# Quantized Surface Complementarity Diversity (QSCD): A Model Based on Small Molecule−Target Complementarity

Edward A. Wintner* and Ciamac C. Moallemi

*NeoGenesis, Inc., 840 Memorial Drive, 4th Floor, Cambridge, Massachusetts 02139*

*Received October 5, 1999*

A model of molecular diversity is presented. The model, termed "Quantized Surface Complementarity Diversity" (QSCD), defines molecular diversity by measuring molecular complementarity to a fully enumerated set of theoretical target surfaces. Molecular diversity space is defined as the molecular complement to this set of enumerated surfaces. Using a set of known test compounds, the model is shown to be biologically relevant, consistently scoring known actives as similar. At the resolution of the model, which examines molecules "quantized" into 4.24 Å cubic units and treats four points of specific energetic complementarity, the minimum number of compounds needed to fully cover molecular diversity space up to volume 1070 cubic Å is estimated to be on the order of 24 million molecules. Most importantly, QSCD allows for individual points in diversity space to be filled by direct modeling of molecular libraries into detailed 3D templates of shape and functionality.

## Introduction

Combinatorial chemistry allows the creation of unprecedented numbers of organic compounds; what was once unthinkable − the rational synthesis of millions of small organic molecules − is now achievable in a matter of days.[1−6] With this newly acquired ability comes a question which, prior to combinatorial chemistry, seemed foolish even to pose: How can one create a set of molecules of such diversity as to contain at least one potent binder to any given target of interest?

This question is central to drug discovery in the "post-genomic era", a scientific midpoint characterized by a growing wealth of DNA sequence information and a relative dearth of corresponding target structures and their functions.[7−9] Soon, there will be many more putative targets than can be studied by X-ray crystallography, multidimensional NMR, or other high-resolution biophysical techniques. Any attempt to generate biologically active ligands to these targets of unknown structure will require general screening libraries: libraries of molecules that cover a high percentage of so-called "diversity space".[10−13]

We define diversity as the measure, based on predefined criteria, of the difference or similarity between all members of a set. In a pharmaceutical setting, it follows that molecular diversity can be defined as the measure, based on biological criteria, of the difference or similarity between small molecules. Largely due to continuing advances in our understanding of the principles of molecular recognition,[14−24] there exist today many methods of calculating biologically relevant diversity of small molecules.[25,26] Each method defines slightly different criteria for molecular comparison, and each thereby presents a different configuration of diversity space as a whole. Examples include low-dimensional diversity space such as BCUT metrics,[27−29] high-dimensional diversity space such as Chem-X/ChemDiverse multiple-point pharmacophores,[30−34] and empirical biological diversity space such as affinity fingerprinting.[35−37]

For the most part, current methods are able to successfully identify compounds of the same pharmacological class as being similar and compounds of different pharmacological classes as being different.[25,38] Given a starting pharmacophore from known ligands and/or the target site of a target crystal structure, such methods interface well with the design of complementary combinatorial libraries.

The design of combinatorial libraries to cover all of diversity space is a rather different problem, however. In this case, it is not enough to be able to compare existing molecules for differences or similarities. In addition to being able to place molecules relative to one another in diversity space, one must be able to point to an absolute area of diversity space *not yet covered* and from its coordinates design a novel set of compounds to fill that uncovered space.

At present, there are no methods which can directly facilitate the above process of filling empty diversity space, because the transformations by which existing models assign molecules to coordinates in diversity space are irreversible: molecules can be mapped to diversity coordinates, but diversity coordinates cannot be mapped directly to molecular structures.

For example, in the well-known BCUT method used to generate 4D to 6D diversity space, molecules are broken down into matrices according to connectivity and molecular interaction properties.[25,27−29] Coordinates in diversity space are assigned through the resulting eigenvalues of these matrices, leading to highly useful multidimensional plots of molecular diversity. However, because the use of eigenvalues is an irreversible transformation (different 3D shapes can map to the same eigenvalues), it follows that an *empty* coordinate in BCUT diversity space cannot be translated into a 3D template of a "missing molecule". Thus, while a model

* To whom correspondence should be addressed. Tel: (617)-868-1500. Fax: (617) 868-1515. E-mail: wintner@neogenesis.com.

10.1021/jm990504b CCC: $19.00 © 2000 American Chemical Society
Published on Web 04/26/2000

such as BCUT diversity is well-validated as a tool for finding combinatorial matches to a lead compound or pharmacophore, it cannot be directly used to populate the entire diversity space which it defines.

Similarly, in the popular Chem-X/ChemDiverse diversity package,[33] molecules are broken down into all accessible three- or four-point pharmacophores of triangular or tetrahedral functionality distances.[30−32,34] If the model is used to display molecular diversity, coordinates in diversity space are assigned through the resulting string of accessible three- or four-point pharmacophores; this method has been shown to be highly effective in classifying molecules by pharmacological similarity. However, the mapping of complex 3D shapes to a set of triangular or tetrahedral functionality distances is an irreversible transformation; empty three- or four-point pharmacophores in Chem-X-derived diversity space cannot be translated into a 3D template of complex shape. Since a set of coordinates in Chem-X is insufficient to define the shape of “missing molecules”, Chem-X cannot be used to directly populate empty molecular diversity space.

A final example of current diversity methods is affinity fingerprinting, in which molecules are empirically assayed against a panel of 10−20 actual proteins selected to be promiscuous in their ability to bind small molecules.[35,36] Position in molecular diversity space is assigned through the resulting string of $IC_{50}$ binding values, and these affinity fingerprints provide unprecedented ability to group similarly active compounds in diversity space. However, because the actual mode of binding in any assay is not incorporated in the resulting $IC_{50}$ value, the mapping of molecules to the selected protein panel is an irreversible transformation. Thus, an empty coordinate in affinity fingerprinting diversity space (an “unmatched” string of $IC_{50}$s to a given protein panel) cannot be back-translated into a 3D molecular template. A similar affinity fingerprinting diversity method has been put into practice using a panel of computational protein surfaces and a modified form of the DOCK program.[37] While this method shows similar promise in its ability to detect pharmacological similarity, it is, like its real-world counterpart, an irreversible mapping.

To rationally and systematically fill diversity space, an *informationally reversible* diversity model is needed. This model must be formulated such that: (1) members (in this case molecules) can be assigned to coordinates for similarity/dissimilarity comparison and (2) empty coordinates retain the information necessary to directly generate coordinate membership.

The path of least resistance to such a model is to use as coordinates the exact information that differentiates one member from another, without intervening, irreversible transformations. To apply this reasoning to molecular diversity, it must first be asked: What are the criteria by which diversity of compounds is to be measured (what information differentiates one molecule from another)? The most fundamental criterion in molecular drug discovery is the extent to which two molecules have similar or different binding affinities to a given target surface. With the assumption that similar binding affinity tracks with a molecule's complementarity to similar target surfaces,[39,40] we have selected as our criterion for diversity *complementarity to a fully enumerated set of theoretical target surfaces*.

Given the above definition of molecular diversity, it remains only to (a) provide a biologically relevant basis set of enumerated theoretical target surfaces and (b) quantify molecular complementarity to a given theoretical target surface at a level which is both in accordance with known principles of molecular recognition and computationally applicable to millions of compounds. With a numerical determination of complementarity and a biologically relevant basis set of surfaces, molecular diversity space is thus absolutely established as the molecular complement to a fully enumerated set of theoretical target surfaces.

## Methods

**Theoretical Target Surfaces.** The model of this paper begins with a set of theoretical target surfaces that approximates all possible binding pockets with volume equal to or less than $V$. To generate a finite set of these surfaces, we consider each theoretical surface to be formed by successively carving cubic units out of an initially flat surface. These cubic units represent “negative space” that a potential ligand could occupy. Given cubic units with sides of length $R$ (the resolution of the model), we use at most $V/R^3$ negative space cubes to describe each theoretical target surface. We note that others have previously employed cubic units to successfully approximate complementarity between small molecules and individual protein surfaces.[41]

The size of a negative space cube is directly related to the resolution and type of diversity data which the user desires as output. In the current formulation of the model, we wished to maximize negative space cube size such that the difference of a single cube in a surface is highly differentiating in terms of molecular recognition (i.e. every surface is orthogonal to every other surface). At the same time, we needed to retain information in each negative space cube sufficient to predict shape and functional complementarity at a ligand/surface interface. The former constraint minimizes overlap of diversity information while the latter constraint maximizes precision of diversity information. Together, the competing constraints result in a basis unit for the enumeration of theoretical target surfaces which minimizes the number of negative space cubes needed to accurately model diversity for a given volume $V$.

A resolution of 4.24 Å negative space cubes was found by computer optimization of test molecules to provide an upper limit of cube size while still maintaining an acceptable level of molecular shape information (see Experimental Section). Interestingly, 4.24 Å is the approximate VDW “cross-section” of a $(CH_2)_n$ chain; a series of 4.24 Å units neatly encapsulates a $(CH_2)_n$ chain in its ground-state conformation (Figure 1).

In the study described herein, we use as a basis set for diversity a set of theoretical target surfaces comprised of all possible shape combinations of 6−14 negative space cubes of resolution 4.24 Å (negative volume between 460 and 1070 cubic Å) subject to the following rules: (1) Surfaces are created by successively “carving out” negative space cubes from a flat block of infinite width and depth (the theoretical target). (2) All negative space cubes of a given surface must share at

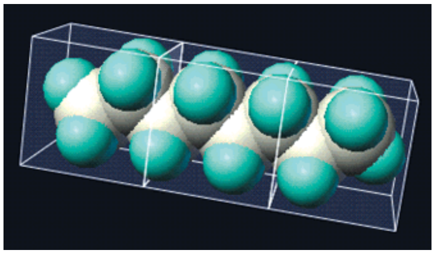


**Figure 1.** A $(CH_2)_n$ chain encapsulated by 4.24 Å cubic units.

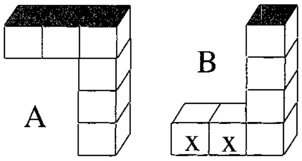


**Figure 2.** Example of surfaces allowed and disallowed by the nonocclusion parameter in the theoretical target surface generation algorithm of the text. Gray shading represents the opening of the theoretical surface: (A) allowed and (B) disallowed due to two occluded negative space cubes (marked X).

least one face with another negative space cube of the surface, and all must be part of a single, contiguous negative surface. (3) No negative space cubes may be occluded in the $+Z$ axis of the infinite surface block; that is, there may be no solid surface between any negative space cube and the surface plane of the infinite block (see Figure 2). (4) Surfaces duplicating a previous surface with respect to rotation in the $X-Y$ plane are discarded.

In this study, rule (3) above was added as a compromise between complete coverage of topological possibilities and desire to maintain computational speed. This compromise was made based on the topological assumption that occlusions of 4.24 Å or more are infrequent in small molecule/target interactions and that their omission would thus have only a small effect on predicting diversity of binding affinities of small molecules.

The above rules generate 49 268 918 unique negative surface shapes including chiral opposites. Covering a negative volume between 460 and 1070 cubic Å, these surface shapes are deemed sufficient to examine diversity of most small molecules. For instance, examining a previously published reference set of pharmaceutically relevant compounds (a filtered Comprehensive Medicinal Chemistry, or CMC, database),[42] 5049 out of 5120 compounds (98.6%) have a volume of 1070 cubic Å or less.

Within each of the 49 268 918 unique negative surface shapes, each negative space cube is assigned a molecular property characteristic $P_m$ that represents the dominant molecular environment which any atoms that are placed within that negative space will experience. Properties used are $P_1$ hydrophobic, $P_2$ polarizable (includes aromatics), $P_3$ H-bond acceptor, $P_4$ H-bond donor, $P_5$ H-bond donor/acceptor, $P_6$ potentially positively charged (basic), and $P_7$ potentially negatively charged (acidic). These seven types of molecular environments are assumed to represent a minimal basis set of factors that contributes to the electrostatic/VDW complementarity of a ligand and a target surface.[14,15,18] In this study, four positions of particular molecular property $P_{1-7}$ are assigned, leading to $7^4*N!/((N-4)!*4!)$ surfaces for each surface shape of $N$ negative space cubes. All other $(N-4)$ cubes not assigned a particular molecular property are given property $P_8$, slightly hydrophobic. The latter assignment is based on an assumption that hydrophobic effects are, on average, the largest single component contributing to ligand/target interaction.[17]

**Table 1.** Numerical Breakdown of the Total Number of Theoretical Target Surfaces Created Using the Algorithm Given in the Text[a]

| vol (no. ($N$) of negative space cubes) | no. of unique surface shapes | approx. no. of functionally different surfaces per unique surface shape: $7^4*N!/((N-4)!*4!)$ | exact no. of unique surfaces |
|---|---|---|---|
| 6 | 212 | 36 015 | 7 163 338 |
| 7 | 885 | 84 035 | 73 271 443 |
| 8 | 3 959 | 168 070 | 655 324 488 |
| 9 | 17 747 | 302 526 | 5 350 917 208 |
| 10 | 81 407 | 504 210 | 40 912 578 322 |
| 11 | 375 897 | 792 330 | 297 622 676 624 |
| 12 | 1 753 218 | 1 188 495 | 2 082 225 979 379 |
| 13 | 8 224 443 | 1 716 715 | 14 116 888 070 845 |
| 14 | 38 811 150 | 2 403 401 | 93 264 917 290 356 |
| total 6−14 | 49 268 918 | | 109 808 653 272 003 |

[a] Surfaces consist of 6−14 negative space cubes and 4 sites of 7 possible molecular property characteristics. Number of functionally different surfaces per surface shape varies for infrequent cases in which a given shape has an axis of symmetry, so actual number of unique surfaces is slightly less than (no. surface shapes)$*7^4*N!/((N-4)!*4!)$.

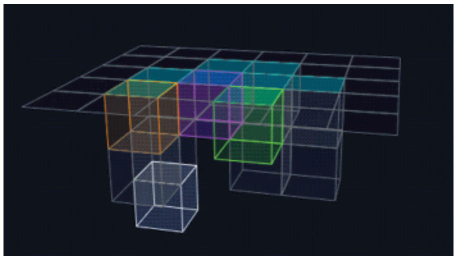


**Figure 3.** Theoretical target surface of 13 negative space cubes and 4 sites of specific molecular property interaction: hydrophobic (white), polarizable (purple), H-bond accepting (green), and H-bond donating (orange). Blue shading indicates the opening of the theoretical surface.

In sum, the above process implies as a basis set for molecular diversity $1.1 \times 10^{14}$ theoretical target surfaces of negative volume between 460 and 1070 cubic Å and having four sites of specific molecular property characteristics $P_{1-7}$. The numerical breakdown of these 110 trillion surfaces is listed in Table 1. One such surface is shown in Figure 3.

**Molecular Quantization.** To measure complementarity of small molecules to the above basis set of theoretical target surfaces, the small molecules to be compared must be formatted in a similar frame of reference. Thus, all molecules used in this study are “quantized” into positive space cubes (“quanta”) of resolution 4.24 Å according to the following algorithm (see also Figure 4):

(1) To represent each molecule, a set of up to 100 minimized conformations within user-defined parameters is created. In this study, Tripos Multisearch modeling was used and all conformations within 10 kcal of the lowest energy conformation found were accepted (see Experimental Section).

(2) For each conformation, a 4.24 Å 3D grid of cubes (quanta) is aligned on top of the 3D structure using the

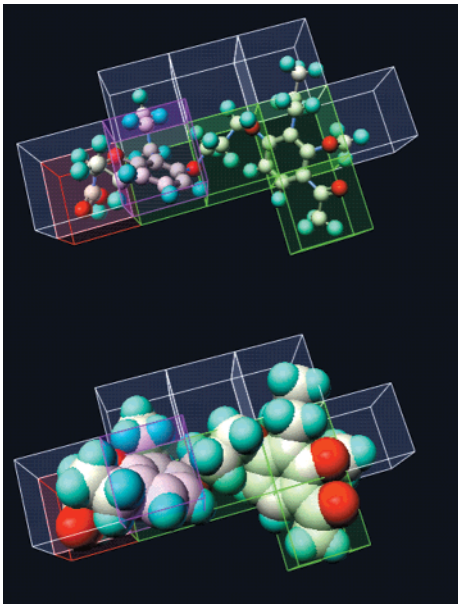


**Figure 4.** “Quantized” representation (Q-file) of one conformation of molecule **6a** superimposed on its atomic structure (ball-and-stick and space-filling model). Molecular property characteristics of the Q-file are hydrophobic (white quanta), polarizable (purple quanta), H-bond accepting (green quanta), and negatively charged (red quanta).

molecule’s principal axes of rotation (calculated with all atoms having mass 1).

(3) To all 4.24 Å quanta which contain at least a user-defined percentage of the VDW radius of any atom, a dominant molecular property characteristic is assigned based on connectivity rules (e.g. R−[C═O]−O−H yields $P_7$, R−O−H yields $P_5$; see definitions of $P_1$−$P_7$ above). Order of dominance is from $P_7$ to $P_1$, in order of maximum complementarity score obtainable by a given characteristic (see Table 2). Minimum percentage of VDW radius parameter allows for a user-defined protrusion beyond the surface of a quantum cube, adding a measure of topological “flexibility” to the quantization process. 32% was found optimal for this study (see Experimental Section).

(4) The total number of 4.24 Å quanta that have been assigned a property characteristic is counted.

(5) The grid alignment is shifted per user-defined parameters and the process is repeated from step (2) until all shift combinations have been searched.

(6) For each conformation in step (1), the “Q-file” (3D configuration of property-assigned quanta) is saved that: (a) has the lowest number of quanta in step (4) and (b) is closest to the principal alignment in step (2).

Thus, an average molecule in this study is represented by 100 Q-files, each file consisting of $N$ positive space cubes or quanta of 4.24 Å resolution having an assigned molecular property characteristic $P_m$ ($m$ = 1−7). A typical Q-file (molecule **6a**) is shown in Figure 4, superimposed upon its corresponding conformation. The process of optimization of quantization parameters is described in the Experimental Section.

**Mapping.** Given molecules which have been rendered into sets of Q-files, each quantized conformation can now be mapped into the diversity space defined by the set of 1.1 × $10^{14}$ theoretical target surfaces above. This is accomplished through the following algorithm:

(1) For each quantized conformation of each molecule, each of its 24 possible $X/Y/Z$ rotations (6 faces * 4 rotations per face) is fit to each of the 49 268 918 available surface shapes.

(2) For a given conformation-to-surface shape fit, if at least a user-defined minimum number of negative and positive space cubes overlap (in this case either 9 quanta *or* $N - 2$ quanta of a conformation of $N$ quanta; see Experimental Section), and if no quanta of the conformation extend beyond the bounds of the surface shape except at the mouth of the surface shape, then the complementarity of the quantized conformation to all theoretical target surfaces of that shape is examined in detail in step (3). If the above conditions are not met, the next conformation is examined.

(3) A score is generated for the complementarity of the given conformation to each theoretical target surface of a given shape from step (2) by employing user-defined parameters. (The process of optimization of the complementarity parameters is described in the Experimental Section.) Complementarity parameters used in the model are as follows: (a) a negative parameter for each rotatable bond of the conformation; (b) if conformational energies are calculated, a negative parameter for the energy of the conformation above the lowest energy conformation from that molecule; (c) a positive parameter for the hydrophobic energy gained by removing “water” from any hydrophobic ($P_1$) or polarizable ($P_2$) surface face of either the conformation or the theoretical surface; (d) a positive parameter for the hydrophobic energy gained by removing “water” from any mildly hydrophobic ($P_8$) surface face of the theoretical surface; (e) a positive or negative molecular property interaction parameter for overlapping negative and positive space cubes as depicted in Table 2.

(4) If and only if the score in step (3) meets a user-defined minimum, then the conformation (and thus the molecule it represents) is said to be complementary to the given theoretical target surface.

The computational advantage inherent in the process of molecule and surface quantization is realized in the speed of complementarity checking. Whereas a traditional docking program must search a high-dimensional configuration space, QSCD resolves the problem to a framework bounded by 24 possible fitting orientations and a finite number of translations. This approximation allows 3D diversity computation on a scale that is applicable to very large sets of molecules.

## Results and Discussion

The above process results in a complementarity map for any molecule that consists of a list of all theoretical target surfaces to which at least one conformation of the molecule is complementary. Comparison of these maps provides a novel method for measuring diversity of small molecules. We term the model on which this process is based “Quantized Surface Complementarity Diversity” (QSCD) because it calculates diversity by measuring complementarity to a quantized representation of theoretical target surfaces.

To maintain a computationally efficient complementarity scoring system, QSCD makes many approximations of molecular recognition. These include cubic units of 4.24 Å resolution, gross approximations of surface

**Table 2.** Relative Magnitudes of Parameters Used in Calculating Molecular Property Interactions between Negative Space Cubes (theoretical target surfaces) and Positive Space Cubes (quantized molecules)[a]

| theoretical target surface properties | quantized molecule properties | | | | | | |
|---|---|---|---|---|---|---|---|
| | $P_7$ neg | $P_6$ pos | $P_5$ H-bond don/acc | $P_4$ H-bond don | $P_3$ H-bond acc | $P_2$ polarizable | $P_1$ hydrophobic |
| $P_7$ neg charged | − − − | +++ | 0 | + | − | − − | − − |
| $P_6$ pos charged | +++ | − − − | 0 | − | + | 0 | − − |
| $P_5$ H-bond don/acc | 0 | 0 | ++ | + | + | − | − − |
| $P_4$ H-bond don | + | − | + | − − | ++ | − | − − |
| $P_3$ H-bond acc | − | + | + | ++ | − − | − | − − |
| $P_2$ polarizable | − − | 0 | − | − | − | ++ | 0 |
| $P_1$ hydrophobic | − − | − − | − − | − − | − − | 0 | + |
| $P_8$ (surface only) | − | − | 0 | − | − | 0 | 0 |

[a] Magnitudes (listed from highest to lowest): +++, ++, +, 0, −, − −, − − −.

contact area, exactly four points of seven finite types of molecular property characteristics, static theoretical surfaces, and a limited set (up to 100) of low-energy conformers. Thus, the final complementarity scores are in no way presumed to give useful *binding energies* for any *individual* match of conformation to target surface. However, taken over all conformations of a molecule and across an enumerated set of theoretical target surfaces, the scoring system is proven to be statistically relevant. This is demonstrated below.

**Model Validation.** To test the validity of the QSCD model, we needed to prove that the method, at a minimum, satisfies the central criterion that it was designed to analyze: the extent to which two molecules have similar or different binding affinities.

Thus, eight sets of test molecules were analyzed (Figure 5), seven of which were known from the literature to have binding affinities to seven distinct targets (in addition to a known overlap between sets **3** and **4**). An eighth set with no known binding affinities was chosen with minor atomic and spatial changes to examine the sensitivity of the model at 4.24 Å resolution. Known activities of the molecules in Figure 5 are listed in Table 3 with references. The bulk of these molecules have previously been used as part of an in-depth study validating molecular descriptor approaches for the prediction of molecular diversity *within* compound classes.[12] This is a more stringent discrimination than the base criterion sought for our model, which seeks at a minimum to show accurate diversity prediction *between* compound classes.

As described above, conformations of all 20 test molecules were "quantized" and then mapped onto the basis set of 1.1 × $10^{14}$ theoretical surfaces. Complementary surfaces are tabulated for each molecule in Table 4. There are many ways to analyze the resulting set of complementarity mappings; since in this case individual molecule comparisons were desired, each of the 20 mappings was compared pairwise for a total of 190 data points. Mappings were scored in similarity from 0 to 1000 based on a function of the number of theoretical surfaces in common:

$$\text{Score} = \text{SS} * \text{FS} = \text{ShapeScore} * \text{FunctionalityScore}$$

$$\text{SS} = 100 * \frac{\text{\# theoretical target surface shapes common to A \&B}}{\text{total \# surface shapes complementary either to A or to B}}$$

$$\text{FS} = 10 * \left(\frac{\text{\# theoretical target surface shapes common to A \& B with at least 1 set of 4 common functionalities}}{\text{total \# theoretical target surface shapes common to A \& B}}\right) \wedge \Phi$$

The first term in this equation gives a percentage measure (0−100) of shape similarity between molecules A and B, while the second term gives a measure from 0 to 10 of functional similarity per given shape overlap. The complete scores are detailed in Table 5. Using this scoring system, the maximum score obtainable by very rigid, structurally similar molecules is 1000. However, many molecules can only be *sampled* by an examination of up to 100 low-energy conformations (an average molecule w/5+ rotatable bonds will have at least $3^5 = 243$ conformations). Thus, for most molecules with more than 100 accessible conformations, similarity scores between 0 and 100 are observed. The scoring constant Φ in the equation above adjusts the influence of functionality on scoring. A value of 0.33 was found to be optimal (see Experimental Section), meaning that shape is the dominant criterion in our measure of diversity.

Figure 6 shows a plot of all 190 pairings ranked by similarity score. Orange circles show "heterogeneous" pairs of expected dissimilarity (e.g. **2a**, **6b**), while blue squares show "homogeneous" pairs of expected similarity (e.g. **2a**, **2b**). Clearly, the model ranks homogeneous pairs almost exclusively higher than heterogeneous pairs; all 15 pharmacologically similar pairs (blue) fell within the top 20 scores out of 190. All homogeneous scores were ranked above 25, while the median score in this experiment was 2.8, showing good "signal-to-noise". QSCD is thus a valid predictor of target binding similarity among these molecules.

A closer look at the pairings reveals further validation. As might be expected from their relative rigidity (low number of accessible conformations) and structural similarity, the highest scoring pairs are **2a**/**2b**, **8a**/**8b**, and **8d**/**8e**. Furthermore, examination of the pairings of **8c** with **8a**,**b**,**d**,**e** (yellow triangles in Figure 6, yellow in Table 5) yields scores that are within the top 20% of the pairing experiment but which are generally lower that the "homogeneous" pairs. This makes sense from a target-binding point of view, considering that one face of **8c** contains a large molecular difference (an extra phenyl substituent). To the extent that this face is not involved in complementarity to a target surface, the molecules are similar; to the extent that this face must be complementary for binding to occur, the molecules are quite different. Figure 7 shows one such case of a surface common to both **8a** and **8c**; the protruding phenyl substituent plays no role in complementarity.

As a rule, we have found close similarity of shape and functionality for molecules which score 25 or higher. In addition to the pairings that one would expect, several other pairs scored between 25 and 35. When these molecules were examined by molecular modeling, significant overlaps were found, suggesting that these high scores are not just "noise" in the QSCD model. Figure 8

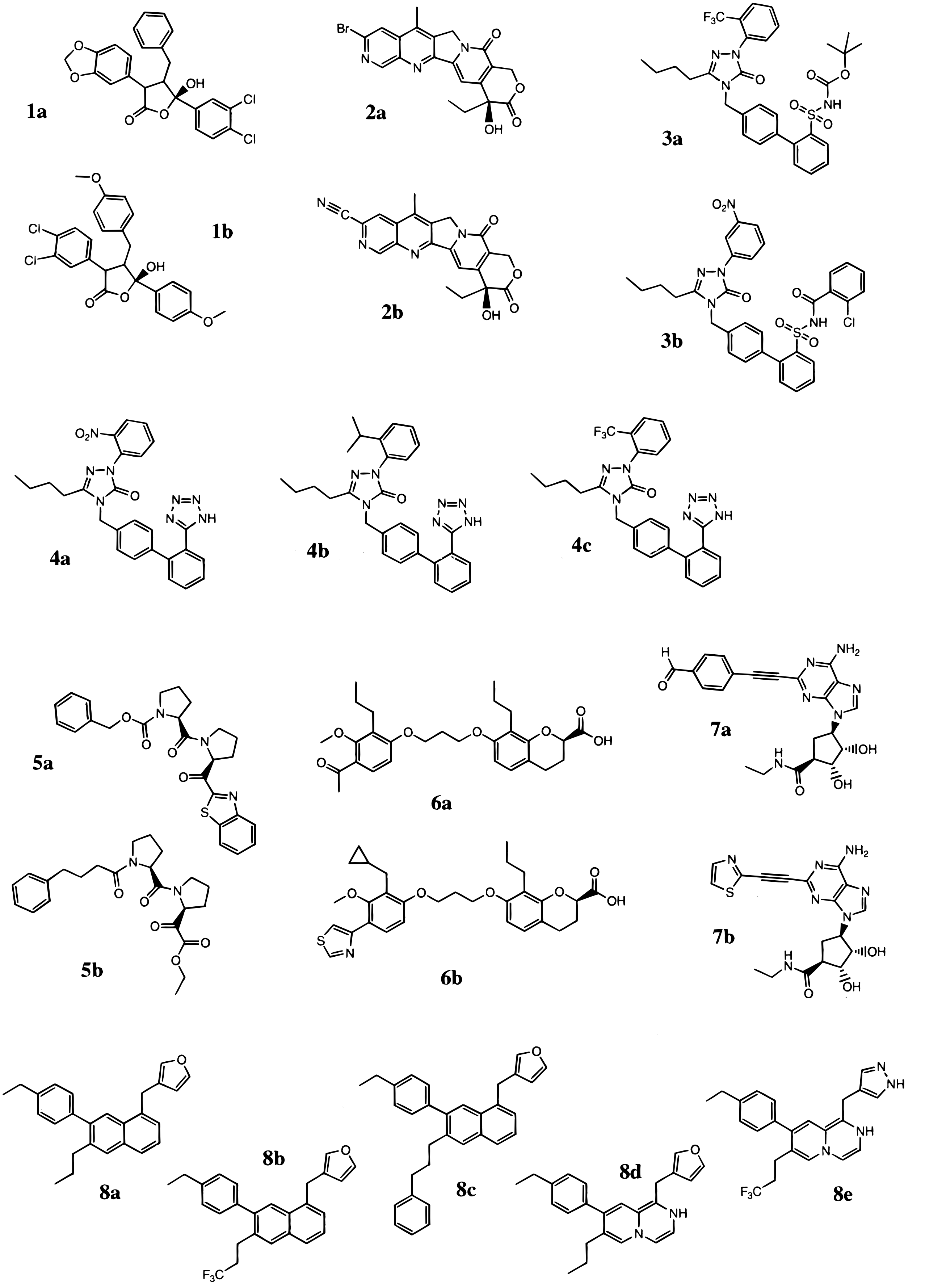


**Figure 5.** Test molecules used in this study.

depicts one such case between **1a** and **5a**; conformations of **1a** and **5a** are displayed that were found in QSCD to be complementary to the same surface (Figure 8A,B). 3D overlays (Figure 8C) confirm correlation of general

**Table 3.** Pharmacological Activities of the Molecules Used in This Study (see Figure 5)

| compd | assay | IC$_{50}$ or $K_i$ (nM) | ref |
|---|---|---|---|
| **1a** | binding to endothelin A receptor | 400 | *a* |
| **1b** | binding to endothelin A receptor | 170 | *a* |
| **2a** | inhibition of DNA fragmentation by topoisomerase I | 28 | *b* |
| **2b** | inhibition of DNA fragmentation by topoisomerase I | 143 | *b* |
| **3a** | binding to AT2 subtype of angiotensin II receptor | 17 (0.45)[h] | *c* |
| **3b** | binding to AT2 subtype of angiotensin II receptor | 173 (31)[h] | *c* |
| **4a** | binding to AT1 subtype of angiotensin II receptor | 0.85 | *d* |
| **4b** | binding to AT1 subtype of angiotensin II receptor | 1.4 | *d* |
| **4c** | binding to AT1 subtype of angiotensin II receptor | 1.2 (23 000)[i] | *d* |
| **5a** | inhibition of prolylendopeptidase protease activity | 5 | *e* |
| **5b** | inhibition of prolylendopeptidase protease activity | 10.3 | *e* |
| **6a** | binding to leukotriene B4 receptor | 320 | *f* |
| **6b** | binding to leukotriene B4 receptor | 3.2 | *f* |
| **7a** | binding to A2A type adenosine receptor | 6.3 | *g* |
| **7b** | binding to A2A type adenosine receptor | 41.3 | *g* |
| **8a** | none | | |
| **8b** | none | | |
| **8c** | none | | |
| **8d** | none | | |
| **8e** | none | | |

[a] Doherty et al. *J. Med. Chem*. **1995,** *38*, 1259−1263. [b] Uehling et al. *J. Med. Chem*. **1995,** *38*, 1106−1118. [c] Chang et al. *J. Med. Chem*. **1994,** *37*, 4464−4478. [d] Chang et al. *J. Med. Chem*. **1993,** *36*, 2558−2568. [e] Tsutsumi et al. *J. Med. Chem*. **1994,** *37*, 3492−3502. [f] Penning et al. *J. Med. Chem*. **1995,** *38*, 858−868. [g] Cristalli et al. *J. Med. Chem*. **1995,** *38*, 1462−1472. [h] Numbers in parentheses indicate IC$_{50}$ in AT1 subtype assay of series **4**. [i] Numbers in parentheses indicate IC$_{50}$ in AT2 subtype assay of series **3**.

**Table 4.** Tabulation of Surface Shapes and Total Number of Theoretical Target Surfaces Complementary to Each Molecule in Figure 5

| compd | complementary surface shapes | complementary surfaces (shape plus functionality) |
|---|---|---|
| **1a** | 376 | 16 127 687 |
| **1b** | 379 | 9 086 768 |
| **2a** | 27 | 545 584 |
| **2b** | 27 | 416 210 |
| **3a** | 414 | 4 970 816 |
| **3b** | 315 | 813 024 |
| **4a** | 487 | 4 542 463 |
| **4b** | 479 | 12 388 826 |
| **4c** | 482 | 7 595 982 |
| **5a** | 337 | 2 080 523 |
| **5b** | 374 | 1 966 837 |
| **6a** | 220 | 192 067 |
| **6b** | 186 | 153 436 |
| **7a** | 362 | 298 927 |
| **7b** | 269 | 22 367 |
| **8a** | 45 | 5 561 654 |
| **8b** | 41 | 3 959 678 |
| **8c** | 333 | 17 324 247 |
| **8d** | 64 | 2 059 546 |
| **8e** | 87 | 1 343 811 |
| average | 265 | 4 572 523 |

shape and four points of functionality, although they also make clear the limits of resolution of complementarity information using 4.24 Å units. As can be seen from Figure 8, the surface in question can detect general shape and functional similarity, but by no means provides a basis to predict atom-for-atom overlap between molecules.

A final result of note comes from examination of the QSCD rankings of sets **3** and **4**. While set **4** is known to bind exclusively to the AT1 subtype of the angiotensin II receptor, set **3** is known to bind to both the AT1 subtype and the AT2 subtype. While QSCD found high similarity within sets **3** (score = 27) and **4** (av score = 54), it found an average similarity of 6.9 between **3a** and set **4** and an average similarity of 3.3 between **3b** and set **4** (green diamonds in Figure 6, green Table 5). On the basis of the model, one would therefore conclude that while sets **3** and **4** share a limited number of complementary theoretical surfaces, they are dissimilar with respect to the majority of theoretical target surfaces. This is in fact the case with the AT2 subtype of the angiotensin II receptor, to which **4c** binds 50 000 times more poorly than **3a** (see Table 3).

**Fundamental Advantages.** Having validated the basis set used for QSCD in the classification of molecular diversity, it must be noted that many other models may do as well or better in detecting target binding similarity/dissimilarity between molecules.[12,29,37,39,43] For instance, Figure 9 and Table 6 show the same set of 20 molecules ranked by Tanimoto similarity[39] of standard 2D UNITY fingerprints (see Experimental Section). The data demonstrate that the 2D model is equally capable of predicting pharmacologically similar pairs; UNITY ranks similarity between AT1 and AT2 subtype binders much higher than QSCD, although it finds unusually high similarity between **8a** and **8c**. In general, such 2D fingerprint descriptors have been found highly effective in clustering pharmacologically similar compounds and are widely used in determining molecular diversity of existing structures.[25]

The great advantage of QSCD, however, lies in the value of its *negative* information: QSCD determines not only diversity of existing structures, but also structure of nonexisting diversity. *Given theoretical surface shapes for which no complements exist in a general screening library, QSCD allows the design of molecules to fill the given diversity void.*

As stipulated in its formulation, the QSCD basis set is created through a reversible process; information *resolution* may be lost in fixing the parameters of a cube's size and functional scope, but information *content* is retained in either direction. Just as a single molecular conformation and orientation corresponds to a defined pattern in QSCD space, likewise, a single point in QSCD space (within the limits of volume $V$, resolution $R$, and $N$ sites of functionality $P_m$) corresponds to a unique 3D shape with a defined 3D array of functionality. Given any starting set of molecules, unoccupied points in QSCD space directly define the molecular shapes and functionalities which those molecules do not cover. Thus, a set of detailed 3D molecular templates (at the resolution of the QSCD model used) is immediately available for the creation of novel molecules.

As an example, Figure 10 shows a plot of all of the theoretical surface shapes covered by all of the conformations of all of the molecules used in this study (see Figure 5). The total volume of the cube in Figure 10 encompasses all 49 268 918 theoretical surface shapes

**Table 5.** Ranking of Molecules in Figure 5 by QSCD Diversity Score[a]

| Surface Shapes | Shape Overlaps | Shape Similarity Score | Functionality Score Per Shape Overlap | Molecule A | Molecule B | Total Similarity Score | Rank |
|---|---|---|---|---|---|---|---|
| 27 | 27 | 100.00% | 10 | **2a** | **2b** | 1000.0 | 1 |
| 57 | 29 | 50.88% | 10 | **8a** | **8b** | 508.8 | 2 |
| 107 | 44 | 41.12% | 9.61 | **8d** | **8e** | 395.0 | 3 |
| 888 | 81 | 9.12% | 9.7 | **4c** | **4a** | 88.5 | 4 |
| 121 | 11 | 9.09% | 9.35 | **8a** | **8e** | 85.0 | 5 |
| 98 | 7 | 7.14% | 10 | **8d** | **8b** | 71.4 | 6 |
| 119 | 9 | 7.56% | 9.2 | **8e** | **8b** | 69.6 | 7 |
| 102 | 7 | 6.86% | 10 | **8a** | **8d** | 68.6 | 8 |
| 657 | 54 | 8.22% | 8.04 | **5a** | **5b** | 66.0 | 9 |
| 376 | 30 | 7.98% | 7.76 | **6a** | **6b** | 61.9 | 10 |
| 716 | 39 | 5.45% | 9.74 | **1b** | **1a** | 53.0 | 11 |
| 402 | 18 | 4.48% | 10 | **8e** | **8c** | 44.8 | 12 |
| 922 | 39 | 4.23% | 10 | **4b** | **4c** | 42.3 | 13 |
| 687 | 26 | 3.78% | 9.31 | **1a** | **5a** | 35.2 | 14 |
| 570 | 24 | 4.21% | 8.15 | **6a** | **5b** | 34.3 | 15 |
| 602 | 29 | 4.82% | 6.77 | **7a** | **7b** | 32.6 | 16 |
| 933 | 33 | 3.54% | 9.12 | **4b** | **4a** | 32.2 | 17 |
| 724 | 27 | 3.73% | 8.22 | **3a** | **5a** | 30.7 | 18 |
| 727 | 23 | 3.16% | 9.38 | **1a** | **5b** | 29.7 | 19 |
| 705 | 24 | 3.40% | 7.94 | **3b** | **3a** | 27.0 | 20 |
| 535 | 22 | 4.11% | 6.49 | **6a** | **5a** | 26.7 | 21 |
| 766 | 22 | 2.87% | 7.94 | **3a** | **5b** | 22.8 | 22 |
| 693 | 16 | 2.31% | 9.79 | **1a** | **8c** | 22.6 | 23 |
| 366 | 8 | 2.19% | 10 | **8b** | **8c** | 21.9 | 24 |
| 507 | 16 | 3.16% | 6.79 | **6b** | **5a** | 21.4 | 25 |
| 389 | 8 | 2.06% | 10 | **8d** | **8c** | 20.6 | 26 |
| 675 | 16 | 2.37% | 8.55 | **3b** | **1a** | 20.3 | 27 |
| 289 | 7 | 2.42% | 8.3 | **7b** | **2a** | 20.1 | 28 |
| 289 | 7 | 2.42% | 8.3 | **7b** | **2b** | 20.1 | 29 |
| 519 | 16 | 3.08% | 6.3 | **3b** | **6a** | 19.4 | 30 |
| 548 | 14 | 2.55% | 7.54 | **1a** | **6b** | 19.3 | 31 |
| 371 | 7 | 1.89% | 10 | **8a** | **8c** | 18.9 | 32 |
| 801 | 15 | 1.87% | 9.28 | **4b** | **5a** | 17.4 | 33 |
| 547 | 13 | 2.38% | 7.27 | **6b** | **5b** | 17.3 | 34 |
| 776 | 14 | 1.80% | 9.5 | **3a** | **1a** | 17.1 | 35 |
| 618 | 16 | 2.59% | 6.3 | **3a** | **6a** | 16.3 | 36 |
| 809 | 15 | 1.85% | 8.74 | **4a** | **5a** | 16.2 | 37 |
| 588 | 12 | 2.04% | 7.47 | **3a** | **6b** | 15.2 | 38 |
| 636 | 16 | 2.52% | 5.73 | **3b** | **5a** | 14.4 | 39 |
| 741 | 12 | 1.62% | 8.74 | **1b** | **5b** | 14.1 | 40 |
| 843 | 12 | 1.42% | 9.71 | **4b** | **1a** | 13.8 | 41 |
| 383 | 6 | 1.57% | 7.94 | **7a** | **2a** | 12.4 | 42 |
| 383 | 6 | 1.57% | 7.94 | **7a** | **2b** | 12.4 | 43 |
| 807 | 12 | 1.49% | 8.36 | **4c** | **5a** | 12.4 | 44 |
| 847 | 11 | 1.30% | 9.35 | **4c** | **1a** | 12.1 | 45 |
| 454 | 9 | 1.98% | 6.06 | **1a** | **8e** | 12.0 | 46 |
| 848 | 10 | 1.18% | 9.66 | **4b** | **1b** | 11.4 | 47 |
| 739 | 17 | 2.30% | 4.9 | **7b** | **4a** | 11.3 | 48 |
| 704 | 8 | 1.14% | 9.57 | **1b** | **8c** | 10.9 | 49 |
| 842 | 11 | 1.31% | 8.17 | **4b** | **5b** | 10.7 | 50 |
| 882 | 11 | 1.25% | 8.6 | **4b** | **3a** | 10.7 | 51 |
| 829 | 15 | 1.81% | 5.85 | **7a** | **4c** | 10.6 | 52 |
| 700 | 7 | 1.00% | 10 | **8c** | **5b** | 10.0 | 53 |
| 853 | 10 | 1.17% | 8.44 | **4a** | **1a** | 9.9 | 54 |
| 740 | 11 | 1.49% | 6.49 | **7b** | **4c** | 9.6 | 55 |
| 589 | 7 | 1.19% | 7.54 | **1a** | **6a** | 9.0 | 56 |
| 663 | 7 | 1.06% | 8.3 | **8c** | **5a** | 8.8 | 57 |
| 417 | 4 | 0.96% | 9.09 | **1a** | **8a** | 8.7 | 58 |
| 435 | 5 | 1.15% | 7.37 | **1a** | **8d** | 8.5 | 59 |
| 853 | 8 | 0.94% | 9.09 | **4c** | **1b** | 8.5 | 60 |
| 709 | 7 | 0.99% | 8.3 | **1b** | **5a** | 8.2 | 61 |
| 833 | 16 | 1.92% | 3.97 | **7a** | **4a** | 7.6 | 62 |
| 398 | 3 | 0.75% | 10 | **2a** | **5b** | 7.5 | 63 |
| 398 | 3 | 0.75% | 10 | **2b** | **5b** | 7.5 | 64 |
| 806 | 6 | 0.74% | 10 | **4b** | **8c** | 7.4 | 65 |
| 809 | 6 | 0.74% | 10 | **4c** | **8c** | 7.4 | 66 |
| 794 | 8 | 1.01% | 7.21 | **4a** | **3b** | 7.3 | 67 |
| 854 | 7 | 0.82% | 8.94 | **4a** | **5b** | 7.3 | 68 |
| 893 | 8 | 0.90% | 7.94 | **4a** | **3a** | 7.1 | 69 |
| 663 | 10 | 1.51% | 4.65 | **4a** | **6b** | 7.0 | 70 |
| 398 | 3 | 0.75% | 8.74 | **8d** | **5a** | 6.6 | 71 |
| 860 | 6 | 0.70% | 9.41 | **4a** | **1b** | 6.6 | 72 |
| 680 | 9 | 1.32% | 4.81 | **3b** | **5b** | 6.4 | 73 |
| 785 | 8 | 1.02% | 6.3 | **3a** | **1b** | 6.4 | 74 |
| 414 | 3 | 0.72% | 8.74 | **1a** | **8b** | 6.3 | 75 |
| 435 | 3 | 0.69% | 8.74 | **8d** | **5b** | 6.0 | 76 |
| 521 | 3 | 0.58% | 10 | **4b** | **8a** | 5.8 | 77 |
| 561 | 4 | 0.71% | 7.94 | **1b** | **6b** | 5.7 | 78 |
| 463 | 3 | 0.65% | 8.74 | **1b** | **8e** | 5.7 | 79 |
| 834 | 7 | 0.84% | 6.59 | **7a** | **4b** | 5.5 | 80 |
| 642 | 6 | 0.93% | 5.51 | **7b** | **1b** | 5.1 | 81 |
| 850 | 6 | 0.71% | 6.94 | **4c** | **5b** | 4.9 | 82 |
| 644 | 4 | 0.62% | 7.94 | **3b** | **8c** | 4.9 | 83 |
| 422 | 2 | 0.47% | 10 | **1b** | **8a** | 4.7 | 84 |
| 512 | 2 | 0.39% | 10 | **4a** | **2a** | 3.9 | 85 |
| 512 | 2 | 0.39% | 10 | **4a** | **2b** | 3.9 | 86 |
| 422 | 2 | 0.47% | 7.94 | **8e** | **5a** | 3.8 | 87 |
| 541 | 2 | 0.37% | 10 | **4b** | **8d** | 3.7 | 88 |
| 695 | 4 | 0.58% | 6.3 | **7a** | **5a** | 3.6 | 89 |
| 743 | 4 | 0.54% | 6.3 | **3a** | **8c** | 3.4 | 90 |
| 597 | 2 | 0.34% | 10 | **1b** | **6a** | 3.4 | 91 |
| 817 | 3 | 0.37% | 8.74 | **4a** | **8c** | 3.2 | 92 |
| 892 | 4 | 0.45% | 6.3 | **4c** | **3a** | 2.8 | 93 |
| 359 | 1 | 0.28% | 10 | **2a** | **8c** | 2.8 | 94 |
| 359 | 1 | 0.28% | 10 | **2b** | **8c** | 2.8 | 95 |
| 377 | 1 | 0.27% | 10 | **8b** | **5a** | 2.7 | 96 |
| 791 | 3 | 0.38% | 6.94 | **4b** | **3b** | 2.6 | 97 |
| 381 | 1 | 0.26% | 10 | **8a** | **5a** | 2.6 | 98 |
| 666 | 2 | 0.30% | 7.94 | **4c** | **6b** | 2.4 | 99 |
| 442 | 1 | 0.23% | 10 | **1b** | **8d** | 2.3 | 100 |
| 508 | 1 | 0.20% | 10 | **4c** | **2a** | 2.0 | 101 |
| 508 | 1 | 0.20% | 10 | **4c** | **2b** | 2.0 | 102 |
| 519 | 1 | 0.19% | 10 | **4b** | **8b** | 1.9 | 103 |
| 565 | 1 | 0.18% | 10 | **4b** | **8e** | 1.8 | 104 |
| 552 | 1 | 0.18% | 10 | **6a** | **8c** | 1.8 | 105 |
| 633 | 12 | 1.90% | 0 | **7b** | **1a** | 0.0 | 106 |
| 492 | 9 | 1.83% | 0 | **3b** | **6b** | 0.0 | 107 |
| 210 | 3 | 1.43% | 0 | **2a** | **6b** | 0.0 | 108 |
| 210 | 3 | 1.43% | 0 | **2b** | **6b** | 0.0 | 109 |
| 449 | 6 | 1.34% | 0 | **7b** | **6b** | 0.0 | 110 |
| 599 | 7 | 1.17% | 0 | **7b** | **5a** | 0.0 | 111 |
| 741 | 7 | 0.94% | 0 | **7b** | **4b** | 0.0 | 112 |
| 543 | 5 | 0.92% | 0 | **7a** | **6b** | 0.0 | 113 |
| 696 | 6 | 0.86% | 0 | **4c** | **6a** | 0.0 | 114 |
| 638 | 5 | 0.78% | 0 | **7b** | **5b** | 0.0 | 115 |
| 678 | 5 | 0.74% | 0 | **7b** | **3a** | 0.0 | 116 |
| 702 | 5 | 0.71% | 0 | **4a** | **6a** | 0.0 | 117 |
| 578 | 4 | 0.69% | 0 | **7a** | **6a** | 0.0 | 118 |
| 771 | 5 | 0.65% | 0 | **7a** | **3a** | 0.0 | 119 |
| 486 | 3 | 0.62% | 0 | **7b** | **6a** | 0.0 | 120 |
| 690 | 4 | 0.58% | 0 | **3b** | **1b** | 0.0 | 121 |
| 354 | 2 | 0.56% | 0 | **7b** | **8e** | 0.0 | 122 |
| 581 | 3 | 0.52% | 0 | **7b** | **3b** | 0.0 | 123 |
| 226 | 1 | 0.44% | 0 | **6b** | **8b** | 0.0 | 124 |
| 459 | 2 | 0.44% | 0 | **8e** | **5b** | 0.0 | 125 |
| 696 | 3 | 0.43% | 0 | **4b** | **6a** | 0.0 | 126 |
| 230 | 1 | 0.43% | 0 | **6b** | **8a** | 0.0 | 127 |
| 738 | 3 | 0.41% | 0 | **7a** | **1b** | 0.0 | 128 |
| 517 | 2 | 0.39% | 0 | **6b** | **8c** | 0.0 | 129 |
| 283 | 1 | 0.35% | 0 | **6a** | **8d** | 0.0 | 130 |
| 306 | 1 | 0.33% | 0 | **6a** | **8e** | 0.0 | 131 |
| 313 | 1 | 0.32% | 0 | **7b** | **8a** | 0.0 | 132 |
| 675 | 2 | 0.30% | 0 | **7a** | **3b** | 0.0 | 133 |
| 332 | 1 | 0.30% | 0 | **7b** | **8d** | 0.0 | 134 |
| 663 | 2 | 0.30% | 0 | **4b** | **6b** | 0.0 | 135 |
| 734 | 2 | 0.27% | 0 | **7a** | **5b** | 0.0 | 136 |
| 405 | 1 | 0.25% | 0 | **2a** | **1b** | 0.0 | 137 |
| 402 | 1 | 0.25% | 0 | **2a** | **1a** | 0.0 | 138 |
| 405 | 1 | 0.25% | 0 | **2b** | **1b** | 0.0 | 139 |
| 402 | 1 | 0.25% | 0 | **2b** | **1a** | 0.0 | 140 |
| 425 | 1 | 0.24% | 0 | **7a** | **8d** | 0.0 | 141 |
| 418 | 1 | 0.24% | 0 | **8a** | **5b** | 0.0 | 142 |
| 414 | 1 | 0.24% | 0 | **8b** | **5b** | 0.0 | 143 |
| 448 | 1 | 0.22% | 0 | **7a** | **8e** | 0.0 | 144 |
| 505 | 1 | 0.20% | 0 | **4b** | **2a** | 0.0 | 145 |
| 505 | 1 | 0.20% | 0 | **4b** | **2b** | 0.0 | 146 |
| 601 | 1 | 0.17% | 0 | **7b** | **8c** | 0.0 | 147 |
| 737 | 1 | 0.14% | 0 | **7a** | **1a** | 0.0 | 148 |
| 796 | 1 | 0.13% | 0 | **4c** | **3b** | 0.0 | 149 |
| 407 | 0 | 0.00% | 0 | **7a** | **8a** | 0.0 | 150 |
| 403 | 0 | 0.00% | 0 | **7a** | **8b** | 0.0 | 151 |
| 695 | 0 | 0.00% | 0 | **7a** | **8c** | 0.0 | 152 |
| 310 | 0 | 0.00% | 0 | **7b** | **8b** | 0.0 | 153 |
| 527 | 0 | 0.00% | 0 | **4c** | **8a** | 0.0 | 154 |
| 546 | 0 | 0.00% | 0 | **4c** | **8d** | 0.0 | 155 |
| 569 | 0 | 0.00% | 0 | **4c** | **8e** | 0.0 | 156 |
| 523 | 0 | 0.00% | 0 | **4c** | **8b** | 0.0 | 157 |
| 532 | 0 | 0.00% | 0 | **4a** | **8a** | 0.0 | 158 |
| 551 | 0 | 0.00% | 0 | **4a** | **8d** | 0.0 | 159 |
| 574 | 0 | 0.00% | 0 | **4a** | **8e** | 0.0 | 160 |
| 528 | 0 | 0.00% | 0 | **4a** | **8b** | 0.0 | 161 |
| 342 | 0 | 0.00% | 0 | **3b** | **2a** | 0.0 | 162 |
| 342 | 0 | 0.00% | 0 | **3b** | **2b** | 0.0 | 163 |
| 360 | 0 | 0.00% | 0 | **3b** | **8a** | 0.0 | 164 |
| 379 | 0 | 0.00% | 0 | **3b** | **8d** | 0.0 | 165 |
| 402 | 0 | 0.00% | 0 | **3b** | **8e** | 0.0 | 166 |
| 356 | 0 | 0.00% | 0 | **3b** | **8b** | 0.0 | 167 |
| 441 | 0 | 0.00% | 0 | **3a** | **2a** | 0.0 | 168 |
| 441 | 0 | 0.00% | 0 | **3a** | **2b** | 0.0 | 169 |
| 459 | 0 | 0.00% | 0 | **3a** | **8a** | 0.0 | 170 |
| 478 | 0 | 0.00% | 0 | **3a** | **8d** | 0.0 | 171 |
| 501 | 0 | 0.00% | 0 | **3a** | **8e** | 0.0 | 172 |
| 455 | 0 | 0.00% | 0 | **3a** | **8b** | 0.0 | 173 |
| 247 | 0 | 0.00% | 0 | **2a** | **6a** | 0.0 | 174 |
| 72 | 0 | 0.00% | 0 | **2a** | **8a** | 0.0 | 175 |
| 91 | 0 | 0.00% | 0 | **2a** | **8d** | 0.0 | 176 |
| 114 | 0 | 0.00% | 0 | **2a** | **8e** | 0.0 | 177 |
| 68 | 0 | 0.00% | 0 | **2a** | **8b** | 0.0 | 178 |
| 364 | 0 | 0.00% | 0 | **2a** | **5a** | 0.0 | 179 |
| 247 | 0 | 0.00% | 0 | **2b** | **6a** | 0.0 | 180 |
| 72 | 0 | 0.00% | 0 | **2b** | **8a** | 0.0 | 181 |
| 91 | 0 | 0.00% | 0 | **2b** | **8d** | 0.0 | 182 |
| 114 | 0 | 0.00% | 0 | **2b** | **8e** | 0.0 | 183 |
| 68 | 0 | 0.00% | 0 | **2b** | **8b** | 0.0 | 184 |
| 364 | 0 | 0.00% | 0 | **2b** | **5a** | 0.0 | 185 |
| 420 | 0 | 0.00% | 0 | **1b** | **8b** | 0.0 | 186 |
| 265 | 0 | 0.00% | 0 | **6a** | **8a** | 0.0 | 187 |
| 261 | 0 | 0.00% | 0 | **6a** | **8b** | 0.0 | 188 |
| 250 | 0 | 0.00% | 0 | **6b** | **8d** | 0.0 | 189 |
| 273 | 0 | 0.00% | 0 | **6b** | **8e** | 0.0 | 190 |

[a] Color code: blue = homogeneous pairs, yellow = +phenyl pairs (**8c**), green = AT1 and AT2 pairs (**3**, **4**).

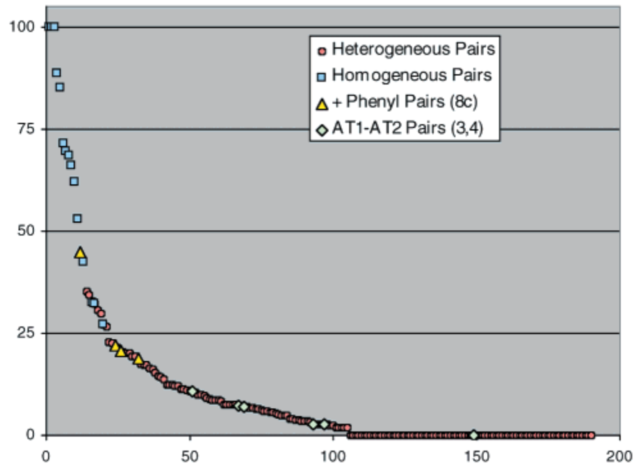


**Figure 6.** Ranking of molecules in Figure 5 by QSCD similarity scores.

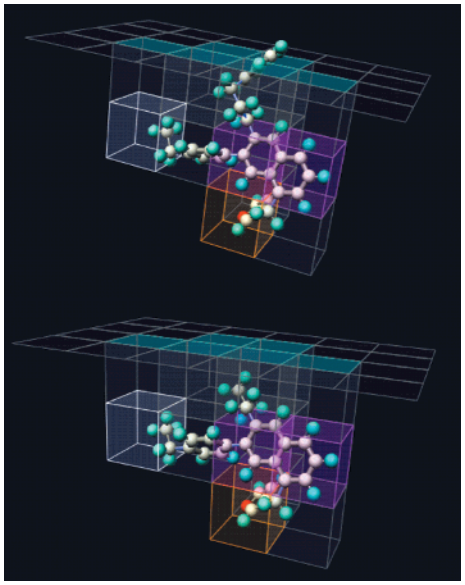


**Figure 7.** Examination of a theoretical target surface common to molecules **8c** (top) and **8a** (bottom). Blue shading indicates opening of the theoretical surface. Specific points of complementarity on the theoretical target surface are hydrophobic (white), polarizable (purple), and H-bond donating (orange). Superimposition of the original molecular conformations onto the theoretical target surface demonstrates that the extra phenyl substituent of **8c** protrudes from the opening of the theoretical surface and is not involved in complementarity to the surface.

as listed in Table 1. As can be seen from the plot and two expanded points, there are many theoretical surface shapes which are "unfilled" by the set of compounds shown in Figure 5. Thus, in searching for molecules or libraries to enhance the diversity of the given set of compounds, the chemist is presented with a set of actual 3D templates into which new compound libraries may be designed. In comparison, mapping the same set of compounds in a "nonreversible" diversity space would also display a set of coordinates to which the molecules map, but there would be no way to visualize the 3D shape of any point that was not filled by one of the compounds in the set. Using BCUT values[27−29] for example (Figure 11), the coordinates specified for an unfilled point leave the chemist with a set of normalized eigenvalues. While these may give an idea of relative abundance of a given functionality (e.g. H-bond donor) at this point in diversity space, the coordinates give no hint of what shape or class of molecules might fill that diversity void.

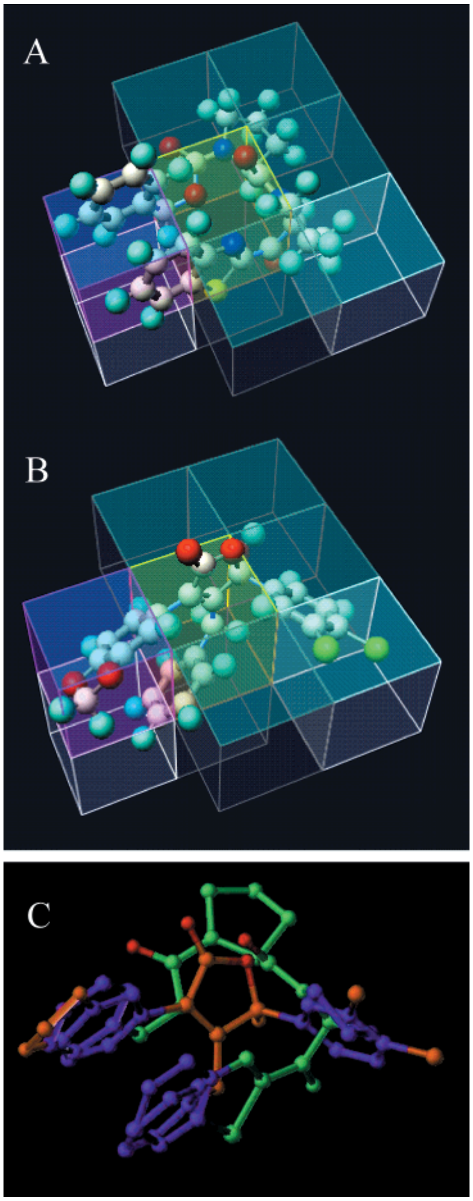


**Figure 8.** Examination of a theoretical target surface common to molecules **1a** (A) and **5a** (B). Blue shading indicates opening of the theoretical surface. Specific points of surface complementarity found by QSCD are hydrophobic (white), polarizable (purple), and H-bond donating/accepting (yellow). Overlay plot (C) of the non-hydrogen backbones of **1a** (orange) and **5a** (green) indicates similar features. Aromatic/hydrophobic overlaps are shown in purple; H-bond donating oxygens are in red. Overlay generated with Sybyl version 6.5 (Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144).

The above example shows how QSCD is a reversible diversity model with respect to molecular shape. Within a given surface shape in QSCD, there are many combinations of functionality, leading to many different theoretical surfaces. If a given library fills only a portion of theoretical surfaces of a given surface shape, by following the same process outlined above and in Figure 10, unfilled surfaces of specific shape and functionality may be identified and filled with complementary libraries. By using data-mining algorithms to analyze and

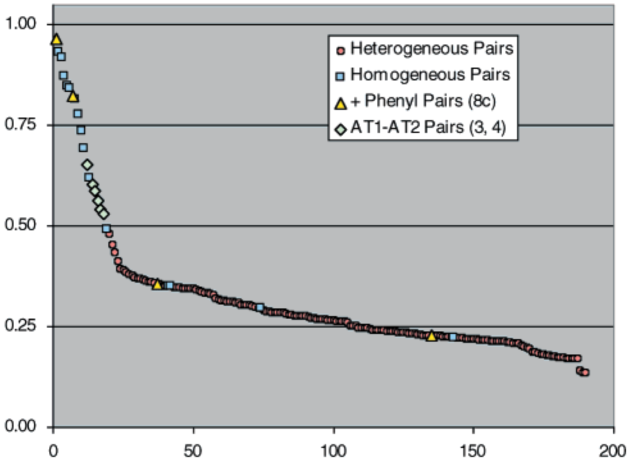


**Figure 9.** Ranking of molecules in Figure 5 by Tanimoto similarity score of 2D UNITY fingerprints.

intersect the shape and functionality of unfilled surfaces, a minimal set of "missing" 3D combinatorial templates can be deduced from the QSCD mapping of a given set of general screening compounds. These templates represent the smallest number of combinatorial syntheses which need to be executed in order to fill out the diversity of the set of screening compounds. One such template is depicted in Figure 12. In conjunction with the efficiency of core-based combinatorial chemistry,[1−4,44,45] QSCD makes possible the contemplation of a "complete" library of screening molecules at a given resolution. The model thus offers a theoretical and practical answer to the problem of generating lead structures for genomic targets of unknown structure and function. Specific uses of the model in combinatorial library design will be the subject of further correspondence.

**Extensions of the Model.** As has been noted previously,[25,46] an obvious extension of any diversity model based on an absolute frame of reference is that the same basis set may be used to classify actual proteins. By mapping onto the QSCD basis set all surfaces of volume *V* of a known protein, actual proteins can be compared and classified by their 3D binding sites. In addition to providing a rough diversity map of known protein binding sites, the theoretical surfaces of QSCD may thus be used to correlate protein classes to complementary molecular core structures. Comparison of known cocrystal structures to QSCD ligand-derived complementary surfaces will give solid benchmarks of the precision and predictive scope provided by a given set of QSCD parameters. An extension of the QSCD model to scan the known protein database is currently underway.

As described under the section on theoretical protein surfaces above, a 4.24 Å cube was found to be the largest predictive unit size of diversity measure for our criteria of designing general screening libraries. For example, both 4.48 and 4.00 Å units gave poorer prediction of homogeneous/heterogeneous pairs than the pairings of Figure 6 (4.24 Å units). This is likely due to the fact that most organic small molecules are themselves quantized by a limited basis set: the VDW radii of H, C, N, O, and a few other atoms (see, for example, Figure 1). If there is no constraint on size of cubic units, however (i.e. if there is no attempt to maximize orthogonality of theoretical target surfaces), other unit measures of diversity can be found. Most obviously, a unit of 2.12 Å should also provide effective diversity information but at a much higher resolution. Such a "high-resolution" adaptation of QSCD brings with it numerical (and thus computational) challenges. 112 negative space cubes (14 × 8) are now required at the upper limit of theoretical target surface size, translating to exponentially greater numbers of theoretical target surfaces and, depending on the stringency of fitting parameters, correspondingly greater numbers of surface fits per molecule as in Table 4. At this resolution, the assumption of no occlusions in theoretical target surfaces becomes far less valid, and removal of this assumption increases computational complexity further. A full analysis of whether the use of "high-resolution" QSCD is practical as a diversity tool compared to other high-resolution methods (e.g. ref 43) is a subject of current research.

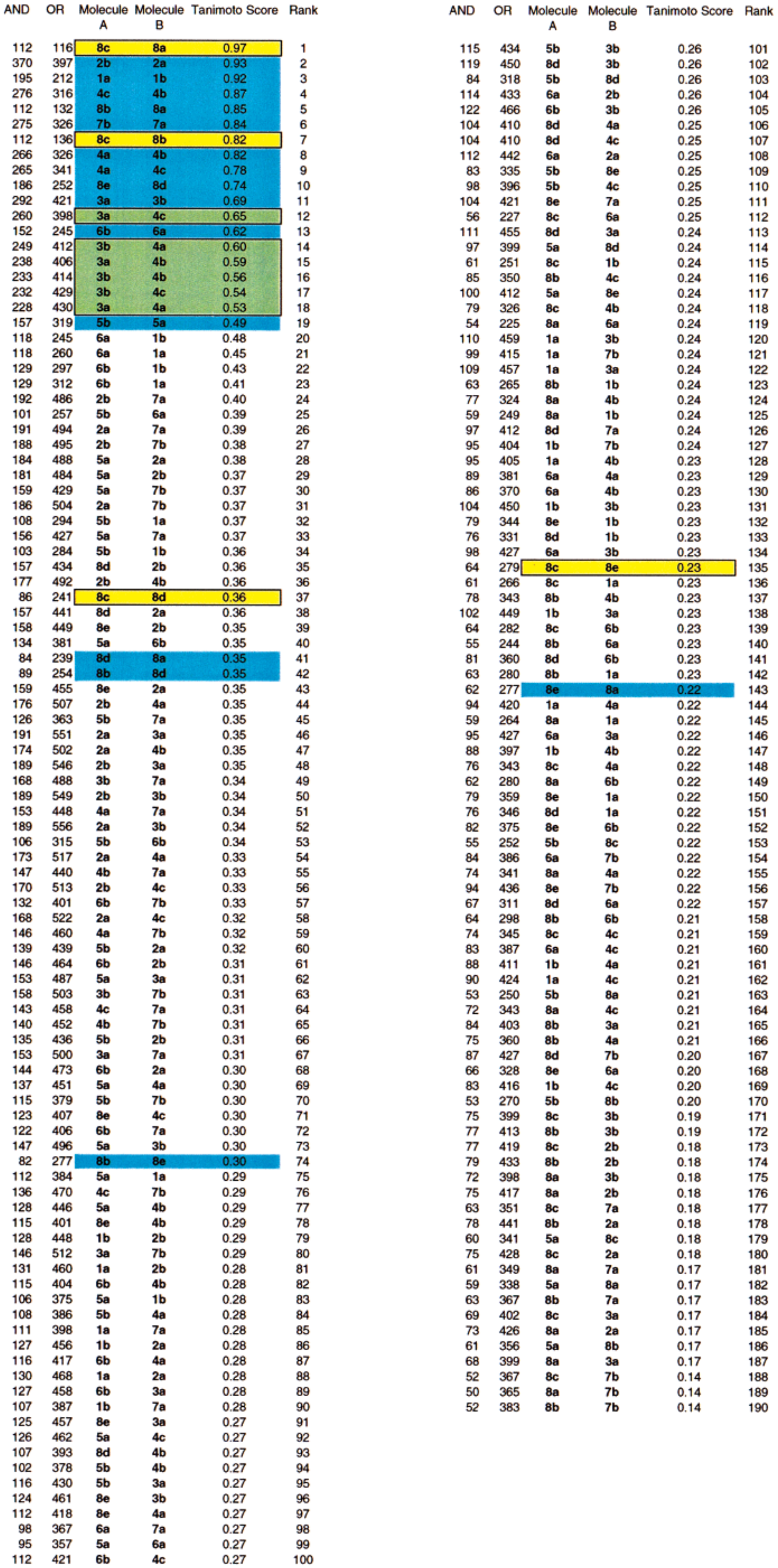
A final corollary of any absolute diversity model is a prediction of the total "size" of diversity space in terms of unique molecular points. In other words, what is the minimum set of molecules needed to fully cover a given diversity space. This calculation is dependent on two factors: the resolution stipulated in the model (e.g. what amount of molecular change is recognized as different) and the maximum values of each dimension of the model's basis axes. In the model of QSCD used herein, resolution is fixed by cubic units of 4.24 Å, and maximum values are fixed at 14 units (molecular volume of 1070 cubic Å) and four points of seven types of molecular property characteristics. As describe above, the result is a set of $1.1 \times 10^{14}$ unique molecular points. Since, using the parameters of this study, an average molecule covers 4.6 million of the unique molecular points bounded by QSCD space (Table 4), the model predicts a minimum of $(1.1 \times 10^{14})/(4.6 \times 10^{6}) = 24$ million molecules would be necessary to completely cover diversity space.

We estimate that an *average* complementary molecule in the context of the QSCD model used herein has a $\Delta G$ of complementarity on the order of −11 kcal (Table 7).[14] In other words, the resolution used to calculate diversity in this study translates roughly to nanomolar binding conditions for an average molecule/target surface pair. Given that some 24 million molecules are needed to completely cover diversity space under these conditions, a general screening library *guaranteed to contain at least one nanomolar binder to any given target of interest* would thus number at least 24 million molecules. This is a large number and will be attenuated by the fact that some molecules have significantly more than 100 conformations available to them. However, the QSCD model suggests that if, in the near future, combinatorial chemistry and high-throughput screening are to generate initial hits primarily in the nanomolar rather than micromolar range, then we must continue to focus our efforts on the development of numerically competent synthesis and screening technologies.[1−4,47,48]

## Conclusions

Like many other diversity models, QSCD is based on accepted tenets of molecular recognition, and its ability to group compounds that bind similar targets is one more piece of evidence that we are slowly progressing in our understanding of small molecule/target interac-

**Table 6.** Ranking of Molecules in Figure 5 by Tanimoto Similarity Score of 2D UNITY Fingerprints[a]



| AND | OR | Molecule A | Molecule B | Tanimoto Score | Rank |
|---|---|---|---|---|---|
| 112 | 116 | **8c** | **8a** | 0.97 | 1 |
| 370 | 397 | **2b** | **2a** | 0.93 | 2 |
| 195 | 212 | **1a** | **1b** | 0.92 | 3 |
| 276 | 316 | **4c** | **4b** | 0.87 | 4 |
| 112 | 132 | **8b** | **8a** | 0.85 | 5 |
| 275 | 326 | **7b** | **7a** | 0.84 | 6 |
| 112 | 136 | **8c** | **8b** | 0.82 | 7 |
| 266 | 326 | **4a** | **4b** | 0.82 | 8 |
| 265 | 341 | **4a** | **4c** | 0.78 | 9 |
| 186 | 252 | **8e** | **8d** | 0.74 | 10 |
| 292 | 421 | **3a** | **3b** | 0.69 | 11 |
| 260 | 398 | **3a** | **4c** | 0.65 | 12 |
| 152 | 245 | **6b** | **6a** | 0.62 | 13 |
| 249 | 412 | **3b** | **4a** | 0.60 | 14 |
| 238 | 406 | **3a** | **4b** | 0.59 | 15 |
| 233 | 414 | **3b** | **4b** | 0.56 | 16 |
| 232 | 429 | **3b** | **4c** | 0.54 | 17 |
| 228 | 430 | **3a** | **4a** | 0.53 | 18 |
| 157 | 319 | **5b** | **5a** | 0.49 | 19 |
| 118 | 245 | **6a** | **1b** | 0.48 | 20 |
| 118 | 260 | **6a** | **1a** | 0.45 | 21 |
| 129 | 297 | **6b** | **1b** | 0.43 | 22 |
| 129 | 312 | **6b** | **1a** | 0.41 | 23 |
| 192 | 486 | **2b** | **7a** | 0.40 | 24 |
| 101 | 257 | **5b** | **6a** | 0.39 | 25 |
| 191 | 494 | **2a** | **7a** | 0.39 | 26 |
| 188 | 495 | **2b** | **7b** | 0.38 | 27 |
| 184 | 488 | **5a** | **2a** | 0.38 | 28 |
| 181 | 484 | **5a** | **2b** | 0.37 | 29 |
| 159 | 429 | **5a** | **7b** | 0.37 | 30 |
| 186 | 504 | **2a** | **7b** | 0.37 | 31 |
| 108 | 294 | **5b** | **1a** | 0.37 | 32 |
| 156 | 427 | **5a** | **7a** | 0.37 | 33 |
| 103 | 284 | **5b** | **1b** | 0.36 | 34 |
| 157 | 434 | **8d** | **2b** | 0.36 | 35 |
| 177 | 492 | **2b** | **4b** | 0.36 | 36 |
| 86 | 241 | **8c** | **8d** | 0.36 | 37 |
| 157 | 441 | **8d** | **2a** | 0.36 | 38 |
| 158 | 449 | **8e** | **2b** | 0.35 | 39 |
| 134 | 381 | **5a** | **6b** | 0.35 | 40 |
| 84 | 239 | **8d** | **8a** | 0.35 | 41 |
| 89 | 254 | **8b** | **8d** | 0.35 | 42 |
| 159 | 455 | **8e** | **2a** | 0.35 | 43 |
| 176 | 507 | **2b** | **4a** | 0.35 | 44 |
| 126 | 363 | **5b** | **7a** | 0.35 | 45 |
| 191 | 551 | **2a** | **3a** | 0.35 | 46 |
| 174 | 502 | **2a** | **4b** | 0.35 | 47 |
| 189 | 546 | **2b** | **3a** | 0.35 | 48 |
| 168 | 488 | **3b** | **7a** | 0.34 | 49 |
| 189 | 549 | **2b** | **3b** | 0.34 | 50 |
| 153 | 448 | **4a** | **7a** | 0.34 | 51 |
| 189 | 556 | **2a** | **3b** | 0.34 | 52 |
| 106 | 315 | **5b** | **6b** | 0.34 | 53 |
| 173 | 517 | **2a** | **4a** | 0.33 | 54 |
| 147 | 440 | **4b** | **7a** | 0.33 | 55 |
| 170 | 513 | **2b** | **4c** | 0.33 | 56 |
| 132 | 401 | **6b** | **7b** | 0.33 | 57 |
| 168 | 522 | **2a** | **4c** | 0.32 | 58 |
| 146 | 460 | **4a** | **7b** | 0.32 | 59 |
| 139 | 439 | **5b** | **2a** | 0.32 | 60 |
| 146 | 464 | **6b** | **2b** | 0.31 | 61 |
| 153 | 487 | **5a** | **3a** | 0.31 | 62 |
| 158 | 503 | **3b** | **7b** | 0.31 | 63 |
| 143 | 458 | **4c** | **7a** | 0.31 | 64 |
| 140 | 452 | **4b** | **7b** | 0.31 | 65 |
| 135 | 436 | **5b** | **2b** | 0.31 | 66 |
| 153 | 500 | **3a** | **7a** | 0.31 | 67 |
| 144 | 473 | **6b** | **2a** | 0.30 | 68 |
| 137 | 451 | **5a** | **4a** | 0.30 | 69 |
| 115 | 379 | **5b** | **7b** | 0.30 | 70 |
| 123 | 407 | **8e** | **4c** | 0.30 | 71 |
| 122 | 406 | **6b** | **7a** | 0.30 | 72 |
| 147 | 496 | **5a** | **3b** | 0.30 | 73 |
| 82 | 277 | **8b** | **8e** | 0.30 | 74 |
| 112 | 384 | **5a** | **1a** | 0.29 | 75 |
| 136 | 470 | **4c** | **7b** | 0.29 | 76 |
| 128 | 446 | **5a** | **4b** | 0.29 | 77 |
| 115 | 401 | **8e** | **4b** | 0.29 | 78 |
| 128 | 448 | **1b** | **2b** | 0.29 | 79 |
| 146 | 512 | **3a** | **7b** | 0.29 | 80 |
| 131 | 460 | **1a** | **2b** | 0.28 | 81 |
| 115 | 404 | **6b** | **4b** | 0.28 | 82 |
| 106 | 375 | **5a** | **1b** | 0.28 | 83 |
| 108 | 386 | **5b** | **4a** | 0.28 | 84 |
| 111 | 398 | **1a** | **7a** | 0.28 | 85 |
| 127 | 456 | **1b** | **2a** | 0.28 | 86 |
| 116 | 417 | **6b** | **4a** | 0.28 | 87 |
| 130 | 468 | **1a** | **2a** | 0.28 | 88 |
| 127 | 458 | **6b** | **3a** | 0.28 | 89 |
| 107 | 387 | **1b** | **7a** | 0.28 | 90 |
| 125 | 457 | **8e** | **3a** | 0.27 | 91 |
| 126 | 462 | **5a** | **4c** | 0.27 | 92 |
| 107 | 393 | **8d** | **4b** | 0.27 | 93 |
| 102 | 378 | **5b** | **4b** | 0.27 | 94 |
| 116 | 430 | **5b** | **3a** | 0.27 | 95 |
| 124 | 461 | **8e** | **3b** | 0.27 | 96 |
| 112 | 418 | **8e** | **4a** | 0.27 | 97 |
| 98 | 367 | **6a** | **7a** | 0.27 | 98 |
| 95 | 357 | **5a** | **6a** | 0.27 | 99 |
| 112 | 421 | **6b** | **4c** | 0.27 | 100 |
| 115 | 434 | **5b** | **3b** | 0.26 | 101 |
| 119 | 450 | **8d** | **3b** | 0.26 | 102 |
| 84 | 318 | **5b** | **8d** | 0.26 | 103 |
| 114 | 433 | **6a** | **2b** | 0.26 | 104 |
| 122 | 466 | **6b** | **3b** | 0.26 | 105 |
| 104 | 410 | **8d** | **4a** | 0.25 | 106 |
| 104 | 410 | **8d** | **4c** | 0.25 | 107 |
| 112 | 442 | **6a** | **2a** | 0.25 | 108 |
| 83 | 335 | **5b** | **8e** | 0.25 | 109 |
| 98 | 396 | **5b** | **4c** | 0.25 | 110 |
| 104 | 421 | **8e** | **7a** | 0.25 | 111 |
| 56 | 227 | **8c** | **6a** | 0.25 | 112 |
| 111 | 455 | **8d** | **3a** | 0.24 | 113 |
| 97 | 399 | **5a** | **8d** | 0.24 | 114 |
| 61 | 251 | **8c** | **1b** | 0.24 | 115 |
| 85 | 350 | **8b** | **4c** | 0.24 | 116 |
| 100 | 412 | **5a** | **8e** | 0.24 | 117 |
| 79 | 326 | **8c** | **4b** | 0.24 | 118 |
| 54 | 225 | **8a** | **6a** | 0.24 | 119 |
| 110 | 459 | **1a** | **3b** | 0.24 | 120 |
| 99 | 415 | **1a** | **7b** | 0.24 | 121 |
| 109 | 457 | **1a** | **3a** | 0.24 | 122 |
| 63 | 265 | **8b** | **1b** | 0.24 | 123 |
| 77 | 324 | **8a** | **4b** | 0.24 | 124 |
| 59 | 249 | **8a** | **1b** | 0.24 | 125 |
| 97 | 412 | **8d** | **7a** | 0.24 | 126 |
| 95 | 404 | **1b** | **7b** | 0.24 | 127 |
| 95 | 405 | **1a** | **4b** | 0.23 | 128 |
| 89 | 381 | **6a** | **4a** | 0.23 | 129 |
| 86 | 370 | **6a** | **4b** | 0.23 | 130 |
| 104 | 450 | **1b** | **3b** | 0.23 | 131 |
| 79 | 344 | **8e** | **1b** | 0.23 | 132 |
| 76 | 331 | **8d** | **1b** | 0.23 | 133 |
| 98 | 427 | **6a** | **3b** | 0.23 | 134 |
| 64 | 279 | **8c** | **8e** | 0.23 | 135 |
| 61 | 266 | **8c** | **1a** | 0.23 | 136 |
| 78 | 343 | **8b** | **4b** | 0.23 | 137 |
| 102 | 449 | **1b** | **3a** | 0.23 | 138 |
| 64 | 282 | **8c** | **6b** | 0.23 | 139 |
| 55 | 244 | **8b** | **6a** | 0.23 | 140 |
| 81 | 360 | **8d** | **6b** | 0.23 | 141 |
| 63 | 280 | **8b** | **1a** | 0.23 | 142 |
| 62 | 277 | **8e** | **8a** | 0.22 | 143 |
| 94 | 420 | **1a** | **4a** | 0.22 | 144 |
| 59 | 264 | **8a** | **1a** | 0.22 | 145 |
| 95 | 427 | **6a** | **3a** | 0.22 | 146 |
| 88 | 397 | **1b** | **4b** | 0.22 | 147 |
| 76 | 343 | **8c** | **4a** | 0.22 | 148 |
| 62 | 280 | **8a** | **6b** | 0.22 | 149 |
| 79 | 359 | **8e** | **1a** | 0.22 | 150 |
| 76 | 346 | **8d** | **1a** | 0.22 | 151 |
| 82 | 375 | **8e** | **6b** | 0.22 | 152 |
| 55 | 252 | **5b** | **8c** | 0.22 | 153 |
| 84 | 386 | **6a** | **7b** | 0.22 | 154 |
| 74 | 341 | **8a** | **4a** | 0.22 | 155 |
| 94 | 436 | **8e** | **7b** | 0.22 | 156 |
| 67 | 311 | **8d** | **6a** | 0.22 | 157 |
| 64 | 298 | **8b** | **6b** | 0.21 | 158 |
| 74 | 345 | **8c** | **4c** | 0.21 | 159 |
| 83 | 387 | **6a** | **4c** | 0.21 | 160 |
| 88 | 411 | **1b** | **4a** | 0.21 | 161 |
| 90 | 424 | **1a** | **4c** | 0.21 | 162 |
| 53 | 250 | **5b** | **8a** | 0.21 | 163 |
| 72 | 343 | **8a** | **4c** | 0.21 | 164 |
| 84 | 403 | **8b** | **3a** | 0.21 | 165 |
| 75 | 360 | **8b** | **4a** | 0.21 | 166 |
| 87 | 427 | **8d** | **7b** | 0.20 | 167 |
| 66 | 328 | **8e** | **6a** | 0.20 | 168 |
| 83 | 416 | **1b** | **4c** | 0.20 | 169 |
| 53 | 270 | **5b** | **8b** | 0.20 | 170 |
| 75 | 399 | **8c** | **3b** | 0.19 | 171 |
| 77 | 413 | **8b** | **3b** | 0.19 | 172 |
| 77 | 419 | **8c** | **2b** | 0.18 | 173 |
| 79 | 433 | **8b** | **2b** | 0.18 | 174 |
| 72 | 398 | **8a** | **3b** | 0.18 | 175 |
| 75 | 417 | **8a** | **2b** | 0.18 | 176 |
| 63 | 351 | **8c** | **7a** | 0.18 | 177 |
| 78 | 441 | **8b** | **2a** | 0.18 | 178 |
| 60 | 341 | **5a** | **8c** | 0.18 | 179 |
| 75 | 428 | **8c** | **2a** | 0.18 | 180 |
| 61 | 349 | **8a** | **7a** | 0.17 | 181 |
| 59 | 338 | **5a** | **8a** | 0.17 | 182 |
| 63 | 367 | **8b** | **7a** | 0.17 | 183 |
| 69 | 402 | **8c** | **3a** | 0.17 | 184 |
| 73 | 426 | **8a** | **2a** | 0.17 | 185 |
| 61 | 356 | **5a** | **8b** | 0.17 | 186 |
| 68 | 399 | **8a** | **3a** | 0.17 | 187 |
| 52 | 367 | **8c** | **7b** | 0.14 | 188 |
| 50 | 365 | **8a** | **7b** | 0.14 | 189 |
| 52 | 383 | **8b** | **7b** | 0.14 | 190 |

[a] Color code: blue = homogeneous pairs, yellow = +phenyl pairs (**8c**), green = AT1 and AT2 pairs (**3**, **4**).

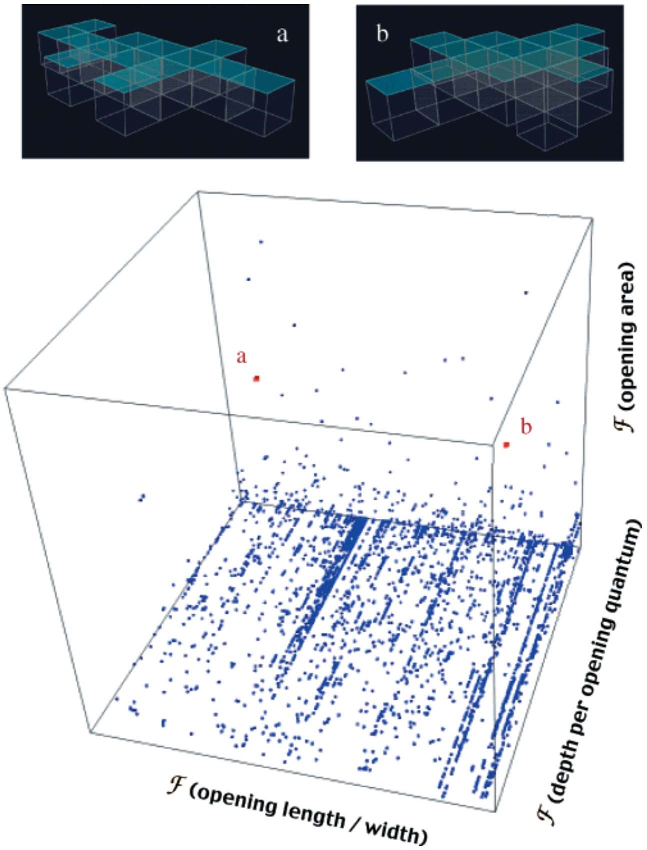


**Figure 10.** QSCD plot of all of the theoretical surface shapes covered by all of the conformations of all of the molecules shown in Figure 5 (blue dots). The total volume of the cube encompasses all 49 268 918 theoretical surface shapes as listed in Table 1. Red dots show two exemplary theoretical surface shapes (a, b) not covered by any of the molecules in Figure 5. Axes used are functions of opening area, opening length/width, and depth per opening quantum.

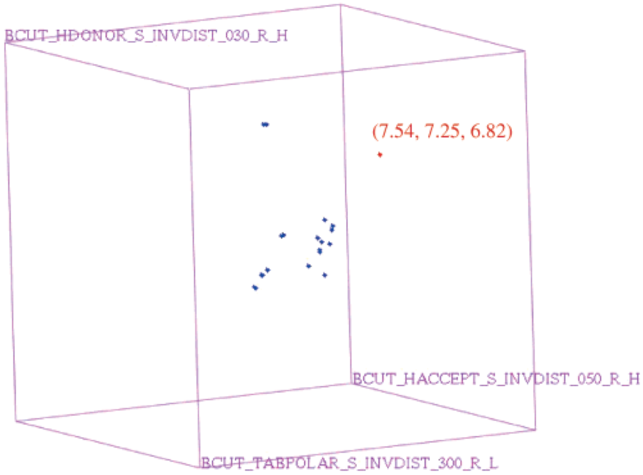


**Figure 11.** Map of the 20 compounds in Figure 5 (blue dots) in a representative BCUT three-axis diversity space.[27-29] BCUT axes used are, respectively: (1) BCUT HACCEPT S INVDIST 050 R H, (2) BCUT HDONOR S INVDIST 030 R H, and (3) BCUT TABPOLAR S INVDIST 300 R L. Red dot shows an unfilled coordinate of diversity space, at (7.54, 7.25, 6.82). The information contained in this BCUT coordinate does not reveal information about the shape of a molecule which might be able to fill this position in diversity space.

tions. QSCD's singular advantage is that it is constructed on a basis set that allows for the reversible flow of diversity information: by defining molecular diversity through a fixed reference frame of spatially and functionally enumerated 3D surfaces, the model determines not only diversity of existing structures but also structures of nonexisting diversity. In addition to numerically defining the absolute percentage of diversity space covered by any given set of screening molecules, QSCD allows for the rational and systematic population of the remaining, *unfilled* diversity space.

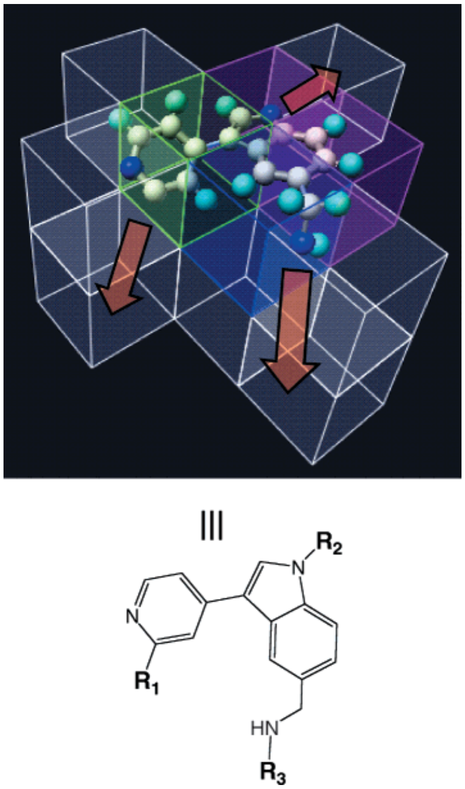


**Figure 12.** Use of QSCD to design complementary combinatorial libraries to unmatched theoretical target surfaces. Note that many conceivable libraries of a given shape and functionality may be designed to fill a given unmet diversity need.

**Table 7.** Summation of Binding Energies for an Interaction of an *Average* Complementary Molecule/Theoretical Target Surface Pair in the Context of the QSCD Model Used Herein[a]

| energetic contribution | av $\Delta G$ (kcal/mol) |
|---|---|
| translational/vibrational entropic loss (constant) | +9 |
| constant +0.7 kcal/mol ($RT$ ln 3) per rotatable bond; assume rigid theoretical target surface | +7 |
| $\Delta\Delta G$ conformation from ground state | +2 |
| constant −0.03 kcal/mol/Å$^2$ nonpolar buried surface = −0.54 kcal/mol/4.24 Å$^2$ nonpolar buried face; total 21 molecular faces + 21 theoretical surface faces | −23 |
| total interaction from Table 2 (four complementary points) | −6.0 |
| sum of binding energies | −11 |
| binding affinity to nearest integer (÷1.363) | $10^{-8}$ = 10 nM |

[a] An average molecule is assumed to have a buried volume of 12 cubic quanta (=915 cubic Å at 4.24 Å resolution), 36 exposed faces (4.24 Å$^2$), 21 nonpolar exposed faces (60%), 10 rotatable bonds, 4 points of complementary electrostatic/VDW potential, and a conformational energy within 2 kcal/mol of ground state. An average complementary theoretical target surface is also assumed to have 60% nonpolar exposed faces. Constants used in the table are taken from Ajay and Murcko.[14]

The model suggests that in order to ensure nanomolar ligands to any given target, a library of at least 24 million molecules will be required. While there is much computational and synthetic effort that stands between current combinatorial libraries and a general screening library of 20+ million molecules, the ever increasing speed of computer processors and the present ability to synthesize million-member libraries leave little doubt that such a feat can be achieved. Combined with the rapid miniaturization and efficiency of screening techniques, diversity models such as QSCD should help to bridge the gap between the myriad achievements of genomics and the next generation of small molecule therapeutics.

## Experimental Section

Molecular conformations were generated with Multisearch in Sybyl (version 6.5; Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144) on an R10000 Silicon Graphics workstation. Conformations were subsequently sorted by energy and conformations within 10 kcal of the lowest energy were accepted. Overlay plots of molecules (Figure 8B) were also generated using Sybyl. UNITY 2D fingerprints (Unity 4.0; Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144) were generated on an R10000 Silicon Graphics workstation. Pairwise Tanimoto coefficients were computed as described by Dixon and Koehler.[39] QSCD software for molecule quantization, mapping of Q-files, and surface complementarity display was developed using the Java programming language (JDK 1.2) and the Java3D graphics API (version 1.1) on Intel-based workstations. Theoretical target surfaces were stored and indexed using an Oracle 7.3.3 database. Parameters for theoretical target surface generation/molecular quantization and parameters for complementarity mapping/scoring were alternately optimized in three successive rounds as below.

The parameters used for theoretical target surface generation and the closely related parameters for quantization of small molecules into quantized files (Q-files) were optimized in the context of the algorithms stated in the text. Parameters were iteratively optimized by varying a given parameter and then quantizing training molecules other than those in Figure 5. Training molecules used were taken from in house structures and two published SAR sets.[49,50] Concomitant with molecular quantization, an enumerated set of theoretical target surfaces was created with corresponding parameters. Using the current optimized complementarity/scoring parameters, molecules were then mapped to theoretical target surfaces and all diversity pairing scores generated as described in the text. Parameters were chosen which accurately predicted known homogeneous/heterogeneous pairs and which maximized “signal-to-noise” of homogeneous scores over heterogeneous scores.

The parameters used for mapping/scoring molecular conformations to theoretical target surfaces were optimized in the context of the algorithm stated in the text. Parameters were iteratively optimized by varying a given parameter and then mapping a constant set of training molecules (see above) to a constant set of theoretical target surfaces, using the most current surface generation and quantization parameters. Diversity pairing scores were generated for all training molecules, and parameters were chosen which accurately predicted known homogeneous/heterogeneous pairs and which maximized “signal-to-noise” of homogeneous scores over heterogeneous scores.

As mentioned in the text, for a given conformation-to-surface shape fit to be accepted, the minimum overlap requirement was set to either 9 quanta *or* $N - 2$ quanta of a conformation of $N$ quanta. This range allows large conformations to fit partially into a theoretical surface (protruding volume must be at the mouth of the surface) while also allowing smaller conformations to be considered for complementarity. It excludes large conformations which do not overlap at least 9 quanta.

Approximate computational speeds of typical QSCD operations are as follows on a single Pentium III 500 MHz workstation: generation of the basis set of theoretical target surface used in the study required 17 min; these data were stored for access by subsequent QSCD functions. Quantization of 100 conformations of a given molecule into 100 Q-files required 250 s. Complementarity mapping of 100 Q-files onto the basis set of theoretical target surfaces used in the study required 40 s.

**Acknowledgment.** We are indebted to Greg Makara for his insights into the nature of molecular diversity and his assistance with the generation of molecular conformations. We thank Tony Hopfinger and Julius Rebek for cogent discussions on the subject of molecular recognition, Huw Nash for critical assessment of the manuscript, and Tomas Revesz for IT support. Finally, we thank Allen Michels for posing the questions of molecular diversity which led to this work.

## References

(1) Tan, D. S.; Foley, M. A.; Shair, M. D.; Schreiber, S. L. Stereoselective synthesis of over two million compounds having structural features both reminiscent of natural products and compatible with miniaturized cell-based assays. *J. Am. Chem. Soc.* **1998**, *120*, 8565−8566.

(2) Boger, D. L.; Jiang, W.; Goldberg, J. Convergent solution-phase combinatorial synthesis with multiplication of diversity through rigid biaryl and diarylacetylene couplings. *J. Org. Chem.* **1999**, *64*, 7094−7100.

(3) Carell, T.; Wintner, E. A.; Sutherland, A. J.; Dunayevskiy, Y.; Vouros, P.; Rebek, J., Jr. New promise in combinatorial chemistry: Synthesis, characterization, and screening of small molecule libraries in solution. *Chem. Biol.* **1995**, *2*, 171−183.

(4) An, H.; Cummins, L. L.; Griffey, R. H.; Bharadwaj, R.; Haly, B. D.; Fraser, A. S.; Wilson-Lingardo, L.; Risen, L. M.; Wyatt, J. R.; Cook, P. D. Solution phase combinatorial chemistry. Discovery of novel polyazapyridinophanes with potent antibacterial activity by a solution phase simultaneous addition of functionalities approach. *J. Am. Chem. Soc.* **1997**, *119*, 3696−3708.

(5) Mitchison, T. J. Towards a pharmacological Genetics. *Chem. Biol.* **1994**, *1*, 3−6.

(6) Bartlett, P. A.; Joyce, G. F. Combinatorial chemistry: The search continues. *Curr. Opin. Chem. Biol.* **1999**, *3*, 253−255.

(7) Drews, J. Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* **1996**, *14*, 1516−1518.

(8) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(9) Gaasterland, T. Structural genomics: Bioinformatics in the driver's seat. *Nat. Biotechnol.* **1998**, *16*, 625−627.

(10) Kauvar, L. M.; Laborde, E. The diversity challenge in combinatorial chemistry. *Curr. Opin. Drug Discov. Dev.* **1998**, *1*, 66−70.

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

(13) Polinsky, A. Combichem and cheminformatics. *Curr. Opin. Drug Discov. Dev.* **1999**, *2*, 197−203.

(14) Ajay; Murcko, M. A. Computational methods to predict binding free energy in ligand−receptor complexes. *J. Med. Chem.* **1995**, *38*, 4953−4967.

(15) Klebe, G.; Böhm, H.-J. Energetic and entropic factors determining binding affinity in protein−ligand complexes. *J. Recept. Signal Transduction Res.* **1997**, *17*, 459−473.

(16) Fersht, A. R. The hydrogen bond in molecular recognition. *Trends Biochem. Sci.* **1987**, *12*, 301−304.

(17) Tokarski, J. S.; Hopfinger, A. J. Prediction of ligand−receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: application to a set of peptidomimetic renin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792−811.

(18) Creighton, T. E. *Proteins: Structures and Molecular Properties*; W. H. Freeman and Co.: New York, 1984.

(19) Mecozzi, S.; Rebek, J., Jr. The 55% solution: A formula for molecular recognition in the liquid state. *Chem. Eur. J.* **1998**, *4*, 1016−1022.

(20) So, S.-S.; Karplus, M. A comparative study of ligand−receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 243−258.

(21) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein–ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791–804.

(22) Burkhard, P.; Taylor, P.; Walkinshaw, M. D. An example of a protein ligand found by database mining: Description of the docking method and its verification by a 2.3 Å X-ray structure of a thrombin-ligand complex. *J. Mol. Biol.* **1998**, *277*, 449–466.

(23) Norel, R.; Petrey, D.; Wolfson, H. J.; Nussinov, R. Examination of shape complementarity in docking of unbound proteins. *Proteins* **1999**, *36*, 307–317.

(24) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884–1897.

(25) Mason, J. S.; Hermsmeier, M. A. Diversity assessment. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342–349.

(26) Warr, W. A. Commercial software systems for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 115–130.

(27) University of Texas. *DiverseSolutions User's Manual version 3.0.2*; Laboratory for Molecular Graphics and Theoretical Modeling, Distributed by Tripos, Inc.: 1699 S. Hanley Rd., St. Louis, MO 63144, 1997.

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

(29) Menard, P. R.; Mason, J. S.; Morize, I.; Bauerschmidt, S. Chemistry space metrics in diversity analysis, library design, and compound selection. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1204–1213.

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

(31) Good, A. C.; Lewis, R. A. New methodology for profiling combinatorial libraries and screening sets: Cleaning up the design process with HARPick. *J. Med. Chem.* **1997**, *40*, 3926–3936.

(32) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity profiling and design using 3D pharmacophores: Pharmacophore derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1204–1213.

(33) Parks, C. A.; Crippen, G. M.; Topliss, J. G. The measurement of molecular diversity by receptor site interaction simulation. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 441–449.

(34) *Chem-X software*; Oxford Molecular: Medawar Center, Oxford Science Park, Oxford OX4 4GA, England.

(35) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107–118.

(36) Dixon, S. L.; Villar, H. O. Bioactive diversity and screening library selection via affinity fingerprinting. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1192–1203.

(37) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401–3408.

(38) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376–380.

(39) Dixon, S. L.; Koehler, R. T. The hidden component of size in two-dimensional fragment descriptors: Side effects on sampling in bioactive libraries. *J. Med. Chem.* **1999**, *42*, 2887–2900.

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

(41) Jiang, F.; Kim, S. H. "Soft Docking": Matching of molecular surface cubes. *J. Mol. Biol.* **1991**, *219*, 79–102.

(42) Bemis, G. W.; Murcko, M. A. The properties of known drugs: 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(43) Mount, J.; Ruppert, J.; Welch, W.; Jain, A. N. IcePick: A flexible surface-based system for molecular diversity. *J. Med. Chem.* **1999**, *42*, 60–66.

(44) Marx, M. A.; Grillot, A.-L.; Louer, C. T.; Beaver, K. A.; Bartlett, P. A. Synthetic design for combinatorial chemistry. Solution and polymer supported synthesis of polycyclic lactams by intramolecular cyclization of azomethine ylides. *J. Am. Chem. Soc.* **1997**, *119*, 6153–6167.

(45) Boojamra, C. G.; Burrow, K. M.; Thompson, L. A.; Ellman, J. A. Solid-phase synthesis of 1,4-benzodiazepine-2,5-diones. Library preparation and demonstration of synthesis generality. *J. Org. Chem.* **1997**, *62*, 1240–1256.

(46) Wallace, A. C.; Borkakoti, N.; Thornton, J. M. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **1997**, *6*, 2308–2323.

(47) MacBeath, G.; Koehler, A. N.; Schreiber, S. L. Printing small molecules as microarrays and detecting protein–ligand interactions en masse. *J. Am. Chem. Soc.* **1999**, *121*, 7967–7968.

(48) Lam, K. S.; Lebl, M.; Krchňák, V. The "one-bead-one-compound" combinatorial library method. *Chem. Rev.* **1997**, *97*, 411–448.

(49) Lewis, R. T.; Macleod, A. M.; Merchant, K. J.; Kelleher, F.; Sanderson, I.; Herbert, R. H.; Cascieri, M. A.; Sadowski, S.; Ball, R. G.; Hoogsteen, K. Tryptophan-derived $NK_1$ antagonists: Confomationally constrained heterocyclic bioisosteres of the ester linkage. *J. Med. Chem.* **1995**, *38*, 923–933.

(50) Depreux, P.; Lesieur, D.; Mansour, H. A.; Morgan, P.; Howell, H. E.; Renard, P.; Caignard, D.-H.; Pfeiffer, B.; Delagrange, P.; Guardiola, B.; Yous, S.; Demarque, A.; Adam, G.; Andrieux, J. Synthesis and structure–activity relationships of novel naphthalenic and bioisosteric related amidic derivatives as melatonin receptor ligands. *J. Med. Chem.* **1994**, *37*, 3231–3239.

JM990504B