Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening

I-Jen Chen* and Nicolas Foloppe*

Vernalis (R&D) Ltd., Granta Park, Abington, Cambridge CB21 6GB, U.K.

Received April 13, 2008

Computational conformational sampling is integral to small molecule pharmaceutical research, for detailed conformational analysis and high-throughput 3D library enumeration. These two regimes were tested in details for the general-purpose modeling program MOE, using its three conformational sampling methods, i.e. systematic search, stochastic search, and Conformation Import. The tests include i) identification of the global energy minimum, ii) reproduction of the bioactive conformation, iii) measures of conformational coverage with 3D descriptors, and iv) compute times. The bioactive conformers are from a new set of 256 diverse, druglike, protein-bound ligands compiled and analyzed with particular care. The MOE results are compared to those obtained from the established program Catalyst. Key parameters controlling the conformational coverage were varied systematically. Coverage and diversity of the conformational space were characterized with unique pharmacophore triplets or quadruplets. Overall, the protocols in both MOE and Catalyst performed well for their intended tasks. MOE performed at least as well as Catalyst for high-throughput library generation and detailed conformational modeling. This work provides a guide and specific recommendations regarding the usage of conformational sampling tools in MOE.

1. INTRODUCTION

Conformational sampling is fundamental to computational chemistry, including the discovery and optimization of small organic compounds of pharmaceutical interest. It underpins three-dimensional (3D) alignments of bioactive small molecules,[1−3] pharmacophore derivation,[4] or 3D QSAR.[5] Many conformations of large numbers of candidate compounds must be produced for virtual screening (VS), including ligand-based pharmacophore searches. VS frequently relies on conformational libraries generated prior to a search being performed, and the outcome depends on the quality of these conformational models.

Many techniques have been developed to generate conformations of small organic molecules.[6,7] Each method requires an algorithm for conformation generation and a scoring scheme to assess the likelihood of the conformers. Early 3D conformation builders such as CONCORD[8] and CORINA[9] were rule-based and produce a small number of conformers per compound. More detailed conformational search may use distance geometry constraints,[10] systematic torsional grids,[11] stochastic moves,[12] assembly of constitutive fragments,[13] directed tweak,[14] genetic algorithms,[1] poling restraints,[15] Tabu search,[16] Monte-Carlo,[17] or even molecular dynamics in explicit solvent.[18] The likelihood of existence of a conformer may be estimated from molecular mechanics force fields.[19−23] In addition to conformers, one would ideally need to generate all stereoisomers and relevant tautomers.[24]

Software to generate conformational libraries for large number of compounds include Catalyst,[25] Omega,[26] Phase,[27] and MOE (Molecular Operating Environment).[28] Such programs need to be computationally efficient, while producing good quality conformational models. Useful models must provide a broad conformational coverage, representing each compound by diverse, nonredundant, and physically plausible conformers.[29] In particular, a key test of a conformational sampling protocol is whether it "reproduces" the experimentally determined bioactive conformation of ligands bound to their targets,[30−33] typically obtained from X-ray crystallography. When evaluated with force fields, the ligands bioactive conformational energies appear significantly strained,[32−35] although it is counterintuitive and a matter for debate.[35−37] This question is complicated by several sources of uncertainty, including limitations in the force field parameters, possible errors in the crystallographic models,[31,38] and definition of the reference unbound state relative to which the strain energy should be calculated.[36,37]

Therefore, the maximum allowed intramolecular conformational energy (ΔE in kcal/mol) is a key parameter when producing conformational models. Other parameters include the maximum number of conformers per compound (MaxConfs) and the optional use of continuum solvation models or energy minimization. A balance between permitted intramolecular energies and conformational diversity needs to be found. With too low an energy threshold one may miss bioactive conformations, but higher thresholds imply the risk of adding noise with irrelevant high-energy conformers, and increased computational resources to handle more conformers. The parameters optimal for conformational coverage need to be estimated empirically, because they depend on the energy model and the algorithm generating the conformations.

Studies of conformational sampling by programs Catalyst,[31−33,39] Confort,[31] Flo99,[31] Macromodel,[31,39] MOE,[39]

* Corresponding author phone: + (44) (0) 1223 895 555; fax: + (44) (0) 1223 895 556; e-mail: i.chen@vernalis.com (I.C.) and phone: + (44) (0) 1223 895 338; fax: + (44) (0) 1223 895 556; e-mail: n.foloppe@vernalis.com.

10.1021/ci800130k CCC: $40.75
Published on Web 09/03/2008

Omega,[30,31,33,39] and Rubicon,[39] for example, have already been presented. Tests in high-throughput mode for Catalyst[32,33] and Omega[30,33] led to specific and helpful recommendations regarding the usage of these packages. Such tests are not available for the widely used program MOE, a general-purpose molecular modeling package for medicinal chemistry,[28] with functionalities for library design and pharmacophore modeling and search. The three conformation generators in MOE for small organic molecules are systematic search, stochastic search, and Conformation Import (details in section 2.2). The systematic and stochastic searches are intended for detailed conformational analysis, while Conformation Import is geared toward high-throughput library enumeration. These algorithms were carefully tested in the present work, with respect to i) identification of the global energy minimum, ii) reproduction of the bioactive conformation, iii) general measures of conformational coverage with 3D descriptors (conformer compactness, pharmacophore triplets and quadruplets), and iv) compute times. The emphasis is on protocols aimed at library generation and the parameters controlling conformational sampling. The MOE results are compared to their counterpart from the established Catalyst FAST and BEST protocols, on the same set of test compounds.

The present effort includes the compilation of a set of diverse test compounds with an X-ray bioactive conformation and druglike properties representative of pharmaceutical research. These stringent criteria led to a set of 256 compounds, smaller than some recent sets[32,33] used to test Catalyst, but still larger than most sets used to investigate the conformational properties of bioactive conformers.[31,34,35,37,39−41] An analysis of the compound properties provides a well-defined context to interpret the conformational sampling results.

We confirm the trends reported previously with Catalyst[32,33] but in a somewhat different light with the less flexible compounds selected here. Catalyst FAST and BEST performed well for their intended tasks, and MOE performed at least as well for high-throughput library generation and detailed conformational modeling. The MOE stochastic search with a generalized Born (*GB*) solvation model reproduced the highest percentage of bioactive conformers, providing surrogate bioactive-like conformers to examine the energetics of the bioactive conformers. Overall, this work provides a detailed guide and specific recommendations regarding the usage of conformational sampling tools in MOE.

2. METHODS

2.1. Preparation of Compound Sets. Three ligand sets were selected, containing only ligands noncovalently bound to proteins. The first two sets were from previous studies[31,35] which also analyzed the ligand bioactive conformations but without testing the software MOE[28] which is central to the present work. The first ligand set, referred to as the Bostrom set,[31] consists of 32 ligands extracted from high-resolution (≤2.0 Å) protein−ligand complexes from the Protein Data Bank (PDB),[42] where most ligands had relatively low temperature factors (<30 Å^2). The second set, the modified-Perola set, was derived from an original ensemble of 150 ligands,[35] 100 are publicly available from the PDB and the remaining 50 are part of the Vertex proprietary collection. Only the 100 public structures were included in the modified-Perola set. This set compiled ligands with a crystallographic resolution <3 Å and available binding constants. In this second set, the ligand molecular weight (MW) ranged between 200 and 600. The number of rotatable bond (NRot) range was 1−11 in the Bostrom set and 1−12 in the modified-Perola set. Ligands from both sets were structurally diverse, relevant to pharmaceutical chemistry, and target multiple protein classes. There are 6 ligands in common between the Bostrom and modified-Perola sets.

The steady increase in protein−ligand complexes in the PDB allowed for the addition of a third set, significantly extending the number and diversity of the test compounds. Some recent studies compiled remarkably large compound test sets,[32,33] partly to represent the ligand property distributions in the PDB.[32] However, we wanted to adhere to the more stringent druglike criteria typical of industrial pharmaceutical research. Thus, the protein−ligand complexes not already in the Bostrom/modified-Perola sets were retrieved from the PDB. Only structures resolved at 2.5 Å resolution or better were kept, a reasonable cutoff.[38] According to their size and flexibility, compounds markedly outside the conventional druglike space were removed, while retaining molecules complex enough to challenge the tested softwares. This led to a tradeoff where overly flexible (MOE descriptor for number of rotatable bonds b_1rotN ≥ 12) or large (MW > 600) ligands were removed (similar criteria used by Perola et al.[35]). To select chemically diverse molecules, compounds were clustered (MACCS fingerprints from MDL,[43] Tanimoto similarity coefficient of 90%) and one member per cluster kept. This provided 130 additional ligands crystallized with a range of protein families including kinases, ATPases, estrogen-binding receptors, HIV protease, and reverse transcriptase (Table S1 in the Supporting Information). These 130 ligands constitute the Vernalis set. Altogether there were 256 unique ligands across the three compound sets. Table 1 provides the average properties for each set.

All ligands were initially prepared with MOE, including assignment of bond orders and standard protonation states at pH=7, which were manually checked. The tautomer states were selected to be consistent with the X-ray binding modes. Prior to conformational analysis with the programs MOE 2006.08 and Catalyst 4.10, all ligands were converted to 2D representations to erase the initial conformational information.

2.2. MOE Conformational Search Algorithms. MOE provides three methods to generate conformers of small organic molecules: systematic search, stochastic search,[12] and a high-throughput protocol called Conformational Import. These algorithms, being extensively tested in this study, are summarized alongside their key default parameters, which were used unless specified otherwise.

The systematic search generates all combinations of rotatable torsions by discrete increments. The default step size is 15 degrees for cyclic single bonds. For noncyclic bonds, it is 60 degrees in general and 120 degrees for bonds connecting two sp3 carbons. Prior to output, each conformation is subjected to a bump check to remove conformers with severe intramolecular clashes, i.e. interatomic van der Waals contact energy greater than 10 kcal/mol. There is an option to energy-minimize the conformers generated by the systematic search. Chiral centers may be inverted as a result of

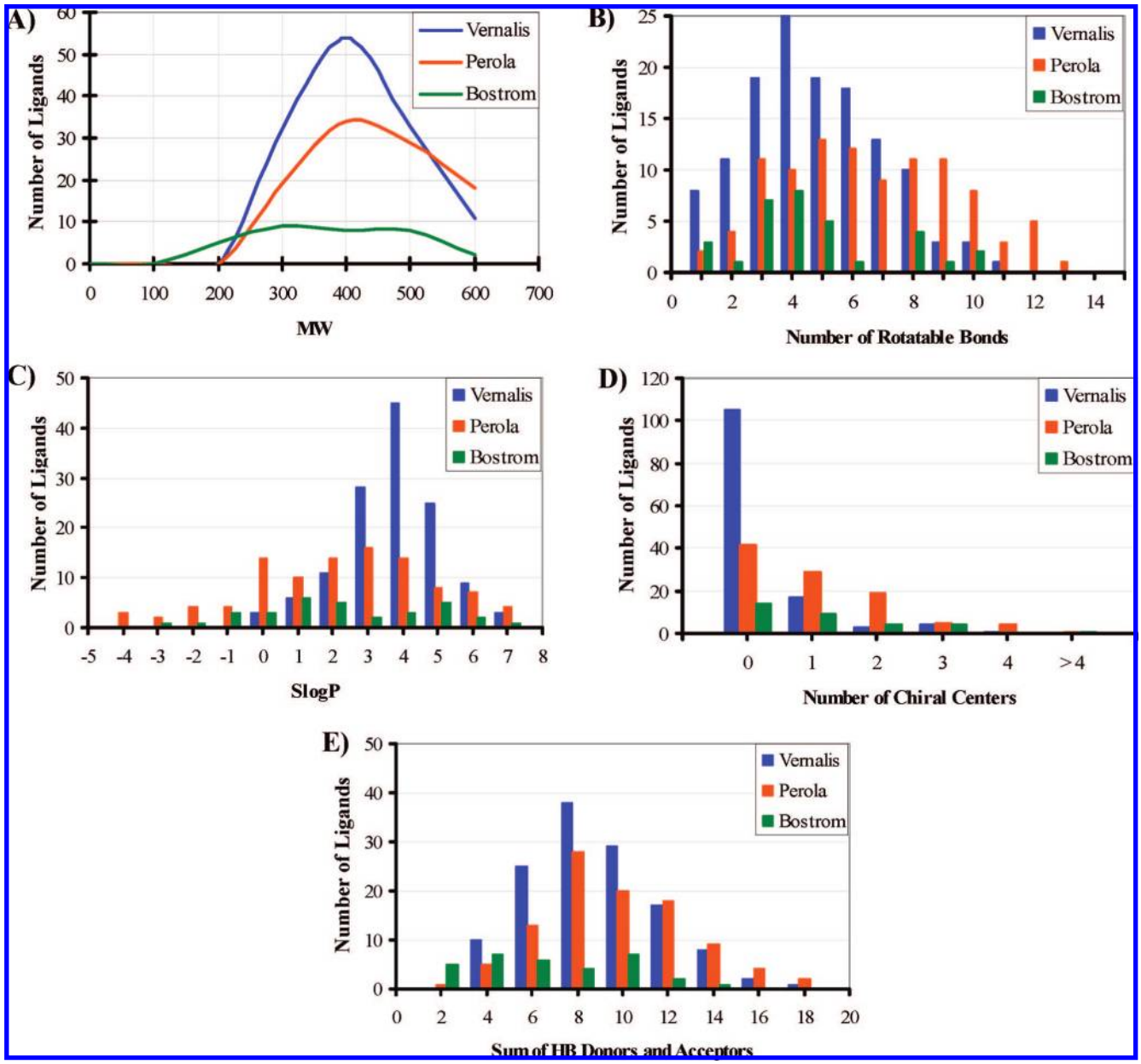

Figure 1. Distributions of selected physicochemical properties for the Vernalis (blue), Bostrom (green), and modified-Perola (orange) sets of compounds. The number of ligands is plotted versus the molecular weight (panel A), the number of rotatable bonds (panel B), the calculated n-octanol/water partition coefficient SlogP (panel C), the number of chiral centers per compound (panel D), and the sum of the number of hydrogen-bond acceptors and donors per compound (panel E).

Table 1. Selected Average Properties of the Bostrom, Modified-Perola, and Vernalis Compound Sets

compound set	MW[a]	NRot[b]	SlogP[c]	chiral centers[d]	donors + acceptors[e]
Bostrom	322.60	4.69	1.69	1.09	6.09
modified-Perola	398.00	6.56	1.71	1.03	9.09
Vernalis	369.49	4.84	3.27	0.29	8.42

[a] Molecular weight. [b] Number of rotatable bonds. [c] n-octanol/water partition coefficient calculated with MOE. [d] Number of chiral centers per compound. [e] Sum of the numbers of hydrogen-bond donors and acceptors per compound.

clashes during energy minimization, but otherwise are not varied. After minimization, only conformers within 10 kcal/mol of the lowest energy are kept, and duplicate conformers are removed (rmsd $\leq$ 0.1 Å after optimal superposition).

The stochastic search generates conformations by randomly sampling the potential energy surface, using an algorithm similar to the Random Incremental Pulse Search.[12] It includes inversions of unconstrained chiral centers, together with random perturbation of the rotatable bonds (including rings) in increments biased around 30 degrees, followed by energy minimization. Two parameters control the number of output conformers: i) the maximum number of retained conformers (default: 10000), called "Conformation Limit" in MOE but referred to as MaxConfs here for consistency across protocols, and ii) the predefined number of attempts at generating new conformers (default: 10000), called

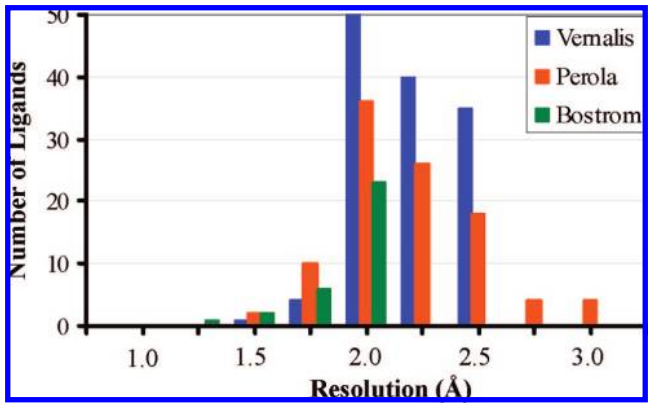

Figure 2. Distribution of the crystallographic resolutions (Å) of the protein−ligand complexes included in the Vernalis (blue), Bostrom (green), and modified-Perola (orange) sets of compounds. The frequencies are shown for resolution bins of 0.25 Å.

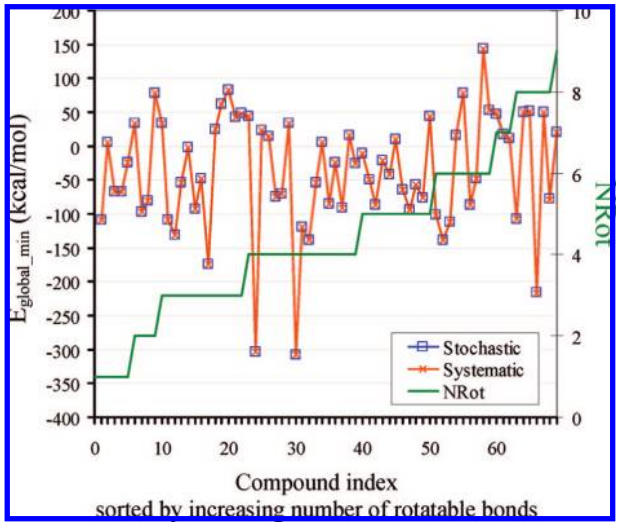

Figure 3. Energies (kcal/mol; left Y axis) of the lowest energy minima obtained by systematic (orange crosses) and stochastic (blue squares) searches with MOE, for 70 selected compounds. On the X axis, the ligands are represented by an individual index number (1 to 70), sorted by increasing number of rotatable bonds (NRot). NRot for each ligand is also plotted as a solid green line, from 1 to 9 (right Y axis). Apart from the use of a *GB* solvation model, these data were obtained at default settings. The lowest energies from systematic and stochastic searches coincide for the vast majority of the compounds. Virtually identical results were obtained with other stochastic searches (section 3.2).

"Iteration Limit". New conformers are kept if they are within an energy cutoff (default: 7 kcal/mol) of the lowest energy structure and if they differ enough (rmsd ≥ 0.1 Å) from previously generated structures. The energy cutoff is applied as the search proceeds, which can influence how higher energy regions are sampled.

Conformational Import is a high-throughput method to generate conformers for large numbers of molecules, emphasizing speed by using pregenerated conformations for template fragments. This approach is recommended when producing multiconformer 3D libraries for pharmacophore search with MOE. The Conformational Import process is summarized in Figure 4. The input molecules are first "washed" to assign protonation states at pH=7 and remove counterions. Preliminary filtering may be applied to the washed molecules based on physicochemical properties (e.g., MW and number of rotatable bonds) and predefined SMARTS

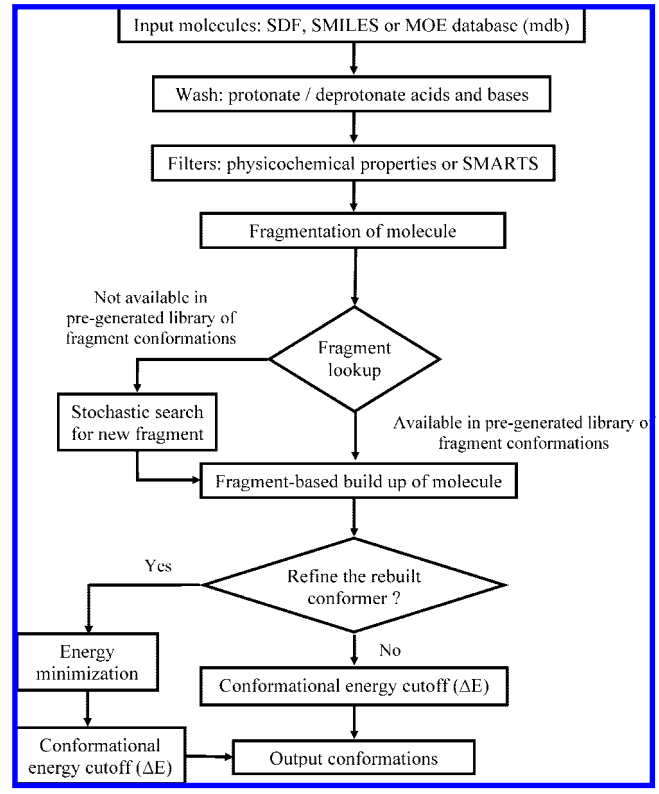

Figure 4. Flowchart summarizing the Conformational Import protocol in MOE.

(e.g., reactive groups). These filters were turned off because selection criteria were already applied to the compound sets. Then, each molecule is divided into overlapping fragments. A pregenerated library of fragment conformations is searched for each fragment. If a fragment cannot be found, its conformations are generated on the fly by the stochastic search described above. Once the conformations for all the constitutive fragments are available, the complete conformers are assembled by systematic combinations of the 3D fragments. This allows the chirality of stereocenters to vary. Complete conformations are rejected if they contain poor van der Waals contacts or unfavorable geometries (e.g., cis amide). The reported strain energy is the sum of the individual fragment contributions and is not equivalent to a typical force field energy, e.g. it does not include interfragment nonbonded energies. By default, no further refinement (energy minimization) is performed, and conformers with a strain energy >4 kcal/mol are discarded. A more complete force field conformational energy may be calculated after energy minimization (refinement option), but the energy window remains 4 kcal/mol by default. A conformation is output to the final database if i) its estimated strain energy is less than a user-defined threshold (4 kcal/mol by default) and ii) the conformation is different from those found previously (rmsd ≥ 0.15 Å by default). The process continues until a specified maximum number of conformations per ligand (250 by default) is reached or no more new conformations can be found. The effect of varying the parameters of Conformation Import was tested in detail.

2.3. Search for the Global Energy Minimum with MOE. The ability of each conformation generation method in MOE (systematic search, stochastic search, and Conformation Import) to locate the global energy minimum was tested, because it provides an energy reference important for

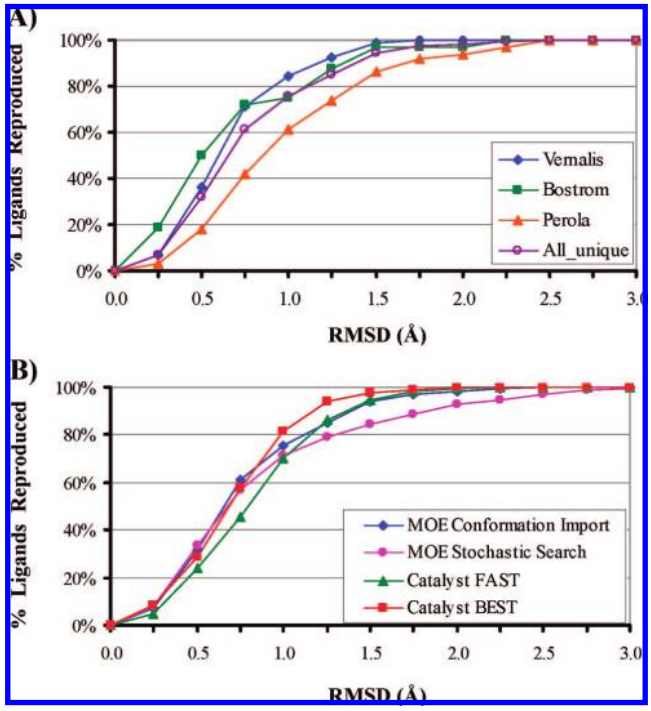

Figure 5. Tests of how well the X-ray bioactive structure of ligands is reproduced by computational methods at their default settings. The bioactive reference is deemed "reproduced" by a computed conformer when it is within a defined rmsd (Å) of this reference. As this rmsd cutoff is increased, more compounds have at least one generated conformer which "reproduces" the bioactive reference. Panel A: cumulative frequencies for the percent of reproduced bioactive structures with Conformation Import versus the rmsd cutoff, for various compound sets: the Vernalis set (blue diamonds), Bostrom set (green squares), modified-Perola set (orange triangles), and combination of all 256 unique compounds (All_unique, open magenta circles). Panel B: cumulative frequencies of the percent of reproduced bioactive structures (all 256 compounds) for MOE Conformation Import (blue diamonds), MOE stochastic search (magenta circles; distance-dependent dielectric constant), Catalyst FAST (green triangles), and Catalyst BEST (red squares).

all aspects of conformational analysis. The Merck Molecular Force Field (MMFF94) developed for small organic molecules[21] combined with the generalized Born (*GB*) model of aqueous solvation implemented in MOE[44] was used. By definition, the systematic search is expected to locate the global energy minimum for a given potential.[7,11] Therefore, the lowest energy minimum from the systematic search was taken as the global energy minimum reference, to which the lowest energy conformers from the stochastic search and Conformation Import were compared.

Since the configuration at chiral centers is varied explicitly in the stochastic search and Conformation Import but may be inverted only erratically during a systematic search (when energy minimizing), the tests for the global energy minima were only performed with 70 achiral compounds. This subset afforded a reasonable coverage of NRot, from 1 to 9 (Figure 3). The systematic search was initiated following 2D to 3D conversion, and the output conformers were energy-minimized. Other parameters were at default.

The stochastic searches were performed on the same 70 compounds with energy cutoffs of 7 (default), 12, 15, and 20 kcal/mol, with the same energy model. Other parameters were at default. Each search was initiated from 2D structures. Thus, four independent stochastic searches were run to test the influence of the energy cutoff and the consistency of the results, given the stochastic element. The lowest energy conformers from each stochastic search were compared against each other and to the global energy minimum from the systematic search.

The lowest energy conformations from Conformation Import were also compared to the global energy minima. In this context, we only considered Conformation Import with refinement of the conformers by energy minimization, with the same potential as used with the other methods (MMFF94 with *GB* solvation model). Other default parameters were kept (maximum 250 conformers per compound, $\Delta E = 4$ kcal/mol).

The energy minima were compared in terms of their energies and structural deviations (rmsd, taking symmetries into account).

2.4. Protocols To Generate Multiconformer Libraries. Thorough conformational analysis toward the identification of energy minima is only one aspect of conformational sampling relevant to pharmaceutical research. Frequently, one is more interested in conformationally diverse 3D libraries of compounds, e.g. for pharmacophore-based screening. This needs to handle a large number of compounds in high-throughput mode. In that perspective, we tested four conformation generators: i) MOE Conformational Import, ii) MOE stochastic search, iii) Catalyst FAST, and iv) Catalyst BEST. Each protocol was tested on the selected 256 ligands, with a known bioactive conformation (see above).

Key parameters influencing the conformational models and under the user control were varied. With the methods in MOE these parameters were i) the maximum number of conformations per compound (MaxConfs), ii) the allowed conformational energy window (ΔE in kcal/mol), iii) refinement or not of the conformers by energy minimization, and iv) the use or not of a solvation model. With Catalyst only MaxConfs and ΔE could be varied.

For these tests, Table 2 summarizes the default settings, the range of additional tested values, and their combinations. Apart from its default settings (no refinement, $\Delta E = 4$ kcal/mol, MaxConfs = 250), MOE Conformation Import was tested with 23 other protocols which would *a priori* allow for greater conformational coverage ($\Delta E > 4$ kcal/mol, MaxConfs >250). We also tested the effect of energy minimization, with both the default MOE distance-dependent dielectric constant and a *GB* solvation model. One motivation was to see if it would be worth spending the additional computational resources inherent to these 23 protocols. Additionally, we investigated 2 Conformation Import protocols more economical than the defaults, reducing ΔE to 3 kcal/mol or MaxConfs to 100.

Conformation Import is the recommended protocol for generation of conformational libraries in MOE. Still, we compared its output to that of stochastic searches. In addition to the default stochastic search ($\Delta E = 7$ kcal/mol, MaxConfs = 10000, distance-dependent dielectric constant), another 17 parameter combinations for this algorithm were tested (Table 2).

It was also of particular interest to compare the conformational coverage obtained with MOE algorithms to the well-established[15,32,33,39,45] FAST and BEST methods in the program Catalyst.[25] FAST and BEST use a modified CHARMm force field devoid of electrostatic component,

Table 2. Tests of Conformation Sampling Protocols[a] with Respect to the Percentage of Reproduced Bioactive X-ray Conformations[b]

	investigated parameters				rmsd (Å) versus bioactive[c]				
method	ΔE[e] (kcal/mol)	MaxConfs[f]	*GB*[g]	*Diel*[h]	0.5	1.0	1.5	2.0	NbConfs[d]
MOE and Catalyst at default settings									
Conformation Import[i]	4	250	no	no	32%	75%	94%	98%	110
Stochastic search[j]	7	10000	no	yes	34%	71%	84%	92%	67
Catalyst FAST	20	255	NA	NA	24%	70%	95%	100%	106
Catalyst BEST	20	255	NA	NA	29%	81%	98%	100%	111
Conformation Import	Conformation Import with MaxConfs varied								
	4	**100**	no	no	30%	70%	91%	96%	60
	4	250	no	no	32%	75%	94%	98%	110
	4	**500**	no	no	32%	79%	96%	98%	171
	4	**1000**	no	no	32%	82%	98%	100%	273
	4	**10000**	no	no	33%	87%	99%	100%	1175
	Conformation Import with Δ*E* varied								
	3	250	no	no	30%	74%	94%	98%	106
	4	250	no	no	32%	75%	94%	98%	110
	7	250	no	no	32%	75%	94%	98%	114
	20	250	no	no	32%	75%	94%	98%	114
	Conformation Import with Δ*E* and MaxConfs increased[k]								
	7	**500**	no	no	32%	79%	96%	98%	177
	7	**1000**	no	no	32%	82%	98%	100%	288
	20	**500**	no	no	32%	79%	96%	98%	177
	20	**1000**	no	no	32%	82%	98%	100%	289
	Conformation Import with refinement								
	4	250	no	**yes**	30%	68%	84%	94%	46
	4	250	**yes**	no	39%	75%	91%	97%	80
	Conformation Import with refinement and Δ*E* increased[k]								
	7	250	no	**yes**	32%	71%	88%	96%	68
	20	250	no	**yes**	34%	73%	93%	98%	92
	7	250	**yes**	no	40%	76%	95%	98%	94
	20	250	**yes**	no	38%	76%	94%	98%	98
	Conformation Import with refinement and MaxConfs increased[k]								
	4	**500**	no	**yes**	30%	68%	84%	95%	49
	4	**500**	**yes**	no	39%	76%	93%	98%	85
	Conformation Import with refinement and Δ*E* and MaxConfs increased[k]								
	7	**500**	no	**yes**	32%	71%	90%	96%	73
	20	**500**	no	**yes**	34%	74%	95%	98%	101
	7	**500**	**yes**	no	40%	78%	96%	99%	102
	20	**500**	**yes**	no	38%	77%	96%	99%	107
Stochastic search	Stochastic search with MaxConfs decreased[l]								
	7	**250**	no	yes	35%	70%	84%	91%	51
	7	10000	no	yes	34%	71%	84%	92%	67
	Stochastic search with Δ*E* increased[k]								
	7	10000	no	yes	34%	71%	84%	92%	67
	12	10000	no	yes	39%	83%	92%	98%	234
	15	10000	no	yes	41%	85%	94%	97%	452
	20	10000	no	yes	41%	89%	97%	100%	760
	30	10000	no	yes	43%	92%	99%	100%	1274
	Stochastic search with Δ*E* increased[k] and *GB*[g]								
	7	10000	**yes**	no	41%	83%	97%	99%	139
	12	10000	**yes**	no	43%	91%	99%	100%	490
	15	10000	**yes**	no	43%	94%	99%	100%	806
	20	10000	**yes**	no	44%	94%	100%	100%	1107
	30	10000	**yes**	no	44%	94%	100%	100%	1276
	Stochastic search with Δ*E* increased[k] and MaxConfs decreased[l]								
	7	**250**	no	yes	35%	70%	84%	91%	51
	12	**250**	no	yes	38%	80%	90%	95%	94
	15	**250**	no	yes	40%	82%	93%	98%	110
	20	**250**	no	yes	66%	82%	95%	98%	122
	Stochastic search with Δ*E* increased,[k] MaxConfs decreased,[l] and *GB*[g]								
	7	**250**	**yes**	no	39%	85%	95%	99%	84
	12	**250**	**yes**	no	44%	88%	98%	100%	123
	15	**250**	**yes**	no	43%	88%	98%	100%	125
	20	**250**	**yes**	no	44%	89%	99%	100%	128
Catalyst FAST	Catalyst FAST with MaxConfs varied								
	20	**50**	NA	NA	19%	59%	88%	99%	33
	20	**100**	NA	NA	23%	63%	92%	99%	57
	20	255	NA	NA	24%	70%	95%	100%	106
	20	**500**	NA	NA	24%	72%	95%	100%	150
	20	**1000**	NA	NA	24%	73%	95%	100%	195

Table 2. Continued

method	investigated parameters				rmsd (Å) versus bioactive[c]				
	ΔE[e] (kcal/mol)	MaxConfs[f]	*GB*[g]	*Diel*[h]	0.5	1.0	1.5	2.0	NbConfs[d]
				Catalyst FAST with Δ*E* varied					
	7	255	NA	NA	20%	58%	89%	98%	52
	12	255	NA	NA	22%	65%	93%	99%	81
	15	255	NA	NA	24%	68%	93%	100%	95
	20	255	NA	NA	24%	70%	95%	100%	106
	30	255	NA	NA	24%	70%	95%	100%	116
	40	255	NA	NA	24%	71%	95%	99%	120
				Catalyst FAST with Δ*E* varied and MaxConfs increased[k]					
	7	**500**	NA	NA	20%	58%	89%	98%	59
	12	**500**	NA	NA	22%	66%	93%	99%	102
	15	**500**	NA	NA	24%	70%	94%	100%	126
	20	**500**	NA	NA	24%	72%	95%	100%	150
	30	**500**	NA	NA	24%	72%	96%	100%	176
	40	**500**	NA	NA	24%	75%	96%	100%	184
	7	**1000**	NA	NA	21%	59%	89%	98%	65
	12	**1000**	NA	NA	23%	67%	93%	99%	123
	15	**1000**	NA	NA	25%	71%	94%	100%	158
	20	**1000**	NA	NA	24%	73%	95%	100%	195
	30	**1000**	NA	NA	25%	73%	96%	100%	250
	40	**1000**	NA	NA	25%	75%	96%	100%	270
Catalyst BEST				Catalyst BEST with MaxConfs varied					
	20	**50**	NA	NA	26%	70%	94%	100%	41
	20	**100**	NA	NA	27%	77%	96%	100%	69
	20	255	NA	NA	29%	81%	98%	100%	111
	20	**500**	NA	NA	29%	82%	98%	100%	131
	20	**1000**	NA	NA	29%	82%	98%	100%	144
				Catalyst BEST with Δ*E* varied					
	7	255	NA	NA	25%	69%	95%	98%	42
	12	255	NA	NA	29%	76%	97%	99%	75
	15	255	NA	NA	28%	79%	96%	100%	90
	20	255	NA	NA	29%	81%	98%	100%	111
	30	255	NA	NA	29%	83%	98%	100%	134
	40	255	NA	NA	29%	82%	98%	100%	150
				Catalyst BEST with Δ*E* varied and MaxConfs increased[k]					
	7	**500**	NA	NA	25%	69%	95%	98%	44
	12	**500**	NA	NA	29%	76%	97%	99%	80
	15	**500**	NA	NA	28%	79%	96%	100%	101
	20	**500**	NA	NA	29%	82%	98%	100%	131
	30	**500**	NA	NA	29%	83%	98%	100%	172
	40	**500**	NA	NA	30%	83%	99%	100%	203
	7	**1000**	NA	NA	25%	69%	95%	98%	45
	12	**1000**	NA	NA	29%	76%	97%	99%	85
	15	**1000**	NA	NA	28%	79%	96%	100%	109
	20	**1000**	NA	NA	29%	82%	98%	100%	144
	30	**1000**	NA	NA	29%	83%	98%	100%	194
	40	**1000**	NA	NA	30%	83%	100%	100%	242

[a] Nondefault parameters are in bold. Some rows with default settings are repeated to facilitate comparison with varied parameters. [b] For the complete set of 256 compounds. [c] rmsd threshold used to define the percent of reproduced X-ray bioactive conformations. [d] Average number of output conformations per ligand. [e] Maximum tolerated intramolecular conformational energy per conformer ("energy window"). [f] Maximum number of conformers generated per compound. [g] Generalized Born solvation model.[44] [h] Screening of electrostatic with the MOE distance-dependent dielectric model. [i] Conformation Import from MOE. [J] Stochastic search from MOE. [k] Increased relative to the default value. [l] Decreased relative to the default value. NA: not applicable, because Catalyst does not offer options in terms of solvation models.

with default settings MaxConfs = 255 and ΔE = 20 kcal/mol. The FAST and BEST approaches have already been presented.[25,32] FAST is recommended to produce conformational libraries because it is faster than the more thorough BEST protocol.[32,33] BEST performs a more rigorous energy minimization in both torsional and Cartesian coordinates. Results from 40 protocols using Catalyst 4.10 FAST or BEST with various combinations of MaxConfs and ΔE (Table 2) were obtained, to put the MOE results in perspective and to compare to other Catalyst tests on different compound sets.[32,33,39]

The conformational models were tested to see if they contained ("reproduced") the bioactive conformation. For selected protocols, we characterized the conformational coverage in broader terms, with more generic 3D descriptors such as radius of gyration and pharmacophoric fingerprints.

2.5. Reproduction of the Bioactive Conformation. A key test of a conformational sampling protocol is whether it produces at least one conformer close to the bioactive conformer. This was tested by finding the lowest rmsd (nonhydrogen atoms only) between the members of a conformational ensemble and the corresponding X-ray bioactive reference, after best fit of the structures using a customized SVL script (symmetry-related conformations were taken into account). Four rmsd thresholds were used: 0.5 Å, 1.0 Å, 1.5 Å, and 2.0 Å (Table 2). RMSDs were calculated in the same

way for the conformers output by MOE or Catalyst. The discussion emphasizes results obtained at rmsd ≤ 1.0 Å, which arguably represents a good quality fit of practical value.

Reproduction of the bioactive conformation was studied with MOE Conformation Import, MOE stochastic search, and Catalyst FAST and BEST. The impact of the following parameters was investigated: MaxConfs and ΔE independently, MaxConfs and ΔE in combination, energy minimization (Conformation Import only), and solvation model (Conformation Import and stochastic search only). The percentage of bioactive conformations reproduced by a protocol is abbreviated as *%BioConf_Rep*. The results are in Table 2.

2.6. Conformational Energies of the Bioactive Conformers. The present work is not focused on an in-depth analysis of the internal energies of the bioactive conformers, which is a difficult question,[34,35,37,46] yet we examined our results in this context, with the conformers from the protocol yielding the highest percent of reproduced bioactive structures (94% within rmsd ≤ 1.0 Å), i.e. MOE stochastic search with *GB* and $\Delta E = 15$ kcal/mol. The energies of the conformers close to experimental bioactive structures (within 1.0 Å or 0.5 Å, two groups considered) were analyzed as surrogates to their experimental counterparts. Indeed, meaningful force field energies cannot be obtained directly from the X-ray coordinates. This is largely because the bond length and valence angles in the crystal structure will frequently differ from the corresponding equilibrium values in the force field, leading to artificially large energies even for small differences. The 1.0 Å or 0.5 Å cutoffs provide only approximations of the bioactive conformations and are not ideal. This is, however, a pragmatic tradeoff to exploit our data to glean insights with respect to the energetics of the bound ligands.

For each group of bioactive-like conformers (within 1.0 Å or 0.5 Å of X-ray reference) only the conformer of lowest energy was kept, providing a single bioactive-like representative conformer per compound. We checked that this conformer had the same stereochemistry as the X-ray bioactive reference. For a given compound, there could be a conformer slightly closer to this reference but of higher energy. With energies being very sensitive to small coordinate displacements, we selected the lowest energy conformer within a certain distance (1.0 Å or 0.5 Å) of the X-ray reference, rather than the closest conformer. The energy of the representative bioactive-like conformer was offset relative to that of the global energy minimum with the same chirality. It is unclear whether consideration of chirality classes was taken into account in similar studies.

This provided relative conformational energies for two groups of bioactive-like conformers, within 0.5 Å or within 1.0 Å of the X-ray reference (Figure 9). These energies include the *GB* contribution. There are arguments to remove this *GB* contribution[37] to obtain pure conformational energies *in vacuo*. For simplicity and clarity, we opted against this approach and its nontrivial practical complications. For instance, it is the total energies including *GB* which were optimized and not their *in vacuo* component. In the context of the other approximations involved, we consider that such analysis is sufficient to comment on the broad trends, as done recently.[35]

2.7. Conformational Coverage Assessed by 3D Descriptors. The extent and diversity of the conformational coverage afforded by a conformational model are important but difficult to characterize unequivocally. The average number of distinct conformers per compounds (NbConfs) and the percent of reproduced bioactive conformers (*%BioConf_Rep*) are two indicators of conformational coverage. Other 3D descriptors are also helpful, which measure the compactness of the conformers[35,40] or their diversity in terms of pharmacophores.[39] The degree of conformational extension/compactness for each compound was characterized with the water accessible surface area (ASA) and the radius of gyration (Rgyr), as calculated with MOE for conformational ensembles generated by MOE and Catalyst. ASA was calculated using a spherical probe of 1.4 Å radius, the expectation being that more extended conformers have greater ASAs. For a given compound, the percent of conformational extension ConfExt was defined with reference to the most compact conformer (Min_descriptor):

$$\text{ConfExt} = 100 \cdot (\text{Max\_descriptor} - \text{Min\_descriptor})/\text{Min\_descriptor} \quad (1)$$

Max_descriptor and Min_descriptors were the maximum and minimum values of the 3D descriptor (ASA or Rgyr) in the conformational ensemble generated for a compound, by a given protocol. Thus, if Max_descriptor (most extended conformer) is twice Min_descriptor, the most extended conformer is 100% more extended than the most compact. ConfExt can be greater than 100%. See Figure S2 in the Supporting Information for illustrations. By comparing the geometric differences between the most extended and the most compact conformers across protocols, ConfExt can be seen as a crude indicator of conformational diversity.

A finer view of the overall conformational diversity afforded by a conformational model was obtained with 3D pharmacophore fingerprints calculated in MOE. These were the Pharmacophore Atom Triangle (piDAPH3, three-point pharmacophores) and the Pharmacophore Atom Quadruplet (piDAPH4, four-point pharmacophores). Every conformer was characterized by fingerprints of 'bits' representing its unique pharmacophore triangles or quadruplets. Each bit was determined by the pharmacophoric features and distances between them. The pharmacophoric features were defined as in MOE, based on hydrophobic groups and hydrogen-bond donors or acceptors. For each conformational model, the total number of unique pharmacophore triangles or quadruplets identified from the 256 compounds was compiled. Comparison of these numbers gave an estimate of the sampling of the pharmacophoric space across methods.

3. RESULTS AND DISCUSSION

Unless stated otherwise, the results relate to the default settings of the tested protocols and the set of 256 unique compounds.

3.1. Properties of the Compound Test Sets. Three sets of ligands were combined, referred to as the Bostrom, modified-Perola, and Vernalis sets (Methods and Table S1 in Supporting Information), with a total of 256 unique compounds. This compound set is larger than others concerned with bioactive conformers[31,35,37,39−41] but smaller than two recent sets[32,33] because it was focused by a ligand

diversity analysis, a cutoff on the crystallographic resolution, and a requirement for more druglike properties.

The Bostrom and Perola sets were presented before[31,35] but not analyzed comparatively. This is done with MOE descriptors in Figure 1, which also puts in perspective the same properties for the newly devised Vernalis set. These properties are related to leadlikeness[47] or druglikeness,[48] that is molecular weight (MW; Figure 1A), number of rotatable bonds (NRot; Figure 1B), the calculated n-octanol/water partition coefficient SlogP (Figure 1C), the number of chiral centers (Figure 1D), and the number of hydrogen-bond donors and acceptors (Figure 1E). For simplicity, the count of hydrogen-bond donors and acceptors per compound is summarized as the sum of these two properties, even though that is not strictly equivalent to separate counts of donors and acceptors.

The limitations of the Bostrom set regarding its relatively small number of compounds and their departures from druglikeness were already noted,[35] e.g. the low MW (<200) of some compounds. Compilation of the Perola set was a major effort toward a larger set enriched in druglike molecules.[35] Here, only the public component of the Perola set is used, therefore called "modified-Perola". Figure 1 shows that the modified-Perola set still contains a significant fraction of compounds with MW > 400 (47%), or NRot > 5 (60%), or more than 1 chiral center (29%). Such larger and more flexible/complex compounds were kept because of their interest when testing conformational sampling. However, such compounds would be rather unattractive starting points when initiating a pharmaceutical discovery project, because smaller and less complex (leadlike) compounds are easier to progress toward drug candidates with the desired properties.[47,49] Hence, we enriched the Vernalis set with a higher proportion of leadlike compounds. Thus, 66% of the Vernalis set compounds have MW < 400, only 25% have NRot >5, and only 6% have >1 chiral center. The more frequent occurrence of peptidomimetics in the Bostrom and modified-Perola sets than in the Vernalis set probably explains the greater polarity of the two former sets, as reflected in the distributions of SlogP. It is consistent with the number of hydrogen-bond donors + acceptors being skewed toward smaller values in the Vernalis set versus the modified-Perola set. These trends are reflected in the properties average values in Table 1.

Another important property regarding the accuracy of the bioactive conformation is the associated crystallographic resolution (Figure 2). All compounds in the Bostrom set were solved at a resolution ≤2 Å, while the Perola set extends to 3 Å. Only structures with a resolution ≤2.5 Å were added to the Vernalis set, compatible with reasonably well-defined ligand structures[38] (42% of these structures were solved at a resolution ≤2 Å). The PDB entries for each compound set are given in Table S1 of the Supporting Information. The Vernalis set is dominated by kinases, reflecting their role as drug targets. Other important targets in the Vernalis set include the HIV reverse transcriptase, estrogen receptors, and the heat shock protein 90 (Hsp90, not represented in the Bostrom/modified-Perola sets).

More relevant to the present work is the chemical diversity of the ligands. Only ligands different enough from those already in the Bostrom/modified-Perola sets were added to the Vernalis set, consistent with the Vernalis compounds having distinct physicochemical properties (e.g., SlogP distribution). Overall, the three compound sets provide a broad coverage of the druglike space with crystallographic information on the bioactive structure. These sets are complementary regarding their physicochemical properties and chemical diversity and are used separately or in combination in what follows. Comparisons across studies are sensitive to the underlying compound sets, as illustrated below.

3.2. Search for the Global Energy Minimum. Conformational analysis relies on intramolecular energy differences (ΔE) between conformers. Ideally, these energy differences should be referenced relative to the global energy minimum relevant to the environment of interest (*in vacuo* or solution). We adopt aqueous solution as reference, considering that it is commonly where binding of a ligand to a receptor takes place. For instance, the global energy minimum in solution is the conformation relative to which the energetics, and possible strain, of the bioactive conformation should be compared. This is a simplification, because a ligand may populate several conformations in solution (and in its bound state, eg groups protruding into the solvent), and it would be more rigorous to represent the unbound ligand with its Boltzmann weighted conformations in solution. However, representing the unbound (and bound) ligand with a single conformation is a reasonable approximation to investigate the practical questions addressed here. The ligand global energy minimum was estimated with the Merck Molecular Force-Field (MMFF) and the generalized Born (*GB*) model of aqueous solvation implemented in MOE.[44]

Given a potential to calculate the energies, one varies the conformations to locate the global minimum with this potential. MOE offers three approaches: systematic search, stochastic search, and Conformational Import. With a systematic search, the lowest energy conformer should in principle be the global energy minimum.[7,11] However, finding the lowest energy via a stochastic search does not *a priori* guarantee that it is the global energy minimum. Yet, the stochastic search is often preferred over the systematic method because it is much faster. Hence, it is important to assess how frequently the global energy minimum is located by the MOE stochastic search, by comparison to the systematic search. Comparing Conformational Import to the systematic search is also of great interest.

A comparison of energy minima found by systematic and stochastic searches was performed on 70 compounds without chiral centers (section 2.3) but spanning a typical Nrot range of 1 to 9. The conformers of lowest energy from systematic and stochastic searches were compared in terms of structural (rmsd) and energy similarity (ΔE between energy minima). The energies of the global minima from the systematic search are compared in Figure 3 to the lowest energy minima from the stochastic search, showing that the lowest energies from both approaches coincide exactly for all but 3 compounds. The three energy differences $E^{\text{stochastic}} - E^{\text{systematic}}$ were 2.3 kcal/mol (PDB entry 1ZZ2), 2.0 kcal/mol (2B52) and −1.9 kcal/mol (966C). Thus, there were only two occasions where the stochastic search missed the global energy minimum (1ZZ2 and 2B52), and one case where it uncovered an energy minimum deeper than found by the systematic search (966C). For 966C, this difference appears related to the tetrahydropyran ring, possibly because the systematic search step of

15 degrees for ring torsions may occasionally be too coarse. Aliphatic rings present particular difficulties for systematic search algorithms.[7,50] Overall, there was excellent agreement between the lowest energies identified by the systematic and stochastic searches for druglike compounds.

Also, we examined the structural similarity between the global energy minima conformers from systematic and stochastic searches (symmetries taken into account). It is complicated by the fact that, occasionally, several distinct "global minima" of the same energy may exist. Only the most structurally similar pairs were considered, to see if the stochastic search found at least one of the global energy minima from the systematic search. The vast majority of these pairs (91.4%) are within rmsd ≤ 0.25 Å of each other, virtually identical. For the remaining 8.6% of the pairs, the RMSDs are distributed essentially evenly between 0.25 and 2.0 Å. Thus, for 91.4% of the compounds, the systematic and stochastic searches identified the same global energy minimum. For 8.6% of the compounds, slightly distinct conformations are associated with the same lowest energy, reflecting the ruggedness of the energy surface for druglike molecules. Figure S1 (Supporting Information) illustrates these situations. Overall, the stochastic search performed remarkably well in locating global energy minima equivalent to those from systematic searches. This is an important validation of the stochastic search implemented in MOE, which complements initial tests performed with cycloalkanes.[12]

To test the ability of stochastic searches to locate the global minimum we also compared various stochastic searches against each other, regarding how consistently they locate the same lowest energy minimum. In addition to the default energy cutoff ($\Delta E = 7$ kcal/mol), three other stochastic searches were run for each compound, with $\Delta E = 12$, 15, and 20 kcal/mol. The energy cutoff is applied as the search proceeds and can therefore influence the search process and the lowest energy minima identified. The different stochastic searches yielded very similar lowest energy minima (not shown). Thus, the stochastic searches proved robust regarding identification of the global energy minimum, with little influence of ΔE. The energy cutoff, however, affects conformational coverage (section 3.5).

We also tested if the high-throughput protocol Conformation Import can identify the global energy minimum, in terms of the RMSDs between lowest energy conformers from Conformation Import and the systematic search. The default Conformation Import is not suited to locate the global energy minimum because it does not energy-minimize the assembled conformers. Hence, only Conformation Import with refinement (energy minimization) with the *GB* model was considered. 76% of the relevant global energy minima pairs had an rmsd ≤ 0.25 Å, 86% had an rmsd ≤ 1.0 Å, and 97% had an rmsd ≤ 2.0 Å. Therefore, this particular usage of Conformation Import is reasonably successful at locating the global energy minimum but less reliable than stochastic search for this task.

Overall, with MOE, the stochastic search is the method of choice to locate the global energy minimum for druglike molecules. It performs virtually as well as the systematic search, in much less computational time. With compounds containing chiral centers of unknown configurations, one expects the stochastic search to be even more helpful than the systematic search, to produce energy minima associated with different stereoisomers. Conformational coverage with the stochastic search is addressed in the following.

3.3. Reproduction of the Bioactive Conformation with Default Settings. The global energy minimum is important in conformational analysis, but other conformations are accessible and relevant at physiological temperatures. In particular, the bioactive conformation is thought to frequently differ from the global energy minimum in solution,[34−37] and it has been proposed that some ligands may tolerate significant intramolecular strain to adopt their bioactive form.[32,34,35] Thus, capturing the bioactive conformation commonly requires extensive conformational sampling, for example when preparing libraries for pharmacophore-based virtual screening. Then, two key variables are the maximum number of conformers generated per compound (MaxConfs) and the maximum accepted intramolecular conformational energy (ΔE in kcal/mol). The optimal values for these parameters need to be estimated empirically. Such tests have not been presented for the MOE Conformation Import workflow (Figure 4), developed to prepare searchable libraries of conformers. Here we test if Conformation Import can produce conformers close (rmsd) to the bioactive structure, using a curated set of 256 ligands (Section 3.1). Results from Conformation Import are compared to those from Catalyst and the MOE stochastic search. For a given compound, when at least one computationally generated conformer is within a defined rmsd cutoff of the bioactive structure, this bioactive conformation is considered "reproduced". The results are presented as the percent of compounds for which the bioactive reference is reproduced (*%BioConf_Rep*), when varying i) the rmsd cutoff (Figure 5, Table 2) and ii) some key parameters of the computational protocol (e.g., ΔE, MaxConfs). This section considers results obtained with the default settings.

Figure 5A presents *%BioConf_Rep* obtained with default Conformation Import for the Vernalis, Bostom, and modified-Perola compound sets as well as the combined 256 unique compounds. Values of *%BioConf_Rep* for the 256 compounds are also presented in Table 2. Overall, >75% of the bioactive conformers are reproduced within rmsd ≤ 1.0 Å. Almost all compounds (98%) are reproduced within rmsd ≤ 2.0 Å. These statistics cover disparities between compound sets, in particular between the Vernalis and modified-Perola sets. At rmsd ≤ 1.0 Å, 84%, 75%, and 61% of the Vernalis, Bostrom, and modified-Perola compounds are reproduced, respectively. At rmsd ≤ 2.0 Å, the corresponding fractions are 100%, 97%, and 94%, respectively. These performance differences presumably reflect different properties of the compound sets (section 3.1). The default Conformation Import performs remarkably well on the Vernalis set of leadlike compounds, but that performance degrades with the statistically larger and more complex compounds in the modified-Perola set. To minimize biases associated with any particular compound set, the following comparisons are made on the 256 unique compounds.

Figure 5B and Table 2 compare Conformation Import to the MOE stochastic search and the Catalyst FAST/BEST protocols (default settings). The test is as in Figure 5A, that is, the percentage of reproduced bioactive conformers within an rmsd cutoff. The stochastic search is not intended for the generation of conformational libraries but provides an

interesting comparison to Conformation Import because it locates energy minima more efficiently (section 3.2). Also, the stochastic search can generate by default a maximum of 10000 conformers per compound versus 250 with Conformation Import. Catalyst FAST provides another important comparison, because it is recommended (rather than Catalyst BEST) for generation of large multiconformer libraries and is well validated.[32,33] Catalyst BEST was also tested (Table 2), although it is significantly slower than FAST[32,33] (section 3.10).

Interestingly, the default settings of Catalyst FAST/BEST use MaxConfs = 255, similar to MaxConfs = 250 in Conformation Import, encouraging a comparison of these protocols. Indeed, they yield an average number of conformers per compound (NbConfs) of the same order (Table 2).

At rmsd ≤ 1.0 Å, *%BioConf_Rep* were 75% (Conformation Import), 71% (stochastic search), 70% (Catalyst FAST), and 81% (Catalyst BEST). At rmsd ≤ 2.0 Å, *%BioConf_Rep* were 98%, 93%, 100%, and 100%, respectively. These *%BioConf_Rep* values are higher than those reported for the programs Omega and Catalyst in a previous study,[33] reflecting the more flexible compounds used previously. For instance Catalyst FAST with MaxConfs = 250 was reported to reproduce 52.7% of the bioactive conformers within 1.0 Å,[33] as compared to 70% here (MaxConfs = 255).

Thus, the four tested methods performed reasonably well in generating the bioactive conformation for a great majority of ligands. At default settings, BEST yielded the highest *%BioConf_Rep* values (81% at rmsd ≤ 1.0 Å), confirming that it has a distinctive advantage in terms of generating diverse conformations. This is clearly at the expense of computational speed (section 3.10). The stochastic search is also computationally demanding but without particular benefits for conformational coverage at default settings. The settings can however be tuned to improve dramatically the conformational coverage afforded by the stochastic search, which outperforms Catalyst BEST (section 3.7). For practical purposes, the much faster Conformation Import and Catalyst FAST performed essentially as well, with slightly better results with Conformation Import at rmsd ≤ 1.0 Å (Table 2 and Figure 5B).

Still, 25% of the bioactive structures were not reproduced by Conformation Import at rmsd ≤ 1 Å. These nonreproduced compounds had on average 7.8 rotatable bonds and 1.3 chiral centers compared to 4.7 rotatable bonds and 0.5 chiral centers for the "reproduced" compounds. Reproduction of the bioactive conformation did not correlate with the crystallographic resolution. This suggests that the compound properties (flexibility, number of chiral centers), rather than possible uncertainties in the crystallographic models, explain the 25% nonreproduced compounds.

Figure 6A shows that NbConfs increases steadily with NRot for Conformation Import and Catalyst FAST/BEST. There is a similar trend regarding the average rmsd between the bioactive conformer and its best matching computational conformer, increasing with NRot (Figure 6B). This confirms that NRot and the attendant NbConfs are key, and probably dominant, parameters regarding the reproduction of the bioactive conformers. Figure 6A shows that NbConfs remains lower than MaxConfs (250 for Conformation Import, 255 for FAST or BEST), even for larger NRot values. Thus, the default values of MaxConfs for Conformation Import and

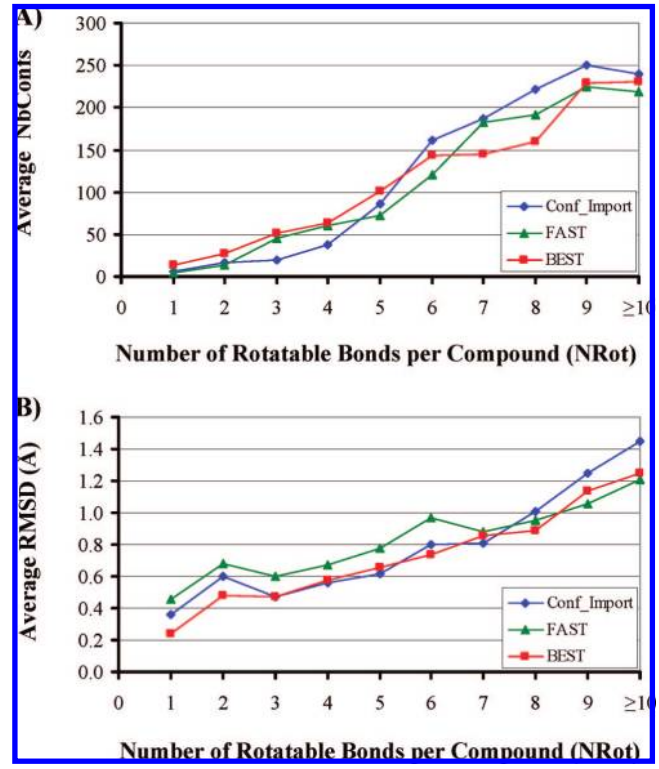

Figure 6. Influence of the number of rotatable bonds per compound (NRot) on i) Panel A: the average number of conformations generated per compound (NbConfs) and ii) Panel B: the average rmsd between bioactive structure and the best matching computational conformer. In each panel, results are plotted for the default settings of MOE Conformational Import (blue diamonds), Catalyst FAST (green triangles), and Catalyst BEST (red squares).

FAST/BEST are appropriate for druglike compounds. For NRot ≥5, Conformation Import generated more conformers per compound than FAST (e.g., at NRot = 6, NbConfs = 161 with Conformation Import, NbConfs = 120 with FAST), contributing to explain why Conformation Import performs slightly better than Catalyst FAST in reproducing the bioactive structures (Figure 5B). However, BEST outperforms both FAST and Conformation Import for reproduction of the bioactive conformation (Figure 5B) despite similar average NbConfs values (Table 2), implying that the performance gain with BEST reflects the diversity of the generated conformers.

Figure 7A shows the influence of the number of chiral centers (NChi) on NbConfs. NbConfs increases with NChi for NChi ≤ 2 but, surprisingly, plateaus or even decreases for higher values of NChi. The average rmsd measuring reproduction of the bioactive conformation also increases with NChi (Figure 7B). Strictly speaking, Figures 6 and 7 do not separate unambiguously the effects of NRot and NChi, but many more compounds would be needed to plot the effect of one variable while keeping the other constant.

Overall, the tested protocols (default parameters) are satisfactory for inclusion of the bioactive conformation in the conformational models. This performance clearly depends on compound flexibility, which correlates with NbConfs. This led to test the impact of NbConfs on *%BioConf_Rep* values, as influenced by the maximum number of conformers allowed per molecule.

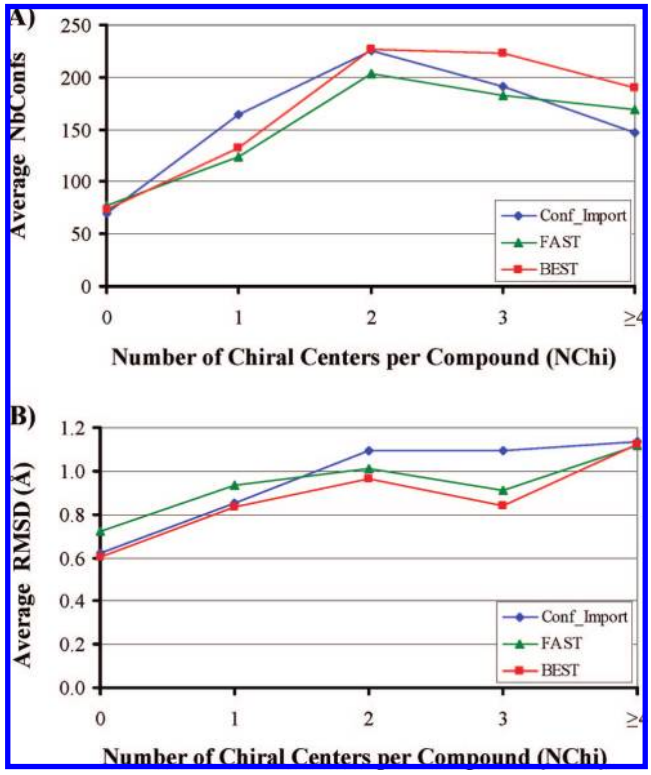

Figure 7. Influence of the number of chiral centers per compound (NChi) on i) Panel A: the average number of conformations generated per compound (NbConfs) and ii) Panel B: the average rmsd between bioactive structure and the best matching computational conformer. In each panel, results are plotted for the default settings of MOE Conformational Import (blue diamonds), Catalyst FAST (green triangles), and Catalyst BEST (red squares).

3.4. Effect of the Number of Generated Conformers.

The maximum number of generated conformers per compound (MaxConfs) influences the average number of conformers generated per compound (NbConfs), and therefore both MaxConfs and NbConfs are key parameters for the conformational coverage (Table 2). MaxConfs, rather than NbConfs, is typically controlled by the program user, hence we analyzed the impact of varying MaxConfs on *%BioConf_Rep*. The emphasis is on MOE Conformational Import, but MOE stochastic search and Catalyst FAST/BEST were also investigated. If not specified otherwise, the results pertain to Conformation Import.

With Conformation Import, MaxConfs was varied from 100 to 10000 (Figure 8A and Table 2). Compared to the default MaxConfs = 250, decreasing MaxConfs to 100 degrades *%BioConf_Rep* by 5% (to 70%) at rmsd ≤ 1.0 Å. Increasing MaxConfs up to 10000 steadily increases *%BioConf_Rep* in the rmsd interval 0.5 to 2.0 Å (Figure 8A). There is a marked effect at rmsd ≤ 1.0 Å, e.g. *%BioConf_Rep* = 87% at MaxConfs = 10000. This influence of MaxConfs is consistent with the above observation that compounds with a higher number of rotatable bonds and chiral centers are less well reproduced at the default settings.

All bioactive conformers are reproduced at rmsd ≤ 2.0 Å when MaxConfs reaches 1000, but the default MaxConfs = 250 already reproduced 98% of these conformations (Table 2). Increasing MaxConfs has a more visible effect in the range 0.75 Å ≤ rmsd ≤ 1.75 Å (Figure 8A). Therefore increasing MaxConfs mostly improves the accuracy with which the bioactive conformers are reproduced for rmsd < 2.0 Å, and the pertinent question is whether one needs such accuracy. Considering that pharmacophoric features are frequently delineated with tolerances <2.0 Å, it would be tempting to increase MaxConfs with Conformation Import. Yet, the default is a reasonable compromise.

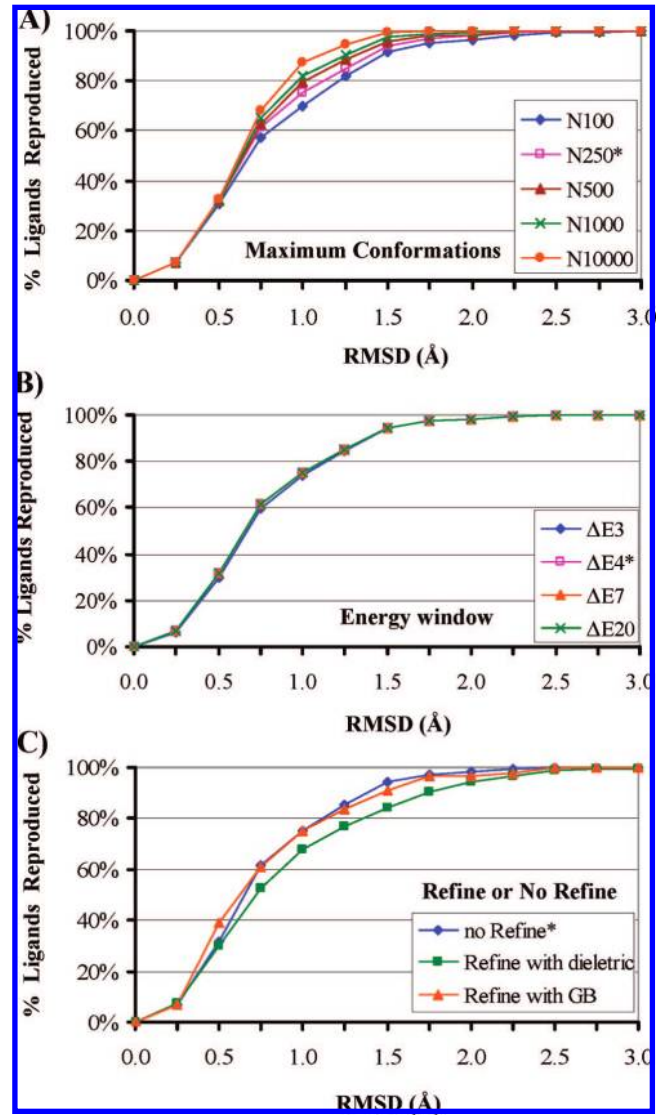

Figure 8. Influence of three variables (maximum number of conformations per compound, conformational energy window, refinement or not) on the performance of MOE Conformation Import with respect to the "reproduction" of the X-ray bioactive structures. A bioactive reference is deemed "reproduced" by a computed conformer when it is within a defined rmsd (Å) of this reference (X axis in all panels). As this rmsd is increased, more compounds have at least one computed conformer which "reproduces" the bioactive reference. Each panel presents cumulative frequencies of the percent of reproduced bioactive structures versus the rmsd, for all the 256 compounds. The default settings are flagged with an asterisk. Panel A: effect of the maximum number of conformers per compound (MaxConfs): 100 (blue diamonds), 250 (magenta squares, default), 500 (brown triangles), 1000 (green crosses), and 10000 (orange circles). Panel B: effect of the energy window (ΔE): 3 (blue diamonds), 4 (magenta squares, default), 7 (orange triangles), and 20 (green crosses) kcal/mol. Panel C: effect of refinement: no refinement (blue diamonds, default), refinement by energy minimization with a distance-dependent dielectric constant (green squares) or by energy minimization with a generalized Born model (orange triangles).

Varying MaxConfs changes NbConfs (Table 2), with implications for storage space requirements. However, Nb-

Confs increases slower than MaxConfs. For instance, at MaxConfs = 10000 (40 times the default 250), NbConfs increases only 10.7-fold (1175) relative to its value at default settings (110). Interestingly, NbConfs is on average significantly less than MaxConfs for Conformation Import and also the Catalyst protocols. At MaxConfs = 250, Conformation Import generates less than half (110) the allowed number of conformers on average. The proportion of added new conformations decreases when MaxConfs increases, as shown by the NbConfs/MaxConfs ratio. This ratio decreases steadily with MaxConfs, from 0.6 (MaxConfs = 100) to 0.1 (MaxConfs = 10000). Thus, increasing demands for computational resources yield diminishing returns, and the appropriate balance depends on the task at hand and the resources available.

MaxConfs was also investigated with Catalyst FAST and BEST, from 50 to 1000 (Table 2). For both FAST and BEST, *%BioConf_Rep* within rmsd ≤ 1.0 Å increases significantly (from 59 to 70% with FAST) when changing MaxConfs from 50 to 255. For rmsd ≤ 1.0 Å, the benefit of increasing MaxConfs levels off above 255 for FAST and BEST. Therefore, if rmsd ≤ 1.0 Å is used as criterion, the default MaxConfs = 255 is an attractive compromise between maximizing *%BioConf_Rep* and computational effort. A different view emerges if one considers the rmsd ≤ 2.0 Å cutoff, for which virtually all the bioactive conformers are reproduced with FAST (99%) and BEST (100%) at MaxConfs = 50. Therefore, the preferred value of MaxConfs depends on the desired accuracy for reproduction of the bioactive conformers.

Previous studies[32,33] have suggested that, with FAST, MaxConfs = 50 is the compromise of choice between quality of the output and resources usage. Our results confirm this if accuracy within rmsd ≤ 2.0 Å is deemed sufficient. However, the present results at rmsd ≤ 1.0 Å, combined with increasing capacity in computational resources, suggest that MaxConfs = 255 may frequently be a more convincing option for FAST.

Previously, MaxConfs with Catalyst was varied on compounds which were statistically larger and more flexible.[32,33] This previous work found that MaxConfs = 255 (ΔE = 20 kcal/mol) yielded i) NbConfs = 124 and *%BioConf_Rep* = 58.2% (rmsd ≤ 1.0 Å) with FAST and ii) NbConfs = 131 and *%BioConf_Rep* = 59.2% (rmsd ≤ 1.0 Å) with BEST. A recent study[33] with the program Omega (MaxConfs = 250, ΔE = 25 kcal/mol) found NbConfs = 109 and *%BioConf_Rep* = 57.4% at rmsd ≤ 1 Å. In our results, *%BioConf_Rep* values at rmsd ≤ 1.0 Å are higher (Table 2), probably reflecting differences in the compound sets. The present tests were performed on more conventionally druglike, less flexible, compounds. Thus, our results are qualitatively consistent with the previous degree of success for reproduction of the bioactive conformers[32,33] and indicate that such success is even more pronounced when tackling druglike compounds, with high-throughput methods such as FAST and Conformation Import. We compared such methods to the stochastic search, which was not developed for conformational library generation but could be of interest when focusing on compounds known to be important.

For default stochastic search MaxConfs = 10000 and NbConfs = 67, lower than NbConfs with other methods at default settings (Table 2). This is consistent with the stochastic search reproducing fewer bioactive conformers than the other methods for rmsd ≥ 1.0 Å (Figure 5B). Reducing MaxConfs dramatically to 250 with the stochastic search reduces NbConfs to 51, with a marginal effect on *%BioConf_Rep*. Thus, varying MaxConfs with the default energy window ΔE = 7 kcal/mol had little impact on the diversity of the conformers. Indeed, the stochastic search combined with energy minimization is primarily geared toward locating energy minima rather than broadening the conformational coverage. Increasing ΔE, however, has a marked influence on *%BioConf_Rep* with the stochastic search (Table 2 and section 3.5).

Overall, the choice of MaxConfs is important for reproduction of bioactive conformers, especially if good agreement is sought (rmsd ≤ 1 Å). However, *%BioConf_Rep* does not increase linearly with MaxConfs and tends to level off beyond a certain threshold. This threshold coincides approximately with default MaxConfs values for Conformation Import and FAST. Even for larger values of MaxConfs, there is a significant fraction of bioactive conformers not reproduced within rmsd ≤ 1.0 Å. This led to investigate the second major conformational sampling parameter, the allowed energy window ΔE.

3.5. Effect of the Energy Window on Reproduction of the Bioactive Conformers. The energy window (ΔE in kcal/mol) within which the conformers have to lie is another key parameter controlled by the program user and defining the conformational models. The default values for ΔE are 4 (MOE Conformation Import), 7 (MOE stochastic search), and 20 kcal/mol (Catalyst FAST and BEST). These energies are not directly comparable because they represent different ways to estimate the conformational energy. Conformation Import reports a "strain energy" which is not a complete force field energy, in contrast to the MMFF94 energies associated with the stochastic search (Methods). Catalyst uses a modified CHARMm force field without electrostatics. The bioactive conformation may adopt relatively high conformational energies,[32−35] at least in the context used to compute these energies. Such computed energies suffer from uncertainties in the force fields and the details of the X-ray structures. It is therefore important to test if ΔE can accommodate these apparently strained bioactive conformations. A broader ΔE increases the risk to add irrelevant high-energy conformers to the model. Therefore a balance has to be found for ΔE. This section investigates the impact of varying solely ΔE on *%BioConf_Rep* for the set of 256 compounds, with Conformation Import, stochastic search, and FAST/BEST.

With Conformation Import, we varied ΔE from 3 to 20 kcal/mol (Table 2, Figure 8B). Remarkably, varying ΔE has very little impact on NbConfs. ΔE = 7 and 20 kcal/mol yielded results virtually identical to those at the default ΔE = 4 kcal/mol. A cutoff of 3 kcal/mol, slightly more stringent than default, started to degrade reproduction of the bioactive conformations by 1% at rmsd ≤ 1 Å. Thus, the default of 4 kcal/mol is appropriate for Conformation Import.

With FAST, varying ΔE has a notable effect on NbConfs (Table 2). Decreasing ΔE below the default 20 kcal/mol degrades the performance. Yet, raising ΔE above 20 kcal/mol does not improve the results. A similar trend was observed with BEST. This is consonant with previous

observations[32] and strengthens the recommendation to keep ΔE at its default with Catalyst.

MOE stochastic search was the only protocol for which *%BioConf_Rep* improved significantly when raising ΔE above the 7 kcal/mol default, reflecting a concomitant increase in NbConfs (Table 2). From $\Delta E = 7$ to 30 kcal/mol, *%BioConf_Rep* (rmsd $\leq$ 1 Å) increases dramatically, from 71 to 92%, respectively (Table 2). At this rmsd cutoff, stochastic search outperforms any other protocol tested here. Even Catalyst BEST with MaxConfs = 1000 and $\Delta E = 40$ kcal/mol yielded only *%BioConf_Rep* = 83% (rmsd $\leq$ 1 Å). The stochastic search performs even better with a *GB* solvation model (section 3.7). Therefore, stochastic search with $\Delta E \geq 15$ kcal/mol is attractive when characterizing carefully a small number of important compounds, perhaps known binders before alignments to derive putative pharmacophores. Of course, deploying the stochastic search for large numbers of compounds remains challenging, given the compute times (section 3.10).

It is difficult to relate the optimal energy windows identified above to the true conformational strain of bioactive conformers because of limitations in the calculated energies and in the details of the X-ray structures.[31,38] Although an apparently low $\Delta E = 4$ kcal/mol yields reasonable results with Conformation Import, one should keep in mind that it does not represent a complete conformational energy. Of the methods tested here, only the stochastic search relies on full molecular mechanics energies (MMFF94). With the stochastic search, improvement of the results upon increasing ΔE to apparently unrealistic values (>15 kcal/mol) echoes similar experiences with Catalyst[32] and Omega.[33] A default energy window of 25 kcal/mol was found appropriate with Omega.[33] Section 3.7 investigates further the effect of ΔE combined with energy minimization and solvation models.

3.6. Varying MaxConfs and ΔE Together. The above studied the effects of MaxConfs and ΔE independently. This section investigates the effect of varying these two parameters together on *%BioConf_Rep* (Table 2). There could *a priori* be an interplay between these two parameters, if additional conformers could only be kept with a higher ΔE. With Conformational Import, $\Delta E = 7$ or 20 kcal/mol was combined with MaxConfs = 500 or 1000. The results were identical to those obtained when varying MaxConfs only, and no synergy between ΔE and MaxConfs was uncovered. It confirmed that there is no benefit in increasing ΔE above its default value with Conformation Import, in absence of refinement. With refinement, the results are more sensitive to ΔE (section 3.7).

With Catalyst FAST and BEST, ΔE was varied from 7 to 40 kcal/mol, with MaxConfs = 500 or 1000 (Table 2). Increasing MaxConfs has more effect for higher ΔE values, presumably because additional conformers are tolerated within the energy window. With FAST at rmsd $\leq$ 1.0 Å, the combination $\Delta E = 40$ kcal/mol with MaxConfs = 1000 achieves *%BioConf_Rep* = 75%, an increase of 5% over the results at default settings. However, at rmsd $\leq$ 1.5 Å the two protocols perform essentially as well. With BEST, there is only a very marginal gain in *%BioConf_Rep* when increasing ΔE and MaxConfs concomitantly. With FAST and BEST, the small improvements are not commensurate with the increase in computational costs.

Overall, with Conformation Import and Catalyst, there was no clear benefit in increasing simultaneously ΔE and MaxConfs.

3.7. Influence of Energy Minimization and Solvation on Reproduction of the Bioactive Conformation. The Catalyst FAST and BEST default protocols already include some energy minimization but without offering ways to adjust the treatment of solvation. Therefore, the influence of energy minimization or solvation was only investigated with the MOE Conformation Import and stochastic search.

With Conformation Import being geared toward computational speed, by default it does not energy-minimize the output conformers. For those to be minimized, the "refine" option has to be selected. This may be done with the MOE distance-dependent dielectric constant ("*Diel*"), a generalized Born model ("*GB*"), or *in vacuo*. These three treatments of electrostatics are also available to MOE stochastic search, which always energy minimizes the conformers. The default is *Diel*, for Conformation Import and stochastic search.

With Conformation Import, energy minimization (other parameters at default) significantly degraded *%BioConf_Rep* (Table 2 and Figure 8C), from 75% (no refine) to 68% (refine with *Diel*) at rmsd $\leq$ 1.0 Å. This degradation is less pronounced when refining with a *GB* model, although the default settings still perform better (Table 2 and Figure 8C). The average number of conformers (NbConfs) decreases from 110 (no refine) to 46 (refine with *Diel*) and 80 (refine with *GB*), explaining somewhat the performance degradation. The lowering of NbConfs upon minimization probably reflects higher-energy conformers converging into the same energy minimum and then being discarded as duplicates. In other words, minimization leads to less diverse conformers.

One has to realize that the nature of the energy scores changes when activating refinement in Conformation Import (Methods), while the energy window remains at 4 kcal/mol by default. The scores calculated after minimization use a more complete force field, which could bring some of the conformational energies outside of the tolerated energies. Thus, some conformers may be discarded based on their rescaled energetics, even if they are not duplicated conformations. Indeed, Conformation Import with refinement yields better results when ΔE is increased, to 7 and 20 kcal/mol (Table 2). These results, however, are virtually equivalent to those obtained at the default settings without refine.

While increasing MaxConfs without refinement increased NbConfs and improved the results (above), it is not the case with energy minimization (Table 2). That is consistent with energy minimization decreasing the number and diversity of the conformers by driving them into similar energy minima, even when more initial conformers are generated. This trend remains when refining with increased ΔE and MaxConfs. When varying ΔE and/or MaxConfs, the *GB* model performed consistently better than *Diel*. However, even with *GB*, improvements over the default settings (no refine) were at best marginal.

Given the significant computational overhead involved in energy minimization (section 3.10), the refinement option of Conformation Import is not attractive. Based on the fraction of reproduced bioactive conformers, the default settings for MOE Conformation Import appear judiciously chosen.

With stochastic search, energy minimization is integral to the algorithm, and only the effect of solvation was tested. As noted above, the results of stochastic search with *Diel* improve significantly when increasing ΔE (Table 2). The equivalent results with *GB* are even better, e.g. 94% of the bioactive conformers are reproduced within 1.0 Å with $\Delta E = 15$ kcal/mol. This is arguably an excellent result in absolute terms and the best *%BioConf_Rep* from the protocols tested here. This suggests that bioactive conformers tend to be at or near an energy minimum. Stochastic search results with *GB* do not improve by increasing ΔE above 15 kcal/mol, even if NbConfs increases. We note that, for $\Delta E < 30$ kcal/mol, NbConfs is significantly higher with *GB* than with *Diel*. With both solvation models the same rmsd test is applied to remove duplicate conformers, thus *GB* acts to promote a greater diversity of conformers. This probably explains the better performance of *GB* over *Diel*. The stochastic search compute times with *GB* are, however, about 2−4 times those with *Diel* (section 3.10). This would not be suitable for large library preparation but is reasonable for careful conformational sampling of selected compounds. Attempts to decrease MaxConfs for the stochastic search with *GB* degraded the results without shortening of the compute times. Decreasing the "Iteration Limit" (section 2.2) decreased the compute time but also NbConfs, yielding significantly degraded results. Therefore, the stochastic search with *GB*, $\Delta E = 15$ kcal/mol and MaxConfs = 10000 emerged as a method of choice to elucidate the bioactive conformation of selected compounds.

3.8. Conformational Energies of the Bioactive Conformers. Estimating the internal energies of the bioactive conformers as determined by X-ray crystallography is of great interest but notoriously difficult. This is because of uncertainties in both the X-ray derived geometries,[31,38] the force fields derived conformational energies,[32] and the reference conformational ensemble for the unbound ligand in solution.[36] Some studies have concluded that a significant fraction of bioactive conformations can be apparently highly strained,[32,34,35] maybe accommodated by compensating favorable interactions with the protein. Others have argued that the bioactive conformations are not so strained.[37,40] From simple thermodynamic considerations, e.g. the exponential relation between equilibrium binding constant and associated free energy, one expects that even small amounts of strain energies would penalize binding. However, calculating accurate conformational energies for the bound conformation, relative to that of the unbound ligand, is challenging for the reasons given above.

The aim of the present work is not to address these points extensively; however, we examine our results in the context of these questions. Thus, we present a simple analysis of the energetics of computed conformers close (within 0.5 or 1.0 Å) to their bioactive counterpart (see Methods). Such analysis was performed on the output of the MOE stochastic search with *GB* and $\Delta E = 15$ kcal/mol, which reproduced the bioactive structures within rmsd ≤ 1.0 Å for 241 compounds (94%) and within rmsd ≤ 0.5 Å for 112 compounds (43%). Thus, for each distance category (within 1.0 Å or 0.5 Å of the X-ray reference), these compounds were represented by one bioactive-like conformer. The advantage of this approach is to provide meaningful force field energies for these bioactive-like conformers, while such energies cannot be derived directly from the X-ray coordinates. For the bioactive-like conformers, the distribution of the internal energies relative to their global energy minimum is shown in Figure 9.

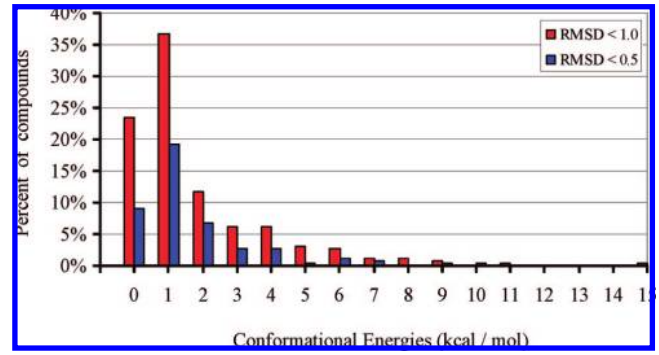

Figure 9. Distribution of the conformational energies for bioactive-like conformers close to the X-ray bioactive structures. The distribution is shown for the conformers within an rmsd of 0.5 Å (112 compounds, blue bars) or 1.0 Å (241 compounds, red bars) of the X-ray bioactive conformation. The frequencies are given as percent of the total number of compounds (256), with a bin width of 1 kcal/mol (e.g., 1−2 kcal/mol for the modes). These data were from the protocol yielding the highest proportion of reproduced bioactive structures, i.e. MOE stochastic search with $\Delta E = 15$ kcal/mol and the MMFF94 potential with a *GB* solvation model. For a given compound, these energies are relative to the conformation of lowest energy, with the same chirality as the bioactive structure.

The frequencies in Figure 9 reflect the larger number of compounds with a conformer within 1.0 Å of the X-ray reference, about twice as large as the group with the shorter cutoff of 0.5 Å. Yet, the distributions of conformational energies for the two groups are essentially similar. In particular, the mode for both distributions is at the 1−2 kcal/mol bin. The majority of the bioactive-like conformers have energies ≤3 kcal/mol relative to their unbound state. This echoes previous observations obtained with smaller sets of ligands,[30,37] which argued that the conformational strain of bound ligand is significantly less than was initially thought.[34] Yet, the matter remains under debate.[32,35] The present results suggest that druglike ligands bind to their targets with relatively low conformational strains, lower indeed than the average 15.9 kcal/mol initially calculated.[34] We find a larger proportion of compounds with conformational energies ≤3 kcal/mol than a recent in-depth study of the question,[35] but that reflects in part differences in the compound sets, i.e. the compounds in the present set are less flexible (Figure 1). The calculated conformational strains of bound compounds typically increase with the number of rotatable bonds.[34,35] A limitation of the present approach is that the bioactive-like conformations were obtained in absence of the protein environment. Thus, some protons which would otherwise engage in hydrogen-bonds with the protein may instead form favorable intramolecular interactions, which would lower artificially the conformational energies. Another caveat in our approach is that it implicitly assumes that the bioactive conformation exists in a local energy minimum. It may be that binding to the protein does force the ligand to deviate enough from such minimum to build significant conformational strain, as implied recently.[35] Alternatively, one may speculate that the ligand in its local energy minimum would fit the crystallographic electron density as well as the strained conformation.

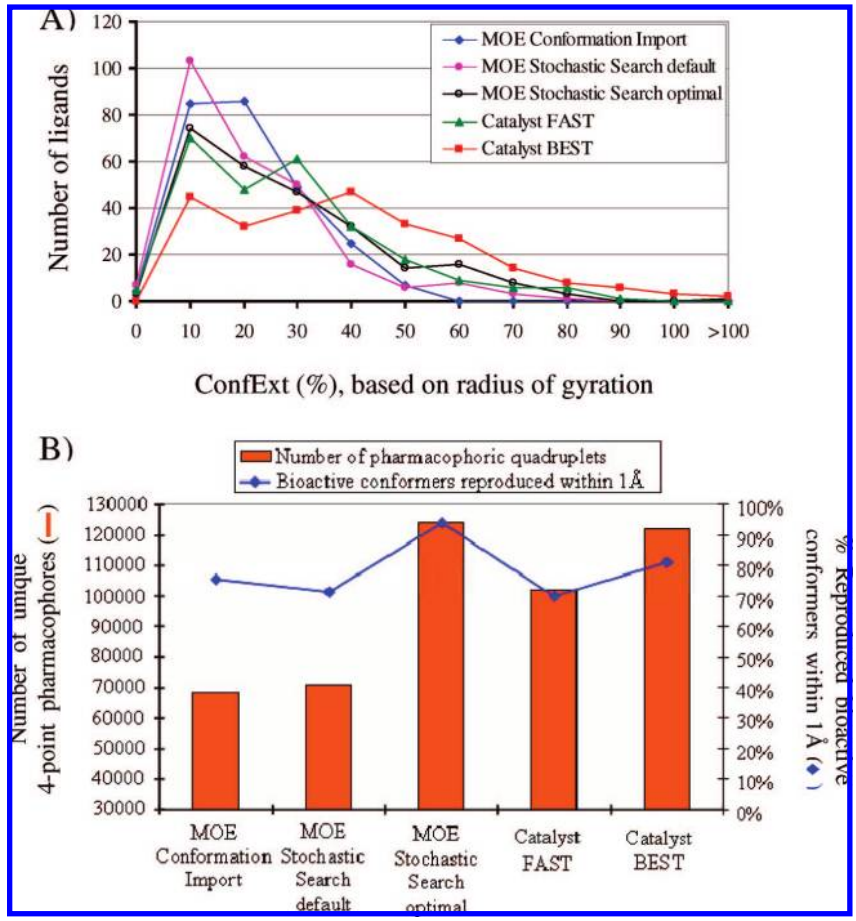

Figure 10. Use of 3D descriptors to characterize the conformational coverage afforded by selected protocols, i.e. MOE default Conformation Import, MOE default stochastic search, MOE optimal stochastic search, Catalyst FAST, and Catalyst BEST. The data are for the set of 256 compounds. Panel A: histogram of the percentage of compound extension ConfExt, based on the radius of gyration. Panel B: number of unique 3D pharmacophore quadruplet per selected protocol (orange bars, left Y axis), compared to the associated fraction of reproduced bioactive conformers at rmsd ≤ 1.0 Å (blue diamonds, right Y axis).

Besides, Figure 9 shows that, even allowing the bioactive-like conformers to be as far as 1.0 Å from their X-ray reference, some bioactive-like conformers are still quite strained. This is consistent with our tests showing that ΔE needs to be at least 15 kcal/mol to yield the largest percent of reproduced bioactive conformers (above). Overall, the present results are compatible with the chemically intuitive idea that most druglike compounds bind to their target without excessive conformational strain. Yet, the tests also show that, practically, one needs to use a broad energy window to maximize relevant conformational coverage with a molecular mechanics force field.

3.9. Conformational Coverage Assessed by 3D Descriptors. Conformational coverage refers to the number and diversity of the conformers generated per compound, including the structural/energetic spread of the conformers across the conformational space. Populating as many relevant areas as possible of the conformational space is a crucial aspect of modeling and 3D library design. The simplest descriptor of conformational diversity may be the average number of distinct conformers per compound NbConfs (Table 2), discussed above together with the influence of MaxConfs. Another key indicator of conformational coverage is how closely the bioactive conformation is reproduced, tested in detail above. Obtaining a more precise picture of conformational coverage and its spread is difficult but can be helped with considering the distribution of 3D descriptors associated with a conformational ensemble.

Among the 3D descriptors, those characterizing the degree of extension (compactness) of the conformers are relevant, because they provide an overall view of the underlying conformational ensemble.[39] Also, this can help assess the general impact of solvation models regarding electrostatic or hydrophobic collapse. Bioactive conformations are more likely to be extended than compact.[40] Here, the water Accessible Surface Area (ASA) and the radius of gyration (Rgyr) were used to characterize the percentage of extension of every compound (ConfExt). This percentage was calculated for the most extended conformer relative to the most compact (see Methods) and can vary from zero (rigid compound) to ≥100% (compound can significantly unfold), as illustrated in Figure S2 of the Supporting Information. The distribution of ConfExt values, based on the radius of gyration, is shown in Figure 10A for selected protocols (set of 256 compounds). Those include Conformation Import and Catalyst FAST/BEST run at their default parameters, as these protocols perform well for their intended tasks. The stochastic search at default parameters is also included, for comparison with its variant run at optimal conditions (*GB*, $\Delta E = 15$ kcal/mol) for reproduction of the bioactive conformers, referred to as "optimal stochastic search" in the following. For those protocols, Figure 10B compares the number of unique 3D

pharmacophore quadruplets generated across all the conformers of the 256 compounds (orange bars). Pharmacophore quadruplets, but not triplets, include a representation of chirality, yet the same trends were observed when comparing the numbers of pharmacophore triplets (not shown). The distribution of ConfExt values based on the ASA yielded trends which mirrored those derived from the radius of gyration (not shown).

Figure 10 shows a different coverage of the conformational space by the associated protocols. BEST yielded the most extended conformations (higher ConfExt values), while the default stochastic search tends to output the most compact (Figure 10A). Combining the stochastic search with *GB* ("optimal stochastic search") yields more extended conformations; the same trend was observed between two stochastic searches differing only by their solvation model, i.e. *Diel* versus *GB*. The conformers from Conformation Import present smaller ConfExt values than those from FAST/BEST. Higher ConfExt values reflect a larger geometric spread between most compact and most extended conformations, an indication of greater conformational diversity. This is consistent with Catalyst FAST and BEST using a poling algorithm[15] to promote conformational variation. It is also consistent with BEST yielding the highest percent of reproduced bioactive conformation, with default protocols (Table 2). However, there is no simple relation between reproduction of bioactive conformers and the distribution of ConfExt values. Indeed, the reproduction of bioactive conformers was maximized with the "optimal stochastic search", for which the ConfExt distribution is skewed toward smaller values as compared to BEST. Also, Conformation Import yields lower values of ConfExt than FAST, while the former performs slightly better for reproduction of the bioactive conformers (default settings).

Figure 10B provides another view of the conformational coverage, via the number of pharmacophore quadruplets (orange bars), alongside the associated *%BioConf_Rep* (blue diamonds). The number of pharmacophores and the underlying conformational diversity clearly vary across protocols. The MOE "optimal stochastic search" produces both the highest numbers of pharmacophores and *%BioConf_Rep* value, followed by BEST. However, there is no clear correlation between number of pharmacophores and *%BioConf_Rep*. For instance, FAST produces many more pharmacophores than MOE Conformation Import but does not perform better regarding reproduction of the bioactive conformers. The large number of pharmacophores associated with FAST and BEST probably reflect the use of the poling algorithm to promote conformational variation.[15] However, comparison across methods in Figure 10B suggests that not all conformational diversity translates into better *%BioConf_Rep* values. Poling could drive structures into regions of space unlikely to be populated.

3.10. Computational Run Times. On an AMD Opteron processor model 248 running at 2.2 GHz with 1Gb of RAM, the average compute times (seconds per compound, default settings) were 6 s (Conformation Import, no refinement), 4 s (Catalyst FAST), 84 s (stochastic search), and 124 s (Catalyst BEST). The 6 s per compound for Conformation Import was when some constitutive fragments were processed for the first time by the algorithm, and conformers for these fragments needed to be computed on the fly (section 2.2). When all fragments were precomputed, Conformation Import without refinement took 1 s/compound, the fastest method. These timings confirm that MOE Conformation Import and FAST combine acceptable accuracy (section 3.3, Table 2) and speed for library enumeration. MOE Conformation Import performed as well as the established Catalyst FAST.

Turning on the optional refinement (energy minimization) with Conformation Import leads to much longer compute times. Keeping other parameters at defaults, refinement took 67 s/compound with the distance-dependent dielectric constant (*Diel*) and 168 s/compound with *GB* (precomputed fragment conformations). Considering that refinement did not yield improvements compared to default settings (sections 3.3 and 3.7), we reiterate that the refinement option is unlikely to be of interest for routine library preparation.

Catalyst FAST at MaxConfs = 50 was recommended for high-throughput library enumeration in other studies[32,33] but is only 4 times faster than with MaxConfs = 255 (default) with our compound set. Considering the significant improvement in *%BioConf_Rep* at rmsd $\leq$ 1.0 Å when increasing MaxConfs from 50 to 255 with FAST, one may consider running FAST at default settings on modern hardware.

The protocol which performed best for reproduction of the bioactive conformers was the MOE stochastic search with *GB* and $\Delta E = 15$ kcal/mol. It was also very time-consuming, at 1763 s/compound. For this type of protocol, the compute time increases with ΔE, e.g. 340, 1200, 1763, 2155, 2563 s/compound for $\Delta E = 7$, 12, 15, 20, 30 kcal/mol, respectively. Results did not improve for $\Delta E > 15$ kcal/mol (Table 2), thus $\Delta E = 15$ kcal/mol emerged as optimal. Decreasing the "Iteration Limit" (see section 2.2), but not MaxConfs, shortened the compute times but also degraded the performance (not shown). The stochastic search with the default dielectric constant (*Diel*) was faster, e.g. 84, 285, 484, 724, 1397 s/compound for $\Delta E = 7$, 12, 15, 20, 30 kcal/mol, respectively. However, when considering only a small number of key compounds, the better results yielded by *GB* over *Diel* suggest that it is worth incurring the additional computational cost.

4. CONCLUSIONS

This work provides a detailed comparative analysis of the conformational sampling tools available in MOE and Catalyst, two widely used computational chemistry softwares. Protocols for both detailed conformational sampling and high-throughput 3D library generation were assessed. This used a carefully compiled and characterized test set of 256 publicly available diverse druglike compounds, for which the bioactive conformation has been elucidated crystallographically. Smaller[31,35,37,39−41] or larger[32,33] sets of bioactive conformers have been studied before, but the present study placed particular emphasis on chemical diversity and druglikeness, attempting to represent current trends in pharmaceutical research.

This well-balanced test set of compounds allowed an in-depth examination of the three conformation generators in MOE, i.e. the systematic and stochastic searches and Conformation Import. Comprehensive tests were carefully performed to address questions pertaining to i) identification of the global energy minimum (given a force field), ii) reproduction of the bioactive conformation in detailed or

high-throughput mode, and iii) more general measures of the conformational coverage. This provides MOE users with a detailed guide regarding the usage, scientific performance, and relative compute times of the conformational sampling tools, with specific recommendations for the adjustable parameters. Our tests also supplement previous and very helpful studies of Catalyst FAST/BEST, where the compounds tended to be larger and more flexible.[32,33,35,39]

Overall, the three MOE conformational sampling algorithms complement each other to tackle the various aspects of small molecule conformational modeling in drug research. As expected, the systematic search is slow but provides a benchmark with respect to completeness of conformational sampling and identification of the global energy minimum. For identification of the global energy minimum, the MOE stochastic search performs well and is much faster than the systematic search. This echos results obtained long ago on alkanes[12,51] and demonstrates the robustness of the approach when applied to diverse druglike molecules. Inclusion of chirality as part of the searched space makes the stochastic approach well-suited to organic molecules considered in drug research. However, the stochastic search at default settings did not perform comparatively well for conformational coverage and reproduction of the bioactive conformers. This, combined with long compute times, indicates that the stochastic search is not a method of choice for generation of 3D libraries. Yet, the conformational coverage afforded by the stochastic search increases with the energy window ΔE and is clearly enhanced when using a generalized Born (*GB*) solvation model. Of all the protocols tested here, the stochastic search with $\Delta E = 15$ kcal/mol and *GB* yielded the highest fraction of accurately reproduced (rmsd $\leq$ 1.0 Å) bioactive conformers and the broadest coverage of the pharmacophore space as measured by the number of pharmacophore quadruplets (and triplets). Therefore, this stochastic search variant is a method of choice for detailed characterization of key compounds. Catalyst BEST at default settings is the other method which stood out for generation of diverse conformers associated with a high degree of accurately reproduced bioactive structures.

For 3D library enumeration with MOE, Conformation Import clearly emerged as the appropriate tool. Regarding the percent of reproduced bioactive conformers and compute times, Conformation Import at default settings performed as well as Catalyst FAST. Both default protocols yield conformers within 1.5 Å of the bioactive X-ray reference for ~95% of the compounds, an excellent basis for retrieval of active compounds via pharmacophore searches. With FAST, a maximum number of conformers per compound (MaxConfs) of 50 has been recommended as optimal compromise.[32,33] Our tests suggest that it may be worth keeping MaxConfs to the default 255, especially if a higher proportion of bioactive conformers reproduced within 1.0 Å is sought.

MaxConfs and the energy window ΔE were varied systematically for Conformation Import and FAST/BEST, to measure the impact on reproduction of the bioactive structures. Increasing ΔE above its default values had no obvious benefit, but a decrease degraded performances. Increasing MaxConfs above its default had a marginal effect with FAST/BEST but a more pronounced influence with Conformation Import. Thus, although the default settings of Conformation Import performed reasonably well, one may envisage increasing its MaxConfs, resources permitting. Increasing MaxConfs and ΔE together did not uncover notable synergies with any protocol. Another key conclusion regarding Conformation Import is that results deteriorate significantly when the refinement option is activated, in addition to requiring significantly longer compute times. This can be understood based on simple considerations, hence we advise against the refinement option with Conformation Import for routine work. In summary, we conclude that Conformation Import at default settings performs well as a high-throughput generator of conformational models for druglike compounds.

Regarding Catalyst FAST and BEST, we largely confirm the trends presented in previous work,[32,33] but additional insights were gleaned. The BEST conformational models are distinctive for their enhanced diversity and conformational coverage, leading to a comparatively high rate of reproduced bioactive conformers. BEST is still too slow for large-scale library generation, for which one needs to resort to FAST. For FAST, we suggest that a MaxConf value (~250) higher than recommended previously[32,33] is tractable and worth the effort. Also, we obtained percentages of reproduced bioactive conformers (e.g., within 1.0 Å) notably higher than reported before,[32,33] probably reflecting differences in the properties of the compound sets.

We reiterate that the default energy window of 20 kcal/mol performs optimally for FAST and BEST. A high energy window ($\geq$15 kcal/mol) is also optimal for conformational coverage with MOE stochastic search. Indeed, some of the bioactive-like conformers close ($\leq$0.5 Å or $\leq$1.0 Å) to their X-ray counterpart retain high conformational strains, even after minimization of their MMFF94 force field energies. However, the majority of such bioactive-like conformers present relatively low ($\leq$3 kcal/mol) conformational energies. Possible caveats in this analysis were pointed out, nevertheless it is compatible with the chemically intuitive notion that most druglike compounds bind to their target without excessive conformational strain. The fundamental question of the conformational energetics of bound ligands is difficult yet crucial to understand for improvements in predictive methods. This suggests directions for future work.

Abbreviations. 3D: three-dimensional; *GB*: generalized Born; *Diel*: distance-dependent dielectric constant; QSAR: quantitative structure−activity relationship; MaxConfs: maximum number of conformers generated per compounds; MMFF: Merck Molecular Force-Field; MOE: Molecular Operating Environment; NChi: number of chiral centers per compound; NRot: number of rotatable bonds per compound; PDB: Protein Data Bank; VS: virtual screening; *%BioConf_Rep*: percentage of reproduced bioactive conformations.

ACKNOWLEDGMENT

The authors thank scientists at Accelrys and Chemical Computing Group for answering our questions and helping us with the best use of the programs Catalyst and MOE.

Supporting Information Available: List of PDB entries used for the ligand bioactive conformations, examples of lowest energy minima from the MOE systematic search overlaid on their stochastic search counterparts, and illustration of compounds with low and high percentage of conformational extension. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532–549.

(2) Lemmen, C.; Lengauer, T.; Klebe, G. FlexS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502–4520.

(3) Labute, P.; Williams, C. Flexible Alignment of Small Molecules. *J. Med. Chem.* **2001**, *44*, 1483–1490.

(4) Green, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297–1308.

(5) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(6) Howard, A. E.; Kollman, P. A. An Analysis of Current Methodologies for Conformational Searching of Complex Molecules. *J. Med. Chem.* **1988**, *31*, 1669–1675.

(7) 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 Publishers: New York, 1991; pp 1−55.

(8) Rusinko, A., III; Sheridan, R. P.; Nilakantan, R.; Haraki, K. S.; Bauman, N.; et al. Using CONCORD To Construct a Large Database of Three-Dimensional Coordinates from Connection Tables. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 251–255.

(9) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.

(10) Crippen, G. M. Rapid Calculation of Coordinates from Distance Matrices. *J. Comput. Phys.* **1978**, *26*, 449–452.

(11) Bruccoleri, R.; Karplus, M. Prediction of the Folding of Short Polypeptide Segments by Uniform Conformational Sampling. *Biopolymers* **1987**, *26*, 137–168.

(12) 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*, 4371–4378.

(13) Klebe, G.; Mietmer, T. A fast and efficient method to generate biologically relevant conformations. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 583–606.

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

(15) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1995**, *16*, 171–187.

(16) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible docking using Tabu search and an empirical estimate of binding affinity. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 367–382.

(17) Chang, G.; Guida, W. C.; Still, W. C. An Internal Coordinate Monte Carlo Method for Searching Conformation Space. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.

(18) Foloppe, N.; Nilsson, L. Toward a Full Characterization of Nucleic Acid Components in Aqueous Solution: Simulations of Nucleosides. *J. Phys. Chem. B* **2005**, *109*, 9119–9131.

(19) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the general purpose tripos 5.2 force field. *J. Comput. Chem.* **1989**, *10*, 982–1012.

(20) Momany, F. A.; Rone, R. Validation of the General Purpose QUANTA 3.2/CHARMm Force Field. *J. Comput. Chem.* **1992**, *13*, 888–900.

(21) Halgren, T. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(22) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(23) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general Amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(24) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W.-D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model* **2006**, *46*, 2342–2354.

(25) *Catalyst*; Accelrys, Inc.: 10188 Telesis Court, Suite 100, San Diego, CA 92121, U.S.A. http://www.accelrys.com(Accessed June 11, 2008).

(26) *Omega*; OpenEye Scientific Software: 9 Bisbee Court, Suite D, Santa Fe, NM 87508. http://www.eyesopen.com (accessed June 11, 2008).

(27) *Phase*; Schrodinger: 120 West 45th Street, 29th Floor, New York, NY 10036-4041. http://www.schrodinger.com (accessed June 11, 2008).

(28) *MOE*; Chemical Computing Group: 1010 Sherbrooke St. W, Suite 910, Montreal, Quebec, Canada H3A 2R7. http://www.chemcomp.com (accessed June 11, 2008).

(29) Borodina, Y.; Bolton, E.; Fontaine, F.; Bryant, S. H. Assessment of Conformational Ensemble Sizes Necessary for Specific Resolutions of Coverage of Conformational Space. *J. Chem. Inf. Model* **2007**, *47*, 1428–1437.

(30) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2002**, *21*, 449–462.

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

(32) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative Analysis of Protein-Bound Ligand Conformations with Respect to Catalyst's Conformational Space Subsampling Algorithms. *J. Chem. Inf. Model* **2005**, *45*, 422–430.

(33) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the Retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46*, 1848–1861.

(34) Nicklaus, M. C.; Shaomeng, W.; Driscoll, J. S.; Milne, W. A. Conformational Changes of Small Molecules Binding to Proteins. *Biorg. Med. Chem.* **1995**, *3*, 411–428.

(35) Perola, E.; Charifson, P. S. Conformational Analysis of Drug-Like Molecules Bound to Proteins: An Extensive Study of Ligand Reorganization upon Binding. *J. Med. Chem.* **2004**, *47*, 2499–2510.

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

(37) Bostrom, J.; Norrby, P. O.; Liljefors, T. Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383–396.

(38) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; et al. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726–741.

(39) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational Sampling of Bioactive Molecules: A Comparative Study. *J. Chem. Inf. Model.* **2007**, *47*, 1067–1086.

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

(41) Gunther, S.; Senger, C.; Michalsky, E.; Goede, A.; Preissner, R. Representation of target-bound drugs by computed conformers: implications for conformational libraries. *BMC Bioinformatics* **2006**, *7*, 293–304.

(42) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; et al. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(43) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL "Keys" as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.

(44) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(45) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285–294.

(46) Ming-Hong, H.; Haq, O.; Muegge, I. Torsion Angle Preference and Energetics of Small-Molecule Ligands Bound to Proteins. *J. Chem. Inf. Model.* **2007**, *47*, 2242–2252.

(47) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference between Leads and Drugs? A historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(48) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3–25.

(49) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 856–864.

(50) Ferguson, D. M.; Glauser, W. A.; Raber, D. J. Molecular mechanics conformational analysis of cyclononane using the RIPS method and comparison with quantum-mechanical calculations. *J. Comput. Chem.* **1989**, *10*, 903–910.

(51) Saunders, M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; et al. Conformations of Cycloheptadecane. A Comparison of Methods for Conformational Searching. *J. Am. Chem. Soc.* **1990**, *112*, 1419–1427.

CI800130K