

Analysis of Conformational Coverage. 2. Applications of Conformational Models

Andrew Smellie,* Scott D. Kahn, and Steven L. Teig

Molecular Simulations Inc., 555 Oakmead Parkway, Sunnyvale, California 94086

Received October 12, 1994[®]

It has previously been demonstrated¹ that the low-energy conformational spaces of small- to medium-sized drug molecules can be adequately represented by a small collection of conformations. This paper demonstrates that these representative conformational models consisting of a small collection of conformations (a) suffice for pharmacophore and/or hypothesis generation and (b) return a significant number of valid geometric hits from a flexible 3D database search. The resolution of terse conformational models is quantified using novel metrics that measure conformational coverage, and the limits of resolution using point conformers for hypothesis generation are investigated. A general methodology for evaluating database search methods is introduced and is used to demonstrate that point conformations can be used successfully in database searching.

INTRODUCTION

Conformational analysis is the study of the relationship between the shape and energetics of molecules. The problem of determining these shapes from topologies is a difficult one but not an end in itself. A conformational model¹ can be used as input to hypothesis² or pharmacophore^{3,4} generators.

Conformational models of small organic molecules generally consist of a collection of one⁵ or more^{6,7} conformations. It is crucial that the resolution of the conformational model, in terms of how well the model represents the low-energy regions of conformational space, be consistent with the magnitude of the tolerances required by any given application.

This paper investigates the relationship between the size (*i.e.*, the number of conformers) of a conformational model and its resolution. An analogy for the analysis presented can be made between the *actual* dot density of a halftoned image and the *perceived* spatial and intensity resolution of the image.⁸ In the case of conformational models, comparisons have been made between a quasi-exhaustive set of conformers (corresponding to a complete image) and smaller sets of conformers (corresponding to a halftoned image) that were chosen to maximize their collective coverage of the larger set.¹

Two experiments are described that address whether, in practice, the use of a compact conformational model is viable when applied to problems in hypothesis generation, 3D pharmacophore identification, and 3D flexible database search.

The first experiment examines the *median hole size* (described later) as a function of the number of conformers generated by the poling method.^{7,9} For a series of aliphatic amino alcohols with chain lengths of 3–20 carbon atoms and for a series of tri-, tetra-, and pentapeptides, the performance of the poling method in generating compact conformational models for use in hypothesis generation and 3D pharmacophore identification is examined. Hypotheses and/or pharmacophores are not generated in this experiment, but the putative conformational model that would be used in the hypothesis generation is studied to determine its

resolution and hence the resolution of any QSAR model that might be derived from it.

The second experiment uses the FAST flexible database searching algorithm to assess how effectively a small collection of conformers represents a larger quasi-exhaustive set of conformers. At the conceptual level, there are only two scientifically reasonable approaches to flexible 3D database searching. In the first approach, only a set of precomputed, low-energy conformations is checked against the query, whereas in the second approach, the precomputed conformations are augmented by additional, low-energy conformations that are generated on-the-fly to fit the query.^{10,11} These two approaches are both flexible in that they both account for the conformational flexibility of the molecule, albeit using different algorithms, so we avoid the misnomers of “rigid” and “flexible” search, respectively, and use FAST and BEST instead to acknowledge the speed-quality tradeoff being made. In the experiment, a set of random queries is automatically generated using the quasi-exhaustive conformations, and these queries are used to determine the ability of the (smaller) extracted conformational model to return “hits” using the FAST search algorithm.

CONFORMATIONAL MODELS FOR HYPOTHESIS GENERATION AND 3D PHARMACOPHORES

Quasi-exhaustive models for each molecule were calculated with a standard distance geometry algorithm using full metrization¹² during random distance selection. The test molecules were all amino alcohols with the general formula $\text{NH}_2(\text{CH}_2)_n\text{OH}$ ($3 \leq n \leq 20$) and a series of tri-, tetra-, and pentaalanines (plus these same polyalanines with one isoleucine substitution). Each distance geometry conformer was fully energy minimized using a standard conjugate gradients minimization¹³ algorithm with a version of the CHARMM molecular mechanics force field.¹⁴ Conformational models were generated that contained 2000–4000 conformers per molecule, subject to an energy cutoff of 20 kcal above the estimated global minimum energy. The quality of the smaller (*i.e.*, non-quasi-exhaustive) conformational models was compared to that of the quasi-exhaustive models using a *two-set hole size metric*. This metric was designed to

[®] Abstract published in *Advance ACS Abstracts*, March 1, 1995.

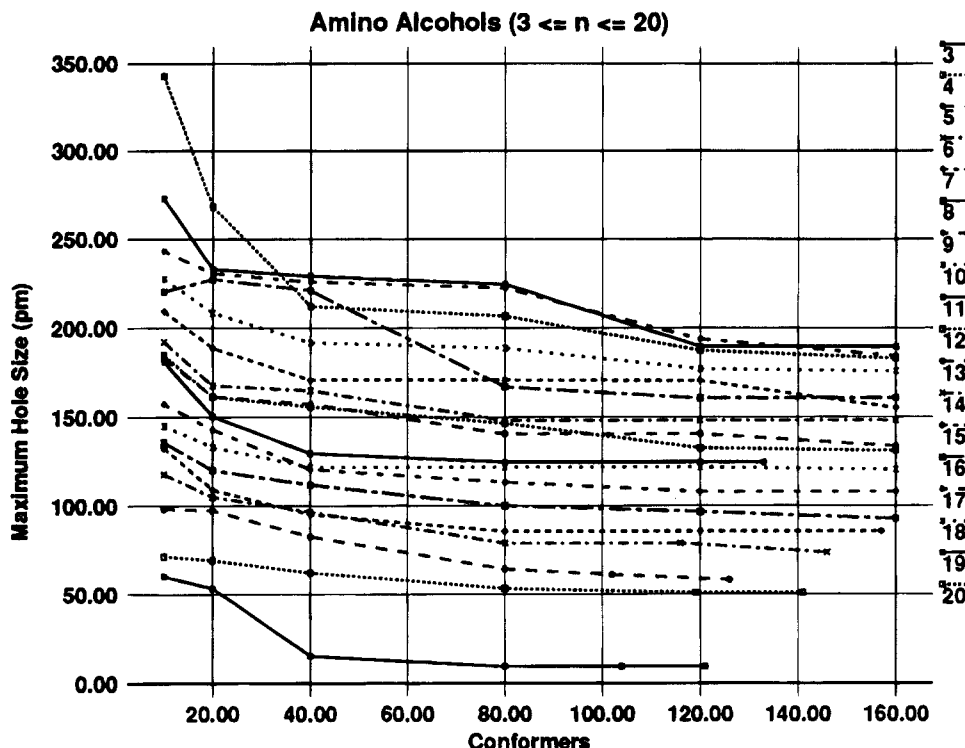


Figure 1. Maximum hole size vs number of conformers for amino alcohols.

measure relative conformational coverage for a given molecule with two independently derived conformational models and has been described fully elsewhere⁷ but is summarized briefly as follows.

Let A and B be two different conformational models (i.e., collections of conformations) of molecule M. Let $d(A_i, B_j)$ be the "distance" (e.g., RMS differences between heavy atom positions) between conformer i (of A) and conformer j (of B), and let $H(A_i, B) = \min_j(d(A_i, B_j))$. The quantity $H(A_i, B)$ is measuring the hole that the i th conformer of A finds in the entire conformational model B. It is the minimum of distances from A_i to all conformers j in B. Thus, a large value of $H(A_i, B)$ means that the i th conformer from A has found a large hole in the conformational model of B. Conversely, a large value for $H(B_j, A)$ means that the j th conformer from B has found a large hole in the conformational model A.

Three quantities can be defined from this generic hole size:

(a) $H_{\max}(A, B)$: the largest hole that the conformational model A finds in the conformational model for B for a given molecule M. This is simply the maximum over all conformations $A_i \in A$ of $H(A_i, B)$ and represents the largest hole that (any conformer of) A finds in the space of B.

(b) $H_{\text{mean}}(A, B)$: the mean hole that the conformational model A finds in the conformational model B, taken over all $H(A_i, B)$.

(c) $H_{\text{median}}(A, B)$: the median hole that the conformational model A finds in the conformational model B, taken over all $H(A_i, B)$.

For a series of terminal amino alcohols in which the amine functionality and the hydroxyl group are separated by a series of methylene groups ranging from 3 to 20, a quasi-exhaustive collection of distance geometry conformers (ranging from 2000 to 4000 conformers/molecule) was compared with conformational models containing 10–160 conformers that were generated using the poling method.⁷ As summarized

in Figures 1–3, the small conformational models seem to approach an asymptote, the ordinate of which increases with the number of rotatable bonds. The plot of *maximum* hole size (H_{\max}) versus the number of poled conformers in the conformational models shows the worst-case behavior of the conformational models, although examination of the median (H_{median}) or mean (H_{mean}) is more meaningful in assessing the typical behavior of the model. These asymptotes are discussed further in the results section.

Further intuition can be gained from Figure 4, where a regular grid is used to represent a simplified 2D schematic of a conformational space in which all conformations are assumed to be energetically accessible. Generated conformations are represented by the centers of the squares of the grid. It can be readily seen that to reduce the hole size by a factor of 2 requires four times as many conformations to be generated. In general, for a d -dimensional conformational space, 2^d times as many conformations are required to halve the hole size if all conformations are low-energy.

Of course, bioaccessible conformations of real molecules are typically of relatively low energies, and the conformational space is highly nonuniform, unlike the simple grid of Figure 4. Nevertheless, the analogy is informative; the exponential relationship between the number of conformations needed and hole sizes explains the apparent asymptotic behavior, which is, in fact, an example of exponentially diminishing returns rather than a true asymptote. The key question is whether hole sizes that are sufficiently small for applications such as hypothesis generation and flexible database searching can be achieved before exponentially diminishing returns set in. Fortunately, substantial empirical evidence, such as that in Figures 1–3, shows that a relatively small collection of conformers is sufficient to cover space for drug-sized molecules to within a hole size of about 1 Å. The significant result that can be gleaned from this data is that *even very flexible systems can be represented by a finite*

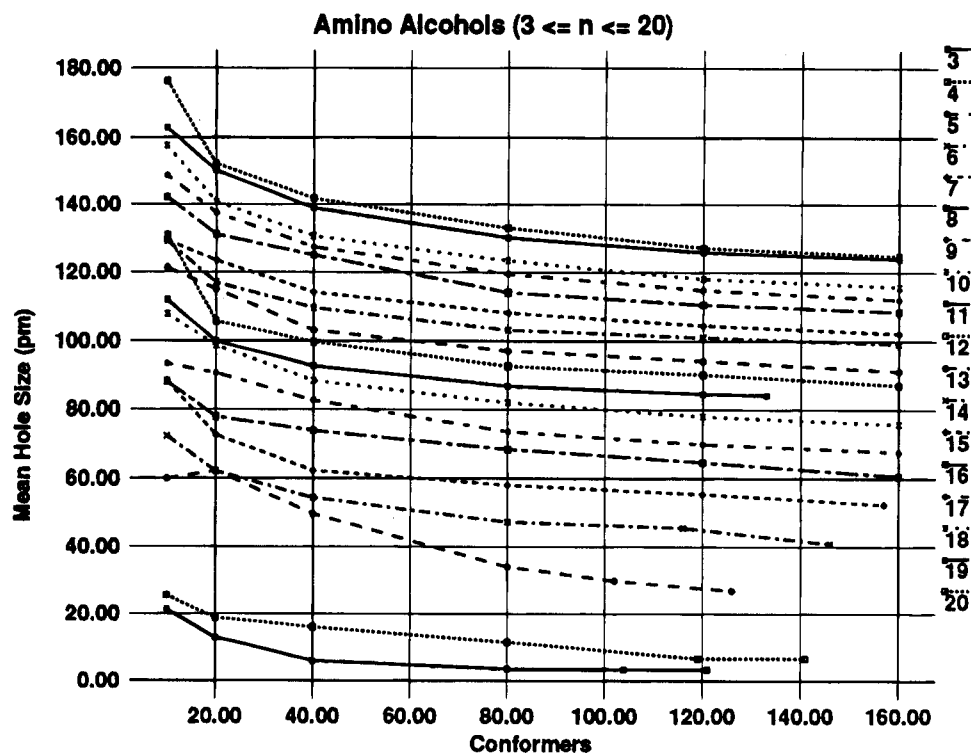


Figure 2. Mean hole size vs number of conformers for amino alcohols.

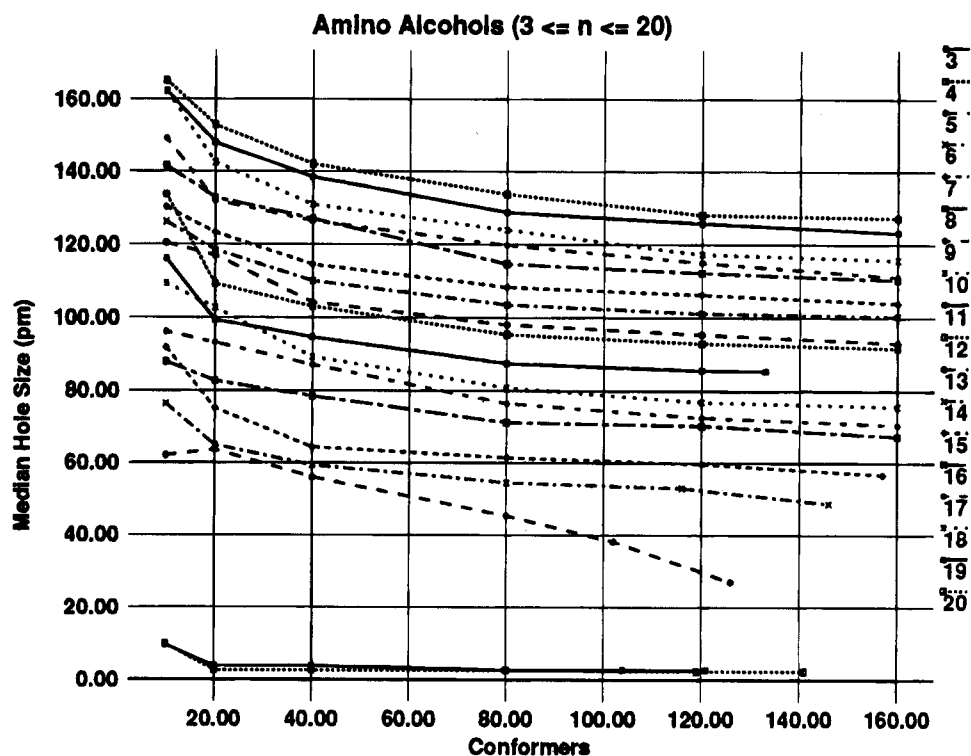


Figure 3. Median hole size vs number of conformers for amino alcohols.

conformational model to a resolution that is of that same magnitude of uncertainties in chemical systems.¹⁵

Similar comparisons of conformational models were made on a series of tri-, tetra-, and pentapeptides, and the data are summarized in Figures 5–7. The maximum hole sizes for the peptides are consistent with the amino alcohol data for the most flexible chains, although the mean and median hole size data for the peptides provide several additional insights. Both the mean and median hole data for the peptides reveal that the average anticipated hole size is larger for the peptides

than for an amino alcohol of the same size even though the peptide bonds are not freely rotatable. For example, the median hole size for 160 conformer model for IAAAA is 0.2 Å larger than that observed for the dodecyl amino alcohol. The larger hole size can be attributed to the (albeit) limited branching, because the “distance” between any two conformers in measuring the hole sizes was the RMS difference in heavy atom positions. The amino alcohols have no branches, but the peptides have short branches off the main chain.

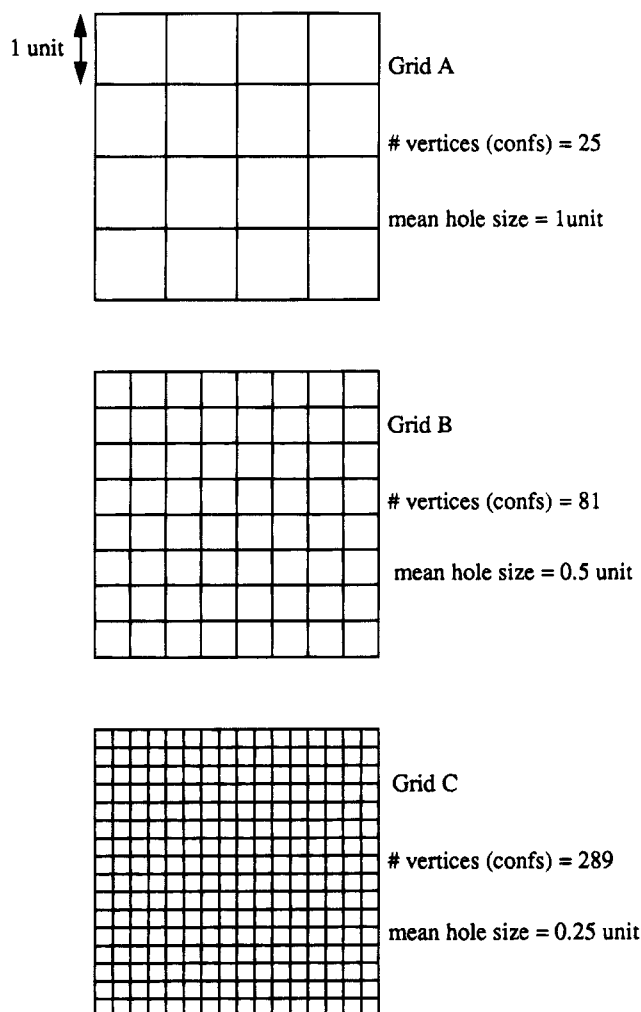


Figure 4. The polynomial nature of conformational coverage for regular grids.

Another trend that is evident in the peptide hole size data is the effect that branching plays in altering the effective conformational space accessible to a compound. For each of the peptide series an N-terminal branched residue (i.e., isoleucine) results in the largest accessible conformational space, and a non-N-terminal branched residue substitution results in a significantly smaller increase in the accessible conformational space compared to the polyalanyl examples. Moreover, in the case of the pentapeptide series, it is apparent that more centrally located branching residues cull the accessible conformational space to the largest degree (cf. AAIAA in Figure 7).

These data place lower bounds on the geometric tolerances of pharmacophoric features if conformational models were used to discover them. Provided the compounds in the training set are "more rigid" than, roughly, pentapeptides, it is possible to accommodate tolerances of 1.0–1.5 Å in the generated hypothesis or pharmacophore with a reasonable number of conformers (i.e., fewer than about 150). Put another way, such a model can specify the proposed "binding" conformer to a precision of about 1.0–1.5 Å RMS away from some conformer in the model. Also, from the analogy of Figure 4, it can be seen that if the desired resolution cannot be achieved with a few hundred conformers, the molecule is too flexible ever to be represented by a reasonably small and finite number of conformers because

of the exponential explosion in the number of conformers required to cover space.

Having demonstrated that fairly small hole sizes can be achieved even for highly flexible small molecules, such as amino alcohols and linear peptides, we next apply discrete conformational models to the problem of flexible 3D database searching.

CONFORMATIONAL MODELS FOR FLEXIBLE 3D DATABASE SEARCHING

A robust procedure to test flexible 3D database searching was developed that uses quasi-exhaustive conformational models to generate satisfiable database queries automatically. This general database test utility is described in Algorithm 2 listed in the appendix. For this study, the diverse molecules shown in Figure 8 were used, and the quasi-exhaustive, energy-minimized conformational models (C_{energy}) were generated using a novel implementation of a systematic search.¹ To determine whether a small subset of the conformations suffices for database searching, a greedy heuristic was then applied to these exhaustive models to extract a subset of conformations (C_{subset}) such that the hole size found by the exhaustive set in the extracted subset was below a user-defined threshold. Queries constructed from the remainder of the quasi-exhaustive set are then used to search the database consisting of the extracted set, as shown in Algorithm 2.

This study is confined to cliques of distance constraints, though arbitrary queries (e.g., involving angles, torsions, etc.) could have been used.¹⁵ The fraction of generated queries hit by the smaller conformational models is presented as a function of (a) varying query tolerances and (b) varying hole sizes used to extract the representative conformational models. For a user-supplied clique size S (i.e., the number of atoms for which pairwise distances will be defined for the query) and a tolerance T on these distances, two conformational models are supplied for each molecule (i.e., C_{energy} and C_{subset}), and it is assumed that the topologies are the same for each model. There is an outer loop over trials, where one trial consists of the generation of a random query Q by selecting a random set of S atoms and topologically defining distance constraints between all pairs of selected atoms. Once the query has been generated, the geometry (i.e., minimum and maximum distances for each distance constraint) of this query is initialized from each reference conformer, but not any conformer that was used to construct the test conformer set (thus keeping the sets disjoint). Thus, no queries are generated from any conformer present in both sets as this would guarantee hitting this query and would not test the quality of the extracted conformational model. Finally, the FAST 3D database search algorithm determines whether the query can be hit by any conformation of the test set. A "hit" means that a conformer in the test set contains interatomic distances that satisfy the query.

Two counts are maintained during the procedure: the total number of queries topologically and geometrically defined (n_{Queries}) and the total number of queries that were hit by a conformer of the test set (n_{Hits}). The percentage of queries hit is reported and given by

$$\% \text{ HitsFound} = \frac{n_{\text{Hits}}}{n_{\text{Queries}}} \times 100\% \quad (1)$$

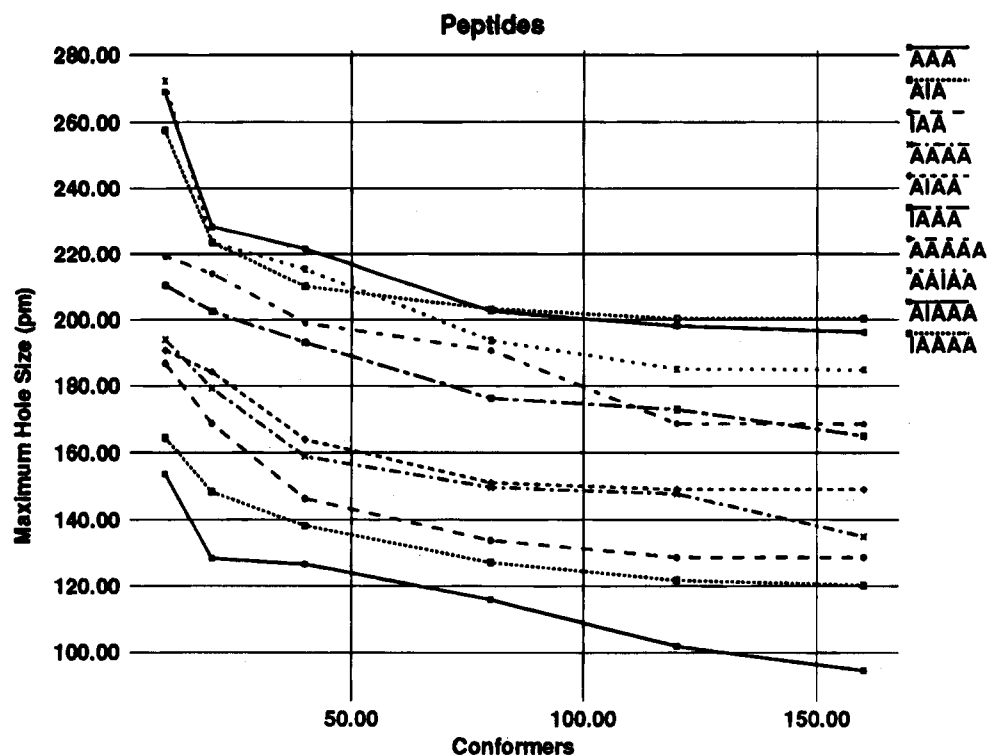


Figure 5. Maximum hole size vs number of conformers for peptides.

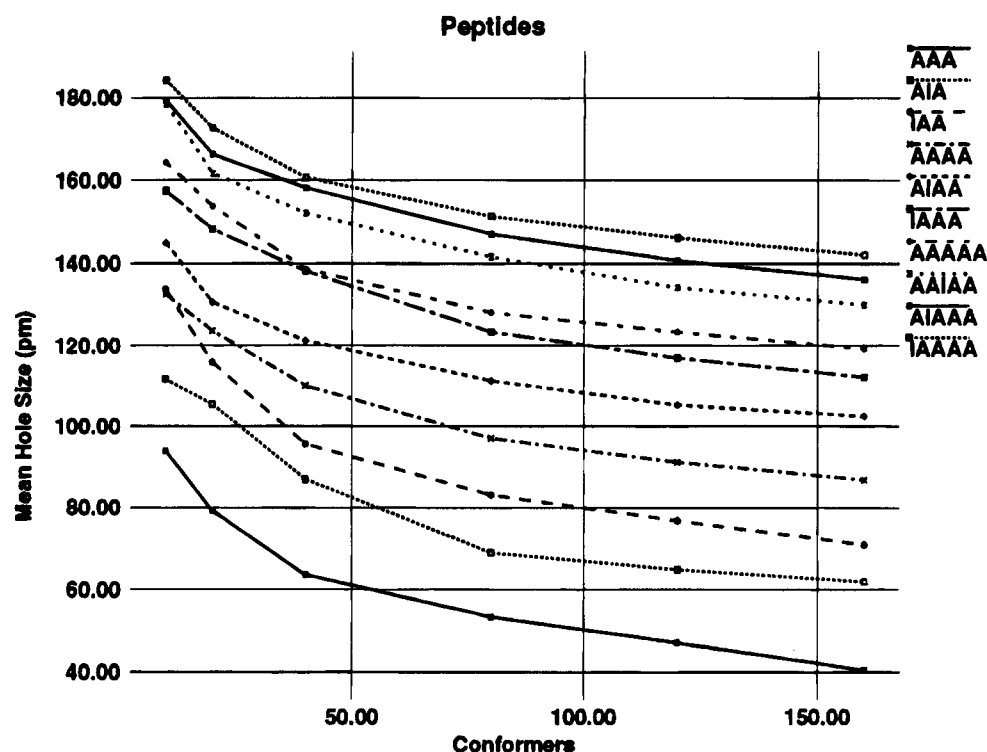


Figure 6. Mean hole size vs number of conformers for peptides.

The reference conformational models used for each molecule were the full energy-minimized set, C_{energy} , which were obtained by a systematic search with 1D minimization and segmented torsion space energy minimization.¹ The test models used were the various conformational models extracted from the full models subject to various hole size requirements, C_{subset} .

Results of this study are presented in tabular form in Tables 1–3. There is one table for every clique size considered. Clique sizes of three, four, and five atoms are presented.

There are $n*(n-1)/2$ distances in the (randomly generated) query produced from a clique of size n , so many of the queries tested are quite complex. Each row of the table represents a different test conformational model used to perform the 3D flexible database search. There is one row for every hole size used, and each column of the table represents a different distance tolerance. After the query's distances are initialized from a reference conformer in C_{energy} , a tolerance on the distance is applied to get minimum and maximum distance bounds, using eq 2. Here, d_i is the

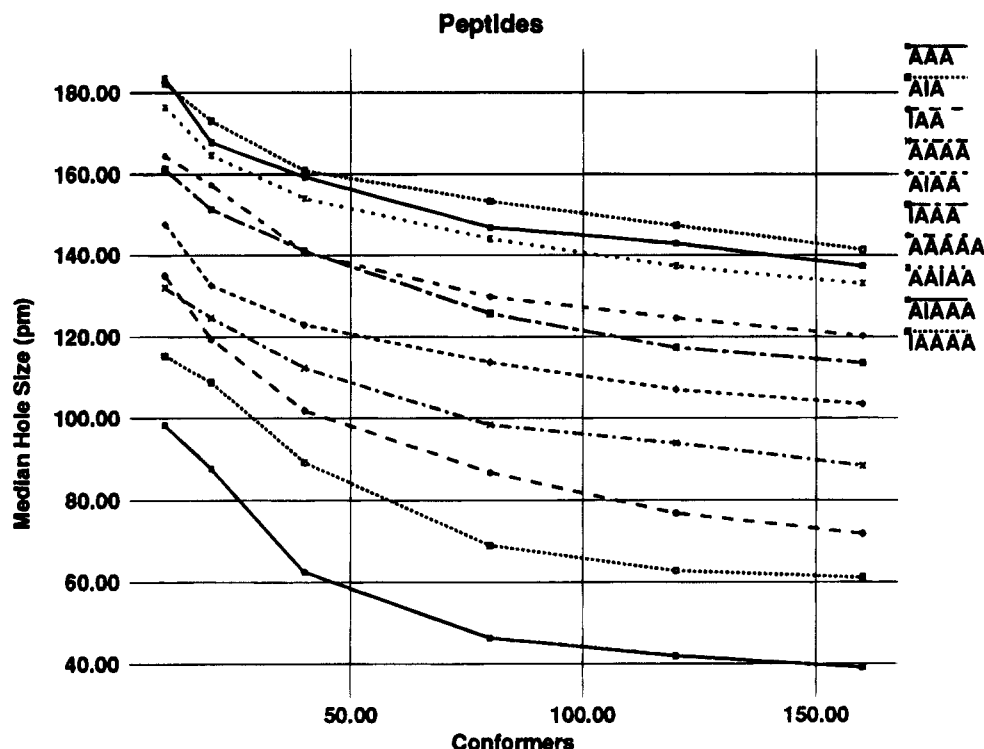


Figure 7. Median hole size vs number of conformers for peptides.

Table 1. Percent Queries Hit wrt Distance Tolerance and Conf. Model Hole Size (Clique Size 3)

hole size	tolerances				
	0.25	0.5	0.75	1.0	1.25
0.5	98.4	99.9	100.0	100.0	100.0
0.6	96.8	99.8	100.0	100.0	100.0
0.7	94.5	99.5	100.0	100.0	100.0
0.8	91.5	99.0	99.9	100.0	100.0
0.9	86.9	97.6	99.7	100.0	100.0
1.0	80.3	95.8	99.3	100.0	100.0
1.1	73.0	93.3	98.9	99.8	100.0
1.2	63.5	89.0	97.6	99.5	99.9
1.3	54.9	84.0	96.3	99.1	99.8
1.4	46.3	75.8	93.0	98.0	99.5
1.5	39.8	68.9	89.2	96.6	99.0

Table 2. Percent Queries Hit wrt Distance Tolerance and Conf. Model Hole Size (Clique Size 4)

hole size	tolerances				
	0.25	0.5	0.75	1.0	1.25
0.5	84.9	97.4	99.8	100.0	100.0
0.6	77.7	94.2	99.2	99.9	100.0
0.7	69.3	90.3	98.0	99.8	100.0
0.8	59.8	85.9	97.1	99.7	100.0
0.9	50.5	79.2	94.3	98.7	99.9
1.0	40.6	70.5	91.2	98.1	99.8
1.1	32.6	61.5	87.2	97.0	99.3
1.2	24.9	51.0	80.9	94.3	98.2
1.3	19.0	42.3	73.6	91.9	97.4
1.4	14.6	33.4	63.6	85.8	94.6
1.5	11.4	26.6	55.4	80.2	92.0

distance computed from the reference conformer, t is the current tolerance on the distance, and l_i and u_i are the lower and upper bounds defined in the query. There is one column for every tolerance.

$$l_i = \max(0, d_i - t), \quad u_i = d_i + t \quad (2)$$

To recap, there is one table for every clique of atoms,

Table 3. Percent Queries Hit wrt Distance Tolerance and Conf. Model Hole Size (Clique Size 5)

hole size	tolerances				
	0.25	0.5	0.75	1.0	1.25
0.5	72.7	89.4	98.3	99.8	100.0
0.6	61.0	81.0	96.6	99.5	99.9
0.7	49.4	71.4	92.6	98.8	99.8
0.8	38.4	60.8	87.9	98.0	99.7
0.9	29.6	49.5	79.8	95.4	99.0
1.0	21.4	37.8	70.3	92.4	98.2
1.1	14.7	27.6	59.1	86.9	96.0
1.2	10.2	20.2	47.9	79.6	93.1
1.3	7.6	15.4	38.7	72.1	89.5
1.4	5.6	11.4	29.4	59.2	81.0
1.5	4.1	8.6	23.3	51.0	74.2

where a random query is defined by all distances between pairs of heavy atoms chosen from a randomly selected set. Each row of the table corresponds to a conformational model extracted from a quasi-exhaustive set subject to a hole size requirement. Each column of the table corresponds to a tolerance applied to each distance in the random query.

The quantity reported in each cell of the table is the mean number of queries hit by a given conformational model (defined by each row) on a query with a given distance tolerance (defined by each column), averaged over all molecules in Figure 8.

Why is this test better than other tests of database search efficacy?^{10,11} Primarily, this test considers *thousands* of queries rather than a small handful of queries extracted from the literature. Also, each query is guaranteed to be *legitimately* satisfiable (i.e., meeting the constraints with a low-energy structure), because it has been generated from a known low-energy conformer. In addition, the test is readily extensible to queries of arbitrary size and complexity and is fully automated. Finally, this test evaluates the quality of the search engine *independent of the quality of the query*. The test simply counts the total number of hits, not how

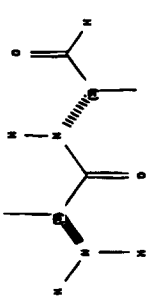
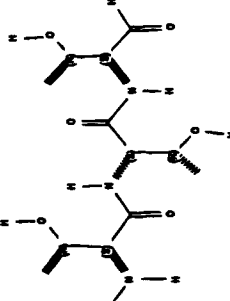
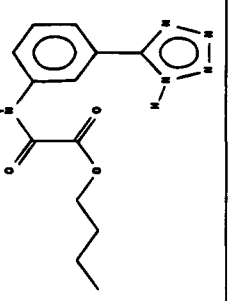
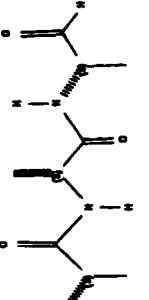
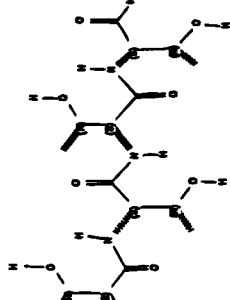
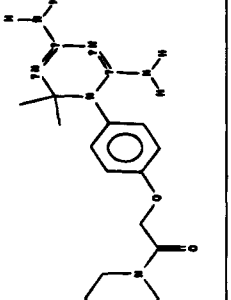
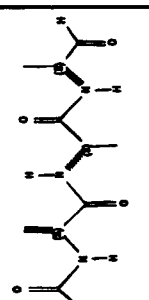
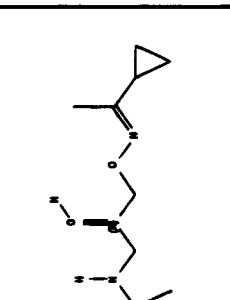
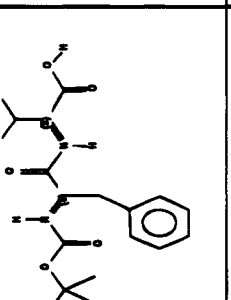
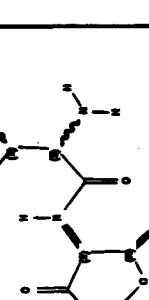
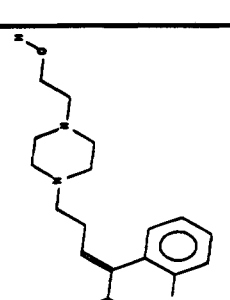
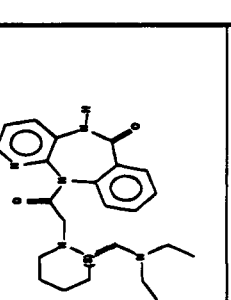
Structure View 12		Name: ala-ala-new		Name: thr-thr-new		Name: thr-thr-new
		Name: ala-ala-ala-new		Name: thr-thr-thr-new		Name: thr-thr-thr-new
		Name: ala-ala-ala-ala-new		Name: thr-thr-thr-thr-new		Name: thr-thr-thr-thr-new
		Name: thr-thr-thr-thr-new		Name: thr-thr-thr-thr-thr-new		Name: thr-thr-thr-thr-thr-new

Figure 8. Molecules used in this study.

many of those hits are "interesting".

The tables of mean hit percentages shown in Tables 1–3 reveal some interesting trends. There is a predictable trend that as (a) the tolerances on distances are loosened and/or (b) the conformational models cover space to finer granularity, the number of hits found increases. This corresponds to moving from the bottom left to the top right of the tables. Also, we observe that as query complexity decreases from a clique size of 5 in Table 3 to a clique size of three in Table 1, the number of hits increases for any given distance tolerance and hole size.

More significantly, the results in these tables show that, contrary to the assertion that "millions of conformations are

required to cover conformational space",¹⁰ quite modest conformational models can, in fact, represent conformational spaces well enough to do flexible database searching at chemically reasonable query tolerances (e.g., ± 1 Å). A summary of these results is given below.

RESULTS

The hole size experiment explored how suitable a terse conformational model is for hypothesis and/or pharmacophore generation.

An interpretation of the asymptotic behavior of Figures 1–3 and Figures 5–7 is straightforward. In line with the

widely held belief that the complexity of conformational spaces increases rapidly with flexibility, the apparent asymptote indicates the practical limitations of using a small number of discrete conformations to represent a conformational space. It is clear from the plots for the amino alcohols that point conformers can readily cover conformational spaces for even quite flexible small molecules at mean hole sizes under 1 Å. Summarizing Figure 2, we can see the following:

*For $n < 6$, mean hole sizes < 0.5 Å are achievable with about 50 conformers.

*For $n < 7$, mean hole sizes < 0.5 Å are achievable with about 100 conformers.

*For $n < 8$, mean hole sizes < 0.5 Å are achievable with about 150 conformers.

*For $n < 13$, mean hole sizes < 1.0 Å are achievable with about 50 conformers.

*For $n < 14$, mean hole sizes < 1.0 Å are achievable with about 100 conformers.

*For $n < 15$, mean hole sizes < 1.0 Å are achievable with about 150 conformers.

Here it was found that even flexible systems such as simple tetra- and pentapeptides can be represented by a conformational model that consists of a couple of hundred conformers, and that the median resolution of these models is ~ 1.4 Å. More typical sets of pharmacologically relevant molecules have hole sizes of about 1 Å. The significance of this result with respect to hypothesis generation and/or pharmacophore identification is that, given the diminishing returns to be had with more conformations, this resolution places a lower bound on the tolerances that can be used by a pharmacophore or hypothesis generator to express spatial relationships if only precomputed conformational models are used. Specifically, *any* tool that generates pharmacophores using conformational models made up of point conformers must take into account the resolution of these models. It has been shown that point conformers cover space to a finite resolution, and *it is not justified to use these conformers to derive a pharmacophore with constraints (usually distances) that are tighter than the resolution of the conformational model*. It is fortunate that chemically meaningful tolerances in hypotheses and pharmacophores are on the same scale as the resolutions that are possible using point conformers for small- to medium-sized drug molecules. These tolerances are discussed in a forthcoming publication,¹⁵ where by gleaning the literature the following tolerances were determined if each interaction could "spend" 1 kcal of energy in any deformation from an ideal geometry:

hydrogen bonding atoms: 0.5–2.1 Å

π – π interactions: 0.5–2.0 Å

hydrophobic interactions: ~ 1.5 Å

The ability of hypothesis generation techniques to use conformational models with resolutions as large as 1.4 Å is discussed more fully elsewhere,² although a brief mention of the high-level results is appropriate for this work. With tolerances used in hypothesis generation of 1.6–2.2 Å, one would anticipate few problems using point conformers with systems of the size described in Figures 1–6, and, in practice, successes have been achieved for tri- and even tetrapeptide datasets.^{16,17}

The database searching experiment explored how effective these terse conformational models are for flexible 3D database searching. A robust methodology was devel-

oped to test the efficacy of database searching by generating a large number of random database queries from the sets of all low-energy conformations found for each molecule from the test set of Figure 8. The queries consisted of cliques of distances between sets of randomly chosen atoms. A check was made to see whether each of these random queries was "hit" by any conformer of a conformational model under test. Several such models were tested that covered space to varying resolutions (measured by the hole size metric).

Results were presented in tabular form in Tables 1–3 that measured the percentage of randomly generated queries hit by a particular conformational model for a particular distance tolerance, averaged over all molecules. The results are intuitive and show that as conformational models cover the space to a higher resolution and/or the tolerances on the distance constraints are loosened, the number of hits increases. It is also observed that the number of hits obtained increases as the complexity of the query (in terms of the clique size) decreases.

We can also use the tables to approximate how well we might expect to do in database searching with a given number of conformations. In Table 3 of ref 1 we found that it takes an average of 184 conformers per molecule to cover conformational space to a hole size of 1 Å (i.e., such that we expect to find no other conformer that is farther away than this) for molecules with up to 10 rotatable bonds. Using this 1 Å hole size, and the tables of hits we find the following:

*For cliques of size 3, distance tolerance 1.0 Å, we expect to find about 100% of the hits.

*For cliques of size 4, distance tolerance 1.0 Å, we expect to find about 98% of the hits.

*For cliques of size 5, distance tolerance 1.0 Å, we expect to find about 92% of the hits.

*For cliques of size 3, distance tolerance 0.5 Å, we expect to find about 95.8% of the hits.

*For cliques of size 4, distance tolerance 0.5 Å, we expect to find about 70.5% of the hits.

*For cliques of size 5, distance tolerance 0.5 Å, we expect to find about 37.8% of the hits.

These numbers support the intuitive result that as distance tolerances are loosened and/or conformational models are more precise, the number of hits expected increases. It has recently been shown⁵ that distance tolerances of ± 0.4 to ± 2.0 Å are appropriate depending on the type of interacting group being constrained. In this case, it can be seen from the results above that we can expect to find $> 90\%$ of hits with multiple rigid conformers for cliques ≤ 5 (i.e., with up to 10 distance constraints) if an average of 184 conformers are spent per molecule. It should be noted that the molecules used in this study are larger than the typical molecule size in, for example, the BioByte database,²¹ which has on average 17 heavy atoms and seven rotatable bonds (including "trivial" bonds such as methyl—which were not considered rotatable in this study), so we would expect to require fewer than 184 conformers for equivalent coverage. In fact, consulting the figures on mean hole sizes found for *n*-amino alcohols (Figure 2) and linear peptides (Figure 6) it can be seen that 50 conformers are sufficient to cover the conformational space to a 1 Å hole size for 8-amino alcohol and AIAA, respectively. These molecules are much larger than the average sized molecule in the BioByte database.

Corroborative evidence for the efficacy of FAST database search (where conformers are precomputed and stored) has already been presented.¹⁰ In this study, a database was seeded with a set of "known" hits, and various search algorithms were used to see if these hits could be rediscovered. It was shown in Table 3 of this paper that performing database searches on a database of 20 *random* conformations found a total of 439/472 (i.e., 93.2%) of the seeded hits. On-the-fly conformer generation, with *no* molecular mechanics energy checking, found 471/472 (i.e., 99.8%) of seeded hits. It should be noted that a more careful population of conformational space ought to improve the number of hits compared to a database of random conformations. Even with this crude algorithm, FAST search found 93% of seeded hits! Further, none of these database search runs was qualified for false positives: other molecules returned that "hit" the query but were not energetically reasonable and were not, in fact, true hits. The false positive and false negative issue will be addressed in a subsequent publication.¹⁸

CONCLUSIONS

This paper has demonstrated that small collections of carefully chosen discrete conformers can cover the conformational spaces of small molecules well enough for applications such as hypothesis generation, pharmacophore generation, and flexible 3D database searching.

The hole size experiment explored how suitable a terse conformational model is for hypothesis and/or pharmacophore generation. We conclude that it is vital to incorporate the concept of resolution into a conformational model before one can meaningfully assess how well a collection of conformations covers accessible low-energy conformational space. It was shown that, when comparing terse conformational models against "exhaustive" ones (generated from distance geometry), the hole size found by the exhaustive set in the terse set reaches an asymptote. This suggests that there is a finite limit to the resolution achievable using point conformations, but these limits are well within the tolerances of common intermolecular interactions.

The database searching experiment explored how effective these terse conformational models are for flexible 3D database searching. We conclude that it is feasible to use point conformations for flexible database searching for molecules of up to 6–7 rotatable bonds and clique complexities of five atoms (i.e., 10 distances) with distance tolerances of ± 1.0 Å on each distance with little risk of missing valid hits, while retaining all the advantages that come with using sets of carefully chosen conformations. These issues will be addressed in a subsequent paper, but some of these advantages include the following.

*A guarantee of no false positives: All hits found with multiple conformers involve at least one energetically reasonable conformation.

*A guarantee of accessibility: It can be easily verified that all hits involve atoms at the surface of the molecule because a valid conformer already exists.

*Speed: Historically, database search with precomputed conformations has been much faster than methods that attempt to account for conformational flexibility only on-the-fly.

In sum, the current paper validates the use of a small collection of carefully chosen conformations as a representa-

tion of a molecule's conformational space for a variety of small-molecule modeling and database applications.

ACKNOWLEDGMENT

The authors would like to thank Jonathan Greene and Peter Towbin for many valuable discussions and insights that contributed to this work.

Appendix 1:

Algorithm 1: FAST flexible database search

```
boolean FAST(Q,D)
input:
  a query, Q
  a database (i.e., a collection of conformations), D
output:
  true if one of the conformations in D satisfies Q
  false, otherwise
// There can be clearly be some sort of filtering scheme here that restricts
// consideration to a subset of the conformations in D. This nuance is
// peripheral to the current work, so we assume that all conformers are
// checked.
for each conformer, Ci ∈ D (
  if (Ci hits Q)
    return(true)
)
return(false)
```

Appendix 2: Robust Evaluation of flexible database searching.

Algorithm 2: Robust validation of flexible database searching

```
input:
  reference conformational set, R
  test conformational set, C
user supplies:
  N = Total number of trials undertaken (20)
  S = Total number of atoms in the query clique (3, 4, 5 or 6)
  T = Tolerance on each distance (0.25, 0.5, 0.75, 1.0 or 1.25)
nQueries = nHits = 0
for each clique size S (
  for each tolerance T (
    for each trial N (
      select a random set of S atoms (≥ 1 rotatable bond between atoms)
      topologically define query, Q, with pairwise distances for all atoms S
      for each reference conformer, Ci ∈ R - C (
        use conf. Ci to set distance constraints, tolerance T, on query Q
        ++nQueries
        if (FAST(Q,Ci)) // See Algorithm 1
          ++nHits
      ) // End loop for reference conformers
    ) // End loop for trials
    // Note: The total number of queries searched = N * |R - C|
    percentHits = (nHits / nQueries) * 100
  ) // End loop for tolerances
) // End loop for cliques
```

REFERENCES AND NOTES

- (1) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage; *J. Chem. Inf. Comput. Sci.* xxxx.
- (2) Hypothesis in Catalyst, Technical White Paper; Molecular Simulations Inc.: 545 Oakmead Parkway, Sunnyvale, CA 94086.
- (3) See, for example: Allen, M. S.; Cook, J. M. Synthetic and Computer-Assisted Analysis of the Pharmacophore for the Benzodiazepine Receptor Inverse Agonist Site. *J. Med. Chem.* **1990**, *33*, 2343–2357.
- (4) See, for example: Bures, M. G.; Danaher, E.; DeLazzer, J.; Martin, Y. C. New Molecular Modeling Tools Using Three-Dimensional Chemical Substructures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 218–223.
- (5) CONCORD, Tripos Associates, St. Louis, MO 63144.
- (6) See, for example: Saunders, M.; Jimenez-Vasquez, H. A. Stochastic Search for Lactone and Cycloalkene Conformers. *J. Comput. Chem.* **1993**, *14*, 330–348.
- (7) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.*, in press.
- (8) Eastman Kodak *Half-tone Methods for the Graphic Arts*; Data Book Q-3, Eastman Kodak Company: Rochester, NY, 1968.
- (9) Catalyst, Molecular Simulations Inc., 545 Oakmead Parkway, Sunnyvale, CA 94086.

- (10) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190–196.
- (11) Moock, T. E.; Henry, D. R.; Ozkabak, A. G.; Alamgir, M. Conformational Searching in ISIS/3D Databases. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 189–198.
- (12) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation. Research Studies Press: Taunton, Somerset, 1988.
- (13) Hestenes, Seifel, Methods of Conjugate Gradients for Solving Linear Systems. *J. Res. Nat. Bur. Standards* **1952**, *49*, 409–436.
- (14) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations*. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (15) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. L. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.*, in press.
- (16) Sprague, P. W. *Building a Hypothesis for Angiotensin Converting Enzyme Inhibition*; Application Note, Molecular Simulations Inc., 555 Oakmead Parkway, Sunnyvale, CA 94086.
- (17) Sprague, P. W. *Building a Hypothesis for HLE Inhibition*; Application Note, Molecular Simulations Inc., 555 Oakmead Parkway, Sunnyvale, CA 94086.
- (18) Smellie, A. S. The Robust Evaluation of 3D Database Searching, manuscript in preparation.

CI940115K