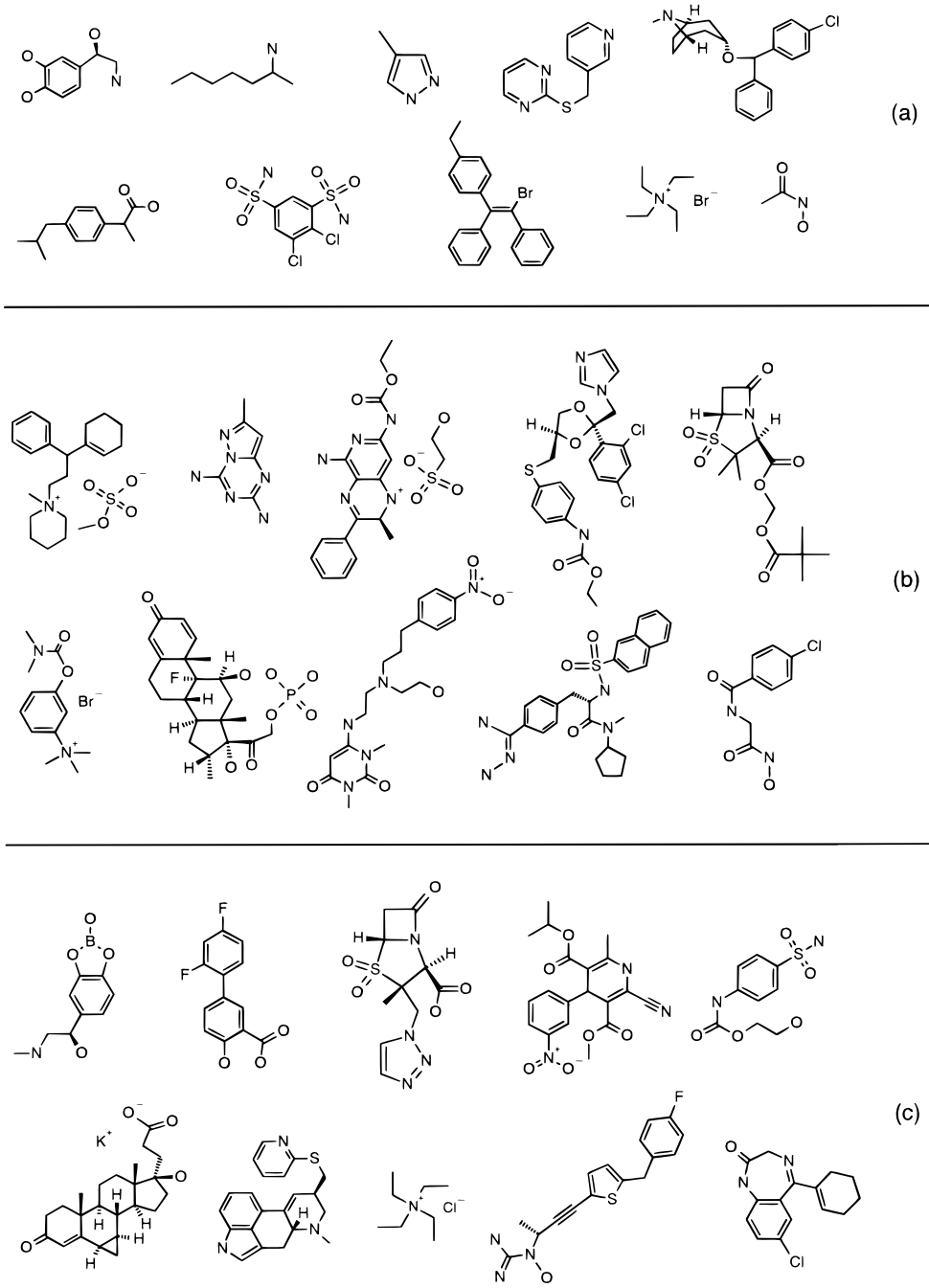

Figure 9. Compounds chosen from the CMC library with the spread design algorithm and different distance functions: (a) 1 − TC; (b) XOR; (c) (1 − TC)·XOR.

Table 3. Comparison of Distances for 10-Member Subsets Selected from the CMC Library Using MOLSKEYS Spread Designs and Random Sampling

subset selection method	average distances within subset (all pairs/nearest neighbors)		
	1 − TC	XOR	(1 − TC)·XOR
1 − TC	**0.863/0.793**	37.6/27.5	32.6/22.3
XOR	0.673/0.581	**59.0/53.2**	40.0/31.4
(1 − TC)·XOR	0.771/0.689	57.2/48.6	**44.1/35.3**
random	0.646/0.470	44.4/28.6	29.7/14.6

continuous fashion, we have established a sliding scale of structural diversity that allows us to directly monitor molecular weight and target coverage as a function of 2D diversity. First, a subset of 50 compounds was selected at random, then one complete pass of the spread algorithm was performed. This corresponds to just a single random restart, which is in contrast to the 100 random restarts used in the previous illustrative example. The compounds in the resulting diverse subset were then replaced, one-by-one, with compounds chosen at random from the complete library. Random selections were made both from within the 50-member subset and from outside it, so that there would be no bias in the replacements. Thus if the randomly chosen compound was already included in the subset, no action was taken, but the number of "diverse" compounds was still considered to drop by one.

With each replacement, the subset becomes progressively more like a random collection, and the overall structural diversity slowly drifts downward. During this

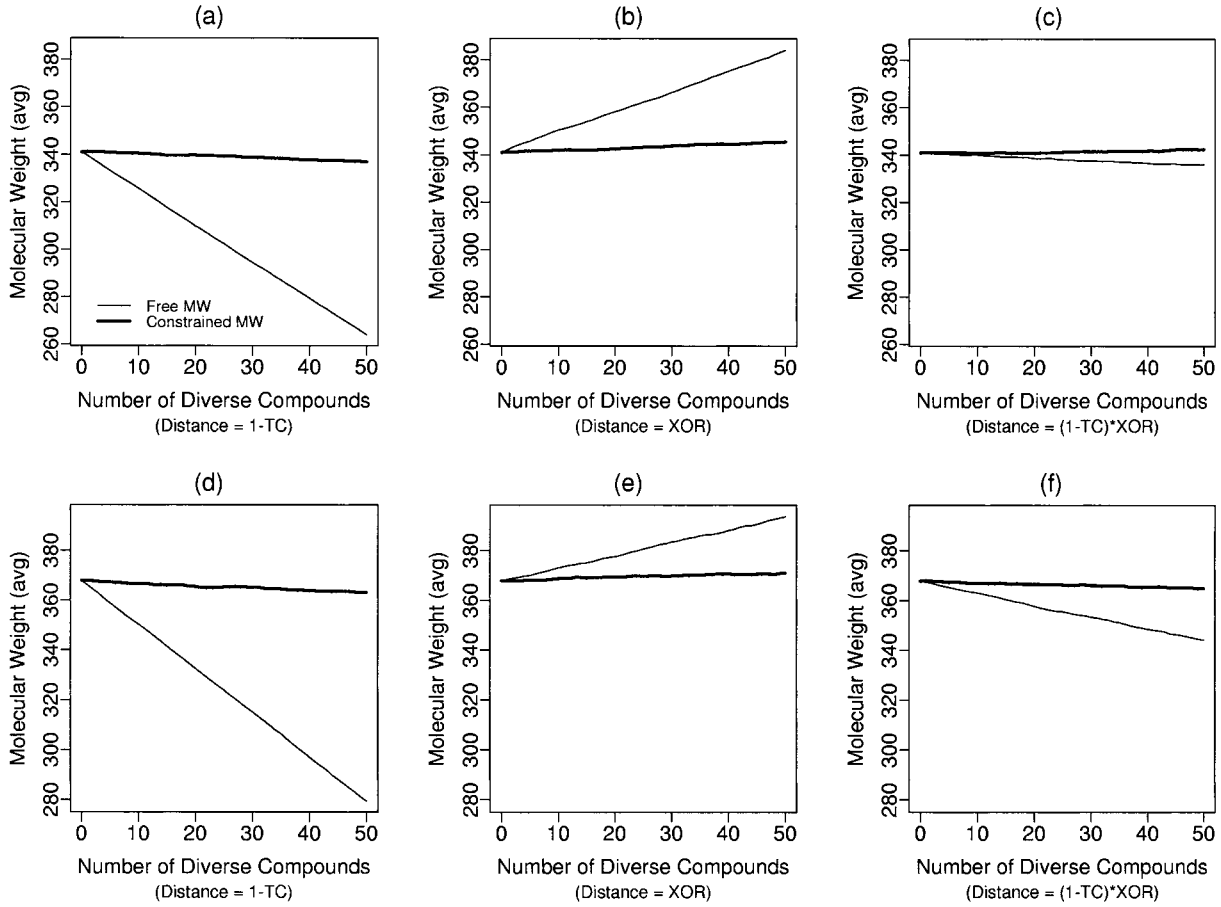

Figure 10. Results for MOLSKEYS. For subsets containing 50 compounds, the average molecular weight is monitored as the number of spread-diverse compounds within the subset is varied: (a−c) RBI library; (d−f) CMC library.

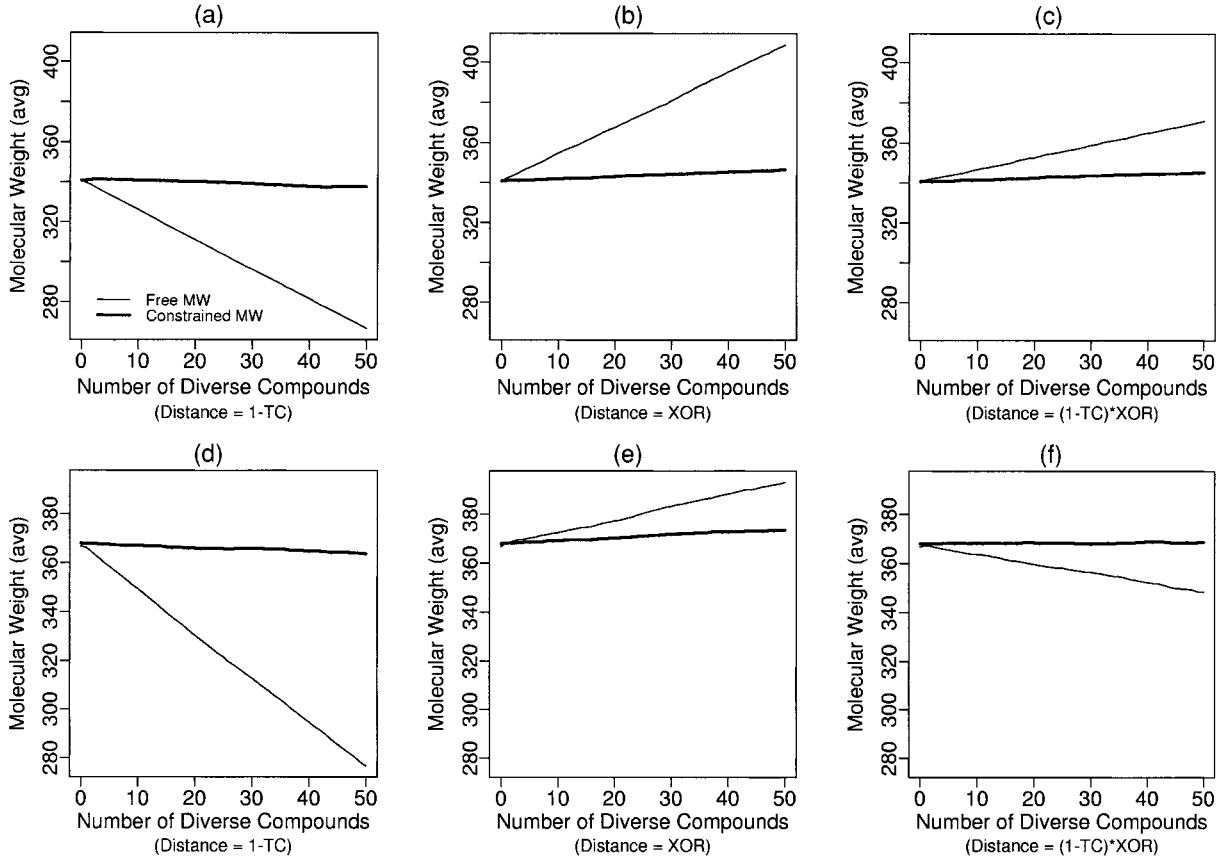

Figure 11. Results for Daylight hashed fingerprints. For subsets containing 50 compounds, the average molecular weight is monitored as the number of spread-diverse compounds within the subset is varied; (a−c) RBI library; (d−f) CMC library.

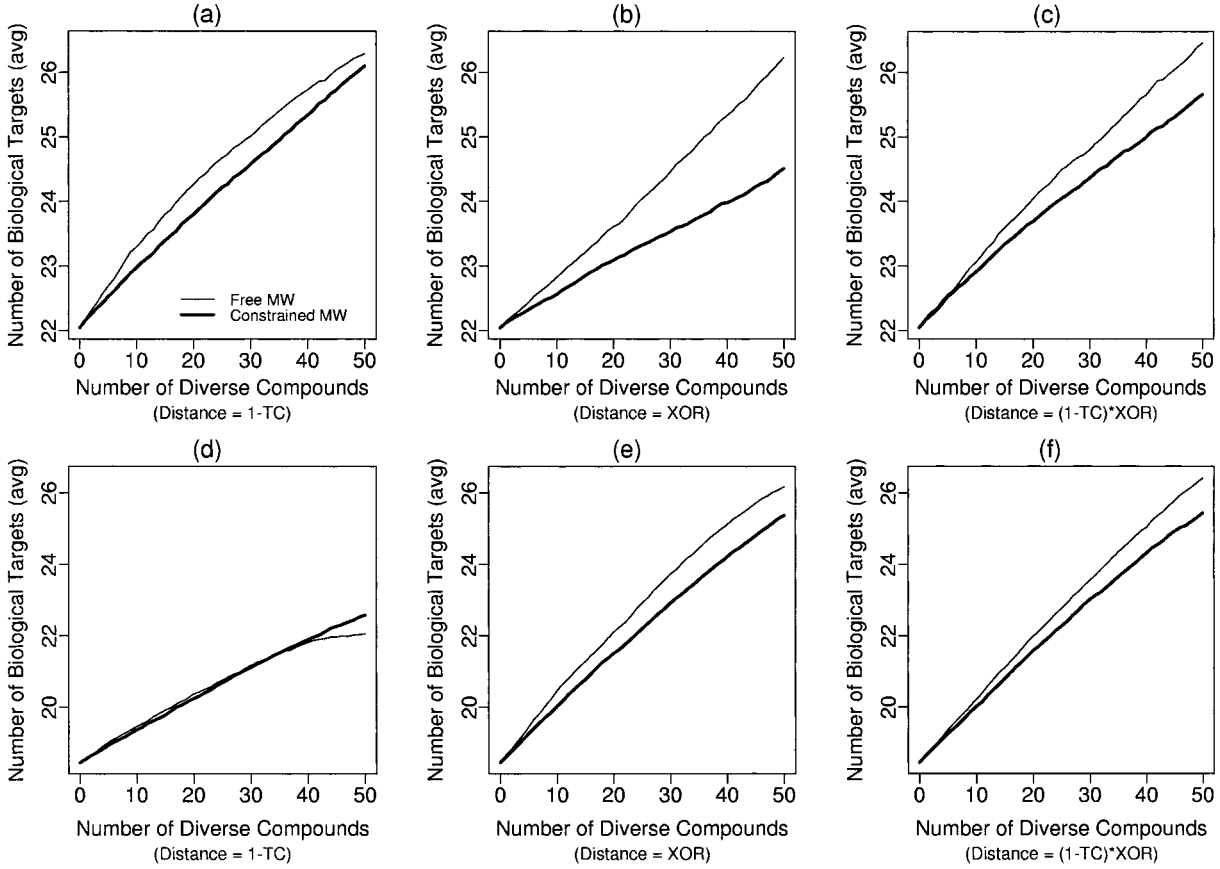

Figure 12. Results for MOLSKEYS. For subsets containing 50 compounds, the average number of biological targets covered is monitored as the number of spread-diverse compounds within the subset is varied: (a−c) RBI library; (d−f) CMC library.

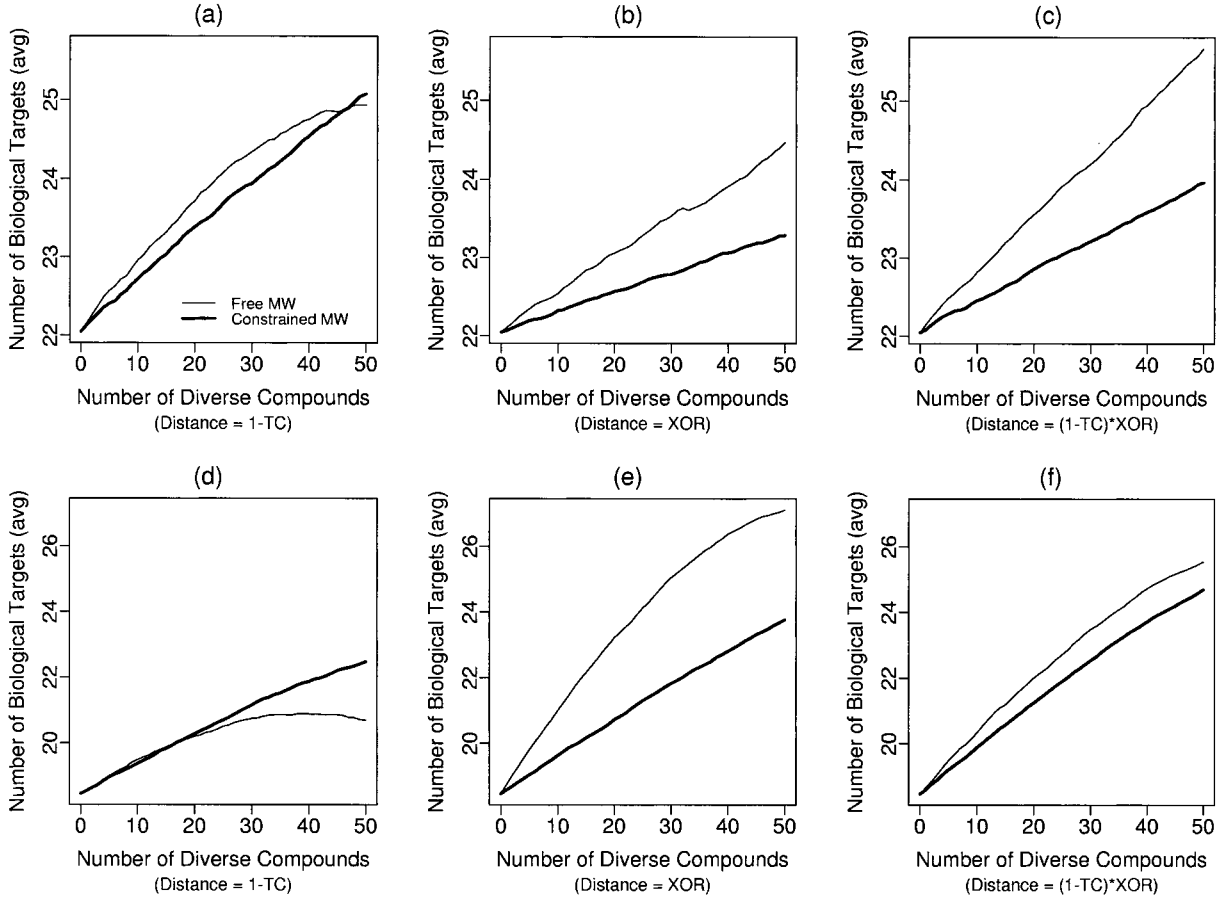

Figure 13. Results for Daylight hashed fingerprints. For subsets containing 50 compounds, the average number of biological targets covered is monitored as the number of spread-diverse compounds within the subset is varied: (a−c) RBI library; (d−f) CMC library.

process, we monitored the average molecular weight and the number of biological targets covered. To generate smooth statistics, the entire random replacement procedure was repeated 1000 times, with a different initial spread-diverse subset in each case. So results reported for, say, subsets with 40 diverse members are actually the average of 1000 different cases, wherein the 50-member spread-diverse subset has had 10 of its compounds replaced with random selections.

To investigate whether size biases could be controlled as compounds were selected via the spread algorithm, a procedure for controlling the average molecular weight was employed. This procedure was directly analogous to the one used when biological target counts were varied systematically. After selecting a random subset in the initial phase of the spread algorithm, each compound that was a candidate for exchange into the subset was required to have a molecular weight that would drive the subset average molecular weight toward that observed for the entire library. All sampling experiments in this section were performed with and without application of this molecular weight constraint.

Figure 10 summarizes for the MOLSKEYS the impact of changes in 2D diversity on ligand size. The horizontal scales in these plots range from zero diverse compounds, i.e., purely random, to 50 diverse compounds, i.e., the fully spread-diverse subset. When no size constraints are exercised in the spread algorithm, molecular weights associated with the three types of distance functions diverge sharply from each other in both the RBI and CMC libraries. Size effects are most pronounced when 1 − TC is used, with downward drifts of about 77 and 74 amu in the two libraries. When the size constraint is enforced, all fluctuations in molecular weight are reduced to less than 5 amu. Analogous results are seen for Daylight hashed fingerprints, Figure 11, where once again 1 − TC leads to the largest unconstrained drifts in molecular weight and enforcing the size constraint attenuates all fluctuations to less than 6 amu.

Figure 12 illustrates the corresponding effects on bioactive diversity when the MOLSKEYS are used to select 2D diversity. With molecular weight free to vary, the numbers of biological targets covered within the RBI collection (12a−c) increase at about the same rate, regardless of which distance function is employed. Enforcing the molecular weight constraint tends to decrease the amount of bioactive diversity accessible, though this reduction is quite small in the case of 1 − TC.

Very different behavior is seen in moving to the CMC library, Figure 12d−f. With unconstrained molecular weight, use of 1 − TC leads to only 3.6 additional targets being covered beyond random selection, which compares to 7.7 and 8.0 additional targets for the XOR and hybrid distance functions, respectively. Constraining molecular weight improves results slightly for 1 − TC, but it does not appear to remove all of the size-related biases that are associated with this distance function.

Results are similar for Daylight hashed fingerprints, Figure 13, except that the molecular weight constraint appears to have a slightly more positive impact on the 1 − TC results than was observed with the MOLSKEYS. Here, as the number of diverse compounds is increased, target counts for constrained sampling overtake those of unconstrained sampling in both the RBI and CMC libraries. It is interesting to note that within the CMC library, the number of biological targets covered by 1 − TC and unconstrained molecular weight actually peaks when only 39 structurally diverse compounds are present and then drops off slightly as more 2D diversity is added. With this method of sampling, fully spread-diverse subsets of 50 compounds cover, on average, only 2.2 more targets than randomly selected subsets of 50.

It is of course natural to wonder what the effects are of varying the number of compounds contained in each subset. In other words, what sorts of results may be expected when larger or smaller diverse subsets are chosen? We have found that the trends observed here using groups of 50 compounds are generally preserved, regardless of how many compounds are selected. However, the effects are most obvious when smaller subsets are analyzed, simply because the sampling characteristics of each distance function are more pronounced. Thus, for example, the bias toward low molecular weights that occurs with 1 − TC is increasingly amplified as smaller and smaller fractions of a library are sampled. In the case of the CMC collection, we have observed that when 1 − TC is used to select subsets of fewer than 40 compounds, the number of biological targets covered is frequently no more than that obtained with random subsets of the same size.

Conclusions

Within the space of 2D fragment descriptors, the method of calculating distance can have a significant impact on the perceived structural diversity of compound subsets, and on the sizes and properties of compounds that are selected using a 2D diversity design. Our experiments have placed a premium on biological target coverage, and we have used this measure of bioactive diversity as the relevant property with which 2D structural diversity should correlate.

Of the three distance functions tested here, 1 − TC shows the most direct and pronounced bias toward size, consistently measuring a higher level of 2D diversity among collections of compounds whose structures are far smaller than the average across the library from which they are sampled. There appear to be no obvious undesirable consequences with regard to bioactive diversity, so long as the library examined exhibits sufficiently wide target coverage among smaller compounds. If, however, target coverage is skewed more toward larger compounds, then 1 − TC appears to be a less reliable measure of bioactive diversity, and selecting compounds on the basis of this distance function is not recommended unless some mechanism is invoked to control average size. A straightforward procedure that constrains the average molecular weight of diverse subsets is observed to generally improve results for 1 − TC.

XOR distance, in contrast to 1 − TC, places a greater emphasis on large compounds when it comes to measuring structural diversity. This bias, however, does not seem to result in as much variability in performance between libraries as observed with 1 − TC. It should be noted that if target coverage within a library were sufficiently biased toward small compounds, then XOR distance might show the same sort of degradation in

performance that is seen with 1 − TC in libraries biased toward large compounds.

References

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

(2) Martin, Y. C.; Brown, R. D.; Bures, M. G. Quantifying Diversity. In *Combinatorial Chemistry and Molecular Diversity*; Kerwin, J. F., Gordon, E. M., Eds; Wiley: New York, 1998; pp 369−385.

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

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

(5) 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.

(6) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(7) Lajiness, M. S. Dissimilarity-based Compound Selection Techniques. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 65−84.

(8) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.

(9) 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.

(10) 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.

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

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

(13) 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.

(14) Nys, G. G.; Rekker, R. F. Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. The Introduction of Hydrophobic Fragmental Constants (f Values). *Chim. Ther.* **1973**, *8*, 521−529.

(15) Klopman, G. Artificial Intelligence Approach to Structure−Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *J. Am. Chem. Soc.* **1984**, *106*, 7315−7321.

(16) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(17) Holliday, J. D.; Ranade, S. S.; Willet, P. A Fast Algorithm for Selecting Sets of Dissimilar Molecules from Large Chemical Databases. *Quant. Struct.-Act. Relat.* **1995**, *14*, 501−506.

(18) RBI HTS Library (LOPAC), Catalog No. SC001, Research Biochemicals, International, 1 Strathmore Rd., Natick, MA 01760-2447.

(19) RBI 1997/1998 Catalog, Research Biochemicals International, 1 Strathmore Rd., Natick, MA 01760-2447.

(20) *Goodman and Gilman's The Pharmacological Basis of Therapeutics*; Gilman, A. G., Goodman, L. S., Rall, T. W., Murad, F., Eds.; MacMillan Publishing Company: New York, 1985.

(21) *The Merk Index*, 12th ed.; Budavari, S., Ed.; Merck & Co., Inc.: Whitehouse Station, NJ, 1996.

(22) Nogrady, T. *Medicinal Chemistry*; Oxford University Press: New York, 1988.

(23) MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

(24) *ISIS/Base 2.1.4*; MDL Information Systems, Inc., San Leandro, CA.

(25) *MACCS−II Menu Reference Version 2.2*; MDL Information Systems, Inc., San Leandro, CA, 1994.

(26) *Daylight 4.51*; Daylight Chemical Information Systems, Inc., 27401 Los Altos, Suite 370, Mission Viejo, CA 92691.

(27) *Daylight Programs Reference Manual*; Daylight Chemical Information Systems, Inc., Mission Viejo, CA, 1997.

(28) Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. Comparison of Algorithms for Dissimilarity-based Compound Selection. *J. Mol. Graphics Mod.* **1997**, *15*, 372−385.

(29) Willett, P. Computational Tools for the Analysis of Molecular Diversity. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 1−11.

(30) Agrafiotis, D. K. Stochastic Algorithms for Maximizing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 841−851.

(31) 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.

(32) Dixon, S. L.; Villar, H. O. Bioactive Diversity and Screening Library Selection via Affinity Fingerprinting. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192−1203.

JM980708C