EXPLORING THE CONFORMATIONAL LANDSCAPE OF BIOACTIVE SMALL MOLECULES

Sanja Zivanovic[1], Francesco Colizzi[1], David Moreno[1], Adam Hospital[1] Robert Soliva[2] and

Modesto Orozco[1,3]*

By using a combination of classical Hamiltonian Replica Exchange with high-level quantum mechanical calculations on more than one hundred drug-like molecules we explored here the energy cost associated with binding of drug-like molecules to target macromolecules. We found that, in general, the drug-like molecules present bound to proteins in the Protein Data Bank (PDB) can access easily the bioactive conformation and in fact for 73% of the studied molecules the "bioactive" conformation is within $3k_bT$ from the most stable conformation in solution as determined by DFT/SCRF calculations. Cases with large differences between the most stable and the "bioactive" conformations appear in ligands recognized by ionic contacts, or very large structures establishing many favorable interactions with the protein. There are also a few cases where we observed a non-negligible uncertainty related to the experimental structure deposited in PDB. Remarkably, the rough automatic force-field used here provides reasonable estimates of the conformational ensemble of drugs in solution. The outlined protocol can be used to better estimate the cost of adopting the bioactive conformation.

---

[1] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology (BIST), Baldiri Reixac, 10, 08028 Barcelona, Spain.

[2] Nostrum Biodiscovery, Nexus II Building. Barcelona

[3] Departament de Bioquímica i Biomedicina. Facultat de Biologia. Universistat de Barcelona.

* Correspondence to Prof. M.Orozco: modesto.orozco@irbbarcelona.org

INTRODUCTION

Binding of a ligand to a macromolecule is a complex procedure involving not only direct ligand-macromolecule interactions, but also changes in solvation and internal free energy of the interacting partners. Based on the conformational changes required for binding three different recognition modes have been suggested [1–3] i) Fisher's lock and key, ii) conformational selection and iii) induced fit. In the first model, the unbound and bound conformations of both ligand and macromolecule are the same and no significant change in internal energy is associated with binding. In the other two models, binding implies rearrangements in the ligand, the protein or both. Within the conformational selection paradigm, the free energy cost associated with such transitions is small and the bioactive conformation is sampled spontaneously within the "unbound" conformational ensemble. On the contrary, according to the induced fit mechanism, the "bioactive" conformation is rarely populated in the unbound state and it appears just as a result of the ligand-macromolecule interactions. The design of efficient binders represents always a compromise between rigidity, which reduces the entropy cost associated to binding and flexibility that increases the possibility of favorable ligand-protein interactions. In general [4,5] bioactive small molecules (such as pharmaceutical drugs) have a limited number of rotatable bonds, suggesting a similarity between the unbound and bioactive conformations, but a non-negligible number of highly potent binders show a large number of rotatable bonds, raising the question of how different is the unbound conformational ensemble with respect to the bioactive conformation. Several authors suggest that adopting bioactive conformation implies a large energy penalty for the drug, while others support the idea that the bioactive conformation is sampled spontaneously in the unbound state as powerful binders are pre-organized to facilitate binding [6–11]. With all these views in mind, we present here an automatized multilevel strategy, which combines fast exploration of the conformational space by means of classical Hamiltonian Replica Exchange (HREX) molecular dynamics (MD) calculations with high-level quantum mechanical (QM) calculations in aqueous solution. The approach is validated on a large set (115) of diverse bioactive small molecules (see Methods). In more than 97% of the cases the bioactive conformation (that one in which the drug binds the protein) is sampled in the HREX simulations (there is at least 1 of the collected snapshots with RMSd< 1 Å from the bioactive conformation). When the conformational space is reduced to a limited number of clusters (from 10 to 40) the bioactive conformation is at less than 2 Å of one cluster in more than 96% of the cases; if the cutoff is reduced to 1.2 Å, the bioactive conformation is close to a cluster in 63% (10 clusters) and 79% (40 clusters) of the cases. In other words, the strategy used here seems to sample quite exhaustively the ligand conformational space.

Our data suggest that the bioactive conformations are typically not the most populated (MD) or the lowest energy (QM) states in aqueous solution, but in around 70% of the cases the bioactive conformation is within 3 k_bT from the most stable QM conformer in solution (60% of the cases difference below 2 k_bT). This indicates that Fisher's lock & key and conformational selection models dominate binding paradigm in our dataset of drug-protein small. In about 10% of the cases, we observed large (> 5 kbT) energy differences between the most populated conformer in solution and

the bioactive one. Cases with large differences between the most stable and the "bioactive" conformations correspond to in general to ligands recognized by ionic contacts and/or establishing a wide network of interactions with the protein. Remarkably, for a few cases where the energy penalty associated with the adoption of the bioactive conformation is especially large, we observed some potential uncertainties related to the experimental structure deposited in PDB, which recommend a critical analysis of PDB-reported conformations.

METHODS

Data set. We have initially selected a data set of a total 123 pharmaceutically relevant ligands whose structures, complexed with at least one protein target, are known experimentally. Selected molecules include a subset (resolution ≤ 2.5 Å) of Perola's dataset (80 ligands [4]), 24 dual binders [12], 14 pharmaceutical compounds with especial complexity (6 of them GPCR ligands) [13] and 5 bioactive macrocycles [14–18]. Analysis of the set reveals that at least in 8 of these cases the bioactive conformation reported in PDB might be partially incorrect (see Suppl. Figure 1 and below). Thus, to avoid bias in the results, these 8 structures were removed leading to a final set of 115 compounds (see below). The set corresponds to a wide range of chemical structures (see Figure 1 for details of database composition), molecular weight (from 200 to 700 g/mol), flexibility (from 1 to 12 rotatable bonds), hydrogen bonding capabilities (from 0 to 12 H-bond donors and/or acceptors) and ionic states (at pH 7.4 46% of the ligands are expected to be neutral, 29 % anions and 25 % cations).

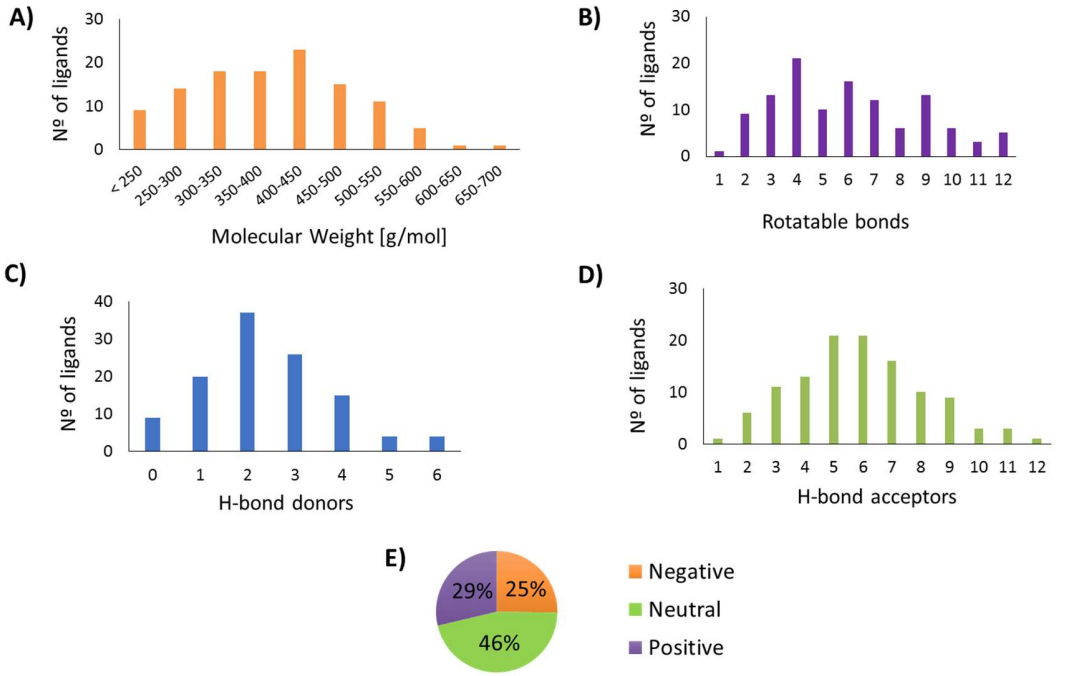


Figure 1 Details of database composition. **A)** Molecular weight of ligands **B)** Distribution of rotatable bonds **C)** and **D)** Hydrogen bonding capabilities **E)** Charge distribution

Ligand preparation. We extract the ligands (with resolution ≤ 2.5 Å) from the corresponding PDB files and determine the placement of the hydrogens at physiological pH using Open Babel [19] and Marvin Sketch [20] checking the consistency of their assignments in view of the drug-protein recognition mode (just a few cases with errors in the annotation were found). Antechamber tool [21] and Acpype [22] open source codes were used to assign rough Generalized Amber Force Field (GAFF) parameters [23] and to define topologies to the ligands. Atomic charges of the ligands were determined at AM1-BCC level [24]. Each ligand was solvated in a truncated octahedron box of TIP3P water molecules [25] (periodic boxes were selected to guarantee a minimum distance greater than 8 Å from the ligand to the closest face). The systems were neutralized by adding suitable ions, minimized, thermalized and equilibrated for 1 ns (at constant pressure and temperature (1 atm, 298 K).

Enhanced sampling simulations. After several tests including unbiased MD, annealing and Temperature replica exchange MD, we decided to accelerate sampling by using Hamiltonian Replica EXchange (HREX) [26,27] using GROMACS 4.6.7 [28] patched with PLUMED 2.1 [29,30] and the Gromacs HREX implementation [31]. We used 16 replicas and scaling all the atoms (charge; epsilon Lennard-Jones parameter; proper dihedral) of the solute (Replica Exchange with Solute Tempering, REST2) [32] with values of λ ranging from 1 to 0.59 following a geometric distribution (1, 0.966086, 0.933324, 0.901672, 0.871095, 0.841554, 0.813015, 0.785442, 0.758806, 0.733075, 0.708214, 0.684196, 0.660993, 0.638578, 0.616921, 0.596). Production runs were evolved at 298 K in the NVT ensemble with the velocity rescaling thermostat. The acceptance rate ranged from 60% to 90%. Exchanges were attempted every 500 steps. Each replica ran for 10 ns with accumulative sampling time of 160ns per compound, taking data every 1 ps. Particle Mesh Ewald (PME) [33] and periodic boundary conditions were used to represent long-range electrostatic effects. All bonds linking hydrogens were frozen using SHAKE [34], which allowed us the use of 2 fs time scale for integration of Newton equations of motion. By default the statistics was extracted from the non-scaled replica (λ=1). Simulations took typically between 6 and 12 hours in a small Intel Xeon E5-2670 @ 2.60GHz cluster using 1 processor per replica. Computational times are compatible with the high-throughput timescale required for drug-design projects. The workflow is shown in Figure 2, while further details on the pipeline will be discussed later.

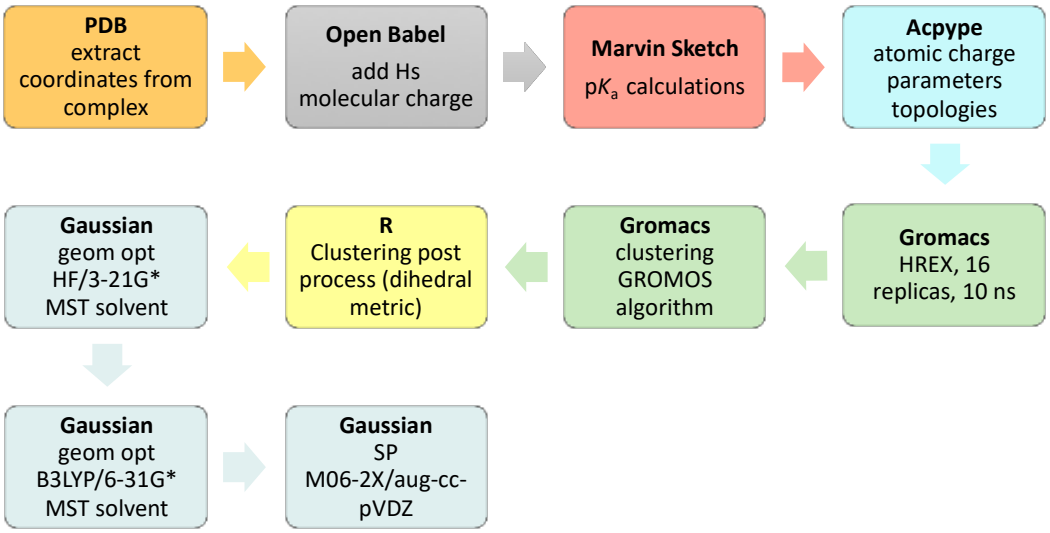


Figure 2 Scheme of the automatic pipeline

Clustering. The ensembles collected from the 10^4 snapshot trajectories were analyzed to define the most populated clusters. After testing different clustering approaches, we follow Daura's algorithm [35] as implemented in GROMOS. Accordingly, internal $RMSd_i$ was used as metrics, counting the number of neighbors (within a given cut-off) for each structure, and taking that with the largest number of neighbors as a center of cluster. Cluster representative was determined as the structure with the lowest RMSD among a family. Clusters were mutually exclusive and were annotated by using an iterative clustering approach, making sure that they represent at least 95% of the total sampled space. Application of this general clustering strategy, which would be very efficient for peptides, leads to some problems for our small drug-like molecule, where regional symmetry can exist. To avoid these problems standard Daura's clustering annotation is refined by: i) using a symmetry-corrected dihedral metrics [36] [37] (dAB; dihMetrics, eq.1) ii) re-group clusters which are identical based on the new metrics, iii) re-annotate snapshots to clusters based on the dihedral metrics measured with respect to the centroid. Detailed explanation of the method and formula are provided in Suppl. Info. All the post-process and refinement methodology was implemented with R scripts [38]. Analysis of several examples showed us that this procedure defines clusters with a robust well-defined population and containing snapshots with similar geometries (see detailed analysis below).

$$d_{AB}^2 = \frac{1}{n}\sum_{i=1}^{n} 2\left(1 - \cos\left(S_i \varphi_{i_{AB}}\right)\right) \quad (1)$$

where n is the number of dihedrals that are used to compute the distance and $\varphi_{i_{AB}} = \varphi_{i_A} - \varphi_{i_B}$, where S_i is the symmetry number of torsion i, φ_{i_A} and φ_{i_B} are the values of the dihedral i in the structure A and B, respectively (see Suppl. Information for additional details).

High level calculations. The geometries of cluster representatives were used as a starting point for quantum mechanical calculations. After preliminary tests we decided to perform a two-steps geometry optimization: first, the geometries of the cluster representatives for each ligand were optimized at IEF-MST/HF/3-21G level, using the output as starting point for further IEF-MST/B3LYP/6-31G(d) geometry refinement as implemented in Gaussian [39,40]. IEF-MST calculations were done considering water as solvent and the associated cavity and van der Waals parameters [41]. The final optimized geometries were used for additional single-point calculations at the M06-2X/aug-cc-pVDZ level [42] taking in this case the solvation contribution from the B3LYP/6-31G(d) calculation [43–45]. To explore potential errors in the DFT treatment, additional calculations were performed at the MP2/aug-cc-pVDZ level [46,47] for a selected set of 14 molecules.

Defining the bioactive conformations. The conformations directly extracted from PDB have in several cases major structural distortions (see examples in Suppl. Figure S1), forcing us to relax the geometry to obtain realistic bioactive conformations. Our first approach was to fix the dihedral angles at the X-ray values, pre-optimize the geometry by AM1/MST [48,49] calculations and perform further refinement by IEF-MST/B3LYP/6-31G(d) calculations [40,41]. This procedure leads to partially relaxed geometries, very close to the X-Ray ones, but with less structural artefacts. Unfortunately, not all obvious distortions were corrected, as some torsional angles led to artefactual internal geometries (see examples in Suppl. Figure S1). We decided then to relax all the degrees of freedom, assuming that the PDB structure is close to a local minimum. For 86% of the ligands the bioactive-relaxed geometry is quite close to the X-ray structure (RMSd< 1 Å; see Suppl. Figure S2); the remaining 18 cases were analyzed in detail. For 4 of these 18 ligands the X-ray conformation is supported by electron density, and we adopt then the partially-relaxed (dihedral fixed) conformation for the remaining analysis; in 6 cases, human inspection shows that relaxed and X-Ray structures are not so different in terms of the bound moiety (i.e. most differences affect solvent-exposed moieties) and we maintain then the relaxed geometry as bioactive conformation. Finally, in 8 cases, X-Ray conformation leads to unlikely geometries, which are not supported by electron densities; to avoid biases these cases were removed for future analysis. As shown in Figure S2, the final dataset contains 115 molecules, the bioactive conformations considered here deviate in general around 0.5 Å from the X-ray structure in the PDB.

The cost of reaching the bioactive conformation. There are many ways of defining the cost of reaching the bioactive conformation. Taken the MD ensembles and assuming that the "bioactive conformation" is reached when the "bioactive cluster" is sampled, the free energy cost associated to moving from the aqueous ensemble to the "bioactive conformation" is given by eq. 2:

$$\Delta G_{dist} = -k_b T ln \frac{P_b}{P_{tot}} \quad (2)$$

where T is the temperature, k_b is the Boltzmann constant and P_b is the population of the "bioactive cluster" referred to the total population of the ensemble, P_{tot}. An additional estimate of the free energy penalty associated to reaching the bioactive state (defined as the "bioactive cluster") is defined from the difference in stability between the most stable cluster and the bioactive one, see eq. 3:

$$\Delta G_{strain} = -k_b T ln \frac{P_b}{P_0} \quad (3)$$

where P_0 is the population of the most stable cluster, i.e. the most populated one in the MD ensemble.

Note that the annotation of the "bioactive cluster" is arbitrary depending on the threshold radii used for the assignment of the bioactive conformation (see below), i.e. on how close a conformation should be to that found in the crystal to be considered bioactive. Accordingly, for threshold radii different to 0 the estimates obtained using eqs. 3 and 4 represent most likely lower boundary values of the (free) energy cost required to adopt the bioactive conformation.

At the quantum level we can obtain a QM/SCRF estimate of the strain (free) energy by using (eq. 4):

$$\Delta G_{strain} = G_b - G_0 \quad (4)$$

where index 0 refers to the most stable conformer in solution (as defined by QM/SCRF calculations) and index b refers to the bioactive conformation (obtained as described above). A discrete quantum alternative to eq. 2 can be formulated assuming the width associated to all clusters is equivalent (i.e. assuming entropy within a cluster is the same) by using eq. 5:

$$\Delta G_{dist} = G_{strain} + k_b T \, ln \sum_k e^{-\frac{G_k - G_0}{k_b T}} \quad (5)$$

where the sum extends for all the k clusters (k≠b) defining the conformational space of the unbound ligand. Note that eqs. 4 and 5 are exact for an infinite number of clusters, but in practice implies an upper boundary limit for the (free) energy cost associated to achieving the bioactive conformation as most likely the bioactive cluster is typically narrower than the others.

Data accessibility. All data is available at the Bioactive Conformational Ensemble (BCE) server http://mmb.irbbarcelona.org/BCE/ (see companion paper ct-2020-00305q). Trajectories were stored for further analysis using MD database recommendations [50].

RESULTS AND DISCUSSION

Robustness of the sampling protocol. After several tests using different enhanced sampling approaches, we chose HREX as sampling method for its efficiency [26,51] and for the easiness to implement it into a general automated workflow. We tested the convergence of the approach to the extension of the simulation trajectory (16 replicas) for 6 representative drugs showing different conformational complexity. Results summarized in Figure 3 demonstrate that extension of the simulation time leads to only small changes in the population of the clusters. This, and the similarity between the cluster representatives obtained from different simulation times (data not shown) suggest the HREX procedure used here in conjunction with 10 ns x replica window is accurate enough to explore conformational space.

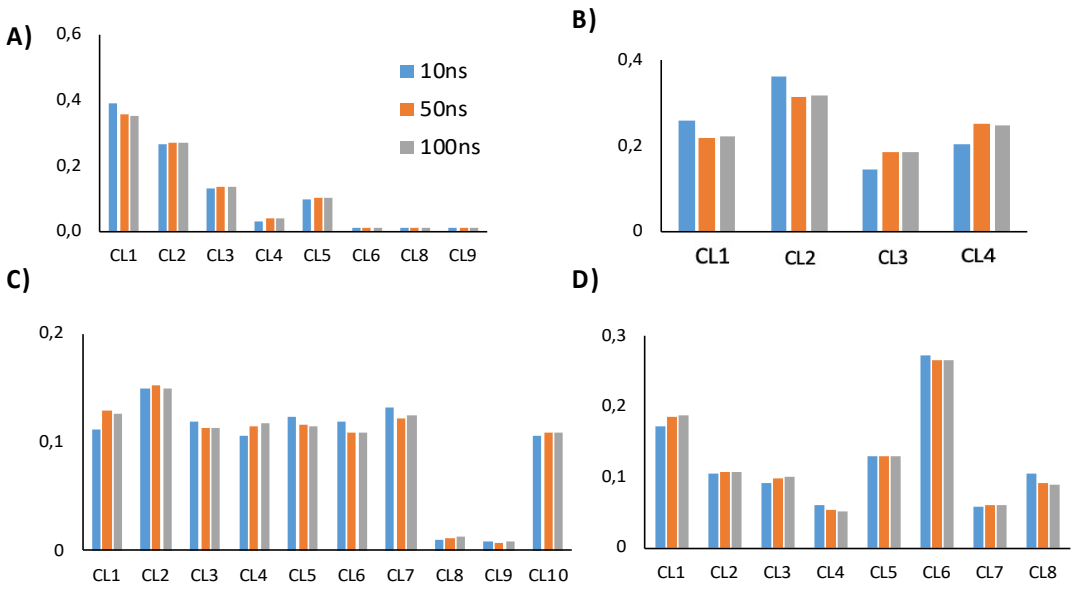


Figure 3 Population of clusters as a function of the simulation length: 10 ns (blue), 50 ns (orange), and 100 ns (gray) is shown. The analysis shows that cluster population is robust with respect to the simulation length and that 10 ns are enough to observe converged cluster. Analysis are performed for (A) the pdb ligand 1afq-0FG with 9 rotatable bonds; (B) 2pix-FLF with 3 rotatable bonds, (C) 1jsv-U55 with 3 rotatable bonds and D) 1qhi-BPG with 6 rotatable bonds. See Suppl. Figures S3-S4 for additional convergence tests.

Analysis of the entire dataset shows that the bioactive conformation is sampled in the HREX simulation (there is at least 1 of the collected snapshots with RMSd < 1 Å from the bioactive conformation; see Figure 4) in more than 97% of the cases. This indicates that the selected strategy (10 ns x HREX replica) provides, in general, an extensive enough sampling of the conformational

space of drugs in aqueous solution, and extension of the simulation times and replica numbers seems justified only for drugs containing many rotatable bonds and potential interacting groups.

Robustness of the clustering protocol. As we cannot re-compute all the sampled conformations at the QM level, clustering is crucial to derive representative structures from the HREX samplings. We applied an iterative clustering algorithm with a maximum of 10 clusters representing at least 95% of the snapshots. Cluster annotation was performed (see Methods) correcting symmetry-related biases. In general, we found that even with a reduced 10-cluster representation, there is one cluster closer than 2 Å to the bioactive conformation in more than 96% of the cases (Figure 4). Even with a very strict cutoff (1.2 Å) the chances to have bioactive conformation close to a cluster are large (from 63% to 79% depending on the number of clusters selected). In summary, the clustering approach seems robust enough to simplify the ligand conformational ensemble.

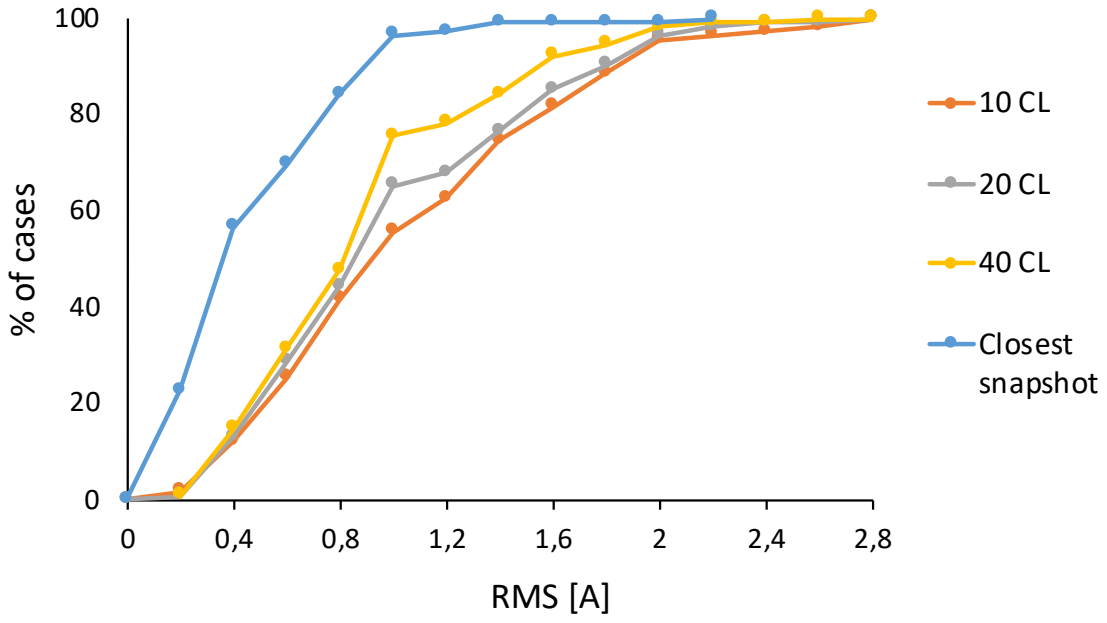


Figure 4 Robustness of the sampling and clustering protocol. The analysis shows that with HREX sampling, there is at least 1 of the collected snapshots with RMSd < 1 Å from the bioactive conformation (blue color). Chances to have bioactive conformation close to a cluster are calculated for different size of ensembles: maximum 10 cluster representatives (orange), 20 (grey) and 40 (yellow) cluster representatives.

Convergence of the QM results. Any high-throughput QM technique requires the use of a medium or low-level QM calculation, whose accuracy needs to be validated against experiments or higher-level calculations. In this paper we used two DFT levels for the entire set of 115 molecules: B3LYP/6-

31G(d), and M06-2X/aug-cc-pVDZ (see *Methods*) and for a selected set of 14 diverse molecules, we repeated calculations at a high correlated level (MP2/aug-cc-pVDZ), always using the same solvation model (MST; see *Methods*). Taken the 14 molecules together, (i.e. more than 120 conformational energies compared), the average energy RMS deviations between the MP2 and the DFT profiles are below 2 k_bT (B3LYP) or 1 k_bT (M06-2X) (see Figure 5). Differences are greater than 3 k_bT only for three molecules at the B3LYP/6-31G(d) level, and none if M06-2X/aug-cc-pVDZ are used. In summary, DFT methods provide results that are reasonably close to the reference MP2 ones, even when lower level B3LYP/6-31G(d) calculations are considered.

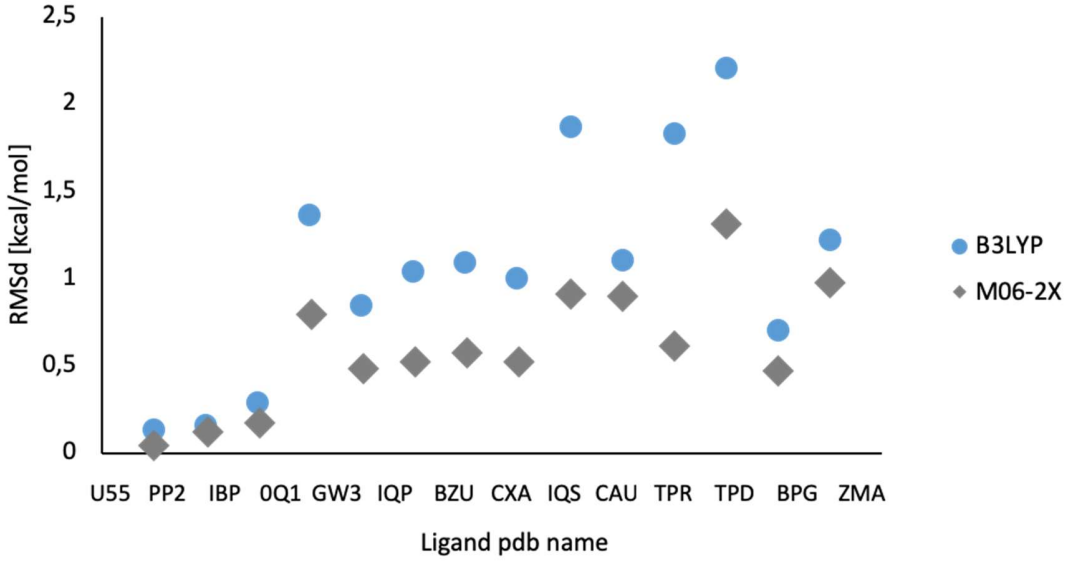


Figure 5 Convergence of the QM results. The average energy RMS deviations between the MP2 and the DFT profiles B3LYP/6-31G(d) (blue) and M06-2X/aug-cc-pVDZ (grey).

Divergence between MM and QM cluster representatives. As described in *Methods,* cluster representatives were selected and re-optimized at the QM level (B3LYP/6-31G(d)) with MST description of the solvent [41]. This optimization leads to new geometries which are in general between 0.2 and 0.8 Å from the MM cluster representative (see Figure 6), with no cluster shift detected during the process. In summary, cluster representatives are close to a local minimum in the SCRF/QM space and accordingly the sampling obtained from Hamiltonian replica exchange can be safely used as seeding for SCRF/QM geometry optimization.

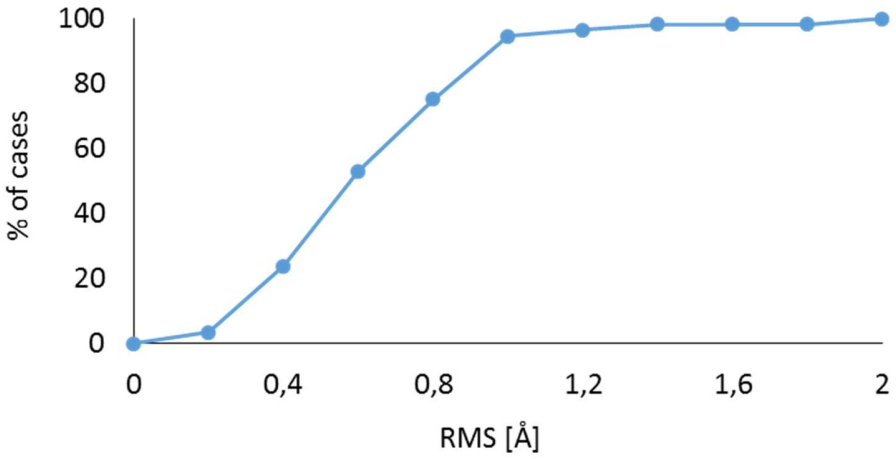
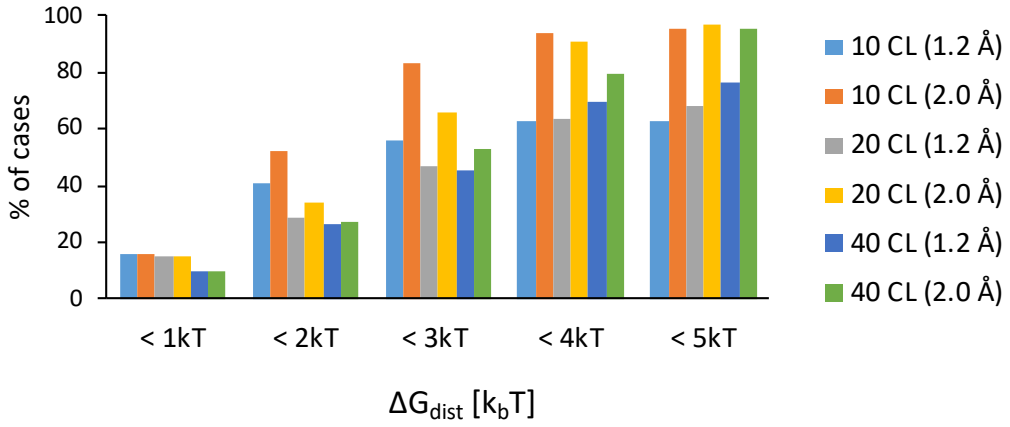


Figure 6 Divergence between MM and QM cluster representatives. RMS [Å] between MM and QM optimized cluster showing there is no cluster shift detected during the SCRF/QM minimization.

Population of the bioactive conformation. As noted above, for more than 96% of molecules we can annotate the bioactive conformation to one of the 10 representative clusters (cutoff 2 Å), but only in ~20% of the cases the "bioactive cluster" is the most populated one. Free energy penalties ΔG_{dist} and ΔG_{strain} (eqs. 1 and 2) depend, obviously, on the threshold used to define the "bioactive cluster" and on the number of clusters used to summarize drug conformational space (see Figure 7). For the default choice of 10 clusters and a cut-off of 2 Å ΔG_{dist} values below 4 k_bT and ΔG_{strain} below 3 k_bT are obtained in more than 90% of the cases (Figure 7). All these findings suggest that reaching the bioactive-like conformation does not require a huge (free) energy investment.



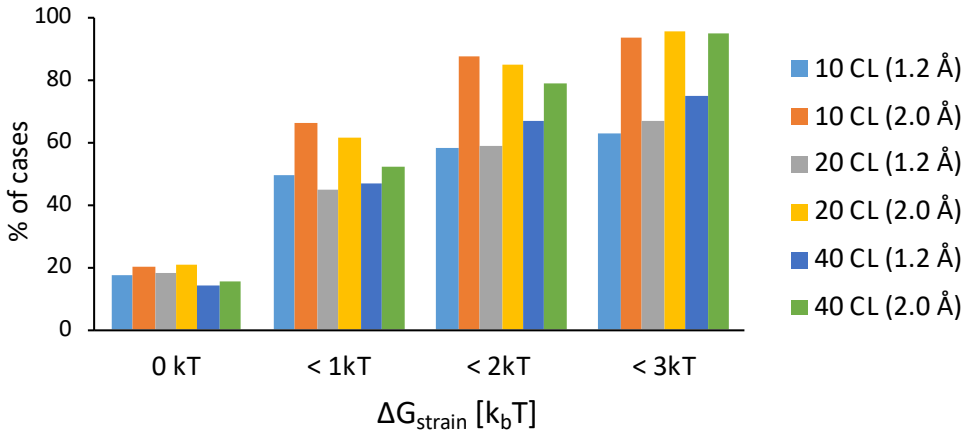
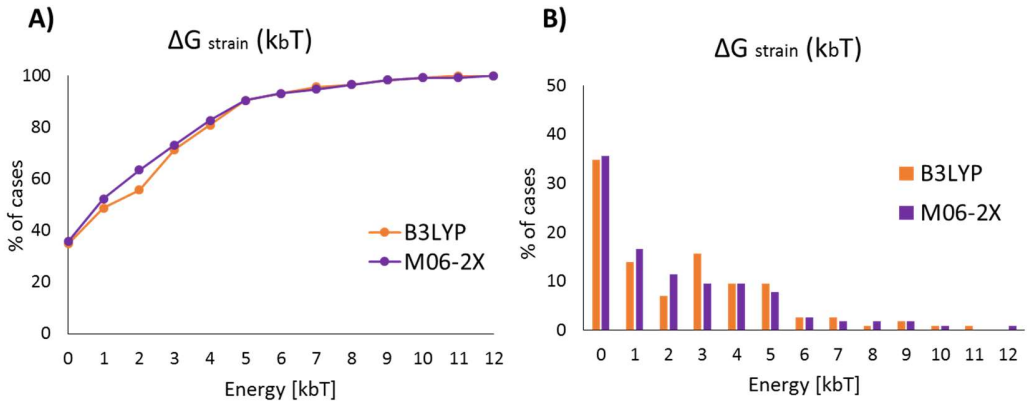


Figure 7 Free energy penalties ΔG_{dist} and ΔG_{strain} with different thresholds used to define the "bioactive cluster" and on the number of clusters used to summarize drug conformational space. **(light blue)** 10 clusters and a cut-off of 1.2 Å ; **(orange)** 10 clusters and a cut-off of 2.0 Å; **(grey)** 20 clusters and a cut-off of 1.2 Å; **(yellow)** 20 clusters and a cut-off of 2.0 Å; **(dark blue)** 40 clusters and a cut-off of 1.2 Å; **(green)** 40 clusters and a cut-off of 2.0 Å

To confirm the small magnitude of the free energy cost associated to the adoption of the bioactive conformation we analyze QM/SCRF estimates of "distortion" and "strain" (free) energies as shown in eqs. 3 and 4. Very interestingly (see Figure 8 A and B), the relaxed bioactive structure (see Methods) is the most stable conformer (ΔG_{strain}=0) in close to 40% of the studied cases. Around 60% of the studied ligands show strain energies below 2 k_bT, and only 10% of the studied ligands show strain energies above 5k_bT. These numbers are quite robust to the DFT functional and basis set used (Figure 8). See Suppl. Figures S5-S7 for examples.



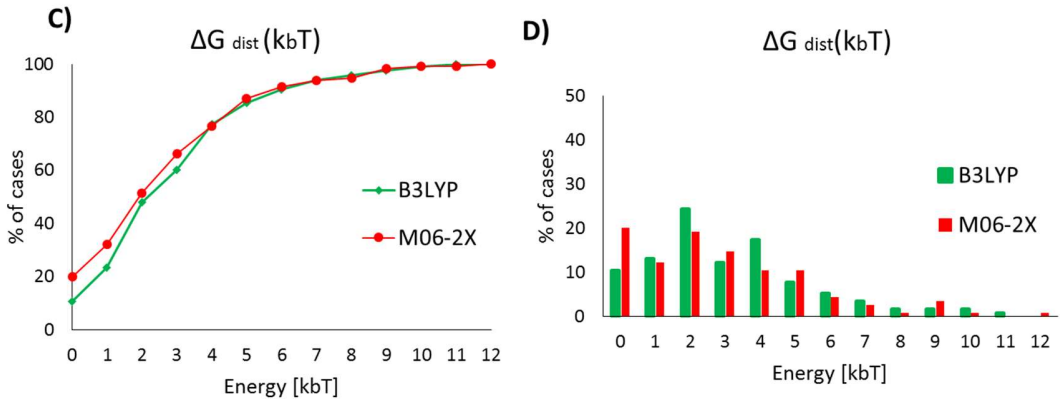


Figure 8 Top: QM/SCRF estimation of ΔG_{strain} at B3LYP/6-31G(d) (orange) and M06-2X/aug-cc-pVDZ level (purple). Figure (A) represents the cumulative plot while (B) is histogram. Bottom: QM/SCRF estimation of ΔG_{dist} at B3LYP/6-31G(d) (green) and M06-2X/aug-cc-pVDZ level (red). Figure (C) represents the cumulative plot while (D) is histogram.

Small changes in the analysis occur if distortions (eq. 4) instead of strain (eq. 3) free energies are considered (Figure 8 C and D). In summary, after analyzing a large dataset of drug-like molecules we can conclude that the cost of moving the ligand from solution to the "bound" conformation is reasonably small and accordingly "bioactive state" can be easily sampled in solution (Figure 8). This suggests that, in the dataset of ligand-protein complexes analyzed here, Fisher's lock & key and conformational selection are the prevalent mechanisms for binding. However, caution is required to extrapolate this finding, as we cannot ignore that we are analyzing a set of high affinity binding compounds, which means that induce-fitting events are probably underrepresented. Finally, it is worth noting that the general agreement between classical and QM/SCRF calculations provide confidence on the quality of the rough force-field used here in the classical MD simulations to describe reasonably well the ligand conformational space.

The exceptions to the rule. As noted above, most of the studied molecules can easily achieve the bioactive conformation, but we cannot ignore that there is a non-negligible number of cases where the bioactive state is a very unlikely conformation of the unbound ligand. We manually analyzed all these cases that escape from the (ligand-wise) Fisher's lock-and-key model (10% of cases with ΔG_{strain} > 5 kbT and/or ΔG_{dist} > 6 kbT at the QM/SCRF level; see Figure 9). Different situations were found (see examples in Figure 9). In some cases, the distorted part of the ligand in the protein-complex is located in a region where no electron density appears (for example the linker in 4dru-0LN, the pentyl arm in 4fgl-CLQ, or the piperidine moiety in 3ebp-CPB) and we cannot discard a problem with the annotated X-Ray conformer. Other cases can be explained by the existence of complex protein-ligand network of hydrogen bonds or salt bridges, which stabilize polar moieties in the ligand at arrangements that are not the optimum ones in the unbound state in solution (ex. 1ohr-1UN, 1eve-E20, 1f0t-PR1, 1ydt-IQB, 1htf-G26, 4dpf-0LG). These interactions are usually shielded from the solvent, thus suggesting an effective way to stabilize the ligand-protein complex upon ligand binding. For example, in the HIV- 1 protease-inhibitor systems 1ohr-1UN and 1htf-G26,

the hydroxyl group of the inhibitors is embedded into the catalytic dyad establishing charge-reinforced H-bond interactions with two Asp (D25) residues while two carbonyl groups (and an amine) interact with the NH backbone via partially buried H-bonds. Clearly, H-bonds and electrostatic ligand-protein interactions can favor ligand conformations that would be otherwise rarely populated in solution. Very interesting is the 1mq6-XLD case, where the amino-oxazolin moiety is in a distorted conformation packed between Trp^{215}, Phe^{174} and Tyr^{99} a site that reproduces a cation-pi pocket, suggesting the ring might be protonated, not neutral as was originally assumed based on rough pKa calculations.

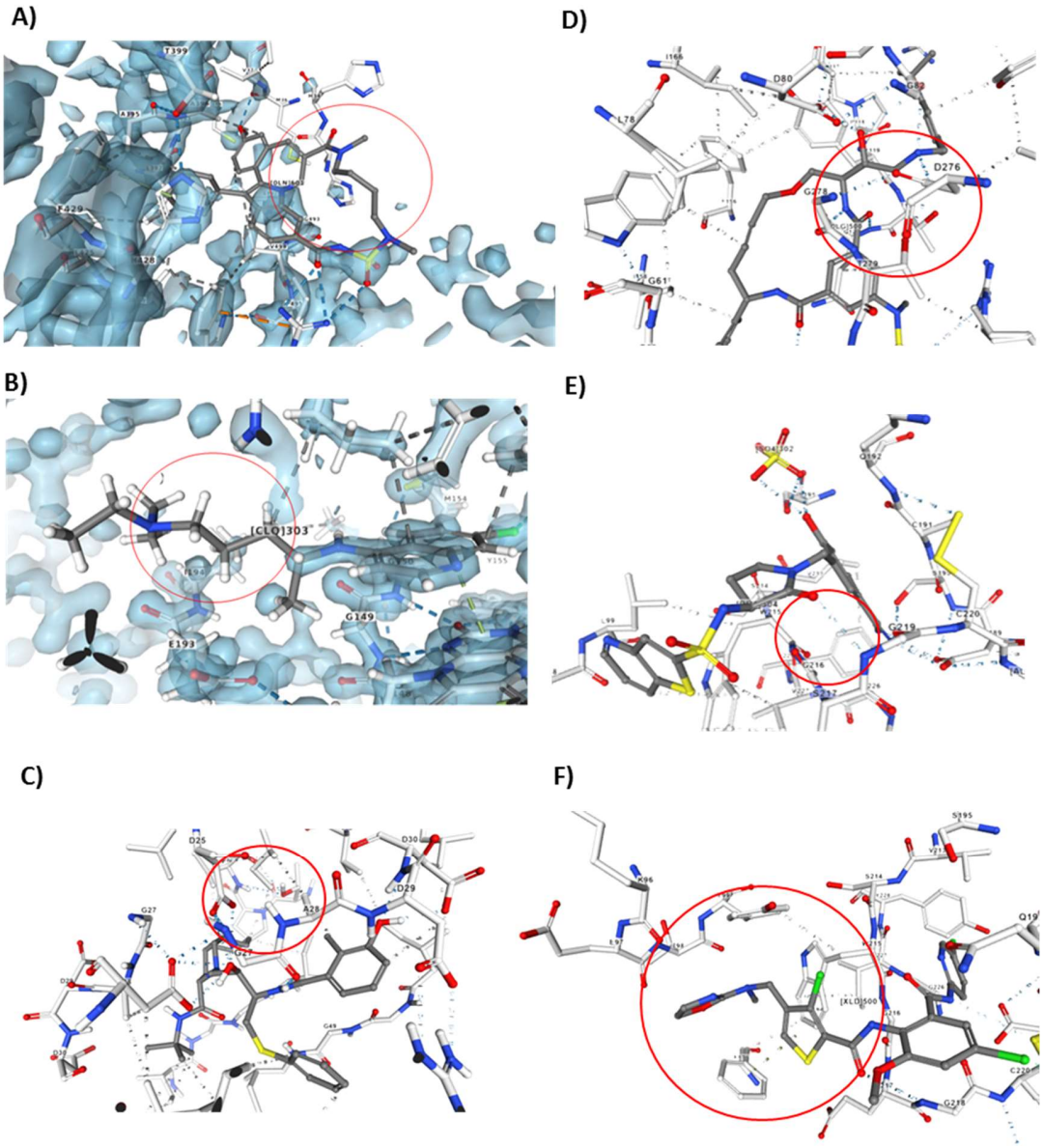


Figure 9 Examples of the exceptions to the rule. **a) 4dru-0LN**, missing electron density **b) 4fgl-CLQ**, missing electron density **c) 1ohr-1UN**, in the complex with the receptor (HIV protease), the OH is

embedded into the catalytic dyad D^{25} (from two monomers), **d) 4dpf-0LG**, our most stable conformation closely resemble the bioactive one except for the protonated nitrogen of the ethyl amino group, typical of BACE1 inhibitors, **e) 1f0t-PR1**, the molecule is well tethered within the receptor with the pirrolidinic group H-bonding the receptor NH of backbone G^{219}, **f) 1mq6-XLD**, where the amino-oxazolin moiety is in a distorted conformation packed between Trp^{215}, Phe^{174} and Tyr^{99} a site that reproduces a cation-pi pocket.

Overall molecules with net charge show slightly larger strain and distortion energies than the neutral ones, but differences are in general small (Figure 10). Molecular weight is not a good descriptor to predict the cost of adopting the bioactive conformation, as heavy molecules can contain very rigid portions (Figure 10). The number of rotatable bonds seems a better descriptor of the cost of moving to the bioactive conformation as seen in Figure 10. Taking molecules with 10 or less rotatable bonds (data for 11-12 rotatable bonds are too limited in our set of compounds) $\Delta G_{strain} = 0.4 \cdot k_b T \cdot N_{rot}$ and $\Delta G_{dist} = 0.5 \cdot k_b T \cdot N_{rot}$; both in kcal/mol. Unfortunately, while these equations are accurate for average values (Pearson's correlation around 0.85 in both cases; n=9), they are not so predictive at the individual level (Pearson's coefficient around 0.4 (strain) and 0.5 (distortion); n= 107), due to the vast dispersions of values in molecules with the same number of rotatable bonds (for example for N_{rot}=5 ΔG_{strain} range from 0 to 10 k_bT). Clearly, well-positioned H-bond and electrostatic ligand-protein interactions can favor ligand conformations that would be otherwise rarely populated in solution (in a companion paper (ct-2020-00305q) we provide the community with a database and webserver that would enable drug discoverers to apply the methodology used in this work to their small molecule(s) of interest). However, these cases are rare and the design of new more active drugs should keep in mind conformational energy considerations, as a very active drug is likely to be one whose bioactive and relaxed conformations are close each other.

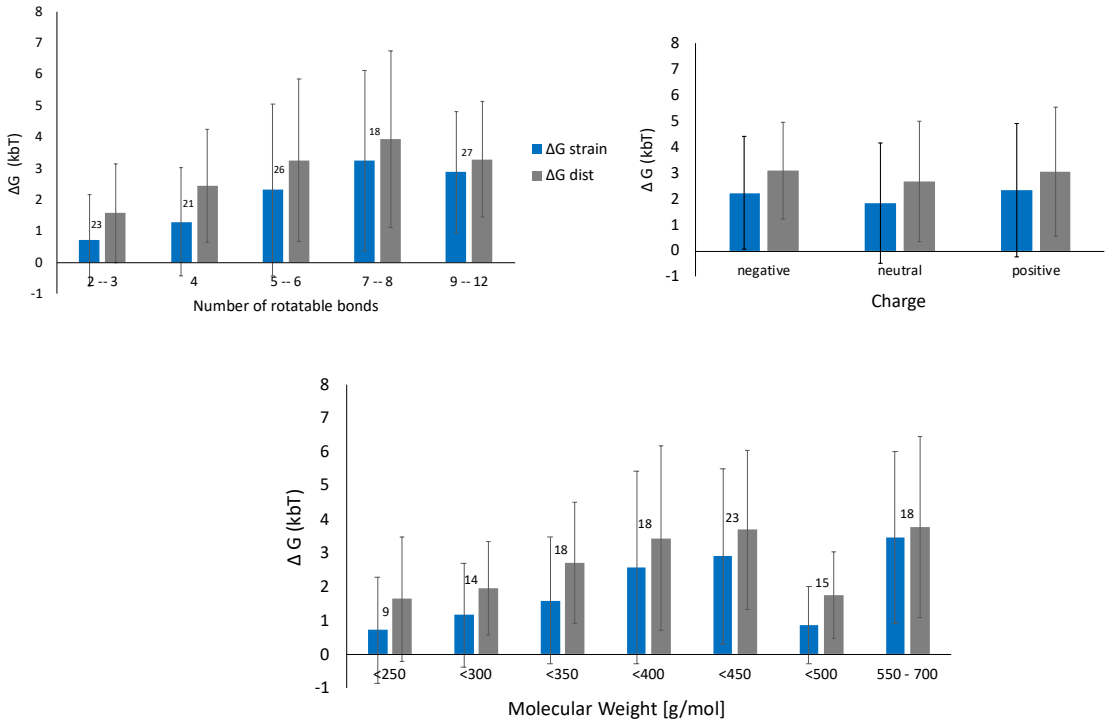


Figure 10 Dependence of the QM/SCRF strain ΔG_{strain} and distortion energy ΔG_{dist} with properties of the ligand. **A)** ΔG_{strain} against number of rotatable bonds (flexibility) of the ligand **(blue)** and ΔG_{dist} against number of rotatable bonds (flexibility) of the ligand **(grey)**; **B)** ΔG_{strain} against charge **(blue)** and ΔG_{dist} against charge **(grey)**; **C)** Dependence of the QM/SCRF strain ΔG_{strain} **(blue)** and distortion energy ΔG_{dist} **(grey)** with molecular weight. Note that when possible, data are grouped to have a similar number of compounds (marked in the histogram bars) for each category. Large standard deviations are related to cases with unusually large strain energies (see "The exceptions to the rule section" and Figure 9).

CONCLUSIONS

The availability of our Bioactive Conformational Ensemble (BCE) server and database allowed us to explore the ability of a very large set of drug-like ligands to adopt the bioactive conformation as described in crystal structures in PDB. The result of a comprehensive analysis showed us that:

- Caution is necessary before assuming structures of ligands deposited in PDB are always accurate representations of bioactive conformations. In some cases, assumed "experimental" structures are incompatible with basic chemical principles. In others, unusual arrangements are not supported by direct experimental observables, but are likely to be the result of a too simplistic modeling. Clearly, refinement of PDB using high-level calculations would benefit an efficient use of structural data for modeling purposes.

- Enhanced sampling techniques such as Hamiltonian Replica Exchange coupled with rough-automatic force-fields allow a comprehensive exploration of the ligand conformational space, the “bioactive” conformation being typically sampled during the simulations.
- Rough force-fields such as GAFF, which are easy to incorporate into automatic pipelines, provide results which are better than anticipated. When coupled to advanced enhanced sampling techniques allow the sampling of bioactive state in most cases.
- Simple B3LYP/6-31G(d)/MST calculations provide results of enough quality as compared with those obtained by more recent DFT functionals or MP2 calculations. We cannot rule out certain SCRF-related biases in the predicted conformations in solution, especially for heavily charged ligands but it seems that simple DFT/SCRF calculations are accurate enough as to provide a good description of conformational states in solution and can be safely used to predict potential bioactive states and eventually to correct automatic force-fields.
- There are many cases where the bioactive conformation is the most stable conformation in QM/SCRF calculations, and in the (free) energy penalty required to adopt the bioactive conformations is typically small. Cases with large distortion energies can be explained by the existence of many favorable protein-ligand contacts.
- Overall, it seems that good ligands, as those present in PDB, have conformational preferences facilitating their binding. In terms of the ligand, Fisher’s lock and key and conformational selection models seem prevalent in the analysed dataset and accordingly, accurate determination of ligand preferred conformational states is crucial for any structure-based drug design exercise. Whereas significant, we note that the analyzed drug-like complexes are biased and not exhaustive, thus not suitable for conclusive remarks on the absolute prevalence of the paradigms driving ligand-protein binding.

ACKNOWLEDGMENTS

We are indebted to the support of the Spanish Ministry of Science [RTI2018-096704-B-100], the Catalan SGR, the Instituto Nacional de Bioinformática; European Union's Horizon 2020 research and innovation program [BioExcel-2 project], Biomolecular and Bioinformatics Resources Platform (ISCIII PT 13/0001/0030) co-funded by the Fondo Europeo de Desarrollo Regional (FEDER) (all awarded to M.O.) as well as the MINECO Severo Ochoa Award of Excellence (Government of Spain) (awarded to IRB Barcelona) and the CDTI (Neotec grant−EXP 00094141/SNEO-20161127) (awarded to Nostrum Biodiscovery). NDB is sponsored by the Fundación Botín (Mind the Gap Program). MO is an ICREA Research Professor. FC has received funding from the Horizon 2020 research and innovation program of the European Union under the Marie Skłodowska-Curie grant agreement No. 752415.

REFERENCES

(1) Boehr, D. D.; Nussinov, R.; Wright, P. E. The Role of Dynamic Conformational Ensembles in Biomolecular Recognition. *Nat. Chem. Biol.* **2009**, *5* (11), 789–796.

(2) Cozzini, P.; Kellogg, G. E.; Spyrakis, F.; Abraham, D. J.; Costantino, G.; Emerson, A.; Fanelli, F.; Gohlke, H.; Kuhn, L. A.; Morris, G. M. Target Flexibility: An Emerging Consideration in Drug Discovery and Design. *J. Med. Chem.* **2008**, *51* (20), 6237–6255.

(3) Changeux, J.-P.; Edelstein, S. Conformational Selection or Induced Fit? 50 Years of Debate Resolved. *F1000 Biol. Rep.* **2011**, *3*.

(4) 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), 2499–2510.

(5) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623.

(6) Boström, J.; Norrby, P.-O.; Liljefors, T. Conformational Energy Penalties of Protein-Bound Ligands. *J. Comput. Aided. Mol. Des.* **1998**, *12* (4), 383.

(7) 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), 5880–5884.

(8) 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), 601–610.

(9) Avgy-David, H. H.; Senderowitz, H. Toward Focusing Conformational Ensembles on Bioactive Conformations: A Molecular Mechanics/Quantum Mechanics Study. *J. Chem. Inf. Model.* **2015**, *55* (10), 2154–2167.

(10) Juárez-Jiménez, J.; Barril, X.; Orozco, M.; Pouplana, R.; Luque, F. J. Assessing the Suitability of the Multilevel Strategy for the Conformational Analysis of Small Ligands. *J. Phys. Chem. B* **2014**, *119* (3), 1164–1172.

(11) Forti, F.; Cavasotto, C. N.; Orozco, M.; Barril, X.; Luque, F. J. A Multilevel Strategy for the Exploration of the Conformational Flexibility of Small Molecules. *J. Chem. Theory Comput.* **2012**, *8* (5), 1808–1819.

(12) Barelier, S.; Sterling, T.; O'Meara, M. J.; Shoichet, B. K. The Recognition of Identical Ligands by Unrelated Proteins. *ACS Chem. Biol.* **2015**, *10* (12), 2772–2784.

(13) Shonberg, J.; Kling, R. C.; Gmeiner, P.; Löber, S. GPCR Crystal Structures: Medicinal Chemistry in the Pocket. *Bioorg. Med. Chem.* **2015**, *23* (14), 3880–3906.

(14) Sandgren, V.; Agback, T.; Johansson, P.-O.; Lindberg, J.; Kvarnström, I.; Samuelsson, B.;

Belda, O.; Dahlgren, A. Highly Potent Macrocyclic BACE-1 Inhibitors Incorporating a Hydroxyethylamine Core: Design, Synthesis and X-Ray Crystal Structures of Enzyme Inhibitor Complexes. *Bioorg. Med. Chem.* **2012**, *20* (14), 4377–4389.

(15) Decroos, C.; Clausen, D. J.; Haines, B. E.; Wiest, O.; Williams, R. M.; Christianson, D. W. Variable Active Site Loop Conformations Accommodate the Binding of Macrocyclic Largazole Analogues to HDAC8. *Biochemistry* **2015**, *54* (12), 2126–2135.

(16) Cummings, M. D.; Lin, T.; Hu, L.; Tahri, A.; McGowan, D.; Amssoms, K.; Last, S.; Devogelaere, B.; Rouan, M.; Vijgen, L. Structure-based Macrocyclization Yields Hepatitis C Virus NS5B Inhibitors with Improved Binding Affinities and Pharmacokinetic Properties. *Angew. Chemie Int. Ed.* **2012**, *51* (19), 4637–4640.

(17) Sund, C.; Belda, O.; Wiktelius, D.; Sahlberg, C.; Vrang, L.; Sedig, S.; Hamelink, E.; Henderson, I.; Agback, T.; Jansson, K. Design and Synthesis of Potent Macrocyclic Renin Inhibitors. *Bioorg. Med. Chem. Lett.* **2011**, *21* (1), 358–362.

(18) Nantermet, P. G.; Barrow, J. C.; Newton, C. L.; Pellicore, J. M.; Young, M.; Lewis, S. D.; Lucas, B. J.; Krueger, J. A.; McMasters, D. R.; Yan, Y. Design and Synthesis of Potent and Selective Macrocyclic Thrombin Inhibitors. *Bioorg. Med. Chem. Lett.* **2003**, *13* (16), 2781–2784.

(19) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminform.* **2011**, *3* (1), 33.

(20) Marvin Was Used for Drawing, Displaying and Characterizing Chemical Structures, Substructures and Reactions, Marvin 17.21.0, ChemAxon (Https://Www.Chemaxon.Com).

(21) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688.

(22) da Silva, A. W. S.; Vranken, W. F. ACPYPE-Antechamber Python Parser Interface. *BMC Res. Notes* **2012**, *5* (1), 367.

(23) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174.

(24) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641.

(25) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926–935.

(26) Fukunishi, H.; Watanabe, O.; Takada, S. On the Hamiltonian Replica Exchange Method for Efficient Sampling of Biomolecular Systems: Application to Protein Structure Prediction. *J. Chem. Phys.* **2002**, *116* (20), 9058–9067.

(27) Sugita, Y.; Okamoto, Y. Replica-Exchange Molecular Dynamics Method for Protein Folding. *Chem. Phys. Lett.* **1999**, *314* (1–2), 141–151.

(28) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4* (3), 435–447.

(29) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New Feathers for an Old Bird. *Comput. Phys. Commun.* **2014**, *185* (2), 604–613.

(30) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G.; Bonas, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16* (8), 670–673.

(31) Bussi, G. Hamiltonian Replica Exchange in GROMACS: A Flexible Implementation. *Mol. Phys.* **2014**, *112* (3–4), 379–384.

(32) Wang, L.; Friesner, R. A.; Berne, B. J. Replica Exchange with Solute Scaling: A More Efficient Version of Replica Exchange with Solute Tempering (REST2). *J. Phys. Chem. B* **2011**, *115* (30), 9431–9438.

(33) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N· Log (N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092.

(34) Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22* (5), 501–508.

(35) Daura, X.; van Gunsteren, W. F.; Mark, A. E. Folding–Unfolding Thermodynamics of a B-heptapeptide from Equilibrium Simulations. *Proteins Struct. Funct. Bioinforma.* **1999**, *34* (3), 269–280.

(36) Rosa, M.; Micciarelli, M.; Laio, A.; Baroni, S. Sampling Molecular Conformers in Solution with Quantum Mechanical Accuracy at a Nearly Molecular-Mechanics Cost. *J. Chem. Theory Comput.* **2016**, *12* (9), 4385–4389.

(37) Rodriguez, A.; Laio, A. Clustering by Fast Search and Find of Density Peaks. *Science (80-.).* **2014**, *344* (6191), 1492–1496.

(38) Team, R. C. R: A Language and Environment for Statistical Computing. **2013**.

(39) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A. Gaussian09 Revision D. 01, Gaussian Inc. Wallingford CT. *See also URL http//www. gaussian. com* **2009**.

(40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. Gaussian 16. Gaussian, Inc. Wallingford, CT 2016.

(41) Soteras, I.; Curutchet, C.; Bidon-Chanal, A.; Orozco, M.; Luque, F. J. Extension of the MST Model to the IEF Formalism: HF and B3LYP Parametrizations. *J. Mol. Struct. THEOCHEM* **2005**, *727* (1–3), 29–40.

(42) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and

Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Function. *Theor. Chem. Acc.* **2008**, *120* (1–3), 215–241.

(43) Tirado-Rives, J.; Jorgensen, W. L. Performance of B3LYP Density Functional Methods for a Large Set of Organic Molecules. *J. Chem. Theory Comput.* **2008**, *4* (2), 297–306.

(44) Becke, A. D. A New Mixing of Hartree–Fock and Local Density-functional Theories. *J. Chem. Phys.* **1993**, *98* (2), 1372–1377.

(45) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785.

(46) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46* (7), 618.

(47) Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 Energy Evaluation by Direct Methods. *Chem. Phys. Lett.* **1988**, *153* (6), 503–506.

(48) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909.

(49) Luque, F. J.; Negre, M. J.; Orozco, M. An Am1-Scrf Approach to the Study of Changes in Molecular Properties Induced by Solvent. *J. Phys. Chem.* **1993**, *97* (17), 4386–4391.

(50) Hospital, A.; Battistini, F.; Soliva, R.; Gelpí, J. L.; Orozco, M. Surviving the Deluge of Biosimulation Data. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* e1449.

(51) Colizzi, F.; Hospital, A.; Zivanovic, S.; Orozco, M. Predicting the Limit of Intramolecular Hydrogen Bonding with Classical Molecular Dynamics. *Angew. Chemie* **2019**, *131* (12), 3799–3803.