Expedited Articles

Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors

David E. Patterson,* Richard D. Cramer, Allan M. Ferguson, Robert D. Clark, and Laurence E. Weinberger†

Tripos, Inc., 1699 South Hanley Road, St. Louis, Missouri 63144, and 882 South Matlack Street, West Chester, Pennsylvania 19380

Received April 18, 1996⊗

When searching for new leads, testing molecules that are too "similar" is wasteful, but when investigating a lead, testing molecules that are "similar" to the lead is efficient. Two questions then arise. Which are the molecular descriptors that should be "similar"? How much "similarity" is enough? These questions are answered by demonstrating that, if a molecular descriptor is to be a valid and useful measure of "similarity" in drug discovery, a plot of differences in its values vs differences in biological activities for a set of related molecules will exhibit a characteristic trapezoidal distribution enhancement, revealing a "neighborhood behavior" for the descriptor. Applying this finding to 20 datasets allows 11 molecular diversity descriptors to be ranked by their validity for compound library design. In order of increasing frequency of usefulness, these are random numbers = log *P* = MR = strain energy < connectivity indices < 2D fingerprints (whole molecule) = atom pairs = autocorrelation indices < steric CoMFA fields = 2D fingerprints (side chain only) = H-bonding CoMFA fields.

Introduction

The very existence of medicinal chemistry teaches that similar molecules will tend to have similar biological properties. Were this not so, the major activity of medicinal chemists, "lead optimization" by the synthesis of individual compounds similar to a lead structure, would be futile! With the increasing use of combinatorial approaches to rapidly synthesize many compounds in parallel, a newer phrase "molecular diversity" has become fashionable. But of course, diversity and similarity are simply opposite sides of the same coin.

More importantly, similarity/diversity are meaningful qualities only with respect to specified molecular descriptors. For example, a synthetic chemist will regard two molecules as similar when their topological descriptions—their networks of atoms and connecting bonds—contain a sufficiently large number of features in common. But, with increasing knowledge of receptor−ligand complex structures, it seems plausible that similarity of pharmacophoric descriptors, critical geometric arrangements of structural features, may be a more important determinant than topology of whether two molecules will bind to the same receptors. And there are very many other properties, such as overall molecular size and lipophilicity, which affect biological potency and may, therefore, be useful diversity descriptors.

Indeed, one can think of the whole process of drug discovery as one of searching through all such possible descriptor dimensions of molecular similarity/diversity. In "lead discovery", the search objective is discovery of an "activity island", a volume whose dimensions correspond to various molecular descriptors, which contains a high frequency of active molecules. In "lead optimization", the two objectives of choosing a clinical candidate or two and constructing a patent are achieved by detailed exploration of a known "activity island".

Starting with this activity island analogy, Figure 1 suggests basic approaches for lead discovery and lead optimization.[1] To keep matters simple, Figure 1 assumes that "molecular diversity space" has only two descriptor dimensions that determine biological properties, its *x*- and *y*-axes. Each star in Figure 1 represents a compound available for testing, located at the *x*,*y* coordinate determined by its structural descriptors. If we further assume that similar compounds in this space tend to have similar biological properties, then the test results for any one compound also provide information about the biological properties of nearby compounds. We can speak of a "neighborhood region" for every molecule within this descriptor space, depicted in Figure 1 by the circles surrounding each star.

In *lead discovery* research,[2] it is inefficient to test compounds whose neighborhood regions overlap, as suggested by the "typical compound library" of the top panel in Figure 1. Each biological test is a probe of the unknown that consumes some resources; even though the "neighborhood behavior" of any diversity property must be probabilistic rather than deterministic (the circles fuzzy rather than hard-edged), if we know that any particular compound-star is inactive, it surely represents a better use of those resources to seek activity from compound-stars that lie outside, rather than inside, that inactive compound's neighborhood region. Conversely, in *lead exploration*, if one discovers that a particular compound-star is active, the same concept of a neighborhood region can be used to efficiently plan a survey of the prospective activity island (shown in Figure 1 as a light gray cloud). This initial phase of lead exploration can be done so rapidly by combinatorial approaches that it is starting to be called "lead explosion", in distinction from the latter phase of

* To whom correspondence should be addressed.
† West Chester, PA.
⊗ Abstract published in *Advance ACS Abstracts,* July 15, 1996.

S0022-2623(96)00290-7 CCC: $12.00

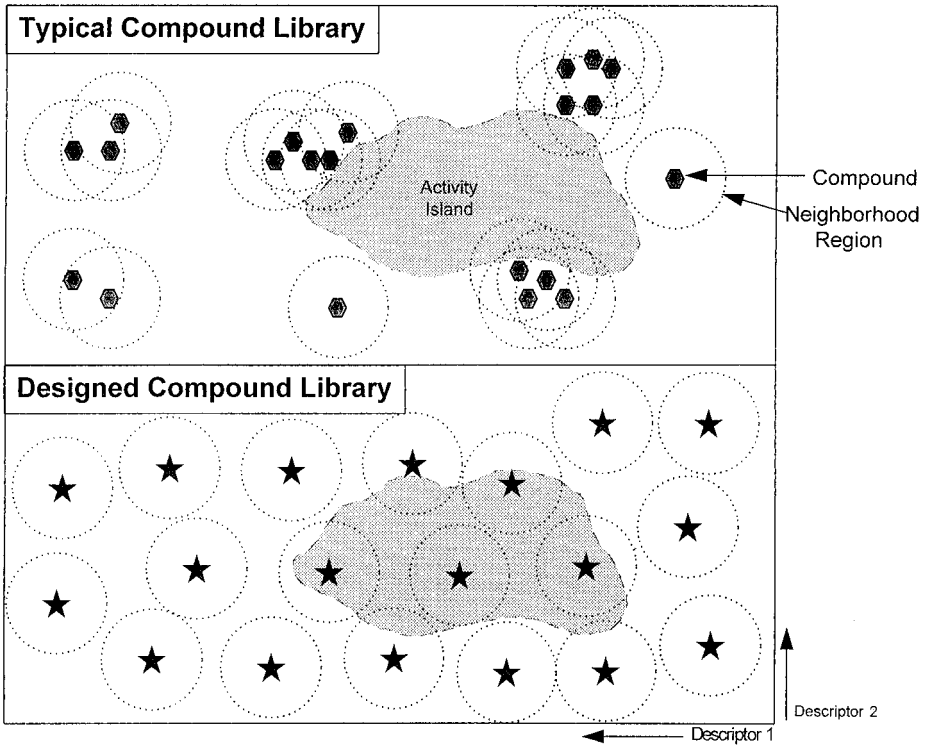


Figure 1. Schematic plots comparing two screening libraries. The two dimensions of each plot are arbitrary diversity descriptors which do exhibit "neighborhood behavior". Therefore the testing of a compound, represented by a star or a hexagon at the middle of a circle, provides probabilistic information about the test results that would be obtained for any other compound lying within that circle. Typical libraries, represented by the top plot, will have many compounds lying within one another's "neighborhood regions". When seeking a new "activity island", represented schematically by the cloud, the chances of success are much better if all the compounds to be tested are spaced as shown in the bottom plot. Conversely, in the earliest stages of lead optimization, the chances of finding other active molecules are greatest for compounds (not shown) which lie within the neighborhood region of the initial lead molecule.

"lead optimization", where ample structure−activity data (SAR) allow QSAR to identify the optimal peak on an activity island.

But the analogy between drug discovery and geographic exploration cannot yet be pushed too far. The geographic descriptors latitude and longitude are useful largely because they do exhibit "neighborhood behavior"—points which are similar in latitude and longitude are similar in other properties such as elevation (the defining descriptor for a geographic island, of course!). What, if any, dimensions of molecular diversity have a "neighborhood behavior" general enough to be useful in biological screening strategies? How large are their "neighborhood regions"? How fuzzy, or diffuse, are the activity islands that may be observed and described within any proposed molecular diversity descriptor space?

To answer these questions, we present a general method for validating molecular diversity descriptors.[3] A key underlying feature of this validation method is explicitly considering absolute differences in, rather than magnitudes of, properties, as both the desired (dependent, biological) and known (independent, molecular diversity) parameters. If such differences are plotted, the data points will tend to concentrate in a characteristically trapezoidal "neighborhood enhancement" whenever the molecular diversity descriptor has the requisite neighborhood behavior.

By applying this validation method (reduced to a particular algorithmic implementation) to 20 datasets taken randomly from the recent literature, we rank 11 molecular diversity measurements currently in use. Brown, Bures, and Martin have also recently used a somewhat different approach to successfully rank several "fingerprint" diversity descriptors,[4] obtaining results consistent with ours.

Methods

All work was done within SYBYL 6.2, enhanced with both SPL and C codes to perform some of the specialized calculations described below.

Descriptor Validation. The data flow for the validation of a diversity descriptor is shown in Table 1 for the (novel) topomeric steric field descriptor. For each of the $n(n-1)/2$ pairings of the n structural variations within a particular dataset (top panel of Table 1), an *absolute* difference (distance) between the diversity descriptors for the two molecules and an *absolute* difference (distance) in the $\log_{10}$s of the two biological activities were recorded (bottom panel of Table 1). More exactly, if the candidate diversity descriptor is a single scalar number (CLOGP, CMR, random, heat of formation/atom), the descriptor difference is simply the absolute value of the arithmetic difference. If the descriptor is a vector or field (connectivity indices, 2D autocorrelation, topomeric steric fields), then the difference is the square root of the sum of the squared differences between corresponding elements. If the descriptor is a set (2D fingerprint, atom pair, topomeric hydrogen bonding), then the difference is the complement of the Tanimoto coefficient for those two sets, defined as: 1 − (no. of bits = 1 in both sets)/(no. of bits = 1 in either set).

Any group of individual diversity descriptors can be combined into one composite descriptor by analogy to Euclidean distance, where composite distance = sqrt(weighted sum of squares of individual descriptors) and the weights—as in any distance-based algorithm—can have a major effect on the results. Often for a vector of scalar descriptors, one autoscales to give each descriptor an equal contribution to distance by

Table 1. Calculation Flow for Validation of a Diversity Descriptor, Showing Data from Row 9 of Table 2A[a]

A. Master Table

	biological activity[b]	strain energy		biological activity[b]	strain energy
1 Doh1	2200.00	2.94	4 Doh4	1200.00	2.42
2 Doh2	1300.00	2.92	5 Doh5	560.00	2.97
3 Doh3	50.00	2.37	6 Doh6	12.00	2.15

B. Scratch Table

	biological distance[b]	strain energy distance		biological distance[b]	strain energy distance
1 2vs1	0.23	0.02	9 5vs3	1.05	0.60
2 3vs1	1.64	0.57	10 5vs4	0.33	0.55
3 3vs2	1.41	0.55	11 6vs1	2.26	0.79
4 4vs1	0.26	0.52	12 6vs2	0.78	0.77
5 4vs2	0.03	0.50	13 6vs3	0.62	0.22
6 4vs3	1.38	0.05	14 6vs4	2.00	0.27
7 5vs1	0.59	0.03	15 6vs5	1.67	0.82
8 5vs2	0.37	0.05			

a Compound distances for each pair of rows in the Master Table become a row in the Scratch Table. A neighborhood plot shows any relationship between the two columns in the Scratch Table. Note in the Scratch Table that all distances are unsigned magnitudes. [b] The distances in biological activities shown in the Scratch Table are the absolute differences in the logarithms of the values in the Master Table.

dividing the descriptor by its observed standard deviation. To ignore scaling is to leave oneself at the mercy of the units and to accept different results if molecular weight is in grams/mole or ounces/mole.

In general, we have analyzed such sets of $n(n-1)/2$ pairwise differences in molecular diversity descriptor and differences in biological activity in two ways:

(1) Visual inspection of 2D scatter plots ("neighborhood plots"), which for a valid diversity descriptor should exhibit a "neighborhood enhancement", that is, a very low frequency of data points in some "upper left triangle" (ULT) region of the scatter plot. *(The Discussion section below starts with explanations of why such a sparsely populated ULT is both expected and also indicative of a neighborhood behavior.)* Many examples of such plots for a valid and an invalid molecular diversity descriptor appear in Figure 2.

(2) Identifying an "optimal" diagonal and comparing the average density of points below that optimal diagonal (within the "lower right trapezoid" or LRT) and within the entire data space with the average density over the entire data space (details below). Here average density is simply the number of points within an area divided by that area, and the entire data space is the rectangular area whose lowest right coordinate is [0,0] and whose upper-right coordinate is [$MAX_{\text{molecular property difference}}$, $MAX_{\text{biological property difference}}$]. A molecular diversity property that is valid for the dataset under consideration should therefore enhance the density of points in some lower right trapezoid, relative to the mean overall point density.

The statistical significance of any enhanced density of points within an LRT can be evaluated using the χ^2 statistic to compare the number of points actually found in the LRT with those that would be found in that area if the true distribution of points is uniformly random.[5] For one degree of freedom, the $P > 0.95$ level for χ^2 is 3.84, so we take any χ^2 value greater than 3.84 for any dataset−descriptor combination as a statistically significant validation of that descriptor with that dataset.

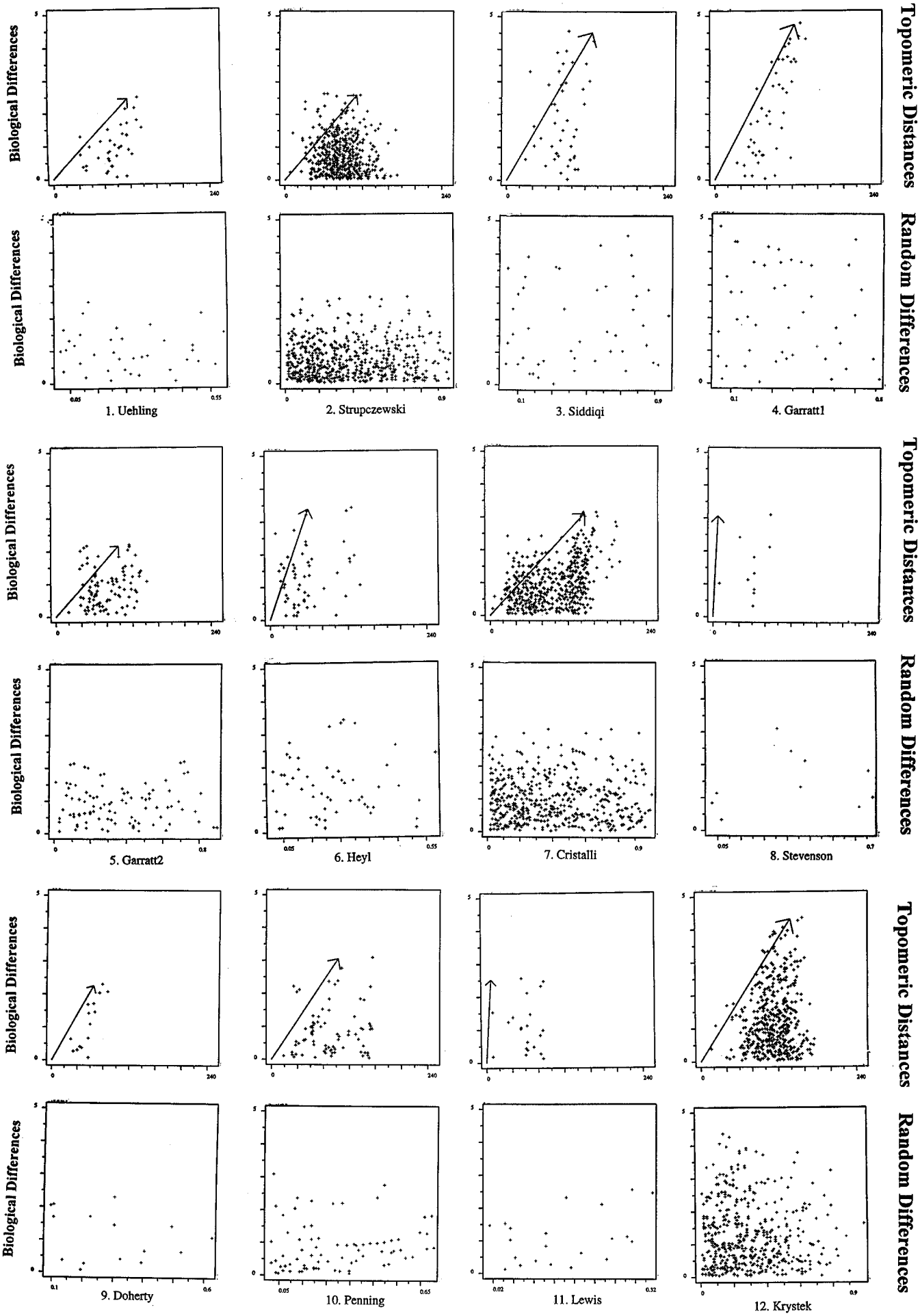
Identification of the Optimal Diagonal. This algorithm begins with the observation that the maximum number of candidates for an optimal diagonal is equal to the number of unique data points in the dataset. Thus each data point in turn is used to generate a candidate diagonal, defined as the line from [0,0] to [x,$MAX_{\text{biological property difference}}$].[6] The diagonal which yields the highest density of points within the triangle bounded by the points [0,0]; [x,0]; [x,$MAX_{\text{biological property difference}}$] is defined as the optimal diagonal for that dataset−descriptor combination. (Thus the algorithm cannot generate a diagonal whose slope is less than $MAX_{\text{diversity descriptor difference}}/MAX_{\text{biological property difference}}$.) The heavy arrow in Figure 3 shows how a particular data point (in black) is used to propose a candidate diagonal, with the shaded area being one of the triangles whose maximal point density defines the optimal diagonal. Note, however, that neighborhood enhancements consider the *entire* area to the right of a diagonal as the LRT. The enhancement then is the ratio of the point density within that LRT to the point density within the whole, where "the whole" is the rectangle with opposite vertices [0,0]; [$MAX_{\text{diversity descriptor difference}}/MAX_{\text{biological property difference}}$]. For these particular definitions, the maximum observable neighborhood enhancement value will be 2.0, resulting from the diagonal bisecting the whole rectangle with no points above it.

Establishing the Neighborhood Distance for a Valid Molecular Diversity Descriptor. The "neighborhood region" concept, as discussed in the Introduction and depicted in Figure 1, implies that any pair of compounds in a screening database should be separated by at least a "neighborhood distance", such that the (in)activity of each member of the pair would not be predicted from the (in)activity of the other. Here we describe how the value of the neighborhood distance for any descriptor is determined from a set of neighborhood plots. Of course, the detection of any particular biological activity depends on the sensitivity of the assay and the administered compound concentration. Typically, however, we suppose an activity island containing a compound having nanomolar potency is being sought by screening compounds at micromolar concentrations, with a safety factor of 10. The resulting "biological potency detection radius" of $(10^{-6}/10^{-9}) \times 10$ or 10^{-2} corresponds to a spacing of 2.0 log units along the y-axis (biology difference) on any of the graphs in Figure 2. The neighborhood distance for a (valid) molecular diversity descriptor can then be obtained by drawing a line from the point [0,2] parallel to the x-axis until the diagonal is intersected and reading off the x coordinate of that intersection. A pair of dashed arrows in Figure 3, beginning at the [0,2.0] coordinate and pointing first toward the diagonal and then to the x-axis, shows graphically how the neighborhood distance is defined. In practice this value is calculated as 2.0/(slope of the diagonal).

Dataset Selection. Twenty datasets were chosen by randomly scanning the recent literature, requiring only that (a) the reported potencies must span at least 2 orders of magnitude, (b) the structural variation must be "monovalent" and contain no prochiral atoms, and (c) no page turning was needed to cross-reference structure and activity data (to help make data entry as reliable and facile as possible). Each dataset−descriptor combination was validated independently. The graphs of Figure 2 indicate the distribution of intercompound distances by the different metrics for each dataset; clearly there is more variability in the Chang topomeric distances than in the Thompson set. Further, no truly inactive compounds were included since the method requires quantitative distances; an active−inactive pairwise comparison would have only a lower bound on the actual biological distance, and an inactive−inactive pairwise comparison could not be assigned a difference at all.

Because the behavior of some diversity descriptors depends on whether the isolated side chain or the whole molecule is considered, the common core and varying side chains were entered separately and joined only where appropriate. Both core and side chains were entered by typing in their Sybyl line notations (SLNs), the former with −Br replacing the side chain and the latter with −SH replacing the core. (The added −SH also provided two additional atoms needed for orientation of 3D models of the isolated side chains.) All structures were checked by visually inspecting their 3D or 2D structures, respectively.

Molecular Descriptors. The molecular descriptors to be validated for these 20 datasets were generated, stored, and manipulated within Molecular Spreadsheets. The negative control descriptor, heats of formation per atom, is the Tripos force field energy[7] for the topomeric conformation of the side chain only, divided by the number of atoms. Another negative


Topomeric Distances
Random Differences
Biological Differences
1. Uehling
2. Strupczewski
3. Siddiqi
4. Garratt1
5. Garratt2
6. Heyl
7. Cristalli
8. Stevenson
9. Doherty
10. Penning
11. Lewis
12. Krystek

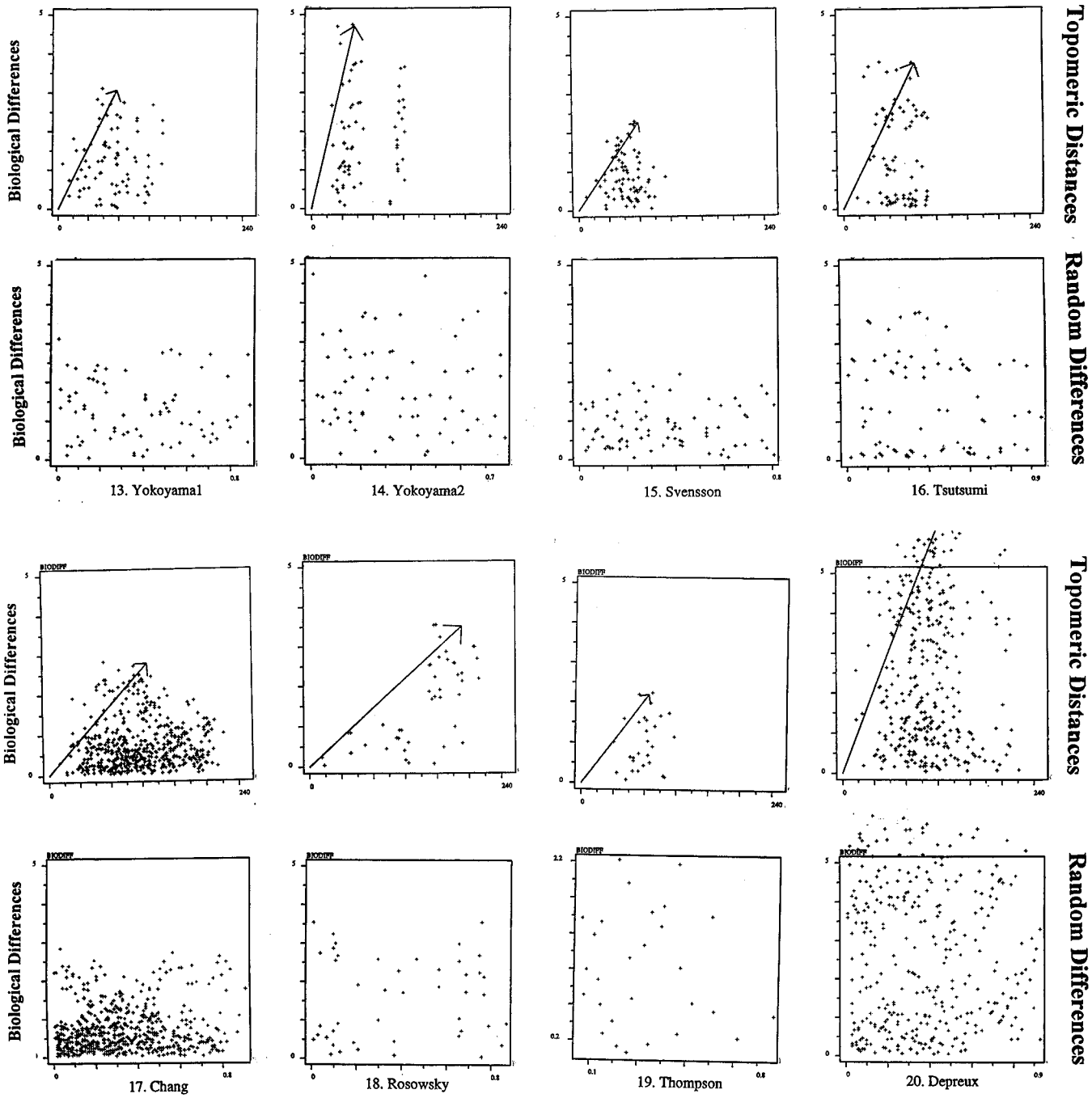


Figure 2. Neighborhood plots, comparing a valid with an invalid diversity descriptor, for 20 randomly selected datasets listed in Table 2A. The plot for the valid "topomeric steric" descriptor is the upper member of each pair; the plot for the invalid "random number" is the lower of the pair. The diagonal line in each upper graph separates the upper left triangle (ULT) of low point density from the lower right trapezoid (LRT) of high point density for that neighborhood plot. The absence of any such ULT/LRT density differentiation within the lower neighborhood plots indicates that the descriptor does not exhibit a "neighborhood behavior".

control, the random number "descriptor", was provided for each molecule by SGI's standard *rand()* generator. 2D "fingerprints", whether for the side chain with −SH attached or for the complete molecule, were generated for all paths of length two to six atoms, excluding hydrogen.[8] Leo/Hansch CLOGP and CMR values were generated, for side chains only, using the standard Pomona software;[9] numerous structures having too low a confidence level were omitted from the validation of these descriptors, and so validation could not be tried for dataset 13 with CLOGP. Kier/Hall connectivity indices, calculated for the whole molecule with HDISQ version 5.0 from Health Design Inc., include sets of four path-type CIs of order 0−6 and cluster-type CIs of order 3−5, each set of four containing an all-sp^3-carbon skeleton, a valence-adjusted skeleton, and the sum and difference of these two skeletons.[10] Connectivity index calculations failed for datasets 6 and 17 (because the whole molecule or the valence of an atom, respectively, was too large), so these were not validated. Carhart's atom pair descriptor[11] is a set, each of whose members corresponds to the occurrence of a specified number of bonds separating atoms of specified classes. In this work the 15 atom classes specified, in SYBYL atom type notation, were C.ar, C.2, C.3, all other carbon, O.2, O.3, all other oxygen, N.2, N.3, N.Ar, all other nitrogen, phosphorus, sulfur, halogen, and all other atoms except hydrogen, and the maximum number of bonds separating such atoms was 10, with atoms separated by more than 10 bonds being specified as 10. Moreau's 2D autocorrelation descriptor[12] is a vector of sums of products, the products being taken between specified atomic properties of pairs of atoms and the sums being over all pairs of non-hydrogen atoms separated by some number of bonds, including "self-pairs" of each atom with itself. In this work,

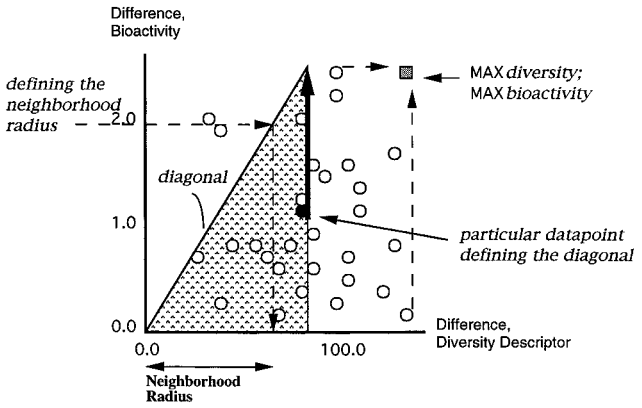


Figure 3. Construction and usage of the diagonal in a neighborhood plot. Each point in the plot is used to construct a trial diagonal, in this instance the point represented by the dark circle. The "optimal diagonal" is the one whose triangle (here shaded) contains the maximal point density. The neighborhood enhancement ratio will then be the ratio of the point density within the LRT (shaded area plus the rectangle to its right, partially bounded by the dashed arrows). The neighborhood distance is defined by assuming a particular sensitivity of the biological assay (2.0 log units here and in practice) and, following the dashed arrows, moving over to the diagonal and then down to where the neighborhood distance is read from the graph's *x*-axis.

eight atomic properties were distinguished as follows: the count of non-hydrogen-attached atoms, the VDW radius, the Pauling electronegativity, the number of lone pairs, and four binary variables indicating whether or not the atom is unsaturated, hydrogen-bond accepting, hydrogen-bond donating, or non-carbon. The maximum number of intervening bonds considered was 15. Methods for generating the topomeric conformation, from which the steric and a particular variety of hydrogen-bonding molecular fields were calculated, are described in an accompanying paper,[13] as are some minor modifications to the usual method of calculating a comparative molecular field analysis (CoMFA) steric field.

In this work, the hydrogen-bonding molecular field comprises two bitsets, accepting and donating, each having one bit for each lattice intersection. A bit is set to "1" if it is sufficiently close[14] to the position where a receptor atom would optimally be located to form a hydrogen bond with a ligand atom. The definitions of these loci for hydrogen-bonding and -accepting receptor atoms were taken from DISCO.[15]

Results

Our most important result is that significant neighborhood enhancements occur frequently for some, but not all, molecular diversity descriptors. The 20 pairs of graphs in Figure 2 show this result. The top members of each pair are the neighborhood plots from the 20 randomly chosen datasets for a diversity descriptor that is very frequently valid, topomeric steric CoMFA fields. Most of these clearly exhibit a neighborhood behavior, with the lower right trapezoid defined by the diagonal arrow having a density of points much higher than the density in the upper left triangle. Conversely, the bottom members of each pair are neighborhood plots from the same datasets for an invalid diversity descriptor, a random number associated with each molecule. None of the bottom 20 graphs displays any hint of a neighborhood behavior.[16]

Which molecular descriptors most frequently display a neighborhood property? Table 2B,C lists the neighborhood enhancement ratios for 11 diversity descriptors. The number in parentheses following each tabulated value is the corresponding χ^2 statistic. Any value of χ^2 greater than 3.84 indicates a significantly nonrandom enhancement, for the particular combination of dataset and descriptor, at the $P > 0.95$ confidence level.

The six descriptors in Table 2B all exhibited some useful frequencies of "neighborhood behavior". The first three of these, the 2D fingerprints of side chains and the two novel topomeric fields, exhibit significant χ^2 in 43/51 or 85% of the possible cases and mean neighborhood enhancements of 1.59, 1.40, and 1.47, respectively. These three descriptors are distinctly superior to the latter three in Table 2B, 2D fingerprints of whole molecules, "atom pairs", and autocorrelation vectors, which have significant χ^2 in 20/60 or 33% of cases and mean neighborhood enhancements of 1.21, 1.15, and 1.13.

The five descriptors in Table 2C exhibited no significant neighborhood behaviors, with only one marginally significant χ^2 associated with a favorable neighborhood enhancement among the 93 results shown, a frequency even lower than would be expected by chance at the $P > 0.95$ confidence level. Thus the four molecular descriptors in Table 2C, connectivity indices, partition coefficient, molar refractivity, and force field strain energy, useful as these can be in lead optimization, appear quite useless for the design of general lead discovery or "lead explosion" experiments, at least as single descriptors. Note that the values in the second column of Table 2B and the last column of Table 2C correspond exactly with the plots at the top and bottom, respectively, of Figure 2.

A few of the neighborhood enhancements, footnoted within Table 2B,C, are "significantly" less than 1.0. Examples are those from dataset 12 for log *P* and molar refractivity. All such "enhancements" were investigated by constructing their neighborhood plots, and in each such plot the distribution of molecular diversity descriptor differences (*x* coordinates of data points) was so far from random as to produce two or three broad "spikes" of data points. With so few discrete *x* values, the homogeneity conditions for applying the χ^2 test are not actually met.[17]

As an overall indicator of merit for neighborhood enhancements, the χ^2 statistic has one disadvantage, in that it is not a very powerful discriminator when the data available are limited. Therefore in practice, we instead use a neighborhood enhancement of 1.1 as the threshold for identifying a useful neighborhood behavior. Within Table 2B,C, there are 17 instances of an enhancement greater than 1.1 associated with a χ^2 value less than 3.84 and one instance of a χ^2 value greater than 3.84 associated with an enhancement between 1.0 and 1.1.

Is any particular descriptor likely to be of value in designing a set of compounds for general lead discovery, where nothing is known in advance about a specific target? To answer to this question, we would ask, "How often does that descriptor exhibit a neighborhood behavior when applied to randomly chosen datasets?" Thus the bottom lines of Table 2B,C, the frequency of distribution enhancements greater than 1.1 for each descriptor, would be our summary statistic for a diversity descriptor.[18] Here again the 11 descriptors fall into three classes, the first three in Table 2B "almost always" having neighborhood behavior, the last three being

Table 2

A. Datasets and References

ref no.	dataset	compd	structure, activity
1	Uehling	9	camptothecin, DNA fragmentation
2	Strupczewski	34	benzisoxazoles, ip behavioral
3	Siddiqi	10	adenosines, brain A1 binding
4	Garratt1	10	tryptamines, melanophore binding
5	Garratt2	14	tyrptamines, melanophore binding
6	Heyl	11	deltorphin, opioid receptor (DAMGO)
7	Cristalli	32	adenosines, A2a agonists
8	Stevenson	5	piperidines, NK1 antagonism
9	Doherty	6	triarylbutenolides, endothelin-A antagonism
10	Penning	13	SC-41930 analogs, LTB4 antagonism
11	Lewis	7	oxazolinediones, NK1 binding
12	Krystek	30	sulfonamides, endothelin-A antagonism
13	Yokoyama1	13	oxamid acids, T3 binding
14	Yokoyama2	12	oxamic acids, T3 binding
15	Svensson	13	benzindoles, 5-HTA agonism
16	Tsutsumi	13	peptidyl heterocycles, endopeptidase inhibition
17	Chang	34	biphenylsulfonamides, AT1 binding
18	Rosowsky	10	trimetrexate analogs, DHFR inhibition
19	Thompson	8	peptidomimetic, HIV-1 protease inhibition
20	Depreux	26	naphthylethyl amides, melatonin displacement

B. Neighborhood Distribution Enhancements for Molecular Diversity Descriptors Having a Neighborhood Behavior

		ratios of point density in optimal LRT to overall point density for various molecular diversity descriptors (number in parentheses is the χ^2 ratio, with 3.84 being the $P > 0.95$ threshold)					
			topomeric				
ref no.	dataset	2D fingerprt (side chain)	steric	H-bond	2D fingerprt (whole mol)	atom pairs	autocorrelation
1	Uehling	1.90 (14.84)	1.69 (9.96)	1.83 (12.50)	1.55 (6.22)	1.55 (6.72)	1.12 (0.48)
2	Strupczewski	1.73 (154.4)	1.39 (57.45)	1.48 (63.54)	1.41 (59.61)	1.40 (57.23)	1.20 (18.66)
3	Siddiqi	1.04 (0.08)	1.50 (6.77)	1.47 (4.90)	1.04 (0.07)	1.00 (0.00)	1.00 (0.00)
4	Garratt1	1.59 (7.97)	1.65 (10.95)	*b*	1.07 (0.19)	0.90 (0.39)	1.14 (0.64)
5	Garratt2	1.85 (32.58)	1.39 (8.76)	*b*	1.08 (0.50)	0.97 (0.06)	1.09 (0.36)
6	Heyl	1.71 (13.83)	1.04 (0.07)	1.24 (1.54)	1.01 (0.00)	1.11 (0.86)	1.01 (0.71)
7	Cristalli	1.76 (148.5)	1.40 (51.16)	1.22 (12.20)	1.31 (30.27)	1.27 (27.84)	1.17 (11.95)
8	Stevenson	1.08 (0.06)	1.06 (0.02)	*b*	1.07 (0.04)	1.02 (0.00)	1.02 (0.23)
9	Doherty	1.72 (4.23)	1.62 (3.57)	1.07 (0.03)	1.06 (0.04)	1.18 (0.39)	1.28 (1.89)
10	Penning	1.92 (33.23)	1.44 (10.04)	1.72 (20.10)	1.53 (12.73)	1.05 (0.19)	1.36 (7.16)
11	Lewis	1.63 (4.64)	1.05 (0.04)	0.57 (1.93)[c]	1.01 (0.00)	0.97 (0.02)	1.14 (1.13)
12	Krystek	1.23 (15.33)	1.63 (104.3)	1.69 (102.7)	1.23 (16.31)	1.43 (49.06)	1.00 (0.00)
13	Yokoyama1	1.48 (10.06)	1.19 (2.12)	0.71 (4.46)[c]	1.01 (0.00)	1.25 (2.73)	0.99 (0.33)
14	Yokoyama2	1.76 (18.94)	1.23 (2.58)	0.33 (14.67)[c]	1.70 (16.03)	1.25 (2.91)	0.99 (1.06)
15	Svensson	1.68 (18.11)	1.26 (3.50)	0.31 (18.69)[c]	1.02 (0.02)	1.31 (4.95)	1.00 (0.00)
16	Tsutsumi	1.74 (21.56)	1.38 (6.50)	1.67 (17.33)	1.58 (14.35)	1.18 (1.80)	0.95 (1.89)
17	Chang	1.68 (144.4)	1.33 (44.72)	1.35 (34.59)	1.13 (8.36)	1.00 (0.16)	1.20 (17.82)
18	Rosowsky	1.02 (0.02)	1.71 (12.15)	1.44 (4.69)	1.01 (0.00)	1.23 (1.89)	1.42 (5.14)
19	Thompson	1.70 (7.57)	1.47 (3.93)	*b*	1.17 (0.68)	0.87 (0.44)	1.01 (0.86)
20	Depreux	1.65 (72.96)	1.22 (11.27)	0.44 (50.40)[c]	1.18 (6.73)	0.99 (0.03)	0.98 (0.04)
mean		1.59 (42.27)	1.40 (20.03)	1.47 (24.93)[c]	1.21 (8.61)	1.15 (7.88)	1.13 (3.95)
standard dev		0.28	0.20	0.24[c]	0.22	0.19	0.18
ratios > 1.1		17/20	17/20	10/11[c]	10/20	11/20	9/20

C. Neighborhood Distribution Enhancements for Molecular Diversity Descriptors *Not* Having a Neighborhood Behavior

		ratios of point density in optimal LRT to overall point density for various molecular diversity descriptors (number in parentheses is the χ^2 ratio, with 3.84 being the $P > 0.95$ threshold)				
ref no.	dataset	connectivity indices	log *P*	molar refractvty	force field energy atom	random number
1	Uehling	1.19 (2.64)	1.09 (0.05)	1.07 (0.15)	1.00 (0.00)	1.01 (0.00)
2	Strupczewski	1.05 (4.91)	1.00 (0.01)	0.99 (0.06)	1.00 (0.00)	1.00 (0.00)
3	Siddiqi	1.07 (2.33)	0.97 (0.03)	0.92 (0.28)	0.98 (0.02)	1.02 (0.01)
4	Garratt1	1.11 (1.06)	1.01 (0.00)	1.01 (0.00)	1.00 (0.00)	1.02 (0.02)
5	Garratt2	1.09 (3.26)	1.01 (0.01)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
6	Heyl	*f*	0.98 (0.02)	0.95 (0.15)	1.00 (0.00)	1.01 (0.00)
7	Cristalli	0.98 (0.55)	1.06 (0.13)	0.99 (0.05)	1.00 (0.00)	1.00 (0.01)
8	Stevenson	1.02 (0.01)	1.03 (0.01)	1.03 (0.01)	1.02 (0.00)	1.03 (0.01)
9	Doherty	1.02 (1.04)	1.00 (0.00)	1.01 (0.00)	1.03 (0.01)	1.09 (0.12)
10	Penning	1.00 (0.00)	1.00 (0.00)	0.97 (0.05)	1.00 (0.00)	1.01 (0.00)
11	Lewis	1.15 (0.75)	1.00 (0.00)	1.02 (0.01)	1.03 (0.01)	1.06 (0.07)
12	Krystek	1.01 (1.85)	0.85 (10.28)[e]	0.85 (10.04)[e]	1.00 (0.00)	1.00 (0.00)
13	Yokoyama1	1.01 (0.00)	*d*	1.01 (0.00)	1.00 (0.00)	1.00 (0.00)
14	Yokoyama2	1.05 (0.58)	1.00 (0.00)	0.99 (0.00)	1.01 (0.00)	1.00 (0.00)
15	Svensson	1.08 (0.97)	1.01 (0.01)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)
16	Tsutsumi	1.00 (0.07)	0.94 (0.19)	0.95 (0.17)	1.00 (0.00)	1.00 (0.00)
17	Chang	*g*	1.00 (0.00)	1.00 (0.09)	1.00 (0.00)	1.01 (0.07)

Table 2. (Continued)

C. Neighborhood Distribution Enhancements for Molecular Diversity Descriptors *Not* Having a Neighborhood Behavior (Continued)

		ratios of point density in optimal LRT to overall point density for various molecular diversity descriptors (number in parentheses is the χ^2 ratio, with 3.84 being the $P > 0.95$ threshold)				
ref no.	dataset	connectivity indices	log *P*	molar refractvty	force field energy atom	random number
18	Rosowsky	1.08 (1.60)	1.03 (0.03)	0.96 (0.07)	1.00 (0.00)	1.00 (0.00)
19	Thompson	1.02 (0.04)	1.12 (0.25)	0.99 (0.00)	1.00 (0.00)	1.05 (0.07)
20	Depreux	1.01 (0.35)	1.02 (0.12)	0.99 (0.05)	1.00 (0.00)	1.00 (0.00)
mean		1.05 (0.06)	1.01 (0.59)	0.98 (0.56)	1.00 (0.01)	1.02 (0.02)
standard dev		0.06	0.05	0.05	0.01	0.03
ratios > 1.1		3/18	1/19	0/20	0/20	0/20

[a] References are as follows: 1. Uehling, D. E.; Nanthakamur, S. S.; Croom, D.; Emerson, D. L.; Leitner, P. P.; Luzzio, M. J.; *et al.* Synthesis, Topoisomerase I Inhibitory Activity, and *in Vivo* Evaluation of 11-Azacamptothecin Analogs. *J. Med. Chem.* **1995** *38*, 1106. (Table 2, with R_2=Et; IC_{50} data). 2. Strupczewski, J. T.; Bordeau, K. J.; Chiang, Y.; Glamkowski, E. J.; Conway, P. G.; *et al.* 3-[[(aryloxy)alkyl]piperidinyl]-1,2-Benzisoxazoles as D2/5-HT2 Antagonists with Potential Atypical Antipsychotic Activity: Antipsychotic Profile of Iloperidone (HP873). *J. Med. Chem.* **1995**, *38*, 1119 (Tables 2 and 3 with $n = 3$, X = O; ED_{50} for inhibition of apomorphine-induced climbing). 3. Siddiqi, S. M.; Jacobson, K. A.; Esker, J. L.; Olah, M. E.; Ji, X.-d.; *et al.* Search for New Purine- and Ribose-Modified Adenosine Analogs as Selective Agonists and Antagonists at Adenosine Receptors. *J. Med. Chem.* **1995**, *38*, 1174 (Table 1, R_2 = H; K_i(A1) values estimated from percent displacement and stereoisomers averaged as needed). 4, 5. Garratt, P. J.; Jones, R.; Tocher, D. A.; Sugden, D. Mapping the Melatonin Receptor. 3. Design and Synthesis of Melatonin Agonists and Antagonists Derived from 2-Phenyltryptamines. *J. Med. Chem.* **1995**, *38*, 1132 (Tables 1 and 2). 6. Heyl, D. L.; Dandabuthla, M.; Kurtz, K. R.; Mousigian, C. Opioid Receptor Binding Requirements for the &-Selective Peptide Deltorphin I: Phe^3 Replacement with Ring-Substituted and Heterocyclic Amino Acids. *J. Med. Chem.* **1995**, *38*, 1242 (Table 1; binding K_1 to DPDPE). 7. Cristalli, G.; Camaioni, E.; Vittori, S.; Volpini, R.; Borea, P. A.; *et al.* 2-Aralkynyl and 2-Heteroalkynyl Derivatives of Adenosine-5′-N-ethyluronamide as Selective A2a Adenosine Receptor Agonists. *J. Med. Chem.* **1995**, *38*, 1462. 8. Stevenson, G. I.; MacLeod, A. M.; Huscroft, I.; Cascieri, M. A.; Sadowski, S.; Baker, R. 4,4-Disubstituted Piperidines: A New Class of NK_1 Antagonist. *J. Med. Chem.* **1995**, *38*, 1264 (Table 1). 9. Doherty, A. M.; Patt, W. C.; Edmunds, J. J.; Berryman, K. A.; Reisdorph, B. R.; *et al.* Discovery of a Novel Series of Orally Active Non-Peptide Endothelin-A (ET_A) Receptor-Selective Antagonists. *J. Med. Chem.* **1995**, *38*, 1259 (Table 3; IC_{50} ET_A). 10. Penning, T. D.; Djuric, S. W.; Miyashiro, J. M.; Yu, S.; Snyder, J. P.; *et al.* Second-Generation Leukotriene B_4 Receptor Antagonists Related to SC-41930; Heterocyclic Replacement of the Methyl Ketone Pharmacophore. *J. Med. Chem.* **1995**, *38*, 858 (Table 1, all; LTB_4 receptor binding). 11. Lewis, R. T.; Macleod, A. M.; Merchant, K. J.; Kelleher, F.; Sanderson, I.; *et al.* Tryptophan-Derived NK1 Antagonists: Conformationally Constrained Heterocyclic Bioisosteres of the Ester Linkage. *J. Med. Chem.* **1995**, *38*, 923 (Table 2, all). 12. Krystek, S. R.; Hunt, J. T.; Stein, P. D.; Stouch, T. R. 3D-QSAR of Sulfonamide Endothelin Inhibitors. *J. Med. Chem.* **1995**, *38*, 659 (Tables 1 and 2; dimethylated isoxazoles with substitution adjacent to oxygen). 13, 14. Yokoyama, N.; Walker, G. N.; Main, A. J.; Stanton, J. L.; Morrissey, M.; *et al.* Synthesis and SAR of Oxamic Acid and Acetic Acid Derivatives Related to L-Thyronine. *J. Med. Chem.* **1995**, *38*, 695 (Table 1, all; nuclear IC_{50}; Table 3, all; nuclear IC_{50}). 15. Haadsma-Svensson, S. R.; Svensson, K.; Duncan, N.; Smith, M. W.; Lin, Ch.-H. C-9 and N-Substituted Analogs of cis-(3aR)-(−)-2,3,3a,4,5,9b-Hexahydro-3-propyl-1H-benz[e]indole-9-carboxamide: 5HT1A Receptor Agonists with Various Degrees of Metabolic Stability. *J. Med. Chem.* **1995**, *38*, 725 (Table 1, R1 = $CONH_2$; 5-HT_{1A} binding). 16. Tsutsumi, S.; Okonogi, T.; Shibahara, S.; Ohuchi, S.; Hatsushiba, E.; *et al.* Synthesis and Structure-Activity Relationships of Peptidyl α-Keto Heterocycles as Novel Inhibitors of Prolyl Endopeptidase. *J. Med. Chem.* **1994**, *37*, 3492 (Table 2, X = CH_2CH_2; IC_{50}). 17. Chang, L. L.; Ashton, W. T.; Flanagan, K. L.; Chen, T.-B.; O'Malley, S. S.; *et al.* Triazolinone Biphenylsulfonamides as Angiotensin II Receptor Antagonists with High Affinity for Both the AT_1 and AT_2 Subtypes. *J. Med. Chem.* **1994,** *37*, 4464 (Table 1, R^3 = (2-Cl)C_6H_5; AT_1 (rabbit aorta) IC_{50}). 18. Rosowsky, A.; Mota, C. E.; Wright, J. E.; Queener, S. F. 2,4-Diamino-5-chloroquinazoline Analogs of Trimetrexate and Piritrexim: Synthesis and Antifolate Activity. *J. Med. Chem.* **1994**, *37*, 4522 (Table 2; rat liver IC_{50}). 19. Thompson, S. K.; Murthy, K. H. M.; Zhao, B.; Winborne, E.; Green, D. W.; *et al.* Rational Design, Synthesis, and Crystallographic Analysis of a Hydroxyethylene-Based HIV-1 Protease Inhibitor Containing a Heterocyclic P1′-P2′ Amide Bond Isostere. *J. Med. Chem.* **1994**, *37*, 3100 (Table 2, X = Boc; apparent K_i). 20. Depreux, P.; Lesieur, D.; Mansour, H. A.; Morgan, P.; *et al.* Synthesis and Structure-Activity Relationships of Novel Naphthalenic and Bioisosteric Related Amidic Derivatives as Melatonin Receptor Ligands. *J. Med. Chem.* **1994**, *37*, 3231 (Table 1, R1 = 7-OCH_3, R2 = H, $x = 1$, R3 = H; K_D). [b] All compounds in the series had identical hydrogen-bonding fields. [c] Only one side chain within the series had a hydrogen-bonding group, so that there were really only two differences possible for the descriptor difference. These datasets are also excluded from the summary statistics for this metric. [d] Descriptor values could not be calculated for more than one compound in the series. [e] Over 80% of the compounds in this series exhibited only three discrete values for these descriptors, so that the diversity descriptor differences possible (*x* values in the neighborhood plot) were fewer than would be expected. [f] Too many atoms for the program to calculate indices. [g] An atom in the central core had too many valences for the program to calculate indices.

useful "one-half the time", and all those in Table 2C showing neighborhood behavior "almost never".

Only one feature of this overall ranking of 11 diversity descriptors really surprised us: the great difference found for 2D fingerprints depending on whether the structure used to generate the fingerprint was the side chain only or the whole molecule (columns 1 and 4 in Table 2B). Careful investigation of individual cases showed that, even though the side chain is the only variable part of the molecules within any of the 20 series, some of these variable side chain bits are all set to 1 when the "core" structure common to all molecules is added to the side chain. The common core is destroying differentiating information about side chains.

The datasets examined here include whole cell assays (such as Garratt) and even a mouse *in vivo* assay (Strupczewski). Nevertheless, it is worth noting that for different biological activity measures, such as efficacy as a local anesthetic or metabolic stability as measured in a cytochrome P450 assay, or for physical properties, such as solubility, it may well be true that other descriptors are preferable, and it might be found that topomeric fields and/or fingerprints would not be valid for those purposes.

Discussion

What does a neighborhood enhancement mean? More exactly, why is a neighborhood enhancement diagnostic for a "neighborhood behavior"? (As discussed in the Introduction, for Figure 1 to be useful, its dimensions must represent descriptors having the neighborhood behavior, such that a particular compound's biological activity is indeed usefully predictive of all others found within its neighborhood distance.) Any contrary point in a neighborhood plot, that is, any point lying within the upper left triangle (or at least somewhere toward

the left and top of the plot), reports that some small change in the diversity descriptor value (small *x* value) produced a large change in biological activity (large *y* value). The more such points, the more often molecules within the neighborhood region of another (Figure 1) will have activities *not* predictable by testing that other molecule. The fewer such points (the emptier the ULT), the more reliable the diversity descriptor is for designing efficient screening libraries.

Why does the neighborhood enhancement exist? One might naively expect that a well-behaved structure−activity difference plot should yield a straight line, similar to a predicted vs actual plot in QSAR, with an empty lower right triangle as well as an empty upper left triangle. But *absolute differences* in descriptor values behave very differently from the descriptor values themselves. A neighborhood enhancement does not imply any causative relation between structural changes and biological changes. In particular, *large* differences in structure may produce *either small or large* differences in biological end-point, so points at the right of a neighborhood plot may lie anywhere, instead of near the diagonal. *A neighborhood enhancement merely indicates that small differences in structure do not (often) produce large differences in biology*.

It should be noted that the algorithm we describe for defining the diagonal in a neighborhood plot is conservative and occasionally arbitrary. Thus some marginally useful descriptors may not be so defined by the algorithm. For example, if one visually examines the neighborhood plots of those three datasets having enhancements less than 1.1 in the second column of Table 2B (datasets 6, 8, and 11 within Figure 2), it is easy to be persuaded that each plot actually contains a rather empty upper left triangle, which the algorithm however failed to discover. In particular the algorithm tends to fail for small datasets when one or more points lie very close to the *y*-axis. Because the "best diagonal" is defined by a maximal density of points under the diagonal, when a single point or two lie near the *y*-axis and there are no areas of high density elsewhere, the maximum density occurs as the diagonal nears the vertical *y*-axis, when the area under the diagonal becomes very small.

The general problem of validating diversity descriptors has also been considered successfully by Brown, Bures, and Martin[4] of Abbott Laboratories, following earlier work by Willett *et al.*[19] Their comparative study of fingerprint descriptors and clustering methods showed that hierarchically determined clusters of 2D Tanimoto differences among MACCS keys are the most enriched in active molecules, across three sets of in-house screening data. From one of their graphs, a "neighborhood distance" of 0.85 for these 2D fingerprints can be deduced, as the finding that 85% of compounds having a Tanimoto similarity of 0.85 to any active compound are themselves active, in remarkably satisfactory accord with the mean value across all 20 datasets of the neighborhood radius in our results for 2D fingerprints.

There are, however, some important methodological differences between our approach and that of Abbott/Willett. For these workers (as for most others), the fundamental indication of whether two molecules are "similar", in terms of any particular descriptor(s), is whether they occupy the same cluster, not whether their separation distance is less than some critical value. Although molecules in any specified cluster will indeed be closer than some computable distance together, it is possible for molecules in different clusters to be closer together than molecules in the same cluster. This point is illustrated by Figure 4. Here the two "natural clusters" shown are the ones which virtually any clustering method would probably identify. Yet the distance from any molecule-point in cluster 2 to the nearest molecule-point in cluster 1 is less than the distances separating the farthest molecule-points within cluster 1. We maintain that diversity-based library design should focus on the underlying distances, not on the clusters, whose occurrence must be somewhat adventitious for any particular collection of molecules.

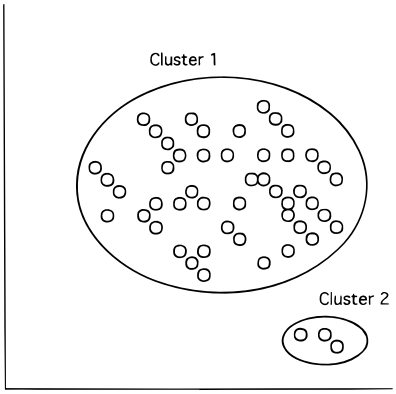


Figure 4. Intracluster distances between data points (within cluster 1) can be larger than intercluster distances (between clusters 1 and 2).

A related concept in diversity design is that of selecting compounds to "fill holes" in some diversity descriptor-defined space, a strategy most commonly put forward in connection with pharmacophorically-based molecular descriptors. In principle this is a perfectly sensible procedure, but if and only if the descriptor(s) defining the space have already been shown to be valid, *e.g.*, have a neighborhood behavior such as we describe here. Otherwise, despite the best of intentions, such a procedure amounts only to an elaborately random selection. Another difficulty with "filling holes" may be that those descriptor spaces which have been successfully validated (Table 2B) have many hundreds of underlying diversity dimensions,[20] while the invalid ones (Table 2C) have few and usually only one dimension. To illustrate this point, Figure 5 shows the (improbably smooth!) relation between the proportion of neighborhood enhancements greater than 1.1 and the log of the descriptor dimensionality for all data in Table 2B,C. One might expect that validity of a descriptor would increase with higher dimensionality if the dimensions are truly relevant, but adding irrelevant new dimensions would probably reduce the apparent validity by making some dissimilar compounds appear to be relatively closer.

A highly dimensional space, the characteristic of valid descriptors, must be mostly empty of compounds, rather than having only a few "holes to fill". (Consider, for example, what happens to NMR peak separations as the experiment dimensionality goes only from 1D to 2D to 3D!) To address this difficulty, some diversity researchers have used factor analysis (PCA) to reduce the number of dimensions[21] (essentially "squeezing down"

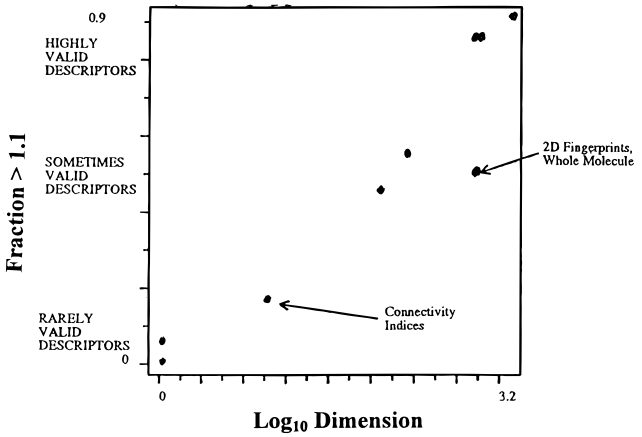


Figure 5. Relationship between dimensionality and frequency of a neighborhood enhancement for 11 diversity descriptors. Although the almost perfectly linear quality of the increase shown in frequency of neighborhood enhancements with the log of descriptor dimensionality is surely artifactual, the increase itself seems real.

all dimensions in which the particular compound library shows relatively little diversity). But, because of the resulting distortion of intermolecular distances, the new diversity descriptors thereby created require validation.

Another difference to be noted between our approach and that of Abbott/Willett is in the nature of the datasets used for validation. Brown *et al.* emphasized structural diversity, by considering three "in-house" Abbott screening datasets, each containing hundreds of active−inactive compound observations, whereas we have emphasized biological diversity, by considering 20 datasets (18 different biological activities), each containing fewer than 40 compounds. The neighborhood validation method is applicable to either situation. However large diverse screening sets with many reported inactive compounds are not readily available outside of large pharmaceutical laboratories. On the other hand, the more variety in the biological datasets used for validation, the more likely that the results will be useful for arbitrary biological test systems, as in design of a screening library.

Although the analytic approaches themselves differ, to the extent they are both successful, they must in some respects be transformations of one another. In the Abbott/Willett approach, a superior diversity descriptor produces some clusters which contain higher than uniform frequencies of active molecules (and perforce other clusters which contain lower frequencies). Put differently, success in a Abbott/Willett experiment implies that the mean intercluster variance in activity must be larger than the mean intracluster variance in activity. Turning to the other (x) dimension of a neighborhood plot, in an Abbott/Willett experiment the mean intercluster diversity differences are always larger than the mean intracluster diversity differences because it is these very differences which control the formation of the clusters. Thus an Abbott/Willett experiment does indeed seek to test whether similar compounds (those in the same clusters) will have a lower frequency of large differences in activities, exactly as is indicated by the empty ULT characteristic of a neighborhood enhancement. Although the two approaches are thus seen as roughly equivalent, the ability of our method to demonstrate significant "neighborhood behavior" for the same metrics as the Abbott/Willett approach but with much smaller datasets bespeaks its much more powerful formulation of the diversity validation problem.

Objective descriptor validation has received little attention yet in the still scanty scientific literature on diversity descriptors. Many publications[22] employ descriptors whose utility for local optimization with QSAR is well-established, such as connectivity indices[23] and log P, but which we here report to be of no significant value in the larger scale diversity problems posed by lead discovery and "lead explosion". Other proposed diversity descriptors have no objective justification.[21,24]

In the Introduction, we propose that lead discovery would be most efficient using screening libraries composed of compounds that are separated in a validated diversity descriptor space. Studies using descriptors and methodologies akin to the Abbott/Willett approach clearly show that "activity-enriched clusters" or "activity-prioritized screening lists" can be constructed retrospectively.[4,11,19,25] But are there yet any data from a prospective experiment, using descriptors known to be valid, which can confirm or deny our proposal? One such result was recently reported by Moreau.[26] Using an autocorrelation vector as the diversity descriptor, followed by PCA and hierarchical cluster analysis, 2000 compounds were chosen to represent the entire Hoechst-Marion-Roussel collection of some 500 000 compounds. He then compared screening results for the "2000" with random and hand-picked selections. In the first comparison, the "2000" scored 98 hits of 1323 tested, a 7.4% rate; the hand-picked compounds scored 11/354 or 3.0%. In a second comparison, the "2000" hit 69 of 1323 or 5.2%, while random-picked hit 0 in 700 tries, and hand-picked had 1 hit of 354. These encouraging results matched our expectations for a screening selection based on a valid diversity descriptor. We therefore decided to validate the autocorrelation vector using our own methodology and obtained the positive results shown as the last column in Table 2B. In summary, a descriptor shown as weakly valid *a posteriori* (the autocorrelation vector) produced an *a priori* screening selection which yielded many more hits than other selection methods. However, more experiments are clearly needed.

Conclusions

Medicinal chemists have always sought to discover and patent "activity islands" and to identify the "peaks" in those islands as potential future drug candidates. In the long run, the appreciation that not all dimensions of molecular diversity are equally relevant in defining those activity islands and the ability to objectively identify and rank the relevant dimensions may be of the greatest importance in molecular discovery research.

References

(1) A commercially available realization of these approaches is the Optiverse screening library of compounds, synthesized by Panlabs (Bothell, WA) and designed and distributed by Tripos, and associated screening, lead explosion, and lead optimization services. For more detail of the library design procedures themselves, see: Ferguson, A. M.; Patterson, D. E.; Garr, C. D.; Underiner, T. L. Designing Chemical Libraries for Lead Discovery. *J. Biomol. Screen.* **1996**, accepted for publication.

(2) An early review distinguishing the computational approaches for lead generation and optimization is Redl, G.; Cramer, R. D.; Berkoff, C. E. Quantitative Drug Design. *Chem. Soc. Rev.* **1974**, *3*, 273−292.

(3) Patents pending.

(4) Brown, R. D.; Bures, M. G.; Martin, Y. C. Similarity and cluster analysis applied to molecular diversity. American Chemical Society Meeting, Anaheim, CA, 1995; COMP 3.

(5) The χ^2 statistic is calculated here as:

$$\frac{(\text{actual LRT count} - \text{expected LRT count})^2}{\text{expected LRT count}}$$

where: expected LRT count = (LRT area/total area) × total count and LRT is an abbreviation of "lower right trapezoid".

(6) The simpler choice of datapoint [*x*,*y*] did not behave well in practice.

(7) Clark, M.; Cramer, R. D.; van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(8) This is the standard "2D screen" generation rules for a UNITY molecular structure database. More exactly, fragments of length 4 do include hydrogen, and 60 of the 988 available bits code for specific atoms or fragments and combinations.

(9) BioByte, Inc., Pomona, CA.

(10) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(11) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(12) Moreau, G.; Broto, P.; Vandycke, C. *Nouv. J. Chim.* **1980**, *4*, 359−360, 757−764; *Eur. J. Med. Chem. Chim. Ther.* **1984**, *19*, 66−78.

(13) 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.

(14) For any receptor atom locus, the nearest lattice point was identified, and then the 27 bits referenced by a cube centered on that point and 4 Å wide were set to 1.

(15) Martin, Y. C.; Bures, M.; Danaher, E.; DeLazzer, J.; Lico, I.; Pavlik, P. A Fast Approach to Pharmacophore Mapping and its Application to Dopaminergic and Benzodiazepine Agonists. *J. Comput.-Aid. Mol. Des.* **1993**, *7*, 93−102.

(16) A careful observer will notice that there is a tendency within the bottom graphs for points to be less dense on the right side of the graph. This skewed distribution is expected for *differences* between the uniformly distributed values within an interval that a random number generator should produce. Note that this innate tendency of value differences *opposes* any tendency for a descriptor to exhibit a neighborhood enhancement.

(17) Such artifactually nonuniform distributions may also produce positive neighborhood distribution enhancements that are actually spurious. This possibility was not directly investigated. However, as can be seen, for example, in the top panels of Figure 2, most of the enhancements appearing in Table 2B are not at all artifactual.

(18) The same ranking of diversity descriptors would be obtained from other figures of merit that might be proposed, for example, the rms χ^2.

(19) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, U.K., 1987.

(20) To address a point of possible confusion, notice that the distance between two objects is always a single scalar number no matter how many dimensions the space containing those objects has. Thus, for example, the Tanimoto "distance" between two fingerprints is effectively a sum over all thousand-odd dimensions represented by the presence/absence of particular structural fragments, albeit normalized not by the total number of bits settable but instead by the bits actually set in either fingerprint.

(21) A good example, which by comparing several libraries addresses the issue of library specific artifacts from factor analysis, is Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.*, accepted for publication. Unfortunately the underlying descriptors used are closely related to those in Table 2C.

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

(23) It should however be cautioned that we have here considered only a small subset of the connectivity indices which can be calculated.

(24) Madden, D.; Krchnak, V.; Lebl, M. Synthetic combinatorial libraries: Views on techniques and their applications. *Perspect. Drug Discovery Des.* **1995**, *2*, 269−285. Chapman, D.; Ross, M. J. Poster at the symposium Chemical and Biolmolecular Diversity, San Diego, CA, Dec 14−16, 1994; lecture at the symposium Exploiting Molecular Diversity: Small Molecule Libraries for Drug Discovery, La Jolla, CA, Jan. 23−25, 1995; conference summary available from Wendy Warr & Assoc., 6 Berwick Ct, Cheshire, U.K. CW4 7HZ.

(25) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural Analysis. A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **1974**, *17*, 533. Hodes, L. Clustering a Large Number of Compounds. 1. Establishing the Method on an Intial Sample. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66−71.

(26) Moreau, G. Proceedings of an IBC Meeting on Combinatorial Chemistry, Oct. 30−31, London, U.K.

JM960290N