Drug-like Bioactive Structures and Conformational Coverage with the LigPrep/ConfGen Suite: Comparison to Programs MOE and Catalyst

I-Jen Chen* and Nicolas Foloppe*

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

Received January 17, 2010

Computational conformational sampling underpins many aspects of small molecule modeling and design in pharmaceutical work. This work examined in detail the widely distributed LigPrep/ConfGen software suite and the conformational models it produces for drug-like compounds. We also compare LigPrep/ConfGen to MOE and Catalyst. Tests of the conformational sampling protocols included the reproduction of known bioactive structures of ligands, characterization of the size, coverage and diversity of the output conformational models, and relative computation times. The present tests will help the user to make informed choices among the predefined ConfGen protocols (Very fast, Fast, Intermediate, and Comprehensive), and the adjustable input parameters. The parameters governing the initial compound preparation (LigPrep) and the subsequent conformational sampling were explored. This analysis has led to a new protocol called "ConfGen Optimized", which improves upon the predefined protocols. ConfGen Optimized is computationally tractable and reproduced 80% of the bioactive structures within 1 Å, versus 66% for the default ConfGen Fast protocol. We also addressed the issue of the reproduction of compact/folded bioactive structures by ConfGen. It involved the compilation of a new set of 50 folded diverse drug-like bioactive structures. This indicates that heuristics penalizing folded conformers hinder reproduction of some binding modes. Overall, ConfGen offers great flexibility of use and provides a valuable addition to the molecular modeling toolbox.

1. INTRODUCTION

Many computational chemistry tools used for the design of small bioactive molecules, e.g. in medicinal chemistry, operate on the three-dimensional (3D) conformations of these molecules. Widely used methods requiring such 3D conformations include pharmacophore modeling,[1–3] 3D QSAR,[4] docking to a receptor,[5] scaffold hopping,[6,7] shape analysis,[8] and modeling of X-ray structures in electron density maps.[9] Thus, generating these 3D conformers is a fundamental step underpinning a broad range of methods. Here, we concentrate on the conformations of small drug-like molecules, where much of the pharmaceutical interest resides.

It is now accepted that usually more than a few conformers per compound need to be considered, and those are typically obtained from a computational conformational sampling algorithm. Broadly speaking, there are two fundamental components to computational sampling methods, a conformer generator algorithm and a scoring scheme which accepts or rejects the generated conformers. Many approaches have been devised to generate conformers of drug-like compounds, including systematic searches, Monte Carlo moves, distance geometry constraints, genetic algorithms, build-up approaches, heuristic rules, and molecular dynamics.[10] The likelihood of the conformers is usually assessed with an energy scheme derived from a molecular-mechanics force field. The field of conformational sampling of drug-like molecules has recently been reviewed,[10] and details can be found in that review and elsewhere.[11,12]

Large conformational libraries of compounds are key to the discovery of novel active chemical entities by virtual screening, e.g. via pharmacophore-based or scaffold-hopping searches, and can also be used in conjunction with docking. So, the methods for conformational analysis and sampling of small organic molecules remain a very active field of research.[10,13–29] New developments have taken advantage of increased computational resources as well methodological advances, e.g. around the distance geometry approach.[19,30] Overall, these efforts have led to the availability of high-throughput conformation generators, such as CAESAR,[31] Catalyst,[32] ConfGen,[33] MOE Conformation Import,[34] Omega,[35] and Rubicon.[36] Other proprietary conformation generators have also been developed,[13,19] as well as some freely available programs.[37–40]

However, the quality and relevance of the generated conformers cannot be taken for granted, as they are dependent on the underlying sampling algorithm and force fields. Therefore, extensive tests of the programs are needed to gain confidence in their output and assess the relevance and coverage of the generated conformational ensembles. Tests studies of commercially available programs have been presented for CAESAR,[31] Catalyst,[18,19,22,24,25,41] MOE Conformation Import,[13,18,19] Omega,[15,24,28,42] and Rubicon.[19] Yet, to our knowledge, an independent detailed analysis of the conformers generated by the LigPrep/ConfGen software suite is not available. The ConfGen algorithm was initially developed as the conformation generator underlying the widely used docking program Glide.[5] Since, ConfGen has also underpinned the pharmacophore perception tool PHASE,[2] and ConfGen is now available as a stand-alone for the generation of 3D compound libraries.[29]

* To whom correspondence should be addressed. Phone: + (44) (0) 1223 895 354 (I.C.); + (44) (0) 1223 895 338 (N.F.). Fax: + (44) (0) 1223 895 556 (I.C.); + (44) (0) 1223 895 556 (N.F.). E-mail: i.chen@vernalis.com (I.C.); n.foloppe@vernalis.com (N.F.).

10.1021/ci100026x
Published on Web 04/27/2010

The aim of the present work is to present detailed tests of the program ConfGen stand-alone, which allows one to focus on the intrinsic quality and breadth of the generated conformers. These tests include the preliminary conversion from a 2D to 3D structure with the program LigPrep, also part of the same software suite. LigPrep offers a number of options regarding the choice of force field and the treatment of tautomers,[43] which can influence the subsequent ConfGen output. Therefore, we investigated a number of combined LigPrep/ConfGen protocols.

For small molecules, a key test of any conformation generator is whether it reproduces known bioactive structures,[10,18,19,22,24,25,28,41,42] as such structures are of obvious relevance to medicinal chemistry. Such tests have recently gained momentum, thanks to the increasing number of X-ray structures of protein−ligand complexes deposited in the Protein Data Bank (PDB). We have recently curated a set of 253 diverse drug-like ligands for which the bioactive X-ray structure was solved at satisfactory resolution[10,18] and is available from the PDB. This is a generic compound set, in that it was compiled to be chemically diverse while retaining conventional drug-like properties (section 2.1). Thus, we refer to this set of 253 compounds as the "Vernalis generic compound set". It was used to test Catalyst (both the FAST and BEST regimes), as well as the MOE conformational search algorithms (systematic and stochastic searches, and high-throughput Conformation Import protocol). Therefore, here, we use again the Vernalis generic compound set to test the LigPrep/ConfGen suite, to allow direct comparison to the results obtained with Catalyst and MOE Conformation Import.

In addition, as part of the present work, we compiled a second set of 50 drug-like compounds, comprised of "folded" X-ray bioactive structures. Here, "folded" means that the ligands adopt compact bioactive conformations, where intramolecular stacking interactions are mediated by flexible linkers. So, such molecules could adopt more extended conformations but bind in a folded state. There were several motivations behind the compilation of this folded set of compounds. First, the possible preferences of ligands regarding their compactness/extendedness when binding to proteins has been a topic of discussion,[19,22,26,44–46] but few studies have addressed the question concretely for smaller drug-like molecules. The distance geometry framework allows one to drive ligand conformations toward greater or lesser extendedness.[19] Some authors have argued that ligands prefer to adopt extended bioactive conformations,[19,26,44,45] which is intuitively appealing as extended conformations allow more (favorable) contacts between the ligand and protein. Yet, compact bioactive conformations have also been reported,[22,46] in particular for flexible natural cofactors.[46] Nevertheless, the perception that ligands "prefer" to bind in extended conformations to proteins dominates. Indeed, the initial description of ConfGen as part of the Glide algorithm,[5] and the more complete ConfGen user manual, stipulate that ConfGen uses heuristic rules which tend to eliminate compact conformers. This provided the second motivation to assemble a set of folded bioactive reference structures, to explore the impact and justification for such heuristic rules. For comparisons, new calculations on this folded compound set were also performed with MOE Conformation Import (default settings) and a previously optimized MOE stochastic search (with a generalized Born solvation model and an energy window of 15 kcal/mol).[18]

Table 1. Average Physico-Chemical Properties of the Vernalis Generic and Folded Compound Sets

compound set	N^a	MW	NRot	SlogP	HBAcc	HBDon	rings
generic	253	373.2	5	2.5	5	2	3
folded	50	498.1	8	3.3	6	3	4

[a] Number of compounds per set. There are 17 compounds in common between the generic and folded sets. Properties are MW (molecular weight), NRot (number of rotatable bonds), SlogP (logarithm of the calculated octanol−water partition coefficient),[62] HBAcc and HBDon (number of hydrogen bond acceptors and donors, respectively), and number of rings. These properties were calculated with MOE.[34]

The distribution of compactness of the computed conformers is in fact also of interest to characterize the diversity of the generated conformers and the corresponding conformational coverage.[13,18,19] A finer measure of conformational coverage relies on the number of 3D pharmacophores visited by a conformational ensemble.[13,18,19] Both measures are used here to characterize the conformational coverage afforded by the tested protocols with the generic compound set.

The generic compound set provides the basis for the majority of the tested ConfGen protocols. We first examined the four predefined "ConfGen Standard" protocols ("Very fast", "Fast", "Intermediate", and "Comprehensive"), as they are presented to the user (with the same nomenclature). We then used the "ConfGen Advanced" functionality of the program to explore how some key parameters influence the output. We found that it is beneficial to adjust some of the default parameters, to construct a new ConfGen protocol that we call "ConfGen Optimized". This new protocol also involves nondefault options in LigPrep.

Overall, the present work provides an independent and comparative assessment of major components of a widely used modeling suite and should therefore be of general interest to users of these programs. In addition, the detailed attention paid to the properties of the reference bioactive conformations, in particular the folded compounds, addresses some fundamental aspects of ligand−protein recognition.

2. METHODS

2.1. The Vernalis Generic Compound Set. The generic set of drug-like compounds was essentially the same as in a previous study, which assessed conformational sampling methods from the programs MOE and Catalyst.[18] This allows a direct comparison between this previous study and the present one. This "Vernalis generic set" contains 253 unique ligands, instead of the 256 initially reported, following the identification of three duplicate compounds. A detailed analysis of these compounds' physicochemical profile has been presented,[18] but the key elements are summarized here and in Table 1. The compounds are drug-like, chemically diverse, and from noncovalent protein−ligand complexes in the Protein Data Bank (PDB),[47] including a broad range of protein families. Most complexes were determined at a crystallographic resolution ≤ 2.5 Å. The updated list of PDB codes is given in Table S1 in the Supporting Information, together with the compounds in the "folded set" (section 2.2).

All ligands were prepared as described before.[18] The assignment of bond orders and standard protonation states

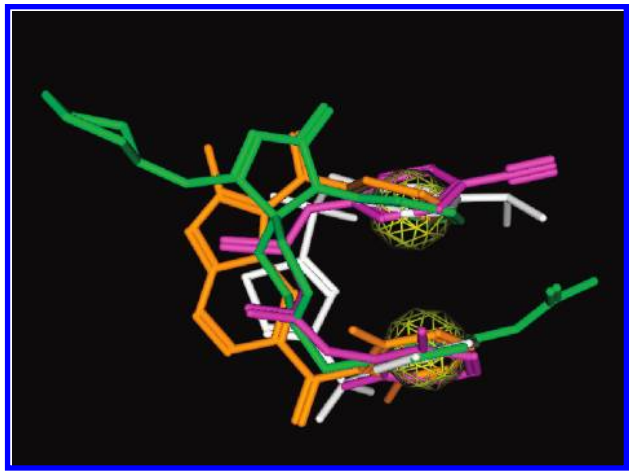


Figure 1. Illustration of the fit of folded bioactive ligand X-ray structures to a two-point pharmacophore associated with intramolecular stacking. Each pharmacophore feature (yellow sphere) can be an aromatic (Aro) or a saturated ring (Hyd). Each feature has a tolerance radius of 1.2 Å, and the distance between the two centroids is 4.0 Å. Four pharmaceutically relevant ligands from the Vernalis folded set are shown, clearly indicating their folded character. White, bound to β-lactamase in 1GA9; magenta, bound to HIV-1 reverse transcriptase in 1S6P; orange, bound to CDK2 in 1PYE; green, bound to β-secretase in 3FKT.

at pH = 7 were checked manually. The tautomers were selected to be consistent with the X-ray binding modes. Prior to conformational analysis with any program, compounds in the generic and folded (see below) sets were converted to 2D representations to erase the crystallographic structural information.

2.2. Selection of the Vernalis Folded Compound Set. The ConfGen manual mentions that "ConfGen also tends to eliminate compact ligand conformations".[33] More generally, the degree of folding (compactness) of bound ligands is of interest regarding molecular recognition and the associated modeling protocols.[19,26,44,46] Thus, as part of the present work, we have compiled a set of drug-like "folded compounds", to investigate the existence and the treatment of such conformers.

The folded bioactive ligand structures were extracted from the PDB.[47] Their X-ray bioactive structures were rigidly tested against a two-point pharmacophore (defined with MOE, Figure 1) aimed at detecting intramolecular stacking interactions. The folded ligands were clustered using MACCS fingerprints at 70% similarity (Tanimoto index), and only one member of each cluster was kept, selecting the compound already in the generic set when applicable. These candidate bioactive structures and their fit to the pharmacophore were inspected manually, in particular to remove compounds with no flexibility between the stacking moieties. Only compounds where these moieties are separated by at least three rotatable bonds, and that can be either stacked intramolecularly or in a more extended conformation, were kept. Thus, the final "folded set" is comprised of 17 compounds from the Vernalis generic set and 33 additional compounds crystallized with proteases, kinases, and protein−protein interaction interfaces (Table S1 in the Supporting Information). In sum, there are 50 nonredundant compounds in the Vernalis set of folded bioactive conformations. Figure 1 illustrates how selected folded bioactive structures fit the two-point pharmacophore.

On average, the folded ligands are larger and more flexible than those in the generic set (Table 1), as may be expected. Yet, the physicochemical properties of those folded ligands remain drug-like.

The crystallographic resolution of a majority of 39 complexes in the folded set is ≤ 2.5 Å. It was decided to keep the other 11 complexes with a crystallographic resolution >2.5 Å, considering their reasonable binding modes. Despite its apparently particularly poor nominal resolution (3.4 Å), it was deemed interesting to keep pactamycin in the set (code 1HNX, resolution of 3.4 Å) as it exemplifies binding to an RNA target. The actual accuracy of RNA X-ray structures is better than expected from their nominal resolution, in part due to the presence of electron-rich phosphates in the RNA backbone.[48]

2.3. Conversion from 2D to 3D with LigPrep. Input to ConfGen required compounds in 3D format.[33] The 2D-to-3D conversion was performed with the Schrodinger LigPrep program,[49] belonging to the same software suite as ConfGen. The LigPrep 2D-to-3D conversion includes options for generating multiple states for possible tautomers, stereoisomers, ionization at a selected pH range (7 ± 2 by default), and ring conformations (1 ring conformer by default). In LigPrep, the ionization states and tautomers can be assigned with Ionizer (default) or Epik.[43] Those default settings were kept for much of the study, but several LigPrep parameters were also explored to test various LigPrep/ConfGen combinations (section 3.4). The final step of a LigPrep preparation is an energy minimization of the 3D conformers with the OPLS_2005 (default) or MMFF94s force field. Most ConfGen experiments (sections 3.1, 3.2, and 3.3) used input conformers energy-minimized with the LigPrep default OPLS_2005. Specific comparisons of input conformers energy minimized with OPLS_2005 or MMFF94s are presented in section 3.4.

2.4. The ConfGen Conformational Search Method. We used ConfGen version 2.1.[33] ConfGen has been the underlying conformation generator for other computational tools from Schrodinger, like Glide (docking[5,50]) and phase (pharmacophore perception[2]). ConfGen was recently introduced as a stand-alone module for general conformational sampling.[29] It offers ready-to-use protocols, one aim being the generation of bioactive conformations while keeping the total number of conformations manageable.[29,33] This section summarizes the main aspects of the ConfGen algorithm (Figure 2), based on the ConfGen and Glide user manuals distributed by Schrodinger.

First, a compound is analyzed to detect the conformational degrees of freedom, i.e., rotatable and amide bonds, asymmetric pyramidal trigonal nitrogens, and flexible rings. Terminal CH_3, NH_2, and NH_3 groups are not treated as rotatable. For rings, a predefined library of conformers is used. Asymmetric pyramidal trigonal nitrogens are inverted.

For each rotatable bond, its potential energy is calculated with the torsional term of the OPLS_2001 force field. The energy of a conformer is estimated as the energy of the ring conformations added to the energy associated to each rotatable bond. This sum of energies is referred to as the "*Cen* energy" (or ConfGen energy).

ConfGen divides a molecule into a core and peripheral "rotamer groups" (Figure 2A). A rotamer group is defined as a terminal appendage that does not contain a rotatable bond but is attached to the rest of the molecule by a rotatable bond, referred to as a terminal rotatable bond (TRB). Thus,

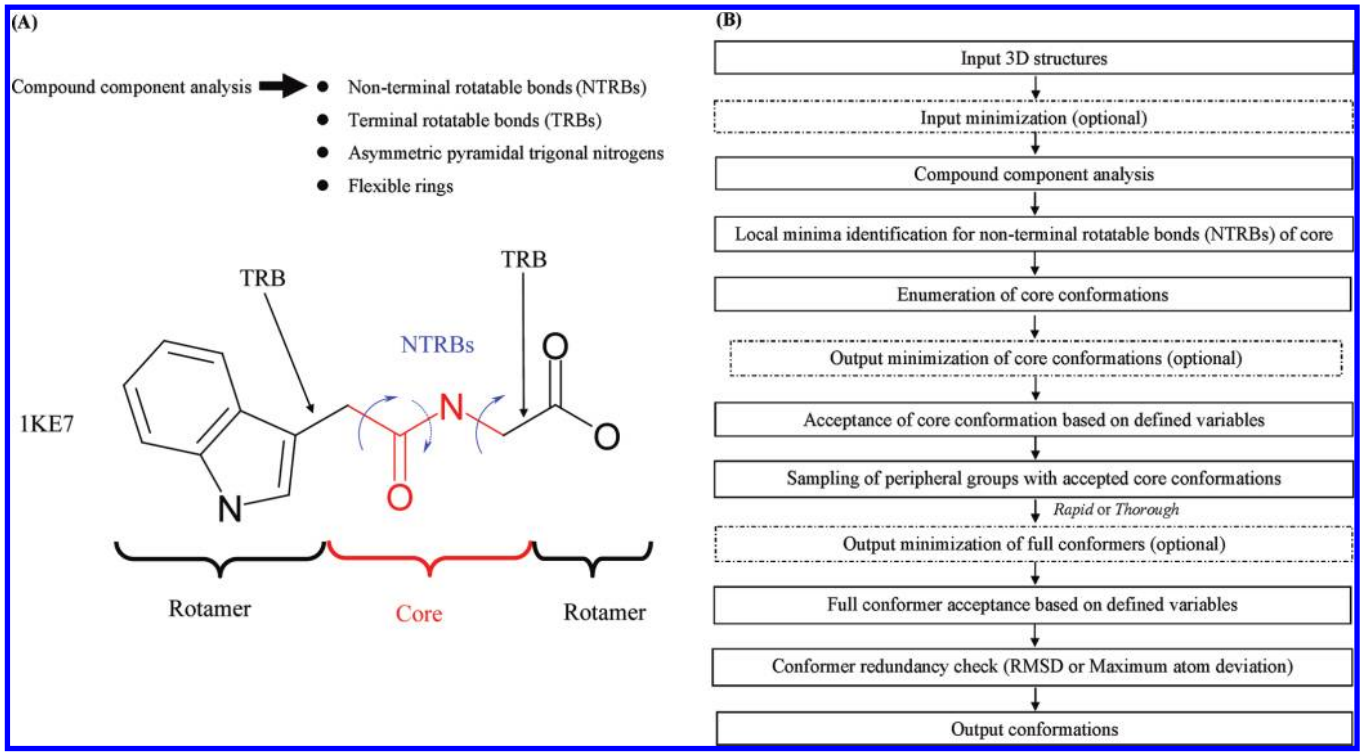


Figure 2. Summary of the ConfGen algorithm and workflow. Prior to ConfGen sampling, each compound is dissected in structural components (panel A). The ligand from PDB entry 1KE7 is used to illustrate this component analysis. The ConfGen overall workflow is in panel B.

all groups connected by nonterminal rotatable bonds (NTRBs) are part of the core.

After partitioning into core and rotamer groups, conformational sampling of the core only is performed, on a fully built molecule. At this stage, the rotamer groups are positioned with their connecting TRB in their lowest energy minima. Conformers with a *Cen* energy greater than a cutoff value are discarded. In the remaining conformers, the torsions are adjusted to minimize the *Cen* energy, under a short-range excluded-volume energy term which reduces atom collisions. Only a set maximum number of core conformers (*maxcore*) is retained. If more conformers are generated, their number is reduced to *maxcore* by eliminating the more compact conformers.

After generation of the core conformers, the terminal rotatable bonds (TRBs) are sampled to position the rotamer groups. Either the TRBs are sampled one after another (*Rapid* mode) or all combinations of the TRB energy minima are tested (*Thorough* mode). Conformations with relative energies exceeding a preset cutoff are eliminated. The remaining structures are energy minimized in the same manner as the core conformations.

Redundant conformers are discarded by one of two criteria, a pairwise RMSD or a maximum deviation between atoms (after best fit of one conformer onto the other). With the maximum atom deviation, if any distance between a pair of corresponding atoms is above a cutoff value, the two conformers are considered different and kept. The default criterion for duplicate removal is a pairwise RMSD $\leq$ 1 Å. The present work used the default RMSD criterion, with several other cutoff values for duplicate removal tested.

According to the user manual, "To better simulate the structures found in protein−ligand complexes, ConfGen also tends to eliminate compact ligand conformations". For example, elimination of the most compact ("folded") conformers reduces the number of conformers below *maxcore* if needed (see above). Penalties screening out folded conformers are also applied after the full conformational search, using a number of heuristics. One heuristic rule is the elimination of conformers with "long-range internal hydrogen bonds",[5] formed between topologically distant parts of the molecule. In addition, conformers with close contacts between nonhydrogen atoms separated by at least three bonds are penalized, particularly if the groups approaching each other are rings (ConfGen manual). This penalty increases with the number of atom pairs in such close contacts, and if the penalty is above a certain threshold, the conformer is removed. These heuristics disfavoring folded conformers can be turned on and off with variables CGO6 and CHYD in the command line mode. CGO6 controls ConfGen capabilities for rejecting conformations that have too many nonhydrogen atoms in close proximity. CHYD influences the formation of intramolecular hydrogen bonds by turning on or off the corresponding electrostatic component. The interested reader is referred to the ConfGen manual for more details. The important point here is that by default ConfGen disfavors folded conformers. The implications of this bias for reproduction of bioactive structures is investigated in section 3.7, based on the compound set described in section 2.2.

2.5. ConfGen Standard or Advanced. Regarding its interface with the user, ConfGen can be run in two ways, called Standard or Advanced. With ConfGen Standard, one chooses one of four predetermined protocols: Very fast, Fast, Intermediate, or Comprehensive. The four protocols differ by the extent of the conformational search, the RMSD cutoff for conformer duplicate removal, and the permitted energy window ΔE (Table 2). For all protocols except the Very fast

Table 2. Default Parameters for ConfGen Standard Protocols

parameters	predefined protocols			
	Very fast	Fast	Intermediate	Comprehensive
sampling steps per degree of freedom (RotSteps)	5	5	75	75
duplicate RMS (Å)[a]	1.25	1	1	0.5
MaxConfs[b]	1000	1000	1000	1000
ΔE (kcal/mol)[c]	n/a	25	25	120
minimization of input structures	none	none	none	none
minimization of output structures	none	none	none	none

[a] RMSD cutoff (Å) to remove duplicate conformers. [b] Maximum number of generated conformers. [c] Allowed energy window.

protocol, conformers are subjected to energy filtering before output. The default ConfGen Standard protocol is ConfGen Fast.

In ConfGen Advanced, the user has greater control of adjustable parameters, which include the force field (OPLS or MMFF), the optional treatment of solvation (GB/SA),[51] minimization procedures, and other ConfGen variables mentioned in Table 2. Parameters accessible from the ConfGen Advanced graphical interface are only a subset of all the parameters. For example, other variables not exposed but differing between ConfGen Standard Fast and ConfGen Advanced default include the permitted interatomic approach distance (0.25 Å for ConfGen Standard and 0.00 Å for ConfGen Advanced). So, the only way to strictly control the computational experiments is by checking and editing the input file parameters and using these command files as input directly from the command line. Thus, to assess the individual impact of selected parameters, we edited the ConfGen Fast input file. The computed output was then compared to that of ConfGen Fast, as it is the default protocol.

2.6. Generation and Evaluation of Multiconformer Libraries with ConfGen. We tested multiple ConfGen protocols, on a previously described set of 253 ligands[18] (section 2.1). ConfGen Standard, ConfGen Advanced, and additional customized protocols were thoroughly tested with respect to the generation of multiconformer libraries. Because it is the default option, ConfGen Standard Fast was used as the benchmark for comparison to other ConfGen protocols.

First, the four predefined ConfGen Standard protocols (Very fast, Fast, Intermediate, and Comprehensive) were compared (section 3.1). In those protocols, the underlying parameters cannot be altered. So, the next step was to explore the key parameters of the conformational search with ConfGen (sections 3.2 and 3.3). That included the "*Rapid*" versus "*Thorough*" search mode, the use of a generalized Born (GB) solvation model,[51] the number of search steps per degree of freedom (RotSteps), the energy window (ΔE, reported in kcal/mol), the RMSD cutoff for removal of duplicate conformers, and the maximum number of allowed conformers (MaxConf). The ConfGen protocols were also combined with various ligand preparations by LigPrep, with OPLS_2005 or MMFFs force fields, with Epik or Ionizer generation of tautomers (section 2.3). In total, 47 LigPrep/ConfGen protocols were tested.

The produced conformational ensembles were tested with respect to several criteria, i.e., the average number of generated conformers (NbConfs), the average computation time per compound, and how well the generated conformational ensemble reproduced the known bioactive structures. A bioactive structure was considered reproduced if the RMSD with at least one computed conformer was within a given threshold. Four thresholds were investigated, 0.5 Å, 1.0 Å, 1.5 Å, and 2.0 Å. The 1.0 Å threshold is emphasized in the analysis of the results, because it arguably represents a good quality fit of practical value. The percentage of bioactive conformations reproduced by a protocol at a given RMSD threshold is abbreviated as *%BioConf_Rep*. NbConfs and *%BioConf_Rep* give simple, but relevant, estimates of the conformational coverage. For selected protocols, a more complete assessment of the conformational coverage based on 3D descriptors was pursued (sections 2.8 and 3.6).

These tests led to the devising of a new combined LigPrep/ConfGen protocol termed "ConfGen Optimized", which yielded significant improvements regarding the reproduction of the bioactive structures (section 3.4), while remaining computationally tractable.

The results obtained with ConfGen are put in context by comparing to those previously obtained on the same compound set with MOE and Catalyst.[18]

2.7. MOE and Catalyst Conformational Search Algorithms. A detailed description of the conformational search algorithms from the softwares MOE (from Chemical Computing Group[34]) and Catalyst (from Accelrys[32]) has been presented before.[18,25] The present work compares the newly obtained ConfGen results on the Vernalis generic set of compounds with their counterpart presented previously with MOE (Conformational Import and Stochastic Search protocols) and Catalyst (FAST and BEST protocols).[18] Thus, the Catalyst calculations, and some of the MOE data reported here, were published before[18] but are summarized again alongside the ConfGen results to allow for comparison. The parameters controlling the algorithms compared across programs are significantly different (Table 3). Therefore, the results must be assessed in this context.

In addition, new conformational sampling calculations were performed with MOE 2008.10[52] on the newly compiled set of folded compounds (sections 2.2 and 3.7), with the default Conformation Import approach and a stochastic search protocol optimized previously.[18] This optimized stochastic search differs from the default settings by using an increased energy window $\Delta E = 15$ kcal/mol and a generalized Born solvation model.[53]

2.8. Conformational Coverage Assessed by 3D Descriptors. An important aspect of conformational sampling is its coverage, i.e., how completely and finely the conformational space of a molecule is represented.[10,20,54] This is already probed to some extent by simple properties, such as the number of generated conformers (NbConfs) and the percent of reproduction of the bioactive structures (*%BioConf_Rep*).

Table 3. Key Parameters Governing the Conformational Sampling Methods Compared Across Programs ConfGen, MOE,[a] and Catalyst[b]

	ConfGen Fast	MOE Conformation Import	MOE optimized stochastic	Catalyst FAST	Catalyst BEST
force field	modified OPLS2001[c]	MMFF94x	MMFF94x	modified CHARMm[d]	modified CHARMm[d]
energy window (ΔE, kcal/mol)	25	4	15	20	20
MaxConfs[e]	1000	250	10000	255	255
duplicate RMS (Å)[f]	1.00	0.15	0.10		
solvation	no	no	yes[g]	no	no

[a] With MOE, either the Conformation Import protocol or a stochastic search was used. [b] Catalyst can use either the FAST or the BEST protocols. [c] Includes only torsional contributions and conformational energies of flexible rings. [d] Electrostatic term turned off. [e] Maximum number of allowed conformers. [f] RMS deviation value above which duplicate conformations are removed. [g] Generalized Born solvation model.[53]

Such properties are straightforward to interpret and allow for quick comparisons across protocols. However, they only provide partial insights into conformational coverage. A much more complete view of conformational coverage can be obtained with 3D descriptors which characterize the overall distribution of conformers.[18,19,26,44] Such 3D descriptors may address the range of compactness of the conformers,[18,19,26,44] or the diversity of the 3D pharmacophores that they cover.[18,19] Here, we used both types of 3D descriptors to quantify further the coverage of selected conformational models.

The compactness/degree of conformational extension for each conformer was characterized with the radius of gyration (Rgyr), calculated with MOE as before.[18] For a given compound, the percent of conformational extension ConfExt was defined with reference to the most compact conformer (Min_Rgyr):

$$\text{ConfExt} = 100(\text{Max\_Rgyr} - \text{Min\_Rgyr})/\text{Min\_Rgyr}$$

Max_Rgyr and Min_Rgyr were the maximum and minimum values of Rgyr in the conformational ensemble generated for a compound, by a given protocol. Thus, if Max_Rgyr (most extended conformer) is twice Min_Rgyr, the most extended conformer is 100% more extended than the most compact. ConfExt can be greater than 100%. By comparing the geometric differences between the most extended and most compact conformers (and their distributions) across protocols, ConfExt can be seen as a crude indicator of conformational diversity.

A finer view of the overall conformational diversity (encompassing all the compounds) afforded by a sampling method was obtained from a count of the visited 3D pharmacophore, calculated via the MOE 3D-pharmacophore fingerprint (piDAPH4, four-point pharmacophores). This fingerprint characterizes every conformer with all its four-point pharmacophores, each pharmacophore corresponding to a distinct "bit".[52,55] Each bit encodes the specific pharmacophoric features and their distances. The pharmacophoric features were hydrophobic groups, hydrogen-bond donors and acceptors, as defined as in MOE. For each conformational protocol, given all different pharmacophores for the 253 compounds, only one occurrence of each (visited) pharmacophore was kept. The total number of such occurrences for a conformational protocol gives the total number of different nonredundant pharmacophores, which is an estimate of the sampling of the pharmacophoric space. Because each visited pharmacophore is counted only once, the total number of pharmacophores associated with a conformational protocol is referred to here as the total number of visited pharmacophores per protocol. Previously, we referred to this number as the total number of "unique" pharmacophores,[18] but it has been pointed out that this terminology is ambiguous, and therefore we now adopt the term "visited" instead of "unique". This total number of visited pharmacophores characterizes the conformational coverage in a way relevant to molecular recognition.

2.9. Docking of Folded Ligands. Ligands from three complexes in the folded set (section 2.2) were subjected to Glide docking, because it uses ConfGen for its conformational search. This is evidently not meant to be an extensive analysis of Glide, but to supplement our investigations of ConfGen with folded bioactive structures. The protein structures from 2ORS, 3FKT, and 3FZM were prepared using the Schrodinger Protein Preparation Wizard. The interaction grids were prepared with all nonprotein residues removed from the binding sites, except with 2ORS for which the cofactor was retained. The docked ligands were prepared with LigPrep at the default setting. The docking was performed with Glide Standard Precision, keeping up to 20 poses per ligand. The docking poses were ranked on the basis of the total docking scores ("docking score"). All poses were compared with their experimental counterparts.

3. RESULTS AND DISCUSSION

We assessed the performance of the program ConfGen with a previously curated generic set of 253 protein−ligand complexes[10,18] and a newly compiled set of 50 "folded" bioactive structures. Considering the intended applications for ConfGen, its assessment emphasizes two criteria. The first is the ability to reproduce the bioactive conformations as determined by X-ray crystallography, available for each test compound. The second is the coverage of conformational and pharmacophore spaces by the generated conformers, using 3D descriptors adapted to this task. For consistency, ConfGen is tested in combination with LigPrep, the associated Schrodinger tool for 2D to 3D conversion, and ionization/tautomer enumeration.[43] We tested in total 47 different LigPrep/ConfGen protocols.

3.1. ConfGen Standard. This section addresses all four predefined ConfGen Standard protocols (Very fast, Fast, Intermediate, and Comprehensive), with ligands prepared by LigPrep at default options. ConfGen Fast is the default of the four ConfGen Standard protocols and is therefore used as a reference for comparisons. The four protocols are subjected to the same maximum number of allowed conformers per compound (MaxConfs = 1000) but differ with

respect to the following parameters (Table 2): sampling steps per degree of freedom (RotSteps), RMSD cutoff below which duplicate conformers are removed (called duplicate RMS in the remaining text), and the energy window ΔE within which the conformers are retained.

Using the generic set of 253 ligands, the ConfGen generated conformer closest (lowest RMSD) to the X-ray bioactive structure was identified and this RMSD recorded. If this RMSD was below a given threshold, the bioactive conformer was deemed "reproduced" for this ligand (Table 4). The four tested thresholds were 0.5 Å, 1.0 Å, 1.5 Å, and 2.0 Å. The percent of compounds for which the bioactive reference is reproduced at a given RMSD threshold is noted as *%BioConf_Rep*.

Table 4 summarizes the results, including (i) *%BioConf_Rep* at the four RMSD thresholds, (ii) the average number of conformers output per ligand (NbConfs), and (iii) the computational speed per protocol (seconds per ligand). Of course, this speed depends on hardware specifications and is only presented for relative comparisons.

The default ConfGen Standard protocol (Fast) reproduced 66% of the bioactive structures within RMSD ≤ 1.0 Å, and 98% within RMSD ≤ 2.0 Å. Comparing ConfGen Standard Fast to the other Standard protocols uncovered a range of *%BioConf_Rep* values at RMSD ≤ 1.0 Å, while at RMSD ≤ 2.0 Å the four protocols plateau close to or at the maximum possible *%BioConf_Rep* values (98−100%). So, at a RMSD threshold of 2.0 Å, virtually all the bioactive conformers are reproduced by ConfGen. For precise pharmacophore modeling, a RMSD threshold of 1.0 Å is arguably more relevant and leads to *%BioConf_Rep* values more sensitive to the selected protocol, as *%BioConf_Rep* ranges from 62% (Very fast) to 77% (Comprehensive).

No energy window filter is applied in ConfGen Standard Very fast, contrary to the other ConfGen protocols (Table 2). This is consistent with Very fast being twice faster than Fast. The energy filtering resulted in a smaller average number of generated conformers (NbConfs) with Fast (50) than with Very fast (60). Thus, Very fast accesses a slightly broader conformational space than Fast. Delineating the conformational space with the Fast energy window (ΔE = 25 kcal/mol) does not penalize the retrieval of bioactive conformations compared to Very fast within RMSD ≤ 1.0 Å, with *%BioConf_Rep* being 4% higher with Fast than with Very fast.

Compared to Fast, the Intermediate protocol has a higher number of sampling steps per degree of freedom (RotSteps), keeping the same energy window (Table 2). Increasing RotSteps from 5 (Fast) to 75 (Intermediate) yields 8% more reproduced bioactive conformations at RMSD ≤ 1.0 Å (Intermediate: 74%; Fast: 66%). The increased sampling resulted in a three-fold increase in NbConfs, so the conformational space is covered more thoroughly by Intermediate than by Fast.

In the Comprehensive protocol, the energy window is extended to ΔE = 120 kcal/mol and the RMS for duplicate removal reduced to 0.5 Å. Consequently, the Comprehensive approach reproduced more bioactive conformations than any other ConfGen Standard protocols (within RMSD ≤ 1.0 Å, *%BioConf_Rep* = 77% for Comprehensive versus 66% for Fast). ConfGen Standard Comprehensive also produced a notably larger NbConfs value (Table 4), 10 times and 3.4 times larger than with ConfGen Standard Fast and ConfGen Standard Intermediate, respectively. ConfGen Comprehensive took 10 times longer than the default ConfGen Fast. As expected, the computation time (Table 4) increased dramatically from Very fast to Comprehensive, corresponding to increasingly extensive searches of the conformational space.

The above highlights the influence of the protocol parameters (RotSteps, ΔE, and the RMS cutoff for duplicate removal) on the recovery of the bioactive structures, although pairwise comparisons between the ConfGen Standard protocols are not sufficient to finely dissect the impact of each parameter. Indeed, the ConfGen Standard predefined protocols tend to differ by more than one parameter (Table 2), so customized tests were required to disentangle the effects of these parameters (section 3.3).

Overall, the four ConfGen Standard protocols offer a good degree of flexibility to the user, covering a range of computation times, with performances which vary accordingly regarding *%BioConf_Rep* and NbConfs. The benchmarks given in Table 4 should help the user choose among the four protocols. For low resolution work (RMSD ≤ 2.0 Å), ConfGen Very fast performs remarkably efficiently. With ligand discovery being increasingly informed by high-quality structural input, the emphasis is on more precise modeling (RMSD ≤ 1.0 Å). In that context, ConfGen Fast is a sensible default option, particularly in a high-throughput regime for large compound libraries. Therefore, we use ConfGen Standard Fast as the reference for subsequent comparisons.

3.2. ConfGen Advanced. ConfGen Advanced allows the use of customizable settings, presented in the ConfGen Advanced GUI. This section presents results obtained via the default settings of this GUI ("default ConfGen Advanced"), or changing some of the parameters via the GUI (as opposed to editing the command-line input files, see section 3.3). Table 5 lists the default parameter values for ConfGen Advanced, which include the *Rapid* search mode.

The *Rapid* and *Thorough* modes influence the treatment of the Terminal Rotatable Bonds (TRBs, see Methods). With *Rapid*, those groups are sampled one after another, while with *Thorough* they are sampled simultaneously. ConfGen Advanced permits both the *Rapid* and *Thorough* search modes, while ConfGen Standard Fast uses the *Rapid* mode (seen in the corresponding command file).

Thus, ConfGen Advanced calculation with *Rapid* was first compared to its *Thorough* counterpart, all other parameters at default. The *Thorough* search mode improved *%BioConf_Rep* by 4% (RMSD ≤ 1.0 Å) over the default *Rapid* (Table 4). *Thorough* increased the average computation time by ∼32% over *Rapid* but surprisingly had virtually no impact on NbConfs. The default ConfGen Advanced *Rapid* performed similarly to ConfGen Standard Fast in all respects.

Applying a GB/SA solvation model was also tested via the ConfGen Advanced GUI. With *Rapid* or *Thorough*, this solvation model had a slightly detrimental effect on *%BioConf_Rep* (1−2% decrease at RMSD ≤ 1 Å), while the computational times increased by >40%. Intriguingly, the solvation model reduced NbConfs relative to the equivalent protocol without solvation (Table 4). Thus, there is no argument in favor of the GB/SA solvation model with ConfGen, based on *%BioConf*_Rep values. This contrasts with results obtained previously with the MOE stochastic search[18] and shows that performing the tests explicitly with

Table 4. Performance[a] of ConfGen in Reproducing the Ligand Experimental Bioactive Conformations[b] and Comparison to Other Methods[c,d]

	investigated parameters					*%BioConf_Rep*[e] RMSD (Å) versus bioactive					
	RotSteps	duplicate RMS (Å)	ΔE (kcal/mol)	Maxconfs	LigPrep	0.5	1	1.5	2	NbConfs[f]	seconds per compound
ConfGen Standard Predefined Protocols											
Very fast	5	1.25	none	1000	Ionizer/OPLS	19%	62%	92%	100%	60	2.6
Fast	**5**	**1**	**25**	**1000**	**Ionizer/OPLS**	**24%**	**66%**	**91%**	**98%**	**50**	**5.4**
Intermediate	75	1	25	1000	Ionizer/OPLS	29%	74%	93%	99%	154	19.1
Comprehensive	75	0.5	120	1000	Ionizer/OPLS	40%	77%	94%	100%	528	39.7
ConfGen Advanced GUI											
Default (*Rapid*[g])	**5**	**1**	**24**	**1000**	**Ionizer/OPLS**	**22%**	**65%**	**90%**	**98%**	**51**	**5.6**
Thorough[g]	5	1	24	1000	Ionizer/OPLS	23%	69%	91%	99%	50	7.4
Rapid/GB[h]	5	1	24	1000	Ionizer/OPLS	21%	64%	88%	98%	46	8.8
Thorough/GB[h]	5	1	24	1000	Ionizer/OPLS	24%	67%	88%	98%	45	10.7
Individual ConfGen Parameters Varied, Based on ConfGen Fast											
ConfGen Fast with RotSteps Varied											
	10	1	25	1000	Ionizer/OPLS	26%	69%	92%	100%	77	6.6
	15	1	25	1000	Ionizer/OPLS	26%	70%	92%	100%	95	7.9
	20	1	25	1000	Ionizer/OPLS	26%	70%	93%	100%	109	9.1
	25	1	25	1000	Ionizer/OPLS	26%	72%	94%	100%	121	10.4
	100	1	25	1000	Ionizer/OPLS	27%	72%	94%	100%	178	20.4
ConfGen Fast with Duplicate RMS Varied											
	5	0.15	25	1000	Ionizer/OPLS	31%	70%	91%	99%	79	5.6
	5	0.25	25	1000	Ionizer/OPLS	31%	70%	91%	99%	77	5.6
	5	0.50	25	1000	Ionizer/OPLS	30%	70%	91%	99%	68	5.5
ConfGen Fast with ΔE Varied											
	5	1	5	1000	Ionizer/OPLS	20%	58%	82%	94%	19	5.2
	5	1	10	1000	Ionizer/OPLS	23%	63%	85%	96%	31	5.3
	5	1	15	1000	Ionizer/OPLS	24%	66%	88%	98%	39	5.3
	5	1	30	1000	Ionizer/OPLS	23%	66%	91%	99%	54	5.4
	5	1	40	1000	Ionizer/OPLS	23%	67%	90%	99%	59	5.5
	5	1	50	1000	Ionizer/OPLS	23%	69%	91%	100%	63	5.7
	5	1	125	1000	Ionizer/OPLS	23%	69%	92%	100%	71	6.0
ConfGen Fast with MaxConfs Varied											
	5	1	25	250	Ionizer/OPLS	24%	66%	91%	99%	50	5.4
RotSteps Varied with RMS = 0.5 Å and ΔE = 25											
	10	0.5	25	1000	Ionizer/OPLS	33%	72%	92%	100%	114	7.1
	15	0.5	25	1000	Ionizer/OPLS	33%	73%	93%	100%	151	8.7
	20	0.5	25	1000	Ionizer/OPLS	34%	74%	94%	100%	181	10.1
	25	0.5	25	1000	Ionizer/OPLS	34%	75%	94%	100%	208	11.6
RotSteps Varied with RMS = 0.5 Å and ΔE = 15											
	10	0.5	15	1000	Ionizer/OPLS	32%	70%	92%	99%	87	6.8
	15	0.5	15	1000	Ionizer/OPLS	32%	72%	93%	98%	115	8.2
	20	0.5	15	1000	Ionizer/OPLS	33%	73%	93%	99%	136	9.6
	25	0.5	15	1000	Ionizer/OPLS	34%	75%	94%	100%	159	11.1
ConfGen Fast Default with Different LigPrep Protocols											
	5	**1**	**25**	**1000**	**Ionizer/OPLS**	**24%**	**66%**	**91%**	**98%**	**50**	**5.4**
	5	1	25	1000	Ionizer/MMFF	24%	72%	90%	100%	47	5.4
	5	1	25	1000	Epik/OPLS	22%	67%	89%	100%	60	6.0
	5	1	25	1000	Epik/MMFF	23%	72%	92%	100%	55	6.0
ConfGen Parameters Varied, with Ionizer/MMFF for LigPrep											
RotSteps Varied with RMS = 0.5 Å and ΔE = 25, with LigPrep Ionizer/MMFF											
	10	0.5	25	1000	Ionizer/MMFF	32%	77%	92%	100%	108	7.1
	15	0.5	25	1000	Ionizer/MMFF	34%	79%	92%	100%	142	8.6
	20	0.5	25	1000	Ionizer/MMFF	34%	80%	93%	100%	170	9.9
	25	0.5	25	1000	Ionizer/MMFF	34%	81%	93%	100%	194	11.2
	10	0.5	15	1000	Ionizer/MMFF	32%	76%	92%	99%	81	6.7
	15	0.5	15	1000	Ionizer/MMFF	33%	78%	91%	99%	105	8.2
	20	0.5	15	1000	Ionizer/MMFF	33%	79%	92%	99%	125	9.4
ConfGen Optimized	25	0.5	15	1000	Ionizer/MMFF	33%	80%	92%	99%	141	10.8
	25	0.5	15	1000	Epik/MMFF	33%	79%	94%	100%	170	12.6
MOE and Catalyst											
MOE Conformation Import	**NA**	**0.15**	**4**	**250**		**32%**	**75%**	**94%**	**99%**	**111**	**10.6/1.3[i]**
Catalyst FAST	**NA**	**NA**	**20**	**255**		**24%**	**70%**	**94%**	**100%**	**105**	**3.6**
Catalyst BEST	**NA**	**NA**	**20**	**255**		**28%**	**81%**	**98%**	**100%**	**110**	**123.9**

[a] Measured as the percentage of reproduced bioactive conformers within a given RMSD threshold and the speed (seconds per compound). [b] Based on the 253 compounds of the generic compound set. [c] Other methods (default settings) used for comparison were obtained from a previous study[18] and are MOE Conformation Import, Catalyst FAST, and Catalyst BEST. [d] Default protocols are shown in bold. [e] *%BioConf_Rep* is the percent of bioactive structures reproduced by a computational protocol, within a defined RMSD threshold (0.5, 1.0, 1.5 or 2.0 Å). [f] NbConfs is the average number of conformers per compound output by a computational protocol. [g] *Rapid* or *Thorough* refers to the treatment of terminal rotatable bonds (TRBs, section 2.4). [h] Generalized Born solvation model. [i] Timing obtained with a pregenerated fragment library.

every piece of software is important. A more rigorous exploration of the ConfGen parameters is possible by editing the input files directly, as presented in the next section.

3.3. Influence of ConfGen Parameters on Reproduction of the Bioactive Structures. We now turn to a detailed analysis of the impact of the ConfGen parameters on the

Table 5. Default and Tested Parameters[a] for ConfGen[b]

	ConfGen Advanced		Customized ConfGen	
parameters[a]	default[d]	tested values[e]	tested values[e]	ConfGen Standard Fast[c]
MaxConfs[f]	1000	1000	1000, 250	1000
RotSteps[g]	5	5	5, 25, 50, 75, 100	5
TRB[h] treatment	*Rapid*	*Thorough*	*Rapid*	*Rapid*
ΔE^i (kJ/mol)	100	100	20.9, 41.9, 62.8, 104.7, 209.3	104.7
ΔE^i (kcal/mol)	23.9	23.9	5, 10, 15, 25, 50	25
duplicate RMS[j] (Å)	1	1	0.15, 0.25, 0.50, 1.00	1
minimization steps of input structures	100	100	none	none
minimization steps of generated structures	none	none	none	none
solvent	none	water with GB[k]	none	none

[a] For parameters not listed, the default values were used throughout. [b] For tests performed via the GUI of ConfGen Advanced, and ConfGen protocols customized at the command-line. [c] To facilitate comparison with the default ConfGen Standard protocol, the parameters used with ConfGen Fast are reiterated here. [d] Default settings in the ConfGen Advanced GUI. [e] Values changed via the ConfGen Advanced GUI; their exact combinations are reported in Table 4 and in sections 3.2 and 3.3. [f] Maximum number of allowed output conformers. [g] Number of sampling steps per degree of freedom. [h] Terminal Rotatable Bond (see section 2.4). [i] The allowed energy window is reported in kJ/mol in the ConfGen GUI; for convenience and comparison to other published numbers, we give the correspondence in kcal/mol. [j] RMSD cutoff for removal of duplicate conformers. [k] Generalized Born solvation model.[51]

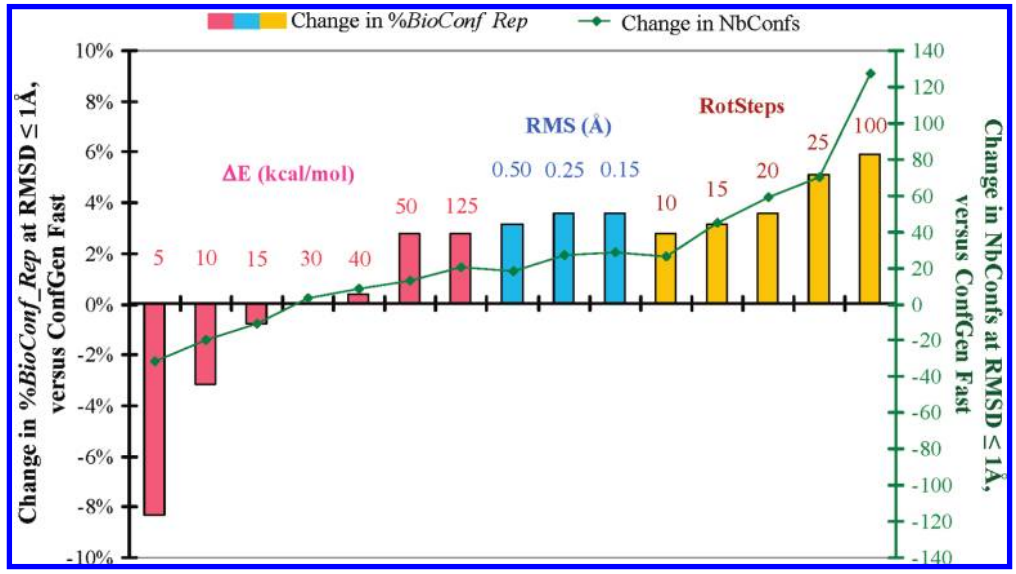


Figure 3. Impact of individual ConfGen parameters (ΔE, duplicate RMS and RotSteps) on the percentage of bioactive conformations reproduced (*%BioConf_Rep*) relative to ConfGen Fast (shown as change in *%BioConf_Ref* on the left *y* axis). The change in NbConfs compared to ConfGen Fast is shown on the right *y* axis. The actual values tried for each parameter are shown on top of each bar. The absolute *%BioConf_Rep* and NbConfs values for ConfGen Fast are 66% and 50, respectively.

reproduction of the bioactive structures. As explained above, it is meaningful to use the default ConfGen Standard Fast as reference for comparisons. Therefore, the parameters were explored by directly editing the ConfGen Standard Fast command file, keeping all its parameters except those varied (see section 2.5). The parameters varied, and their tested values are summarized in Table 5.

In Table 4, all customized ConfGen protocols derived from Fast are under the subheading "Individual ConfGen Parameters Varied". The parameters investigated in this section are the number of sampling steps per degree of freedom (RotSteps), the allowed energy window ΔE, the RMS cutoff for removal of duplicate conformers (duplicate RMS), and the maximum allowed number of conformers (MaxConfs). These parameters were first investigated independently of each other. Table 4 summarizes the influence of these parameters on NbConfs, *%BioConf_Rep*, and the computation times.

Increasing RotSteps yielded a significant increase in *%BioConf_Rep* (72% at RotSteps = 25), although this effect saturated for RotSteps values above 25 (Table 4). Increasing RotSteps above 25 keeps increasing NbConfs and the computing time, but without improvements on *%BioConf_Rep*. RotSteps = 25 improved *%BioConf_Rep* by 6% (RMSD ≤ 1.0 Å) relative to ConfGen Standard Fast (RotSteps = 5; Figure 3). This 5-fold increase in RotSteps more than doubled NbConfs from 50 to 121 and doubled the computing time.

Reducing the duplicate RMS also improved *%BioConf_Rep*, as observed before with the program OMEGA.[42] Reducing the duplicate RMS from a default of 1.0 Å to 0.5 Å increased *%BioConf_Rep* from 66% to 70% (RMSD ≤ 1.0 Å, Table 4, Figure 3), with very little impact on the computing time. The improved *%BioConf_Rep* was again associated with an increased average NbConfs, from 50 to 68. Smaller duplicate RMS values, e.g. 0.15 Å as used in MOE Conformation Import, did not yield further improvements in *%BioConf_Rep*.

Reducing the ConfGen Fast energy window from 25 kcal/mol (default) to 15 kcal/mol retained the same performance in reproducing the bioactive structures, while reducing

Table 6. Impact of LigPrep Protocols on Preparation of Compound 3D Structures[a]

ionization and tautomerization protocol	Ionizer		Epik	
force field	OPLS 2005	MMFFs	OPLS 2005	MMFFs
total average number of structural variants per compound[b,c]	3.5	3.5	3.7	3.7
average number of tautomeric and ionization states only[c]	1.7	1.7	1.8	1.8
average NbConfs from ConfGen Fast	50	47	60	55

[a] For the generic set of compounds. [b] The total number of structural variants covers ring conformations, stereoisomers, tautomers, and ionic states. [c] Ionic states were calculated using the default pH range from 5 to 9.

NbConfs by 11. Reducing ΔE below 15 kcal/mol degraded *%BioConf_Rep* (Table 4). Extending ΔE to values as high as 125 kcal/mol increased *%BioConf_Rep* only marginally, while increasing NbConfs by 42% compared to the default (NbConfs = 50). Thus, increasing ΔE to "large values" may introduce noise in terms of generating conformers outside the energetically accessible bioactive space, although that is a matter for debate.[10,18,26,56,57] Varying ΔE did not impact significantly the computing time, consistent with the energy filtering being applied as a post sampling step.

Decreasing the maximum allowed number of conformations per ligand MaxConfs from 1000 to 250 had no influence on *%BioConf_Rep*, NbConfs, or the computing time, compared to ConfGen Standard Fast. This is consistent with the average NbConfs = 50 for ConfGen Fast being lower than 250. It reflects the drug-like sizes of the compounds in the generic set, most of them having six or less rotatable bonds. For much larger/flexible compounds, MaxConfs may have an impact.

The above showed that, when changed independently, RotSteps and the duplicate RMS improved *%BioConf_Rep*, while a lower ΔE had a beneficial impact on NbConfs. Therefore, we tested protocols where these parameter values were changed in combination (Table 4). In these combinations, RotSteps was varied systematically again from 10 to 25, with a duplicate RMS decreased to a constant value of 0.5 Å, with ΔE = 25 or 15 kcal/mol. A (quasi) additivity of the effect of combining RotSteps and the duplicate RMS was observed. Indeed, for both ΔE = 25 and 15 kcal/mol, setting RotSteps = 25 and the duplicate RMS at 0.5 Å yielded *%BioConf_Rep* = 75% at RMSD ≤ 1.0 Å, better than when changing any parameter alone. Importantly, the protocol with RotSteps = 25, duplicate RMS = 0.5 Å, and ΔE = 15 kcal/mol performed almost as well as ConfGen Standard Comprehensive regarding *%BioConf_Rep* (75 versus 77%) but required 3.6 times less computing time and 3.3 times fewer conformers. Such conformers are expected to have lower energies than those from Comprehensive because of the vastly different energy windows (15 versus 120 kcal/mol). Below, we show how one can improve further, to provide a protocol that we call "ConfGen Optimized".

3.4. Impact of Compound Preparation, and ConfGen Optimized. Before running ConfGen, the compounds were prepared with LigPrep, the Schrodinger tool to convert compounds from 2D to 3D geometry. For a compound, LigPrep can generate various states in terms of stereochemistry, ring conformations, tautomerization, and ionization, in a given pH range.[33,43] The tautomeric and ionization states of a compound may be assigned with one of two options, Ionizer or Epik. Each of these compound states is energy minimized at the end of the LigPrep process, with OPLS_2005 (default) or MMFF94s. Since ConfGen does not vary the bonded geometries (bond length and valence angles), nor the tautomers/ionization states, the energy minima identified by ConfGen are sensitive to the input structures obtained from LigPrep. Therefore, we investigated the impact of the ligand preparation on the performance of subsequent ConfGen calculations. In addition to the default LigPrep combination (Ionizer/OPLS), the three other combinations (Epik/OPLS, Ionizer/MMFF, and Epik/MMFF) were tested. All four sets of input ligand structures were submitted to ConfGen Standard Fast, to investigate their impact on *%BioConf_Rep* (Table 4).

The average numbers of structural variants generated by LigPrep are in Table 6, where NbConfs covers all the states for a given compound. Each compound is represented by 3.5 to 3.7 structural variants when generated by Ionizer or Epik, respectively (Table 6). Epik increased the average NbConfs (output from ConfGen Fast) by 10 (OPLS) or 9 (MMFF) relative to Ionizer. Among those, only tautomeric and ionic variants can be attributed to differences between Ionizer and Epik. There are on average 1.7 (Ionizer) and 1.8 (Epik) tautomeric and ionic variants per compound, roughly half the total number of variants. Thus, the remaining half of the structural variants generated by LigPrep came from the combination of ring conformations and stereochemistries. So, on average, Ionizer and Epik produced similar numbers of structural variants. We note, however, that there are compounds for which these variants can be quite different from one protocol to the other (not shown).

Table 4 shows that the percent of reproduced bioactive structures (RMSD ≤ 1.0 Å) can be improved depending on the selected LigPrep options for compound preparation. We again use the default options (Fast protocol with Ionizer/OPLS, *%BioConf_Rep* = 66%, Table 4) as reference for comparisons. On the basis of this, switching the force field to MMFF had a significant beneficial impact, increasing *%BioConf_Rep* to 72%. This increase in *%BioConf_Rep* was as large as that obtained when increasing RotSteps to 25 (section 3.3). The possible influence of the initial geometries on *%BioConf_Rep* had already been noted with OMEGA.[42] Switching from Ionizer to Epik had virtually no effect on *%BioConf_Rep*, consistent with Epik producing on average a similar number of structural variants (Table 6). Yet, at a constant force field, the input structures prepared with Epik led to slightly increased computation times and NbConfs, with no apparent benefit regarding *%BioConf_Rep*. So, in the present context, no argument emerged to replace the default Ionizer by Epik.

The next step was to combine the optimal options from LigPrep (Ionizer/MMFF) and from ConfGen (RotSteps = 25, duplicate RMS = 0.5 Å, ΔE = 15 kcal/mol, see section 3.3). This protocol retrieved 14% more bioactive conformations (*%BioConf_Rep* = 80% at RMSD ≤ 1.0 Å) than the

default LigPrep/ConfGen Fast (Table 4). Thus, the improvements brought independently by LigPrep and ConfGen were also additive. It produced the best ConfGen protocol identified in the present study for reproduction of the bioactive structures and is referred to as "ConfGen Optimized". Ionizer was selected for ConfGen Optimized, because substituting it with Epik increased NbConfs without improving *%BioConf_Rep* at RMSD ≤ 1.0 Å. ConfGen Optimized remains competitive if one takes into account the computational tradeoffs involved. ConfGen Optimized generates only 2.8 as many conformers as the default ConfGen Fast on average and requires only twice as much computation time (Table 4). This performance is indeed much optimized if compared to its counterpart with ConfGen Intermediate and ConfGen Comprehensive, in terms of balancing *%BioConf_Rep* with the computation time and size of the conformational ensembles. ConfGen Optimized was 3.7 times faster than ConfGen Comprehensive, generated 3.7 times fewer conformers, and added 3% of reproduced bioactive structures within RMSD ≤ 1.0 Å. ConfGen Optimized is compared to other programs in section 3.5. Also, section 3.6 shows that the conformational coverage afforded by ConfGen Optimized is competitive.

3.5. Comparison of ConfGen, MOE, and Catalyst for Reproduction of the Bioactive Structures. This section compares ConfGen to two other established high-throughput conformer generators at default settings, MOE Conformation Import and Catalyst. The default settings are representative for MOE and Catalyst, as they performed well in a number of tests,[18,24,25] including reproduction of bioactive structures of the generic set of compounds. This test is clearly relevant to medicinal chemistry, so it is used again for comparison here. A detailed study of MOE versus Catalyst has already been presented;[18] therefore, we focus on their comparison to ConfGen. The compared programs differ with respect to their search algorithms and force fields; however, their conformational sampling is subjected to the same types of parameters (Table 3). The interplay of these parameters controls the resulting *%BioConf_Rep* and NbConfs, also reported in Table 4 for MOE and Catalyst.

The default ConfGen Fast reproduced 66% of the bioactive conformations within RMSD ≤ 1 Å, compared to 75% by MOE, 70% by Catalyst FAST, and 81% by Catalyst BEST. ConfGen Fast was as fast as MOE Conformation Import or Catalyst FAST but yielded fewer reproduced bioactive conformers within RMSD ≤ 1 Å. On the other hand, ConfGen Comprehensive yielded *%BioConf_Rep* = 77% (RMSD ≤ 1.0 Å), slightly higher than MOE Conformation Import or Catalyst FAST. However, Comprehensive is not the ConfGen default and is much slower than MOE Conformation Import or Catalyst FAST (Table 4). Section 3.4 showed that improvement was achieved with ConfGen Optimized (*%BioConf_Rep* = 80% at RMSD ≤ 1.0 Å), performing virtually as well as Catalyst BEST, but at a tenth of the computational time. This is a notable result, as Catalyst BEST has outperformed many of its rivals regarding *%BioConf_Rep*,[10,18,24,25] but Catalyst BEST is too slow for high-throughput work.

Within RMSD ≤ 2 Å, the four predefined ConfGen Standard protocols yielded *%BioConf_Rep* values between 98% and 100%, as good as MOE Conformation Import (99%) and Catalyst FAST (100%) or BEST (100%).

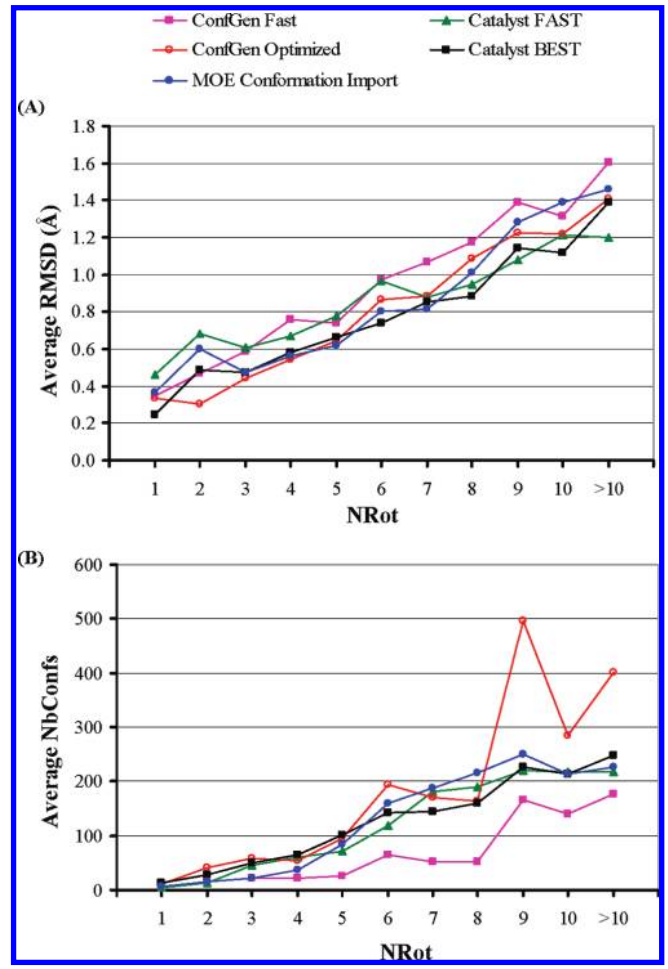


Figure 4. Influence of the number of rotatable bonds (NRot) per compound on (A) the average RMSD (Å) between X-ray bioactive structures and the best matching computed conformer and (B) the average number of conformers generated per compound (NbConfs). Except for ConfGen Optimized defined in section 3.4 (red open circles), all results were generated using the default settings of ConfGen Fast (magenta squares), MOE Conformation Import (blue filled circles), Catalyst FAST (green triangles), and Catalyst BEST (black squares).

As expected, reproduction of the bioactive structure becomes less successful as the number of rotatable bonds per compound (NRot) increases, as monitored by the average lowest RMSD between the X-ray and computed conformers for increasing values of NRot (Figure 4A). Typically, with MOE and Catalyst, the bioactive structures were reproduced within 1 Å for NRot up to 8. With ConfGen Fast, the reproduction of the bioactive structures within 1 Å was only maintained up to NRot = 6 on average and degraded somewhat compared to MOE or Catalyst for NRot > 6. From this perspective, ConfGen Optimized is again an improvement over ConfGen Fast and follows more closely MOE and Catalyst.

The overall increase of the average best distance between the X-ray and computed conformers as a function of NRot, observed across methods (Figure 4A), suggests that conformational sampling is still a factor limiting the reproduction of some drug-like bioactive structures. Thus, the development of existing and new sampling methods remains important.[13,21,27,31]

The smaller NbConfs values associated with ConfGen Fast relative to MOE Conformation Import and Catalyst (Table 4) affect essentially the entire NRot range (Figure 4B). This

echoes the lesser performance of ConfGen Fast regarding *%BioConf_Rep* at RMSD ≤ 1.0 Å. ConfGen Fast yielded NbConfs = 50, significantly less than the counterpart with Conformation Import (NbConfs = 111), Catalyst FAST (105), or ConfGen Optimized (141). The smaller NbConfs value between ConfGen Fast and the other methods is striking for NRot from 5 to 8 (Figure 4B). On the other hand, ConfGen Optimized yielded NbConfs values in step with MOE and Catalyst.

Together, the various protocols tested here confirm that moderate NbConfs values are sufficient to reproduce a majority of drug-like bioactive structure, as suggested before.[20,54] For instance, the protocols with 100 ≤ NbConfs ≤ 150 are very likely to reproduce the bioactive structures within 1 Å when NRot ≤ 8. Generating only tens of conformers, as done with ConfGen Fast, can only afford reproduction of the bioactive structures at a lower resolution, as the compound flexibility increases (Figure 4A). Although NbConfs is a useful indicator, it only captures one facet of conformational coverage. This led us to examine the conformational coverage in more detail in the next section.

3.6. Conformational Coverage Assessed by 3D Descriptors. Conformational coverage represents the extent and resolution with which a conformational model covers the conformational space.[10,20,54] In the previous sections, NbConfs and *%BioConf_Rep* already gave indications of the conformational coverage. In addition, a number of 3D descriptors can characterize further the coverage of conformational space by a conformational model.[13,18,19,26]

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. Such analysis is particularly relevant for the ConfGen program, as the manual mentions that the compact conformers tend to be eliminated from the ensemble. Thus, the radius of gyration (Rgyr) was used as a basis to calculate the maximum percentage of extension (ConfExt) of every compound in a given conformational model. 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% (significant unfolding relative to a very compact conformation). A greater value of ConfExt corresponds to a greater spread in terms of conformational compactness and overall shape. So, larger values of ConfExt also suggest a greater coverage of the conformational space. Another, and finer, measure of coverage is the number of different 3D pharmacophores visited by a conformational model. This measure has also the advantage of being relevant to molecular recognition.

Figure 5 shows the minimal and maximal Rgyr, as well as the average values for ConfExt, as a function of NRot, for representative protocols run at default settings (except ConfGen Optimized, see section 3.4). As expected, ConfExt increases with the compound's intrinsic flexibility NRot (Figure 5A). Yet, the range of extension varies across sampling protocols, further supporting the relevance of the ConfExt descriptor. Catalyst FAST and BEST promote the greatest range of compound extension/compactness, especially as NRot increases. This may reflect the influence of the poling restraints implemented in Catalyst to promote conformational diversity.[58] The overall average ConfExt values for Catalyst FAST and BEST are 23% and 34%,

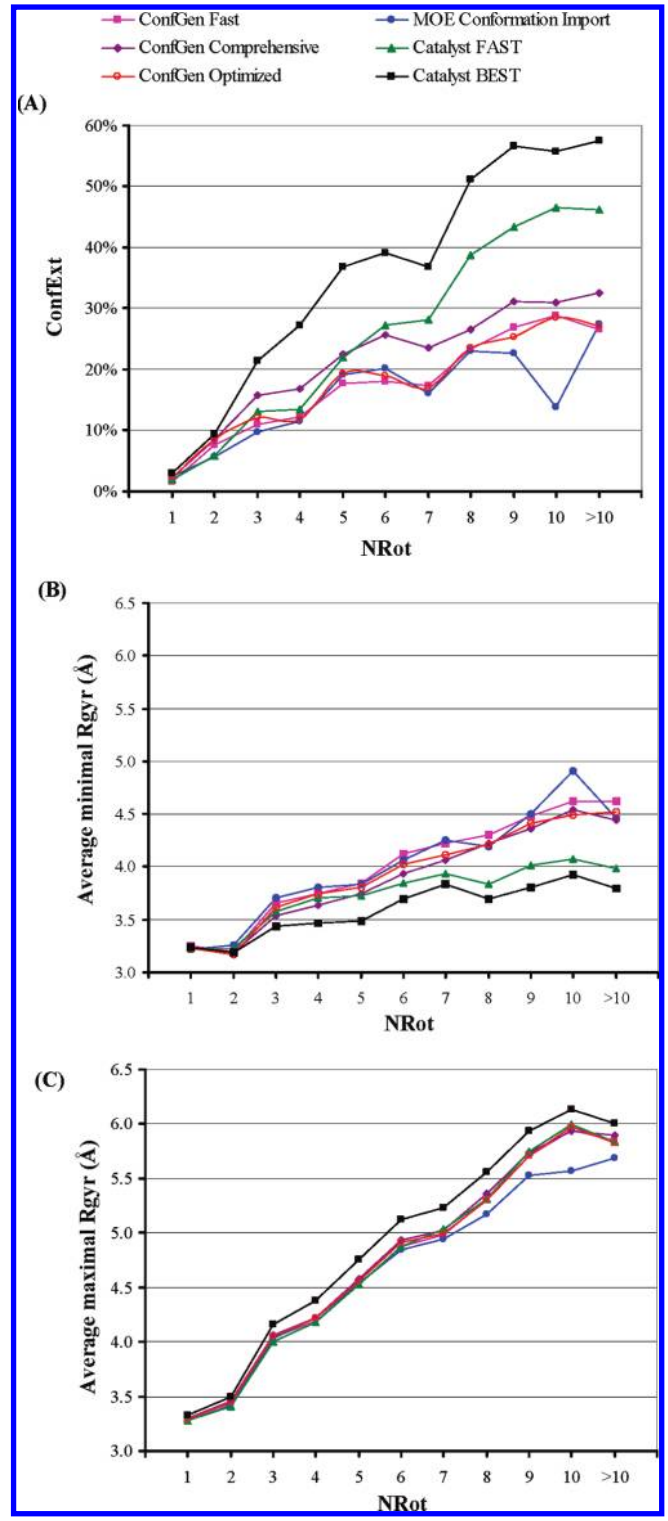


Figure 5. Use of descriptors of the compactness of computed conformers as a function of their intrinsic flexibility (number of rotatable bonds, NRot), for selected protocols at default settings, except for ConfGen Optimized. (A) The percent of conformational extension (ConfExt) of compounds based on the radius of gyration (Rgyr, see section 2.8). The average minimal (B) and maximal (C) Rgyr values for compounds as a function of NRot. The legends for ConfGen Fast (magenta squares), ConfGen Optimized (red open circles), MOE Conformation Import (blue filled circles), Catalyst FAST (green triangles), and Catalyst BEST (black squares) are consistent throughout the panels.

respectively, compared to 15−21% of other methods (see below). Both Catalyst BEST and ConfGen Optimized reproduced ~80% of the bioactive structures (Table 4), yet

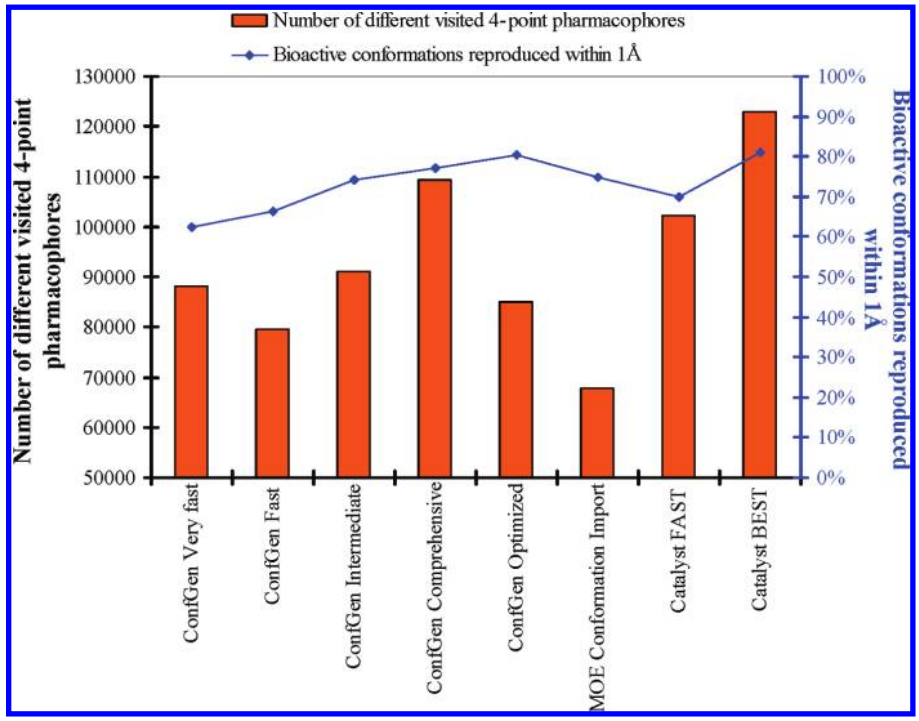


Figure 6. Estimate of conformational diversity based on the total number of different visited four-point pharmacophores (left *y* axis, orange bars) from ConfGen, MOE, and Catalyst alongside the corresponding *%BioConf_Ref* at RMSD ≤ 1 Å (right *y* axis, blue diamonds). These data are for the generic compound set.

Catalyst BEST conformers have ConfExt values twice those of ConfGen Optimized. It suggests that a significant fraction of the Catalyst BEST conformers may not be directly relevant for reproduction of the bioactive structures.

As the ConfExt value only gives an idea of the relative range of extension/compactness of a compound, we also examined the underlying minimal (Figure 5B) and maximal (Figure 5C) Rgyr values as a function of NRot, averaged across compounds for each NRot value. Figure 5C shows that most protocols produce similar maximal conformation extensions, except for Catalyst BEST, which consistently produces the largest Rgyr values. On the other hand, the minimal Rgyr values show more variations across methods (Figure 5B). Catalyst FAST and BEST generated the most compact conformers. The two methods associated with the least compact conformers are MOE and ConfGen Fast. The underlying minimal and maximal Rgyr plots are consistent with the resulting ConfExt values. They also suggest that the consistently larger ConfExt effect of Catalyst BEST is largely contributed by its greater ability to generate particularly compact conformers.

Another view of the conformational coverage was obtained by counting the number of visited four-point pharmacophores across all compounds of the generic set for selected computational protocols (Figure 6). This count of visited pharmacophores provides a fine-grained estimate of the generated conformational diversity and is plotted alongside the corresponding *%BioConf_Rep* values (RMSD ≤ 1 Å, Table 4). Figure 6 shows that this count and the associated conformational diversity vary dramatically across protocols. This allows one to investigate if a greater conformational diversity helps for the reproduction of the bioactive structures. In principle, such discussion should be put in the context of the physical relevance of the computed conformers, expected to be related to their allowed energy window ΔE. Of course, the energy windows applied across different programs are not directly comparable because they refer to different energy models. In addition, the conformational diversity will depend on other factors such as algorithms biasing the sampling (e.g. poling in Catalyst[58]), the number of search steps, and the RMS applied for duplicate removal. Nevertheless, considering its potential importance as a physical parameter, it is interesting to reflect on the role of ΔE in influencing the resulting conformational diversity.

The energy window varies across the ConfGen Standard protocols (Table 2), and it was 15 kcal/mol for ConfGen Optimized, 4 kcal/mol for MOE Conformation Import, and 20 kcal/mol for Catalyst FAST or BEST. Overall, the smaller energy windows associated with Conformation Import probably contributes to the associated smaller number of pharmacophores. Nevertheless, comparing the various ConfGen protocols shows that increasing ΔE to very large values (e.g., 120 kcal/mol with ConfGen Comprehensive versus 15 kcal/mol with ConfGen Optimized) does not increase the conformational diversity in the same proportion. In fact, the largest number of pharmacophores was produced with Catalyst BEST ($\Delta E = 20$ kcal/mol), probably reflecting the effect of poling.[58] Thus, the conformational coverage estimated by the number of visited pharmacophores is not dominated by a single parameter.

A striking result from Figure 6 is that, for drug-like ligands, there is no simple correlation between the generated conformational diversity (orange bars) and *%BioConf_Rep* (blue diamonds). This can be illustrated by several comparisons. First, MOE Conformation Import reproduced more or almost as many bioactive structures as the other protocols, despite a markedly smaller number of pharmacophores visited by MOE Conformation Import. Indeed, MOE Conformation Import and ConfGen Comprehensive visited 67 753 and 109 292 pharmacophores, respectively, yet both

Table 7. Reproduction of Folded Bioactive Structures[a] by Selected Computational Protocols

	folded compound set				generic compound set[e]	
MOE	%BioConf_Rep (RMSD ≤ 1 Å)	NbConfs			%BioConf_Rep (RMSD ≤ 1 Å)	NbConfs
conformation import	42%	172			75%[d]	110[d]
stochastic optimized	63%	1615			94%[d]	806[d]
	compactness disfavored[b]		compactness allowed[c]			
ConfGen	%BioConf_Rep (RMSD ≤ 1 Å)	NbConfs	%BioConf_Rep (RMSD ≤ 1 Å)	NbConfs		
Fast	14%	145	22%	145	66%	50
Comprehensive	33%	1490	37%	1581	77%	528
Optimized	35%	387	35%	379	80%	141

[a] Within RMSD ≤ 1 Å. [b] ConfGen algorithm at default disfavors compact conformations by activating options CGO6 and CHYD (section 2.4).[33] [c] For the folded set, some ConfGen protocols were run after turning off options CGO6 and CHYD, everything else kept the same. [d] Data from a previous study.[18] [e] Data for the generic set are listed for comparison to the folded set.

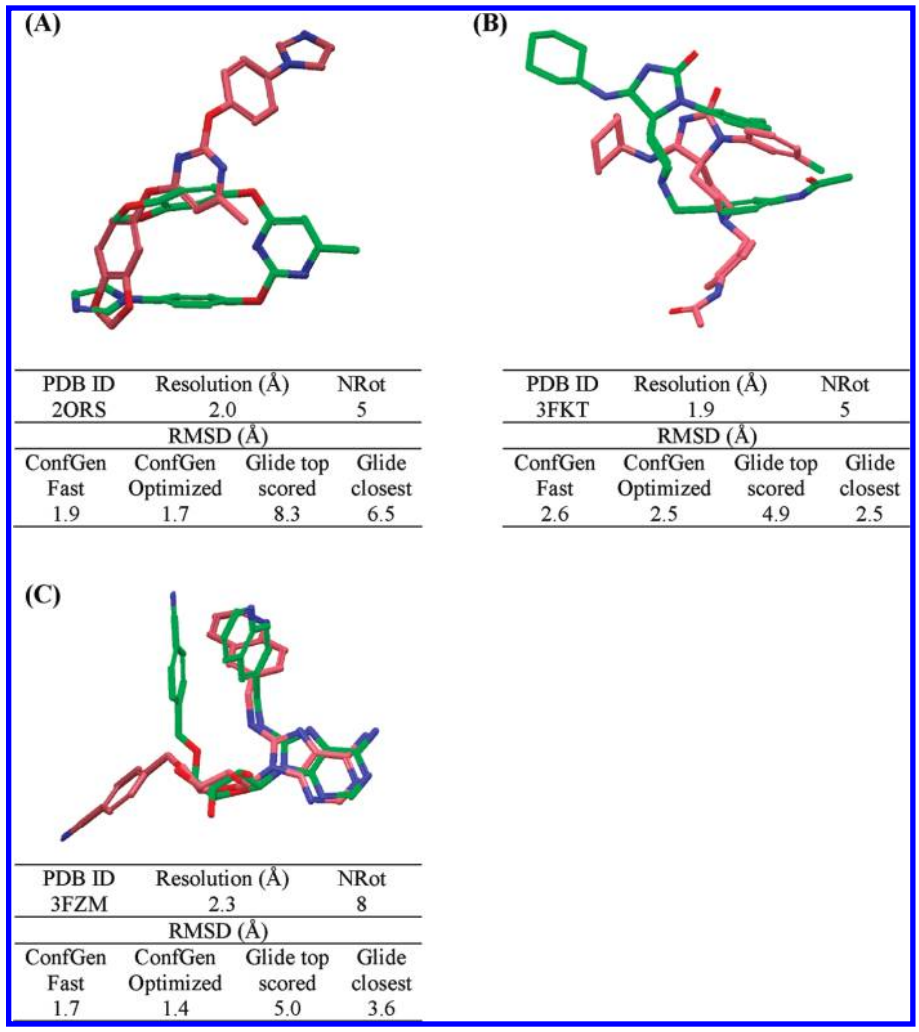
protocols yielded similar *%BioConf_Rep* values. A second example is the dip in the number of visited pharmacophores when going from ConfGen Very fast to Fast, while *%BioConf_Rep* is larger with Fast than with Very fast. Incidentally, the dip in the number of pharmacophores is consistent with the lack of energy filtering in Very fast, keeping more conformers. The weeding out of high energy conformers with ConfGen Fast ($\Delta E = 25$ kcal/mol) decreases the conformational diversity, yet *%BioConf_Rep* is increased versus Very fast. A third example is the comparison of ConfGen Optimized (*%BioConf_Rep* = 80%, number of visited pharmacophores = 85 155) to ConfGen Comprehensive (*%BioConf_Rep* = 77%, number of visited pharmacophores = 109 292). In sum, generating greater conformational diversity does not guarantee a greater number of reproduced bioactive structures. Thus, it is not enough to generate as large a conformational spread as possible. Although there is no consensus regarding the conformational energetics of bound ligands,[10,17,18,26,57] it is possible that many high energy conformations kept in the most diverse conformational ensembles tend not to be relevant to the populated bioactive space. Importantly, this analysis confirmed that ConfGen Optimized offers an attractive balance between conformational diversity and coverage of bioactive space.

3.7. Folded Bioactive Structures. It has been suggested that flexible ligands are more likely to bind to proteins in extended conformations.[26,44,45] However, there are reports that natural cofactors[46] or marketed drugs[22] can also adopt compact/folded bioactive conformations. This variety of situations is not surprising considering the diverse and heterogeneous nature of receptor binding sites. Thus, methods have been developed to drive conformational sampling over the whole range of conformational compactness.[19] Nevertheless, the common perception remains that bioactive conformations ought to favor extended shapes, as that intuitively implies more extensive interactions with the receptor. Such a perception may have contributed to the introduction in ConfGen of heuristic rules penalizing folded conformers[5] (see also section 2.4). Indeed, the ConfGen manual states that "*ConfGen also tends to eliminate compact ligand conformations*".

To investigate these questions, we have compiled a new set of 50 drug-like ligands (section 2.2) with bioactive structures which fit the pharmacophore in Figure 1. This pharmacophore is one way to detect folded conformers. Of course, this approach does not provide a statistically rigorous way to estimate the general frequency of folded bioactive conformers. However, it is sufficient to make the point that even relatively small, drug-like compounds (Table 1) can adopt folded bioactive states. We call these 50 ligands the "folded set" of compounds. This set confirms the relevance of folded bioactive conformations and provides a pertinent test for ConfGen.

Table 1 shows that compounds in the folded set are on average larger (MW) and more flexible (NRot) than compounds in the Vernalis generic set. This complicates a comparison of the performance of conformational sampling applied to the folded versus generic sets. Indeed, the average distance between the X-ray and closest computed conformers increases with NRot (Figure 4A). Thus, it is not surprising that the *%BioConf_Rep* values for the folded set are lower than their counterparts with the generic set, for all tested protocols (Table 7).

However, with the folded set of compounds, the main aim is to compare how different methods perform for these particular conformations, monitored via *%BioConf_Rep* values (RMSD ≤ 1 Å; Table 7). Default ConfGen Fast reproduced 14% of the folded set (RMSD ≤ 1 Å), compared to 42% by default MOE Conformation Import. The equivalent difference based on the generic set of compounds is 9% (Table 7), showing that the gap between ConfGen Fast and Conformation Import is 19% larger with folded compounds. This marked increased disparity between ConfGen and Conformation Import partly reflects the deliberate penalties against folded conformations in ConfGen, although this could not be completely disentangled from differences in compound flexibility. Improved sampling with ConfGen Comprehensive or ConfGen Optimized essentially doubled *%BioConf_Rep* with the folded set (Table 7). However, for the folded set, the performance of ConfGen Comprehensive or Optimized still fell behind that of MOE Conformation Import, reproducing 7−9% less bioactive structures than Conformation Import (RMSD ≤ 1 Å). In contrast, for the generic set, ConfGen Comprehensive and Optimized performed as well as or better than Conformation Import (Table 7). Therefore, the performance of ConfGen for the folded set was comparatively poorer. This further points to disfavored compact conformations as a likely cause for ConfGen's lesser



PDB ID	Resolution (Å)	NRot
2ORS	2.0	5

RMSD (Å)			
ConfGen Fast	ConfGen Optimized	Glide top scored	Glide closest
1.9	1.7	8.3	6.5

PDB ID	Resolution (Å)	NRot
3FKT	1.9	5

RMSD (Å)			
ConfGen Fast	ConfGen Optimized	Glide top scored	Glide closest
2.6	2.5	4.9	2.5

PDB ID	Resolution (Å)	NRot
3FZM	2.3	8

RMSD (Å)			
ConfGen Fast	ConfGen Optimized	Glide top scored	Glide closest
1.7	1.4	5.0	3.6

Figure 7. Results of Glide docking on three selected ligands from the folded compound set. The ligands from PDB entries (A) 2ORS, (B) 3FKT, and (C) 3FZM were docked into their native protein binding site. Each panel shows the ligand X-ray structure (green carbons) and the corresponding best scored docking pose (orange carbons). Each panel also presents annotations regarding the X-ray resolution (Å) and the NRot value for the ligand. In addition, the RMSD (Å) is given, between the bioactive X-ray structure and computed conformers from (i) ConfGen Fast stand-alone (no protein), (ii) ConfGen Optimized stand-alone (no protein), (iii) the top scored pose after Gilde docking, and (iv) the docked pose with the closest RMSD to the X-ray structure.

performance for the folded set, because considerable additional sampling afforded by ConfGen Comprehensive or Optimized is not sufficient to match the performance of MOE Conformation Import.

There are two ConfGen command-line options, CGO6 and CHYD, which disfavor compact conformations (section 2.4). The options CGO6 and CHYD were turned off while running ConfGen Fast and ConfGen Optimized on the folded set ("compactness allowed" in Table 7). As anticipated, ConfGen Fast clearly benefited from deactivating CGO6 and CHYD, with 8% additional bioactive structures recovered. With ConfGen Comprehensive, the improvement was modest (4%) but still notable. This demonstrates that penalizing the folded conformers in ConfGen can degrade reproduction of the bioactive structures in some situations relevant to drug design.

Importantly, deactivating CGO6 and CHYD with the generic set of 253 compounds did not degrade the results. In fact, 67% of the bioactive structures were reproduced within 1 Å with ConfGen Fast and CGO6/CHYD turned off, compared to 66% when CGO6/CHYD are turned on. Therefore, no gain was associated with the use of CGO6/CHYD in the present experiments.

The best performing protocol on the folded compounds set is the so-called optimized stochastic search from MOE (with generalized Born and $\Delta E = 15$ kcal/mol) described previously,[18] which reproduced 63% of the folded bioactive structures (Table 7). This approach provides a very thorough search of the conformational space and performed particularly well with the generic set of compounds.[18] The present results further validate the relative competitive performance of the optimized MOE stochastic protocol with a qualitatively different set of compounds.

ConfGen is the conformational generator underpinning other Schrodinger tools (e.g., docking with Glide); therefore, its penalization of folded structures raises questions for systems which accommodate or favor such structures. Addressing these questions in detail is beyond the scope of the present work. However, to illustrate the pertinence of such questions, we tested Glide docking on three complexes from the folded set (PDB entries 2ORS, 3FKT, and 3FZM). These complexes were chosen because ConfGen Fast stand-alone did not reproduce these bioactive structures within 1 Å (Figure 7). In addition, their crystal structures are of good resolution and the ligands are not overly flexible (Figure 7), suggesting that the extent of conformational sampling is

unlikely to explain failure. These ligands were docked to their native protein X-ray structures with Glide Standard Precision. For each complex, Figure 7 shows the top scored pose alongside its X-ray counterpart. For these examples, none of the docking poses reproduced the X-ray folded bioactive structures within 2 Å. This is intriguing because Glide is otherwise a widely used and well-validated docking program.[5,50,59–61] Visual inspections showed that the docked poses were not folded. This suggests possible limitations of Glide with folded bioactive structures, related to the underlying ConfGen search heuristics. The present examples do not firmly demonstrate that it is the case; however, they provide a hypothesis for further work. Thus, for systems conducive to folded bioactive ligands, Glide docking may benefit by being more tolerant of compact ligand conformations.

4. CONCLUSIONS

ConfGen is a general purpose conformational sampling engine distributed by Schrodinger, available in a stand-alone mode, and underpinning other widely used modeling tools (e.g., docking with Glide). ConfGen is accessible via a user-friendly graphical user interface, and its input/output easily integrated with the rest of the Schrodinger software suite, such as its 2D to 3D converter LigPrep. Thus, it was of general interest to examine thoroughly the performance of ConfGen in the context of drug-like compounds. This relied on two sets of compounds with publicly available X-ray structures, a generic set of 253 diverse ligands, and a specialized set of 50 folded ligands. The generic compound set allowed for a direct comparison of ConfGen to MOE and Catalyst,[18] two other widely used softwares for conformational sampling. The compilation of the folded set of compounds is a new contribution, of interest considering that one commonly meets the perception that drug-like ligands bind to proteins in extended conformations. The deliberate penalization of compact/folded conformers by ConfGen documents this perception.

Our tests of conformational sampling addressed the ability to reproduce the bioactive structures (*%BioConf_Rep*), the size of the conformational ensembles (NbConfs), their diversity and coverage, and the corresponding relative compute times. We first tested the four predefined ConfGen Standard protocols (Very fast, Fast, Intermediate, and Comprehensive). This was followed by investigating individually the impact of important ConfGen and LigPrep parameters on the nature and relevance of the conformers. It led to propose a new ConfGen protocol that we call "ConfGen Optimized", which puts more weight on the quality of the result (percent of reproduced bioactive structures *%BioConf_Rep* and relevant conformational coverage) than the predefined commercial protocols, while remaining computationally tractable.

The four predefined ConfGen Standard protocols provide a range of performances (e.g., *%BioConf_Rep* and NbConfs) with different run times. This offers good flexibility to the user depending on the computational resources. For high-throughput conformation generation, the default ConfGen Fast is a sensible compromise between performance, computation resources, and storage space. It reproduced 66% of the generic bioactive structures (RMSD $\leq$ 1 Å) with a comparatively small number of conformers per compound (average NbConfs = 50). Smaller NbConfs values facilitate the management of 3D compound libraries; however, they also tend to yield a less frequent reproduction of the bioactive structures within 1 Å. The tested protocols produced a maximum average NbConfs = 528 (ConfGen Comprehensive), confirming that less than a thousand conformers are typically sufficient to describe drug-like compounds. Actually, the largest NbConfs values are no guarantee of getting the most relevant conformational coverage, and explicit tests are required to determine an appropriate balance.

Investigation of ConfGen parameters showed that the number of sampling steps per degree of freedom (RotSteps), the RMSD cutoff to discard duplicate conformers (duplicate RMS), and the energy window (ΔE) largely control the number of output conformers. After optimization, in particular with respect to *%BioConf_Rep* and NbConfs, these parameters take the values RotSteps = 25, duplicate RMS = 0.5 Å, and ΔE = 15 kcal/mol, which are proposed for the new ConfGen Optimized protocol. We also showed that the performance of ConfGen is sensitive to the force field used during the ligand preparation by LigPrep. An additional 6% of the bioactive structures were recovered with ConfGen Fast just by changing the default OPLS_2005 force field to MMFF94s in LigPrep. Thus, different representations of the input geometries influence the quality of the ConfGen output. The choice of Ionizer or Epik did not affect *%BioConf_Rep*. These findings about compound preparation also contributed to the performance of ConfGen Optimized. ConfGen Optimized reproduced 80% of the generic bioactive structures (RMSD $\leq$ 1 Å) within a competitive computation time. Among the computational protocols tested so far on the generic compound set,[18] only Catalyst BEST yielded such a high rate of reproduced bioactive structures, but ConfGen Optimized achieves this result 10 times faster. Of course, performance depends on the compound flexibility and topology, and the present conclusions apply to relatively small drug-like compounds. There are chemotypes such as macrocycles for which the ConfGen algorithm may not perform as well.[16]

For high-throughput 3D library generation, compared to MOE Conformation Import and Catalyst FAST, ConfGen Fast produces reasonably good conformers, with competitive computation times and smaller library sizes. However, MOE Conformation Import at default settings reproduced more drug-like bioactive structures within 1 Å. Yet, ConfGen Optimized emerges as a promising (albeit somewhat slower) alternative for medium- to high-throughput conformer generation, especially in the context of increasing computer speed.

The generated conformers were also assessed in terms of their diversity, estimated via 3D descriptors such as radius of gyration and four-point pharmacophores. The conformational diversity associated with the predefined ConfGen Standard protocols varied as anticipated; i.e., it increased with the amount of sampling and energy window (or lack of). Interestingly, a greater conformational diversity did not automatically correlate with greater reproduction of bioactive structures. Overpromoting conformational diversity may drive some conformers into high energy states which may not be accessible to bioactive structures, although this is an area of debate.[10,17,18,26,41,56] Thus, it is worth optimizing the parameter combinations with respect to objective criteria,

rather than increasing the parameter tolerances in a somewhat *ad hoc* fashion. ConfGen Optimized arguably yields a good balance of conformational diversity, reproduction of bioactive structures, and computational tractability.

An intriguing aspect of ConfGen is that it tends to penalize and remove compact/folded conformers. This was investigated with 50 ligands with folded bioactive structures. The assembly of the Vernalis folded compound set shows that folded bioactive conformations need to be taken into account, and it is dangerous to disfavor them. The compounds in the folded set were more flexible than those in the generic set, and comparisons took this difference into account. With the folded set, the performance of ConfGen degraded comparatively more than that of MOE Conformations Import, consistent with ConfGen disfavoring the folded conformers. This was confirmed by turning off the options enabling this ConfGen bias, leading to a marked improvement in reproduction of the folded bioactive structures. In addition, turning off these options did not degrade the ConfGen performance with the generic set of compounds.

In addition, we presented anecdotal docking failures with Glide which may be attributed to the penalization of folded conformers by ConfGen. A more complete study of this hypothesis is beyond the scope of the present work, but the present results suggest directions for future work.

In summary, LigPrep and ConfGen offer many parameters to adjust the trade-offs between computational resources and the quality/completeness of the output conformational models. The present analysis should help the LigPrep/ConfGen user to make informed choices among all these protocols. The ConfGen GUI is user-friendly, and the command line approach allows easy customization of the protocols. In fact, the proliferation of parameters could maybe benefit from some consolidation. ConfGen Optimized only requires the adjustment of a few key parameters from the default settings, making it straightforward to implement. It may be worth revisiting the deliberate penalization of folded conformers in ConfGen.

ABBREVIATIONS

3D, three-dimensional; ConfExt, percent of conformational extension; GB, generalized Born; GUI, graphical user interface; MaxConfs, maximum number of conformers generated per compounds; MMFF, Merck Molecular Force-Field; MOE, Molecular Operating Environment; MW, molecular weight; NbConfs, number of conformations; NRot, number of rotatable bonds; NTRB, nonterminal rotatable bonds; OPLS, Optimized Potentials for Liquid Simulations; PDB, Protein Data Bank; QSAR, quantitative structure−activity relationship; Rgyr, radius of gyration; RotSteps, number of search steps per degree of freedom; TRB, terminal rotatable bonds; %BioConf_Rep, percentage of reproduced bioactive conformations.

ACKNOWLEDGMENT

We thank the scientists from Schrodinger and Chemical Computing Group for answering our questions about ConfGen and MOE.

Supporting Information Available: List of PDB entries used for the Vernalis generic and folded sets of X-ray bioactive structures. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Patel, Y.; Gillet, V. J.; Bravi, G.; Leach, A. R. A Comparison of the Pharmacophore Identification Programs: Catalyst, DISCO and GASP. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 653–681.

(2) Dixon, S.; Smondyrev, A.; Knoll, E.; Rao, S.; Shaw, D.; Friesner, R. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 647–671.

(3) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional pharmacophore methods in drug discovery. *J. Med. Chem.* **2010**, *53*, 539–558.

(4) 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.

(5) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(6) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations. *J. Chem. Inf. Model.* **2007**, *47*, 390–399.

(7) Cheeseright, T. J.; Mackey, M. D.; Melville, J. L.; Vinter, J. G. FieldScreen: virtual screening using molecular fields. Application to the DUD data set. *J. Chem. Inf. Model.* **2008**, *48*, 2108–2117.

(8) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; Nevins, N.; Jain, A. N.; Kelley, B. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, Article ASAP.

(9) Wlodek, S.; Skillman, A. G.; Nicholls, A. Automated ligand placement and refinement with a combined force field and shape potential. *Acta Crystallogr.* **2006**, *D62*, 741–749.

(10) Foloppe, N.; Chen, I. Conformational Sampling and Energetics of Drug-like Molecules. *Curr. Med. Chem.* **2009**, *16*, 3381–3413.

(11) Leach, A. R. A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules. In *Rev. Comp. Ch.*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH: New York, 1991; Vol. 2, pp 1−55.

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

(13) Tresadern, G.; Agrafiotis, D. K. Conformational sampling with stochastic proximity embedding and self-organizing superimposition: establishing reasonable parameters for their practical use. *J. Chem. Inf. Model.* **2009**, *49*, 2786–2800.

(14) Musafia, B.; Senderowitz, H. Bioactive conformational biasing: a new method for focusing conformational ensembles on bioactive-like conformers. *J. Chem. Inf. Model.* **2009**, *49*, 2469–2480.

(15) Takagi, T.; Amano, M.; Tomimoto, M. Novel Method for the Evaluation of 3D Conformation Generators. *J. Chem. Inf. Model.* **2009**, *49*, 1377–1388.

(16) Bonnet, P.; Agrafiotis, D. K.; Zhu, F.; Martin, E. Conformational analysis of macrocycles: finding what common search methods miss. *J. Chem. Inf. Model.* **2009**, *49*, 2242–2259.

(17) Butler, K. T.; Luque, F. J.; Barril, X. Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.* **2009**, *30*, 601–610.

(18) Chen, I.; Foloppe, N. Conformational Sampling of Druglike Molecules with MOE and Catalyst: Implications for Pharmacophore Modeling and Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1773–1791.

(19) 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.

(20) Borodina, Y. V.; 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.

(21) Sadowski, J.; Boström, J. MIMUMBA revisited: torsion angle rules for conformer generation derived from X-ray structures. *J. Chem. Inf. Model.* **2006**, *46*, 2305–2309.

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

(23) Tirado-Rives, J.; Jorgensen, W. L. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. *J. Med. Chem.* **2006**, *49*, 5880–5884.
(24) 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.
(25) 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.
(26) 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.
(27) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational analysis by intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10–20.
(28) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, Article ASAP.
(29) Watts, K. S.; Dalal, P.; Murphy, R.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, Accepted for publication.
(30) Izrailev, S.; Zhu, F.; Agrafiotis, D. K. A distance geometry heuristic for expanding the range of geometries sampled during conformational search. *J. Comput. Chem.* **2006**, *27*, 1962–1969.
(31) Li, J.; Ehlers, T.; Sutter, J.; Varma-O'brien, S.; Kirchmair, J. CAESAR: A New Conformer Generation Algorithm Based on Recursive Buildup and Local Rotational Symmetry Consideration. *J. Chem. Inf. Model.* **2007**, *47*, 1923–1932.
(32) *Catalyst*, version 4.10; Accelrys: San Diego, CA. http://www.accelrys.com/ (accessed March 29, 2010).
(33) *ConfGen*, version 2.1; Schrödinger LLC: New York. http://www.schrodinger.com/ (accessed March 29, 2010).
(34) *MOE Conformation Import*, version 2008.10; Chemical Computing Group: Montreal, Canada. http://www.chemcomp.com/ (accessed March 29, 2010).
(35) *OMEGA*, version 2.3.2; OpenEye Scientific Software, Inc.: Santa Fe, NM. http://www.eyesopen.com/ (accessed March 29, 2010).
(36) *Rubicon*, version 4.9; Daylight Chemical Information Systems: Aliso Viejo, CA, 2008.
(37) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.
(38) Leite, T. B.; Gomes, D.; Miteva, M. A.; Chomilier, J.; Villoutreix, B. O.; Tuffery, P. Frog: a FRee Online druG 3D conformation generator. *Nucleic Acids Res.* **2007**, *35*, W568–W572.
(39) Liu, X.; Bai, F.; Ouyang, S.; Wang, X.; Li, H.; Jiang, H. Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinf.* **2009**, *10*, 101.
(40) Smi23D. http://depth-first.com/articles/2007/12/12/simple-3d-conformer-generation-with-smi23d/ (accessed March 29, 2010).
(41) Boström, 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.
(42) Boström, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Model.* **2003**, *21*, 449–462.
(43) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M. Epik: a software program for pKa prediction and protonation state generation for drug-like molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.
(44) Diller, D. J.; Merz, K. M. Can we separate active from inactive conformations. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 105–112.
(45) Moodie, S. L.; Thornton, J. M. A study into the effects of protein binding on nucleotide conformation. *Nucleic Acids Res.* **1993**, *21*, 1369–1380.
(46) Stockwell, G. R.; Thornton, J. M. Conformational Diversity of Ligands Bound to Proteins. *J. Mol. Biol.* **2006**, *356*, 928–944.
(47) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(48) Knowles, D.; Foloppe, N.; Matassova, N. B.; Murchie, A. I. H. The bacterial ribosome, a promising focus for structure-based drug design. *Curr. Opin. Pharmacol.* **2002**, *2*, 501–506.
(49) *LigPrep*, version 2.3; Schrödinger LLC: New York. http://www.schrodinger.com/ (accessed March 29, 2010).
(50) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.
(51) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
(52) *MOE (The Molecular Operating Environment)*, version 2008.10; Chemical Computing Group Inc.: Montreal, Canada, 2008.
(53) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.
(54) 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.
(55) Chen, I.; Hubbard, R. E. Lessons for fragment library design: analysis of output from multiple screening campaigns. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 603–620.
(56) Nicklaus, M. C.; Shaomeng, W.; Driscoll, J. S.; Milne, W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411–428.
(57) Boström, J.; Norrby, P. O.; Liljefors, T. Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383–396.
(58) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting Conformational Variation. *J. Comput. Chem.* **1995**, *16*, 171–187.
(59) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599–1608.
(60) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(61) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.
(62) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

CI100026X