Conformational Sampling with Stochastic Proximity Embedding and Self-Organizing Superimposition: Establishing Reasonable Parameters for Their Practical Use

Gary Tresadern*,† and Dimitris K. Agrafiotis‡

Johnson & Johnson, Pharmaceutical Research & Development, Janssen-Cilag S.A., Calle Jarama 75, Poligono Industrial, Toledo 45007, Spain, and Johnson & Johnson, Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive, Exton, Pennsylvania 19341

Received May 29, 2009

Stochastic proximity embedding (SPE) and self-organizing superimposition (SOS) are two recently introduced methods for conformational sampling that have shown great promise in several application domains. Our previous validation studies aimed at exploring the limits of these methods and have involved rather exhaustive conformational searches producing a large number of conformations. However, from a practical point of view, such searches have become the exception rather than the norm. The increasing popularity of virtual screening has created a need for 3D conformational search methods that produce meaningful answers in a relatively short period of time and work effectively on a large scale. In this work, we examine the performance of these algorithms and the effects of different parameter settings at varying levels of sampling. Our goal is to identify search protocols that can produce a diverse set of chemically sensible conformations and have a reasonable probability of sampling biologically active space within a small number of trials. Our results suggest that both SPE and SOS are extremely competitive in this regard and produce very satisfactory results with as few as 500 conformations per molecule. The results improve even further when the raw conformations are minimized with a molecular mechanics force field to remove minor imperfections and any residual strain. These findings provide additional evidence that these methods are suitable for many everyday modeling tasks, both high- and low-throughput.

INTRODUCTION

Conformational sampling is an essential part of the molecular modeling process. Three-dimensional (3D) conformations are used in molecular overlays, pharmacophore hypothesis generation, pharmacophore and shape searching, 3D QSAR, field-based virtual screening, docking, and other related applications. These analyses are often facilitated through the construction of databases containing a finite number of precomputed conformations for each molecule of interest. The underlying aim of conformational sampling is to generate 3D molecular structures, which represent the chemical compound as it exists in a biologically relevant context. A molecule can often adopt many energetically accessible conformations in solution, but it generally assumes a distinct conformation, referred to as the bioactive conformation, when bound to a target protein. Clearly, to be useful in drug design, a conformational search method must be able to reproduce the bioactive conformation. Because that conformation is, in turn, dependent on the structure and dynamics of the protein host and some ligands adopt different binding conformations with different proteins,[1] it is important that the sampling of conformational space is as complete and unbiased as possible.

Conformation generation algorithms generally fall into two broad categories: deterministic methods, which exhaustively enumerate all possible torsions at certain discrete intervals, and stochastic methods, which use a random element to explore the molecule's conformational space.[2] Although systematic search can be very effective for molecules with limited conformational flexibility, the exponential growth of the search space with the number of rotatable bonds, as well as problems associated with ring closures, limit its utility as a general conformational sampling technique.[3−6] For flexible molecules, stochastic methods designed to sample low energy conformations represent a viable alternative. In its simplest form, a stochastic method randomly perturbs the current conformation of the molecule, minimizes it in energy, and repeats the process to generate a sequence of minimized conformations.[7,8] Standard simulation techniques, such as molecular dynamics and Monte Carlo methods, have been used to generate an ensemble of conformations that lie in the low energy regions on the potential energy surface.[9−11] All of these methods generate conformations in a continuous trajectory, in that each trial conformation is derived from the preceding one by a relatively small change. Because of this continuity, a large number of conformations are generated between the important low energy ones, and a considerable amount of computer time is spent on the calculation and minimization of potential energies for these transitional conformations.

There have been numerous studies comparing computationally and experimentally determined conformations as a means of validating different search methods and protocols. Some of these studies focus on small molecule crystal structures and others on bioactive conformations.[12−14] In

* Corresponding author phone: +34-925-24-5782; e-mail: gtresade@its.jnj.com.
† Johnson & Johnson, Spain.
‡ Johnson & Johnson, PA.

10.1021/ci9001926 CCC: $40.75
Published on Web 11/17/2009

general, the conformation of a molecule in its crystal rarely matches the protein-bound structure.[15] Earlier work on a small data set of 10 ligands cocrystallized with a target protein showed that the bioactive conformation is similar to the lowest energy conformation in solution,[16] but subsequent studies on larger data sets suggest that it may be significantly higher in energy as compared to the global minimum.[14,17] In fact, the relative conformational energies derived with current molecular mechanics force fields introduce considerable errors, which hinder the accurate calculation of free energies of binding.[18] This emphasizes the need for caution and good judgment when applying energy cutoffs to filter out irrelevant conformations. As for the differences in geometry between free and bioactive conformations, Diller and Merz have shown that bound conformations tend to be more extended than the average random conformations generated in vacuo.[19] Together, these studies cast doubt over the efficiency of conventional search methods in locating bioactive conformations.

Two newer techniques that have shown considerable promise in sampling the entire conformational space accessible to a given molecule are stochastic proximity embedding (SPE) and self-organizing superimposition (SOS). Both methods are based on a distance geometry formalism that generates conformations that satisfy a set of geometric constraints derived from the molecular connectivity table.[20,21] There are two forms of constraints: (1) distance constraints encoded in the form of upper (u_{ij}) and lower (l_{ij}) bounds for every interatomic distance, d_{ij} (such that $l_{ij} \leq d_{ij} \leq u_{ij}$), and (2) volume constraints that prevent the signed volume V_{ijkl} formed by four atoms i, j, k, l from exceeding certain limits. The latter are used to enforce planarity of conjugated systems and correct chirality of stereocenters. By generating coordinates that satisfy these constraints, one should in theory be able to sample the entire conformational space. The advantage of distance geometry is that it generates chemically sensible conformations without any direct energy calculation. The conformations can then be rapidly minimized with any suitable force field to identify the corresponding energy minima.

SPE starts from a random initial conformation and gradually refines it by repeatedly selecting an individual constraint at random and updating the respective atomic coordinates toward satisfying that specific constraint. This procedure is performed repeatedly until a reasonable conformation is obtained.[22−24] SOS utilizes a similar algorithm, but makes use of precomputed fragment geometries, which contributes to reduced CPU times and more physically realistic geometries.[25] A more detailed description of these algorithms can be found in the references and in the Methods section.

Because of the nature of the embedding procedure and the use of random initial atomic coordinates, both of these methods tend to produce conformations that are relatively compact. To overcome this problem, we introduced a simple boosting heuristic that can be used in conjunction with SOS and SPE to bias the search toward more extended (or more compact) geometries.[26] This is accomplished through a series of embeddings, each seeded on the result of the previous one. By probabilistically promoting extended geometries, the boosting heuristic biases the search toward the likely bioactive region of conformational space, but maintains sufficient sampling power to ensure that important conformations will not be missed.[22−30]

In the present work, we study in greater detail the performance of SPE and for the first time the SOS algorithm in their ability to sample biologically relevant conformations. In a previous paper, we compared SPE to seven commercially available programs by generating 10 000 conformations for each of a set of 59 ligands, and showed excellence performance for SPE and conformational boosting.[29] The aim of the present study is not to repeat such an extensive comparison under idealized conditions, but rather to establish a reasonable set of parameters that would allow reliable sampling of bioactive space using considerably fewer trials, and to provide practical guidelines on how to best use these methods in everyday modeling problems. To have more confidence in our findings, we utilize a larger data set of protein−ligand crystal structures, which combines data from the work of Bostrom et al.,[12] Diller and Merz,[19] Perola and Charifson,[14] and Chen and Foloppe (which, for the sake of brevity, will hereafter be identified by the name of their first author).[31] The performance of the search is assessed in five ways: (1) the ability to retrieve the bioactive conformation, (2) the diversity of the generated conformations, (3) comparison to a gold standard ensemble, (4) the strain energy of the conformations, and (5) the computational cost required. More specifically, we use as metrics the rmsd of the generated conformations as compared to their respective bioactive structures, the number of unique 3D pharmacophore bits present in the resulting conformation families, the Tanimoto similarity of the 3D pharmacophore bits to the gold standard, the strain energy between each conformation and its nearest local minimum, and the CPU time required to perform the search. We show that the use of SPE and SOS with subsequent energy refinement can lead to excellent performance with respect to all of these criteria.

METHODS

Data Sets. The data set used in this study was assembled from several previous related works.[12,14,19,31] The Bostrom subset contained 32 ligands from cocrystal structures with resolution ≤2.0 Å, most of which had temperature factors below 30 Å^2. The Diller subset contained 59 ligands with resolution ≤3.0 Å, of which 39 were ≤2.0 Å. This is the same data set that we used in our previous comparison of SPE to several widely used conformational analysis porgrams.[29] The third subset consisted of the 100 publically available structures in the Perola data set, all of which had resolution ≤3.0 Å and 48 of them ≤2.0 Å. Our validation suite was completed with 130 ligands from the Chen data set, all of which had resolution ≤2.5 Å and 54 of them ≤2.0 Å. The Chen subset included a diverse selection of structures from the PDB with resolution ≤2.5 Å, MW ≤ 600, and less 12 rotatable bonds, which were not already present in Bostrom and Perola collections.

The entire data set was made up exclusively of noncovalently bound ligands. The structures were downloaded and extracted from their respective Protein Data Bank (PDB) entries, protonation states were assigned using the MOE wash function, tautomer states were manually assigned, and the results were visually inspected. All ligands were converted into 2D and used in that form as input to the different conformational search programs.

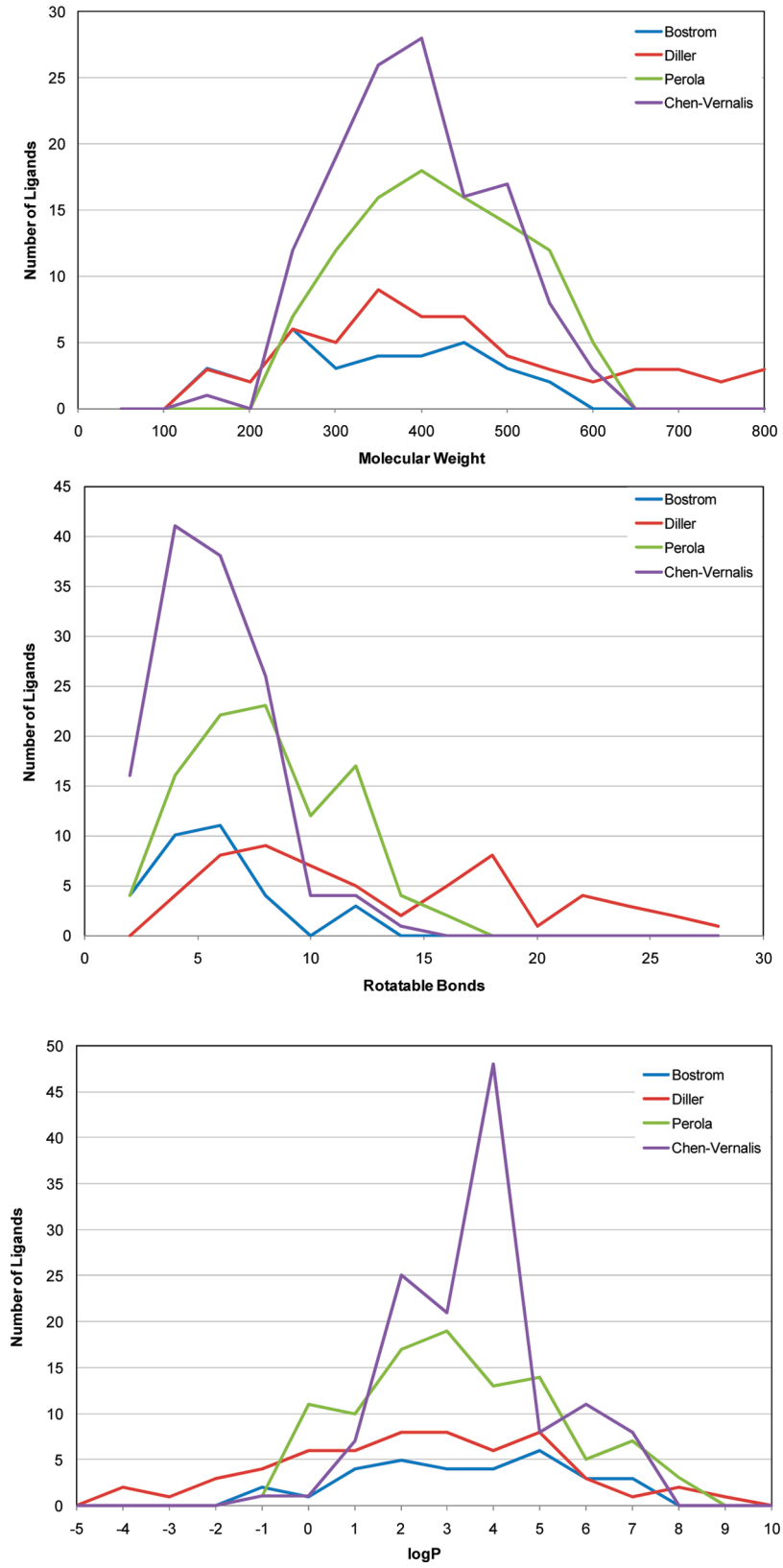

Figure 1. Distributions of molecular weight, rotatable bonds, and log *P* for each subset of ligands.

More detailed analyses of the physicochemical properties of the molecules in these data sets have been presented elsewhere.[14,31] To highlight the key differences, the distributions of molecular weight, number of rotatable bonds, and computed log *P* for each data set are illustrated in Figure 1. It can be seen that the Diller set contains a larger proportion of ligands with high MW and number of rotatable bonds (the latter ranged between 4 and 27 and had a median of 11). This data set has been criticized for containing compounds that are too large and too flexible, and its relevance for drug discovery has been called into question. We would like to reiterate that according to 2005 sales figures, 15 small-molecule "blockbuster" drugs (with sales in excess of $1 billion) contained 10 or more rotatable bonds, with 5 more

Table 1. Details of All Calculations Performed in This Work[a]

name	maximum number of conformations	description
SPE	50, 100, 500, 1000, 5000, and 10 000	SPE with bidirectional boosting toward both compact and extended conformations in a 3:5 ratio
SPE_COM	50, 100, 500, 1000, 5000, and 10 000	SPE with three levels of boosting toward more compact geometries
SPE_EXT	50, 100, 500, 1000, 5000, and 10 000	SPE with five levels of boosting toward more extended geometries
SPE_MIN_0.1	50, 100, 500, 1000, 5000, and 10 000	same as "SPE" above, followed by MMFF94s minimization and removal of duplicates within 0.1 Å rmsd
SPE_COM_MIN_0.1	50, 100, 500, 1000, 5000, and 10 000	same as "SPE_COM" above, followed by MMFF94s minimization and removal of duplicates within 0.1 Å rmsd
SPE_EXT_MIN_0.1	50, 100, 500, 1000, 5000, and 10 000	same as "SPE_EXT" above, followed by MMFF94s minimization and removal of duplicates within 0.1 Å rmsd
SPE_0.1	50, 100, 500, 1000, 5000, and 10 000	same as "SPE" above, followed by removal of duplicates within 0.1 Å rmsd
SPE_0.4	50, 100, 500, 1000, 5000, and 10 000	same as "SPE" above, followed by removal of duplicates within 0.4 Å rmsd
SPE_0.8	50, 100, 500, 1000, 5000, and 10 000	same as "SPE" above, followed by removal of duplicates within 0.8 Å rmsd
SOS	50, 100, 500, 1000, 5000, and 10 000	SOS without boosting
SOS_EXT	50, 100, 500, 1000, 5000, and 10 000	SOS with five levels of boosting toward more extended geometries
MOE_CI	50, 100, 500, 1000, 5000, and 10 000	MOE conformational import with default settings
MOE_BR	50, 100, 500, 1000, 5000, and 10 000	MOE bond rotation with default settings
MOE_SS	10 000	MOE stochastic search with Chen settings

[a] See main text for details on parameter settings. Minimization performed for a maximum of 1000 cycles using the MMFF94s force field ignoring the electrostatic term.

of these molecules (like clarithromycin) having just as much flexibility due to large rings.[29] Also, a recent study of 147 antibacterial compounds that are either on the market or in clinical development showed noticeably different MW and polar surface area properties as compared to drugs in other therapeutic areas.[32] Therefore, in the antibacterial field, where finding initial hits is a known challenge,[33] drug discovery efforts may be forced toward higher molecular weights. In addition, larger and more flexible compounds present a challenge to many conformational search methods, and, as will be shown in the sequel, allow for differences in performance to be more clearly discerned. The Diller data set is thus a useful complement to the compounds found in the Bostrom, Perola, and Chen collections, the last two being perhaps the most "drug-like", with 88% of the compounds having MW between 200 and 500, and 88% of them having $\leq$10 rotatable bonds.

The protein structures from which these ligands were extracted span a range of different targets, including kinases, HIV reverse transcriptase, nuclear receptors, phosphotases, oxido-reductases, metalloproteases, isomerases, and lyases. In summary, our data set consists of drug-like ligands found in multiple binding site environments, and contains a small fraction of challenging, more flexible molecules that test the limits and range of applicability of the conformational search methods.

Comparison Metrics. The main objective of the present study was to assess how many conformations need to be generated to have sufficient confidence that the biologically active conformation (or some other conformation similar to it) is identified. Stated differently, we wanted to examine how much does the performance of the various algorithms decline as fewer and fewer conformations are generated, and establish sensible trade-offs between quality of sampling and computational speed. To address these questions, each method was asked to produce 50, 100, 500, 1000, 5000, and 10 000 conformations for each molecule, in six independent runs. Several variations of these methods were evaluated, using different parameter settings as described in the sequel. A list of these parameter settings along with their abbreviated names are shown in Table 1. Five basic methods were compared: SPE with and without conformational boosting, SOS with and without conformational boosting, MOE conformational import, MOE bond rotation, and MOE stochastic search.

Stochastic Proximity Embedding. As mentioned in the Introduction, SPE attempts to generate conformations that satisfy a set of distance and volume (chiral) constraints. The first step of the process is to establish the lower and upper distance bounds for every pair of atoms, as well as the desired chiral volumes for every chiral center and planar system in the molecule. These bounds are derived from the molecular connectivity table and standard covalent geometry rules, in a manner similar to other distance geometry methods. Once the constraints are established, the program assigns random coordinates to all of the atoms, and uses a self-organizing scheme to rapidly refine the atomic positions so as to satisfy all of the input constraints.[24,26,27] This refinement is accomplished by randomly selecting one constraint at a time, adjusting the positions of the respective atoms so as to better satisfy that specific constraint, and repeating the process many times until a sensible geometry is obtained. The magnitude of the atomic adjustments is controlled by a parameter called the learning rate, which is gradually reduced during the course of the refinement to improve convergence and avoid oscillatory behavior. Distance and volume constraints are selected with different probabilities, with the former being sampled more frequently than the latter. More details about the algorithm can be found in several previous publications.[24,26,27]

Self-Organizing Superimposition. The self-organizing superimposition (SOS) algorithm employs an iterative scheme similar to that SPE, but makes use of precomputed conformational templates of rigid fragments to enforce the desired geometry.[25] In SPE, the embedding is carried out one constraint at a time, and the coordinate adjustments are

applied only to the atoms involved in that specific constraint. However, most organic molecules contain locally rigid fragments connected through freely rotatable bonds. For instance, a phenyl ring can be thought of as a rigid fragment that assumes a regular hexagonal geometry regardless of its surrounding environment. SOS takes advantage of this by using precomputed conformational templates for each rigid part of the molecule, and by adjusting the positions of all of the atoms in each fragment at once using a least-squares fitting procedure.

More specifically, the SOS algorithm starts by identifying all rotatable bonds in the molecule, and then removes these bonds to generate a set of disconnected fragments. Bonds are considered rotatable if they are single, nonterminal, and they are not part of a delocalized system or a ring of size 6 or less. Each resulting fragment is encoded as a canonical SMILES string, which is then used as a key to retrieve the corresponding conformation from a precomputed 3D fragment library. If the requested fragment is not present in the library, the program will generate a reasonable conformation using standard SPE.

While the coordinates of core atoms in a template are taken directly from the retrieved fragment conformation, the coordinates of the attached atoms are not immediately available because they are not present in the fragment molecule. However, all hydrogen atoms of the fragment conformations in the library are explicitly represented with 3D coordinates. The coordinates of each attached atom are determined by replacing a corresponding hydrogen atom in the fragment conformation and adjusting the bond length accordingly. With the steps described above, the templates can be quickly constructed for any input molecule during the initialization phase.

The actual embedding is carried out by an alternating series of template fitting operations and pairwise distance adjustments. Template fitting involves two basic steps. In the first step, a rigid-body transformation (by translation and rotation) is performed on the reference template, which essentially places it in the closest position to the corresponding fragment in the molecule. This can be done using the standard rigid-body rmsd superimposition technique or the improved method described in ref 34. In the second step, the coordinates of the fragment atoms in the molecule are replaced by the new coordinates of the respective atoms in the superimposed reference template. In essence, the fitting operation applies the smallest possible adjustment to the fragment atoms in the molecule to achieve the geometry of the reference template. Each template fitting operation is followed by a number of pairwise distance adjustments between randomly chosen atoms from different fragments to remove steric clashes (similar to SPE). This process is carried out for each fragment in the molecule, and the entire process is repeated until a predetermined convergence criterion is met.

Because rigid fragments are precomputed, planarity and chirality constraints are automatically satisfied after the template superimposition process, and local geometry is perfectly restored. Furthermore, because each embedding starts from completely random initial atomic coordinates, each new conformation is independent of those generated in the previous runs, resulting in greater diversity and more effective sampling. As demonstrated in the sequel, because the algorithm only involves pairwise distance adjustments and superimposition of relatively small fragments, it is impressively efficient.

In this study, the required precomputed 3D library was generated by breaking each molecule in an earlier version of PubChem into fragments using the method outlined above, generating a conformation for each of these fragments using SPE, and finally refining the resulting raw conformations through MMFF94s energy minimization.

Conformational Boosting. Boosting is a simple heuristic that can be used in conjunction with SPE and SOS to generate increasingly extended or compact conformations through iterations.[26,28] In the first iteration, a normal SPE or SOS embedding is performed as described in the previous sections, generating a chemically sensible conformation c_1. The lower bounds of all atom pairs l_{ij} are then replaced by the actual interatomic distances d_{ij} in conformation c_1 and used along with the unchanged upper bounds u_{ij} and volume constraints, V^l_{ijkl} and V^u_{ijkl}, to perform a second embedding to generate another conformation, c_2. This process is repeated for a prescribed number of iterations. The lower bounds are then restored to their original default values, and a new sequence of embeddings is performed using a different random number seed. Because the distance constraints in any iteration are always equal to or greater than those in the previous iterations, successively more extended conformations should be generated. This process will never yield a set of distance constraints that are impossible to satisfy, because there exists at least one conformation (i.e., the one generated in the preceding iteration) that satisfies them. Therefore, the conformational space defined by the distance constraints will shrink but not vanish over the iterations, thus effectively biasing the SPE sampling toward more extended geometries. An analogous procedure can be used to generate increasingly compact conformations[26] and has been employed in this work as described in the following paragraphs.

SPE and SOS Parameters. The present study was initiated before the SOS algorithm was available; therefore, the majority of our experiments centered on SPE. Several independent conformational searches with different parameter settings were performed. For the SPE method, the adjustable parameters involved the level and direction of boosting, the use of minimization, and the use of different rmsd thresholds to eliminate duplicate conformations. The various parameter settings and the abbreviated names of the resulting search protocols are listed in Table 1.

Three different levels of boosting were evaluated for SPE. The first, referred to simply as SPE, used boosting in both directions to generate a combination of compact and extended conformations in a ratio of 3:5, following the same protocol used previously.[29] To generate the compact conformations, SPE was used with the options "*-boost 3 -inverse -keepall*". These options instructed the program to generate a series of three conformations for each trial, $\{c_{i,1}, c_{i,2}, c_{i,3}\}$, where i denotes the ith trial. Conformation $c_{i,1}$ was generated using the default distance bounds, $c_{i,2}$ was generated using the interatomic distances in $c_{i,1}$ as upper bounds, and $c_{i,3}$ was generated using the interatomic distances in $c_{i,2}$ as upper bounds. This process was repeated for $N/3$ trials, where N was the desired number of conformations. The first conformation in each trial was generated completely independently using a different random number seed. The extended

conformations were generated in a similar fashion using the options "*-boost 5 -keepall*". The resulting conformations from each run were concatenated to form the final ensemble. By default (see exceptions below), no duplicate check was performed, and all conformations were retained.

To further probe the impact of boosting, calculations were also performed using only compact or only extended conformations (denoted as SPE_COM and SPE_EXT, respectively). The parameter settings were virtually the same as those described above, except that the number of trials was increased accordingly to generate the desired number of conformations.

The SPE and SOS algorithms generate conformations based on purely geometric considerations and do not require any energy calculations. However, both programs have a built-in energy minimization option to allow conformation generation and subsequent energy refinement in the same command. To assess the impact of minimization, the SPE, SPE_COM, and SPE_EXT protocols were repeated using identical settings but with the additional *-minimize* flag. This approach minimizes all generated conformations for a maximum of 1000 steps using the MMFF94s force field and the BFGS variable metric minimization algorithm. Because minimization forces some raw conformations to converge to the same local minima, we used an additional flag to remove duplicate conformations from the final output, using an rmsd cutoff of 0.1 Å. The resulting ensembles are referred to as SPE_MIN_0.1, SPE_COM_MIN_0.1, and SPE_EXT_MIN_0.1, respectively.

The effect of applying a duplicate check to the raw SPE geometries was also tested. Three different rmsd thresholds for removing duplicates were examined: 0.1, 0.4, and 0.8 Å. The resulting ensembles are denoted by SPE_*cutoff*, SPE_COM_*cutoff*, and SPE_EXT_*cutoff*, where *cutoff* is the value of the rmsd threshold employed in the corresponding run (i.e., 0.1, 0.4, and 0.8).

While SPE forms the majority of this work, we also included two SOS protocols to assess the general utility of the method. The first used the default parameters without any boosting (denoted as SOS), and the second used five levels of boosting toward more extended geometries, in a manner similar to the one described above (denoted as SOS_EXT). No inverse boosting was attempted, because at that point there was increasing evidence that both SPE and SOS had a tendency to generate relatively compact geometries. In all cases, the number of trials was set such that the total number of output conformations was equal to 50, 100, 500, 1000, 5000, and 10 000. The resulting conformations were not subjected to rmsd duplicate filtering or energy minimization.

MOE Conformational Import. The MOE conformational import algorithm (MOE_CI) breaks the molecule into overlapping fragments, retrieves precomputed conformations of those fragments, and reassembles them by rigid body superimposition. The first time a fragment conformation is encountered, should its conformation not exist in the fragment database, a stochastic approach is used to generate suitable conformations. The coordinates for the fragments are then appended to the fragment database and used for subsequent molecules containing the same structure. In the present study, fragment conformations with MMFF94x strain energy greater than 5 kcal mol^{-1} were ignored. As with SOS and SPE, six different conformational ensembles were generated for each molecule by setting the maximum number of conformations to 50, 100, 500, 1000, 5000, and 10 000, respectively. All other parameters were left at their default values. Chen and Foloppe found that modification of other parameters did not lead to significant improvements in performance.[31]

MOE Bond Rotation. The MOE bond rotation (MOE_BR) sampling approach is based on the conformation generator used in the MOE docking and pharmacophore elucidation routines. This relatively simple approach treats all ring systems as rigid and generates new conformations by bond rotation. The algorithm starts by converting the input molecule into 3D and minimizing its energy using the MMFF94x force field. From the resulting 3D configuration, new conformations are generated by modifying the torsion angles of the rotatable bonds based on a set of predefined rules. There is no subsequent external refinement/minimization of the resulting conformations. The process generates a user-defined maximum number of conformations, which was set to 50, 100, 500, 1,000, 5000, and 10 000, in accordance with the protocol outlined above.

MOE Stochastic Search. MOE's stochastic search (MOE_SS) is similar to Ferguson's Random Incremental Pulse Search method[8] in that new conformations are generated via random perturbation of the parent conformation but differs in that the perturbation is of bonds rather than Cartesian coordinates. In this study, we have used stochastic search largely as a reference method, using the protocol settings identified by Chen and Foloppe, which gave the best results in their comparative analysis of the MOE and Catalyst conformational search algorithms. However, it was deemed to be a relatively low-throughput approach that was best suited for "detailed characterization of key compounds".[31] Our goal was to see if we could approach (or even improve) the performance of this protocol while reducing the computational cost. A 15 kcal mol^{-1} energy window was used along with the MMFF94x force field and the generalized Born (GB) solvation model, as implemented in MOE.[35] The internal dielectric constant was set to 1 and the external to 80. All other parameters were left at their default values. A maximum of 10 000 conformations were requested for each molecule using an rmsd threshold of 0.1 Å to filter out duplicate structures (i.e., a new conformation was added to the output set only if it differed by more than 0.1 Å rmsd to all previously identified conformations of the same molecule).

Analysis of Conformations. The first measure of search performance was to test if at least one of the generated conformations was similar to the bioactive conformation of the molecule in question. Here, similarity was quantified by the root mean squared distance (rmsd) between the generated conformations and the corresponding bioactive structure. Each generated conformation was first aligned to the bioactive one using MOE's least-squares superimposition procedure (SVL "superpose" function), and the rmsd between the overlaid atoms was computed (heavy atoms only). (We note, parenthetically, that our group has recently developed an improved algorithm to determine the optimal rotation for least-squares superposition using a Newton−Raphson quaternion-based method and an adjoint matrix, that is at least an order of magnitude faster than conventional inversion/decomposition methods.[34]) The total number of ligands with

a conformation within an rmsd of 0.5, 1.0, 1.5, and 2.0 Å to the bioactive conformation was determined and was used to measure the ability of each conformational search protocol to identify bioactive conformations.

Recently, the use of rmsd to assess docking poses to crystal structures has been compared to the real space R-factor (RSR).[36] That work was based on the observation that experimentally determined electron densities are sometimes incomplete and the exact coordinates of certain parts of the bound ligand can be equivocal. In such cases, comparing docking poses to the electron density itself rather than exact atomic positions could help eliminate some of that ambiguity. These measures, however, are not entirely uncorrelated, in that conformations with low rmsd also tend to have low RSR. Even when the electron density is inconclusive, the fitted structure is still in the general vicinity of the true bioactive conformation, and rmsd remains a relevant metric, and one that is much easier to implement and faster to compute.

The diversity of the conformations produced by each search protocol was quantified using the number of unique pharmacophore triplets and quadruplets in the resulting ensembles, as encoded by MOE's 3D pharmacophore fingerprints. This approach is similar to that used in a recent study of library diversity.[37] Both 3-point (piDAPH3) and 4-point (piDAPH4) pharmacophores were computed, using the default definitions of hydrogen-bond donors, acceptors, and hydrophobic centers. These descriptors encode the 3D pharmacophores of each conformation in a fixed-length binary set, where each bit represents the presence or absence of a particular pharmacophore feature. The union (inclusive OR) of the binary sets of all of the conformations generated for a given molecule by a given method represents the entire pharmacophore space sampled by that method, with the total count of "on" bits serving as the ultimate measure of pharmacophore diversity. To normalize for conformational flexibility across different ligands, we report the average number of pharmacophore bits per generated conformation, obtained by dividing the total number of pharmacophore bits by the number of unique conformations identified by each search protocol.

In addition, a "gold standard" conformational ensemble was generated by combining the output of all conformational searches and minimizing with the MMFF94x force field in MOE. As before, an rmsd threshold of 0.1 Å was used to filter out duplicate structures. The same 3- and 4-point pharmacophore fingerprints were computed, and the output from each protocol was compared to the gold standard using the Tanimoto coefficient. This approach penalizes methods with missing pharmacophores from energetically favorable conformations as well as excess pharmacophores from strained conformations.

The strain energy of the raw conformations produced by the various methods was calculated using the E_strain descriptor in MOE. The strain energy is defined as the energy of the raw conformation minus the energy of the nearest local minimum, obtained by molecular mechanics force field minimization using the raw conformation as a starting point. To prepare the conformations for energy minimization, implicit hydrogen atoms were added and charges were calculated. Energies were computed using the MMFF94x force field as implemented in MOE using default settings. Distributions and simple statistics such as the mean strain energy were used to identify which methods produced more highly strained and unrealistic conformations. Preliminary analysis showed that the differences in strain energy between different search protocols were not sensitive to the maximum number of requested conformations; therefore, the strain statistics were computed only for the runs involving 500 conformations per ligand.

Finally, the computational time required to execute the searches was recorded for a random selection of 10 ligands from the combined data set (1GWX, 1I7Z, 1PXI, 1QPE, 1URW, 1UYD, 2BAL, 7EST, 1LAH, and 3ERK; see Table 1). CPU cost was evaluated in a separate postprocessing step, by repeating the calculations for the aforementioned ligands using the same settings that were used for the production runs. All timings were performed on a Windows 2000 PC equipped with a single core 32-bit 2.0 GHz Intel Pentium M processor and 2 GB of RAM. The SPE and SOS programs were executed from the DOS shell (DOS command prompt) and were timed using the *timethis.exe* utility available through the Windows 2000 resource kit. The MOE searches were spawned from within the MOE GUI using custom SVL scripts, and the CPU times were extracted from the SVL command window.

RESULTS AND DISCUSSION

Retrieval of Bioactive Conformations/Differences Across Data Sets. As discussed in the Methods section and illustrated in Figure 1, the Diller collection contains compounds with generally higher molecular weight and more rotatable bonds than the Chen data set. One obvious question is whether these differences have any significant impact on the ability of the various conformational search methods to sample bioactive space. The ability to retrieve the bioactive conformation at different levels of sampling is graphically illustrated in Figure 2 for the Diller and Chen data sets and the SOS and MOE conformational import methods, respectively. Each of these plots shows the percentage of ligands (*y*-axis) for which the generated ensemble contained at least one conformation close to the bioactive structure as determined by different rmsd cutoff values (*x*-axis). Clearly, the greater is the rmsd cutoff, the greater is the probability of finding at least one conformation close to the crystal structure. For the Diller data set, both SPE and MOE conformational import improved as the number of generated conformations increased regardless of the rmsd cutoff employed (Figure 2a and b). More importantly, both methods were able to outperform MOE stochastic search with the Chen settings, which served as the reference method throughout this study. Strikingly, even with only 100 trials, SPE was able to identify the bioactive conformation for more ligands than MOE stochastic search with 10 000 trials. This is despite the fact that the latter method employed an elaborate solvation model and consumed vastly more computational resources than the former (see computational times in Table 2 and detailed discussion in the following sections). MOE conformational import with 500 trials also showed improved retrieval of bioactive conformations over the stochastic search method. This is in contrast to the findings of Chen and Foloppe, who found that increasing the maximum number of conformations requested from conformational import made little appreciable difference in per-

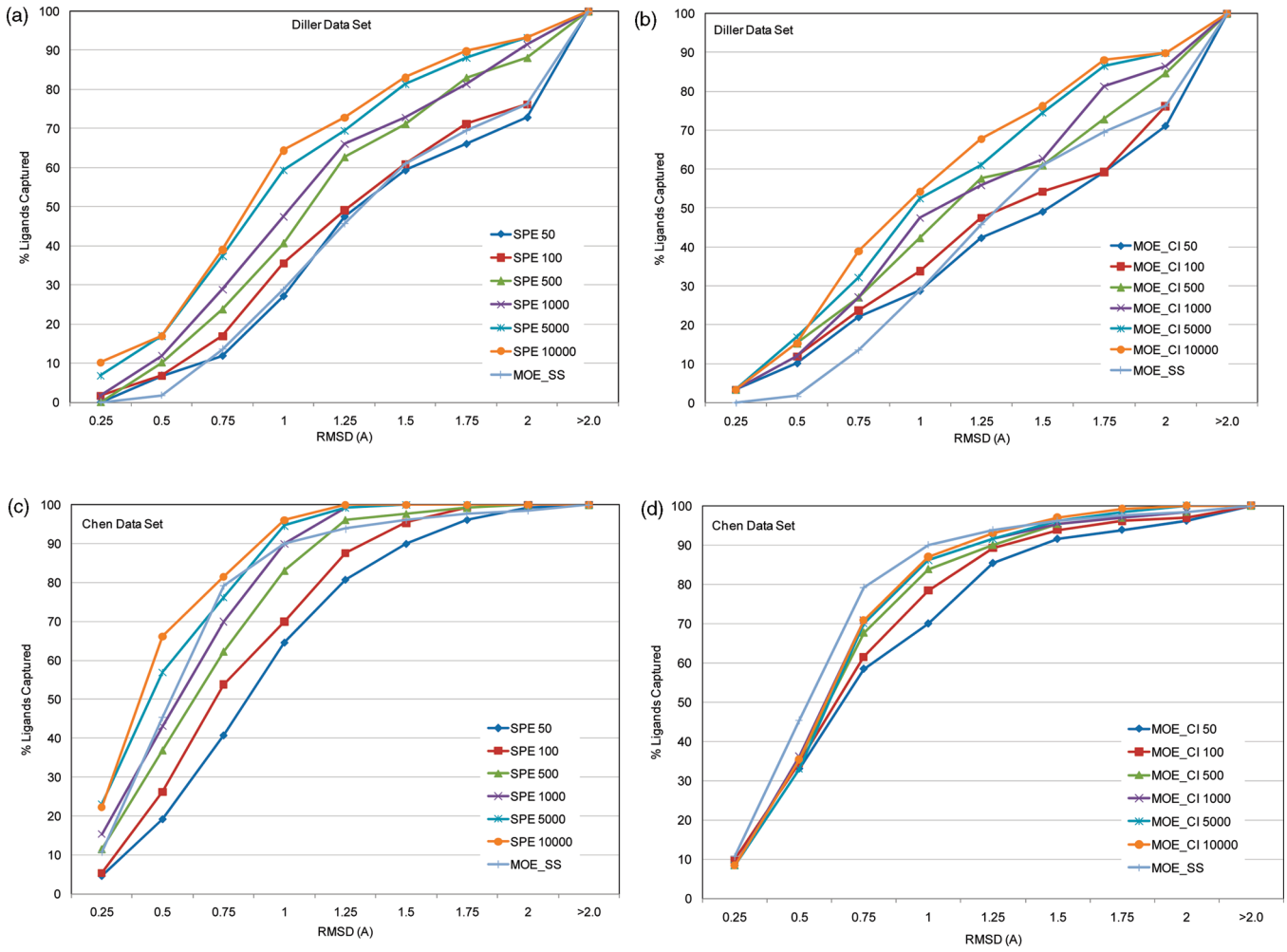

Figure 2. The percentage of ligands reproduced within a particular rmsd from the bioactive conformation for two different methods and data sets. (a) Diller data set using SPE; (b) Diller data set using MOE conformational import; (c) Chen data set using SPE; and (d) Chen data set using MOE conformational import. For comparison, each plot also shows the results from the MOE stochastic search using the optimal settings reported by Chen.

formance. In our hands, both SPE and MOE conformational import performed better than MOE stochastic search despite requiring significantly less sampling. The total number of conformations generated for all 59 ligands in the Diller data set by SPE_100, MOE_CI_500, and MOE_SS_10000 were 5900, 24 016, and 58 596, respectively. These results seem to suggest that when working with larger, more flexible molecules, there may be better and faster alternatives to MOE stochastic search with the Chen settings.

For the Chen data set, both SPE and MOE conformational import again improved with increasing number of requested conformations. However, conformational import reached a limit, which was not the case for SPE. From Figure 2d it can be seen that as the number of requested conformations increases from 1000 to 5000 to 10 000, there is no discernible difference in the number of bioactive conformations retrieved with conformational import. Conformational import was unable to surpass MOE stochastic search, which is in agreement with the Chen study and their general conclusion that MOE stochastic search performed well for lower molecular weight and less flexible compounds. Nevertheless, SPE was able to outperform stochastic search with both 5000 and 10 000 trials. The total number of conformations identified for all 130 ligands in the Chen data set by SPE_5000, MOE_CI_10000, and MOE_SS_10000 was 65 000, 57 250, and 76 831, respectively.

Taken together, these results suggest that the relative performance of conformational search methods depends on the size and flexibility of the molecules under investigation. Overall, SPE outperformed MOE conformational import and MOE stochastic search for both the Diller and the Chen data sets.

Retrieval of Bioactive Conformations/Performance on the Entire Data Set. A summary of the performance of all of the search protocols vis-à-vis their ability to identify bioactive conformations is illustrated in Figure 3. This figure shows the percentage of ligands for which there was at least one conformation identified within 0.5, 1.0, 1.5, and 2.0 Å rmsd to the corresponding crystal structure by each search protocol under evaluation. The ordinate axis is sorted such that the protocols with the best retrieval of ligands within 1.0 Å rmsd of the bioactive conformation are at the top of the plot. SPE_MIN_10000, which used the SPE algorithm to generate 10 000 conformations per molecule followed by molecular mechanics force field minimization, performed the best. As expected, sampling and therefore performance improve as more conformations are requested, and this is true for every method examined. The interest lies in

Table 2. Key Statistics for the Different Conformational Search Protocols[a]

		bioactive Conf retrieval rate at different rmsd cutoffs (Å)				no. of unique fingerprint bits and Confs						
search protocol	max Confs	0.5	1.0	1.5	2.0	nFP3	nFP4	nConfs	nFP3/ nConfs	nFP4/ nConfs	CPU time per mol (s)	strain energy (kcal/mol)
SPE	50	13%	50%	78%	93%	1 162 617	95 174	16 050	72.4	5.9	1.6	271.5
	100	19%	57%	84%	94%	1 527 817	103 168	32 100	47.6	3.2	3.2	
	500	26%	69%	90%	98%	2 607 418	119 220	160 500	16.2	0.7	16.0	
	1000	30%	74%	93%	98%	3 069 776	124 496	321 000	9.6	0.4	43.9	
	5000	42%	83%	96%	99%	4 060 352	135 588	1 605 000	2.5	0.1	177.6	
	10 000	47%	87%	97%	99%	4 443 897	140 370	3 210 000	1.4	0.0	359.5	
SPE_COM	50	14%	46%	69%	85%	1 160 773	85 717	16 050	72.3	5.3	2.3	188.6
	100	16%	49%	73%	89%	1 481 384	90 814	32 100	46.1	2.8	5.3	
	500	28%	60%	83%	94%	2 196 284	99 341	160 500	13.7	0.6	22.7	
	1000	34%	65%	86%	94%	2 476 483	102 400	321 000	7.7	0.3	32.1	
	5000	38%	69%	90%	96%	3 015 065	108 436	1 605 000	1.9	0.1	170.1	
	10 000	42%	73%	91%	96%	3 216 354	111 157	3 210 000	1.0	0.0	335.7	
SPE_EXT	50	16%	49%	79%	93%	642 470	74 028	16 050	40.0	4.6	2.0	231.0
	100	19%	56%	85%	96%	894 310	82 520	32 100	27.9	2.6	3.7	
	500	28%	70%	92%	97%	1 599 881	95 759	160 500	10.0	0.6	17.9	
	1000	31%	74%	95%	98%	1 915 094	99 931	321 000	6.0	0.3	34.6	
	5000	42%	85%	96%	99%	2 581 853	107 515	1 605 000	1.6	0.1	171.6	
	10 000	47%	86%	97%	99%	2 838 564	110 529	3 210 000	0.9	0.0	343.7	
SPE_MIN_0.1	50	26%	62%	87%	97%	648 702	59 292	12 277	52.8	4.8	18.4	6.7
	100	28%	69%	89%	96%	799 513	61 955	20 671	38.7	3.0	37.3	
	500	32%	77%	94%	99%	1 140 380	66 354	75 638	15.1	0.9	188.2	
	1000	36%	80%	96%	100%	1 264 090	68 299	130 657	9.7	0.5	375.2	
	5000	38%	88%	98%	100%	1 489 386	71 493	457 977	3.3	0.2	1881.1	
	10 000	40%	89%	99%	100%	1 566 971	73 024	779 136	2.0	0.1	3784.5	
SPE_COM_MIN_0.1	50	23%	64%	85%	94%	699 664	61 247	11 638	60.1	5.3	20.0	6.9
	100	26%	68%	89%	96%	868 514	63 686	20 911	41.5	3.0	38.1	
	500	33%	75%	93%	99%	1 214 601	68 123	77 657	15.6	0.9	195.9	
	1000	36%	78%	94%	99%	1 332 957	69 618	135 003	9.9	0.5	399.0	
	5000	39%	84%	98%	100%	1 551 806	73 494	473 633	3.3	0.2	1985.9	
	10 000	38%	86%	98%	100%	1 627 819	74 565	805 195	2.0	0.1	3990.1	
SPE_EXT_MIN_0.1	50	26%	61%	85%	96%	552 983	56 251	10 692	51.7	5.3	18.0	6.8
	100	29%	69%	90%	98%	705 144	59 780	19 163	36.8	3.1	36.1	
	500	33%	78%	95%	99%	1 035 645	64 925	71 235	14.5	0.9	180.7	
	1000	34%	82%	97%	99%	1 164 002	66 434	123 030	9.5	0.5	361.3	
	5000	38%	87%	98%	100%	1 404 755	69 920	425 927	3.3	0.2	1838.0	
	10 000	39%	89%	99%	100%	1 488 299	71 597	720 345	2.1	0.1	3715.9	
SPE_0.1	50	14%	49%	81%	94%	1 157 903	94 861	15 955	72.6	5.9	1.8	294.0
	100	16%	54%	83%	96%	1 529 687	102 936	31 440	48.7	3.3	3.7	
	500	25%	67%	91%	98%	2 600 298	118 597	148 618	17.5	0.8	28.3	
	1000	31%	76%	93%	98%	3 063 139	124 309	292 920	10.5	0.4	77.9	
	5000	42%	83%	96%	99%	4 053 081	135 785	1 429 044	2.8	0.1	1332.9	
	10 000	45%	86%	97%	99%	4 436 728	140 122	2 813 952	1.6	0.0	4956.8	
SPE_0.4	50	15%	50%	79%	93%	1 141 550	94 489	14 833	77.0	6.4	1.8	345.0
	100	15%	56%	83%	96%	1 518 708	102 895	27 069	56.1	3.8	3.6	
	500	27%	69%	92%	98%	2 583 540	118 540	120 391	21.5	1.0	26.9	
	1000	32%	72%	94%	98%	3 036 811	123 941	227 739	13.3	0.5	75.4	
	5000	38%	83%	96%	99%	4 021 255	134 896	999 633	4.0	0.1	1370.1	
	10 000	44%	86%	96%	99%	4 401 537	139 888	1 875 306	2.3	0.1	4662.7	
SPE_0.8	50	10%	48%	78%	93%	1 092 574	92 401	11 529	94.8	8.0	1.7	409.2
	100	10%	56%	83%	94%	1 464 056	101 236	19 950	73.4	5.1	3.5	
	500	15%	66%	91%	98%	2 482 745	115 807	77 820	31.9	1.5	24.9	
	1000	18%	73%	92%	98%	2 917 989	121 097	138 999	21.0	0.9	63.2	
	5000	21%	83%	96%	98%	3 836 905	131 778	526 695	7.3	0.3	783.2	
	10 000	23%	86%	96%	99%	4 190 474	135 851	929 011	4.5	0.1	2387.4	
SOS	50	13%	43%	68%	88%	1 014 346	76 606	16 050	63.2	4.8	0.1	107.0
	100	15%	46%	74%	91%	1 277 139	81 030	32 100	39.8	2.5	0.2	
	500	23%	60%	83%	95%	1 874 707	90 291	160 500	11.7	0.6	0.9	
	1000	23%	61%	86%	95%	2 110 687	94 103	321 000	6.6	0.3	1.6	
	5000	31%	68%	90%	96%	2 653 479	103 597	1 605 000	1.7	0.1	8.0	
	10 000	35%	71%	92%	97%	2 860 172	106 522	3 210 000	0.9	0.0	16.8	

Table 2 Continued

search protocol	max Confs	bioactive Conf retrieval rate at different rmsd cutoffs (Å) 0.5	1.0	1.5	2.0	no. of unique fingerprint bits and Confs nFP3	nFP4	nConfs	nFP3/ nConfs	nFP4/ nConfs	CPU time per mol (s)	strain energy (kcal/mol)
SOS_EXT	50	16%	50%	82%	94%	628 884	68 856	16 050	39.2	4.3	0.2	107.7
	100	19%	56%	87%	96%	856 681	75 793	32 100	26.7	2.4	0.3	
	500	26%	69%	90%	97%	1 445 027	88 056	160 500	9.0	0.5	1.3	
	1000	31%	72%	91%	97%	1 720 402	92 721	321 000	5.4	0.3	2.0	
	5000	39%	81%	93%	98%	2 301 005	102 394	1 605 000	1.4	0.1	8.4	
	10 000	41%	82%	96%	99%	2 546 276	107 043	3 210 000	0.8	0.0	26.6	
MOE_CI	50	24%	57%	78%	91%	259 866	39 266	11 856	21.9	3.3	11.2	39.7
	100	25%	62%	83%	93%	327 388	43 102	21 138	15.5	2.0	11.9	
	500	28%	69%	88%	96%	480 491	49 292	68 831	7.0	0.7	13.3	
	1000	27%	72%	89%	97%	608 775	53 070	119 935	5.1	0.4	13.2	
	5000	28%	77%	93%	98%	823 604	57 822	381 249	2.2	0.2	28.9	
	10 000	29%	79%	94%	98%	884 862	58 817	644 565	1.4	0.1	28.9	
MOE_BR	50	15%	54%	83%	95%	669 451	68 655	14 511	46.1	4.7	0.6	150.0
	100	20%	59%	83%	96%	887 337	73 759	28 017	31.7	2.6	0.7	
	500	24%	67%	91%	98%	1 352 914	81 477	122 812	11.0	0.7	0.7	
	1000	27%	68%	94%	99%	1 542 185	83 648	225 446	6.8	0.4	0.8	
	5000	29%	74%	93%	99%	1 934 722	87 563	902 290	2.1	0.1	2.3	
	10 000	30%	75%	95%	99%	2 012 879	87 607	1 588 277	1.3	0.1	4.0	
MOE_SS	10 000	30%	70%	88%	93%	1 097 628	70 979	240 770	4.6	0.3	5647.1	0.0

[a] nFP3 and nFP4 refer to the number of unique 3-point (piDAPH3) and 4-point (piDAPH4) pharmacophore bits as calculated in MOE. nConfs refers to the number of conformations identified by each protocol. FP3/nConfs and FP4/nConfs represent the average number of unique 3- and 4-point pharmacophores per conformation. CPU time per molecule represents the amount of time required to complete the search for a single molecule, averaged over 10 randomly selected molecules from the data set. Strain energy represents the median average local strain energy in kcal/mol between each conformation and its nearest local minimum, calculated on the basis of the protocols involving 500 trials.

identifying the methods that, for the same number or fewer trials, perform better. Some of the methods are highlighted in Figure 3 with different colors. The MOE stochastic search using the Chen and Foloppe settings, which is highlighted in yellow, showed average performance at all rmsd cutoffs.

Looking again at Figure 3, it appears that for protocols with 5000 and 10 000 conformations, SPE slightly outperforms SOS with or without boosting. (In the following discussion, we use the term "boosting" to refer to biasing toward more extended geometries, "inverse boosting" to refer to biasing toward more compact conformations, and "bidirectional boosting" to refer to biasing in both directions, as described in the Methods section.) Moreover, both SPE and SOS with boosting (SPE_EXT and SOS_EXT) rank higher than SOS without boosting (SOS). However, as shown in Table 2, these differences are rather marginal; for 5000 trials, the difference in the bioactive conformation retrieval rate with SPE_EXT and SOS_EXT is only 4% (85 and 81%, respectively). In fact, SOS_EXT performs within 0−2% of the SPE method for protocols involving 50, 100, 500, 1000, and 5000 trials. These data suggest that the use of precomputed template geometries in the SOS method has a negligible impact on the ability to retrieve bioactive conformations.

The positive impact of boosting on sampling bioactive space is evident at all levels. Comparing the SOS and SOS_EXT protocols, where all parameters were identical except for the use of boosting, the latter method showed 9% improvement over the former for the protocols involving 500 trials (60% for SOS_500, and 69% for SOS_EXT_500). The same is true for SPE, where we observed a 10% improvement from SPE_COM to SPE_EXT at 1.0 Å rmsd for protocols involving 500 trials (60% for SPE_COM_500, and 70% for SPE_EXT_500). The results with bidirectional boosting, which represents a 3:5 mix of compact and extended conformations, were very similar to those obtained from regular boosting regardless of the number of trials requested. These results are fully consistent with our previous observation that SPE tends to generate relatively compact conformations and that inverse boosting offers little practical benefit. It also reflects the fact that bioactive conformations tend to be more extended than random ones, and therefore biasing the search toward extended geometries increases the likelihood of identifying the biologically active ones.

One parameter that seems to have a significant impact is energy minimization. Comparing SPE_COM_MIN, SPE_EXT_MIN, and SPE_MIN with the corresponding nonminimized ensembles shows a notable change in the rate of retrieval of bioactive conformations. Even with just 100 trials per molecule, these three methods outperformed all other approaches in this study and were within 1−2% of the MOE stochastic search using the best settings identified by Chen and Foloppe. This is particularly encouraging considering that these methods generated an order of magnitude fewer conformations than MOE stochastic search (∼20 000 as compared to ∼240 000; see Table 2). Interestingly, minimization had the largest effect on the protocols with the fewer trials. For example, SPE with 50 trials retrieved 50% of the bioactive conformations within 1.0 Å rmsd, whereas with minimization this percentage increased to 62%. Generally, at lower rmsd cutoffs, minimization seems to work better when the number of trials is small, but gets progressively

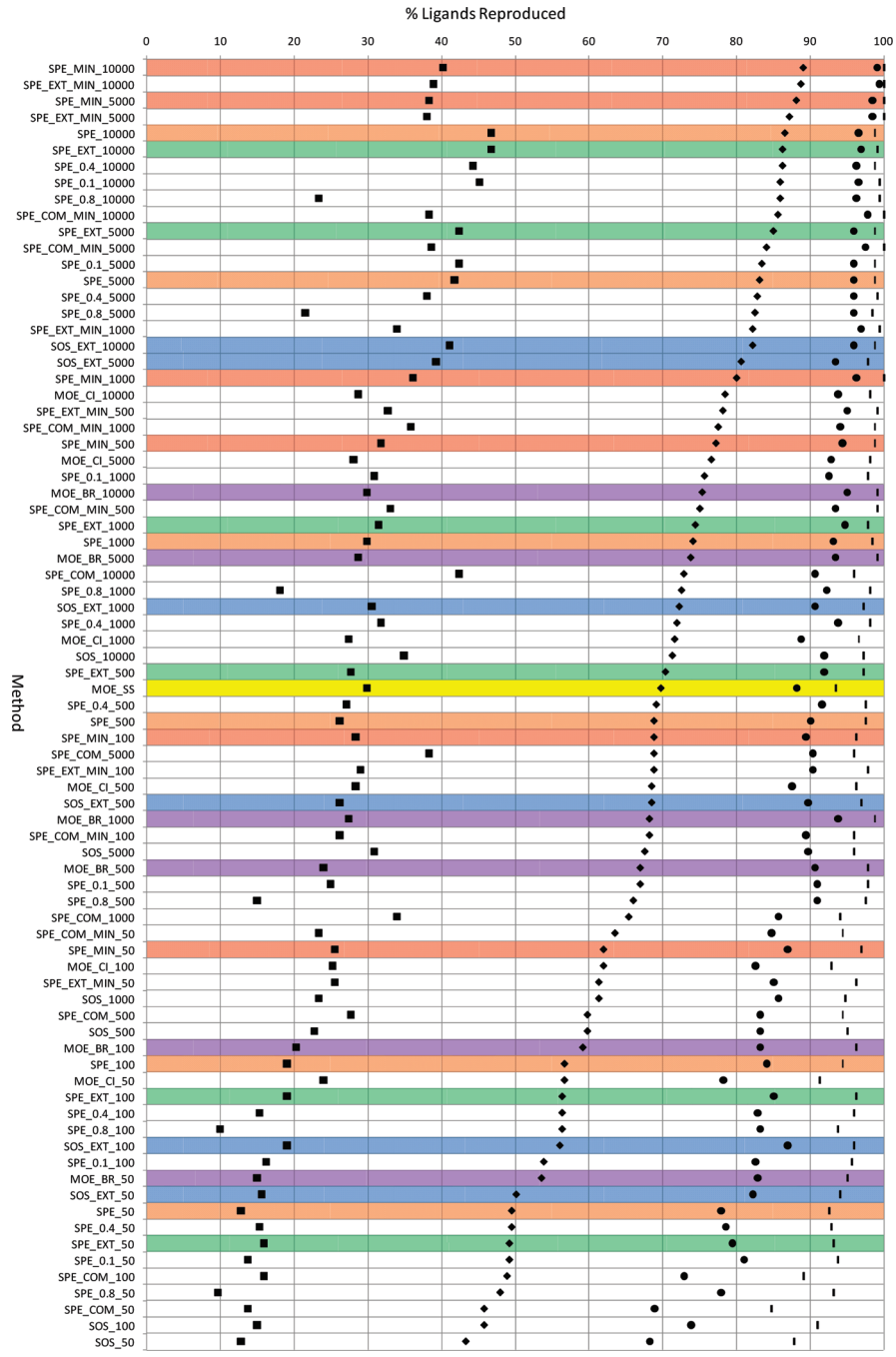

Figure 3. The percentage of ligands reproduced within 0.5 (■), 1.0 (◆), 1.5 (●), and 2.0 (−) Å rmsd from the bioactive conformation for different methods and settings. The search protocols are sorted top to bottom in decreasing fraction of ligands reproduced within 1.0 Å, with the best performing methods placed at furthest left. Color-coding is used to identify selected methods: MOE_BR (purple), SPE (orange), SPE_MIN (red), SPE_EXT (green), SOS_EXT (blue), and MOE_SS (yellow).

worse as more and more conformations are generated. At higher rmsd cutoffs, minimization almost always has a positive impact, although the effect becomes less pronounced as the number of trials increases.

To further examine the effects of minimization, we also subjected the MOE conformational import (MOE_CI) and bond rotation (MOE_BR) methods to similar post processing. Without minimization, these two methods with 50 conformations per molecule identified at least one conformation within 1.0 Å rmsd from the bioactive structure for 57% and 54% of the ligands, respectively. Upon minimization, the retrieval rate was slightly reduced for conformational import (55%) and slightly increased for bond rotation (57%). Improvement, if any, was much less pronounced than seen with SPE. This suggests that it is not sufficient to minimize the raw conformations from any method and expect such improvements. We attribute the improvement in performance to the synergistic effects of minimizing more diverse conformations as generated by SPE, as discussed in the following section.

The rmsd cutoff used to eliminate duplicate conformations had minimal impact on the ability to retrieve the bioactive structure. Duplicate filtering at 0.1 Å rmsd had almost no effect, and at 0.4 and 0.8 Å the difference was no more than 2−4% for virtually any level of sampling (number of trials). The only exception is that there were significantly fewer ligands identified within 0.5 Å rmsd to the bioactive structure

when filtering duplicates at 0.8 Å rmsd. This result, however, is expected because duplicate elimination is susceptible to the order in which the conformations are generated. Indeed, a conformation within 0.5 Å from the crystal structure would have been eliminated if it were within 0.8 Å to a previously generated conformation that was itself more than 0.5 Å from the crystal structure. Therefore, we recommend the use of a sensible rmsd cutoff that is large enough to eliminate redundant structures and minimize storage requirements, yet not too large so as to preclude the sampling of important regions of conformational space due to artifacts associated with the sampling sequence.

Considering the MOE conformational import and bond rotation methods, in general they performed well for fewer conformations but relatively worse as the number of trials increased. Indeed, MOE conformational import requesting 50 conformations per molecule outperformed the standard, nonminimized SPE and SOS protocols. Performance continued to improve with increasing number of conformations, but not to the same extent as with the other methods. The difference in percent retrieval between 50 and 10 000 conformations was 22% for conformational import and 21% for bond rotation, but 37% for SPE. With 500 conformations per molecule, SPE and conformational import showed comparable performance, both identifying 69% of ligands within 1.0 Å rmsd of the crystal structure. On average, bond rotation was 2−4% worse than conformational import at all levels of sampling within the same rmsd cutoff.

Conformation Diversity Captured with 3D Fingerprints. The extent of conformational coverage was the second important aspect considered. A good conformational search method must cast a wide net over the potential energy landscape and sample as broad a range of molecular geometries as possible. In some respects, the ability to retrieve the bioactive conformation is itself a measure of the extent of sampling. We have previously studied the performance of SPE using the radius of gyration with 10 000 conformations per molecule.[29] The boosting heuristic was partly designed to promote the generation of more extended geometries, and we have shown that, when applied to SPE and SOS, they both performed very well in this regard. In this work, the extent of sampling was assessed using the number of unique 3-point and 4-point pharmacophores combined with the overall number of conformations generated.

The numbers of unique 3- and 4-point pharmacophores identified with each protocol are listed in Table 2. The SPE method yields by far the greatest pharmacophore diversity of all approaches evaluated. With 10 000 trials per molecule, SPE generated ∼4.4 million 3-point and over 140 000 4-point pharmacophores. In comparison, the MOE stochastic search identified ∼1.1 million and ∼71 000, respectively. However, upon force field minimization, the number of unique conformations and pharmacophores identified by SPE was reduced by a factor of 2−4, due to the fact that many of the raw conformations were minimized to similar geometries containing equivalent pharmacophores. This suggests that SPE produces some conformations and pharmacophores, which are physically unrealistic, and that additional refinement or minimization is necessary if this method is to be useful in a practical setting. Eliminating duplicate conformations prior to minimization reduced the number of conformations significantly (particularly as the number of trials increased), but the impact on the number of unique pharmacophores was minimal even at relatively high rmsd cutoffs (SPE vs SPE_0.1, SPE_0.4, and SPE_0.8 in Table 2). Interestingly, the average number of unique pharmacophores per conformation (nFP3/nConfs and nFP4/nConfs in Table 2) decreases upon minimization for searches involving very few trials, but increases as more and more conformations are generated. The effect is more pronounced for extended geometries, and less pronounced for compact ones. These results compare very favorably with MOE stochastic search; indeed, a mere 500 minimized conformations generated by SPE (SPE_MIN_0.1) captured just as much pharmacophore diversity as MOE_SS, whereas given the same number of trials (10 000), SPE produced nearly 40% more 3-point pharmacophores.

The SOS method did not yield as much pharmacophore diversity as raw SPE, but produced twice as many pharmacophores as compared to SPE coupled with minimization. Furthermore, SOS was nearly 4 times as effective as MOE_CI, and almost 1.5−2 times as effective as MOE_BR at comparable levels of sampling. More impressively, even 50 SOS conformations were enough to capture almost the same amount of pharmacophore diversity as 10 000 conformations obtained with MOE stochastic search using the Chen settings.

Comparison to the Gold Standard Ensemble. A good search method must produce conformations that are not only diverse but also energetically plausible. It was shown above that minimization reduced the number of pharmacophores generated by SPE, suggesting the presence of physically unrealistic conformations in the output of the sampling procedure. Although irrelevant high energy conformations can be removed via energy minimization, this step imparts additional computational cost. To assess the quality of the sampling protocols, we compared them to a "gold standard" ensemble constructed by combining the output from all protocols, minimizing the conformations, and removing duplicates within an rmsd threshold of 0.1 Å. The resulting ensemble comprised over six million conformations. The 3- and 4-point pharmacophores were calculated in the manner described above, and the output of each protocol was compared to the gold standard via the Tanimoto similarity coefficient. This metric effectively penalizes for missing conformations from incomplete sampling and high energy conformations from unproductive sampling. Similar trends were observed for 3- and 4-point pharmacophores; therefore, details are given only for the 3-point pharmacophore fingerprints.

Figure 4 shows the Tanimoto similarity for each protocol with increasing number of requested conformations. As expected, the overall similarity is generally high as each individual method captures a reasonable fraction of the total pharmacophore space. The worst protocol was MOE conformational import (MOE_CI) with 50 requested conformations per molecule, with a similarity of 0.86. In Figure 4, the similarity curves are separated into three different plots according to the pattern of change in the similarity score as a function of the number of requested conformations. In the first plot, SPE_COM, SPE_EXT, SOS_EXT, and MOE_BR show relatively little change, with the Tanimoto score slightly increasing for the first few hundred conformations, and slightly decreasing afterward. This is gratifying because it

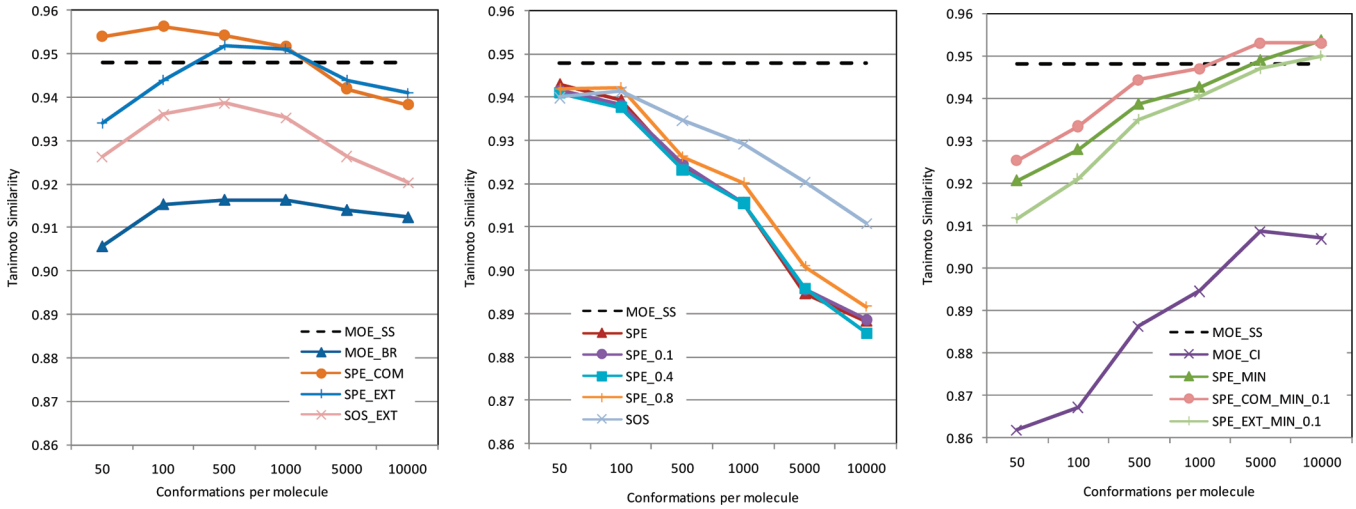

Figure 4. The Tanimoto similarity between each protocol and the gold standard ensemble. The data were separated into three different plots according to the change in similarity as a function of the number of conformations requested. The MOE stochastic search results (horizontal dashed line) are shown as reference in each plot.

suggests that these protocols continue to produce reasonable (although not necessarily novel) geometries and pharmacophores as the level of sampling increases. The SPE_EXT and SOS_EXT protocols show the most notable improvement from 50 to 500 conformations, and SPE_COM and SPE_EXT even outperform the more expensive MOE stochastic search method, which again confirms the positive effect of boosting.

In contrast, the SPE, SPE_0.1, SPE_0.4, SPE_0.8, and SOS protocols show a decline in the Tanimoto scores with increasing number of conformations, which suggests that the new conformations introduce pharmacophores that are not present in the gold standard ensemble. We conjecture that these are irrelevant pharmacophores stemming from poorly optimized, high-energy geometries. The comparable behavior between the SPE methods with and without rmsd filtering is also expected, given that removal of duplicates reduces the absolute number of conformations but has little impact on the diversity of the resulting pharmacophores. The SOS method shows a more modest decline in the similarity score as compared to SPE, due to the use of precomputed templates that enforce more optimal geometries.

Finally, the four protocols that made use of force field refinement, MOE_CI, SPE_MIN, SPE_COM_MIN_0.1, and SPE_EXT_MIN_0.1, showed increased similarity to the gold standard with increasing number of generated conformations. This is the most desirable behavior, but it comes at a considerably higher computational cost.

Conformation Strain Energy. Strain energy is another criterion that can be used to assess the quality of the generated conformations. The strain energy of a conformation is defined as the difference between the energy of the raw conformation produced by the search method and its nearest local minimum. As discussed in the Methods section, this was calculated only for the protocols involving 500 trials per molecule. The median value for each protocol across all ligands is reported in the last column of Table 2. Given that typical force fields involve very steep-welled potentials where small changes in geometry can lead to large changes in energy, one should not be surprised that some of the strain energies reported in Table 2 are very large. Naturally, methods that do not use any energy-based refinement produce more strained conformations. This is particularly true for the SPE protocols, which show the higher energies of all methods examined. Interestingly, the strain energies are somewhat reduced when boosting is employed in either direction, but the overall numbers suggest that SPE raw conformations depart from ideal covalent geometry. This situation becomes particularly acute when the raw conformations are filtered for duplicates, especially at higher rmsd thresholds. At 0.8 Å rmsd, for example, the average strain energy increases to 409.2 kcal/mol, suggesting that filtering duplicate conformations at the raw SPE level makes the method more vulnerable to "exotic" geometries, which displace more reasonable structures that happen to be visited later during the sampling process. What is not obvious from this analysis is what is the minimum effort required to "clean up" these raw conformations and make them comparable to those obtained from competing methods.

Of note is the significant improvement with SOS as compared to SPE. This is clearly attributable to the use of precomputed fragment templates, which enforce locally ideal geometry. We believe that the residual strain is at the fragment interfaces, which are optimized using pairwise distance adjustments similar to SPE.

MOE_CI shows an average strain energy of ~40 kcal/mol, which is very competitive as compared to the other methods. Naturally, methods that use energy minimization during or after the search did very well in this respect (SPE_MIN, MOE_SS). Interestingly MOE_BR, which perturbs preoptimized conformations through simple bond rotation, produces conformations with considerably higher strain energies than SOS (150 kcal/mol), due to the inability of the molecule to relax to accommodate any steric clashes resulting from the assigned torsions.

CPU Times. The final metric used to compare the various methods is the CPU time required to perform the search. Given infinite computing resources and time, one would perform a full systematic conformational search on every molecule of interest. However, such an approach is only practical for the smallest and least flexible molecules. A good conformational search method must not only be effective, it must also be efficient.

The second-to-last column in Table 2 lists the mean CPU time per molecule for each method and each level of sampling, averaged over the 10 ligands as described in the Methods section. Dividing this by the maximum number of requested conformations yields an estimate of the time per generated (more accurately, retained) conformation. These timings reveal several interesting trends. First, MOE stochastic search, which serves as our reference method, was the slowest of all methods, requiring an average of 0.565 s per conformation using the computer system described in the Methods section. Clearly, this is a low-throughput technique that is best suited for detailed studies of relatively few compounds or small libraries.[31]

SPE was an order of magnitude faster, requiring 0.036 s per conformation. As expected, the computational time for the SPE and SOS methods scales linearly with the number of conformations, and this is true with or without boosting. Boosting introduces a negligible overhead to the overall process, which corresponds to the time required to measure all interatomic distances and update the bounds matrix. The rate-limiting step is the embedding itself, which is what makes boosting so appealing; it is simple, effective, and efficient. In contrast, minimization increases the CPU time by a factor of 10, to an average of 0.36−0.40 s per conformation, which is comparable to MOE stochastic search. This time, however, also includes checking for and eliminating duplicate conformations, a process that is quadratic in nature and becomes the limiting factor as the number of conformations grows beyond a certain point. The lower is the rmsd threshold, the more conformations are retained, and the longer it takes to detect duplicates.

SOS is remarkably fast, requiring an average of only 0.002 s per conformation. MOE conformational import and MOE bond rotation are the methods that show the most unusual scaling, with the average cost per conformation becoming smaller and smaller as the total number of generated conformations increases. This reflects a relatively large initialization time for both methods, which becomes less of a factor as sampling increases. At 10 000 conformations, MOE_CI is comparable to SOS with an average cost of 0.003 s per conformation, but at 500 conformations (which is a common practical limit for large multiconformation database construction), it is more than 10 times slower (0.027 s for MOE_CI versus 0.003 s for SOS). MOE_BR is the fastest of all methods examined, particularly at large sampling levels, requiring less than 0.001 s per conformation at 10 000 trials. However, as discussed above, the efficiency of both MOE_CI and MOE_BR comes at the expense of conformational diversity. For MOE_BR, it also comes at the expense of the quality of conformations, which have relatively large strain energies (although they are still significantly better than those generated by SPE without minimization).

CONCLUSIONS

The intent of the present study was to examine the utility of the relatively newer SPE and SOS conformational search algorithms in everyday modeling tasks. Our aim was not to perform an exhaustive evaluation of all possible combinations of options and parameter settings, but to assess whether the settings recommended in the original publications would lead to sensible outcomes in a practical setting. The practical application of any modeling tool always involves compromise.

Our analysis is reminiscent of the classic engineering triangle, which describes the interplay between the three core elements of an engineering project, quality, time, and cost. These three attributes describe how well, how fast, and how cheaply a solution is obtained. In our context, the quality of a conformational search method encompasses a number of factors such as the ability to sample energetically sensible and biologically relevant space while maintaining sufficient diversity to accommodate hitherto unobserved (but highly probable) binding modes. Time is equally important. A timely result is far more valuable than one that is obtained after the fact and has no ability to influence experiments. On the surface, the third element, cost, may appear the least relevant, but it too has a subtle but profound impact. Licensing costs and prior familiarity are often the determining factors in what method is chosen in a particular environment. More importantly, licensing costs are often related to the development costs required to build and maintain the solution, and ultimately reflect the complexity of the solution itself. Indeed, simplicity is often an under-appreciated factor.

The results presented above give us hope that the SPE and SOS methods hold great promise for a variety of modeling applications. Comparing the bioactive conformation retrieval rates between SPE and MOE conformational import on the Diller and Chen data sets showed that SPE performed relatively better for more flexible molecules. SPE outperformed an elaborate variant of MOE stochastic search on the Diller data set while generating significantly fewer conformations in a much shorter amount of time. In addition, SPE improved with increasing number of trials, whereas MOE conformational import reached a limit beyond which additional conformations did not help. Other work from our group has shown that SPE and SOS perform excellently on macrocyclic compounds, where most systematic search methods collapse.[38] Taken together, these results suggest that SPE should be considered among the best approaches for challenging, flexible molecules.

While SOS and SPE performed similarly in terms of retrieving bioactive conformations, SOS was an order of magnitude faster and generated much better geometries with lower strain energies. Confirming our previous studies, we also found that the boosting heuristic had a positive impact in directing the search toward bioactive relevant space.

The question of most interest to us when we embarked on this study was whether the SPE and SOS methods had adequate performance at lower sampling levels, which would determine their suitability for building multiconformation databases for pharmacophore and shape-based searching. Our results confirmed that SOS coupled with boosting and 500 conformations per molecule had comparable performance to MOE stochastic search with 10 000 conformations, while having vastly lower CPU demands.

Applying an rmsd filter to reduce the number of output conformations did not adversely affect the bioactive retrieval rate, and the 3D pharmacophore diversity of the resulting ensemble improved. However, the resulting conformations were more dissimilar to the gold standard and had significantly higher strain energies, so we would not recommend rmsd filtering prior to minimization.

The high strain energies of the SPE conformations are of some concern. When minimized, the conformational ensembles produced by SPE improved greatly, but this came

at a significant cost in terms of speed. However, the key result of this study is that the sampling capacity of SPE is such that fewer minimized conformations obtained by SPE capture more biologically relevant conformational and pharmacophore diversity than much larger ensembles obtained by other methods. Thus, generating fewer conformations with SPE and paying the additional price of minimization may still be preferred over spending an equivalent amount of time sampling more extensively with alternative methods. These other methods, like MOE conformational import and MOE bond rotation, did not improve to the same extent with minimization, which suggests that they have significantly lower sampling potential.

The above also suggests a reasonable compromise. In this work, we used a relatively long minimization procedure involving up to 1000 steps. Yet, experience with gradient minimizers suggests that most of the strain is relieved in the earlier cycles and that the bulk of the effort is spent fine-tuning the geometry to perfection. We believe that most practical applications do not require "perfect" conformations, and therefore the number of minimization steps (and computational time) can be significantly reduced without affecting the modeling outcome. This hypothesis is currently under investigation.

REFERENCES AND NOTES

(1) Stockwell, G. R.; Thornton, J. M. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **2006**, *356*, 928–944.

(2) Leach, A. R. Survey of methods for searching the conformational space of small and medium-sized molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1991.

(3) Lipton, M.; Still, W. C. The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space. *J. Comput. Chem.* **1988**, *9*, 343.

(4) Bruccoleri, R. E.; Karplus, M. Chain closure with bond angle variations. *Macromolecules* **1985**, *18*, 2767.

(5) Bruccoleri, R. E.; Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **1987**, *26*, 137.

(6) Go, N.; Scheraga, H. A. Ring closure and local conformational deformations of chain molecules. *Macromolecules* **1970**, *3*, 178.

(7) Saunders, M. Stochastic explorations of molecular mechanics energy surfaces. Hunting for the global minimum. *J. Am. Chem. Soc.* **1987**, *109*, 3150.

(8) Ferguson, D. M.; Raber, D. J. A new approach to probing conformational space with molecular mechanics: random incremental pulse search. *J. Am. Chem. Soc.* **1989**, *111*, 4371.

(9) Chang, G.; Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(10) Li, Z.; Scheraga, H. A. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611–6615.

(11) Jorgensen, W. L.; Tirado-Rives, J. Monte Carlo vs molecular dynamics for conformational sampling. *J. Phys. Chem.* **1996**, *100*, 14508–14513.

(12) Bostrom, J. Reproducing the conformations of protein-bound ligands: A critical evaluation of several popular conformational searching tools. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1137–1152.

(13) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21*, 449–462.

(14) Perola, E.; Charifson, P. S. Conformational analysis of drug-like molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499–2510.

(15) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411–428.

(16) Vieth, M.; Hirst, J. D.; Brooks, C. L. Do active site conformations of small ligands correspond to low free-energy solution structures. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 563–572.

(17) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45*, 422–430.

(18) Tirado-Rives, J.; Jorgensen, W. L. Contribution of conformation focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **2006**, *49*, 5880–5884.

(19) Diller, D. J.; Merz, K. M. Can we separate active from inactive conformations. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 105–112.

(20) Crippen, G. M. Rapid calculation of coordinates from distance matrices. *J. Comput. Phys.* **1978**, *26*, 449–452.

(21) Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. Conformational analysis using distance geometry methods. *J. Mol. Graphics Modell.* **1997**, *15*, 18.

(22) Agrafiotis, D. K. Stochastic proximity embedding. *J. Comput. Chem.* **2003**, *24*, 1215–1221.

(23) Rassokhin, D. N.; Agrafiotis, D. K. A modified update rule for stochastic proximity embedding. *J. Mol. Graphics Modell.* **2003**, *22*, 133–140.

(24) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational sampling by self-organization. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1186–1191.

(25) Zhu, F.; Agrafiotis, D. K. Self-organizing superimposition algorithm for conformational sampling. *J. Comput. Chem.* **2007**, *28*, 1234–1239.

(26) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *J. Comput. Chem.* **2006**, *27*, 1962–1969.

(27) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869–15872.

(28) Agrafiotis, D. K.; Gibbs, A.; Zhu, F.; Izrailev, S.; Martin, E. Conformational boosting. *Aust. J. Chem.* **2006**, *59*, 874–878.

(29) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: A comparative study. *J. Chem. Inf. Model.* **2007**, *47*, 1067–1086.

(30) Agrafiotis, D. K.; Bandyopadhyay, D.; Carta, G.; Knox, A. J.; Lloyd, D. G. On the effects of permuted input on conformational sampling of drug-like molecules: an evaluation of stochastic proximity embedding. *Chem. Biol. Drug Des.* **2007**, *70*, 123–133.

(31) Chen, I. J.; Foloppe, N. Conformational sampling of druglike molecules with MOE and Catalyst: Implications for pharmacophore modeling and virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773–1791.

(32) O'Shea, R.; Moser, H. E. Physicochemical properties of antibacterial compounds: Implications for drug discovery. *J. Med. Chem.* **2008**, *51*, 2871–2878.

(33) Payne, D. J.; Gwynn, M. N.; Holmes, D. J.; Pompliano, D. L. Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discovery* **2007**, *6*, 29–40.

(34) Pu, L.; Agrafiotis, D. K.; Theobald, D. L. Fast determination of the optimal rotational matrix for weighted superpositions. J. Comput. Chem., in press.

(35) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(36) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An alternative method for the evaluation of docking performance: RSR vs RMSD. *J. Chem. Inf. Model.* **2008**, *48*, 1411–1422.

(37) Martin, E. J.; Hoeffel, T. J. Oriented substituent pharmacophore property space (OSPPREYS): A substituent-based calculation that describes combinatorial library products better than the corresponding product-based calculation. *J. Mol. Graphics Modell.* **2000**, *18*, 383–403.

(38) Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. Conformational analysis of macrocycles: Finding what common search methods miss. J. Chem. Inf. Model., Article ASAP.

CI9001926