Conformational Sampling of Bioactive Molecules: A Comparative Study

Dimitris K. Agrafiotis,*† Alan C. Gibbs,† Fangqiang Zhu,† Sergei Izrailev,†,§ and Eric Martin‡

Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 665 Stockton Drive, Exton, Pennsylvania 19341, and Novartis, 4560 Horton Street, Emeryville, California 94608

Received December 4, 2006

The necessity to generate conformations that sample the entire conformational space accessible to a given molecule is ubiquitous in the field of computer-aided drug design. Protein−ligand docking, 3D database searching, and 3D QSAR are three commonly used techniques that depend critically upon the quality and diversity of the generated conformers. Although there are a wide range of conformational search algorithms available, the extent to which they sample conformational space is often unclear. To address this question, we conducted a robust comparison of the search algorithms implemented in several widely used molecular modeling packages, including Catalyst, Macromodel, Omega, MOE, and Rubicon as well as our own method, stochastic proximity embedding (SPE). We found that SPE used in conjunction with conformational boosting, a heuristic for biasing conformational search toward more extended or compact geometries, along with Catalyst, are significantly more effective in sampling the full range of conformational space compared to the other methods, which show distinct preferences for either more extended or more compact geometries.

INTRODUCTION

Most organic molecules of nontrivial size are not just three-dimensional, they are "four-dimensional", because they exist as an ensemble of three-dimensional conformations interchanging over time (or equivalently, consist of a distribution of conformations at any time). Their properties and reactivities depend intimately on this ensemble. Knowing the "structure" of a molecule requires knowing all the structures in this ensemble. Conversely, an incomplete ensemble of conformations amounts to an incomplete molecular structure. Identifying all of the conformations which are relatively stable and likely to be populated at room temperature has been the subject of countless studies 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] 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. 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.

Completeness is important in conformational sampling because our knowledge of pharmacologically relevant conformational space is very limited. Reproducing known ligand geometries is insufficient because these represent an extremely limited and biased sampling of all bound ligand conformations. Most ligands have never been crystallized in their own targets, and even fewer have been crystallized in important countertargets. In fact, many drugs target protein classes that have never been crystallized. Moreover, conformational sampling impacts many aspects of drug design modeling outside of ligand binding, such as the ensembles of solution phase conformations, the statistical mechanics of partitioning, membrane diffusion, nonspecific binding to serum albumin, metabolism, chiral induction, and other pharmaceutically relevant physical and chemical processes.

While diversity is sometimes a goal in its own right, as in many approaches to library design, thoroughness of conformational sampling is usually not an end in itself, nor is it the sovereign virtue for a conformational search method. For example, Omega is extremely fast, with sampling suitable for many applications. However, thorough sampling is an important means to many further ends, and any practicing computational chemists would want to know which methods sample the full ensemble of accessible conformations.

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. 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 limits its utility as a general conformational sampling technique.[4−7] For flexible molecules, stochastic methods, such as molecular dynamics and Monte Carlo sampling, represent a viable alternative.[8−10] 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

* Corresponding author phone: (610)458-6045; fax: (610)458-8249; e-mail: dagrafio@prdus.jnj.com.

† Johnson & Johnson Pharmaceutical Research & Development, L.L.C.

‡ Novartis.

§ Current address: Tribeca Global Management, 731 Lexington Ave., New York, NY 10022.

10.1021/ci6005454 CCC: $37.00
Published on Web 04/06/2007

local minima, and a considerable amount of computer time is spent calculating and minimizing the potential energies for these transitional geometries. An alternative approach, known as distance geometry,[11,12] is to generate conformations that satisfy a set of geometric constraints derived from the molecular connectivity table. 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 conjugate 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 minimized with any suitable force field to identify the corresponding energy minima. In a recent study, distance geometry was shown to identify conformations that were missed by alternative systematic search methods.[13] Many studies comparing various implementations of several of these approaches have been published.[14−17]

Stochastic proximity embedding (or SPE) is a recent technique for finding molecular geometries that satisfy distance constraints in an efficient manner.[18−20] The method 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. A detailed description of the algorithm can be found in ref 20 and in the Methods section below.

As we demonstrate in the sequel, SPE tends to produce conformations that are relatively compact. To alleviate this problem, we introduced a simple boosting heuristic that can be used in conjunction with SPE to bias the search toward more extended (or more compact) conformations.[24] The method generates increasingly extended conformations through a series of embeddings, each seeded on the result of the previous one. Here, we extend the literature comparing conformational sampling methods by comparing this "boosted SPE" approach with seven widely used and representative conformational sampling techniques implemented in the Rubicon, Catalyst, Macromodel, Omega, and MOE software packages. We find that boosting is very effective in sampling the full range of geometric sizes attainable by any given molecule as well as in identifying low-energy minima.

METHODS

The comparison was based on a published standard data set of 59 ligands[2] extracted from the Protein Data Bank,[22,23] containing between 3 and 23 rotatable bonds, with a median of 8 (Table 1). For comparison, Lipitor, which is by far the world's best selling drug, has 13 rotatable bonds. 2005 sales figures indicate that 15 small-molecule "blockbuster" drugs (sales > $1 billion) have 10 or more rotatable bonds, and 5 more like clarithromycin have just as much flexibility due to large rings.[24] In addition, blockbuster peptides like the nonapeptide leuprorelin have far more rotatable bonds than any of the molecules in our test set. Flexible ligands are obviously important to practicing drug designers.

Table 1. PDB Codes, Sequence Numbers, and Number of Rotatable Bonds of the Molecules in the Data Set

PDB code	mol. id	no. rot. bonds	PDB code	mol. id	no. rot. bonds
1APT	1	18	1HVR	31	8
1B0H	2	15	1ICN	32	15
1B1H	3	16	1IDA	33	15
1B2H	4	16	1LAH	34	4
1B3H	5	15	1LIC	35	15
1B4H	6	15	1LMO	36	6
1B5H	7	14	1LNA	37	8
1B6H	8	15	1LST	38	5
1B7H	9	16	1MCR	39	5
1CBX	10	5	1PGP	40	7
1ETR	11	10	1POC	41	23
1HVJ	12	19	1PPC	42	9
1HVL	13	19	1PPH	43	7
1STP	14	5	1RNE	44	21
1TMN	15	13	1SME	45	22
1TNI	16	4	2CGR	46	9
1AAQ	17	17	2CMD	47	5
1APU	18	7	2LGS	48	4
1ATL	19	9	2PLV	49	15
1EAP	20	10	2R07	50	8
1EED	21	19	2SIM	51	5
1EPO	22	15	2YHX	52	3
1ETA	23	5	3CPA	53	5
1ETT	24	7	3TMN	54	6
1FKG	25	10	3TPI	55	6
1GLP	26	10	4DFR	56	9
1HEF	27	19	4PHV	57	12
1HFC	28	9	5P2P	58	20
1HRI	29	9	8GCH	59	7
1HTF	30	12			

Nine representative conformational sampling techniques implemented in six different software packages were evaluated. These include SPE, boosting SPE, Rubicon, Catalyst, poling Catalyst, Macromodel, Omega, MOE stochastic search, and MOE systematic search. It is important to realize that the package names are used as shorthand for a particular conformational sampling protocol employed within that package. These results cannot be taken as a general endorsement or an indictment of that package for conformational sampling for several reasons. First, several of these packages include other algorithms besides the particular ones used in this study, which would give different results. For example, "Macromodel" here refers to their serial torsional/low-mode conformational sampling method as detailed below. Macromodel also supplies 8 other search algorithms: torsional sampling, serial torsional sampling, systematic torsional sampling, mixed torsional/low-mode sampling, low-mode sampling, serial low-mode sampling, large-scale low-mode sampling, and mixed torsional/large-scale low-mode sampling. Furthermore, most methods have adjustable parameters that affect the results. Every effort was made to use sensible methods and parameters, and all nondefault parameters are described below. Still, other choices would obviously affect the outcome, so users' results may vary.

Ten thousand conformers were requested from each of these methods except Omega and poling Catalyst, which use a different strategy for selecting conformations (vide infra). First the raw conformers were analyzed. Then, for consistency, the raw conformers generated by each method were minimized using the MMFF94s force field[25−29] and the BFGS variable metric minimization algorithm, as implemented in the DirectedDiversity[30] software suite. This software has been tested thoroughly and has passed the entire MMFF94s

validation suite.[31] The argument has been put forth for many years that there may be many additional geometries near the raw conformations that are of low-energy but not near any local minima. While this is possible in principle, no satisfactory systematic way to identify such geometries has been presented, that is not plagued with serious flaws, so these hypothetical additional pharmacophores must remain an intriguing conjecture for the present. In some cases (Omega and MOE stochastic), it was not easy to decouple energy minimization from the actual search, so a 500 kcal/mol energy cutoff was applied during the search in order to retain all remotely plausible conformers before secondary minimization, in an effort to maximize the diversity of final conformations. The extendedness of the resulting conformations was measured by the radius of gyration, computed with DirectedDiversity. Unless noted otherwise, all calculations were carried out on a series of IBM Intellistations running Windows XP Professional and Red Hat Enterprise Linux and equipped with two 3.2 GHz Xeon processors and 2048 Mb of RAM. These nine methods are described below.

SPE. SPE generates molecular conformations that satisfy a set of geometric constraints. These constraints are derived from the molecular connectivity table and fall into two categories: (1) distance constraints that require the distance between two atoms i and j, d_{ij}, to fall within a certain range $l_{ij} \le d_{ij} \le u_{ij}$ and (2) volume constraints that require the signed volume V_{ijkl} formed by four atoms i, j, k, and l to fall within a certain range $V^l_{ijkl} \le V_{ijkl} \le V^u_{ijkl}$. Collectively, these constraints are sufficient to define all possible 3-dimensional geometries attainable by a given molecule. SPE starts from a random initial configuration and uses a self-organizing scheme to rapidly refine the atomic positions so as to satisfy all the input constraints. Since this is a relatively newer method, the algorithm is summarized below:

1. Establish distance and volume constraints. In this step, distance bounds matrices $\{l_{ij}\}$ and $\{u_{ij}\}$ are populated, where elements l_{ij} and u_{ij} denote the lower and upper bounds for the distance between atoms i and j, respectively. $\{l_{ij}\}$ and $\{u_{ij}\}$ are determined from the connectivity table and geometric parameters such as bond lengths, angles, and atomic radii. In addition, upper $\{V^u_{ijkl}\}$ and lower $\{V^l_{ijkl}\}$ volume bounds are established for every set of four atoms i, j, k, and l that are attached to tetrahedral atoms with explicit chirality or required to have a planar configuration.

2. Randomly place the atoms in a box of appropriate size.

3. Select a distance learning rate λ_d, a volume learning rate λ_v, and a relative frequency for enforcing a distance or a volume constraint, ν.

4. With probability ν, do (5); otherwise, do (6).

5. Randomly select a pair of atoms, i and j, and compute their distance $d_{ij} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$. If $l_{ij} \le d_{ij} \le u_{ij}$, leave the atomic positions unchanged. Otherwise, update the coordinates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ by

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i + \lambda_d \frac{1}{2} \frac{t_{ij} - d_{ij}}{d_{ij} + \epsilon} (\boldsymbol{x}_i - \boldsymbol{x}_j)$$

and

$$\boldsymbol{x}_j \leftarrow \boldsymbol{x}_j + \lambda_d \frac{1}{2} \frac{t_{ij} - d_{ij}}{d_{ij} + \epsilon} (\boldsymbol{x}_j - \boldsymbol{x}_i)$$

where t_{ij} is the nearest bound to d_{ij} (i.e., $t_{ij} = l_{ij}$ if $d_{ij} < l_{ij}$, or $t_{ij} = u_{ij}$ if $d_{ij} > u_{ij}$), and ϵ is a small number used to avoid division by zero.

6. Randomly select a volume constraint k and the four atoms involved, p, q, s, and t. Compute the signed volume V_{pqst} formed by the 4 atoms. If $V^l_k < V_{pqst} < V^u_k$, leave the atom positions unchanged. Otherwise, compute the gradient of the signed volume with respect to the atomic positions, $\mathbf{g}_\mu = \nabla_\mu V_{pqst}$, where $\mu = p$, q, s, and t, and update the atomic coordinates by

$$\mathbf{x}_\mu \leftarrow \mathbf{x}_\mu + \lambda_v (V^0_k - V_{pqst}) \frac{\mathbf{g}_\mu}{\sum_{\beta=p,q,s,t} |\mathbf{g}_\mu|^2}$$

where V^0_k is the nearest bound to V_{pqst} (i.e., $V^0_k = V^l_k$ if $V_{pqst} < V^l_k$ or $V^0_k = V^u_k$ if $V_{pqst} > V^u_k$).

7. Repeat (4)−(6) for a prescribed number of steps, S.

8. Decrease the learning rates λ_d and λ_v by prescribed decrements $\delta\lambda_d$ and $\delta\lambda_v$.

9. Repeat (4)−(8) for a prescribed number of cycles, C.

In the present study, a total of 10 000 conformers were generated for each hydrogen-depleted molecule 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 $\nu = \min\left(0.5, 8.0 \times \frac{||V||}{N(N+1)/2+8||V||}\right)$, 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, υ is the ratio of volume to distance refinements, N is the number of atoms in the molecule, and $||V||$ is the total number of volume constraints. Each conformation was derived by running the algorithm with a different random number seed. We refer the reader to ref 23 for a complete description, validation, and performance benchmark of the algorithm as well as a comparison to other distance geometry methods.

SPE w/ Boosting. Boosting is a simple heuristic that can be used in conjunction with SPE to generate increasingly extended (or compact) conformations through iterations. In the first iteration, a normal SPE embedding is performed as described in the previous section, 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, $\{V^u_{ijkl}\}$ and $\{V^l_{ijkl}\}$, 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. More details can be found in ref 24.

In this study, a total of 10 000 conformations were generated for each molecule using the same parameter settings as outlined above. 6250 of these conformations were generated by tightening the lower bounds (i.e., toward more extended conformations), using 1250 independent trials and 4 boosting iterations (i.e., 5 embeddings) per trial. The remaining 3750 conformations were generated by tightening the upper bounds (i.e., toward more compact conformations), using 1250 trials and 2 boosting iterations (i.e., 3 embeddings) per trial.

Rubicon. Rubicon[32] uses a distance geometry method very similar in principle to SPE. Like SPE, Rubicon sets upper and lower bounds for distances and volumes but uses a different algorithm to find atomic coordinates that satisfy those bounds. Rather than selecting random starting coordinates, Rubicon selects a random set of distances within the bounds and uses the "metric matrix" algorithm[36] to generate approximate 3-D coordinates, followed by conjugate gradient minimization to refine these coordinates so they satisfy the original bounds. It also has its own set of chemical rules to set the distance and volume bounds, which differ somewhat from those implemented in the SPE method above. Because SPE samples random pairs of atoms and adjusts their coordinates, it only considers a small fraction of the $(N^2 - N)/2$ interatomic distances and scales linearly with the number of atoms, N. The conjugate gradient algorithm in Rubicon minimizes all pairwise distance errors and therefore scales with N^2. Thus, SPE is much faster for large molecules.

In this study, a total of 10 000 conformers were generated for each molecule, using one trial per conformation, 0.5 Å maximum distance violation, 0.5 Å^3 maximum volume violation (and volume constraints enabled), 1−4 bump checking enabled, redundancy checking disabled, and default values for all the remaining parameters. Since subsequent minimization was to be performed on all conformations, hydrogens were ignored even though their inclusion is known to generate better geometries (similarly to SPE).

Catalyst. Two sets of conformers were generated using the command-line tool "catConf" supplied with Catalyst version 4.10.[34] The first used was Catalyst's poling technique, which introduces an artificial potential during the sampling to repel similar conformers, thereby promoting conformational variation.[35] Five hundred conformers were requested for each molecule, though the actual number produced ranged from 43 to 456. The second set of conformers was obtained with poling disabled (confAnalysis.best.no.poling parameter set to true). In this case, 10 000 conformers were requested, and the program indeed produced as many for most molecules, although for some small molecules it generated only as few as 115 conformers. Duplicate checking was disabled (confGen.removeTopSymDuplicate parameter set to false), and all other parameters were kept at their default values. Both sets of conformers (with and without poling) were generated using the BEST conformer generation routine, which produces conformers of better quality and diversity compared to the FAST method, which achieves better speed.

Omega. Omega version 1.8.1[36] was used to generate conformations using a systematic rule-based approach. This algorithm divides each molecule into component fragments, which may contain up to five contiguous rotatable bonds. A library of predefined angles is used to generate conformations for each fragment, which are then assembled together to construct the conformations of the whole molecule using a depth-first, divide-and-conquer approach, driven by the fragment energies. Ten thousand conformers were requested for each molecule. Gas-phase conditions were mimicked by eliminating the Coulombic and attractive van der Waals terms during the final refinement. Default values were used for all the remaining parameters.

Macromodel. A total of 10 000 conformers were generated for each molecule using the serial torsional/low-mode conformational sampling method in Macromodel version 9.0016.[37] This is a hybrid technique that combines broad Monte Carlo sampling of torsional space[38] with local low-frequency eigenvector sampling in the vicinity of the current conformation.[39] Since minimization could not be decoupled from the actual search, and since we wanted to apply a consistent minimization technique across all methods, a single cycle of truncated Newton conjugate gradient[40] minimization was performed on each conformer using the MMFF94s force field. No energy cutoff was applied to discard unreasonable conformations. Default values were used for all other parameters.

MOE. Two sets of conformers were generated with MOE version 2005.6.[41] The first was obtained using MOE's stochastic conformational search, which is similar to Ferguson's Random Incremental Pulse Search method[42] 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. Ten thousand conformations were requested for each molecule, and the resulting geometries were minimized in MOE using MMFF94s with the distance-dependent dielectric disabled. Ninety successive failed attempts at generating new conformations prior to termination were applied in order to maximize the number of conformers produced within the specified energy window. A 500 kcal/mol energy cutoff was used in order to retain all sensible conformations for subsequent minimization with DirectedDiversity, and all other parameters were set to their default values. The second set of conformers was generated using a systematic fragment-based approach. This algorithm breaks the molecule up into overlapping fragments, retrieves precomputed conformations of those fragments, and reassembles them by rigid body superimposition. Fragment conformations with strain energy greater than 5 kcal/mol were ignored. Once again, 10 000 conformations were requested for each molecule. All other parameters were set to their default values.

Pharmacophore Analysis. For each conformation, all unique 3-point pharmacophores were computed using the five atom type definitions in Table 2[43] and 16 distance bins spanning the intervals [0,1], (1, 2], (2, 3], ..., (14, 15] and (15, +∞). (Positively and negatively charged centers were not taken into consideration, as all the molecules in our training set were neutralized during preprocessing.) Each triplet was canonicalized using a simple vertex ordering

Table 2. Smarts[45] Definitions of Pharmacophore Atoms Types[a]

```
#FEATURE D
[OH;!$([OH][C,S,P]=O)] 0 include
[S;!H0] 0 include
[CH]#C 0 include
[N;!H0;!$([N;!H0]#C);!$([NH](C(F)(F)(F))S(=O)=O);!$([nH]
   1nnnc1);!$([nH]1nncn1);!$([nH]1ncnn1)] 0 include
#FEATURE A
[#7,#8,#16;-0,-1;!$([o,s,nX3]);!$([Nv5,Pv4,Pv5,Sv4,Sv6])]
   0 include
#FEATURE P
[+;!$([+]~[-])] 0 include
#FEATURE N
[-;!$([-]~[+])] 0 include
#FEATURE R
a1aaaaa1 0,1,2,3,4,5 include
a1aaaa1 0,1,2,3,4 include
```

[a] Each entry following the #FEATURE keyword consists of 3 identifiers: the SMARTS definition, the zero-based indices of the atoms in the SMARTS pattern defining the centroid of the pharmacophore feature, and a keyword indicating whether the pattern should be included or excluded. The rules were kept intentionally simple.

algorithm and assigned a unique id or bit location in a binary fingerprint. The fingerprints of all the conformers generated by a given method for a given molecule were combined into a method-specific, multiconformer molecular fingerprint. The number of '1' bits in these multiconformer fingerprints was used to quantify the pharmacophore space covered by that specific method and molecule. The precise binning strategy and atom type definitions were chosen to provide sufficient resolution and discriminatory power and were not intended to be exhaustive.

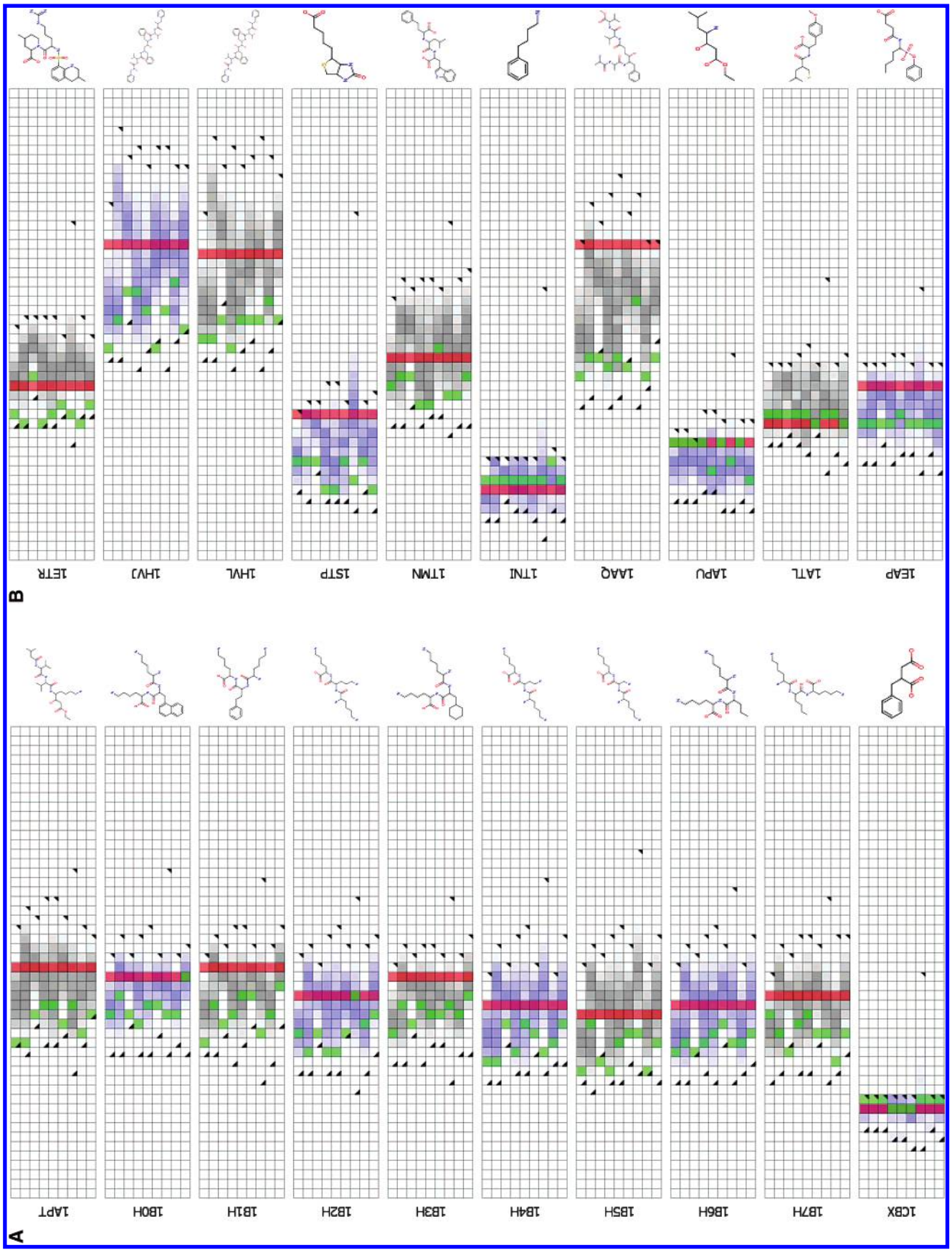
RESULTS AND DISCUSSION

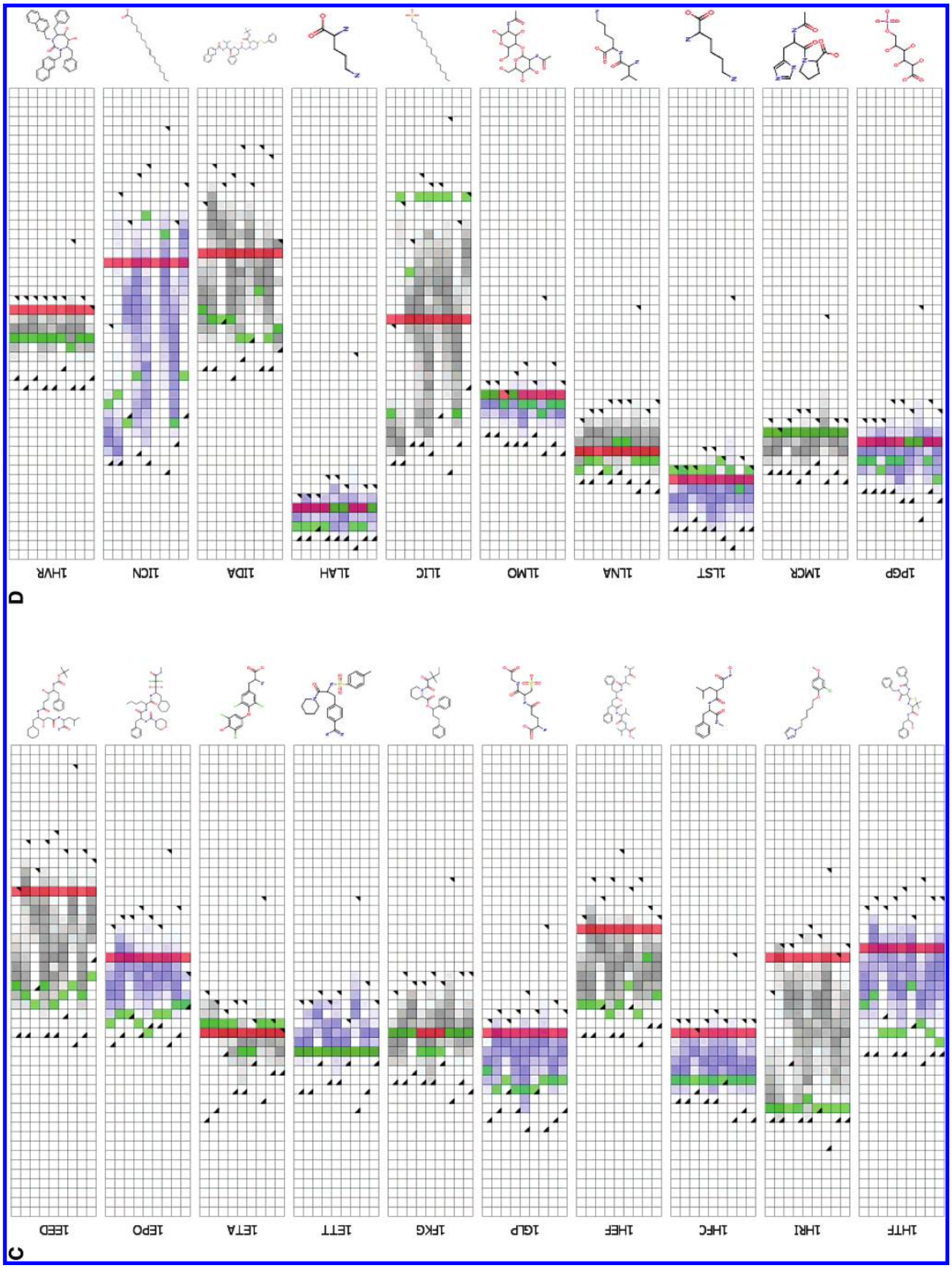
Conformational sampling is often not an end in itself but rather the means to some further end. One way to assess the quality of conformational sampling, therefore, would be to assess success in some specific drug design problem. However, this "case study" approach, i.e., showing that conformations from one method perform well on this or that specific application compared to other methods, suffers from questions of mechanism and idiosyncrasy. Can superior performance on this or that particular application be isolated and definitively attributed to superior conformational sampling? Will it apply widely to other problems? Because identifying the full ensemble of accessible conformations is such a fundamental and ubiquitous molecular property, more direct, less problem-specific strategies are needed. In contrast to the "downstream endpoint" of a case study, directly measuring the underlying mechanism is analogous to a translational medicine endpoint. In this case, the thoroughness of conformational sampling can be directly measured, at least relatively among widely used available methods. This study employs three different metrics to measure the thoroughness of conformational sampling: (1) the distribution of the radius of gyration of the conformations identified by each method (a measure of compactness/extendedness), (2) the distribution of energies of the minimized conformations, and (3) the pharmacophore diversity in the resulting conformational ensembles. The results are illustrated in Figures 1−3 and Tables 3 and 4.

Figures 1 and 2 contain a series of panels, one for each molecule, showing the distribution of the radius of gyration for all the conformers obtained by each method before (Figure 1) and after (Figure 2) minimization. For ease of comparison, all the heatmaps are identically scaled from 1.5 to 8.5, which covers the full range of radii encountered across all conformers, all molecules, and all methods. As a frame of reference, the bin containing the lowest energy conformer identified by each method is highlighted in green, and that containing the bioactive conformation is highlighted in red.

SPE tends to produce significantly more compact conformations compared to the other methods, including its close relative, Rubicon. This is not entirely surprising if one considers the way in which SPE refines the atomic coordinates. Because the embedding starts from a random initial conformation, many topologically neighboring atoms have to approach each other from opposite directions. A typical SPE optimization proceeds through an initial contraction phase where atoms cross each other to get close to their neighbors, followed by an expansion phase where the coordinates are relaxed to satisfy all the distance constraints. There is no intrinsic force in SPE driving the system to an extended geometry, as there are typically many geometrically feasible compact conformations that are more easily accessible from these intermediate "imploded" states. The algorithm will generate extended conformations only if these are enforced by the distance constraints. There are three apparent ways to alleviate this problem within the general SPE framework. The first is to use constraints that can be satisfied only by extended geometries (such as the boosting strategy described herein), the second is to use partially extended (but still random) starting geometries, and the third is to introduce additional moves that probabilistically stretch the molecule during the course of the embedding. An example of the latter two approaches will be presented in due course. MOE stochastic search is the only other method that produces similarly compact geometries to SPE.

Interestingly, Rubicon does not suffer from this problem, presumably because it starts from the nearest "linear" 3-D embedding of initial random distances, rather than from random initial coordinates. This scheme ensures that conformations with large interatomic distances are sampled. Also, Rubicon's conjugate gradient minimization works in two phases: it first minimizes bounds violations in 4 dimensions (which allows atoms to pass through each other) and then minimizes the 4th dimension toward zero, collapsing the conformation into 3 dimensions. The use of a fourth dimension increases the effective volume where atoms can move, causing the molecule to expand more freely. It will be interesting to see if 4-dimensional minimization works equally well with the SPE stochastic minimizer (work in progress). In general, the two methods explore different, complementary regions of conformational space. The range of radii sampled by either of them is relatively narrow, with SPE offering a slight advantage except for long, linear molecules such as 1ICN (Figure 1/2d), ILIC (Figure 1/2d), and 2PLV (Figure 1/2e) and to a lesser extent 1SME (Figure 1/2e). In the first three of these cases, the bioactive conformation lies outside the range of radii sampled by the raw SPE geometries (Figure 1), but this changes upon minimization (Figure 2), which relaxes the molecules and expands their size. For smaller molecules, this expansion is less pronounced. In contrast, Rubicon geometries contract substantially upon minimization, indicating significant strain


B
1ETR
1HVJ
1HVL
1STP
1TMN
1TNI
1AAQ
1APU
1ATL
1EAP
A
1APT
1B0H
1B1H
1B2H
1B3H
1B4H
1B5H
1B6H
1B7H
1CBX


D
1PGP
1MCR
1LST
1LNA
1LMO
1LIC
1LAH
1IDA
1ICN
1HVR
C
1HTF
1HRI
1HFC
1HEF
1GLP
1FKG
1ETT
1ETA
1EPO
1EED

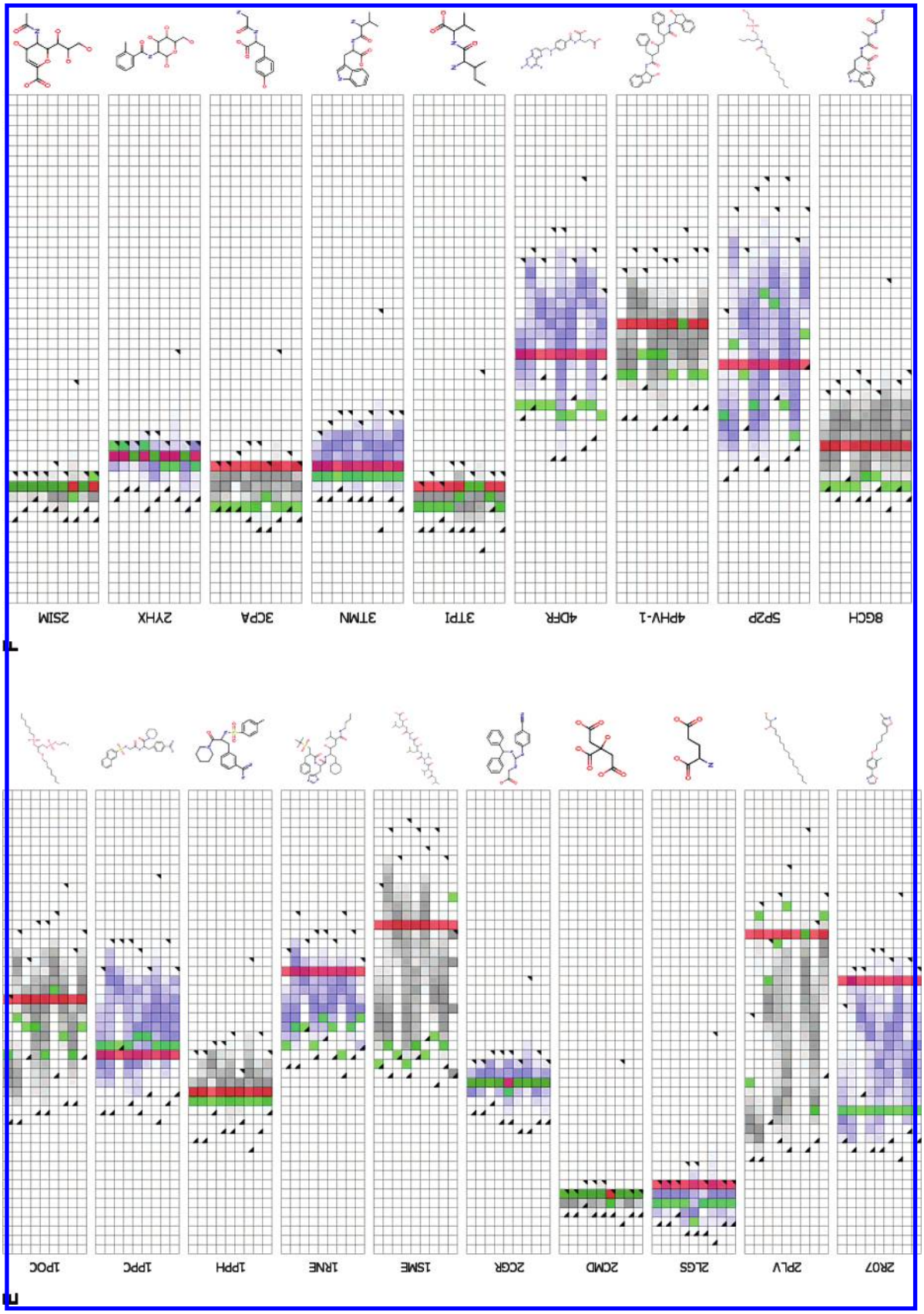
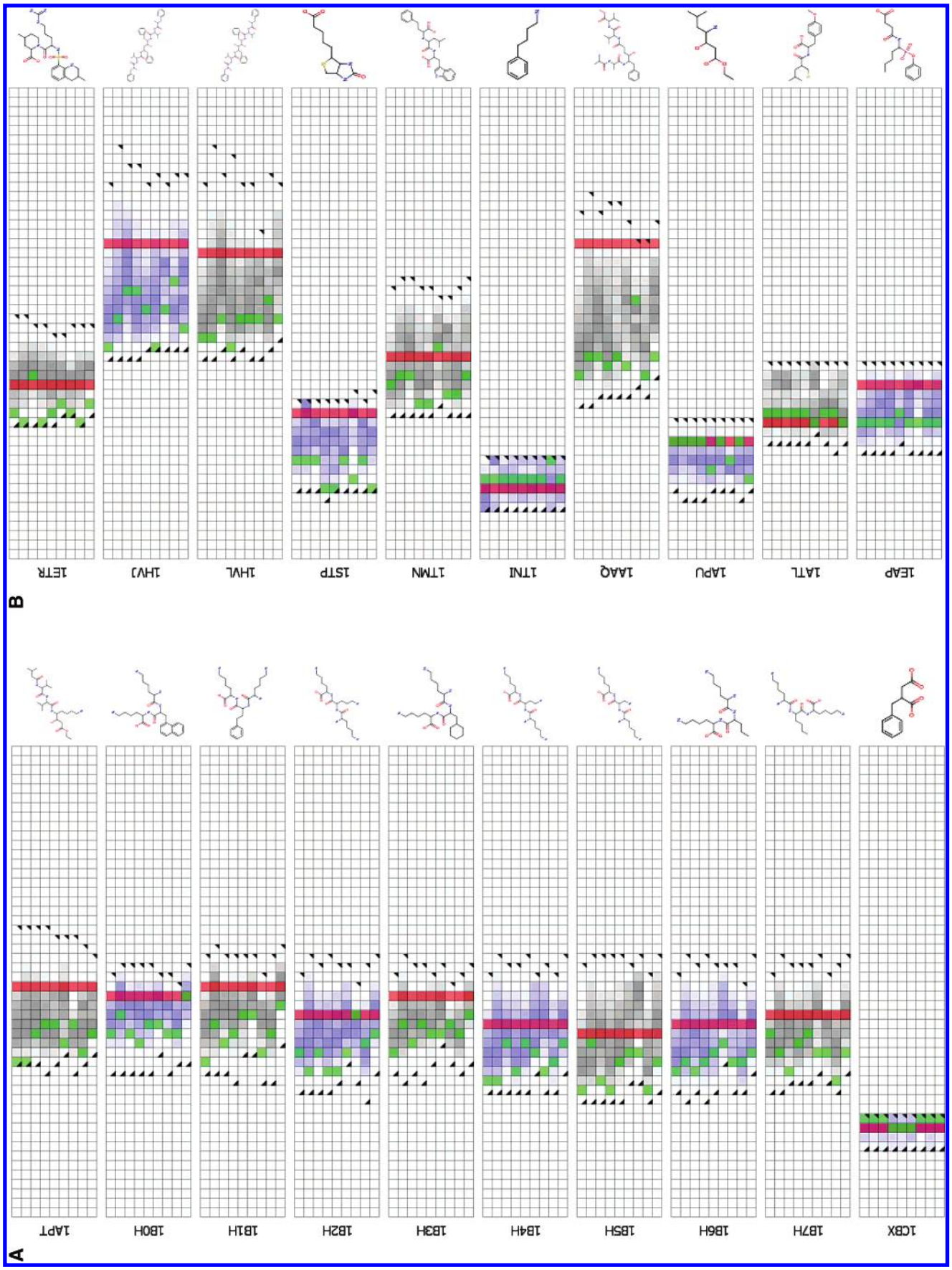
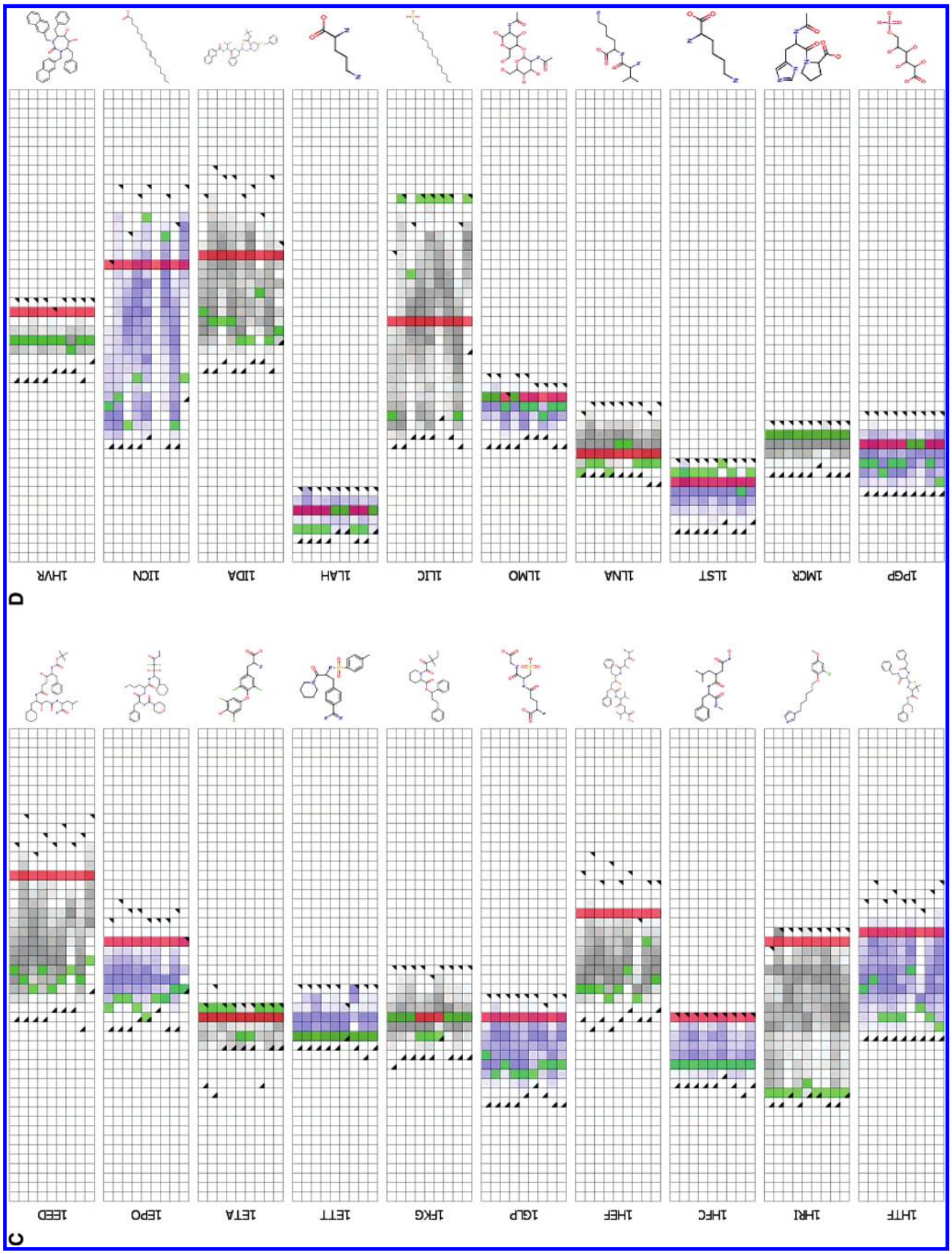


Figure 1. Distribution of radius of gyration of the raw conformations generated by each method for each molecule. All the heatmaps are identically scaled from 1.5 to 8.5, which covers the full range of radii encountered across all conformers, all molecules, and all methods. Grey and blue colors alternate for greater clarity. The intensity of the color is proportional to the fraction of conformations that fall in each bin. The black triangles indicate the bins containing the minimum and maximum values. The red and green squares indicate the bins containing the bioactive and lowest energy structure conformation, respectively. Within each molecule panel, the methods appear in the following order: SPE, boosting SPE, Rubicon, Catalyst, poling Catalyst, Omega, Macromodel, MOE stochastic, and MOE systematic.


B
1ETR
1HVJ
1HVL
1STP
1TMN
1TNI
1AAQ
1APU
1ATL
1EAP
A
1APT
1B0H
1B1H
1B2H
1B3H
1B4H
1B5H
1B6H
1B7H
1CBX


1PGP
1MCR
1LST
1LNA
1LMO
1LIC
1LAH
1IDA
1ICN
1HVR
D
1HTF
1HRI
1HFC
1HEF
1GLP
1FKG
1ETT
1ETA
1EPO
1EED
C

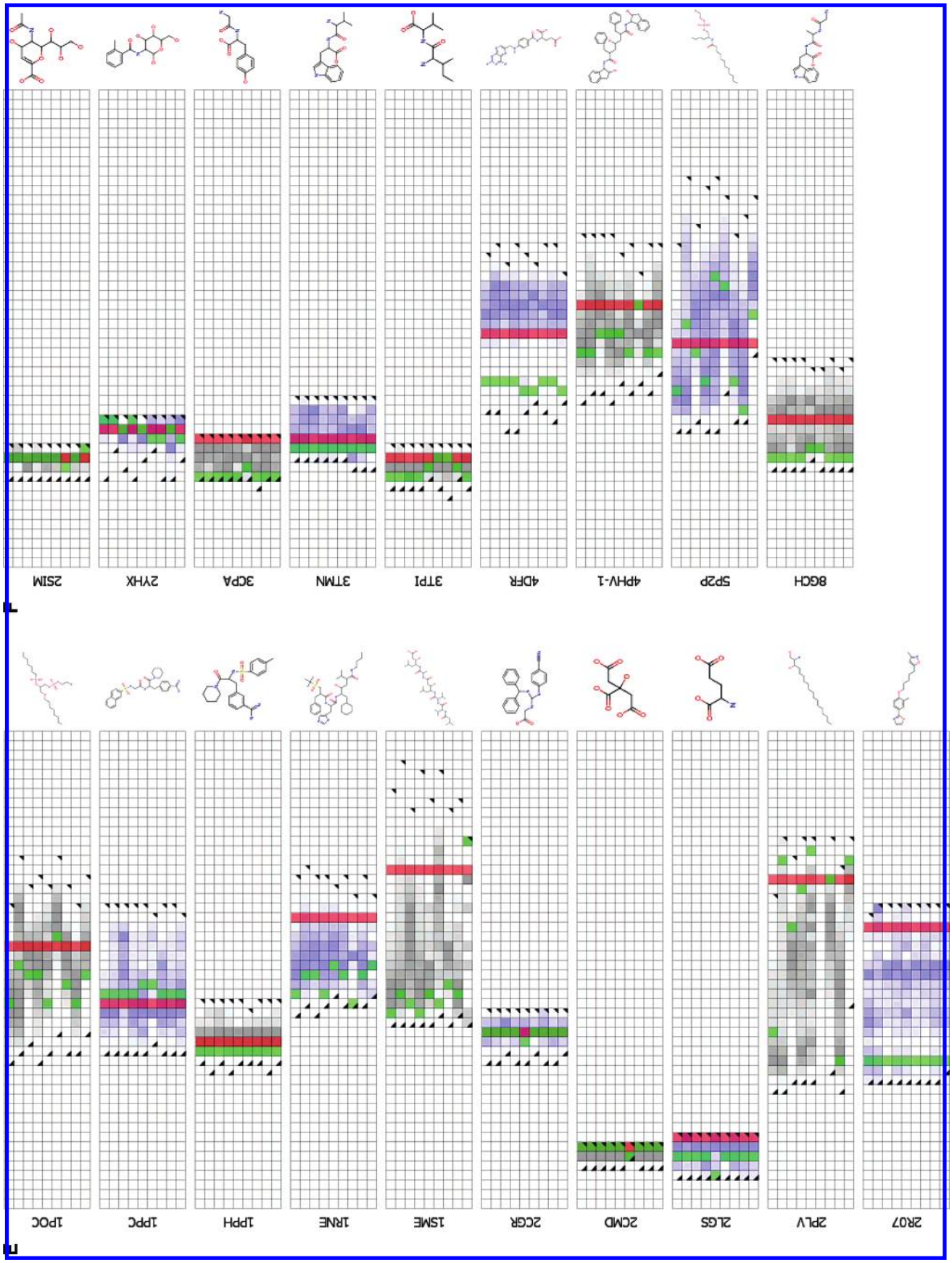
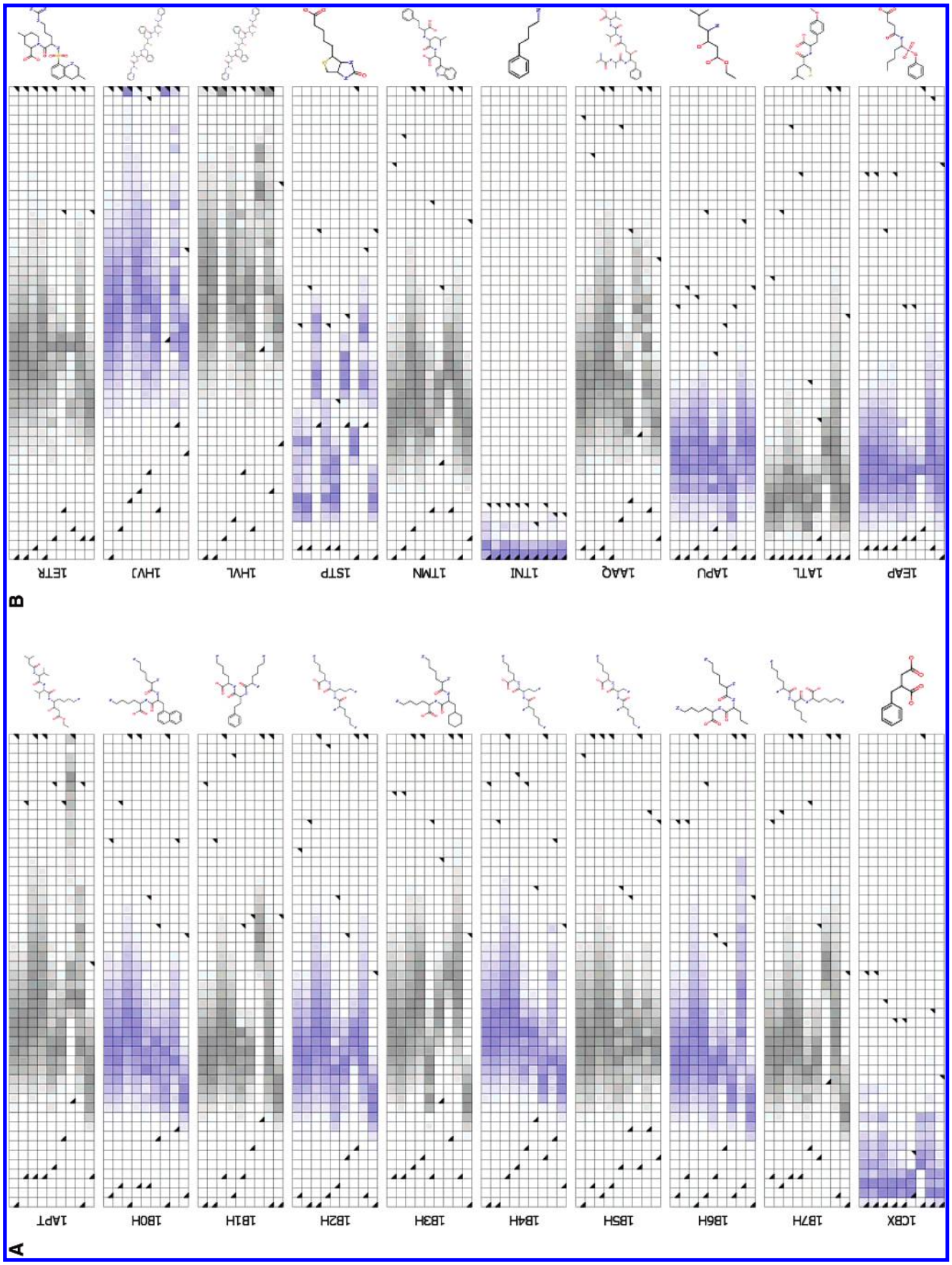
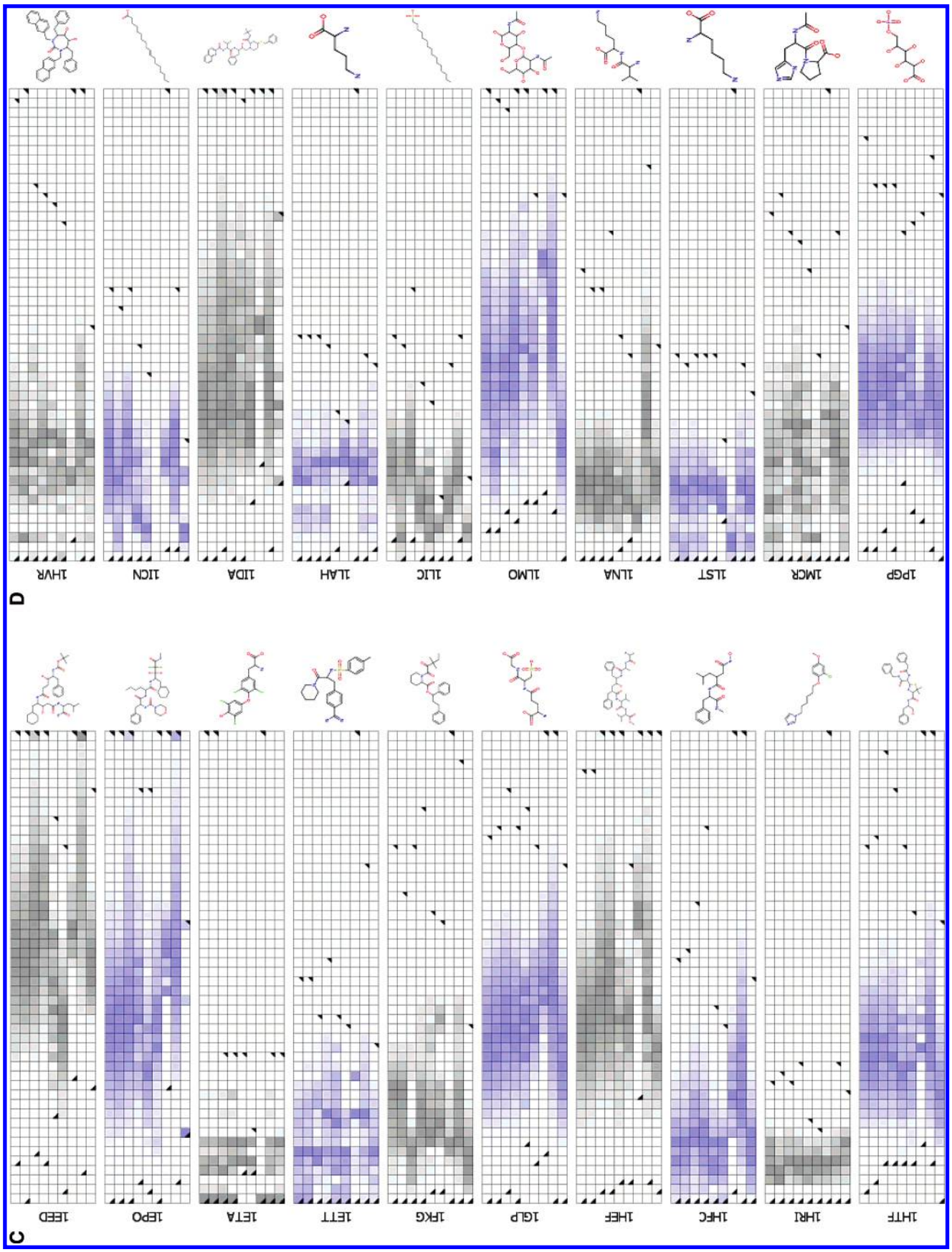


Figure 2. Same as Figure 1, with the exception that the radii correspond to the MMFF94s-minimized conformations.


B
1ETR
1HVJ
1HVL
1STP
1TMN
1TNI
1AAQ
1APU
1ATL
1EAP
A
1APT
1B0H
1B1H
1B2H
1B3H
1B4H
1B5H
1B6H
1B7H
1CBX


1PGP
1MCR
1LST
1LNA
1LMO
1LIC
1LAH
1IDA
1ICN
1HVR
D
1HTF
1HRI
1HFC
1HEF
1GLP
1FKG
1ETT
1ETA
1EPO
1EED
C

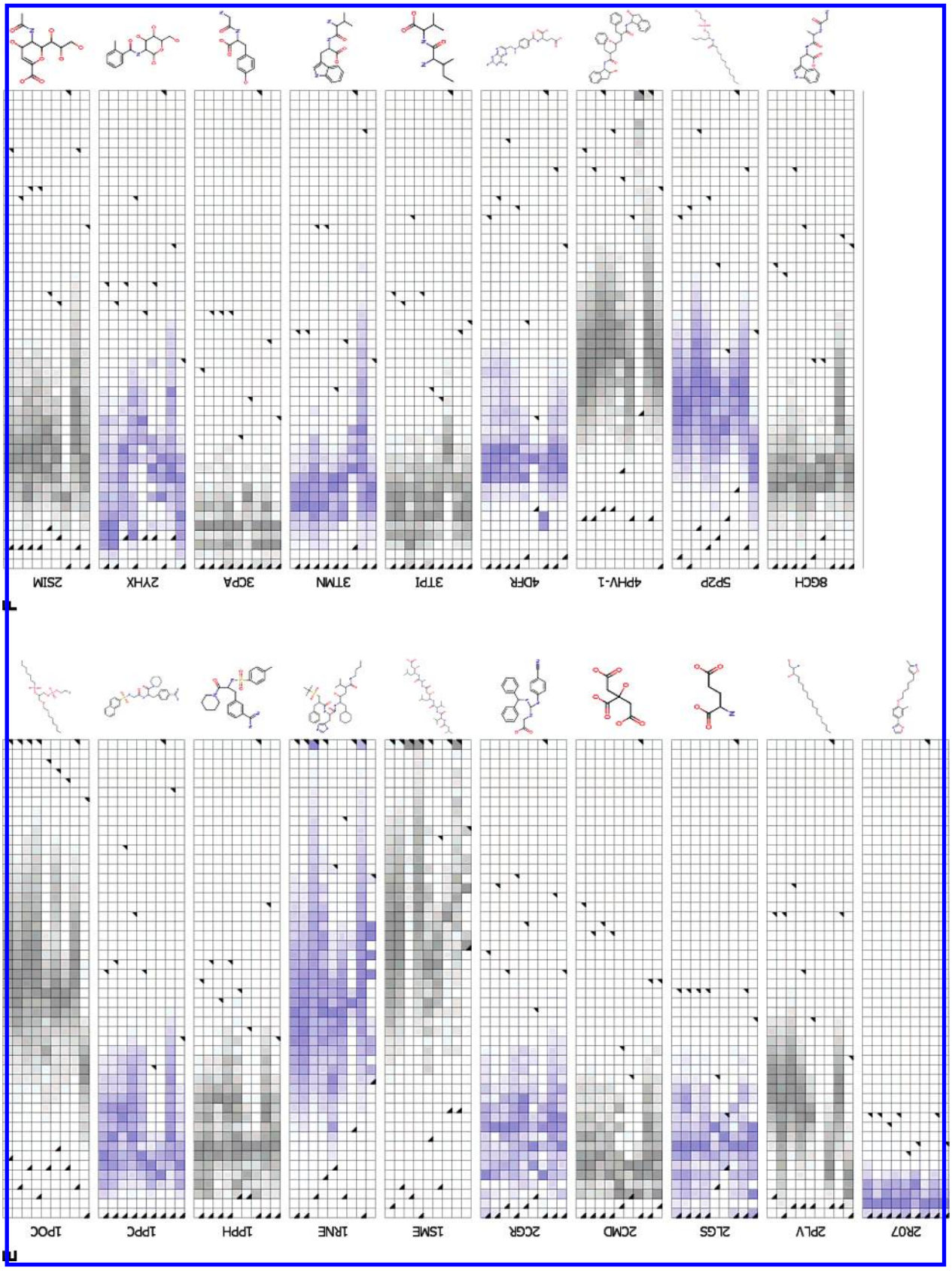


Figure 3. Distribution of relative energies (in kcal/mol) of the minimized conformations generated by each method for each molecule. Energies are relative to the lowest energy structure identified by any method for a given molecule. All the heatmaps are identically scaled from 0 to 50 kcal/mol. The black triangles indicate the bins containing the minimum and maximum values. Within each molecule panel, the methods appear in the following order: SPE, boosting SPE, Rubicon, Catalyst, poling Catalyst, Omega, Macromodel, MOE stochastic, and MOE systematic.

Table 3. Number of Unique Pharmacophore Triplets in the Conformational Ensembles Produced by Each Method

PDB code	SPE[a]	SPE_B[b]	RUB[c]	CAT[d]	CAT_P[e]	OMEGA[f]	MAC[g]	MOE STO[h]	MOE SYS[i]
1APT	2844	3113	3061	2977	2311	2356	2432	2168	2153
1B0H	6518	6960	6806	6891	5645	6251	6591	5587	6636
1B1H	6513	6902	6725	6739	5269	5815	6448	6618	6405
1B2H	4097	4360	4262	4283	3677	4230	4090	4237	4376
1B3H	3455	3630	3644	3600	2935	3309	3420	3627	3338
1B4H	3923	4183	4071	4076	3478	3838	4070	3258	4144
1B5H	3753	4017	3906	3944	3322	3762	4027	3211	3939
1B6H	3480	3742	3703	3671	3153	3247	3575	3209	3447
1B7H	3517	3712	3720	3680	3063	3350	3617	3670	3410
1CBX	226	226	223	200	189	129	257	215	201
1ETR	3882	4060	4022	3967	3249	3459	3754	3955	3743
1HVJ	10148	10380	10396	10439	8569	9916	9791	7573	9118
1HVL	10351	10562	10432	10503	8515	10083	8583	9444	9142
1STP	608	616	535	579	540	449	648	520	640
1TMn	3242	3284	3279	3351	2641	2684	2861	3378	2918
1TNI	0	0	0	0	0	0	0	0	0
1AAQ	6521	6783	6622	6707	5203	6175	5951	5196	5373
1APU	134	134	135	136	116	135	132	136	147
1ATL	1284	1277	1342	1342	1064	1092	1309	1404	1246
1EAP	1105	1112	1124	1127	967	1007	1053	1177	1106
1EED	6006	6096	5869	5838	4773	5103	5989	6102	5730
1EPO	5020	5038	5101	4474	4028	4642	4068	5119	1120
1ETA	707	788	597	621	528	493	695	614	514
1ETT	1088	1087	999	1112	894	739	1097	1111	922
1FKG	378	381	364	385	310	249	359	399	370
1GLP	2192	2343	2273	2317	2088	2040	2042	2257	2128
1HEF	8659	8845	8681	8797	6781	8218	6697	7462	7471
1HFC	1735	1730	1773	1802	1523	1527	1659	1939	1733
1HRI	251	266	253	257	213	230	272	263	239
1HTF	5413	5610	5471	5645	4202	4765	4827	5521	5046
1HVR	1881	1866	1700	1845	1303	1607	1647	1925	1736
1ICN	0	0	0	0	0		0	0	0
1IDA	10121	10349	10276	10224	7647	9514	9286	10400	3464
1LAH	189	192	192	188	153	162	197	179	169
1LIC	1	1	1	1	1	1	1	1	1
1LMo	1627	1582	1552	1693	1468	1546	1746	1819	1514
1LNA	962	984	988	988	842	915	944	910	1026
1LST	270	265	264	267	252	210	265	257	222
1MCr	905	913	916	893	791	647	994	924	859
1PGP	1037	1090	1008	1087	1008	927	1054	1037	1057
1POC	652	684	669	697	581	581	591	687	503
1PPC	4478	4474	4434	4504	3732	3963	4370	4707	4014
1PPH	1431	1434	1362	1442	1244	1040	1405	1436	1299
1RNE	6841	7036	6785	5610	4601	5750	5133	6909	2234
1SME	4162	4278	4095	4166	3675	4100	4006	3773	2263
2CGR	1709	1723	1916	1847	1484	1405	1783	1841	1572
2CMd	305	303	304	305	271	263	320	305	283
2LGS	203	207	207	207	177	127	215	203	194
2PLV	45	45	45	45	32		45	45	39
2R07	454	473	436	469	389	432	465	470	405
2SIM	842	836	774	831	758	721	823	848	855
2YHX	727	721	733	891	612	542	867	828	662
3CPA	1094	1095	1191	1069	671	936	1255	1020	766
3TMn	608	615	618	629	543	572	646	645	601
3TPI	167	167	175	177	160	163	201	162	167
4DFR	6587	6674	6571	6744	6081	5137	6775	6850	5959
4PHV	6411	6638	6455	6719	4849	6334	6231	6559	6500
5P2P	555	556	601	574	527	352	579	620	509
8GCH	1837	1859	1895	1864	1417	1645	2013	1821	1876

[a] SPE. [b] SPE with boosting. [c] Rubicon. [d] Catalyst. [e] Catalyst with poling. [f] Omega. [g] Macromodel. [h] MOE stochastic. [i] MOE systematic.

in the raw geometries.

SPE's tendency for compact geometries is compensated by the use of boosting, which greatly expands the range of conformations sampled and covers the broadest spectrum of geometric sizes among all the methods examined, except Catalyst, which was equally effective. Interestingly, the heatmaps in Figure 1 show a characteristic bimodal distribution for every molecule of considerable flexibility. This is not an intrinsic limitation of our method but rather an artifact of our decision to explore the effect of boosting in both directions (toward more extended as well as more compact conformations). As it turns out, the latter is unnecessary, since SPE already shows a bias toward compact conformations. The bimodal effect is not observed when boosting is used only toward more extended conformations (results not shown) and disappears upon minimization (Figure 2).

Perhaps the most striking feature of Figure 1 is the substantially wider range of radii sampled by Macromodel.

Table 4. Fraction of Unique Pharmacophore Triplets in the Conformational Ensembles Produced by Each Method over the Maximum Identified by Any Method[m]

PDB code	SPE[a]	SPE_B[b]	RUB[c]	CAT[d]	CAT_P[e]	OMEGA[f]	MAC[g]	MOE STO[h]	MOE SYS[i]
1APT	0.91	1.00	0.98	0.96	**0.74**	0.76	0.78	**0.70**	**0.69**
1B0H	0.94	1.00	0.98	0.99	0.81	0.90	0.95	0.80	0.95
1B1H	0.94	1.00	0.97	0.98	0.76	0.84	0.93	0.96	0.93
1B2H	0.94	1.00	0.97	0.98	0.84	0.97	0.93	0.97	1.00
1B3H	0.95	1.00	1.00	0.99	0.81	0.91	0.94	1.00	0.92
1B4H	0.94	1.00	0.97	0.97	0.83	0.92	0.97	0.78	0.99
1B5H	0.93	1.00	0.97	0.98	0.82	0.93	1.00	0.80	0.98
1B6H	0.93	1.00	0.99	0.98	0.84	0.87	0.96	0.86	0.92
1B7H	0.95	1.00	1.00	0.99	0.82	0.90	0.97	0.99	0.92
1CBX	0.88	0.88	0.87	0.78	**0.74**	**0.50**	1.00	0.84	0.78
1ETR	0.96	1.00	0.99	0.98	0.80	0.85	0.92	0.97	0.92
1HVJ	0.97	0.99	1.00	1.00	0.82	0.95	0.94	**0.73**	0.87
1HVL	0.98	1.00	0.99	0.99	0.81	0.95	0.81	0.89	0.87
1STP	0.94	0.95	0.83	0.89	0.83	**0.69**	1.00	0.80	0.99
1TMn	0.96	0.97	0.97	0.99	0.78	0.79	0.85	1.00	0.86
1TNI									
1AAQ	0.96	1.00	0.98	0.99	0.77	0.91	0.88	0.77	0.79
1APU	0.91	0.91	0.92	0.93	0.79	0.92	0.90	0.93	1.00
1ATL	0.91	0.91	0.96	0.96	0.76	0.78	0.93	1.00	0.89
1EAP	0.94	0.94	0.95	0.96	0.82	0.86	0.89	1.00	0.94
1EED	0.98	1.00	0.96	0.96	0.78	0.84	0.98	1.00	0.94
1EPO	0.98	0.98	1.00	0.87	0.79	0.91	0.79	1.00	**0.22**
1ETA	0.90	1.00	0.76	0.79	**0.67**	**0.63**	0.88	0.78	**0.65**
1ETT	0.98	0.98	0.90	1.00	0.80	**0.66**	0.99	1.00	0.83
1FKG	0.95	0.95	0.91	0.96	0.78	**0.62**	0.90	1.00	0.93
1GLP	0.94	1.00	0.97	0.99	0.89	0.87	0.87	0.96	0.91
1HEF	0.98	1.00	0.98	0.99	0.77	0.93	0.76	0.84	0.84
1HFC	0.89	0.89	0.91	0.93	0.79	0.79	0.86	1.00	0.89
1HRI	0.92	0.98	0.93	0.94	0.78	0.85	1.00	0.97	0.88
1HTF	0.96	0.99	0.97	1.00	**0.74**	0.84	0.86	0.98	0.89
1HVR	0.98	0.97	0.88	0.96	**0.68**	0.83	0.86	1.00	0.90
1ICN									
1IDA	0.97	1.00	0.99	0.98	**0.74**	0.91	0.89	1.00	**0.33**
1LAH	0.96	0.97	0.97	0.95	0.78	0.82	1.00	0.91	0.86
1LIC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1LMo	0.89	0.87	0.85	0.93	0.81	0.85	0.96	1.00	0.83
1LNA	0.94	0.96	0.96	0.96	0.82	0.89	0.92	0.89	1.00
1LST	1.00	0.98	0.98	0.99	0.93	0.78	0.98	0.95	0.82
1MCr	0.91	0.92	0.92	0.90	0.80	**0.65**	1.00	0.93	0.86
1PGP	0.95	1.00	0.92	1.00	0.92	0.85	0.97	0.95	0.97
1POC	0.94	0.98	0.96	1.00	0.83	0.83	0.85	0.99	**0.72**
1PPC	0.95	0.95	0.94	0.96	0.79	0.84	0.93	1.00	0.85
1PPH	0.99	0.99	0.94	1.00	0.86	**0.72**	0.97	1.00	0.90
1RNE	0.97	1.00	0.96	0.80	**0.65**	0.82	0.73	0.98	**0.32**
1SME	0.97	1.00	0.96	0.97	0.86	0.96	0.94	0.88	**0.53**
2CGR	0.89	0.90	1.00	0.96	0.77	**0.73**	0.93	0.96	0.82
2CMd	0.95	0.95	0.95	0.95	0.85	0.82	1.00	0.95	0.88
2LGS	0.94	0.96	0.96	0.96	0.82	**0.59**	1.00	0.94	0.90
2PLV	1.00	1.00	1.00	1.00	**0.71**		1.00	1.00	0.87
2R07	0.96	1.00	0.92	0.99	0.82	0.91	0.98	0.99	0.86
2SIM	0.98	0.98	0.91	0.97	0.89	0.84	0.96	0.99	1.00
2YHx	0.82	0.81	0.82	1.00	**0.69**	**0.61**	0.97	0.93	**0.74**
3CPA	0.87	0.87	0.95	0.85	**0.53**	**0.75**	1.00	0.81	**0.61**
3TMn	0.94	0.95	0.96	0.97	0.84	0.89	1.00	1.00	0.93
3TPI	0.83	0.83	0.87	0.88	0.80	0.81	1.00	0.81	0.83
4DFR	0.96	0.97	0.96	0.98	0.89	**0.75**	0.99	1.00	0.87
4PHV	0.95	0.99	0.96	1.00	**0.72**	0.94	0.93	0.98	0.97
5P2P	0.90	0.90	0.97	0.93	0.85	**0.57**	0.93	1.00	0.82
8GCH	0.91	0.92	**0.94**	**0.93**	0.70	**0.82**	**1.00**	**0.90**	**0.93**
min[j]	0.82	0.81	**0.76**	**0.78**	**0.53**	**0.50**	**0.73**	**0.70**	**0.22**
max[k]	1.00	1.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
av[l]	0.94	0.96	**0.95**	**0.96**	**0.80**	**0.82**	**0.93**	**0.93**	**0.84**

[a] SPE. [b] SPE with boosting. [c] Rubicon. [d] Catalyst. [e] Catalyst with poling. [f] Omega. [g] Macromodel. [h] MOE stochastic. [i] MOE systematic. [j] Minimum over all molecules. [k] Maximum over all molecules. [l] Average over all molecules. [m] Bold font indicates values ≤ 0.75.

This range is reduced sharply upon minimization (Figure 2), indicating raw (nonminimized) geometries that are far from local minima. A dramatic example is 1CBX (Figure 1/2a), a small molecule with limited conformational flexibility, where Macromodel produced a geometry three times as extended as those identified by the other methods. As shown in Figure 4, this conformation is severely distorted, most likely as a result of low-mode sampling. By looking at these geometries and considering the significant computational overhead associated with computing the eigenvectors, it is

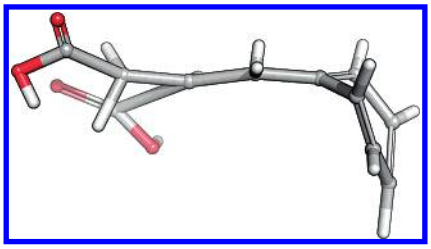


Figure 4. Most extended raw conformation produced by Macromodel for 1CBX (prior to minimization).

unclear whether this form of sampling offers any significant advantages over much simpler and faster procedures, such as Saunders' stochastic search. Macromodel relies primarily on energy minimization to generate chemically sensible geometries, an approach that we believe is computationally inefficient. Energy minimization did not materially affect the distributions produced by the other methods except in rare cases (vide infra). Interestingly, minimization did have a minor impact on the conformations derived from MOE stochastic search, which were already minimized in MOE using the same force field. This is probably due to different minimizer thresholds and to different default MMFF94s parameters in MOE. In contrast to our own and other commercial implementations, we have found that the default settings in MOE do not reproduce the energies in the MMFF94s validation suite by a wide margin.[31]

Of all the other methods, Catalyst emerges as the most competitive to boosting SPE. It covers an equally broad range of geometric sizes and identifies comparable minima and maxima in terms of extendedness for most molecules. Poling works as expected, producing just as broad of a sampling but with significantly fewer conformations. Omega shows a strong preference for extended conformations, in contrast to MOE's stochastic search, which is comparable to conventional SPE in favoring compact geometries. We should note, parenthetically, that we had considerable difficulty getting Omega to work for many of these molecules and failed to generate any conformers for two ordinary compounds, 1ICN (Figure 1/2d) and 2PLV (Figure 1/2e).[44] MOE's systematic search yields lesser diversity, has a distinct preference for extended conformations, and fails to sample large regions of space on the compact end of spectrum, particularly for long, linear molecules (1B1H in Figure 1/2b, 1AAQ in Figure 1/2b, 1EED in Figure 1/2c, 1EPO in Figure 1/2c, 1ICN in Figure 1/2d, 1LIC in Figure 1/2d, 2PLV in Figure 1/2e, and 5P2P in Figure 1/2f). Despite its preference for extended geometries, MOE systematic search fails to produce conformations that are as extended as those discovered by boosting SPE (e.g., 1APT in Figure 1a, 1HVL in Figure 1b, 1AAQ in Figure 1b, 1HEF in Figure 2c, 1RNE in Figure 2e, and most notably 1EPO in Figure 2c, 1IDA in Figure 2d, 1SME in Figure 2e, and 4DFR in Figure 2f).

A particularly intriguing case is 1ETA (Figure 2c). Although this molecule has limited conformational flexibility, SPE and Macromodel produced two conformations that were substantially more compact than those identified by the other methods (Figure 6). These highly compact structures reveal a problematic, though extremely rare, scenario: a bond penetrating a ring system, yielding a structure that cannot be corrected by energy minimization. These local minima are of very high energy and can be easily detected and eliminated from further consideration either before or after minimization.

The heatmaps in Figures 1 and 2 also highlight as a reference point the radius of gyration of the bioactive conformation (red) and the lowest energy conformation identified by each method (green). These plots confirm the previous observation[2,21] that bound conformations tend to be more extended than random ones. This is true for most methods regardless of their individual preference for compact or extended geometries, though not without exceptions. For example, the bioactive conformation of 5P2P (Figure 1/2f) is significantly more extended than the majority of the conformations generated by regular SPE and stochastic MOE, significantly more compact than those generated by Rubicon, Omega, and systematic MOE, and of about average size for

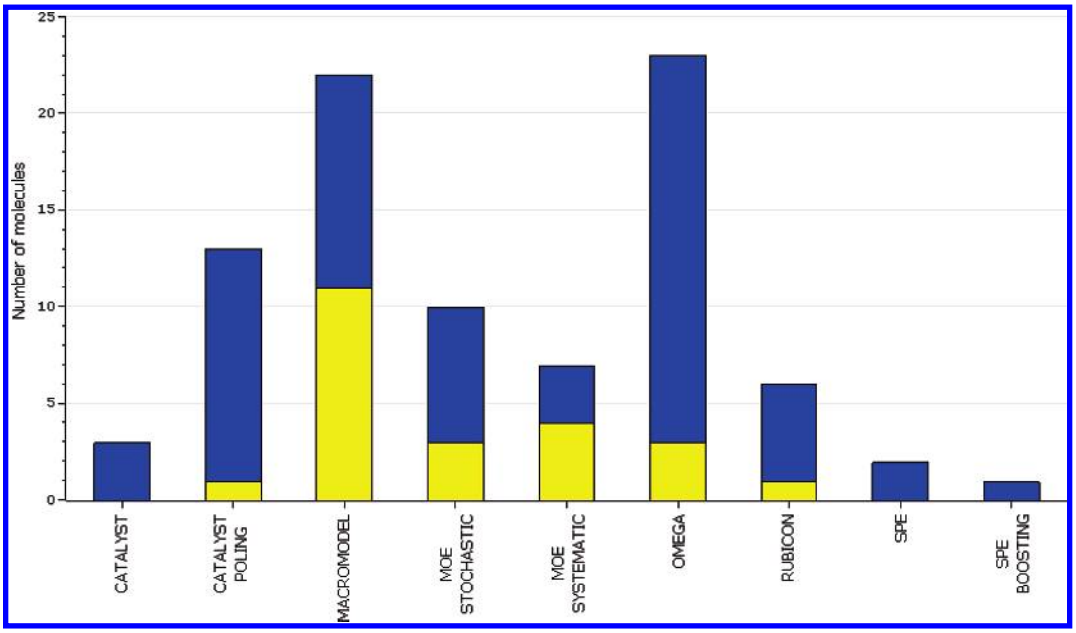


Figure 5. Number of molecules for which a particular method failed to produce a conformation within 5 (blue) and 10 (yellow) kcal/mol from the global minimum.

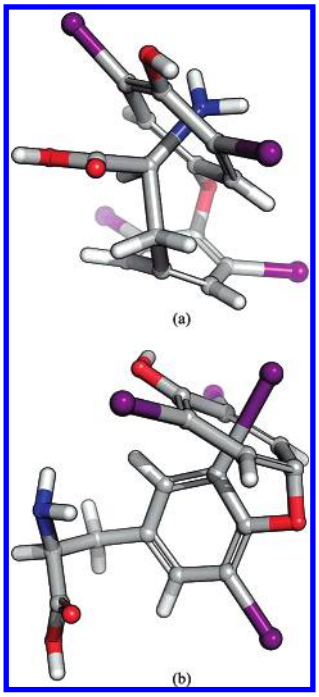


Figure 6. Most compact minimized conformation for 1ETA produced by (a) boosting SPE and (b) Macromodel.

the remaining methods. Similarly, for 1B5H (in Figure 1/2a), Macromodel tends to produce more extended geometries relative to the bound conformation, in contrast to the other methods. In some cases, the bioactive conformation lies outside the range sampled by some methods, both preminimization (1LIC with SPE in Figure 1d, 1PPC with Rubicon in Figure 1e, 1SME with MOE systematic in Figure 1e, 2PLV with SPE and Rubicon in Figure 1e, and 2RO7 with SPE in Figure 1e) and postminimization (1HEF with Macromodel in Figure 2c, 1HRI with SPE in Figure 2c, and 2PLV with SPE in Figure 2e). Although these are exceptions, the question of what constitutes a "random" conformation remains open; the answer is clearly method dependent.

In contrast to the bound state, the lowest-energy conformers tend to be more compact due to intramolecular interactions, which are favored in the gas phase but can be offset by interactions with the active site of the receptor. SPE (with or without boosting) consistently identifies the lowest or nearly the lowest energy structure for most of the molecules in this study (Figure 3). While poling appears to be an excellent method for exploring the diversity of the energy landscape, it is not the most suitable approach for finding the lowest energy minima, as those located in the vicinity of visited states are eliminated from further consideration through the artificial poling potential (and the core algorithm does not intrinsically favor the lowest energy state over alternative conformations in that same neighborhood). MOE's systematic search is very effective in identifying low-energy minima but not as consistently as SPE and Catalyst (e.g., 1B5H in Figure 3a, 1HVJ and 1HVL in Figure 3b, 1EED and 1EPO in Figure 3c, 1IDA in Figure 3d, and 1RNE and 1SME in Figure 3e). MOE systematic search and Omega focus on similar regions of extendedness, but MOE samples geometries of considerably lower energy (Figure 3).

The number of molecules for which each method failed to produce a conformer within 5 (blue) and 10 (yellow) kcal/mol from the global minimum (here defined as the lowest-energy structure identified by any method) is illustrated in Figure 5. Boosting SPE failed to identify a conformation within 5 kcal/mol from the global minimum in only one case, and, even then, the best solution was only 5.8 kcal/mol higher in energy. Regular SPE and Catalyst were equally impressive, producing at least one conformation within 6.5 and 8.0 kcal/mol from the global minimum for every single molecule. Rubicon, MOE systematic, MOE stochastic, and poling Catalyst showed intermediate performance, while Macromodel and Omega failed to produce a competitive low-energy solution (within 5 kcal/mol) in nearly 40% of the molecules. The largest energy difference between the global minimum and the lowest energy structure identified by each method across all molecules was 10 kcal/mol for poling Catalyst, 23.8 kcal/mol for Macromodel, 28.1 kcal/mol for MOE systematic, and 14.8 kcal/mol for Rubicon, Omega, and MOE stochastic.

SPE and Catalyst also produced the greatest pharmacophore diversity as measured by the number of unique 3-point 3D pharmacophores found in the conformational ensembles generated by each method (Tables 3 and 4). Two of the molecules, 1ICN and 1TNI, had only 2 pharmacophore centers (the two oxygens on the carboxylic acid moiety in 1ICN and the nitrogen atom and aromatic ring in 1TNI) and therefore no 3-point pharmacophores, while one molecule, 1LIC, had a single 3-point pharmacophore defined by the three oxygen atoms on the conformationally rigid sulfonic acid. Boosting SPE produced the richest set of pharmacophore triplets for nearly one-third of the molecules and more than 90% of the triplets for the vast majority of the remaining

Table 5. General Trends of the Nine Conformational Search Algorithms Examined[j]

performance criterion	SPE[a]	SPE_B[b]	RUB[c]	CAT[d]	CAT_P[e]	OMG[f]	MAC[g]	MOE_STO[h]	MOE_SYS[i]
sampling of compact conformations (measured by radius of gyration)	***	***	**	***	**	**	*	***	*
sampling of extended conformations (measured by radius of gyration)	**	***	**	**	***	**	***	**	**
identification of low-energy minima (measured by MMFF94s)	***	***	**	***	*	**	*	**	***
failure to identify conformers within 5−10 kcal/mol from the global minimum (measured by MMFF94s)	***	***	**	***	*	*	*	**	**
pharmacophoric diversity (measured by the number of unique 3-point 3D pharmacophores found in the ensemble).	**	***	**	***	*	*	**	**	*

[a] SPE. [b] SPE with boosting. [c] Rubicon. [d] Catalyst. [e] Catalyst with poling. [f] Omega. [g] Macromodel. [h] MOE stochastic. [i] MOE systematic. [j] The number of asterisks represents the relative performance of the algorithm against the listed criteria.

structures. The few exceptions (e.g., 1HFC, 3TPI) represent small molecules with a limited number of accessible conformations, where one or two missing conformations may have contributed significantly to the relative number of triplets. Poling had a significant impact on Catalyst's performance, reducing the number of 3-point pharmacophores by nearly 20%. Poling Catalyst produced the fewest pharmacophore triplets in most of the cases, which is not surprising given the relatively high resolution of our pharmacophore bins (1 Å) and the sparse sampling that is inherent in the poling algorithm. Of the remaining methods, Omega exhibited the worst performance, producing the lowest number of triplets in 31 out of 59 cases. For some small molecules, the number was strikingly smaller compared to the other methods (e.g., 1CBX, 1FKG, 2LGS), which indicates that several conformations are missing from the respective ensembles. Equally striking are some of the results obtained with MOE systematic search (e.g., 1EPO, 1IDA, 1RNE), where the number of triplets was one-third to one-fifth of those discovered by other methods. The remaining methods showed intermediate performance.

Although the precise number of 3-point pharmacophores depends on the atom type definitions and distance cutoffs, the qualitative trends are likely to hold. This metric is important in applications like docking, 3D database searching, and 3D QSAR, which depend critically upon the number and diversity of the pharmacophores that are being considered during the search. For example, given that a ligand may bind in a conformation that may be several kcal/mol higher in energy than the global minimum, it is imperative that a 3D database be populated with the most pharmacophorically rich set of conformations within a reasonable energy window, in order to minimize the number of false negatives.

The general trends of the nine conformational search algorithms examined as they pertain to the sampling of compact and extended conformations, the number of 3-point pharmacophores generated, the ability to identify low-energy minima, and the failure to identify conformers within 5−10 kcal/mol from the global minimum are summarized in Table 5.

CONCLUSIONS

The results outlined above demonstrate that boosting greatly increases the sampling capacity of SPE and renders it an extremely competitive method for conformational analysis of small organic molecules. Boosting is a general strategy that is likely to benefit any distance geometry method. It can be easily implemented in existing distance geometry codes and provides a way of sampling regions of conformational space that may not be easily accessible by alternative methods. Since bioactive conformations tend to be extended and often fall outside the range sampled by an unbiased search, this heuristic significantly improves the chances of finding such conformations.

ACKNOWLEDGMENT

We thank Dr. David J. Diller of Pharmacopeia, Inc. for providing the data set and Dr. Huafeng Xu of D. E. Shaw & Co. for his earlier work on SPE.

Note Added after ASAP Publication. This article was released ASAP on April 6, 2007, with the wrong journal abbreviation in ref 2. The correct version was posted on April 11, 2007.

REFERENCES AND NOTES

(1) Leach, A. R. A 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; Vol. 2.
(2) Diller, D. J.; Merz, K. M., Jr. Can we separate active from inactive conformations? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 105−112.
(3) 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.
(4) Lipton, M.; Still, W. C. The multiple minimum problem in molecular modeling. Tree searching internal coordinate conformational space. *J. Comput. Chem.* **1988**, *9*, 343.
(5) Bruccoleri, R. E.; Karplus, M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers* **1987**, *26*, 137.
(6) Bruccoleri, R. E.; Karplus, M. Chain closure with bond angle variations. *Macromolecules* **1985**, *18*, 2767.
(7) Go, N.; Scheraga, H. A. Ring closure and local conformational deformations ofchain molecules. *Macromolecules* **1970**, *3*, 178.
(8) Saunders, M. Stochastic explorations of molecular mechanics energy surfaces. Hunting for the global minimum. *J. Am. Chem. Soc.* **1987**, *109*, 3150.
(9) 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.
(10) 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.
(11) Crippen, G. M. Rapid calculation of coordinates from distance matrices. *J. Comput. Phys.* **1978**, *26*, 449.
(12) 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.
(13) 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.
(14) Bostrom, J.; Greenwood, J. R.; Gottfries, J. *J. Mol. Graphics Modell.* **2003**, *21* (5), 449-62.
(15) Bostrom, J. *J. Comput.-Aided Mol. Des.* **2001**, *15* (12), 1137−52.
(16) Feuston, B. P.; Miller, M. D.; Culberson, J. C.; Nachbar, R. B.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 754−63.
(17) Perola, E.; Charifson, P. S. *J. Med. Chem.* **2004**, *47* (10), 2499−510.
(18) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869.
(19) Agrafiotis, D. K. Stochastic proximity embedding. *J. Comput. Chem.* **2003**, *24*, 1215−1221.
(20) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational sampling by self-organization. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1186−1191.
(21) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A distance geometry heuristics for biasing conformational sampling towards extended or compact conformations. *J. Comput. Chem.* **2006**, *27* (16), 1962−1969.
(22) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235.
(23) Berman, H. M.; Bhat, T. N.; Bourne, P. E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J. The PDB and the challenge of structural genomics. *Nat. Struct. Biol.* **2000**, *7*, 957.
(24) "Blockbuster drugs (over US$ 1bn sales)." Retrieved Nov. 14, 2006, from Thomson Pharma, Guided Search − Competitive Intelligence − Drugs (www.thomson-pharma.com).
(25) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 616.
(26) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 490.
(27) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 520.
(28) Halgren, T. A. *J. Comput. Chem.* **1996**, *17*, 553.
(29) Halgren, T. A.; Nachbar, R. B. *J. Comput. Chem.* **1996**, *17*, 587.
(30) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System and method for automatically generating chemical compounds with desired properties. United States patents 5,463,564; 5,574,656; 5,684,-711; 5,901,069; 6,421,612; 6,434,490.
(31) http://www.ccl.net/cca/data/MMFF94s/.
(32) *Rubicon*; Daylight Chemical Information Systems, Inc., 120 Vantis, Suite 550, Aliso Viejo, CA 92656, U.S.A. http://www.daylight.com.

(33) Crippen, G. M.; Havel, T. F. *Acta Crystallogr.* **1978**, *A34*, 282
(34) *Catalyst*; Accelrys, Inc., 9685 North Scranton Road, San Diego, CA 92121, U.S.A. http://www.accelrys.com.
(35) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting conformational variation. *J. Comput. Chem.* **1995**, *16*, 171−187.
(36) *Omega*; OpenEye Scientific Software, 3600 Cerrillos Rd., Suite 1107, Santa Fe, NM 87507, U.S.A. http://www.eyesopen.com.
(37) *Macromodel;* Schrodinger, Inc., 1500 S. W. First Avenue, Suite 1180, Portland, OR 97201, U.S.A. http://www.schrodinger.com.
(38) Chang, G.; Guida, W. C.; Still, W. C. *J. Am. Chem. Soc.* **1989**, 111, 4379−4386.
(39) Kolossvary, I.; Guida, W. C. *J. Am. Chem. Soc.* **1996**, *118*, 5011−5019.
(40) Ponder, J. W.; Richards, F. M. *J. Comput. Chem.* **1987**, *8*, 1016.
(41) *MOE*; Chemical Computing Group, 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec, Canada H3A 2R7. http://www.chemcomp.com.
(42) Ferguson, D. M.; Raber, D. J. *J. Am. Chem. Soc.* **1989**, *111*, 4371−4378.
(43) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.
(44) Openeye Scientific was not contacted to resolve this problem.
(45) *SMARTS*; Daylight Chemical Information Systems, Inc., 120 Vantis, Suite 550, Aliso Viejo, CA 92656, U.S.A. http://www.daylight.com.

CI6005454