MOLMAKER: De Novo Generation of 3D Databases for Use in Drug Design

David E. Clark,* Mike A. Firth, and Christopher W. Murray

Proteus Molecular Design Ltd., Proteus House, Lyme Green Business Park, Macclesfield, Cheshire, SK11 0JL, United Kingdom

Received September 3, 1995[⊗]

A program, MOLMAKER, is described which, in conjunction with a 2D−3D conversion program and 3D database software, can generate *de novo* 3D databases to aid in drug design. MOLMAKER is based upon graph-theoretical techniques for vertex degree set generation and constructive enumeration of molecular graphs. The generated molecular graphs are then functionalised in a probabilistic manner but in accordance with various constraints specified by the user. The resulting connection tables can be converted into 3D structures by commercial software and loaded into a 3D database for pharmacophore searching. The utility of MOLMAKER is illustrated by two examples of interest from the recent scientific literature: the design of novel protein kinase C agonists and of a bridging ligand for cyclophilin-calcineurin.

INTRODUCTION

The last five years have witnessed a continuing rise in the development and application of novel computational technologies to the process of drug design and discovery. One of the technologies which has excited particular interest is that of searching algorithms which enable rapid access to databases of three-dimensional chemical structures, particularly with the recent advent of "flexible searching" in which the conformational flexibility of the stored structures is accounted for.[1−6] 3D database searching has already shown potential as a useful tool in molecular design aiding in the discovery of novel HIV-1 protease inhibitors,[7] protein kinase C agonists[8] and angiotensin II antagonists,[9] to give just a few recent examples from the literature.

The effectiveness of 3D searching as a drug discovery tool is dependent not merely upon the quality of the searching algorithms but also upon the content of the databases available for searching. The great majority of 3D databases in current use have been generated from commercial or proprietary compound collections by fast, approximate rule-based methods such as CONCORD[10] or Chem-X.[11] The main exception to this rule is the Cambridge Structural Database (CSD)[12] which contains 3D structures determined by X-ray crystallography or neutron diffraction. While current databases are attractive in that one can be fairly certain that any retrieved "hits" will be synthetically accessible or even available for testing, they are also subject to certain limitations.

Firstly, searching current databases will never discover a "new" compound *per se*, it merely suggests a new use for an existing one. Secondly, even the largest collections will be biased toward particular types of compound. The bias may result from, among other things, the historic interest of a company in a given therapeutic area or the academic interests of those depositing crystal structures. Taken together, these represent a considerable restriction on the present potential of 3D searching for drug discovery. Such limitation may be part of the reason for the burgeoning interest in combinatorial chemistry.[13,14] The concept of "molecular diversity" is becoming popular in the search for novel lead compounds as researchers realize what a small fraction of "compound space" is represented even by the union of all available databases.

It has been recently stated[15] that "the utility of 3D database searching is greatly enhanced if databases of hitherto-unknown compounds can be searched". In this paper, we present a method for generating databases of such "hitherto-unknown compounds" and, in so doing, seek to address some of the current limitations in 3D databases. Specifically, we present a program, MOLMAKER, for the *de novo* generation of sizeable structural databases rich in structural novelty but also tailorable to the user's specifications. MOLMAKER forms an integral part of our in-house software system for molecular design and simulation, PROMETHEUS. The paper begins with the history behind MOLMAKER and an overview of the program's functionality. Each step in the process of database generation is then explained in detail before two examples which demonstrate the usefulness of MOLMAKER in molecular design applications.

OVERVIEW OF MOLMAKER PROGRAM.

The idea of generating novel chemical structure databases as an aid in drug discovery has been previously investigated by a few groups of workers. Nilakantan *et al*. developed a method based upon the random fusion of 2D chemical fragments.[16] The resulting 2D structures were then filtered using statistical structure−activity models to ensure that the retained molecules were of relevance to the design problem under study and stored in a MACCS-II database for future retrieval and reference. In terms of 3D databases, an early attempt at novel molecule generation was developed by Martin and Van Drie in their MODSMI program.[17] MODSMI automatically modifies the SMILES[18] strings of "scaffold" molecules retrieved from a database so that they meet the functional as well as geometric requirements of the specified pharmacophore.[19] The modified SMILES strings were then submitted to CONCORD[10] for 3D structure generation. The program proved effective in the development of novel D2 agonists.[20] Mason has recently reported using a slightly different SMILES manipulation procedure for elaborating upon an initial hit compound for angiotensin II antagonism,

* To whom all correspondence should be addressed.

[⊗] Abstract published in *Advance ACS Abstracts,* January 1, 1996.

0095-2338/96/1636-0137$12.00/0

again with encouraging results.[6] The SMILES notation has also formed the basis for Ho and Marshall's DBMAKER, a set of programs for the generation of 3D databases based upon user-specified criteria.[21] DBMAKER generates random compounds within constraints supplied by the user and outputs these as SMILES strings which are then converted to 3D coordinates by CONCORD.[10] Supporting modules, DBCYCLE and DBCROSS, allow the generation of cyclic compounds and the "splicing" of novel fragments to create larger entities, respectively. In addition, Glen and Payne[22] have recently described a genetic algorithm-based program which can be used to generate novel, structurally-diverse 3D databases, although no detailed illustrations of its use in this regard were given. Finally, the CAVEAT program developed by Bartlett and co-workers[23] also employs 3D databases to aid in the process of *de novo* design and has proved effective in a number of examples, e.g., refs 24 and 25.

The inspiration for MOLMAKER came from a recent paper by Hall and Fisk[26] in which they presented a novel algorithm for the generation of all *vertex degree sets* consistent with a given number of nodes and cycles. A vertex degree set describes the *degree* of the nodes in a graph, where the degree of a node is the number of nodes to which it is connected. Thus, for instance, assuming a maximum degree of 4 and a descending order of degree from left to right, the vertex degree set {0,1,3,1} describes all graphs having 1 node of degree 3, 3 nodes of degree 2, and 1 node of degree 1. Note that in many cases, there will be more than one graph that satisfies a given vertex degree set. Hall and Fisk suggested that coupling their method with an algorithm for the generation of molecular graphs could prove of utility in *de novo* drug design *inter alia*. On the basis of this suggestion, we developed MOLMAKER. The program itself consists of three basic functionality blocks which interface with commercial software for 2D−3D conversion and database creation and searching. The steps involved in the use of MOLMAKER are thus the following (1) vertex degree set generation, (2) molecular graph generation, (3) molecular graph functionalization, (4) 2D−3D conversion, and (5) database creation and searching. This process is illustrated in the flowchart given in Figure 1. In what follows, each of these steps will be described in detail after the input parameters to the program have been introduced.

INPUT PARAMETERS

MOLMAKER's operation is controlled by a number of user-specified parameters which are read from an input file. A typical input file is shown in Figure 2. These parameters will be explained fully under the relevant section of the program description.

VERTEX DEGREE SET GENERATION

The paper of Hall and Fisk[26] is built upon earlier work on the so-called "inverse QSAR" process[27−29] which seeks to invert QSAR equations to generate structures consistent with a target activity range. Drawing on equations developed in the earlier papers, Hall and Fisk developed and published a simple algorithm to generate all vertex degree sets for a given number of nodes and cycles. This algorithm was encoded directly from the paper and used in MOLMAKER.

The vertex degree set generation algorithm takes as input the values of the parameters MIN_HEAVY_ATOMS and MAX_HEAVY_ATOMS which control the number of nodes in the generated degree sets and also the values of MIN_RINGS and MAX_RINGS which determine the minimum and maximum number of rings permitted in the graphs implied by the degree sets. The vertex degree sets generated are stored in an array and are processed either sequentially (if RANDOM_VSET_SELECT is specified as "N") or randomly (RANDOM_VSET_SELECT is specified as "Y") by the remainder of the program which operates in a cascade-like manner as shown in the pseudocode description in Figure 3. Thus each vertex degree set gives rise to a number of graphs which are then passed on to the functionalization routine where each graph may give rise to a number of molecules. A simple illustration of this is given in Figure 4.

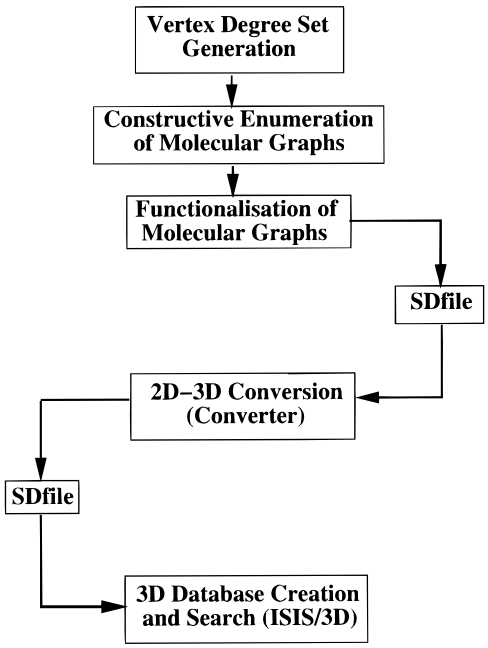

Figure 1. Flowchart giving an overview of MOLMAKER and its interface with Converter and ISIS/3D.

```
MAX_STRUCTURES       2500          MIN_EXO_TRPL_BONDS   0
MAX_GRAPHS_PER_VSET  100000        MAX_EXO_TRPL_BONDS   0
MAX_MOLS_PER_GRAPH   25            NUM_HETERO_TYPES     3
MIN_HEAVY_ATOMS      15            MAX_PERCENT_HETERO   25.0
MAX_HEAVY_ATOMS      15            PERMITTED_HET_TYPE   O  3  4
MIN_RINGS            2             PERMITTED_HET_TYPE   N  0  2
MAX_RINGS            3             PERMITTED_HET_TYPE   S  0  1
ALLOW_BRIDGED_RINGS  N             SUBS_FILE            SMILES.SUB
ALLOW_SPIRO_RINGS    N             H_DONOR_FILE         H_DONOR.SUB
MIN_RING_SIZE        5             H_ACCEPTOR_FILE      H_ACCEPTOR.SUB
MAX_RING_SIZE        6             MINIMUM_ACCEPTORS    0
MIN_AROM_RINGS       0             MINIMUM_DONORS       0
MAX_AROM_RINGS       3             OUTPUT_FILE          pkc15.sd
MIN_EXO_DBLE_BONDS   1             RANDOM_SEED          1
MAX_EXO_DBLE_BONDS   3             PRINT_OPTION         1
MIN_ALI_DBLE_BONDS   0             MAX_ATTEMPTS_FUNC    500
MAX_ALI_DBLE_BONDS   2             RANDOM_VSET_SELECT   N
```

Figure 2. Input file for MOLMAKER job showing program parameters.

```
generate_vertex_sets(vsets,nvsets)
do i=1, nvsets
  generate_graphs(vsets(i),graphs,ngraphs)
  do j=1, ngraphs
    generate_mols(graphs(j))
  enddo
enddo
```

Figure 3. Pseudo-code description of MOLMAKER algorithm.

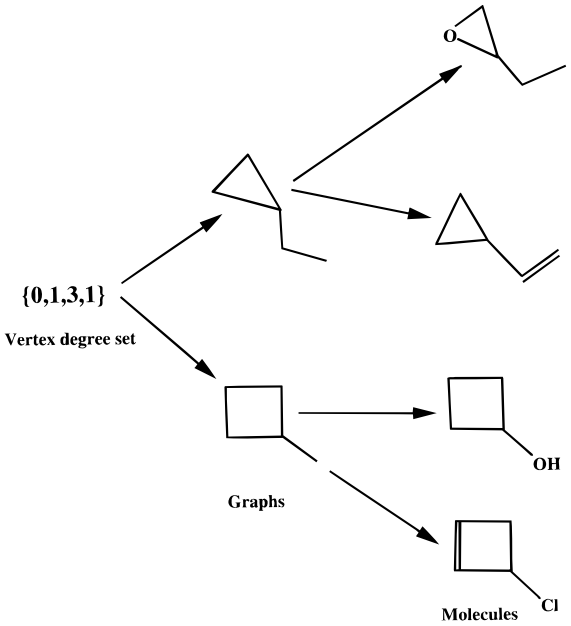

Figure 4. Schematic illustration of MOLMAKER's operation.

GENERATION OF MOLECULAR GRAPHS

Given a number of vertex degree sets, the next step is to generate from each vertex degree set all the molecular graphs consistent with it; in graph-theoretical terms that is known as *constructive enumeration*. For a given vertex degree set, up to MAX_GRAPHS_PER_VSET graphs may be generated and stored.

One of the main difficulties in the constructive enumeration of molecular graphs is that a given graph of n nodes may be labeled in $n!$ ways. In other words, there are $n!$ possible adjacency matrices corresponding to a particular graph which differ only in the arrangement of the node labels. A brute force algorithm for constructive enumeration will necessarily discover all these, but in general one is interested in only one representative adjacency matrix for each graph.

The algorithm for constructive enumeration employed in MOLMAKER is based on that of Kvasnička and Pospíchal.[30−32] A crucial element of their approach is the concept of *semicanonical* labeling of graphs (see Appendix). Enforcing the criterion that all generated adjacency matrices be semicanonical places a strong constrain upon the number of graphs that need to be tested for canonical labeling. In the work of Kvasnička and Pospíchal, this canonicity check is an integral part of the graph generation algorithm. At present in MOLMAKER, we have not implemented such an integral check; instead we generate a number of semicanonically labeled graphs for a given vertex set and check each one for uniqueness using the extended adjacency matrix indices (EAMIs) of Yang *et al*.[33] While this approach is comparatively unsophisticated and somewhat inefficient, it is at least sufficient to allow us to test the utility of the method overall.

As well as checking for duplicate graphs, checks are also present to allow the discarding of graphs based on ring size and the nature of ring systems present. Thus the user can specify that spiro and/or bridged (i.e., rings with two or more edges in common) systems be filtered out using the ALLOW_SPIRO_RINGS and ALLOW_BRIDGED_RINGS parameters and also that only rings over a specified range of sizes be included in the graphs produced by means of the MIN_RING_SIZE and MAX_RING_SIZE parameters.

FUNCTIONALIZATION OF MOLECULAR GRAPHS

The molecular graphs produced by the process described above can be considered as saturated, hydrogen-depleted carbon skeletons. Thus, for most applications, one will be interested in functionalizing these in some way to produce potential candidates for further study. MOLMAKER adopts a multistage, probabilistic approach to functionalization; this will be described in what follows.

Molecular Perception. When a molecular graph is passed to the functionalization routine, its structure is first perceived and ring and exocyclic bonds distinguished and stored. Ring bonds are perceived by creating a breadth-first spanning tree from an arbitrary root node and then detecting all the chords. Backtracking from each of the chords identifies all bonds involved in rings. Each vertex is assigned as initial valence equal to its degree and a maximum valence of 4. All bond orders are set to 1.

Aromatic Rings. If the MIN_AROM_RINGS parameter is greater than zero, the algorithm then examines each of the rings in the graph to see if they have the potential to be made aromatic. This is judged using fairly simple criteria: the ring must be of size 5 or 6 and have no vertex of degree greater than 3. Each potential aromatic ring is marked accordingly. If the number of possible aromatics, *NPOSS*, is less than MIN_AROM_RINGS, the algorithm aborts the current graph and moves on to the next available one.

To decide which of the potential aromatic rings are to be aromatized, the program calculates a probability, P_{arom}, based upon the number of potential aromatic rings and the maximum and minimum number permissible in a structure, as defined by the user

$$P_{arom} = \frac{MAX\_AROM\_RINGS + MIN\_AROM\_RINGS}{2 \times NPOSS} \quad (1)$$

Thus, if there are two potential aromatic rings and the user specifies that the structure should contain only one (MIN_AROM_RINGS and MAX_AROM_RINGS both equal 1), P_{arom} for each ring will be 0.5.

The program then loops over all the potential rings. For each ring, if P_{arom} exceeds a randomly generated number, the ring is made aromatic by the appropriate positioning of double bonds and, in the case of five-membered rings, a heteroatom. Any ring so aromatized is marked appropriately. Bond orders and the values of the valence and maximum valence for each affected node are updated accordingly. Once all rings have been tested in this way, the program checks that the number of rings aromatized does not exceed MAX_AROM_RINGS. If it does, the functionalization process is restarted and the algorithm tries again until the

MAX_FUNC parameter governing the number of attempts to be made at functionalizing a graph is exceeded.

Alicyclic Double Bonds. If the user has requested that one or more alicyclic double bonds be present (i.e., if MIN_ALI_DBLE_BONDS is greater than zero), the remaining nonaromatic rings are examined for edges which could be made into double bonds. If the number of possible edges is too low, the functionalization process is restarted as described above. Otherwise, double bonds are placed by a probabilistic procedure akin to that used for the aromatic rings and the bond order and valence arrays updated. Once this is finished, a check is made to ensure that MAX_ALI_DBLE_BONDS has not been exceeded and also that no new aromatic rings have been created by the placement of the alicyclic double bonds. If either case is found to be true, the functionalization is restarted.

Exocyclic Double and Triple Bonds. The positioning of exocyclic double and triple bonds is handled by a single routine. Double bonds are positioned first. The atoms involved in exocyclic bonds are examined to see which have sufficient free valences to allow a double bond to be placed between them. Double bonds are then added probabilistically, updating valences and bond orders appropriately, and the number of double bonds formed checked in the usual manner. The process is then repeated to place any required triple bonds. The parameters governing the number of exocyclic double bonds are MIN_EXO_DBLE_BONDS and MAX_EXO_DBLE_BONDS. Similarly, the parameters MIN_EXO_TRPLE_BONDS and MAX_EXO_TRPLE_BONDS delimit the number of triple bonds to be placed.

Heteroatoms. The addition of heteroatoms is controlled by three parameters. NUM_HETERO_TYPES simply defines the number of elements (other than C and H) which the user is prepared to allow in the generated molecules. The MAX_PERCENT_HETERO parameter specifies the maximum percentage of the heavy atoms in molecule which can be other than carbon. Finally, for each of the NUM_HETERO_TYPES the user specifies the element type and the minimum and maximum number of permitted occurrences in a molecule. This is done through the PERMITTED_HET_TYPE parameter.

The positioning of heteroatoms is carried out as follows. An initial list of "forbidden" nodes is created. These are defined as those nodes in nonaromatic five-membered rings whose conversion to a heteroatom would cause the creation of a new aromatic ring. The algorithm then loops over each of the permitted heteroatom types and looks for suitable, nonforbidden nodes for conversion to that type. These are then converted using the usual probabilistic scheme with concurrent updating of valences and maximum valences. Two passes are carried out and then the distribution of heteroatoms is checked to ensure that it coincides with the user's specifications. If not, the functionalization routine is restarted as described earlier.

Duplicate Checking. To ensure that duplicate molecules are not generated for a given molecular graph, as each molecule is generated it is assigned two identifiers, EAMax and EASum, which are calculated according to the method of Yang *et al*.[33] These are then compared with the identifiers corresponding to any molecules previously generated from the graph and if any is found to possess the same values for EAMax *and* EASum, then the current molecule is rejected and the functionalization process restarted.

Adding Hydrogens. Having placed all specified multiple bonds and heteroatoms, hydrogen atoms are then added to the molecular graph to fill any empty valences.

Substructure Search. Having generated a fully functionalized molecule, the user may then opt for it to be examined for the presence of various substructures. This is of utility in two respects: it allows the rejection of molecules containing user-specified undesirable substructures and it permits the user to specify desired functional groups which must be present.

Firstly, a check is made for any undesirable substructures. These are held in the file specified by the SUBS_FILE parameter. The substructures are stored in a simple, SMILES-like notation so it is easy for the user to formulate a description of the undesirable molecular features to be detected. If any of the specified substructures is discovered, the functionalization process is restarted. The second check enables the user to specify the number and/or type of hydrogen-bond acceptor and donor groups which must be present in any generated molecule. These are stored in the same SMILES-like notation in the files specified by the parameters H_ACCEPTOR_FILE and H_DONOR_FILE respectively. The acceptor/donor check can take two forms which we have termed "specific" and "nonspecific".

In nonspecific mode, the user simply specifies the minimum number of acceptors and donors that must be present in the generated molecule by means of the MINIMUM_ACCEPTORS and MINIMUM_DONORS parameters. Those substructures which the user decides constitute acceptors and donors are stored in the relevant substructure files. The substructure search then simply adds up all the occurrences of the specified acceptors and donors and checks that the MINIMUM_ACCEPTORS and MINIMUM_DONORS parameter values are attained. In specific mode, indicated by the user assigning a value of zero to both the MINIMUM_ACCEPTORS and MINIMUM_DONORS parameters, the program reads in a specified minimum number for each specified acceptor or donor substructure. The substructure search then checks that each specified substructure occurs at least as many times as the specified minimum in each structure.

By means of these substructural checks, the user is able to ensure that the program generates molecules that are not only chemically sensible but which also have the desired functionality, perhaps to match a pharmacophore.

Output. Once a molecule has been generated and has passed all the structural checks, it is saved to disk in Molecular Design's SD file format.[36] This format was chosen for compatibility with the available 2D−3D conversion software. Details of each molecule's construction can be saved to a log file if the user requires. The level of output sent to the log file is determined by the value of the PRINT_OPTION parameter.

For each graph, up to MAX_MOLS_PER_GRAPH molecules can be generated. The process of molecule generation continues until either all the vertex degree sets have been exhausted or the maximum number of required structures (specified by MAX_STRUCTURES) has been attained.

2D−3D Conversion. The generation of 3D atomic coordinates from a 2D or 2.5D connection table has been an active research area in the last 10 years, and, indeed, the advent of programs capable of performing this task reliably and rapidly is largely responsible for the rise in interest in

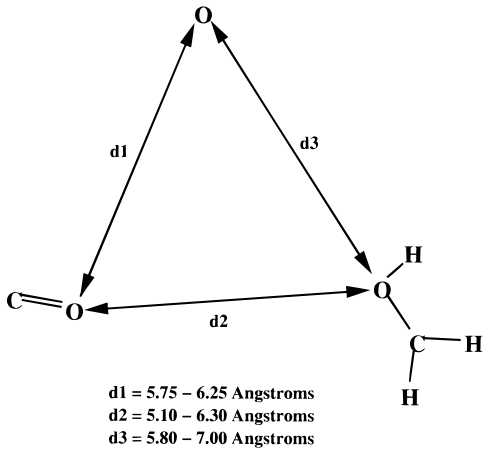

Figure 5. Protein Kinase C pharmacophore.

3D searching. There are several programs currently available, and these have been reviewed and compared in a number of articles.[37−39]

The package available in-house at the time of these studies was Converter,[40] although MOLMAKER could be used in conjunction with any of the available programs. Converter uses distance geometry techniques (as embodied in Havel's DG-II program package[41]) to generate 3D coordinates from 2D structures in SD format. A full description of the methods involved is given in ref 41 but in outline the method involves the following: (1) generation of distance bounds and chirality and planarity constraints from the input structure, (2) triangle inequality bounds smoothing, (3) sampling using metrization, (4) Embedding in four dimensions, and (5) four-dimensional minimization followed by three-dimensional minimization.[42]

In all our studies, we used Converter in batch mode by means of the db_convert utility.[42] The default options were used resulting in the generated structures being biased toward containing extended chains and chair conformations for sp^3 six-membered rings. As MOLMAKER does not make any specification concerning the stereochemistry of chiral centers, Converter was instructed to generate a single stereoisomer for each molecule with the chiral centers randomly assigned as *R* or *S*. Converter outputs the generated structures in SD file format for easy integration with 3D database packages.

3D Database Creation and Searching. The software used for creating and searching 3D databases was ISIS/3D.[43] ISIS/3D is capable of performing flexible searches over large databases using algorithms described by Moock *et al.*[5] Databases were created from the SD files generated by Converter.

EXAMPLES

Having described in detail the functionality underlying the MOLMAKER program and indicated how it interfaces with the commercial software for 2D−3D conversion and database creation and searching, we now proceed to illustrate its utility in drug design by means of two examples.

Protein Kinase C Agonists. In a recent paper,[8] Wang and co-workers carried out a flexible search over the "open" NCI database of 206 876 compounds which yielded 535 hits for the protein kinase C (PK-C) pharmacophore shown in Figure 5. Five of the retrieved molecules possessed PK-C binding affinities in the low micromolar range with six others showing marginal binding affinities.[8] We were interested to see if we could generate novel compounds containing the PK-C pharmacophore. Specifically, the aim is to use MOLMAKER to generate 2D structures containing the pharmacophore elements, build the structures with Converter, and then create a 3D database and conduct a flexible search over the generated structures using ISIS/3D.

Table 1. Statistics for MOLMAKER PK-C Runs

no. of heavy atoms	vertex degree sets generated (used)	CPU s/structure
15	65 (13)	10.38
16	73 (10)	11.52
17	80 (10)	20.36
18	88 (9)	20.99

Four separate MOLMAKER jobs were run, each of which generated 2500 structures having 15, 16, 17, and 18 heavy atoms, respectively. An input file for the 15 heavy atom job is shown in Figure 2. From this it can be seen that two or three rings were specified, but no bridged or spiro systems were allowed. Aromatic rings were permitted but not specified. In order for the generated structures to contain the appropriate substructural features to match the PK-C pharmacophore, sufficient exocyclic double bonds and heteroatoms were specified to allow the formation of the carbonyl group, the free acceptor oxygen, and the terminal CH_2OH group. These substructures were specified using the specific mode of substructure searching by placing the appropriate SMILES strings in the H_ACCEPTOR_FILE. The CPU times for the generation of the four subsets are given in Table 1. These times refer to one node of a convex exemplar (which runs at 5−6 times the speed of an R3000 Indigo in these experiments). It should be noted that in each case only a few of the possible vertex sets were sampled in generating the 2500 structures. All 10 000 structures were successfully converted to 3D by Converter and loaded into an ISIS/3D database.

Using the ISIS/3D CFS (conformationally flexible searching) option, a search was carried out on this database using the pharmacophore shown in Figure 5. Bump checking was enabled at both search and view times, and the hits were ranked in ascending order according to the number of rotatable bonds separating the features matching the pharmacophore.[5] All other search parameters were left at their default values. This search retrieved 264 hits. The top 100 ranked hits were browsed through manually, and ten were selected as being of potential interest. These structures are shown in Figure 6. Of these, **1** and **2** were subjected to more detailed analysis to assess the likelihood of their showing PK-C agonist activity. In particular, the Search_Compare[44] software was used to establish that both structures could adopt the pharmacophore geometry with an energy less than 10 kcal/mol above the "global" minimum established by systematic search with energy minimization. Ten kcal/mol represents the upper limit for the conformational energy penalty able to be incurred by a structure binding to PK-C suggested by Wang *et al*. The details of these computations are given in what follows.

For structure **1**, a systematic search was carried out over the three nonterminal acyclic bonds with a search increment of 30°. Each of the conformers thus discovered was minimized under the CVFF forcefield[45] for 500 iterations or until a maximum derivative of 0.01 was attained.

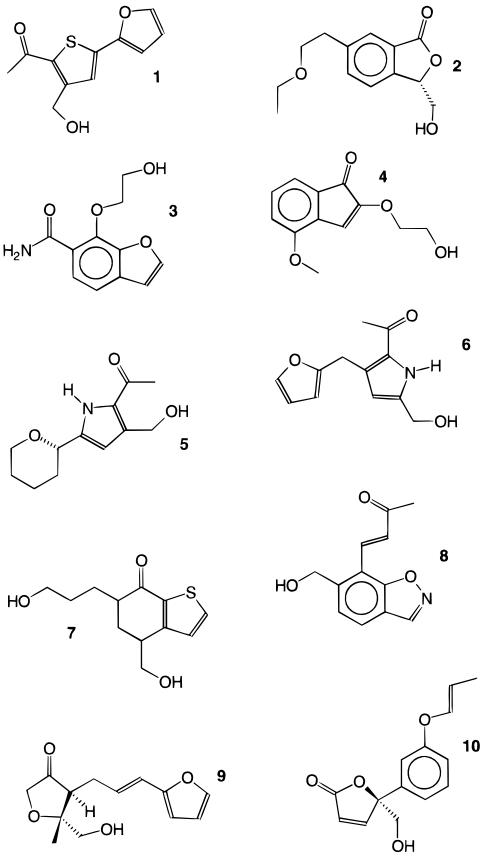

Figure 6. Ten MOLMAKER structures which are hits for the PK-C search query.

Minimized confomers were eliminated as duplicates if they fell with 0.01 kcal/mol or 0.01 Å RMS of each other. Only conformers within 10 kcal/mol of the lowest energy conformer found at any point in the search were retained. This search resulted in 12 conformers, ranging in energy from −0.1 to 6.1 kcal/mol. One of the conformers, of energy 5.2 kcal/mol was found to fit the PK-C pharmacophore. This conformer is clearly within 10 kcal/mol of the lowest energy conformer found by the systematic search (−0.1 kcal/mol). In addition, it was observed that the lowest-energy conformers were those which favored the intramolecular hydrogen bond formation between the carbonyl and hydroxyl groups. This type of geometry is inconsistent with the pharmacophore. However, in solution, it is likely that such an interaction would be much less strongly favored.

A similar search was carried out over the five nonterminal, acyclic bonds of structure **2**. This search resulted in the default maximum of 100 conformers being stored. These conformers had an energy range of 56.7−59.6 kcal/mol. None of the stored conformers matched the PK-C pharmacophore. Thus, a constrained systematic search was carried out using the pharmacophore distance ranges as search constraints. A torsion angle increment of 30° was used, and only the three acyclic, single bonds directly affecting the relative positions of the pharmacophore elements were specified as rotatable. This search resulted in eight conformers, all of which, of necessity, fit the pharmacophore. The lowest energy conformer of the eight was of 63.1 kcal/mol. Again, this falls within 10 kcal/mol of the lowest energy conformer located by systematic search (56.7 kcal/mol).

Thus, for both structures, the results are in accordance with the observation of Wang *et al*. that "10 kcal/mol is probably

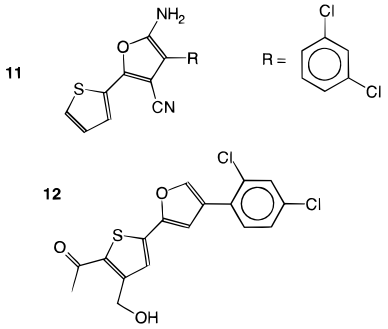

Figure 7. Modification of structure **1** suggested by database search.

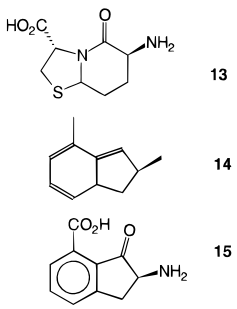

Figure 8. Design of a cyclophilin-calcineurin bridging ligand.

the upper limit for the conformational energy penalty when a ligand binds to PK-C".[8]

Wang and co-workers also established that hydrophobicity was an important criterion for PK-C agonist activity. The log *P* values for **1** and **2**, as calculated using the method of Viswanadhan *et al*.,[46] are 0.51 and 1.47, respectively. These values may indicate that these structures are not sufficiently hydrophobic *per se*. However, it should be relatively straightforward to add hydrophobic substituents to the structures to alter the log *P* values without interfering with the basic scaffold holding the pharmacophore elements. As an indication of this, a search of the ACD[53] found the structure **11** (Figure 7) suggesting a modification to **1**. When **1** is modified by the addition of the dichlorophenyl moiety to form **12** (Figure 7), the log *P* value increases to 3.23. Such a modification would seem likely to increase any binding to PK-C.

Design of a Cyclophilin-Calcineurin Bridging Ligand. Cyclophilin (CyP) is a ubiquitous intracellular protein that binds the immunosuppressive drug, cyclosporin A (CsA). The CyP−CsA complex binds to calcineurin, a calcium-activated phosphatase which is essential for T-cell signalling. The formation of the trimer thus inhibits T-cell activation.[47] Whilst CsA is a powerful immunosuppressant, the search for more potent analogues is ongoing. Studies by Alberg and Schreiber have demonstrated that the incorporation of an appropriate, rigid template molecule into CsA can help to lock it into its bound conformation and thus produce a modified molecule with enhanced binding affinity.[24] In their work, Alberg and Schreiber used the CAVEAT program[23] to identify the heterocycle **13** (Figure 8) which fixes an amine substituent and a carboxyl substituent in the same orientations as the C_α−N bond of L-Ala[7] in bound CsA and the C_α−C bond of D-Ala[8] in bound CsA. The modified CsA molecule produced by the incorporation of **13** was shown to bind to cyclophilin A with three times the affinity of the native CsA.

MOLMAKER was used to help suggest an alternative to **13** which could perform the same operation. The program

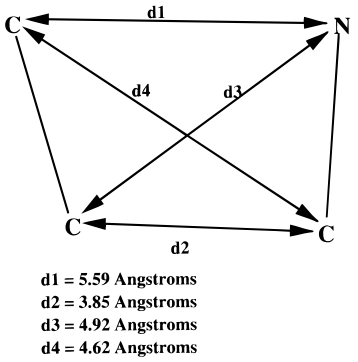

Figure 9. Search query to find possible cyclophilin-calcineurin bridging ligands.

was instructed to build structures with between 10 and 12 heavy atoms and which contained two nonaromatic rings of either five or six atoms. Bridged and spiro ring systems were not permitted. In order to incorporate some rigidity into the "scaffolds" produced, between one and three alicyclic double bonds were requested. To ensure that the structures had suitable exocyclic vectors for attaching the functional groups, the specific substructure specification functionality was used to ensure all structures generated contained two methyl groups. Finally, since the primary purpose of this exercise is the construction of a suitable scaffold, no heteroatoms were permitted.

With these parameters, MOLMAKER generated 355 structures in 2718 CPU s (R3000 Indigo). These were all successfully converted to 3D by Converter and loaded into an ISIS/3D database for searching. The pharmacophore for the database search was derived from the structure of cyclosporin A as bound to cyclophilin A by NMR (PDB structure 1CYA[48]) and is shown in Figure 9. This pharmacophore represents the respective orientations in which two vectors must be held in order for them to fix the amine and carboxyl functions in the same orientations as in CsA. To allow for the fact that part of the pharmacophore was derived from a C−N vector while the database consists of purely hydrocarbon frameworks, a tolerance of ±0.1 Å was added to all the distances. As in the previous example, a database search was carried out using the CFS option within ISIS/3D. Default settings were used except for the specification of ring flexibility for five- and six-membered nonaromatic rings and bump checking at both search and view times.

Of the two hits retrieved by the search, the most attractive in terms of its rigidity was **14** which is shown in Figure 8. While this structure, *per se*, is likely to be unstable, it immediately suggested an alternative that would keep the methyl groups in a similar orientation. The structure, **15** (Figure 8), an indanone derivative, was built in Insight II[49] and energy minimized under the CVFF force field using the BUILD/OPTIMISE functionality with default settings. When the vectors holding the functional groups were superimposed back onto their respective vectors in CsA using the TRANSFORM/SUPERIMPOSE option in InsightII, the RMSD was 0.077 Å. This is better than the value of 0.13 Å reported by Alberg and Schreiber for their bridging ligand.[24] It would be very interesting to see if the incorporation of this bridging ligand into CsA resulted in a similar or better increase in binding affinity over the native structure.

DISCUSSION

The examples presented above indicate that MOLMAKER can be a useful aid to molecular design, particularly in generating novel molecules to meet pharmacophore constraints or to anchor bond vectors in particular orientations.

The PK-C example demonstrates that MOLMAKER is able to suggest diverse sets of structures to meet a standard three-point pharmacophore. Many of the structures would seem to be synthetically accessible although such a judgment would be facilitated and validated by the use of a program such as CAESA[50] for the automatic estimation of synthetic accessibility. Detailed analysis of two of the hits returned from a conformationally flexible search indicated that they were indeed able to adopt the pharmacophore geometry in a low-energy conformation. This result would seem to indicate both that the structures generated by Converter are of reasonably low energy and that the bump checking in the database search is an effective screen for bad van der Waals's energies. MOLMAKER is also useful in suggesting bridging ligands to link fragments or rigidify existing structures. The rapid design of a novel bridging ligand for cyclophilin−calcineurin exemplifies this. In this instance, MOLMAKER functions to serve a purpose analogous to that of CAVEAT.[23]

The *de novo* design of 3D databases has also been reported by Ho and Marshall.[21] Their program, DBMAKER, manipulates SMILES strings to generate random compounds within user-specified constraints concerning molecular size, composition, connectivity, and so forth. The resulting SMILES strings are converted to 3D by CONCORD and then stored in databases for searching. In general, DBMAKER works in a stepwise manner, first generating smaller partial solutions to a given design problem and then joining them together in a subsequent step to form the final set of solutions. The program seems to be fast in operation and have a fairly general application in rational design situations.

MOLMAKER takes a rather different approach to structure generation being based very firmly upon graph-theoretical principles. We believe that this is advantageous in terms of the diversity of structures that can be produced by the program although the inherent combinatorial nature of the constructive enumeration process imposes an upper limit of about 20 upon the number of heavy atoms that can be considered by our program at present. However, we anticipate that the improvements to the graph generation process will produce substantial time-savings in the use of MOLMAKER. Indeed, a recent re-engineering of the graph generator resulted in savings of up to 30% on the times published here. Furthermore, constructive enumeration remains an active research area, and the fundamental algorithms are being continually improved.[51] This, along with the ongoing increase in the power of computer hardware, gives us much confidence that MOLMAKER will be able to be applied to systems of increasing size very soon.

The novelty of the MOLMAKER approach lies in coupling the vertex degree set generation algorithm of Hall and Fisk[26] with an algorithm for the constructive enumeration of molecular graphs. Although this was suggested by Hall and Fisk, there was no description in their paper of how this might be achieved other than a reference to an obscure Russian paper by Faradzhev.[52] We have implemented our own molecular graph generator based on more recent work by Kvasnička and Pospíchal.[30−32] Furthermore, we have

developed novel algorithms for functionalizing the generated molecular graphs in a probabilistic manner to satisfy functional constraints specified by the user. The linking of these three functional blocks (which comprise the MOLMAKER program) with commercial software for 2D−3D conversion and 3D database creation and searching constitutes a novel methodology for the generation of molecular structures satisfying pharmacophoric constraints.

CONCLUSION

We have described a novel approach to *de novo* molecular design based upon established graph theoretical techniques, model building, and 3D database technology. In addition, we have demonstrated how this approach may be used in the generation of diverse and novel molecules to satisfy pharmacophore constraints derived from a variety of molecular systems.

APPENDIX. SEMICANONICAL LABELING OF GRAPHS

Given an adjacency matrix, **A** of dimension $p \times p$, for any s such that $1 \leq s < p$, **A** can be decomposed into block matrices thus[30−32]

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} \mathrm{A}_{12} \\ \mathbf{A}_{12}^{T} \mathbf{A}_{22} \end{pmatrix} \tag{2}$$

where $\mathbf{A}_{11}$ ($\mathbf{A}_{22}$) corresponds to the top-left (bottom-right) corner submatrix of **A** of dimension $s \times s$ ($p - s \times p - s$). The rectangular submatrix $\mathbf{A}_{12}$ is of dimension $s \times p - s$ and may be expressed in terms of its s-dimensional column vectors

$$\mathbf{A}_{12} = (c_1 \geq \mathbf{c}_2 \ldots \geq \mathbf{c}_{p-s}) \tag{3}$$

Consider two such s-dimensional vectors $\mathbf{a} = (\mathrm{a}_i)$ and $\mathbf{b} = (\mathrm{b}_i)$. These vectors are equal ($\mathbf{a} = \mathbf{b}$), if $\mathrm{a}_i = \mathrm{b}_i$, for $i = 1, 2 \ldots, s$. Alternatively, $\mathbf{a} > \mathbf{b}$ if there exists such an integer $1 \leq i \leq s$ such that $\mathrm{a}_i > \mathrm{b}_i$ and $\mathrm{a}_j = \mathrm{b}_j$ for $j = 1, 2, \ldots, i - 1$.

An adjacency matrix, **A**, is defined as semicanonical if for each $1 \leq s < p$, the column vectors $c_1, c_2, \ldots, c_{p-s}$ of the submatrix $\mathbf{A}_{12}$ satisfy

$$\mathbf{c}_1 \geq \mathbf{c}_2 \ldots \geq \mathbf{c}_{p-s} \tag{4}$$

A graph determined by a semicanonical adjacency matrix is termed a semicanonically labeled graph.

REFERENCES AND NOTES

(1) Bures, M. G.; Martin, Y. C.; Willett, P. Searching Techniques for Databases of Three-Dimensional Chemical Structures. *Topics in Stereochemistry* **1994**, *21*, 467−511.

(2) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.

(3) Clark, D. E.; Jones, G.; Willett, P.; Kenny, P. W.; Glen, R. C. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures Comparison of Conformational-Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197−206.

(4) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.

(5) Moock, T. E.; Henry, D. R.; Ozkabak, A. G.; Alamgir, M. Conformational Searching in ISIS/3D Databases. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 184−189.

(6) Mason, J. S. Experiences with Searching for Molecular Similarity in Conformationally Flexible 3D Databases. In *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Blackie, Academic & Professional: London, 1995; pp 138−162.

(7) Lam, P. Y. S.; Jadhav, P. K.; Eyermann, C. J.; Hodge, C. N.; Ru, Y.; Bacheler, L. T.; Meek, J. L.; Otto, M. J.; Rayner, M. M.; Wong, Y. N.; Chang, C.-H.; Weber, P. C.; Jackson, D. A.; Sharpe, T. R.; Erickson-Viitanen, S. E. Rational Design of Potent, Bioavailable, Nonpeptide Cyclic Ureas as HIV Protease Inhibitors. *Science* **1994**, *263*, 380−384.

(8) Wang, S.; Zaharevitz, D. W.; Sharma, R.; Marquez, V. E.; Lewin, N. E.; Du, L.; Blumberg, P. M.; Milne, G. W. A. The Discovery of Novel, Structurally Diverse Protein Kinase C Agonists through Computer 3-D Database Pharmacophore Search. Molecular Modelling Studies. *J. Med. Chem.* **1994**, *37*, 4479−4489.

(9) Kiyama, R.; Homma, T.; Hayashi, K.; Ogawa, M.; Hara, M.; Fujimoto, M.; Fujishita, T. Novel Angiotensin II Receptor Antagonists: Design, Synthesis and In Vitro Evaluation of Dibenzo[a,d]cycloheptene and Dibenzo[b,f]oxepin Derivatives. Searching for Bioisosteres of Biphenylyltetrazole using a Three-Dimensional Search Technique. *J. Med. Chem.* **1995**, *38*, 2728−2741.

(10) Rusinko, III, A.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *Concord: A Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*; The University of Texas at Austin and Tripos Associates: St. Louis, MO, 1988.

(11) *Chem-X*; Chemical Design Ltd.: Roundway House, Cromwell Park, Chipping Norton, Oxfordshire, OX7 5SR.

(12) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, E. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.

(13) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Applications of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.* **1994**, *37*, 1233−1251.

(14) Gordon, E. M.; Barrett, R. W.; Dower, W. J., Fodor, S. P. A.; Gallop, M. A. Applications of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(15) van Drie, J. H. 3D Database Searching in Drug Discovery. *Network Science (http://www.awod.com/netsci)* **1995**, *1*, Issue 4 (October).

(16) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. A Method for Automatic Generation of Novel Chemical Structures and Its Potential Applications to Drug Discovery. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 527−530.

(17) Martin, Y. C.; Van Drie, J. H. Identifying Unique Core Molecules from the Output of a 3D Database Search. In *Chemical Structures 2. The International Language of Chemistry*; Warr, W. A., Ed.; Springer Verlag: Berlin, 1993; pp 315−326.

(18) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−38.

(19) Martin, Y. C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145−2154.

(20) Martin, Y. C. Computer Design of Potentially Bioactive Molecules by Geometric Searching with ALADDIN. *Tetrahedron Comput. Methodol.* **1990**, *3*, 15−25.

(21) Ho, C. M. W.; Marshall, G. R. DBMAKER: A Set of Programs to Generate Three-Dimensional Databases Based Upon User-Specified Criteria. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 69−86.

(22) Glen, R. C.; Payne, A. W. R. A Genetic Algorithm for the Automated Generation of Molecules within Constraints. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 181−202.

(23) Lauri, G.; Bartlett, P. A. CAVEAT: A Program to Facilitate the Design of Organic Molecules. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 51−66.

(24) Alberg, D. G.; Schreiber, S. L. Structure-Based Design of a Cyclophilin-Calcineurin Bridging Ligand. *Science* **1993**, *262*, 248−250.

(25) Weiss, G. A.; Collins, E. J.; Garboczi, D. N.; Wiley, D. C.; Schreiber, S. L. A Tricyclic Ring System Replaces the Variable Regions of Peptides by Three Alleles of Human MHC Class I Molecules. *Chem. Biol.* **1995**, *2*, 401−407.

(26) Hall, L. H.; Fisk, J. B. Computer Generation of Vertex Degree Sets for Chemical Graphs from a Number of Vertices and Rings. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1184−1188.

(27) Kier, L. B.; Hall, L. H.; Frazer, J. W. Design of Molecules from Quantitative Structure-Activity Relationship Models. 1. Information Transfer between Path and Vertex Degree Counts. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 143−147.

(28) Hall, L. H.; Kier, L. B.; Frazer, J. W. Design of Molecules from Quantitative Structure-Activity Relationship Models. 2. Derivation and Proof of Information Transfer Relating Equations. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 148−152.

(29) Hall, L. H.; Kier, L. B. Design of Molecules from Quantitative Structure-Activity Relationship Models. 3. Role of Higher Order Path Counts. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 598−603.

(30) Kvasnička, V.; Pospíchal, J. An Improved Method of Constructive Enumeration of Graphs. *J. Math. Chem*. **1992**, *9*, 181−196.
(31) Kvasnička, V.; Pospíchal, J. An Improved Version of the Constructive Enumeration of Molecular Graphs with Prescribed Sequence of Valence States. *Chemom. Intell. Lab. Syst*. **1993**, *18*, 171−181.
(32) Pospíchal, J.; Kvasnička, V. An Alternative Approach for Constructive Enumeration of Graphs. *Collect Czech. Chem. Commun*. **1993**, *58*, 754−774.
(33) Yang, Y.-Q.; Xu, L.; Hu, C.-Y. Extended Adjacency Matrix Indices and Their Applications. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1140−1145.
(34) Reference deleted in proof.
(35) Reference deleted in proof.
(36) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Ltd. *J. Chem. Inf. Comput. Sci*. **1992**, *32*, 244−255.
(37) Pearlman, R. S. 3D Molecular Structures: Generation and Use in 3D Searching. In *3D QSAR in Drug Design: Theory, Method and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 41−79.
(38) Ricketts, E. M.; Bradshaw, J.; Hann, M.; Hayes, F.; Tanna, N. Comparison of Conformations of Small Molecule Structures from the Protein Data Bank with Those Generated by Concord, Cobra, ChemDBS-3D and Converter and Those Extracted from the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci*. **1993**, *33*, 905−925.
(39) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1000−1008.
(40) *Converter, Version 2.3;* Biosym/MSI: San Diego, CA.
(41) Havel, T. F. An Evaluation of Computational Strategies for Use in the Determination of Protein Structure from Distance Constraints Obtained by Nuclear Magnetic Resonance. *Prog. Biophys. Mol. Biol*. **1991**, *56*, 43−78.
(42) *Sketch* and *Converter User Guide, Version 2.3*; Biosym/MSI: San Diego, CA, 1993.
(43) *ISIS/3D, Version 1.2*; MDL Information Systems Inc.: San Leandro, CA.
(44) *Search_Compare, Version 2.3*; Biosym/MSI: San Diego, CA, 1993.
(45) As implemented in the program *Discover, Version 2.9.5*; Biosym/MSI: San Diego, CA, 1993.
(46) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibodies. *J. Chem. Inf. Comput. Sci*. **1989**, *29*, 163−172.
(47) Ke, H.; Mayrose, D.; Belshaw, P. J.; Alberg, D. G.; Schreiber, S. L.; Chang, Z. Y.; Etzkorn, F. A.; Ho, S.; Walsh, C. T. Crystal Structures of Cyclophilin A Complexed with Cyclosporin A and *N*-methyl-4-[(*E*)-3-butenyl]-4,4-dimethylthreonine Cyclosporin A. *Structure* **1994**, *2*, 33−44.
(48) Fesik, S. W.; Gampe Jr., R. T.; Eaton, H. L.; Gemmecker, G.; Olejniczak, E. T.; Neri, P.; Holzman, T. F.; Egan, D. A.; Edalji, R.; Simmer, R.; Helfrich, R.; Hochlowski, J.; Jackson, M. NMR Studies of (u-13c)Cyclosporin A Bound to Cyclophilin: Bound Conformation and Portions of Cyclosporin Involved in Binding. *Biochemistry* **1991**, *30*, 6574−6583.
(49) *Insight II, Version 2.3.0*; Biosym/MSI: San Diego, CA.
(50) Johnson, A. P. New Developments in the SPROUT Program for De Novo Design and the CAESA System for Estimation of Synthetic Accessibility. Presented at the 13th Annual Conference of the Molecular Graphics Society, Evanston, IL, July 1994.
(51) Grund, R. Construction of molecular graphs with given hybridization and non-overlapping fragments. *Bayreuther Mathematische Schriften* **1995**, *49*, 1−113. (In German).
(52) Faradzhev, I. Generation of Nonisomorphic Graphs with Given Partition of Vertex Degrees. In *Algorithmic Investigations in Combinatories*; Faradzhev, I., Ed.; Nauka: Moscow, 1978; pp 11−19 (in Russian).
(53) ACD: Available Chemicals Directory Release 94.2, MDL Information Systems Inc.: San Leandro, CA.CI9502055

CI9502055