

On the Effects of Permuted Input on Conformational Sampling of Drug-like Molecules: an Evaluation of Stochastic Proximity Embedding

Dimitris K. Agrafiotis^{1,*}, Deepak Bandyopadhyay¹, Giorgio Carta², Andrew J. S. Knox² and David G. Lloyd²

¹Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive, Exton, PA 19341, USA

²Molecular Design Group, School of Biochemistry and Immunology, Trinity College Dublin, Dublin 2, Ireland

*Corresponding author: Dimitris K. Agrafiotis, dagrafio@prdu.jnj.com

Conformational sampling is a problem of central importance in computer-aided drug design. A good conformational search method must not exhibit any intrinsic bias, and must provide confidence that important regions of conformational space are not missed during the search. A recent study by Carta *et al.* showed that this is not always the case, and that several popular conformational search methods, such as Omega, are very sensitive to the relative ordering of atoms and bonds in the connection table. Here, we examine the performance of a newer method known as stochastic proximity embedding, or SPE, using five diverse bioactive ligands extracted from the PDB. Our results confirm that the conformational ensembles produced by SPE using different permuted inputs are statistically indistinguishable, and well within the range of variability that would be expected from the stochastic nature of the method itself. This, along with the results of a more comprehensive comparative study (Agrafiotis *et al.*, *J. Chem. Info. Model*, 2007, in press), provides further evidence that SPE is one of the most robust and competitive conformational search methods described to date.

Key words: bioactive conformation, boosting, conformational analysis, conformational sampling, distance geometry, SPE, stochastic proximity embedding

Received 10 May 2007, revised 15 June 2007 and accepted for publication 17 June 2007

The properties and reactivity of organic molecules depend intimately on their three-dimensional shape. Most organic molecules of non-trivial size can assume a multitude of three-dimensional conforma-

tions. Identifying which of these conformations are relatively stable and likely to be populated at room temperature has been the subject of extensive study in the computational chemistry literature (1). This problem is particularly critical in computer-assisted drug design. Recent studies of crystal structures of protein–ligand complexes have shown that bioactive conformations tend to be more extended than random ones (2), and may lie several kcal/mol higher in energy than their respective global minima (3). Several applications depend critically on the diversity of conformations sampled during the search, including protein docking, pharmacophore modeling, 3D database searching, and 3D-QSAR, to name a few. Since the bioactive conformation of a ligand also depends on the geometry of its host, it is imperative that the search for conformational minima casts a wide net over the potential energy surface.

Conformation generation algorithms fall into two broad categories: deterministic, which exhaustively enumerate all possible torsions at certain discrete intervals, and stochastic, which use a random element to explore the molecule's conformational space. Validation studies (4–7) and practical experience suggest that no individual method is decisively better than the others. Systematic search can be very effective for molecules with limited conformational flexibility, but the exponential growth of the search space with the number of rotatable bonds and problems associated with ring closures limit its utility as a general conformational sampling technique (8–11). Stochastic methods, particularly those based on iterative perturbation such as molecular dynamics and Monte Carlo sampling (12–14), have broader applicability, but tend to generate many transitional conformations between local minima and, as a result, they are generally slower.

An alternative approach, known as distance geometry (15,16), is to generate conformations that satisfy a set of interatomic distance constraints derived from the molecular connectivity table and defined in the form of lower and upper distance bounds $\{l_{ij}\}$ and $\{u_{ij}\}$.[†] Distance geometry methods involve four basic steps: (i) generation of interatomic distance bounds, (ii) assignment of a random value to each distance within the respective bounds, (iii) conversion of the resulting distance matrix into a starting set of Cartesian co-ordinates, and (iv) refinement of the co-ordinates through minimization of an error function measuring the violation of the input

[†]In addition to distance constraints, distance geometry methods also employ volume constraints that prevent the signed volume formed by four atoms from exceeding certain limits. These constraints are used to enforce planarity of conjugate systems and correct chirality of stereocenters.

constraints. To ensure that reasonable conformations are generated, the original upper and lower bounds are usually refined using an iterative triangular smoothing procedure. Although this process improves the initial guess, the randomly chosen distances may still be inconsistent with a valid 3-dimensional geometry, necessitating expensive metrization schemes (16,20,21) or higher dimensional embeddings (13) prior to error refinement, or lengthy refinement procedures if random starting co-ordinates are used.

Stochastic proximity embedding (SPE) is a relatively recent technique for finding molecular geometries that satisfy distance constraints in an efficient manner (17–19). The method starts from random initial atomic positions, and gradually refines them by repeatedly selecting a pair of atoms at random, and updating their respective co-ordinates toward satisfying the corresponding distance bounds. This procedure is performed repeatedly until a reasonable conformation is obtained.[‡]

While SPE was originally shown to provide a good sampling of conformational space, it was observed that 'extreme' conformations located near the periphery of conformational space were not as likely to be visited, and in principle some important conformations could be missed. To remedy this problem, a boosting heuristic was introduced that can be used in conjunction with SPE to bias the search toward more extended or more compact geometries (20). The method generates increasingly extended (or compact) conformations through a series of embeddings, each seeded on the result of the previous one. In the first iteration, a normal SPE embedding is performed, 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 $\{u_{ij}\}$ to perform a second embedding to generate another conformation, c_2 . This process is repeated for a prescribed number of iterations, yielding additional conformations. 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. This process will never yield a set of distance constraints that are impossible to satisfy, because there exist at least one conformation (i.e. the one generated in the preceding iteration) that satisfies them. An analogous procedure can be used to generate increasingly compact conformations. The method was validated against seven widely used conformational sampling techniques, and was found to be significantly more effective in sampling the full range of geometric sizes attainable by any given molecule compared to the other methods (21).

However, it is well known that many stochastic 3D modeling techniques are very sensitive to starting configurations and random

[‡]The actual algorithm involves stochastic refinement of all geometric constraints, both distance and volume. The co-ordinate adjustments differ for each type of constraint, for example, a volume constraint is minimized by moving the positions of the four atoms associated with that constraint in the direction opposite to the gradient of an error function that measures the deviation from the target value. The algorithm also uses a number of parameters such as learning rates and sampling frequencies to control the self-organization process. We refer the reader to ref. 19 for a complete description, validation, and performance benchmark of the algorithm, as well as a comparison to other distance geometry methods.

number effects, and the results are often difficult to reproduce when the search is repeated under slightly different initial conditions. Carta *et al.* (22) demonstrated that this reproducibility problem plagues systematic methods as well. More specifically, they examined how different permutations of the connection table affected the conformations generated by Corina^a, Omega^b, Catalyst^c, and Rubicon.^d The authors used Daylight and in-house utilities to generate different (non-canonical) variants of SMILES and SDF representations for 17 bioactive ligands, effectively changing the order of the atoms and bonds while keeping the topology intact. Each variant was subjected to conformational search, using the same set of parameters for conformer generation. The results were evaluated by comparing the distributions of several 3D descriptors in the conformational ensembles produced by each permuted input. While Catalyst showed negligible sensitivity, Omega and Rubicon produced substantially different ensembles, suggesting that these methods have an intrinsic bias that is highly dependent on the atom and bond ordering. The work clearly indicated the benefit of including multiple input permutations in improving the sampling of conformational space by established methodologies.

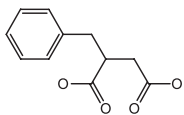
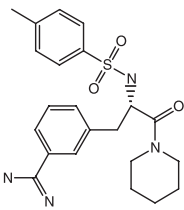
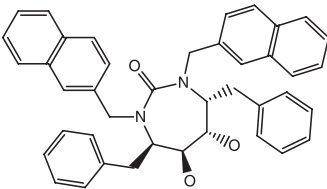
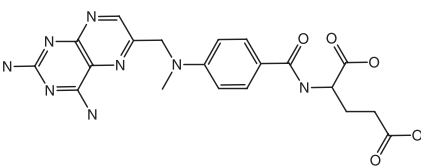
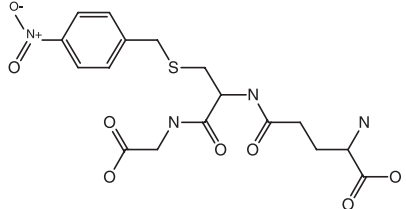
Here, we examine the sensitivity of SPE to permutation of input structure representation using a diverse subset of five bioactive ligands from our reference collection. Our results confirm that the conformational ensembles produced by different permuted inputs are statistically indistinguishable, and within the range of variability that would be expected from the stochastic nature of the method.

Methods

The comparison was based on five bioactive ligands extracted from the Protein Data Bank (23,24), containing between 5 and 13 rotatable bonds. Their structures, PDB codes, and number of rotatable bonds are listed in Table 1, along with the number of SPE trials and the number of conformational boosts employed in each trial (vide infra). Each ligand was neutralized and the proper chirality was assigned to each chiral center, as inferred from the PDB 3D co-ordinates. For 1CBX, 4DFR, and 1GLQ, these chiral flags were subsequently cleared to study the performance of SPE in the presence and absence of stereochemistry. For 1ETT and 1HVR, the chiral flags remained as shown in Table 1.

For each ligand, five different SMILES and five different SDF variants were generated, using the Daylight utility `permsmi` (25) for the former and an internally developed program for the latter. Each variant was used as input to SPE, which was instructed to produce between 1000 and 10 000 conformations for each one, depending on the size of the molecule (1000 conformations for each 1CBX variant, 5000 for each 1ETT and 1HVR variant, and 10 000 for each 4DFR and 1GLQ variant). The SPE conformer generator reads both SMILES and SDF files, and does not necessitate explicit file conversion. The input structures are read into an internal C++ data structure, which preserves the order in which the atoms and bonds were specified in the input stream (i.e. no canonicalization takes place). Thus, we are certain by virtue of the

Table 1: PDB codes, structures, number of rotatable bonds, number of SPE trials, and number of boosts for each of the molecules in the data set

PDB code	Structure	Rotatable bonds	Trials	Boosts ^a
1CBX		5	1000	0
1ETT		7	5000	1
1HVR		8	5000	1
4DFR		9	10 000	1
1GLQ		13	10 000	1

^aNumber of boosts per trial. For example, the 5000 conformations for 1ETT were obtained by generating 2500 conformations using standard distance bounds, and boosting each of these conformations once to obtain 2500 additional conformations.

architecture of our program that each variant resulted in a unique internal atom and bond numbering order as well.

Given the size and flexibility of the molecules in our data set, we used one level of boosting for 1ETT, 1HVR, 4DFR, and 1GLQ, and no boosting for 1CBX (Table 1). For example, the 10 000 conformations for 4DFR were obtained by generating 5000 random conformations using the standard distance bounds derived from the connection table, and boosting each of these conformations once to obtain 5000 additional conformations (10 000 = 2 × 5000). As SPE proper favors compact geometries, boosting was applied to the lower bounds only, leading to more extended conformations.

All SPE embeddings were carried out on hydrogen-depleted graphs using the following parameters: $\lambda_d = \lambda_v = 1$, $C = 50$, $\delta\lambda_d = \delta\lambda_v = 0.9/(C - 1)$, $S = 50 \times N$, and

$$v = \min\left(0.5, 8.0 \times \frac{N_V}{N(N+1)/2 + 8N_V}\right),$$

where λ_d and λ_v are the initial distance and volume learning rates, C is the number of cycles, S is the number of steps, $\delta\lambda_d$ and $\delta\lambda_v$ are the decrements of the distance and volume learning rates in each cycle, v is the ratio of volume to distance refinements, N is the number of atoms in the molecule, and N_V is the total number of volume constraints [for a detailed description of these parameters, see Ref. (19)]. Each conformation was derived by running the algorithm with a different random number seed.

All the raw conformations generated by SPE were minimized using the MMFF94s force field (26–30) and the BFGS variable metric

minimization algorithm, as implemented in the DIRECTEDDIVERSITY (31) software. This software has been tested thoroughly and has passed the entire MMFF94s validation suite.⁹ For each permuted input, duplicate conformations were identified and eliminated in a post-processing step using an RMSD threshold of 1.0 Å. More specifically, each new minimized conformation c was compared to all previously unique conformations, and if the RMSD to all of them was >1.0 Å, c was added to the unique set. If, on the other hand, there was an existing conformation c^* in the unique set such that $\text{RMSD}(c, c^*) \leq 1.0$ Å, the one with the lowest energy between the two was retained, and the other was discarded as a duplicate. The RMSD was computed by taking topological symmetry into account, i.e. all possible mappings of topologically equivalent atoms were considered, and the alignment with the lowest RMSD was used to measure the similarity of two conformations. From the resulting non-duplicate lists, the lowest energy conformation for each variant was identified, and all conformations with energies >20 kcal/mol from that conformation were discarded.

All calculations and postprocessing were performed within the DirectedDiversity (31) and Third Dimension Explorer (32) environments, except ANOVA, which was carried out in Microsoft Excel. All calculations were carried out on an IBM Intellistation running Windows XP Professional, and equipped with two 3.2 GHz Xeon processors and 2048 Mb of RAM.

Results and Discussion

Following the approach of Carta *et al.* (22) we analyzed the conformational ensembles produced by the different permuted inputs through both statistical and visual means. Being a stochastic method, SPE is not guaranteed to produce the same results when run under different initial conditions. Different runs, starting from different random number seeds, are expected to produce slightly different sets of conformations, but these should converge as sampling becomes more exhaustive and the search space becomes saturated. This is illustrated in Figure 1, which shows the number of unique conformations discovered at the end of each trial for each SMILES (black) and SDF (red) permuted input for each of the five molecules. When contrasted with multiple independent random runs of only one of the permuted inputs (Figure 1E for 4DFR), it becomes clear that the input representation does not have any discernible impact, and that the variation in the sampling profiles is within the range expected by the stochastic nature of the method.

The lowest energy and the number of unique conformations identified after the prescribed number of trials for each molecule and each permuted input are summarized in Table 2. These numbers represent the conformations within 20 kcal/mol from the lowest energy structure identified by each permuted input. For 1CBX, which is a small molecule with few conformational states, every run identified the same global minimum ($E = -50.58$ kcal/mol). However, there were some differences in the number of local minima identified by each variant, which ranged from 21 to 25, with no apparent dependence on the input format. As the size of the molecule and the number of accessible conformations increa-

ses, differences between variants become more likely, and this is reflected in both the lowest energy minimum and the number of local minima identified by each method (170–189 for 1ETT, 340–353 for 1HVR, 411–433 for 4DFR, and 3103–3543 for 1GLQ). Again, there is no apparent interdependence between the input representation (SMILES versus SDF) and the number of conformers generated.

The ensembles produced by each permuted input were further compared using three different conformational properties: (i) the potential energy computed by MMFF94s; (ii) the root mean square deviation (RMSD) from the X-ray crystal structure; and (iii) the radius of gyration, which measures a conformation's degree of compactness (or extendedness). The cumulative distributions of the RMSD from the crystal structure of the low-energy conformations identified for each molecule and each permuted input are illustrated in Figure 2. In all cases, the conformational ensembles seem to be drawn from the same probability distribution and the variation appears to be random and not dependent on the input representation.⁵ To obtain a more robust statistical estimate, we compared these distributions using single-factor ANOVA. ANOVA is a well-established technique that attempts to test the null hypothesis that two or more mean values are equal, by comparing the variation observed within each class against the one observed across the entire sample. The key parameters of an ANOVA study include the F -statistic which is compared to the critical value of the F distribution, F_{crit} , and the p -value which represents the smallest level of significance for which the observed sample information becomes significant, provided the null hypothesis is true. The results, which are summarized in Table 3, show that the differences in the conformational property distributions of the permuted inputs are statistically insignificant, except for the energy distribution of the conformations of 1GLQ.[†] This result is not surprising in light of the fact that the conformational space of that molecule is grossly undersampled. As seen in Figure 1, the rate at which new conformations are discovered decreases exponentially with the number of trials, as the probability of visiting a previously seen conformation increases. While for 1CBX, 1ETT, 1HVR, and 4DFR the search appears to have converged, for 1GLQ new minima continue to be discovered at a fast rate even after the 10 000 trials. These plots also show that the permuted inputs explore the same range of conformational space throughout the sampling process, i.e. the number of unique conformations identified by each variant is comparable after any number of trials.

These results contrast with the original findings of Carta *et al.*, who showed that the low-energy conformations generated by Omega and Rubicon for each permuted SMILES string exhibited a

⁵Similar distributions were observed for the potential energy and radius of gyration as well. The results are not shown here to conserve space.

[†]Strictly speaking, ANOVA tests the equality of mean values and not the general similarity between two distributions, for which other statistics such as Kolmogorov–Smirnov's may be more suitable. Although one could imagine different distributions with the same mean but very different shapes, visual inspection of the plots in Figure 2 (and those of the other conformational properties which are not included here) show that their shapes are virtually identical.

Effects of Permuted Input on Conformational Sampling of Drug-like Molecules

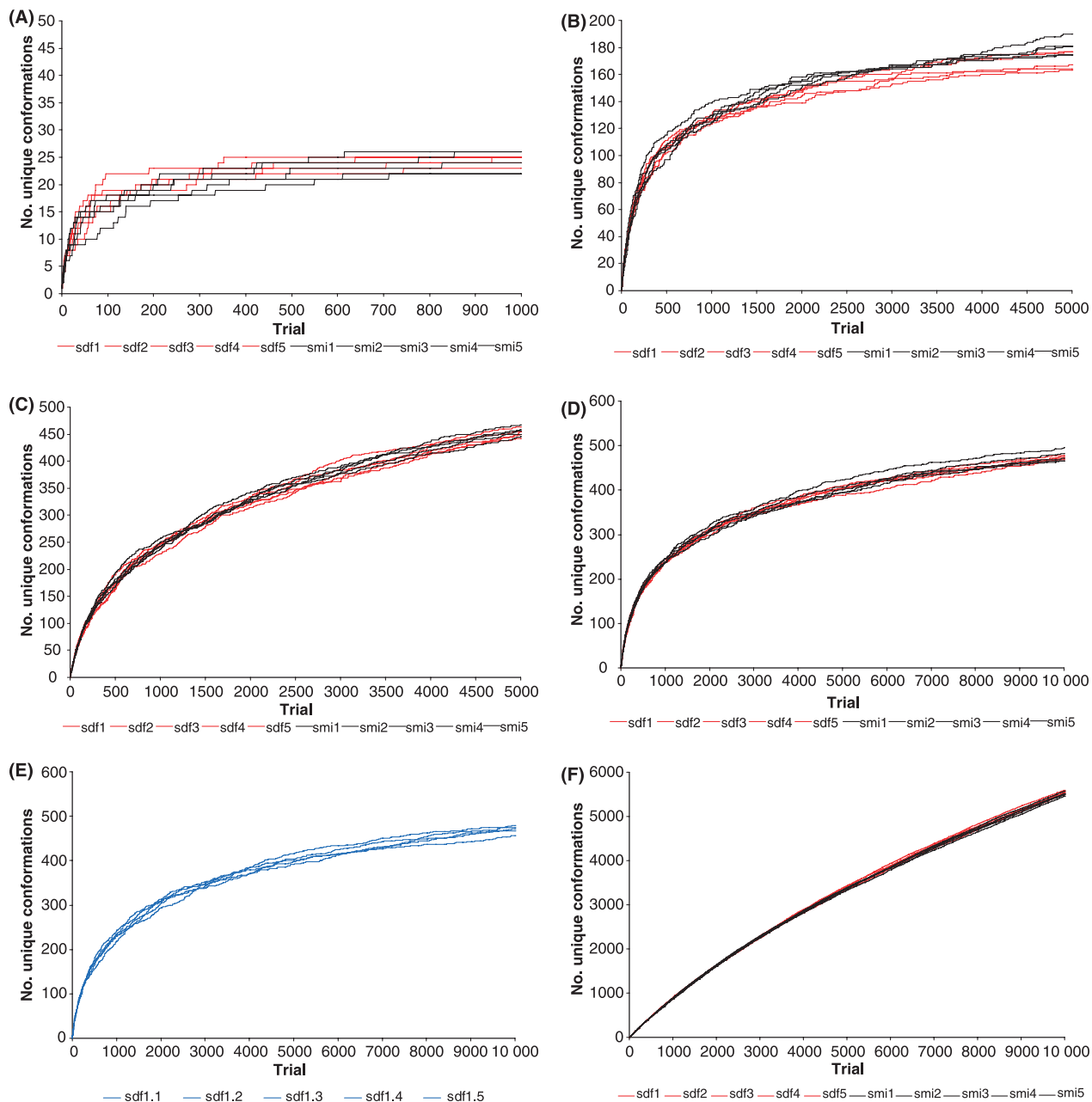


Figure 1: Number of unique conformations (within 1 Å RMSD) discovered at the end of each SPE trial for the SDF (red) and SMILES (black) permuted inputs of each molecule under investigation. (A) 1CBX; (B) 1ETT; (C) 1HVR; (D) 4DFR; and (F) 1GLQ. (E) Five independent runs of the first SDF permuted input of 4DFR, each starting from a different random seed.

different distribution of RMSD from the X-ray co-crystal conformation (22). Catalyst, on the other hand, was found to be insensitive to permuted input, yielding identical distributions for all SMILES variants. This is consistent with our recent comparison of SPE to seven other widely used conformational analysis programs, which showed that Catalyst's sampling capacity was comparable to SPE's, and significantly higher than all the other methods evaluated (21).

We should point out, parenthetically, that for the two chiral molecules (1ETT and 1HVR), not all trials resulted in valid geometries. SPE uses volume constraints to generate conformations that preserve the chirality of all explicitly defined stereocenters in the input file. As we noted in our original publication (19), the probability of failing to generate a correct 3D geometry increases with the number and proximity of chiral centers. As shown in Table 4, this has not been a serious issue for 1ETT and 1HVR. On average,

Table 2: Global minimum and number of unique conformations within 20 kcal/mol from the 'global' minimum identified for each permuted input

PDB code	Variant	E_{\min}	Number of conformations
1CBX	SDF 1	-50.58	23
1CBX	SDF 2	-50.58	22
1CBX	SDF 3	-50.58	22
1CBX	SDF 4	-50.58	25
1CBX	SDF 5	-50.58	22
1CBX	SMI 1	-50.58	22
1CBX	SMI 2	-50.58	22
1CBX	SMI 3	-50.58	23
1CBX	SMI 4	-50.58	21
1CBX	SMI 5	-50.58	22
1ETT	SDF 1	-34.37	173
1ETT	SDF 2	-36.05	182
1ETT	SDF 3	-35.60	178
1ETT	SDF 4	-34.46	179
1ETT	SDF 5	-33.98	170
1ETT	SMI 1	-33.90	174
1ETT	SMI 2	-36.05	189
1ETT	SMI 3	-35.55	181
1ETT	SMI 4	-33.95	182
1ETT	SMI 5	-34.43	189
1HVR	SDF 1	102.37	341
1HVR	SDF 2	102.32	346
1HVR	SDF 3	102.32	344
1HVR	SDF 4	102.32	347
1HVR	SDF 5	102.32	351
1HVR	SMI 1	102.32	348
1HVR	SMI 2	102.32	352
1HVR	SMI 3	102.32	340
1HVR	SMI 4	102.32	342
1HVR	SMI 5	102.37	353
4DFR	SDF 1	-30.02	416
4DFR	SDF 2	-30.00	422
4DFR	SDF 3	-28.80	420
4DFR	SDF 4	-28.19	421
4DFR	SDF 5	-27.55	423
4DFR	SMI 1	-30.00	414
4DFR	SMI 2	-30.00	420
4DFR	SMI 3	-29.38	433
4DFR	SMI 4	-30.00	418
4DFR	SMI 5	-31.16	425
4DFR	SDF 1.1	-29.36	427
4DFR	SDF 1.2	-28.34	411
4DFR	SDF 1.3	-30.02	419
4DFR	SDF 1.4	-31.20	416
4DFR	SDF 1.5	-30.00	412
1GLQ	SDF 1	59.70	3543
1GLQ	SDF 2	58.62	3225
1GLQ	SDF 3	58.40	3103
1GLQ	SDF 4	58.40	3226
1GLQ	SDF 5	58.62	3197
1GLQ	SMI 1	58.62	3197
1GLQ	SMI 2	58.40	3151
1GLQ	SMI 3	58.40	3154
1GLQ	SMI 4	59.25	3421
1GLQ	SMI 5	59.25	3418

only 17 of 5000 trials failed to produce the correct stereochemistry for 1ETT, and only 38 of 5000 for 1HVR (which contains four adjacent chiral centers), and in both cases that failure rate was insensitive to permuted input. Note that energy minimization cannot correct such problems because of the enormous energy barriers involved in inverting a stereocenter. Because SPE is extremely fast (it can generate hundreds of conformations per second on a modern PC for a typical drug-like molecule), the raw conformations produced by SPE can be tested for the correct stereochemistry in a postprocessing step, and discarded if they fail to reproduce the input specification.

To provide a more visual assessment of the overlap in the conformational space covered by each permuted input, the unique low-energy conformations produced by the 10 variants of 4DFR were concatenated into a single list and were embedded into a two-dimensional nonlinear map using SPE. As mentioned in the Introduction, SPE was originally devised as a method to generate low-dimensional Euclidean embeddings that best preserve the similarities between a set of related observations (17,18). Unlike classical multidimensional scaling and nonlinear mapping, SPE scales linearly with respect to sample size, and can be applied to very large data sets that are intractable by conventional embedding procedures. In the case at hand, the map was constructed in such a way that the distances of the points on the map approximated as closely as possible the (dis)similarities of the respective conformations, as measured by their pairwise RMSD. As 4DFR is a flexible molecule with many conformational degrees of freedom, this projection leads to some loss of information, in the sense that the distances cannot all be preserved to perfection (the intrinsic dimensionality of the conformational space of 4DFR is roughly equivalent to the number of rotatable bonds, which is considerably >2). That loss of information is measured by the stress function, which in this particular case was 0.262.

Embedding all the conformations produced by all the 4DFR variants (4221) into a single map allows them to be placed on a common reference frame. The 10 panels in Figure 3 show the conformations generated by each permuted input on that common reference frame, color-coded by energy (from blue to red, in order of decreasing energy). Unlike other conformational search methods (22), the ensembles generated by SPE for each permuted input cover the same regions of conformational space and with the same relative density. Minor differences in the maps are attributed to two factors related to the stochastic nature of the method. First, if each unique conformation is seen as a representative of a cluster of related conformations within an RMSD radius of 1 Å, different ensembles may contain different representatives from each cluster, which are closely related but not identical. Secondly, because of the 20 kcal/mol energy cap, an ensemble that contains a 'global' minimum of lower energy than the other ensembles will be missing conformations on the high end of the energy spectrum. For example, the 'global' minimum identified by the fifth SMILES variant of 4DFR was -31.16 kcal/mol, whereas that identified by the first SDF variant was only -27.55 kcal/mol (Table 2). As seen in Figure 4, there

Effects of Permuted Input on Conformational Sampling of Drug-like Molecules

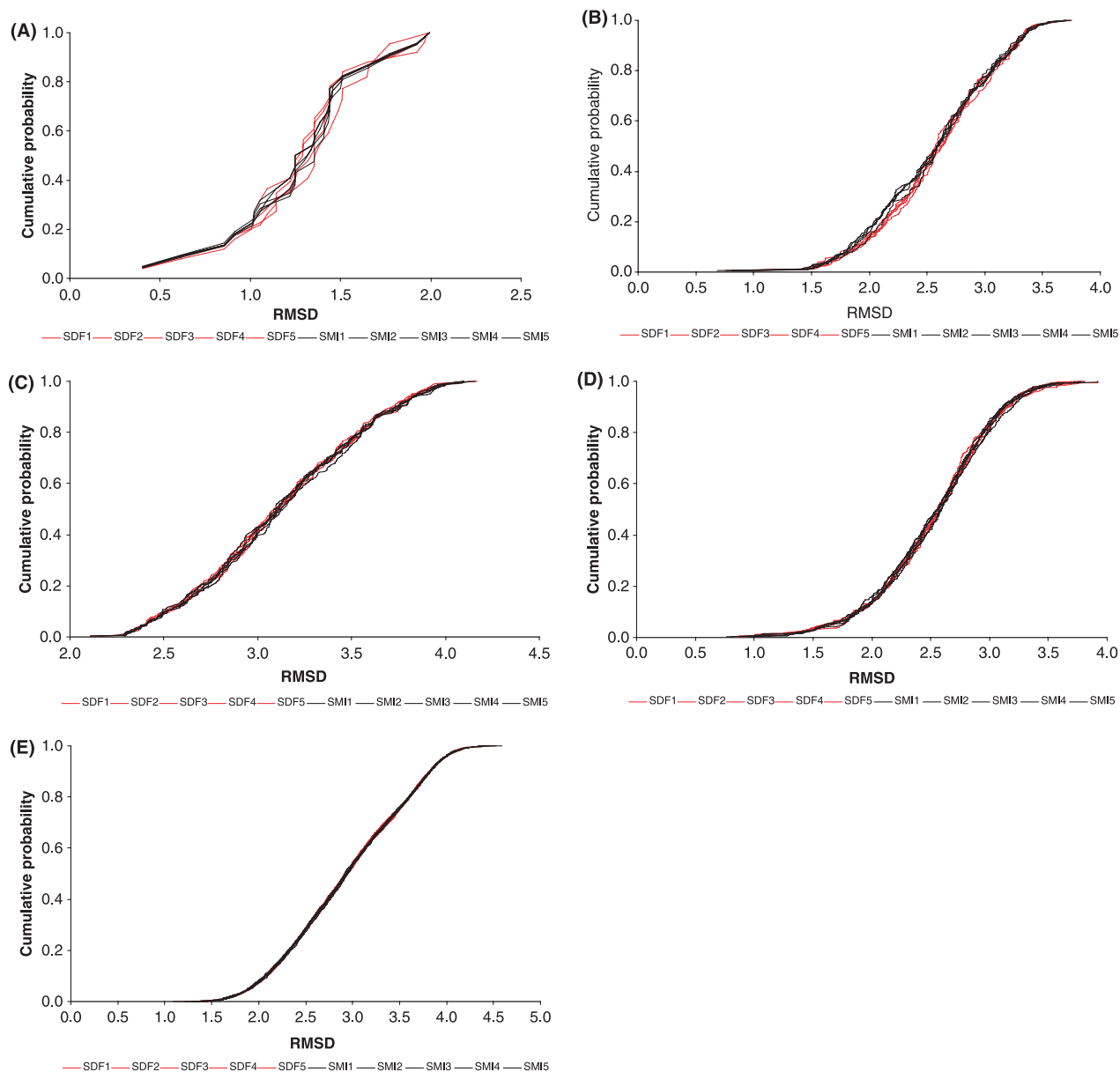


Figure 2: Cumulative distribution of RMSDs from the bioactive conformation of the conformational ensembles generated from the SDF (red) and SMILES (black) permuted inputs of each molecule under investigation. (A) 1CBX; (B) 1ETT; (C) 1HVR; (D) 4DFR; and (E) 1GLQ.

Table 3: ANOVA of the conformational properties of the ensembles produced by the permuted inputs for each molecule under investigation (when F is less than F_{crit} , the differences are considered to be statistically insignificant)

PDB code	Energy			RMSD			Radius of gyration		
	F	F_{crit}	p-Value	F	F_{crit}	p-Value	F	F_{crit}	p-Value
1CBX	0.269	1.924	0.982	0.027	1.924	0.999	0.073	1.924	0.999
1ETT	0.124	1.885	0.999	0.214	1.885	0.992	0.189	1.885	0.995
1HVR	0.281	1.883	0.980	0.189	1.883	0.995	0.283	1.883	0.980
4DFR	0.834	1.693	0.632	0.316	1.693	0.992	0.185	1.693	0.999
1GLQ	13.337	1.880	1.45E-21	0.346	1.880	0.960	0.499	1.880	0.876

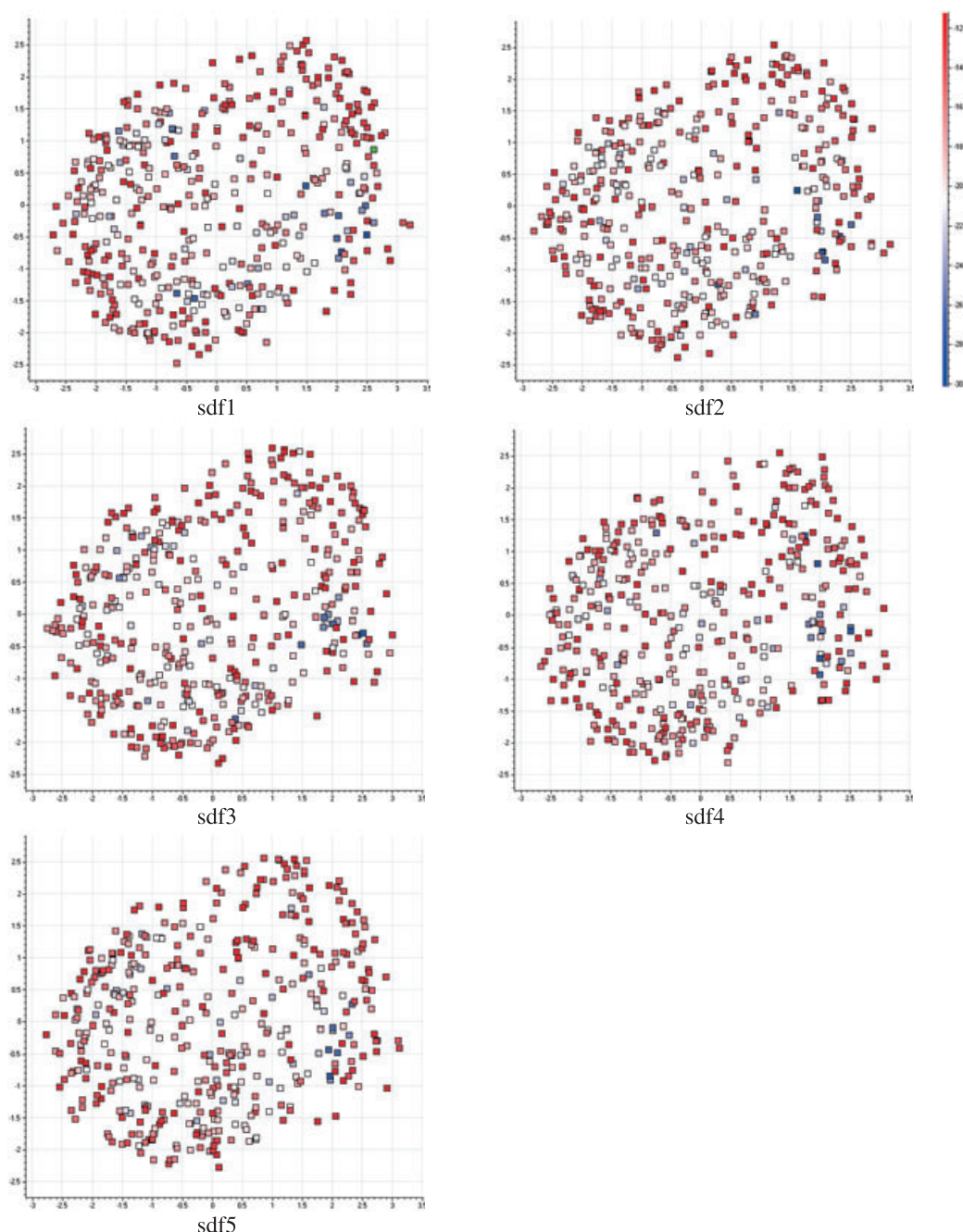


Figure 3: Two-dimensional SPE maps of the conformational ensembles generated from each SDF and SMILES permuted input of 4DFR, color-coded by their MMFF94s potential energy. These maps are constructed in such a way that the distances of the points in the map match as closely as possible the RMSDs of the respective conformations.

are considerably more conformations on the high end of the energy scale, which explains some of the observed differences in the maps^{**}.

These results are not at all surprising, if one considers the way in which the SPE conformer generator works. The process starts by

^{**}To conserve space, the SPE maps for the other four molecules are not included, but they do exhibit similar qualitative trends.

establishing a set of distance and volume constraints from the molecular connectivity table. This is accomplished using a set of rules encoded in the form SMARTS patterns. These rules define groups of atoms that form a local substructure, along with the geometric constraints imposed by the local geometry of that substructure. The rules are supplied to the program in a separate file and can easily be extended and customized by the users. When a new molecule is read in, the patterns in the rule base are mapped onto the target molecule in sequence, and the corresponding constraints

Effects of Permuted Input on Conformational Sampling of Drug-like Molecules

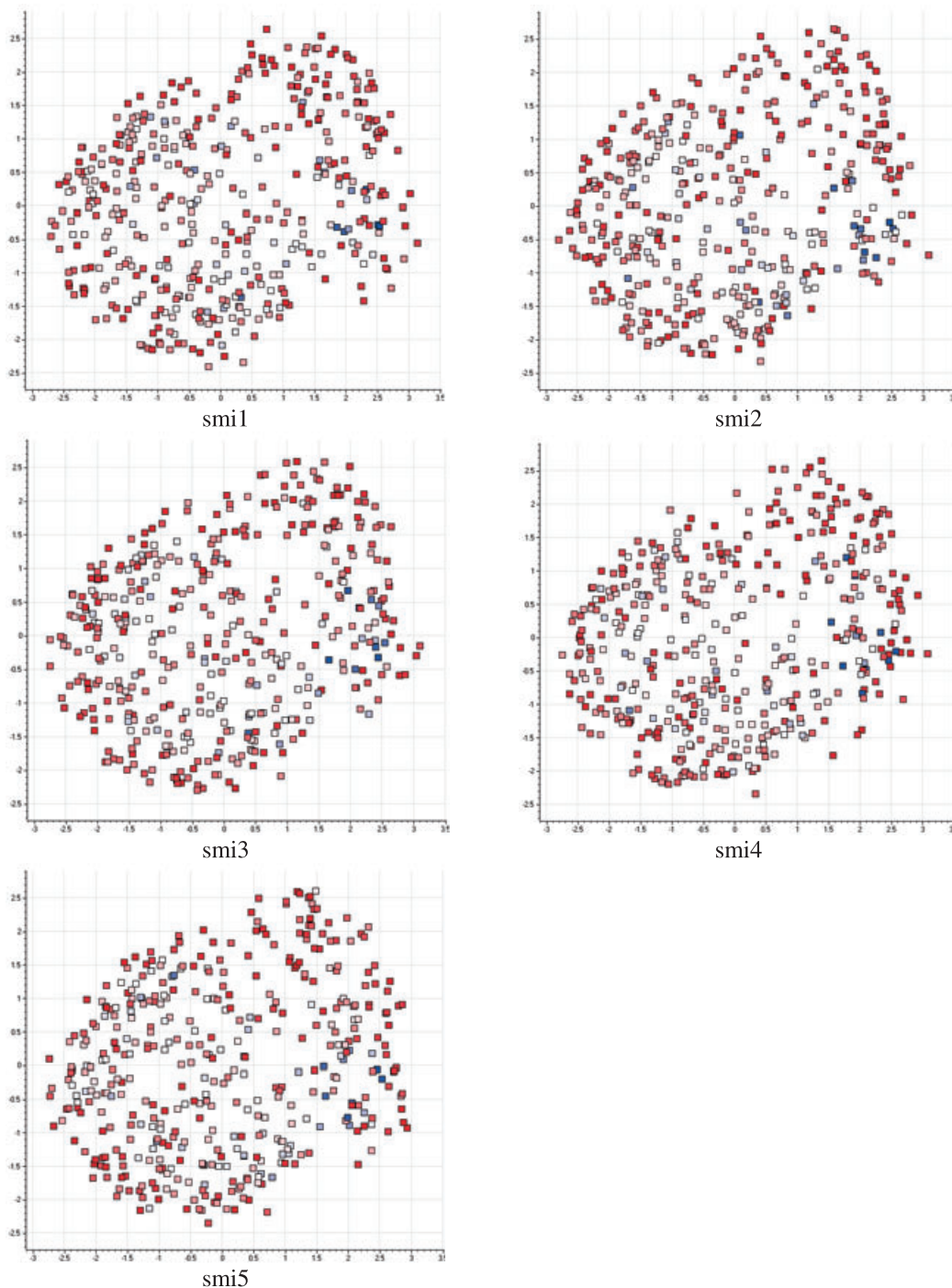


Figure 3: (Continued)

are recorded in the distance and volume constraint lists. The rules are applied in sequence, with those appearing later on in the sequence taking precedence over earlier ones, if they are more restrictive. For example, a generic rule may define some broad minimum and maximum bond lengths for a carbon–carbon bond, but these limits may be tightened up by subsequent rules for specific types of carbon–carbon bonds, such as single, double, triple, or aromatic. Additional rules may tighten these constraints even further for more restrictive environments, such as strained rings or other

unusual connectivity patterns. The following examples highlight five different types of rules. Each rule consists of four elements: (i) the type of the constraint; (ii) the SMARTS pattern; (iii) the zero-based indices of the atoms in the SMARTS pattern to which the constraint is to be applied; and (iv) the minimum and maximum value for the constraint.

- Set Van der Waals radius of C to 1.65 Å
radius {[#6]} {0} {1.65 1.65}

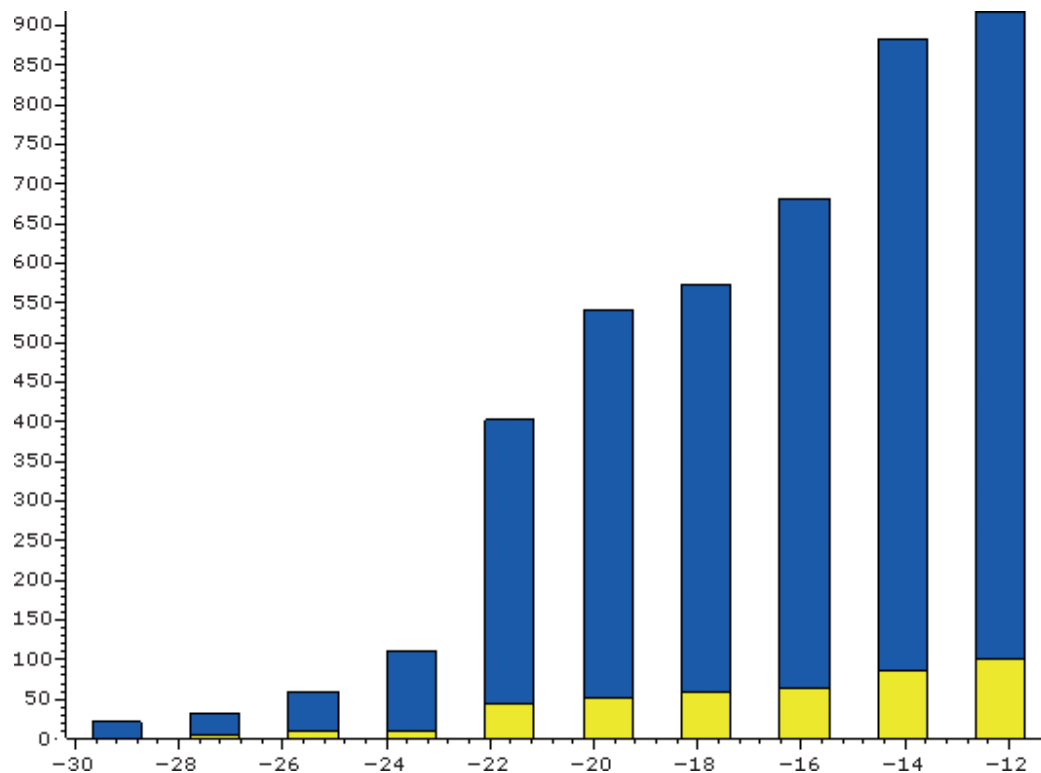


Figure 4: Histogram of the MMFF94s potential energies of the conformations generated by all permuted inputs of 4DFR (blue). The conformations generated by the first SDF permuted input are highlighted in yellow.

Table 4: Number of conformations with correct stereochemistry identified after 5000 SPE trials for each permuted input of 1ETT and 1HVR

Variant	1ETT	1HVR
SDF 1	4975	4951
SDF 2	4980	4961
SDF 3	4988	4964
SDF 4	4988	4966
SDF 5	4983	4958
SMI 1	4983	4949
SMI 2	4986	4970
SMI 3	4984	4968
SMI 4	4985	4963
SMI 5	4975	4970

- Set bond length of C-N triple bond to 1.15 Å
distance {N#C} {0 1} {1.15 1.15}
- Set bond angle of 6-membered aromatic ring to 120°
angle {a1aaaaa1} {0 1 2} {120.0 120.0}
- Set minimum sp³-sp³ torsion to 60°
torsion {*~[sp3]-[sp3]~*} {0 1 2 3} {60.0 180.0}
- Set all aromatic rings and substituents to be in-plane.
volume {*~*(.*)~*} {0 1 2 3} {0.0 0.0}

Since the mapping of substructures is unambiguous and independent of the order of the atoms and bonds in the connection table, and since the same set of rules are applied in the same sequence, different permuted inputs will produce exactly the same set of geometric constraints (and consequently lead to statistically equivalent sampling of the molecule's conformational space, which is exactly what is observed in our experiments).

Conclusions

Carta *et al.* demonstrated that some systematic conformational search methods, such as Omega, can be seriously impacted by minor rearrangements of the connection table, and recommended the use of multiple permuted inputs to improve their performance and enhance their sampling capacity. While this approach only requires a way to generate permuted connection tables, and therefore can be used with any conformational search program to circumvent its intrinsic bias, it remains symptomatic. It does not guarantee that the additional permutations will lead to an exhaustive sampling of conformational space, and that important regions of conformational space will not be missed. The preceding analysis demonstrates that stochastic proximity embedding does not exhibit such bias, and that its sampling capacity is limited solely by the geometric constraints encoded in the rule base. Coupled with our recent extensive comparative study (21), this provides further evidence that SPE is one of the most competitive conformational sampling methods invented to date.

References

- Leach A.R. (1991) A survey of methods for searching the conformational space of small and medium-sized molecules. In: Lipkowitz K.B., Boyd D.B., editors. *Reviews in Computational Chemistry*, Vol. 2. New York: VCH; p. 1–55.
- Diller D.J., Merz K.M. Jr. (2002) Can we separate active from inactive conformations? *J Med Chem*;16:105–112.
- Kirchmair J., Laggner C., Wolber G., Langer T. (2005) Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms. *J Chem Inf Model*;45:422–430.
- Bostrom J., Greenwood J.R., Gottfries J. (2003) Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model*;21:449–462.
- Bostrom J. (2001) Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. *J Comput.-Aided Mol Des*;15:1137–1152.
- Feuston B.P., Miller M.D., Culberson J.C., Nachbar R.B., Kearsley S.K. (2001) Comparison of knowledge-based and distance geometry approaches for generation of molecular conformations. *J Chem Inf Comput Sci*;41:754–763.
- Perola E., Charifson P.S. (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem*;47:2499–2510.
- Lipton M., Still W.C. (1988) The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space. *J Comput Chem*;9:343–355.
- Brucoleri R.E., Karplus M. (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*;26:137–168.
- Brucoleri R.E., Karplus M. (1985) Chain closure with bond angle variations. *Macromolecules*;18:2767–2773.
- Go N., Scheraga H.A. (1970) Ring closure and local conformational deformations of chain molecules. *Macromolecules*;3:178–187.
- Saunders M. (1987) Stochastic explorations of molecular mechanics energy surfaces. Hunting for the global minimum. *J Am Chem Soc*;109:3150–3152.
- Ferguson D.M., Raber D.J. (1989) A new approach to probing conformational space with molecular mechanics: random incremental pulse search. *J Am Chem Soc*;111:4371–4378.
- Chang G., Guida W.C., Still W.C. (1989) An internal coordinate Monte Carlo method for searching conformational space. *J Am Chem Soc*;111:4379–4386.
- Crippen G.M. (1978) Rapid calculation of coordinates from distance matrices. *J Comput Phys*;26:449–452.
- Spellmeyer D.C., Wong A.K., Bower M.J., Blaney J.M. (1997) Conformational analysis using distance geometry methods. *J Mol Graphics Model*;15:18–36.
- Agrafiotis D.K., Xu H. (2002) A self-organizing principle for learning non-linear manifolds. *Proc Natl Acad Sci USA*;99:15869–15872.
- Agrafiotis D.K. (2003) Stochastic proximity embedding. *J Comput Chem*;24:1215–1221.
- Xu H., Izrailev S., Agrafiotis D.K. (2003) Conformational sampling by self-organization. *J Chem Info Comput Sci*;43:1186–1191.
- Izrailev S., Zhu F., Agrafiotis D.K. (2006) A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *J Comput Chem*;27:1962–1969.
- Agrafiotis D.K., Gibbs A., Zhu F., Izrailev S., Martin E. (2007) Conformational sampling of bioactive molecules: a comparative study. *J Chem Info Model*;47:1067–1086.
- Carta G., Onnis V., Knox A.J. S., Fayne D., Lloyd D.G. (2006) Permuting input for more effective sampling of 3D conformer space. *J Comput Aided Mol Des*;20:179–190.
- Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*;28:235–242.
- Berman H.M., Bhat T.N., Bourne P.E., Feng Z., Gilliland G., Weissig H., Westbrook J. (2000) The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol*;7(Suppl):957–959.
- Daylight Chemical Informations Systems Inc. (<http://www.daylight.com>).
- Halgren T.A. (1996) Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J Comput Chem*;17:490–519.
- Halgren T.A. (1996) Merck molecular force field. II. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. *J Comput Chem*;17:520–552.
- Halgren T.A. (1996) Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J Comput Chem*;17:553–586.
- Halgren T.A., Nachbar R.B. (1996) Merck molecular force field. IV. conformational energies and geometries for MMFF94. *J Comput Chem*;17:587–615.
- Halgren T.A. (1996) Merck molecular force field. V. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. *J Comput Chem*;17:616–641.
- Agrafiotis D.K., Lobanov V.S., Salemme F.R. (2002) Combinatorial informatics in the post-genomics era. *Nat Rev Drug Discov*;1:337–346.
- Agrafiotis D.K., Alex S., Dai H., Derkinderen A., Farnum M., Gates P., Izrailev S. *et al.* Advanced biological and chemical discovery (ABCD): centralizing discovery knowledge in an inherently decentralized world. Submitted.

Notes

- CORINA 3.6, distributed by Molecular Networks GmbH (<http://www.mol-net.de>).
- OMEGA 1.8.1, distributed by Openeye Scientific Software (<http://www.eyesopen.com>).
- Catalyst v4.9.1, distributed by Accelrys, Inc. (<http://www.accelrys.com>).
- RUBICON, distributed by Daylight Chemical Information Systems Inc. (<http://www.daylight.com>).
- <http://www.ccl.net/cca/data/MMFF94s/>