JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

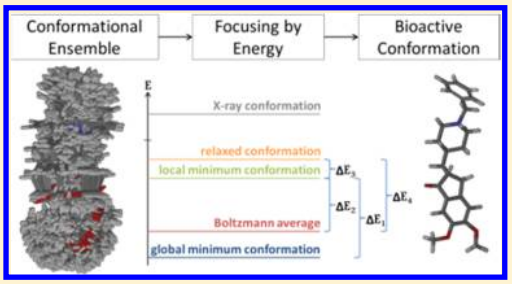
Toward Focusing Conformational Ensembles on Bioactive Conformations: A Molecular Mechanics/Quantum Mechanics Study

Hannah H. Avgy-David and Hanoch Senderowitz*

Department of Chemistry, Bar Ilan University, Ramat-Gan 52900, Israel

Supporting Information

ABSTRACT: The identification of bound conformations, namely, conformations adopted by ligands when binding their target is critical for target-based and ligand-based drug design. Bound conformations could be obtained computationally from unbound conformational ensembles generated by conformational search tools. However, these tools also generate many nonrelevant conformations thus requiring a focusing mechanism. To identify such a mechanism, this work focuses on a comparison of energies and structural properties of bound and unbound conformations for a set of FDA approved drugs whose complexes are available in the PDB. Unbound conformational ensembles were initially obtained with three force fields. These were merged, clustered, and reminimized using the same force fields and four QM methods. Bound conformations of all ligands were represented by their crystal structures or by approximations to these structures. Energy differences were calculated between global minima of the unbound state or the Boltzmann averaged energies of the unbound ensemble and the approximated bound conformations. Ligand conformations which resemble the X-ray conformation (RMSD < 1.0 Å) were obtained in 91%−97% and 96%−98% of the cases using the ensembles generated by the individual force fields and the reminimized ensembles, respectively, yet only in 52%−56% (original ensembles) and 47%−65% (reminimized ensembles) as global energy minima. The energy window within which the different methods identified the bound conformation (approximated by its closest local energy minimum) was found to be at 4−6 kcal/mol with respect to the global minimum and marginally lower with respect to a Boltzmann averaged energy of the unbound ensemble. Better approximations to the bound conformation obtained with a constrained minimization using the crystallographic B-factors or with a newly developed Knee Point Detection (KPD) method gave lower values (2−5 kcal/mol). Overall, QM methods gave lower energy differences than force field methods. These energy thresholds could be used for focusing conformational ensembles on bound conformations. For example, when using energy cutoffs which corresponded to retaining 50% and 70% of the ensembles, QM methods and CHARMm offer 60−65% and 80−84% probability of obtaining the bound conformation, respectively. In contrast, none of the structural criteria considered in this work was able to differentiate between bound and unbound conformations.



1. INTRODUCTION

The methods most often used in computer-aided drug design could be broadly classified as target-based and ligand-based. Both approaches critically depend on the ability to identify for each ligand its bound conformation (also known as the bioactive conformation), namely the conformation(s) adopted by the ligand when bound to its biotarget. In target-based modeling, seeding a docking algorithm with a relevant set of conformations (i.e., conformations which approximate the bioactive one) is likely to result in better predictions of the free energies of binding of the complexes and therefore in more accurate pose ranking. Moreover, identifying the bioactive conformation is essential for the calculation of the energy difference between the bound and unbound ligand conformations. This energy difference, often referred to as conformational energy, is frequently ignored by current scoring functions despite being an important part of the binding free energy. As pointed out by Tirado-Rives and Jorgensen, inaccurate conformational energies may compromise the ability of docking methods to rank-order diverse compounds during high-throughput virtual screening campaigns.[1] Bioactive conformations are also important for any 3D ligand-based approach which implicitly assumes such conformations are included in the input conformational ensemble (e.g., pharmacophore models).

Bound conformations of ligands can be determined experimentally by X-ray crystallography or NMR. However, crystal structures are limited for a few reasons.[2] First they represent the outcome of a strong selection procedure forced by the crystallization conditions and consequently may not reliably represent the solution conformational ensemble and perhaps not even the solid state ensemble. Second, crystal structures are subjected to errors both at the data acquisition and at the data refinement stages (e.g., fitting the atomic coordinates into experimental electron densities). Finally crystal structures convey static molecular models which can be misleading.[3]

Received: May 6, 2015
Published: September 25, 2015

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

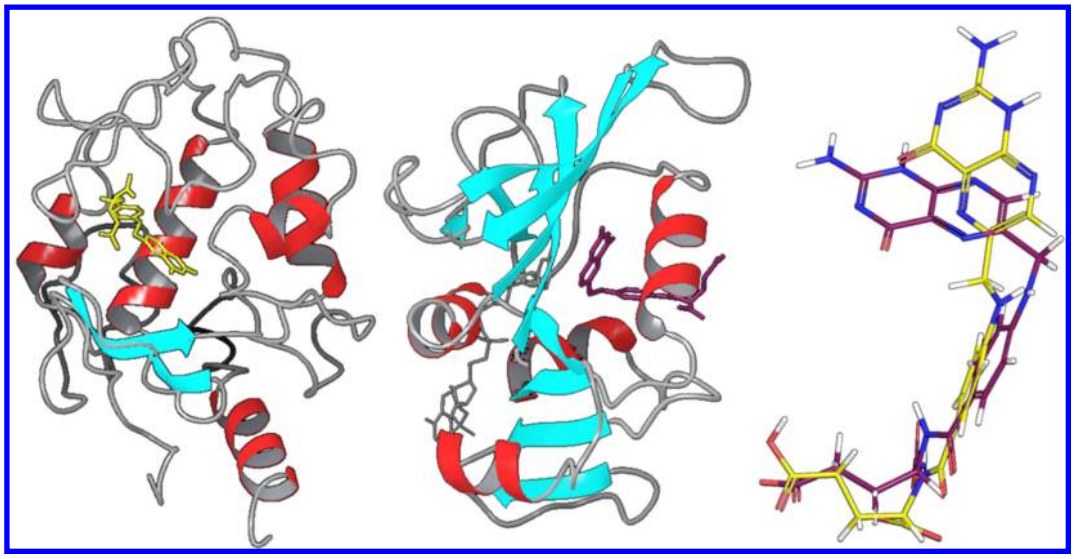


Figure 1. Bioactive conformations are target dependent. The conformations of folic acid when bound to human folate receptor alpha (PDB code: 4LRH, left) and to dihydrofolate reductase (PDB code: 3QL3, middle) are shown. The RMSD between these conformations (right) is 2.54 Å.

Alternatively, bound conformations can be obtained using a variety of computational methods. When the structure of the biotarget is known, docking methods could be used, even though their scoring functions have long been shown to be limited.[4] When the target structure is unknown, only the Potential Energy Surface (PES) of the unbound ligand can be sampled. Based on the conformational selection hypothesis,[5] this is a viable approach because the bound conformation is also accessible from the unbound PES. Sampling the PES of the unbound ligand could be performed using conformational search methods. However, this approach faces numerous challenges. First, bioactive conformations are target-dependent. Thus, a ligand may adopt different conformations when bound to different targets, as seen in Figure 1.[6]

Second, there is not necessarily a single bioactive conformation, not even for a specific target. Rather there may be multiple bioactive conformations whose number depends on the flexibility of the ligand and the characteristics (e.g., shape and flexibility) of the binding site. A third challenge is related to the parametrization of force fields, which are typically parametrized to reproduce (global) energy minima rather than bioactive conformations. The same holds true for quantum mechanical (QM) methods which although are not based on empirical parameters, cannot be expected to reproduce the bound state structure of a ligand in the absence of information on the biotarget. Finally, conformational search methods are mostly designed to produce a diverse set of conformations (with the hope that at least one will resemble the bound one). Consequently, the bound conformation, if at all located, is accompanied by many irrelevant conformations. Thus, identifying the bound conformation from within a conformational ensemble of the unbound ligand requires a focusing mechanism.[7]

Criteria for focusing the conformational ensemble of an unbound ligand on its bioactive conformation can be based on structural or energetic characteristics. Structural criteria assume that structures of bound ligands are different from those of unbound ligands in some consistent way. Energy criteria assume that bound ligand conformations are found within a reasonably low energy window from the energy of the unbound ligand. This assumption is clearly plausible since binding would only occur if the conformational energy associated with the unbound-to-bound transition is compensated for by ligand interactions with the binding site which in turn are likely to be capped at some value (for example, the binding of avidin to biotin which is one of the strongest in nature has an absolute free energy of 20.4 kcal/mol).[8]

Much work has been previously published proposing different structural criteria for focusing conformational ensembles on bioactive conformations.[9] A pioneering work by Diller and Merz suggested that bound conformations tend to have larger polar and apolar solvent accessible surface areas (SASA), fewer internal interactions, and a larger Radius of Gyration (ROG) than random conformations.[10] This was explained by the hypothesized tendency of small molecules to unfold when binding to a protein in order to maximize favorable interactions with key functionalities within the protein binding site. A similar conclusion was reached by Perola and Charifson based on an analysis of ligands from 150 crystal structures.[11] In a later study, Auer and Bajorath used a mining of emerging chemical patterns algorithm to identify patterns occurring in bound but not in modeled conformations of 18 sets of inhibitors, each targeting a specific protein.

Other studies have tried to focus conformational ensembles on bioactive conformations using an energy criterion, namely, the energy difference between the bound and unbound conformations of ligands. This conformational energy (sometimes referred to as strain energy, although this term is not entirely accurate since even at their global energy minima, molecules have some strain) has been extensively evaluated for its ability to distinguish bound and unbound conformations, but its accurate prediction is still challenging. Nicklaus et al. reported a study of 27 flexible ligands, where the calculated strain energy of the bound ligands varied between 0 and 39.7 kcal/mol, with an average of 15.9 kcal/mol.[12] Later, Perola and Charifson reported a study of 150 ligands in their protein-bound complexes.[11] For most ligands, they obtained strain energies of 2−9 kcal/mol. For a rather significant minority of ligands, the strain energies calculated were larger than 10 kcal/mol. Even more recently, Nicklaus et al. reported the calculation of strain energies quantum mechanically obtaining in most cases values below 25 kcal/mol.[2a]

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

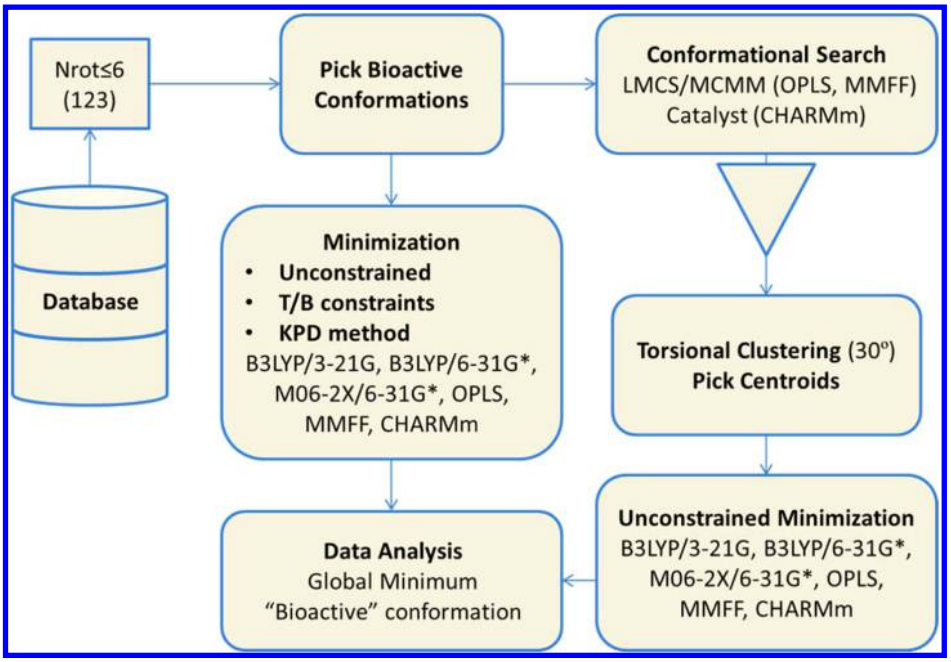


Figure 2. Project Workflow.

Some of the strain energies reported above, if correct, are likely to prevent ligand binding.[13] Such unrealistically high energies could result from inaccuracies in the definition of the bioactive conformation or the poor quality of the energy functions used. Even with a good energy model, one cannot compute meaningful energies directly from X-ray coordinates (i.e., the resulting energies are unrealistically high; see Table S1). These coordinates are primarily refined to fit the electron density map rather than be consistent with calibrated energy functions, although progress toward a better balance is being made.[14] Better energy functions are thus a necessity for these calculations, as well as better treatment of the crystallographic coordinates.

For this purpose, Butler et al. have proposed to relax X-ray derived bound conformations of ligands using harmonic constraints derived from the crystallization B-factors.[15] Energy differences (between the bound conformation and its closest local minimum) were calculated using quantum mechanics and provided a reasonable threshold for this energy difference. However, the strain energy upon binding was not estimated, as this would require the identification of the conformational ensemble in aqueous solution and not just the local minimum conformation.

More recently, Sitzmann et al. suggested a QM-based stepwise relaxation mechanism of bound conformations derived from ligand-protein crystal structures.[2a] They found many cases with energy differences larger than 20 kcal/mol, either between a partially relaxed (hydrogen atoms, bond lengths, bond angles) and a fully relaxed (hydrogen atoms, bond lengths, bond angles, torsional angles) bioactive conformation or between a partially relaxed bioactive conformation and the global energy minimum.

The present work focuses on a comparison between the conformational ensembles of ligands and their bioactive conformations with the aim of identifying suitable focusing criteria. The ligands studied in this work are all FDA approved drugs whose complexes are available in the PDB. Conformational ensembles were obtained by first subjecting each ligand to multiple conformational search procedures and then by merging and clustering the resulting ensembles. The resulting cluster centers forming reliable ensembles of the unbound state were reminimized using three force fields and four QM methods. The bound conformation of each ligand was represented by either its crystal structure or an approximated conformation generated from it through constrained or unconstrained energy minimization. The structural properties of bound and unbound conformations were compared in order to identify systematic differences between them. The strain energy was evaluated by subtracting the global energy minimum or the Boltzmann averaged energy of the unbound ensemble from the energy of the approximated bioactive conformations. We found that none of the structural descriptors considered in this work, either alone or in combinations, could distinguish between unbound and bound conformations, yet some focusing could be obtained using energy thresholds.

2. METHODS

2.1. Workflow. The overall workflow of the present study is depicted in Figure 2. Briefly, 100 FDA approved drugs with 1−6 rotatable bonds were extracted from their respective protein−ligand PBD complexes and submitted to a conformational search procedure using three different force fields: OPLS-AA,[16] MMFF,[17] and CHARMm.[18] The resulting conformational ensembles were combined and clustered. The centroid conformations from each cluster together with the X-ray conformation were minimized using the three above-mentioned force fields, and quantum mechanically, using two basis sets with the B3LYP functional and the larger basis set with the M06-2X functional.

2.2. Data Set. The ligands studied in this work are all FDA approved drugs whose target bound structures were available in the PDB on October 2011.[19] This work focused on 123 drugs with 1−6 rotatable bonds, which are not covalently bound to their target. Due to the relative rigidity of this subset, it was deemed amenable to conformational searches using high level

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

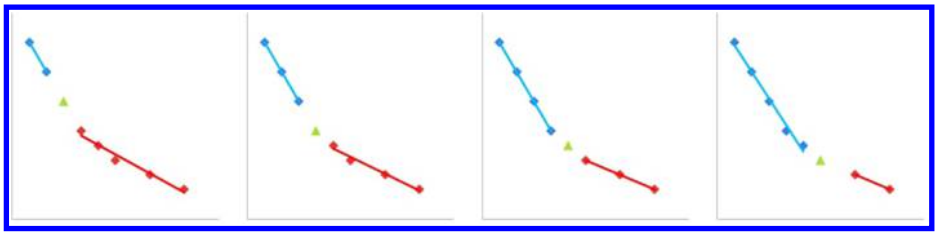


Figure 3. All possible pairs of best-fit lines for a graph that contains seven data points. Of the four possible line pairs, the third pair fits their respective data points with the smallest error and is therefore used to define the knee point of the graph (green triangle).

QM methods. Ligands were pretreated with the Epik software to assign tautomeric forms and protonation states in aqueous solution at pH 7.0. Following this procedure, 20%, 25%, and 55% of the ligands were found to be negatively charged, positively charged, or neutral, respectively. PDB codes and resolutions for the final set of ligand considered in this work are provided in Table S2.

2.3. Conformational Search. Conformational searches were performed using the Mixed Monte Carlo Multiple Minimum/Low-Mode Conformational Search (MCMM/LMCS) method as implemented in MacroModel[20] and the polling algorithm as implemented in Discovery Studio (DS) 3.1.[21] MCMM/LMCS was performed with the OPLS-AA and MMFF force fields. The conformational searches were performed in aqueous solution using the Generalized Born/solvent accessible surface area (GB/SA) solvent model developed by Still et al.[22] Energy minimizations were carried out with the Truncated Newton−Raphson Conjugate Gradient algorithm using a convergence criterion of 0.02 kcal/mol. High energy conformers (outside an energy window of 5 kcal/mol above the current global minimum) were discarded. Structures produced during a conformational search were compared to see if they are unique. Two structures were considered different if their RMSD exceeded a 0.5 Å threshold. If the RMSD was lower than 0.5 Å, the higher energy conformation was removed.

DS calculations employed the BEST method with the CHARMm force field and the Generalized Born with Molecular Volume (GBMV) solvent model developed by Lee et al.[23] Redundant conformations (where redundancy was calculated as a combination of global RMSD and local differences between atom pairs) or conformations outside an energy window of 20 kcal/mol were discarded.

In all cases, conformational ensembles were found to be independent of the ligand's starting conformation (i.e., using a 2D starting point or the crystal structure). For each molecule, the conformations obtained by the different conformational searches were combined into a single conformational ensemble.

2.4. Torsional Clustering. Combined conformational ensembles were subjected to agglomerative hierarchical clustering in torsional space (developed in-house). The maximum torsional distance within each cluster was set to 30°, that is, in each cluster all torsions were within 30° from each other. Dihedral angles which are not fully rotatable such as in cyclohexane were also taken into account. Clusters were represented by a member of the cluster closest to the centroid. The new conformational ensembles consisted of the representative conformers of the clusters (results shown in Figure S1).

2.5. Minimization. *2.5.1. Force Field Minimization.* The clustered conformational ensembles for all ligands, as well as their crystal structures, were minimized with three different force fields. The Polak Ribière Conjugate Gradient (PRCG) method was used for minimization with OPLS-AA and MMFF.[24] The Smart Minimizer algorithm as implemented in DS was used for CHARMm minimization. The same solvation models as before were used.

2.5.2. QM Minimization. Quantum mechanical calculations were performed with the Gaussian09 package.[25] The B3LYP density functional method was used with the 3-21G and 6-31G* basis sets. The M06-2X method was used with the 6-31G* basis set. Calculations were done in water using the integral equation formalism of the polarizable continuum model (IEF-PCM).[26] The geometry and initial guess for the B3LYP/6-31G* calculations were read from the checkpoint file of the B3LYP/3-21G calculations.

The B3LYP/6-31G* chemistry model is often criticized for thermochemical problems.[27] Therefore, the B3LYP/6-31G*-gCP-D3 model proposed by Grimme was used as a more robust and physically sound option.[28] This scheme removes two of the major deficiencies of the B3LYP/6-31G* model. The D3-DFT term accounts for the missing London dispersion effects, while the gCP (geometrical counterpoise correction) term is treating the basis set superposition error (BSSE).

2.5.3. Minimization of Bioactive Conformations. As discussed above, calculating energies of bound conformations directly from their crystal structures is meaningless. In this work, different procedures for obtaining conformations which are energetically relaxed yet geometrically compatible with the X-ray coordinates were therefore attempted. The procedures for relaxing the bound conformation are as follows:

2.5.3.1. Unconstrained Minimization. First, the X-ray conformation was minimized toward its closest local minimum conformation. This approach may of course provide a poor approximation to the bioactive conformation since it may be structurally remote from its closest local minima.

2.5.3.2. T/B Constrained Minimization. Following the method proposed by Butler et al.,[15] crystal conformations were minimized using harmonic constrains derived from the crystallographic temperature B-factors (eq 1)

$$\kappa = 4\pi^2 k_B T/B \tag{1}$$

where κ is the force constant of the harmonic potential, k_B is the Boltzmann constant, T is the data collection temperature (around 100 K), and B denotes the atomic B-factor.

2.5.3.3. Knee Point Detection (KPD). Minimizing a structure toward its closest (local) energy minimum typically generates a series of conformations with decreasing energies (except for when the minimization algorithm takes an "uphill" step). The KPD approach developed in this paper takes advantage of this behavior to locate better approximations to the geometry of the bioactive conformation than those provided by an unconstrained minimization end-point. The best approximation is taken to be the one that best balances low energy (i.e., closeness to the energy of the corresponding local minimum)

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

and low RMSD with respect to the starting point of the minimization (i.e., the X-ray conformation). This conformation typically corresponds to a knee point on an "energy vs RMSD" graph (Figure S2). The "knee" of such graph is loosely defined as the point of maximum curvature and can be determined with various methods.[29] In this work, the knee point was detected using the L-method (Figure 3).[30] According to this method, the data points are repeatedly divided into two sets with a minimum of two points per set. Each set is approximated by a linear function. and the knee point is defined as the division point which leads to the best overall fit. The main advantage of the KPD method, in comparison with several constraints-based approaches, is that it locates an approximation to the bioactive conformation in an unbiased manner.

2.6. Data Analysis. *2.6.1. Calculation of Conformational Energy Differences.* For each ligand, the conformational energy difference between its bound and unbound states was calculated by subtracting the energy of its global minimum from the energy of its approximated bound conformation, where both structures were solvated. Approximating the bound conformation by its closest local minimum may lead to a zero difference if this local minimum happens to coincide with the global energy minimum.

2.6.2. Calculation of Boltzmann Averaged Conformational Energy Differences. The true conformational energy of the unbound ligand is not determined by its global minimum conformation only but rather by the whole conformational ensemble. Thus, a more accurate estimate of the energy of the unbound state could be obtained by Boltzmann averaging the energies of the conformational ensemble. Interestingly, this method was not used in previous studies. Energy differences based on Boltzmann averaged energies were calculated according to eq 2

$$\Delta E_i = E_i^{Boltz} - E_i^{bio} \tag{2}$$

where ΔE_i, E_i^{Boltz}, and E_i^{bio} denote, respectively, the energy difference, the Boltzmann-averaged energy, and the energy of the (approximated) bound conformation of ligand i. Boltzmann-averaged energies were calculated according to eq 3

$$E_i^{Boltz} = \sum_j P_i^j * E_i^j \tag{3}$$

where E_i^j denotes the energy of conformer j of ligand i, and P_i^j denotes the probability of occurrence of conformer j of ligand i within the ensemble. P_i^j was calculated according to eq 4

$$P_i^j = \frac{e^{(-\Delta E_i^j)/(k_B T)}}{\sum_k e^{(-\Delta E_i^k)/(k_B T)}} \tag{4}$$

where T is the absolute temperature, k_B is the Boltzmann constant, and the summation runs over all members of the ensemble of conformations.

2.7. Calculation of Structural Descriptors. Radii of gyration (ROG), Jurs descriptors, shadow indices, molecular volume, dipole magnitude, and principal moments of inertia were calculated for the conformational ensembles obtained with the M06-2X/6-31G* chemistry model and for the bound conformations. Descriptors were calculated with Accelrys Discovery Studio 3.1. Descriptor values for bound conformations were compared with their mean values across the conformational ensembles. Student's t-test was performed for each descriptor to determine whether there is a significant difference between its values for bound and unbound conformations.

In addition, each descriptor was characterized by its information gain. The information gain of a descriptor reflects its ability to split a data set into two "homogeneous" groups (e.g., bound vs unbound). Homogeneity is determined by Shannon's entropy measure.[31]

2.8. Calculation of RMSD. RMSD calculations were performed in order to evaluate the ability of the "standard" conformational analysis procedures as well as of the conformational search workflow developed in this work to reproduce bound conformations. These calculations were therefore carried out separately for each of the clustered conformational ensembles resulting from the six minimization methods (OPLS-AA, MMFF, CHARMm, B3LYP/3-21G, B3LYP/6-31G*, and M06-2X/6-31G*) as well as for the ensembles resulting from the three initial conformational searches. For the purpose of RMSD calculations, the bioactive conformation was taken directly from the crystal.

Two RMSD calculation procedures were carried out for each ligand. First, the RMSD between the bound conformation and the global energy minimum of the ensemble was calculated. Additionally, the RMSD was calculated between the bound conformation and the closest conformation in the ensemble, termed here "best RMSD". All RMSD calculations were performed with the python script rmsd.py which is part of Schrödinger's Maestro suite 2011.

3. RESULTS

This work presents systematic force field and QM-based conformational searches for a large set of drug compounds in order to sample their unbound ensembles and focus these ensembles on bioactive conformations. The search was implemented by first generating a large conformational ensemble for each compound using three different force fields (OPLS-AA, MMFF, CHARMm) and then by reminimizing a representative (yet large) subset of the ensemble using both the original force fields and three QM methods (B3LYP/3-21G, B3LYP/6-31G*, M06-2X/6-31G*). It is therefore a relevant question whether the entire representative subset should have been subjected to QM minimization or would it have sufficed to only reminimize its low energy conformations resulting from the less time-consuming force field calculations. For this purpose the correlation between force field-calculated and QM-calculated conformational energies was reviewed. The results for a representative subset of six ligands (with 1−6 rotational bonds) are presented in Figure S3 and Table S3 of the Supporting Information. These results demonstrate a correlation between energy differences resulting from different QM methods but almost no correlation between the different force fields or between the force fields and QM methods. Therefore, force field calculations cannot be used to filter conformational ensembles prior to the more time-consuming QM calculations.

3.1. Can Conformational Search Methods Generate Bioactive Conformations? The RMSDs between the bound conformation and the closest conformation in the ensemble, termed here "Best RMSDs", are shown in Tables 1a and 1b. At the level of the individual force fields, CHARMm clearly performs best, being able to identify for 78% of the ligands at least a single conformation within 0.5 Å of the bound conformation. Reminimization of the clustered ensembles improved the performances of the OPLS and MMFF force fields up to the same level as CHARMm. QM-based

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

Table 1a. Best RMSDs for the Conformational Search Results Using the Individual Force Fields

	<0.5 Å[a]	<1.0 Å	<1.5 Å	<2.0 Å	>2.0 Å
OPLS[b]	0.62[c]	0.91	0.97	0.99	0.01
MMFF	0.63	0.92	0.98	1.00	0.00
CHARMm	0.78	0.97	1.00	1.00	0.00

[a]An upper limit of RMSD. [b]The computational model. [c]The fraction of ligands with Best RMSD within this limit.

Table 1b. Best RMSDs for the Conformational Analysis Results Using the Workflow

	<0.5 Å[a]	<1.0 Å	<1.5 Å
OPLS[b]	0.71[c]	0.97	1.00
MMFF	0.73	0.96	1.00
CHARMm	0.74	0.96	1.00
B3LYP/3-21G	0.72	0.98	1.00
B3LYP/6-31G*[d]	0.78	0.97	1.00
M06-2X/6-31G*[d]	0.77	0.98	1.00

[a]An upper limit of RMSD. [b]The computational model. [c]The fraction of ligands with Best RMSD within this limit. [d]The number of ligands in the data set is different for each method as some ligands failed the minimization in some force fields.

minimization led to marginally better results, particularly for the 6-31G* basis set.

Low resolution structures may result in inaccurate bound conformations. To test this hypothesis, the correlation between the crystallographic resolution and the best RMSD value for all ligands was calculated for the reminimized ensembles (Figure 4). The results clearly demonstrate that there is no influence of the X-ray resolution of a protein−ligand complex on the Best RMSD observed for this ligand.

3.2. Can Bioactive Conformations Be Identified from within Conformational Ensembles? The data in Tables 1a and 1b demonstrate that conformational search methods can produce bioactive (or bioactive-like) conformations for most ligands, yet the RMSD values (Table 2) suggest that these methods also produce many conformations with little resemblance to the bound one. In this table, as well as in subsequent tables where average RMSD values are presented, we also provide median RMSD values, to demonstrate that the results are not biased by "outlier RMSD" values. Thus, even after a thorough clustering procedure, most compounds have relatively large ensembles, as presented in Figure S1. The ensemble size is proportional to molecular flexibility.

Clearly, reducing the number of irrelevant conformations will benefit any application which relies on the availability of bioactive conformations as an input (e.g., docking, pharmacophore). Filtering the conformational ensemble by some reasonable means is therefore essential.

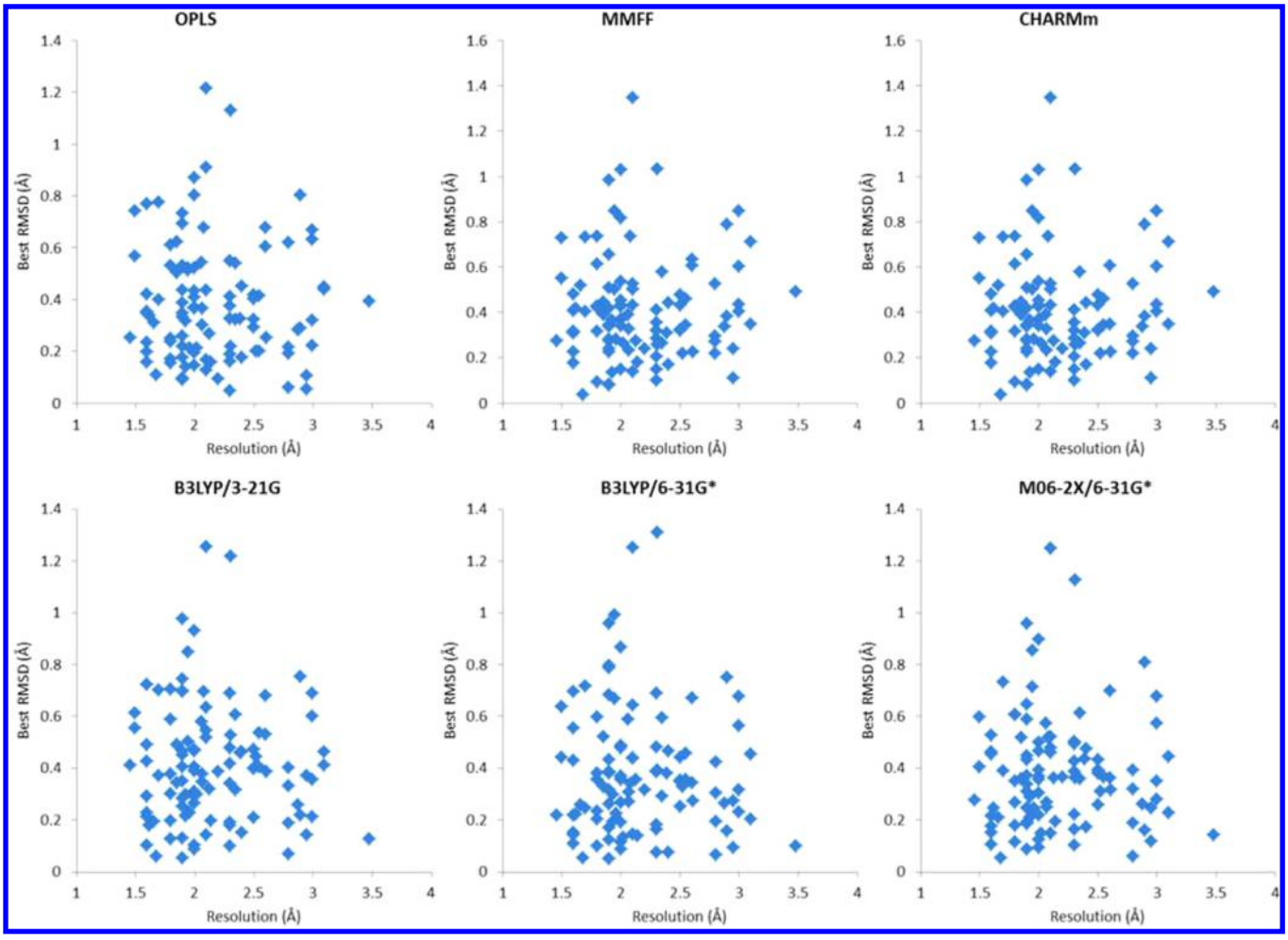


Figure 4. Correlations between the X-ray resolution and the Best RMSD values for the different methods considered in this work.

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

Table 2. Average RMSD Values from the Bound Conformation of the Conformational Ensembles Generated by the Different Methods Considered in This Work[a]

OPLS	MMFF	CHARMm	B3LYP/3-21G	B3LYP/6-31G*	M06-2X/6-31G*
1.37 ± 0.67 (1.24)	1.40 ± 0.67 (1.30)	1.40 ± 0.67 (1.30)	1.39 ± 0.65 (1.29)	1.35 ± 0.66 (1.23)	1.41 ± 0.69 (1.29)

[a]The median is shown in parentheses, and its overall similarity to the average shows that the latter is not greatly biased by "outlier RMSD" values.

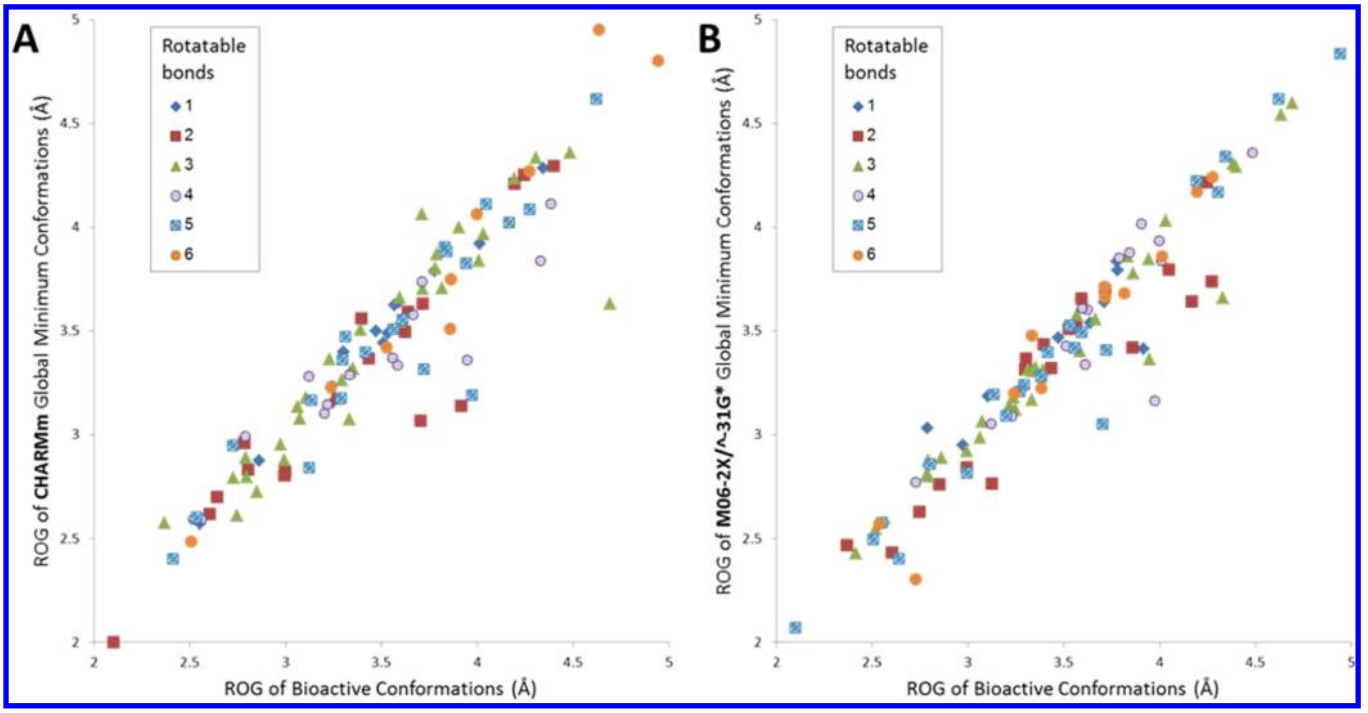


Figure 5. Comparison between the radius of gyration of bound conformations (crystal structure) and global minima ones for ligands with different numbers of rotatable bonds. (A) CHARMm generated global minima. (B) M06-2X/6-31G* generated global minima.

3.2.1. Can Bioactive Conformations Be Identified Based on Their Structures? Previously, several studies used measurements such as solvent accessible surface area (SASA) or radius of gyration (ROG) to suggest that bioactive conformations are more elongated than global minima ones.[11] These findings led to the hypothesis that small molecules tend to unfold when binding to a protein, in order to maximize favorable interactions with key functionalities within the protein binding site.

To further probe this hypothesis, the radii of gyration (ROG) of bound conformations were compared to those of their corresponding CHARMm and M06-2X/6-31G* generated global minimum conformations (Figure 5). The average ROGs were found to be 3.48 Å for the bound conformations and 3.43 Å and 3.44 Å for the global minima generated by CHARMm and M06-2X/6-31G*, respectively. Thus, in contrast with previous reports, the current set of data does not support the elongation of bound conformations upon binding and precludes the usage of measures such as ROG for focusing conformational ensembles on such conformations.

Other 3D descriptors were tested to determine whether their values differ between bound and unbound conformations. Mean values of these descriptors across conformational ensembles and across bound conformations were calculated and compared by Student's *t*-test. The results (Table 3) demonstrate that none of the descriptors considered in this study can be used, by itself, to identify bound conformations. Furthermore, the information gain for all these descriptors was found to be zero.

3.2.2. Can Bioactive Conformations Be Identified Based on Their Energies? 3.2.2.1. Can Conformational Search Methods Generate Bioactive Conformations As Global Minima? The conformational energy is theoretically a good criterion for focusing conformational ensembles on bioactive conformations. Tables 4a and 4b show the RMSDs between the bound conformation and the global energy minimum obtained with the individual methods evaluated in this paper (Table 4a) compared with the workflow (Table 4b). The workflow produces bioactive-like conformations as global minima somewhat more often than standard conformational search methods. These results are not surprising as the workflow samples a wider conformational space. OPLS produces bioactive-like conformations as global minima most often, with 65% and 83% of the global minima within 1.0 and 1.5 Å of the bound conformation, respectively.[34] In this respect, QM methods do not perform better than force fields.

Using the most demanding criterion (RMSD of the global energy minimum with respect to the bound conformation <0.5 Å), the 6-31G*/B3LYP chemistry model with or without the Grimme correction performs best, albeit only marginally better than OPLS. As noted above, for higher RMSD values, force field based methods are as good as or even slightly better than QM based ones. Interestingly, in all cases the performances of the M06-2X functional were inferior to those of B3LYP and to some of the force fields. These data clearly demonstrate that global energy minima are poor predictors of bioactive conformations. Other methods for focusing conformational ensembles on bioactive conformations are therefore needed.

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

Table 3. p-Values for the 3D Descriptors Considered in This Work[b]

descriptor	p-value	descriptor	p-value
ROG	0.81	Jurs_RPCG	0.32
Jurs_DPSA_1[a]	0.98	Jurs_RPCS	0.32
Jurs_DPSA_2	0.95	Jurs_RPSA	0.32
Jurs_DPSA_3	0.54	Jurs_SASA	0.88
Jurs_FNSA_1	0.32	Jurs_TASA	0.96
Jurs_FNSA_2	0.32	Jurs_TPSA	0.64
Jurs_FNSA_3	0.32	Jurs_WNSA_1	0.60
Jurs_FPSA_1	0.98	Jurs_WNSA_2	0.88
Jurs_FPSA_2	0.32	Jurs_WNSA_3	0.65
Jurs_FPSA_3	0.32	Jurs_WPSA_1	0.80
Jurs_PNSA_1	0.81	Jurs_WPSA_2	0.86
Jurs_PNSA_2	0.96	Jurs_WPSA_3	0.26
Jurs_PNSA_3	0.78	Shadow_Xlength	0.30
Jurs_PPSA_1	0.93	Shadow_Ylength	0.30
Jurs_PPSA_2	0.99	Shadow_Zlength	0.31
Jurs_PPSA_3	0.26	Molecular_Volume	0.98
Jurs_RASA	0.32	Dipole_mag	0.34
Jurs_RNCG	0.32	PMI_mag	0.75
Jurs_RNCS	0.30		

[a]Jurs descriptors combine shape and electronic information to characterize molecules.[32] Shadow indices are calculated by projecting the molecule on three mutually perpendicular planes defined by the principal moments of inertia and by characterizing the resulting projection. The shadow indices calculated herein are the length of the molecule in the three possible dimensions.[33] PMI is the magnitude of the principle moment of inertia. [b]The lowest p-value is 0.26 indicating no significant difference between the values of bound conformations and the rest of the ensemble.

Table 4a. RMSD of the Global Minimum from the Bound Conformation for the Conformational Search Results

	<0.5 Å[a]	<1.0 Å	<1.5 Å	<2.0 Å	>2.0 Å
OPLS[b]	0.22[c]	0.56	0.73	0.88	0.12
MMFF	0.20	0.54	0.79	0.93	0.07
CHARMm	0.20	0.52	0.74	0.91	0.09

[a]An upper limit of RMSD. [b]The force field. [c]The fraction of ligands with RMSD of global minimum within this limit.

Table 4b. RMSD of the Global Minimum from the Bound Conformation for the Conformational Analysis Results

	<0.5 Å[a]	<1.0 Å	<1.5 Å	<2.0 Å	>2.0 Å
OPLS[b]	0.25[c]	0.65	0.83	0.95	0.05
MMFF	0.25	0.47	0.76	0.93	0.07
CHARMm	0.23	0.55	0.76	0.94	0.06
B3LYP/3-21G	0.21	0.48	0.75	0.95	0.05
B3LYP/6-31G*	0.27	0.58	0.78	0.95	0.05
B3LYP/6-31G*-gcP-D3	0.27	0.55	0.78	0.95	0.05
M06-2X/6-31G*	0.20	0.52	0.72	0.95	0.05

[a]An upper limit of RMSD. [b]The force field. [c]The fraction of ligands with RMSD of global minimum within this limit.

3.2.2.2. Could an Energy Threshold Be Defined within Which Conformational Search Methods Identify Bioactive Conformations? As discussed in the Introduction, several attempts have been made to define an energy threshold above the global energy minimum within which the bound conformation can be found. Once identified, this threshold can serve as a criterion for focusing conformational ensembles on bioactive conformations. To date, this energy threshold was typically based on the calculation of conformational energies, namely the energy difference between (an approximation to) the bioactive conformation and the global minimum. Such energy differences for the set of compounds considered in this study are presented in Figure 6. These differences are calculated by subtracting the energy of the global minimum conformation from the energy of the bound conformation approximated as its closest local energy.

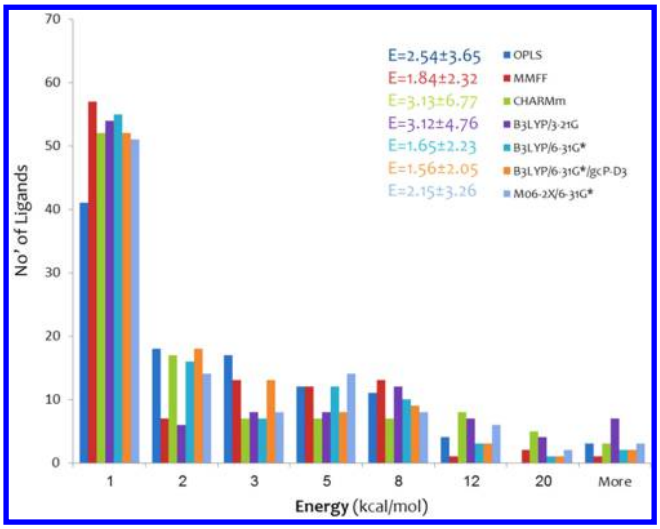


Figure 6. Energy differences for the data set (ΔE_1 in Figure 8; see below). Energy differences were calculated by subtracting the energy of the global minimum conformation from the energy of the bound conformation approximated as its closest local energy minimum.

The B3LYP/6-31G* chemistry model yielded lower energy differences than B3LYP/3-21G, with an average of 1.65 kcal/mol compared with the 3.12 kcal/mol of 3-21G. These differences were reduced to an average of 1.56 kcal/mol when adding the Grimme correction to the B3LYP/6-31G* chemistry model. Moreover, these results have a smaller standard deviation than the uncorrected results. The considerably more accurate method M06-2X yielded higher energy differences than those obtained by B3LYP at the same level, with an average of 2.15 kcal/mol. It also showed a much larger standard deviation. Overall, the majority of the energy differences are below 8 kcal/mol, and the data support energy cutoffs of 4−6 kcal/mol for focusing conformational ensembles on bioactive conformations (under the closest local minimum approximation to the bound conformation).

Based on average values, the performance of force field based and QM based methods seems similar. However, analyzing the results on a per-ligand basis conveys a significant difference. Table 5 presents the correlation (in terms of R^2 values) between energy differences obtained by the different methods over all ligands in the data set. A correlation between two methods would support their reliability, while the lack of correlation would point out an inaccuracy of at least one of the methods. The results in Table 5 show that good correlations were obtained among all QM methods, whereas the force field methods are uncorrelated. Most notable is the CHARMm force field, which shows no correlation whatsoever with the results of other methods. It can therefore be concluded that QM-based methods provide more reliable energy differences than force field based methods.

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

Table 5. Correlation (R^2) for the Energy Differences Calculated with the Various Methods Considered in This Work[a]

	OPLS	MMFF	CHARMm	B3LYP/3-21G	B3LYP/6-31G*	B3LYP/6-31G* -gcP-D3	M06-2X/6-31G*
OPLS	1.00	0.39	0.00	0.10	0.13	0.13	0.30
MMFF	0.39	1.00	0.00	0.23	0.27	0.24	0.22
CHARMm	0.00	0.00	1.00	0.00	0.00	0.00	0.00
B3LYP/3-21G	0.10	0.23	0.00	1.00	0.86	0.84	0.62
B3LYP/6-31G*	0.13	0.27	0.00	0.86	1.00	0.98	0.71
B3LYP/6-31G*-gcP-D3	0.13	0.24	0.00	0.84	0.98	1.00	0.74
M06-2X/6-31G*	0.30	0.22	0.00	0.62	0.71	0.74	1.00

[a]Energy differences were calculated by subtracting the energy of the global minimum conformation from the energy of the bound conformation approximated as its closest local energy minimum.

3.2.2.3. The Energy Cost of the Bioactive Conformation. In the previous section, it was suggested that bioactive conformations approximated by closest local minima are typically found within an energy threshold of 4−6 kcal/mol above the global energy minimum. This value can therefore be used to focus conformational ensembles on bioactive conformations and concurrently report on the energy cost a ligand pays to attain the bioactive conformation. However, these energy thresholds are inadequate estimators of the bioactive energy cost for two main reasons: (1) They equate the energies of the bound conformation with its closest local energy minimum. (2) They equate the energies of the unbound ligand with its global energy minimum. Neither of these assumptions is correct. The energy of the unbound ligand can be better approximated by a Boltzmann averaged energy calculated for the entire ensemble rather than the energy of the global minimum conformation alone. The first inaccuracy concerning the energy of the bioactive conformation is more difficult to improve and requires a better approximation of this conformation.

Boltzmann Averaged Energy Costs. The Boltzmann averaged energies calculated with respect to the energy minima closest to the bound conformation are presented in Figure 7. Not surprisingly, these penalties are on average lower than the above-described numbers but are still in the range of 5 kcal/mol. As before, better correlations were observed between the QM-based methods than between the force field based methods (Table 6).

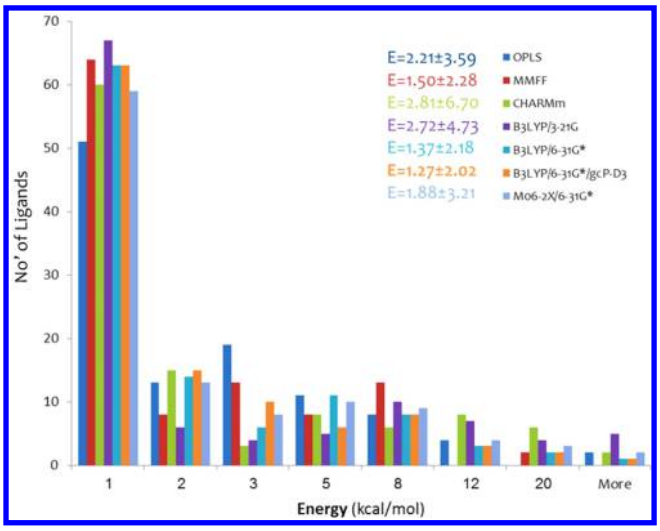


Figure 7. Boltzmann energy differences for the data set (ΔE_2 in Figure 8; see below). Energy differences were calculated by subtracting the Boltzmann averaged energy over the conformational ensemble from the energy of the bound conformation approximated as its closest local energy minimum.

The Effect of Better Approximations to the Bioactive Conformations on Energy Costs. For the data set considered in this work, averaged RMSD values between crystal bound structures and their corresponding closest local minima are between 0.42 and 0.46 Å (Table 7) suggesting that closest energy minima are often poor approximations to the bound conformations. Other means are thus required to relax the bound conformation to allow better evaluation of its energy (Figure 8).

T/B Constrained Minimization of Bound Conformations. Crystal bound conformations were relaxed using constraints derived from B-factors (T/B constraints) as previously proposed by Butler et al. (see eq 1).[15] The resulting energy differences (ΔE_3 in Figure 8) are presented in Figure 9, and the RMSD values with respect to the bound conformation are presented in Table 7.

The Knee Point Detection (KPD) Method. Crystal bound conformations were also relaxed using the KPD method (see Methods section). The energy differences calculated by this method are presented in Figure 10. Force fields show lower energy differences than QM calculations, presumably because in many cases (20%, 50%, and 34% for OPLS, MMFF, and CHARMm, respectively) negative energy differences were obtained. In contrast negative energy differences were rarely obtained with QM methods (less than 5%). The relaxed bound conformations are within 0.22−0.37 Å of the X-ray conformation, closer than those obtained with the unconstrained minimization (0.42−0.46 Å) and similar to those obtained with the T/B minimization (0.13−0.37 Å; see Table 7).

Negative penalties found with both the KPD method and the T/B constrained minimization indicate that nonstationary conformations with energies lower than the local minima were visited in the course of the minimization process. This in turn implies that additional minima, lower in energy than the minimization end points, exist in the vicinity of the bound conformations but were not identified by the minimizers. Changing the minimization parameters (e.g., step size) or alternatively switching to a different minimizer may be required to locate these minima. Thus, the numbers given in Figures 9−12 are lower-bound.

Figures 11 and 12 provide the best estimates for the energy differences between bound and unbound ligand conformations obtained in this work.

3.3. Can Conformational Ensembles Be Focused on Bioactive Conformations? An energy threshold can only be used to focus conformational ensembles on bioactive conformations if it is high enough to include the bioactive

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

Table 6. Correlation (R^2) for the Energy Differences Claculated with the Various Methods Considered in This Work[a]

	OPLS	MMFF	CHARMm	B3LYP/3-21G	B3LYP/6-31G*	B3LYP/6-31G*-gcP-D3	M06-2X/6-31G*
OPLS	1.00	0.36	0.00	0.06	0.11	0.11	0.28
MMFF	0.36	1.00	0.00	0.15	0.24	0.21	0.20
CHARMm	0.00	0.00	1.00	0.00	0.00	0.00	0.00
B3LYP/3-21G	0.06	0.15	0.00	1.00	0.71	0.71	0.52
B3LYP/6-31G*	0.11	0.24	0.00	0.71	1.00	0.98	0.70
B3LYP/6-31G*-gcP-D3	0.11	0.21	0.00	0.71	0.98	1.00	0.73
M06-2X/6-31G*	0.28	0.20	0.00	0.52	0.70	0.73	1.00

[a]Energy differences were calculated by subtracting the Boltzmann averaged energy across the conformational ensemble from the energy of the bound conformation approximated as its closest local energy minimum.

Table 7. Average RMSD Values (Å) of Minimized Conformations from the Crystal Conformation[d]

	local minimum[a]	T/B constraints[b]	knee point[c]
OPLS	0.43 ± 0.33 (0.34)	0.38 ± 0.25 (0.33)	0.36 ± 0.26 (0.30)
MMFF	0.44 ± 0.31 (0.36)	0.20 ± 0.11 (0.17)	0.36 ± 0.27 (0.29)
CHARMm	0.42 ± 0.28 (0.35)	0.13 ± 0.07 (0.13)	0.22 ± 0.18 (0.17)
B3LYP/3-21G	0.45 ± 0.35 (0.37)	0.36 ± 0.22 (0.30)	0.38 ± 0.30 (0.29)
B3LYP/6-31G*	0.44 ± 0.37 (0.35)	0.30 ± 0.20 (0.24)	0.36 ± 0.35 (0.28)
B3LYP/6-31G*-gcP-D3	0.44 ± 0.37 (0.35)	0.30 ± 0.20 (0.24)	0.34 ± 0.33 (0.28)
M06-2X/6-31G*	0.42 ± 0.32 (0.36)	0.30 ± 0.18 (0.27)	0.35 ± 0.28 (0.29)

[a]RMSD between the X-ray conformation and its closest local minimum. [b]RMSD between the X-ray conformation and its minimized conformation under T/B constraints. [c]RMSD between the X-ray conformation and the knee point detection method during its minimization. [d]Median values are given in parentheses and support the lack of a biasing effect by high RMSD values. The median and average values follow similar trends, both being lower for the T/B and KPD methods than for the unconstrained minimization.

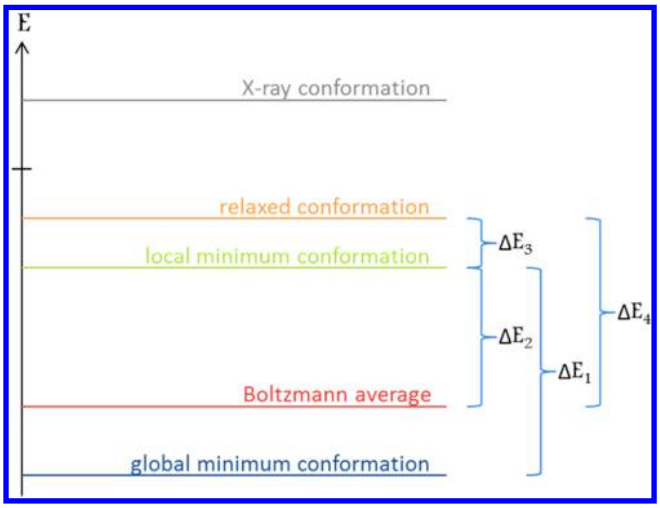


Figure 8. Diagram shows schematic levels of the different energies calculated in this work. The relaxed conformations can be obtained by different constrained minimization methods. ΔE_1 is obtained by calculating the energy difference between the global minimum and the local minimum closest to the bioactive conformation (see Figure 6 above). ΔE_2 is obtained by calculating the energy difference between the Boltzmann averaged energy and the local minimum closest to the bioactive conformation (Figure 7). ΔE_3 is obtained by calculating the energy difference between approximations to the bound conformations obtained through T/B constrained minimization or the KPD method (Figures 9 and 10, respectively). $\Delta E_4 = \Delta E_2 + \Delta E_3$ is the best approximation to the energy difference between the bound and unbound conformation (see Table 8).

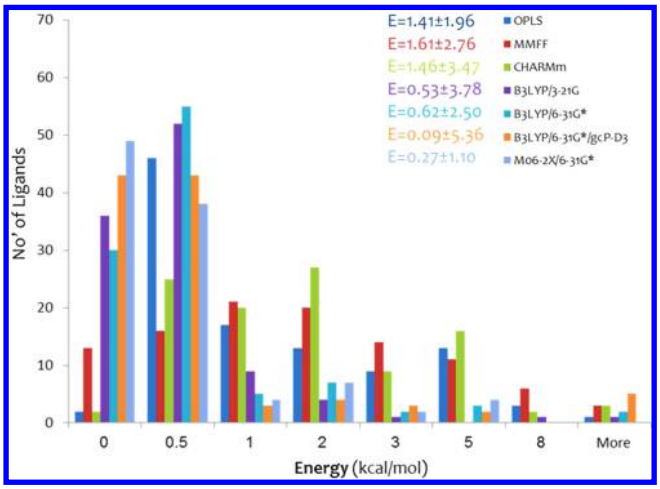


Figure 9. Energy differences obtained from the relaxation of the bound conformation with T/B constraints (ΔE_3 in Figure 8). (*) A few QM jobs did not converge. Therefore, this histogram shows energy differences for 98 and 99 ligands for the B3LYP methods and the M0602X/6-31G* model, respectively.

conformation yet low enough to exclude many irrelevant conformations. Using small energy thresholds will undoubtedly reduce the size of the ensemble but at the same time may miss the bioactive conformation altogether. This trade-off is illustrated in Figure 13 where each energy cutoff defines a point on a "probability of obtaining the bound conformation vs. ensemble size" graph. For this purpose, energy differences between the Boltzmann averaged energy and the local minimum closest to the bioactive conformation (ΔE_2) were used (because they do not require T/B or KDP minimization). These data suggest that QM methods and CHARMm perform better than the other force fields. The B3LYP/3-21G chemistry model is somewhere in the middle but is still better than OPLS and MMFF. When retaining 50% of the ensembles, QM methods offer 60−65% probability of obtaining the bound conformation (termed accuracy), while OPLS and MMFF only offer 45% and 50%, respectively. This trend continues when going up to 70% of the ensemble. At this level, OPLS, MMFF,

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

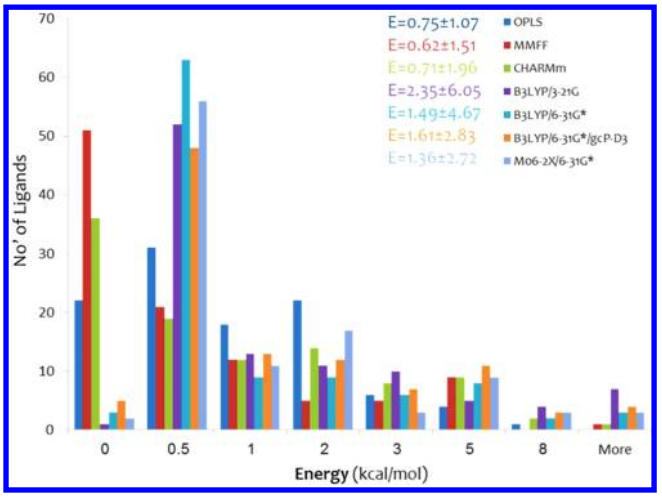


Figure 10. Energy differences for the relaxation of the bound conformation by the KPD method. (ΔE_3 in Figure 8.)

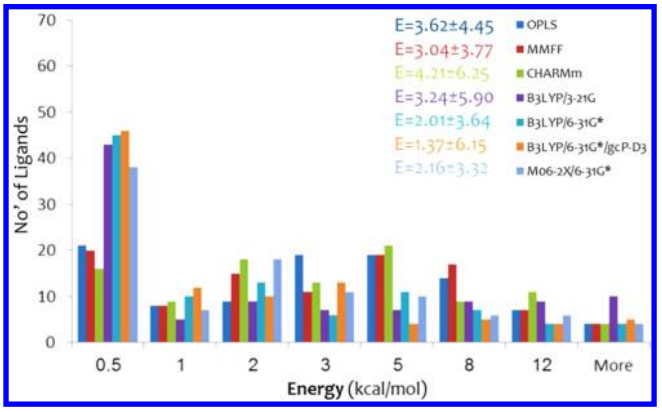


Figure 11. Energy differences, calculated as $\Delta E_4 = \Delta E_2 + \Delta E_3$, between bound and unbound ligand conformations, with ΔE_3 calculated using the T/B constraints.

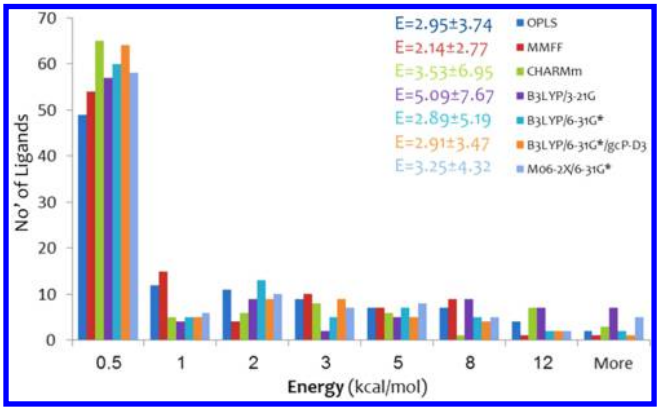


Figure 12. Energy differences, calculated as $\Delta E_4 = \Delta E_2 + \Delta E_3$, between bound and unbound ligand conformations, with ΔE_3 calculated using by the KPD method.

and B3LYP/3-21G offer 70% accuracy or less when other QM methods and CHARMm offer 80−84%. Table 8 summarizes these data and the energy cutoff at which they are observed.

4. DISCUSSION AND CONCLUSIONS

Developing a method for the focusing of unbound conformational ensembles on bound conformations requires, on one hand, reliable conformations of the unbound state and, on the other hand, reliable bound conformations. Bound conformations are typically obtained from protein−ligand crystal structures. To date, unbound conformations have been primarily generated using various force fields due to the large computational resources required for high level QM calculations on a large set of conformations.[2a,11,15] However, the reliability of conformational energies of molecules for which specific parameters have not been necessarily developed is questionable.

This work focuses on a set of relatively rigid (1−6 rotatable bonds) FDA-approved drugs whose X-ray bound-conformations are available in the PDB. Solution experiments may provide a better approximation to the true bound conformation of ligands,[2b] yet such conformations are unavailable in sufficient numbers to produce meaningful statistics. More flexible drugs were not considered for this paper, as their QM calculations require significantly larger computational resources, yet conclusions drawn from this work will facilitate future research on more flexible molecules. This work describes the largest collection of drug compounds ever subjected to QM-based conformational searches. The advantage of studying drug compounds is that any conclusion drawn may have relevance to other compounds of pharmaceutical interest. At the same time, however, this choice may flaw the data set by including low resolution structures. The list of crystal structures studied in this work along with their atomic resolutions is provided in Table S2 and demonstrates that for most of the structures (>80%) the resolution is higher than 2.5 Å. Moreover, as pointed out by Sitzmann et al., below the 1.3 Å threshold there is no correlation between important characteristics of bound conformations and the resolution of the crystal structure from which they were obtained.[2a]

This work uses RMSD from crystal structures as a metric to evaluate the difference between a computationally derived conformation and the X-ray structure. Such a metric is potentially problematic in cases where the ligand coordinates are misfitted into the experimental electron density.[2b] Using the COOT program,[35] we visually inspected the electron density maps available for our data set and found no cases where the deposited ligand coordinates largely deviate from the map. We then optimized the fitting of the coordinates with COOT and found an average RMSD of 0.22 Å between the deposited and the refined ligands. This implies that the ligands are well-fitted into the electron densities in the original, deposited coordinates (Table S3).

In this work we sought to derive reliable conformational ensembles for 100 FDA-approved drugs using seven energy functions (rather than assessing the performances of conformational search algorithms in terms of their ability to reproduce bioactive conformations, as has been reported by others). Assessing the reliability of computationally derived conformational ensembles is a difficult task further complicated by the lack of reliable experimental data for complete conformational spaces. Thus, we have limited our assessment to analyzing correlations between conformational energies ($E^i_{\mathrm{conf}} - E_{\mathrm{global\ minimum}}$ where E^i_{conf} is the energy of the i^{th} conformation and $E_{\mathrm{global\ minimum}}$ is the global energy of the conformational ensemble) obtained with the different methods (Figure S3 and Table S4). As a first observation we note the low correlation between the results obtained with different force fields versus the higher correlation between results obtained with different QM methods. This strongly suggests that the latter calculations

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

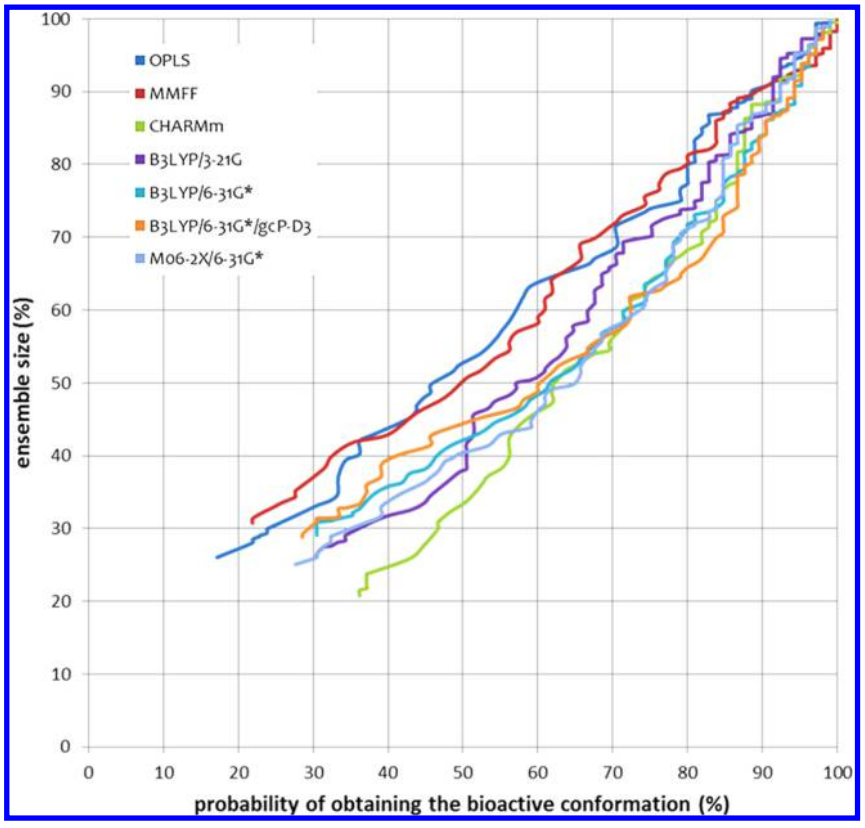


Figure 13. Graph shows the probability of obtaining the bound conforamtion vs the size of the conforamtional ensemble. Each point on this graph is defined from a different energy cutoff using ΔE_2.

Table 8. Energy Cutoffs (Top, kcal/mol) Defining Ensemble Size and Accuracy (Bottom) When Focusing Conformational Ensembles[a]

ensemble size	OPLS	MMFF	CHARMm	B3LYP/3-21G	B3LYP/6-31G*	B3LYP/6-31G*-gcP-D3	M06-2X/6-31G*
50%	1.6	0.8	1.8	2.0	1.6	1.5	2.1
	46%	50%	62%	57%	62%	60%	65%
60%	2.3	1.6	2.8	3.6	2.6	2.4	3.4
	57%	61%	71%	67%	74%	72%	74%
70%	3.0	2.5	4.0	6.0	3.6	3.6	4.5
	70%	67%	82%	75%	79%	85%	79%
80%	4.2	3.4	6.0	9.0	5.5	5.4	6.5
	80%	79%	87%	83%	88%	89%	85%
90%	7.0	5.7	10.1	14.6	9.4	8.9	9.4
	89%	90%	92%	91%	94%	94%	92%
95%	9.6	7.9	15.5	21.8	13.3	12.7	13.9
	95%	97%	96%	93%	95%	96%	94%
98%	15.1	11.7	28.6	30.7	19.2	17.5	20.3
	97%	99%	98%	98%	98%	98%	97%
99%	19.0	15.4	44.5	41.2	21.1	20.3	23.3
	97%	100%	99%	98%	99%	99%	98%

[a]Accuracy is defined as the probability of obtaining the bound conformation.

are more reliable. Similar observations are made for energy differences calculated between bound conformations approximated by closest energy minima and either global energy minima or Boltzmann-averaged energies (Tables 5 and 6, respectively). Next, the data in Figure S3 and Table S4 clearly suggest that none of the force field derived energy differences are correlated with the QM derived ones. Thus, using force field calculations to filter conformational ensembles prior to the more time-consuming QM calculations may result in loss of low energy conformations. In contrast, correlations were found between lower and higher level QM results, in particular those that use the same basis set (e.g., B3LYP/6-31G* and MO6-2X/6-31G*), thereby suggesting that lower level QM calculations can be used to prefilter conformational ensembles prior to their submission to higher level calculations. Taken together these data demonstrate that QM-based conformational searches are required for obtaining reliable unbound ensembles of drug-like compounds.

All methods considered in this work were able to generate a conformation within RMSDs of 0.5 and 1.0 Å from the bound

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

one for >60% and >90% of the ligands, respectively (Tables 1a and 1b). Overall the best results were obtained with "stand alone" CHARMm and with the ensembles reminimized with QM methods.

Having obtained considerably reliable conformational ensembles of the unbound state, the task at hand was to focus them on bioactive conformations. As noted in the Introduction, focusing mechanisms may be based on structural or energetic criteria. In contrast with previous reports, for the present set of compounds we found no consistent differences in structure between bound conformations and other conformations. Moreover, all attempts to combine different descriptors into a single bound conformation predictor in the form of a decision tree have failed (results not shown). It can thus be concluded that either bound conformations are not structurally different in a consistent manner from other conformations or if they are, these differences were masked by the inclusion of compounds targeting a variety of protein families with diverse binding site characteristics. In accord with the second hypothesis, Auer and Bajorath have found structural patterns distinguishing bound and modeled conformations for 18 sets of inhibitors each targeting a different protein.[36] Generalizing these findings will require a larger set of ligand−protein complexes coupled with different classification schemes of either ligands or proteins, the latter preferably according to their binding sites.

The previously observed correlation between structural parameters (e.g., ROG, SASA) and bioactivity may be due to the limited flexibility of the compounds considered in this work (1−6 rotational bonds). To test this hypothesis, we have performed CHARMm-based conformational searches for a set of 71 compounds with 7−19 rotatable bonds and compared the resulting conformational ensembles with the corresponding bioactive conformations. The results (Tables S5 and S6) demonstrate that also for this more flexible data set, structural parameters are indistinguishable between bioactive and other conformations. A more comprehensive analysis of flexible ligands is the subject of future research.

Having failed to identify structural criteria able to focus unbound conformational ensembles on bioactive conformations for the current set of ligands, we tried to identify energy criteria. We first examined the resemblance between global energy minima and bound conformations. Surprisingly, we found that 21−27% of the ligands have their global energy minima within an RMSD of 0.5 Å from their corresponding bioactive conformation. These numbers increase to 47−58% when considering a 1.0 Å RMSD, yet there is no clear correlation between RMSD values and conformational energies (data not shown). As before, results obtained with the workflow were marginally better than those obtained with individual force fields. Best results at the lower threshold were obtained with QM methods (Tables 4a and 4b), while at the higher threshold best results were obtained with OPLS. Not surprisingly, a closer examination of the data revealed a correlation between the number of rotational bonds and the fraction of ligands having their global energy minima in close proximity to the bioactive conformation (R^2 = 0.72−0.89 at the 0.5 Å level for the different methods). Thus, we expect these numbers to decrease for larger and more flexible ligands. Taken together, these data suggest that, as expected, global energy minima of the unbound ensembles are poor approximations to the bound conformation.

Next we sought to define an energy threshold within which it is likely to locate bioactive conformations using several approximations to the unbound and bound states. The T/B method and the KPD method gave similar results both in terms of energy (ΔE_3) and RMSD from the crystal structure (Figures 8−10 and Table 7 and 8). The advantage of the KPD method is that it does not require the B-factors.

Overall, the present work supports lower-bound energy differences between solution ensembles and bioactive conformations in the range of 2−5 kcal/mol depending on the computational model and the nature of the relaxed conformation. These numbers are overall lower than previous estimates, but a direct comparison cannot be made due to differences in both the data set and the exact nature of the relaxed bound conformations.

It is important to note that irrespective of the nature of the approximation, ligands with high (>10 kcal/mol) energy differences still persisted. Similar to the work of Sitzmann et al., these high energy ligands did not necessarily come from low resolution crystal structures. Thus, the potential limitations of the crystallographic data discussed by these authors may also be relevant for the current data set.

Finally, we propose a partial focusing of unbound conformational ensembles on bound conformations using the energy criterion. The data in Figure 11 indicate a higher tendency of bound conformations to reside within low energy regions of the unbound conformational ensemble. Thus, the lowest 50% of conformational ensembles calculated at the M06-2X/6-31G* level contain 65% of the bound conformations (enrichment factor of 1.3). Overall better enrichment factors were obtained with QM methods than with force field methods. However, due to the high similarity in enrichment factors among the different methods, we do not anticipate a major improvement upon moving to even higher level QM calculations. Clearly, focusing unbound conformational ensembles on bound conformations is still a largely open question.

In conclusion, this work presents a detailed force field-based and QM-based computational study of the bound conformations and unbound conformational ensembles of a large set of relatively rigid FDA-approved drugs. The work was undertaken with the hope of identifying suitable parameters for focusing the unbound ensembles on bound conformations. The main conclusion in this respect is that structural parameters cannot differentiate between unbound and bound conformations (at least when ligands are treated together and not classified into target-related groups). Yet, some focusing can be obtained using energy criteria. Overall we note that the performance of QM methods is marginally better than force fields in terms of their ability to reproduce bound conformations either in general or as lowest energy minima (in particular when using stringent RMSD cutoffs) and in terms of the enrichment factors resulting from the focusing mechanism.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.5b00259.

Figures S1−S3 and Tables S1−S6 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hsenderowitz@gmail.com.

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167

Notes

The authors declare no competing financial interest.

REFERENCES

(1) 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* (20), 5880−4.

(2) (a) Sitzmann, M.; Weidlich, I. E.; Filippov, I. V.; Liao, C.; Peach, M. L.; Ihlenfeldt, W. D.; Karki, R. G.; Borodina, Y. V.; Cachau, R. E.; Nicklaus, M. C. PDB ligand conformational energies calculated quantum-mechanically. *J. Chem. Inf. Model.* **2012**, *52* (3), 739−56. (b) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 179−90.

(3) Weng, Z. F.; Motherwell, W. D.; Allen, F. H.; Cole, J. M. Conformational variability of molecules in different crystal environments: a database study. *Acta Crystallogr., Sect. B: Struct. Sci.* **2008**, *64* (3), 348−362.

(4) Graves, A. P.; Shivakumar, D. M.; Boyce, S. E.; Jacobson, M. P.; Case, D. A.; Shoichet, B. K. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J. Mol. Biol.* **2008**, *377* (3), 914−34.

(5) (a) Boehr, D.; Nussinov, R.; Wright, P. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5* (11), 789−796. (b) Csermely, P.; Palotai, R.; Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **2010**, *35* (10), 539−546.

(6) (a) Lu, J.-J.; Pan, W.; Hu, Y.-J.; Wang, Y.-T. Multi-target drugs: the trend of drug research and development. *PLoS One* **2012**, *7* (6), e40262. (b) Imming, P.; Sinning, C.; Meyer, A. Drugs, their targets and the nature and number of drug targets. *Nat. Rev. Drug Discovery* **2006**, *5* (10), 821−834.

(7) Musafia, B.; Senderowitz, H. Biasing conformational ensembles towards bioactive-like conformers for ligand-based drug design. *Expert Opin. Drug Discovery* **2010**, *5* (10), 943−59.

(8) Green, N. Thermodynamics of the binding of biotin and some analogues by avidin. *Biochem. J.* **1966**, *101* (3), 774−780.

(9) Musafia, B.; Senderowitz, H. Bioactive conformational biasing: a new method for focusing conformational ensembles on bioactive-like conformers. *J. Chem. Inf. Model.* **2009**, *49* (11), 2469−80.

(10) Diller, D. J.; Merz, K. M., Jr. Can we separate active from inactive conformations? *J. Comput.-Aided Mol. Des.* **2002**, *16* (2), 105−12.

(11) 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* (10), 2499−510.

(12) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3* (4), 411−28.

(13) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (18), 9997−10002.

(14) (a) Wlodek, S.; Skillman, A. G.; Nicholls, A. Automated ligand placement and refinement with a combined force field and shape potential. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2006**, *62* (7), 741−749. (b) Foloppe, N.; Chen, I. J. Conformational sampling and energetics of drug-like molecules. *Curr. Med. Chem.* **2009**, *16* (26), 3381−413.

(15) Butler, K. T.; Luque, F. J.; Barril, X. Toward accurate relative energy predictions of the bioactive conformation of drugs. *J. Comput. Chem.* **2009**, *30* (4), 601−10.

(16) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225−11236.

(17) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17* (5−6), 490−519.

(18) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4* (2), 187−217.

(19) Kinnings, S. L.; Xie, L.; Fung, K. H.; Jackson, R. M.; Bourne, P. E. The Mycobacterium tuberculosis drugome and its polypharmacological implications. *PLoS Comput. Biol.* **2010**, *6* (11), e1000976.

(20) (a) Chang, G.; Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111* (12), 4379−4386. (b) Kolossváry, I.; Guida, W. C. Low Mode Search. An Efficient, Automated Computational Method for Conformational Analysis: Application to Cyclic and Acyclic Alkanes and Cyclic Peptides. *J. Am. Chem. Soc.* **1996**, *118* (21), 5011−5019.

(21) Accelrys Software Inc., *Discovery Studio Modeling Environment, Release 3.1*; Accelrys Software Inc.: San Diego, CA.

(22) (a) 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* (16), 6127−6129. (b) Qiu, D.; Shenkin, P. S.; Hollinger, F. P.; Still, W. C. The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii. *J. Phys. Chem. A* **1997**, *101* (16), 3005−3014.

(23) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., 3rd New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. *J. Comput. Chem.* **2003**, *24* (11), 1348−56.

(24) Polak, E.; Ribiere, G. Note sur la convergence de méthodes de directions conjuguées. *Revenue Francaise Informat. Recherche Operationelle, Serie Rouge* **1969**, *3* (R1), 35−43.

(25) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B. Gaussian 09, Revision B.01. In *Gaussian 09, Revision A.02*; Gaussian, Inc.: Wallingford, CT, 2009.

(26) Tomasi, J.; Mennucci, B.; Cances, E. The IEF version of the PCM solvation method: an overview of a new method addressed to study molecular solutes at the QM ab initio level. *J. Mol. Struct.: THEOCHEM* **1999**, *464* (1−3), 211−226.

(27) Kruse, H.; Goerigk, L.; Grimme, S. Why the standard B3LYP/6-31G* model chemistry should not be used in DFT calculations of molecular thermochemistry: understanding and correcting the problem. *J. Org. Chem.* **2012**, *77* (23), 10824−34.

(28) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104.

(29) Zhao, Q. P.; Hautamaki, V.; Franti, P. Knee Point Detection in BIC for Detecting the Number of Clusters. *Lect Notes Comput. Sc* **2008**, *5259*, 664−673.

(30) Salvador, S.; Chan, P. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *Proc. Int. C Tools Art* **2004**, 576−584.

(31) Mitchell, T. M. *Machine Learning*; McGraw-Hill Education: 1997.

(32) Stanton, D.; Jurs, P. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Anal. Chem.* **1990**, *62* (21), 2323−2329.

(33) Rohrbaugh, R. H.; Jurs, P. C. Descriptions of Molecular Shape Applied in Studies of Structure Activity and Structure/Property Relationships. *Anal. Chim. Acta* **1987**, *199*, 99−109.

(34) 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* (4), 1428−37.

(35) Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Features and development of Coot. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2010**, *66* (4), 486−501.

(36) Auer, J.; Bajorath, J. Distinguishing between bioactive and modeled compound conformations through mining of emerging chemical patterns. *J. Chem. Inf. Model.* **2008**, *48* (9), 1747−53.

DOI: 10.1021/acs.jcim.5b00259
J. Chem. Inf. Model. 2015, 55, 2154−2167