JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Conformational Heterogeneity of Unbound Proteins Enhances Recognition in Protein−Protein Encounters

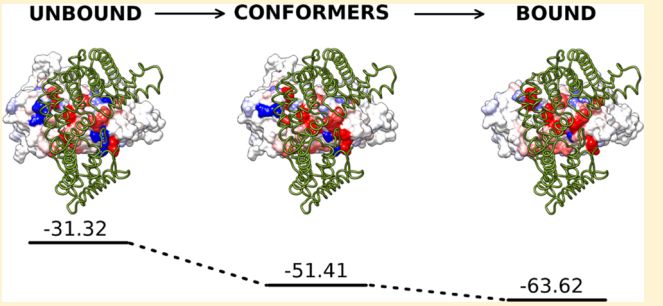
Chiara Pallara,[†] Manuel Rueda,[‡] Ruben Abagyan,[‡] and Juan Fernández-Recio*,[†]

[†]Joint BSC-CRG-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, C/Jordi Girona 29, Barcelona 08034, Spain

[‡]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, United States

Supporting Information

ABSTRACT: To understand cellular processes at the molecular level we need to improve our knowledge of protein−protein interactions, from a structural, mechanistic, and energetic point of view. Current theoretical studies and computational docking simulations show that protein dynamics plays a key role in protein association and support the need for including protein flexibility in modeling protein interactions. Assuming the conformational selection binding mechanism, in which the unbound state can sample bound conformers, one possible strategy to include flexibility in docking predictions would be the use of conformational ensembles originated from unbound protein structures. Here we present an exhaustive computational study about the use of precomputed unbound ensembles in the context of protein docking, performed on a set of 124 cases of the Protein−Protein Docking Benchmark 3.0. Conformational ensembles were generated by conformational optimization and refinement with MODELER and by short molecular dynamics trajectories with AMBER. We identified those conformers providing optimal binding and investigated the role of protein conformational heterogeneity in protein−protein recognition. Our results show that a restricted conformational refinement can generate conformers with better binding properties and improve docking encounters in medium-flexible cases. For more flexible cases, a more extended conformational sampling based on Normal Mode Analysis was proven helpful. We found that successful conformers provide better energetic complementarity to the docking partners, which is compatible with recent views of binding association. In addition to the mechanistic considerations, these findings could be exploited for practical docking predictions of improved efficiency.



INTRODUCTION

Proteins are key components in the cell and function through intricate networks of interactions[1] that are involved in virtually all relevant biological processes, such as gene expression and regulation, enzyme catalysis, immune response, or signal transduction.[2,3] Understanding such interactions at the molecular level is essential to target them for therapeutic or biotechnological purposes. X-ray crystallography and NMR techniques have produced a wealth of structural data on protein−protein complexes, which has largely extended our knowledge on molecular recognition and protein association mechanism and has fostered drug discovery. However, such structural data covers only a tiny fraction of the estimated number of protein−protein complexes formed in cell,[4] and, therefore, computational approaches that can complement such experimental efforts are strongly needed. In recent years, a variety of protein−protein docking methods have been reported, based either on template modeling[5−7] or on *ab initio* algorithms. Geometry-based methods try to find the best surface complementarity between interacting proteins, using simplified structural models and approximate scoring functions. A popular strategy is to discretize the proteins into grids and use Fast Fourier Transform (FFT) algorithms[8] to accelerate search on the translational space, such as in FTDock,[9] PIPER,[10] GRAMM-X,[11] ZDOCK,[12] or on the rotational space, as in Hex[13] or FRODOCK.[14] Another strategy to explore surface complementarity is geometric hashing, as used in PatchDock.[15] Docking methods based on energy optimization use a variety of sampling strategies based on molecular mechanics, such as molecular dynamics in HADDOCK,[16] or Monte Carlo minimization in RosettaDock[17] or ICM-DISCO.[18] The function used to identify the best orientations is an important aspect of docking, and dedicated scoring schemes have been developed, based on energy terms, such as in pyDock,[19] or on statistical potentials as in SIPPER[20] or PIE.[21] The Critical Assessment of PRediction of Interactions (CAPRI; http://www.ebi.ac.uk/msd-srv/capri/) experiment has indeed shown that accurate models can be produced by docking in many of

Received: February 25, 2016

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

the cases.[22] However, there are other cases in which all docking methods systematically fail, typically the most flexible ones.[23,24]

Thus, one of the major challenges in docking is how to deal with molecular flexibility and conformational changes that happen upon association.[23,24] A major hurdle is the computational cost of integrating docking and conformational search, aggravated by our limited knowledge of the protein−protein association mechanism. Different mechanisms for flexible protein−protein binding have been proposed. Perhaps the most widespread view is the induced-fit mechanism, in which the interacting partners are involved in initial encounters that evolve toward the final specific complex by adjusting their interfaces. Most of the reported methods for flexible docking try to mimic this mechanism, typically using an initial rigid-body search followed by a final refinement of the interfaces as in ICM-DISCO,[25] HADDOCK,[16] RosettaDock,[17] or FiberDock[26] or by integrating small deformations of the global structures during the sampling based on normal modes as in ATTRACT[27,28] or SwarmDock.[29]

An alternative mechanism is conformational selection, which was initially proposed for systems in which the ligand selectively binds one of the conformers of the dynamically fluctuating receptor protein.[30,31] This was generalized to the "conformational selection and population shift" concept, which postulated that flexible proteins in solution naturally sample a variety of conformational states, and the ligand protein preferentially binds to a pre-existing subpopulation of such conformers, thus adjusting the equilibrium in favor of them.[32−34] Recently, the conformational selection model has been extended to include different mutual conformational selection and adjustment steps,[35] so that the unbound conformational states that are available for mutual selection might not be initially in the bound conformation. The conformational selection model has been largely supported by several structural studies including MD, NMA, X-ray crystallography, and NMR experiments,[35−39] **and later strongly** confirmed by theoretical analysis based on the correlation between complex association/dissociation rates and several molecular descriptors detailing specific features of protein intrinsic flexibility and complex formation.[40] This mechanism can be implemented in a computational docking strategy by using precomputed ensembles of unbound proteins, which ideally contain conformers that are suitable for binding the interacting partner. However, to date this strategy has not been really used for practical protein−protein docking predictions. Most of the prior studies were limited to the use of a few selected conformers and/or applied to specific cases of interest.[41−43] Unexpectedly, the few systematic analyses published so far[44−46] failed to improve the structural prediction of protein complexes with respect to the unbound structure. This could be related to an unrealistic representation of the motions occurring in the time scale of molecular association or to the use of only a few conformers.[44−46] Indeed, for small proteins like ubiquitin it is possible to obtain more representative ensembles, based on RDC data, which are definitely useful in docking predictions.[47] However, this approach is difficult to generalize for large scale predictions due to experimental limitations. Therefore, it would be important to find practical ways of generating ensembles that include conformers that improve binding. This could help not only to improve docking predictions but also to advance toward a better understanding of flexible protein−protein association mechanism. With this purpose in mind, here we used three different computational approaches to represent the conformational heterogeneity of the unbound proteins and tested them on a standard protein−protein docking benchmark. Our analysis clearly shows that a simple molecular mechanics minimization approach provides sufficient conformational heterogeneity to improve docking predictions in medium-flexible cases, which are the most likely to follow the conformational selection mechanism.

METHODS

Generation of Protein Conformational Ensembles. We used three different computational techniques to generate conformational ensembles starting from the unbound protein structures: MODELER minimization (MM), molecular dynamics simulations (MD), and Normal Modes Analysis (NMA).

Conformational search based on the optimization of a molecular probability density function (PDF) was performed with the comparative modeling program MODELER version 9v10,[48] using as template the unbound X-ray structure of the same protein, and default parameters. Cofactors and small-molecule ligands, if present in the template structure, were taken into account during the modeling procedure. MODELER minimization (MM) is based on an optimization step using the variable target function method (VTFM) with restrained conjugate gradients (CG), followed by a refinement step using short (a few ps) molecular dynamics (MD) and simulated annealing (SA), with CHARMM parameters and distance-dependent dielectric constant.

Conformational search based on Molecular Dynamics (MD) was performed by a 10 ns long explicit solvent unrestrained MD simulation on the unbound structure using the force field AMBER parm99[49] and the AMBER8 package.[50] As a first preparation step, all the missing loops in the protein structures were modeled using the MODELER program. The parametrization of each system was performed using AMBER's module LEAP, whereas the cofactor and small-molecule ligand libraries, when needed, were written with the AMBER modules ANTECHAMBER and LEAP. Each system was then minimized, solvated, and equilibrated at similar conditions to those previously described for the MoDEL database,[51] as follows. First, original PDB coordinates were stripped of hydrogen atoms, monovalent ions, and all water molecules. Noncovalent ligands were kept and parametrized with the GAFF force-field using standard procedures,[52] and missing side-chain atoms and hydrogen atoms were added from AMBER residue libraries using the LEAP AMBER tool. Each system was relaxed by a short restrained energy minimization (20 steps steepest descent, 80 steps conjugate gradient, restraining all the heavy atoms with a 20 kcal/mol Å^2 to the initial structure) to relieve highly unfavorable sterical clashes. Then each minimized structure was immersed in a periodic truncated octahedron box containing a 12 Å buffer of TIP3P water molecules, and Na^+ and Cl^- counterions were added to the solvent bulk to maintain neutrality of the system and reach 50 mM NaCl ionic strength.

Each solvated system underwent a short solvent minimization and five-step equilibration protocol. First, a 500-cycle steepest descent and a 500-cycle conjugate gradient minimization were performed, applying harmonic restraints with a force constant of 50 kcal/(mol·Å^2) to all protein atoms in order to minimize the solvent molecules. Then, the five-step equilibration was performed by applying periodic boundary

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

conditions and computing long-range electrostatics by the particle-mesh Ewald method. At each stage, the integration time step was set to 2 fs, the system pressure to 1 atm, and the nonbonding cutoff distance to 12 Å. The five steps are

Step 1: A 40 ps MD simulation was run applying harmonic restraints to all the protein atoms with a force constant of 25 kcal/(mol·Å^2), raising the temperature to 300 K by Langevin dynamics approach with a collision frequency of 1 fs.

Step 2: A 20 ps step was performed, setting the temperature to 300 K and reducing system restraints to 10 kcal/(mol·Å^2).

Step 3: Another 20 ps simulation was run with 10 kcal/(mol·Å^2) restraints only to the protein backbone atoms.

Step 4: A further 20 ps simulation was performed, decreasing protein backbone restraints to 5 kcal/(mol·Å^2).

Step 5: A final 100 ps unrestrained MD simulation was run without any restraint.

Finally, a 10 ns MD simulation was performed in isothermal−isobaric ensemble, setting pressure to 1 atm and temperature to 300 K. From each MD simulation, two conformational ensembles were created by extracting trajectory snapshots every 10 or 100 ps. Additionally, a random subset of 11 benchmark cases (1ACB, 1AY7, 1D6R, 1E6J, 1GCQ, 1IRA, 1JMO, 1PXV, 2HRK, 2CFH, 2C0L) was selected for longer simulations. Each protein underwent two 100 ns long explicit solvent unrestrained NPT-MD simulations, at the temperatures of 300 and 340 K, respectively, using the same force field as above.

Conformational search based on Normal Mode Analysis (NMA) was performed by an in-house protocol on a small subset of the 6 flexible benchmark cases that show strong binding affinity (in which potential errors in the docking scoring function have a minimal impact). NMA is a powerful modeling technique that allows for a fast and accurate description of the intrinsic movements of biomolecules. Modern interpretations of the procedure use the elastic network model (ENM), first described by Tirion as an all-atom version[53] and later reformulated as coarse-grained.[54] In the ENM, the biomolecule is represented as a network of connected atoms, where each node is connected to all the atoms within a cutoff, and the springs represent the interactions between the nodes. Here we used the Anisotropic Network Model[54] that describes the protein as a Cα model, and we assigned the spring constants by a term that assumes an inverse exponential relationship with the distance,[55] analog to that from Hinsen.[56] We tried to enhance the conformational space by introducing an iterative exploratory search. The proposed method is called eNMA (enhanced NMA) and creates enriched structurally diverse ensembles. The algorithm works as follows:

Step 1: Starting from the unbound Cα atoms, we created 100 discrete Cartesian conformers from random combinations of displacements along the first 10 normal modes (as described elsewhere).[57] The average Cα displacement with respect to the original structure was set to ∼1 Å.

Step 2: The resulting conformers were then clustered hierarchically via *average linkage* method (as implemented in *ptraj10*)[50] to obtain 100 diverse conformations.

Step 3: Each conformer from the cluster was sent to Step 1, and the whole cycle was started.

Step 4: The process was ended up after 10 iterations.

In total, around 100,000 intermediate structures were created per protein, but we only kept the ones resulting from the clustering (i.e., 100 × 10 iterations = 1,000 discrete conformers). The final structures underwent a last optimization step with MODELER 9.10. All-atom models were rebuilt by adding missing atoms and side-chains and were atomically refined with MODELER (using the original Cα model as template) to fix incorrect bond distances.[48,58] In addition, 100 discrete conformers were randomly selected, and for each of them 10 MODELER models were built. The whole procedure took ∼2 h per protein (ranging from 40 min for 1ACB ligand, with 70 Cα atoms, to 5 h for 1IBR ligand, with 440 Cα atoms) on 1 core of an Intel Xeon 3.5 GHz CPU (16GB RAM) Linux workstation. Note that our conformational search was unguided, but it could be also guided in future applications (i.e., selecting the combination of models that provides the best score on a given fitness function).

Docking Simulations. For all the dockings experiments, the FTDock docking program[9] was used to generate 10,000 rigid-docking poses based on surface complementary and electrostatics at 0.7 Å grid resolution, and then, each docking solution was evaluated by the energy-based pyDock scoring scheme,[19] based on desolvation, electrostatics, and van der Waals energy contributions. All energy values are shown as arbitrary units. Cofactors, small-molecule ligands, and ions were excluded during the sampling and the scoring calculations in docking.

Benchmark. In order to validate the approach proposed here, we used the protein−protein docking benchmark 3.0,[59] comprising a total of 124 test cases in which the structures of both the free components and the complex are known. We have classified these cases according to the conformational variation of the proteins from the unbound to the bound state (based on the RMSD of Cα atoms of the interface residues as defined in the mentioned protein−protein benchmark 3.0), which resulted in the following categories: "rigid" (I-RMSD$_{C\alpha}$ < 0.5 Å), "low-flexible" (0.5 Å < I-RMSD$_{C\alpha}$ < 1.0 Å), "medium-flexible" (1.0 Å < I-RMSD$_{C\alpha}$ < 2.0 Å), "flexible" (2.0 Å < I-RMSD$_{C\alpha}$ < 3.0 Å), and "highly flexible" (I-RMSD$_{C\alpha}$ > 3.0 Å). The quality of the docking predictions was evaluated according to the ligand protein Cα-RMSD with respect to the complex crystal structure (after superimposing the Cα atoms of the receptor molecules). A docking experiment was considered successful if a near-native solution (a docking pose with ligand Cα-RMSD < 10 Å) was ranked among the top 10 predictions according to the pyDock scoring function. Structural analyses of proteins, including RMSD and clashes calculations, were performed using the ICM program[60] (www.molsoft.com).

RESULTS

Unbound Conformational Ensembles from Molecular Mechanics Contain Conformers with Better Binding Capabilities than the Unbound Structure. Here we explored in a systematic way whether a minimal description of the conformational heterogeneity of the interacting proteins could significantly improve their binding capabilities. For that purpose, we generated conformational ensembles from the unbound proteins of the complexes in the protein−protein benchmark 3.0.[59] Ensembles of 100 conformers were initially generated by using two distinct molecular mechanics procedures: a fast restricted conformational optimization, as implemented in MODELER, and a much more computationally demanding molecular dynamics method, as implemented in the AMBER package (see Methods). Figure 1 shows examples of the typical conformational heterogeneity (at backbone and side-chain level) generated by MODELER minimization (MM). The deviation of the interface atoms from the initial

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

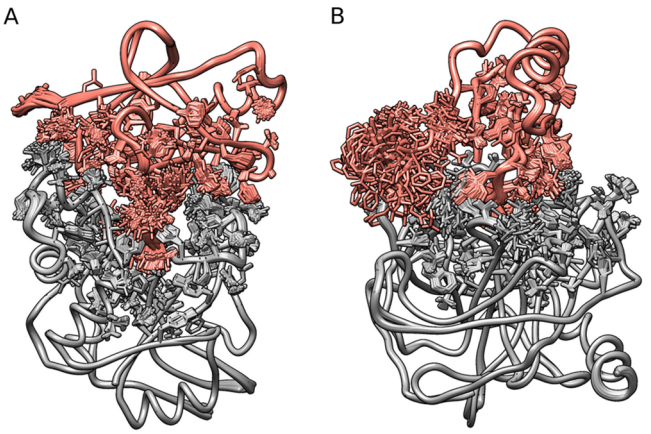


Figure 1. Representative conformational ensembles generated by MODELER minimization. 100 conformers independently generated by MODELER for receptor and ligand proteins are shown for two benchmark cases: (A) 1PXV and (B) 1ACB. Conformers were superimposed onto the corresponding molecules in the reference complexes for visualization. Only interface side chains are shown for the sake of clarity.

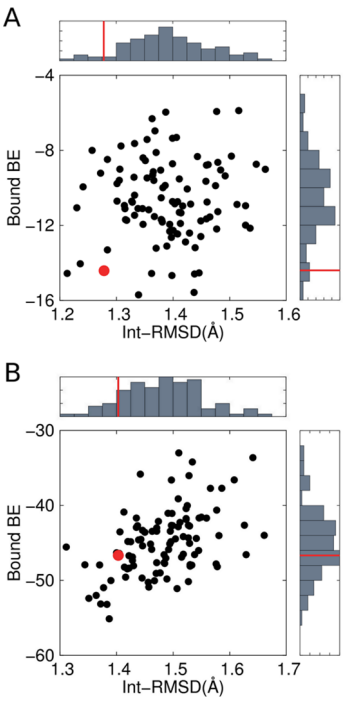


Figure 2. Distribution of geometrical and energetic values for ensemble conformers. The correlation between the full atom interface RMSD (Int-RMSD) with respect to the bound state and the pyDock binding energy (au) toward the bound partner in the native orientation (bound BE) for all conformers in MODELER ensembles are shown for two benchmark cases: (A) 2F0R (1S1Q receptor) and (B) 1MKF (1ML0 receptor). Distribution of Int-RMSD and bound BE values are shown as histograms. Data for the unbound X-ray structure are shown in red.

unbound structure was 1.2 Å RMSD on average (ranging from 0.6 Å for the 1R0R receptor to 7.7 Å RMSD for the 2QFW receptor).

We compared the unbound models to their corresponding structures in the complex to structurally characterize these conformers and to estimate their capabilities for binding. In order to do that, we first structurally superimposed each model into the corresponding native complex structure (considering only the Cα atoms) and then computed the following parameters: (i) the RMSD for all Cα atoms (Cα-RMSD) with respect to the complex structure; (ii) the RMSD for all interface atoms (Int-RMSD) after superimposing only those interface atoms; (iii) the pyDock binding energy (au) with the bound partner; (iv) the pyDock binding energy (au) with the unbound partner; and (v) the number of clashes with the bound partner. The values for these parameters in the different conformers generated by MODELER are randomly distributed following a Gaussian function (Figure S1). Except for a few cases, such as the viral chemokine binding protein M3 (1ML0 receptor), there is no significant correlation between the binding energy of the different conformers in the native orientation and their similarity with respect to the bound structure (Figure 2). Perhaps the main reason for this is that, in general, due to the limited sampling used here, these conformers are not exploring the vicinity of the bound state. Indeed, only 20% of the benchmark proteins contain conformers within 1.0 Å Int-RMSD from the bound state (and in virtually all of these cases the unbound state already had Int-RMSD < 1.0 Å from bound).

Ensembles generated by short MD trajectories showed larger conformational variability, but in general they were not closer to the bound state (Figure S1 and S2). Increasing the number of conformers to 1,000 (Figure S3) did not significantly modify the range of conformational variability for either sampling method.

We next aimed to identify which conformers of the ensemble seemed more promising for binding. Thus, we selected the best conformers of the ensemble according to the criteria analyzed in the previous paragraphs. Figure 3 shows the best conformers for receptor and ligand proteins identified according to each parameter as compared to the unbound receptor and ligand structures for all benchmark cases. Regarding the RMSD with respect to the complex structure, only in a few cases (21% and 6%, according to Cα-RMSD and Int-RMSD, respectively) the best pair of conformers were significantly better (i.e., more than 10% change; averaged for receptor and ligand proteins) than the unbound X-ray structures. These cases were not particularly enriched in conformers with Int-RMSD < 1.0 Å. Actually, in some cases (14% and 36%, according to Cα-RMSD and Int-RMSD, respectively) the best conformers were even farther from the bound structure than the unbound one.

Interestingly, we found a much higher number of cases in which the best conformers showed significantly better binding energy (in 46% and 51% of cases, when considering the bound or unbound structure as partner, respectively) or fewer clashes (in 69% of cases) than the unbound X-ray structure. It is remarkable that the improvement in binding energy was independent of the structural similarity to the bound structure. Again, one of the reasons is that in the majority of cases the limited conformational sampling used here does not permit reaching the vicinity of the bound state, and, therefore, in such unbound minima any small improvement toward the bound state is not relevant in terms of binding energy.

Although MD ensembles showed larger conformational variability (Figures S1 and S2), the percentage of cases with conformers that became significantly better than the unbound state according to each of the above-mentioned criteria (12%, 7%, 37%, 62%, and 69%, respectively) was very similar to those observed for the MODELER ensembles. However, the

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

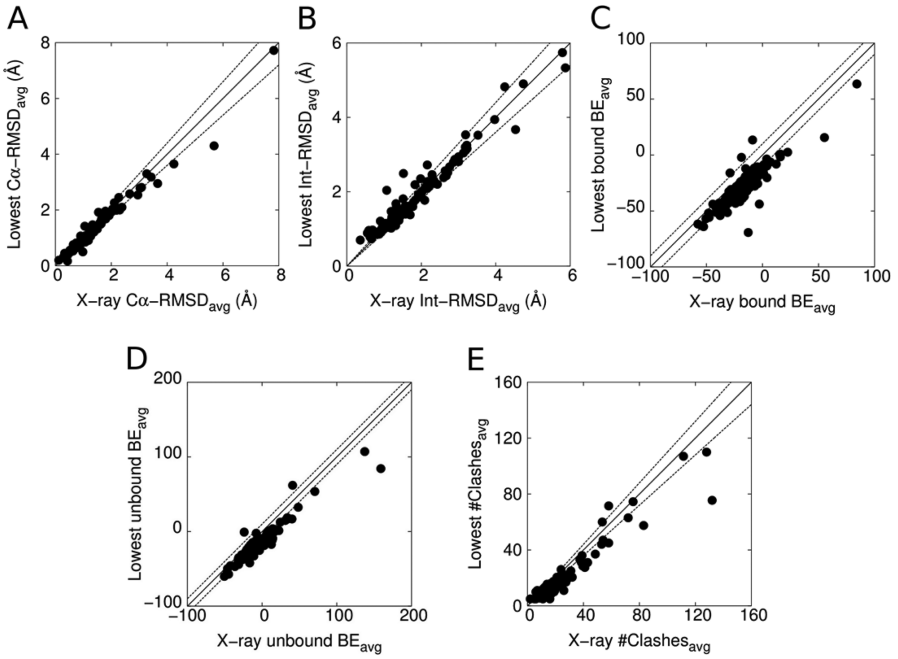


Figure 3. Best ensemble conformers according to quality criteria based on the complex native orientation. For each benchmark case, it is shown the best pairs of receptor and ligand conformers in the conformational ensemble according to the following criteria: (A) Cα-RMSD, (B) Int-RMSD, (C) pyDock binding energy with the bound partner (au), (D) pyDock binding energy with the unbound partner (au), (E) number of clashes with respect to the bound partner. The above-described descriptors were calculated independently for the best receptor and ligand conformers and then averaged. These are compared to those of the unbound X-ray structures. Dashed lines represent the (arbitrary) range of variation that we used to consider a change as significant, and it was defined as 10% in the RMSD- and clash-based criteria or 10 au in the energy-based criteria.

ensembles generated from the short MD trajectories showed even more limited conformational sampling in the vicinity of the bound state, since less than 4% of the benchmark proteins had conformers with Int-RMSD < 1.0 Å with respect to the bound state (as compared to 20% in MODELER). Moreover, in most of the cases (74% and 71%, according to Cα-RMSD and Int-RMSD, respectively) the best conformers from MD were even farther from the bound structure than the unbound one. This could be related to the limited sampling of the short MD simulations used here, as well as to the fact that the protein structures in solution typically deviates 1−2 Å RMSD from that in the crystal.[61] Indeed, this is what we observe in our MD ensembles (Figures S1 and S2). Of course, we should not disregard possible inaccuracies in the force-field, but they are usually more relevant in very long trajectories but not so much in the suboptimal sampling used here.[62]

Selected Conformers Can Yield Significantly Better Docking Results than Unbound Subunits. The fact that in the majority of cases the conformational ensembles contained conformers that showed better binding energy capabilities than the unbound X-ray structure encouraged us to evaluate their use for docking. Since the systematic cross-docking of all conformers for receptor and ligand proteins would be impractical, we tried instead to estimate the expected performance of the unbound ensembles for docking in the best-case scenario. Therefore, based on the native orientation, we selected those conformers that seemed the best candidates to improve the docking predictions according to the criteria described in the previous section: (i) the lowest Cα-RMSD with respect to the bound state, (ii) the lowest Int-RMSD, (iii) the best binding energy with the bound partner, (iv) the best binding energy with the unbound partner, and (v) the smallest number of clashes with the bound partner. These conformers were used in protein−protein docking as described in the Methods section.

Figure 4A shows the docking success rates for the top 10 predictions when using these selected conformers, with all the details in Table 1. Interestingly, the results do not significantly change when using a larger number of conformers (1,000) generated by MODELER (and applying the same procedure for selecting the best expected conformers), or when conformers were generated by short MD trajectories, either using 100 or 1,000 conformers (Figure S4). Strikingly, when we used the best conformers based on Cα- or Int-RMSD with respect to the complex structure, the docking results were slightly worse than those of unbound docking, as can be seen in Figure 4A (the results did not significantly change when selecting only those cases in which the best conformer had significantly better Cα- or Int-RMSD than that of the unbound structure). This can be due to the fact that either MODELER minimization or a short MD trajectory cannot generally sample too far from the unbound structure, and therefore cannot reach the vicinity of the bound state in most of the cases. However, when using the conformers that would give the best binding energy or the smallest number of clashes when in the native orientation, the docking results significantly improved with respect to those of the unbound structures, as can be seen in Figure 4A. Again, this did not correspond to an improvement in geometrical terms. Indeed, in 99% of the cases in which the best-energy conformer improved the docking predictions, such conformer did not have significantly better Int-RMSD from the complex structure than the unbound conformation. For comparison, we show the success rates that we would obtain when using the bound structures, which establishes the upper limit for the expected docking success with this approach. The success rates of the selected conformers based on binding energy are more than

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

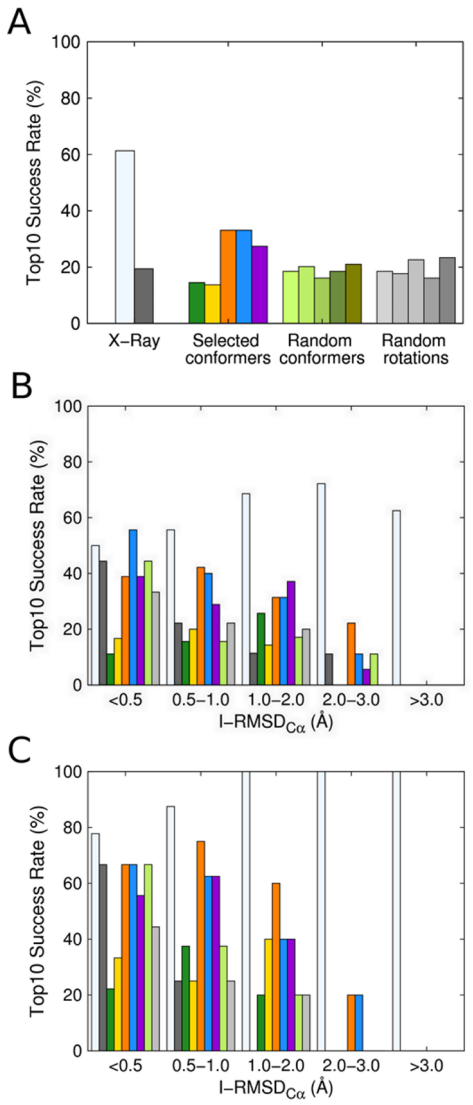


Figure 4. Docking performance for selected conformers. (A) Docking success rates for the top 10 predicted models on the protein−protein docking benchmark when using selected conformers according to specific criteria: Cα-RMSD (green), Int-RMSD (yellow), binding energy toward the bound partner (orange), binding energy toward the unbound partner (blue), number of clashes with respect to the bound partner (magenta). For comparison, the docking success rates for bound (white) and unbound (dark gray) X-ray structures are also shown. To show the significance, docking rates for five random conformers pairs (green gradations) and five random initial rotations of the unbound docking partners (gray gradations) are also shown. (B) Docking success rates according to the conformational variability between the unbound and bound structures for selected conformers (same color code as above). For comparison, docking success rates for the bound and unbound X-ray structures, as well as for one random pair of conformers (light green) and one random initial rotation of the unbound docking partners (light gray) are also shown. (C) Docking success rates according to unbound−bound conformational variability on the 28 cases of the benchmark with reported high affinity (ΔG < −12.0 kcal/mol) when using selected conformers, as well as the bound and unbound X-ray structures, one random pair of conformers and one random initial rotation of the unbound docking partners (same color code as above).

half of the maximum expected success rates when using the bound structures. It is important to mention that this approach only proves the existence of some conformers that can provide better docking predictions than the unbound structures. However, in a realistic scenario, it would be impossible to accurately identify these optimal conformers from the unbound ensemble. The success rates obtained in this section can be considered as a rough estimation of the potential predictive rates that could be obtained if all ensemble conformers were used in a docking strategy. One could hypothesize that if these conformers are able to find docking orientations with good scoring, such orientations would automatically appear well-ranked in a general docking pool, even if they include docking poses obtained with other conformers. Preliminary tests show that merging the results of 100 docking runs with one conformer from the receptor protein and another from the ligand protein, each time a different one, would provide similar success rates than the ones obtained here when using the best-energy (with the unbound partner) conformers (data not shown). There could be other strategies that can be devised for using all unbound conformers in docking. However, the systematic evaluation of these different approaches in a practical docking procedure is beyond the scope of this work. The findings here show some potentiality for future developments in ensemble protein−protein docking, but the practical problem is still unsolved.

In order to provide a statistical significance for these results, we randomly selected five conformers from the conformational ensemble. The results for each random conformer were similar (within experimental error) to those of the unbound structure (Figure 4A) and show that the conformers selected according to the optimal binding energy improved significantly the docking results with respect to the randomly chosen conformers. An alternative possible explanation for the docking improvement when using the ensemble conformers might be related to the limited discrete sampling in FTDock derived from the fix number of ligand rotations (which makes coarser surface sampling for large proteins) and the grid resolution of 0.7 Å (which introduces inaccuracies in the atomic coordinates). This creates a stochastic dependence of the FTDock docking algorithm on the initial rotation of the interacting subunits, given that each initial rotation of the interacting proteins could be mapped on different cells of the 3D grids. This is a limitation of any FFT-based algorithm, and it was already shown that performing parallel docking runs using several initial rotations provided more consistent docking results than using just a single one.[14] To evaluate the possibility that the extensive sampling in the atomic positions provided by the use of unspecific conformers prior to docking could compensate the suboptimal grid-based sampling of FTDock, we performed five different docking runs with random initial rotations for the unbound receptor and ligand proteins. The results from the individual random rotations were similar, within experimental error, to the unbound docking results (Figure 4A).

These results suggest that the selected conformers according to specific criteria (i.e., optimal binding energy, number of clashes) were more beneficial for docking than just a random selection of conformers or initial rotations. Overall, this clearly shows that conformational heterogeneity in the interacting subunits can improve the binding capabilities of the unbound X-ray structures.

A Simple Description of Conformational Heterogeneity Is Particularly Beneficial for Low- and Medium-Flexible Cases. We have analyzed whether the docking improvement when using ensembles depends on the conforma-

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. Docking Performance of Conformers Selected from MODELLER Ensembles[a]

PDB	bound	unbound	Cα-RMSD	Int-RMSD	bound BE	unbound BE	clashes
Rigid (I-RMSD$_{C\alpha}$ < 0.5 Å) (18 Cases)							
1AVX	**2**	**102**		**33**	**40**	**1**	**231**
1FSK	**3**	**3**	**39**	**34**	**1**	**1**	**514**
1GHQ	7455			6528		1878	
1IQD	**1**	**8**	**64**	**3**	**6**	**6**	**3**
1KLU	18	1246	6002	4468	2587	6498	1647
1KTZ	48	3725	6333				309
1NCA	14	7	1269		1332	7	1
1NSN	**405**	**500**	**254**	**5587**	**33**	**33**	**1085**
1PPE	**28**	**6**	**12**	**2**	**5**	**1**	**4**
1R0R	**1**	**3**	**258**	**230**	**9**	**17**	**37**
1SBB	161	298	73				
1WEJ	**1**	**274**	**5**	**456**	**64**	**2**	**9**
2JEL	1	42	16	25	12	2	1
2MTA	2	78	61	187	48	3	554
2PCC	12	6	91	12	6	4	11
2SIC	**1**	**8**	**3378**	**1**	**2**	**249**	**1**
2SNI	**1**	**3**	**1**	**16**	**1**	**1**	**1**
2UUY	69	4472	4801	64	159	11	1997
Low-Flexible (I-RMSD$_{C\alpha}$ 0.5−1.0 Å) (45 cases)							
1AHW	1043	4049	6796	838	431	836	2974
1AY7	**1**	**24**	**130**	**118**	**4**	**2**	**7**
1AZS	1	30			6	6	
1BJ1	9				18	9	25
1BUH	71	66	209	426	36	24	119
1BVN	**1**	**2**	**2**	**1**	**1**	**1**	**687**
1DQJ	216	**604**	**261**	**3363**	**75**	**25**	**223**
1E96	113	1	59	168	73	5	130
1EAW	**8**	**622**	**297**	**86**	**42**	**25**	**1**
1EFN	6	166	197	1684	203	97	172
1EWY	4	8	200	5	10	10	1
1F34	**1**	**139**	**174**	**226**	**52**	**280**	**2**
1F51	2	7	13	375	1505	130	8
1FQJ	14	309	396	482	218	438	101
1GCQ	274	1091	574	1540	5	5	364
1GLA	61	50		12	6	21	131
1GPW	1	1	1	1	1	1	1
1HE1	1	3958	102	4506	2425	523	2629
1HE8	138	2917	2612	1503	277	242	3437
1IJK	16	1309	69	61	493	487	388
1J2J	46	19	303	18	2	3	5
1JPS	**709**	**481**		**2135**	**1**	**2**	
1K4C			3036	3369	2275		2379
1K74	150	14	172	82	1	1	24
1KAC	4737	1286	3545	990	107	19	917
1KXQ	1	250	8	4	4	1	1
1MAH	**1**	**19**	**2**	**4**	**4**	**1**	**1**
1MLC	2	37	50	10	1	97	144
1N8O	3	53			5		90
1QA9	3253	7378	5902	6152	1546	37	7973
1QFW	81	239	234		26	21	72
1RLB	1319	4094		7917			
1S1Q	147	1211	2994	541	164	175	87
1T6B	**3**	**56**	**802**	**1464**	**2**	**11**	**3**
1TMQ	1	1	27	4	54		4
1UDI	1	1	2	47	1	1	420
1YVB	1	19	1	2	3	21	7
1Z0K	2	8	523	57	42	11	44
1ZHI	5	3	7450		196	5	5
2AJF	5	1788		311	562	2268	2122
2B42	**1**	**1**	**2**	**37**	**1**	**2**	**21**

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. continued

PDB	bound	unbound	Cα-RMSD	Int-RMSD	bound BE	unbound BE	clashes
Low-Flexible (I-RMSD$_{C\alpha}$ 0.5−1.0 Å) (45 cases)							
2BTF	1	33	120	26	60	9	250
2OOB	588	112	131	217	106	547	432
2VIS	64						
7CEI	1	19	11	1	1	1	20
Medium-Flexible (I-RMSD$_{C\alpha}$ 1.0−2.0 Å) (35 Cases)							
1A2K	36	114	5641	284			782
1AK4	2420	2040	3983	3619		2721	1055
1AKJ	89	656	345	261	204	162	1168
1B6C	1	3	6	11	1	1	21
1BGX	1						
1BVK	7	18	4	146	87	85	2
1D6R	1050	2128	227	888	669	785	102
1DFJ	**6**	**557**	**2**	**1**	**1**	**1**	**4**
1E6E	1	3	2	1	1	8	1
1E6J		33	34	3	1	2	5
1EZU	**1**	**2048**	**3633**		**1449**	**1547**	**102**
1FC2	127		233	326	1256	683	171
1GP2	1			842	85		87
1GRN	2	858	184	1184	450	23	2909
1HIA	99	40	415	42	23	166	7
1I4D	1			642		44	132
1I9R	15	846	568	212		99	
1K5D	**1**	**360**	**85**		**2**	**610**	
1KXP	**1**	**16**	**14**	**1**	**1**	**1**	**1**
1ML0	1	173	80	140	1	1	9
1NW9	1	9	181	36	43	39	181
1OPH	59	14		469			2584
1VFB	37	59	86	59	128	31	95
1WQ1	4	2448	5	1077	16	6	6
1XD3	1	1	3	13	2	1	1
1XQS	1	14	55	628	1	8564	7
1Z5Y	1	16	320		4	39	17
2CFH	1	1904	202	1394	4066	43	5
2FD6	68	31				81	1
2H7V	1		734		1091		
2HLE	1	13	1	1	2	1	3
2HQS	1	30	2	30	146	146	129
2I25	**1**	**40**	**443**	**1520**	**15**	**948**	**3599**
2O8V	1	60	5	186	220	1	
2QFW	1		19			7	73
Flexible (I-RMSD$_{C\alpha}$ 2.0−3.0 Å) (18 Cases)							
1ACB	**1**	**361**	**144**	**668**	**6**	**4**	**15**
1BKD	2	522	157	1050	99	114	646
1CGI	1	19	98	13	1	12	5
1DE4	1			366			
1E4K	104	1215	722	148	200	4249	74
1EER	**3**	**1821**	**91**	**21**	**81**	**37**	**675**
1I2M	**1**		**683**	**632**	**50**	**149**	**247**
1IB1	34		2116	7028	255	2775	1626
1IBR	**1**						
1KKL	88	49	271	176	1	2	289
1M10	1	81	5742	574		21	2873
1N2C	1					16	
1PXV	**1**	**2073**	**100**	**429**	**673**	**1498**	**2375**
2C0L	83	3958	1024	1589		5105	3834
2HMI	2						
2HRK	49	16	23	47	83	83	241
2NZ8	1	10	5509	247	2	168	5848
2OT3	1	5	212	14	91		131

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 1. continued

PDB	bound	unbound	Cα-RMSD	Int-RMSD	bound BE	unbound BE	clashes
Highly Flexible (I-RMSD$_{Cα}$ > 3.0 Å) (8 Cases)							
1ATN	7	**2568**				**665**	
1FAK	41	5327			43	41	
1FQ1	6	3865	4315	7901	927		4833
1H1V	537						
1IRA	1						
1JMO	1	5325		5398	2969	5510	5547
1R8S	1					4043	
1Y64	1420					1329	

[a]For each case, the best rank of any near-native docking solution is shown, when docking the bound and unbound structures of the interacting proteins, or the best conformers selected from the MODELER ensembles according to the different criteria, based on the native orientation, as described in the main text. In bold are shown the high-affinity cases.

tional rearrangement of the interacting proteins upon binding (see Methods). The largest docking improvement when using the selected conformers is observed in the low- and medium-flexible cases, i.e. those with I-RMSD$_{Cα}$ between 0.5 and 2.0 Å (Figure 4B). The ensemble success rates are particularly good in the low-flexible cases, for which they reach predictive docking values similar to the optimal ones when using the bound structures. This could be related to the limited sampling used here, which did not explore too far from the unbound (1.2 Å of Int-RMSD as average) and therefore they can only deeply explore the vicinity of the bound state in low-flexible or rigid cases. Indeed, in the rigid cases (I-RMSD$_{Cα}$ < 0.5 Å), the selected conformers yield similar results to the unbound structures. In these cases, unbound structures already produced optimal results, similar to the optimal success rates obtained when using the bound structures. In flexible or highly flexible cases (I-RMSD$_{Cα}$ > 2.0 Å), the docking results for the ensembles are as poor as those for the unbound structures, very far from the optimal success rates when using the bound structures. Using MD or more conformers does not significantly change the results (Figure S5).

We noted that the results of bound docking are not as good as one would expect, mostly due to the above-mentioned low-resolution FFT-based discrete searching algorithm. In this method, proteins are represented in 3D grids with 0.7 Å of grid cell size, and thus the exact atomic positions of the interacting proteins are not explicitly considered during the simulations. Rotations of the ligand protein are also discretely sampled, which makes it unlikely to find the exact native orientation. This would be particularly critical in low-affinity cases, in which the small number of interactions would make them less tolerant to small errors in the atomic positions. To minimize the impact of this technical limitation in our evaluation, we have performed the same analysis as above but focusing only on the 28 cases of the benchmark that have been experimentally defined as high-affinity ($\Delta G < -12.0$ kcal/mol), for which the results of bound docking are close to optimal (Figure 4C). Under these conditions, we can observe even more clearly that the selected conformers largely improved the docking success rates in the low-flexible cases. This analysis helps to explain the observed general improvement in docking when using the optimal conformers and aims to be useful for future development of a practical docking protocol. Given that it would be very difficult to identify *a priori* which cases can be more benefitted from this approach, any future docking protocol using this strategy should be of general applicability to all cases.

DISCUSSION

Conformers Providing Better Binding Energy in the Native Orientation Are More Likely To Improve Docking. We have shown that a set of discrete conformers representing the conformational heterogeneity of the unbound structure yielded better docking results than the unbound structures alone. It would be important to analyze the reasons for the success of such conformers. Surprisingly, the conformers that were structurally more similar to the reference structures did not yield better docking results than the unbound structures. More interestingly, selected conformers with the best binding energy in the native orientation yielded better docking results than the unbound structures. Thus, the capacity to provide favorable binding energy in the native orientation seems to be a major determinant for the success of docking, as opposed to the criterion of structural similarity to the native conformation. This might be due to the fact that in the majority of our cases, ensembles are not exploring the conformational space close to the bound state, because sampling is limited to a region in the vicinity of the unbound.

Figure 5A shows, for each case, the best ranked near-native solution obtained when docking the conformers that had the best native-oriented binding energy with the bound partner (i.e., best near-native rank in ordinates; average native-oriented energy of best pair of conformers in abscissas). As we can see in Figure 5A, 90% of the successful cases (i.e., near-native solution ranked within top 10) have an average conformer binding energy < −20.0 au in the native orientation. Actually, 71% of the docking cases with conformers with binding energy in the native orientation < −40.0 au were successful. This confirms that the existence of conformers with good optimal energy in the native orientation is determining the success of docking. Figure 5A highlights the cases that significantly improved, i.e. which had a near-native ranked ≤10 when using the energy-based selected conformers but not when using the unbound structures. In many of these cases, the unbound structures in the native orientation had binding energy < −20.0 au (Figure 5B) but were not successful in unbound docking. In these cases, a minimal amount of conformational sampling seems to be sufficient to generate conformers that significantly improve the docking results.

Ensembles in Docking: Does Size Really Matter? For a practical use in docking, the conformational ensembles should provide a reasonable coverage of the conformational space, using a minimal number of conformers. We have shown here that the selected conformers (based on the reference complex structure) from the 1,000-member ensembles generated by

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

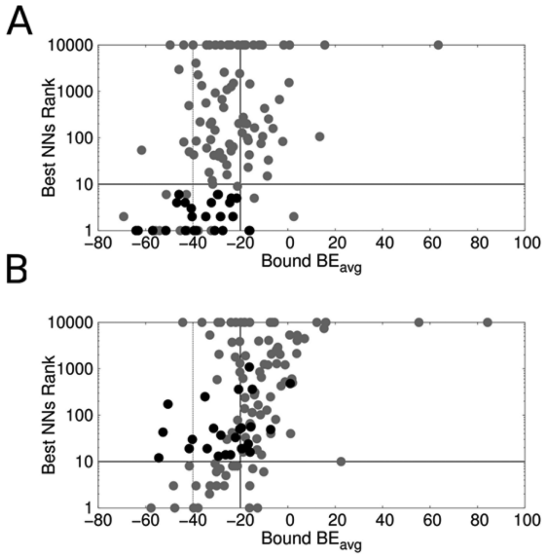


Figure 5. Docking performance dependence on energetic complementarity of the docking partners. Best rank of any near-native docking solution vs average native-oriented pyDock binding energy (au) toward the bound partner calculated for (A) best pair of conformers according to binding energy toward the bound state and (B) unbound X-ray structures. Highlighted in black are the cases that largely improve docking performance (from near-native rank >10 to rank ≤10) using the energy-based selected conformers.

MODELER or MD yielded similar results to those selected from the 100-member ensembles (Figure S4). This suggests that the "extra" 900 conformers are exploring the same conformational region and that the initial ensembles of 100 conformers already achieved convergence and provided reasonable sampling, at least at the conditions of our experiment. Perhaps the resolution of the docking or the energy-based scoring is not sufficient to appreciate subtle changes in the conformational ensembles, so adding more models, without changing the conditions of the simulation, is not going to help. This is exactly what happens in the more rigid cases, in which a larger conformational ensemble does not seem to help to find better conformers to improve the docking results. However, we can observe a small improvement in the flexible cases when using the larger ensembles (Figure S5). Perhaps, in addition to larger ensembles, higher conformational variability would be needed in order to see further improvement in the flexible cases. In this sense, we have performed longer MD simulations (100 ns), at different temperatures (300 and 340 K), on a random selection of 11 cases with no missing long loops (comprising all ranges of flexibility values). The 1,000-member ensembles from these extended MD simulations showed larger conformational variability as compared to the shorter simulations. However, these larger ensembles did not increase the number of cases with conformers significantly more similar to the bound structure, neither provided better docking success rates (S1 Table). Given the known convergence issues in MD,[63] it seems that more exhaustive sampling of the unbound conformational space is needed for most of the cases. This could be achieved by much longer MD trajectories, multiple MD runs in parallel, or enhanced sampling methods like metadynamics,[64] replica-exchange,[65] or MD with excited normal modes.[66] Future work on ensemble docking would need to explore the use of these enhanced ensembles.

Binding Mechanism: What Can We Learn from Docking? The different possible mechanisms that have been proposed for protein−protein association could be described by existing computational approaches. In this context, we can consider several possible scenarios. For protein complexes following a rigid association (similar to "lock-and-key" mechanism), the use of rigid-body docking with the unbound subunits could be a suitable approach to describe the binding process and obtain good predictive models. Indeed, this seems to be the case for complexes with small conformational changes between the unbound and bound states (I-RMSD$_{C\alpha}$ < 0.5 Å), in which unbound docking already gives similar success rates as bound docking (Figures 4B and 4C). In these rigid cases, the optimal conformers from the unbound ensembles also yielded similarly good docking rates as those obtained with the unbound and bound structures. Indeed, Figure 6 shows one

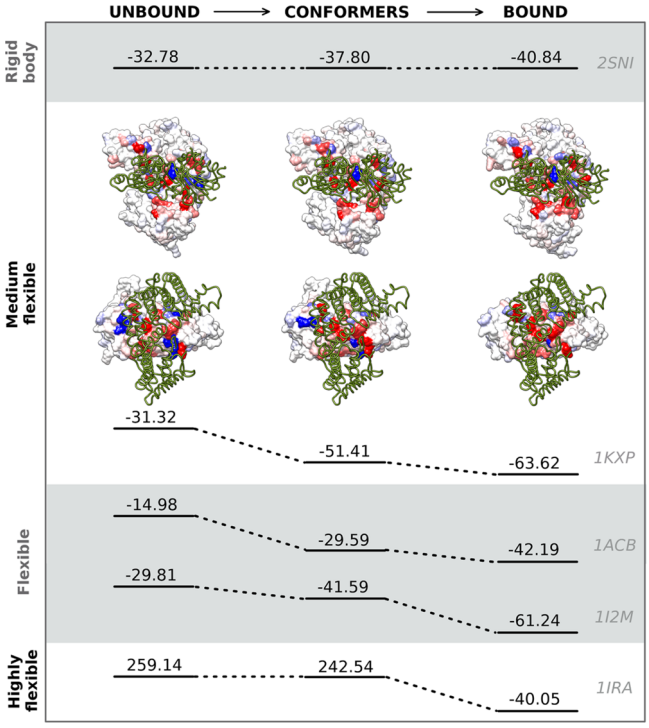


Figure 6. Average pyDock binding energy (au) of the unbound, best conformer according to native-oriented bound energy and bound conformations, in the native orientation, with the bound structure. A few examples with different degrees of unbound-to-bound conformational changes are shown. Similar binding energies for the unbound, conformer and bound structures suggest a lock-and-key binding mechanism (as in 2SNI). An average binding energy for conformers better than that of unbound structures and similar to that of bound structures suggests a conformational selection model (as in 1KXP, 1ACB). Conformer binding energy similar to that of the unbound conformation and worse than that of the bound structure could be compatible with conformational selection (1I2M; see main text) or induced fit mechanisms (1IRA).

example of rigid-body docking (2SNI) in which the unbound proteins in the native orientation showed good average binding energy toward the bound partner (−32.8 au), not far from that of the bound structures (−40.8 au). Consistently, the average binding energy of the best conformers were similar to that of the unbound or bound pairs (−37.8 au). However, when conformers were selected by criteria of structural similarity to bound state, docking success rates were much worse than

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Table 2. Docking Performance of Conformers Selected from NMA-Based Ensembles

PDB	bound	unbound	MM (100 conf)	MM (1,000 conf)	NMA (100 conf) * MM (1 conf)	NMA (100 conf) * MM (10 conf)	NMA (1,000 conf) * MM (1 conf)
1ACB	1	361	4[d]	34[d]	46[c]	3[d]	1[d]
1ATN	7	2568	665[d]		292[c]	3245[a]	1788[a]
1EER	3	1821	21[b]	13[c]	17[d]	3[c]	3[b]
1I2M	1		50[c]	13[c]	23[c]	1[c,d]	1[d]
1IBR	1			1108[e]	87[a]	146[e]	88[a]
1PXV	1	2073	100[a]	822[c]	168[d]	168[d]	232[d]

[a]Cα global RMSD. [b]Full-atom interface RMSD. [c]Native-oriented binding energy with bound partner. [d]Native-oriented binding energy with unbound partner. [e]Number of clashes with bound partner in the native orientation.

unbound or bound docking, because in these cases conformational heterogeneity is more likely to produce conformers that are further from the bound state than the unbound one (given that the unbound was already close to the bound state). Indeed, in none of these cases there were a single conformer that was significantly closer (in terms of Int-RMSD) to the bound state than the unbound structure.

On the other side, we know that in complexes involving flexible association, rigid-body docking with the unbound structures is not going to produce correct models. For such cases, different binding mechanisms have been proposed, such as conformational selection or induced fit. For cases following the conformational selection mechanism, the hypothesis is that the unbound proteins naturally sample a variety of conformational states, a subset of which are suitable to bind the other protein. Therefore, for these cases the use of precomputed unbound ensembles describing the conformational variability of the free proteins in solution should generate conformers that would improve the rigid-body docking predictions with respect to those with the unbound structures. Indeed, this is the case for the complexes undergoing unbound to bound transitions between 0.5 and 1.0 Å I-RMSD$_{C\alpha}$. In these cases, selected conformers from the unbound ensembles yielded much better docking predictions than the unbound structures, virtually achieving the success of bound docking (Figure 4B). For cases undergoing unbound-to-bound transition between 1.0 and 2.0 Å I-RMSD$_{C\alpha}$, the use of unbound ensembles also improved the predictions with respect to the unbound docking results, although to a lesser extent (Figure 4B). Figure 6 shows one of these cases, 1KXP, in which the binding energy of the selected pair of conformers in the native orientation (−51.4 au) is better than that of the unbound structures (−31.3 au) and similar to that of the bound structures (−63.6 au). Some residues in the best pair of conformers show better energy contribution than in the unbound state, which explains why this specific pair of conformers improves docking results. In these cases, the existence of a subpopulation of "active" conformers, i.e. with good binding capabilities toward the bound partner, would be consistent with a conformational selection mechanism. The fact that these conformers with improved binding capabilities are not geometrically closer to the bound state seems counterintuitive. However, recent views of binding mechanism suggest that active conformers that are selected by the partner (initial encounters) do not necessarily need to be in the bound state, as they can adjust their conformations during the association process.[35] Our docking poses are likely to represent these initial encounters between the most populated conformational states of the interacting proteins and would be compatible with this extended conformational selection view.[35] However, in other cases the limited conformational sampling used here might not be sufficient to explore all conformational states available in solution, and therefore the specific binding mechanism cannot be easily identified.

On the other extreme, in cases following an "induced-fit" mechanism the bound complexes would only be obtained after rearrangement of the interfaces when interacting proteins are approaching to each other, in which case the use of precomputed conformational ensembles in docking (even if generated by exhaustive sampling) would not produce favorable encounters around the native complex structure. This seems the case for complexes undergoing unbound to bound transitions above 3.0 Å I-RMSD$_{C\alpha}$. In all these cases, rigid-body docking, either with unbound structures or with selected conformers, fails to reproduce the experimental complex structure. Figure 6 shows one of these highly flexible cases, 1IRA, in which the binding energy of the selected pair of conformers is similar to that of the unbound structures and much worse than that of the bound conformation. For these complexes, the use of precomputed unbound ensembles does not seem to be advantageous, and they would probably need to include flexibility during the docking search, mimicking the induced fit mechanism. However, in the flexible category (i.e., unbound to bound transitions between 2.0 and 3.0 Å I-RMSD$_{C\alpha}$), there are cases like 1ACB, which seem to follow the (extended) conformational selection mechanism, since the use of conformers helps to improve the docking results, and the conformers show better energy than the unbound structures (Figure 6). Again, there might be other complexes under this category that could still follow the conformational selection mechanism, but for which our conformational search is perhaps not sufficient to sample the productive conformations that may exist in solution. This seems to be the case of 1I2M, in which the ensembles based on MODELER did not produce any pair of conformers with sufficiently good binding energy in the native orientation (Figure 6), but the use of extended sampling based on NMA was able to improve the docking rates (see later).

Obviously, the use of docking calculations to learn about the binding mechanism has additional limitations. The time scale of transitions between inactive and active conformers can play an important role in controlling the binding mechanism.[67] In the present work, we can only assume that our ensembles are formed by the most populated conformers in solution, so the existence of active conformers that can be preferentially selected by the bound partner would be compatible (but not exclusively) with a mainly conformational selection mechanism. However, in a situation in which the active conformers are not highly populated, as those that would need extended sampling to be identified, we could not define the type of mechanism unless transition rates between conformers are considered.

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

Future Perspectives: Improving Sampling with Normal-Mode Analysis for Flexible Cases. We have shown here that cases with large conformational changes after binding (I-RMSD$_{C\alpha}$ > 2 Å) do not generally benefit from the use of conformers from unbound ensembles generated by MODELER or short MD simulations. These complexes could follow the induced fit binding mechanism, in which the use of precomputed unbound ensembles would not be appropriate to describe association. However, we should not disregard that some of these complexes could still follow a conformational selection mechanism, but for some reason a dramatically larger conformational sampling would be needed to find suitable conformers. One way to extend conformational sampling is by using Normal Mode Analysis (NMA). When generating 100 conformers for this group of cases (strong and flexible I-RMSD$_{C\alpha}$ > 2.0 Å) with an *ad-hoc* Monte Carlo sampling method based on Cα NMA and full-atom rebuilding with MODELER (Figure S6; see Methods), the results were not better than those obtained with the conformers directly generated by MODELER (Tables 2 and S2). However, when generating 1,000 conformers based on NMA (either formed by 1,000 NMA-based conformers rebuilt with MODELER, or by 100 NMA-based conformers with 10 MODELER models for each of them), the success rates largely improved with respect to those obtained when generating conformers only with MODELER (either 100 or 1,000 conformers). It is interesting to analyze the flexible case 1I2M, in which docking failed when using the unbound structure or the best conformers from either MODELER or MD ensembles, but yielded successful results with the 1,000-member NMA-based ensembles. This shows that new sampling approaches based on NMA could produce the type of enhanced sampling needed for the most flexible cases that follow a conformational selection mechanism. Our findings could help to develop future strategies to integrate NMA-sampling in a practical docking protocol, but this would need extensive evaluation on the entire benchmark (including cases in which the conformational difference between bound and unbound structures is small) and algorithmic optimization, which is beyond the scope of current work.

CONCLUSIONS

We present here the most complete systematic study so far about the potential capabilities of using precomputed unbound ensembles in docking. The results show that considering conformational heterogeneity in the unbound state of the interacting proteins can improve their binding capabilities in cases of moderate unbound-to-bound mobility. In these cases, the existence of conformers with better binding energy in the native orientation is associated with a significant improvement in the docking predictions. It seems that protein plasticity increases the chances of finding conformations with better binding energy capabilities, not necessarily related to similar-to-bound geometries, which is compatible with the extended conformational selection mechanism. Other moderately flexible cases have conformers that look promising from a binding energy perspective but do not provide good docking predictions. These cases could also follow a conformational selection mechanism, but they would need extensive sampling to find suitable conformers for binding. The most flexible cases would show larger induced fit effects and therefore would not be well described by ensemble binding. In a realistic scenario, optimal conformers would not be easy to identify *a priori*, and, as a consequence, new ways of efficiently including all conformers in a docking protocol should be devised. This work helps to set guidelines for future strategies in practical docking predictions based on unbound ensembles generated by molecular mechanics minimization.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.6b00204.

Figure S1: Distribution of conformers according to different quality criteria in the 100-member ensembles. Figure S2: Representative conformational ensembles generated by 10 ns MD simulations. Figure S3: Distribution of conformers according to different quality criteria in the 1,000-member ensembles. Figure S4: Docking performance for the best conformers of different ensembles. Figure S5: Docking performance for the different ensembles according to unbound-to-bound variability. Figure S6: Representative conformational ensembles generated by NMA-based sampling. Table S1: Docking performance (best rank of any near-native solution) with conformers selected from extended MD ensembles. Table S2: Docking performance of conformers selected from NMA-based ensembles (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: juanf@bsc.es.

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by grant BIO2013-48213-R from the Spanish Ministry of Economy and Competitiveness. R.A. was partially supported by NIH (R01 GM071872) and is a founder and shareholder in Molsoft, LLC.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Dr. Iain Moal for the extended molecular dynamics trajectories.

ABBREVIATIONS

MD, molecular dynamics; NMA, normal modes analysis; eNMA, enhanced normal modes analysis; RMSD, root-mean-square deviations; CAPRI, Critical Assessment of PRediction of Interactions; NMR, nuclear magnetic resonance; RDC, Residual Dipolar Coupling; MM, modeling minimization; BE, binding energy; FFT, Fast Fourier Transform

REFERENCES

(1) Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **1998**, *92*, 291−4.

(2) Rual, J. F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G. F.; Gibbons, F. D.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem, M.; Milstein, S.; Rosenberg, J.; Goldberg, D. S.; Zhang, L. V.; Wong, S. L.; Franklin, G.; Li, S.; Albala, J. S.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, R. S.; Vandenhaute, J.; Zoghbi, H. Y.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, M.

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

E.; Hill, D. E.; Roth, F. P.; Vidal, M. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **2005**, *437*, 1173−8.
(3) Stelzl, U.; Worm, U.; Lalowski, M.; Haenig, C.; Brembeck, F. H.; Goehler, H.; Stroedicke, M.; Zenkner, M.; Schoenherr, A.; Koeppen, S.; Timm, J.; Mintzlaff, S.; Abraham, C.; Bock, N.; Kietzmann, S.; Goedde, A.; Toksoz, E.; Droege, A.; Krobitsch, S.; Korn, B.; Birchmeier, W.; Lehrach, H.; Wanker, E. E. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **2005**, *122*, 957−68.
(4) Mosca, R.; Ceol, A.; Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **2013**, *10*, 47−53.
(5) Szilagyi, A.; Zhang, Y. Template-based structure modeling of protein-protein interactions. *Curr. Opin. Struct. Biol.* **2014**, *24*, 10−23.
(6) Kundrotas, P. J.; Zhu, Z.; Janin, J.; Vakser, I. A. Templates are available to model nearly all complexes of structurally characterized proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 9438−41.
(7) Kundrotas, P. J.; Vakser, I. A. Global and local structural similarity in protein-protein complexes: implications for template-based docking. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 2137−42.
(8) Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 2195−9.
(9) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* **1997**, *272*, 106−20.
(10) Kozakov, D.; Brenke, R.; Comeau, S. R.; Vajda, S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins: Struct., Funct., Genet.* **2006**, *65*, 392−406.
(11) Tovchigrechko, A.; Vakser, I. A. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* **2006**, *34* (WebServer issue), W310−W314.
(12) Chen, R.; Li, L.; Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 80−7.
(13) Ritchie, D. W.; Kemp, G. J. Protein docking using spherical polar Fourier correlations. *Proteins: Struct., Funct., Genet.* **2000**, *39*, 178−194.
(14) Garzon, J. I.; Lopez-Blanco, J. R.; Pons, C.; Kovacs, J.; Abagyan, R.; Fernandez-Recio, J.; Chacon, P. FRODOCK: a new approach for fast rotational protein-protein docking. *Bioinformatics* **2009**, *25*, 2544−51.
(15) Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* **2005**, *33* (Web Server issue), W363−W367.
(16) Dominguez, C.; Boelens, R.; Bonvin, A. M. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **2003**, *125*, 1731−7.
(17) Lyskov, S.; Gray, J. J. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.* **2008**, *36* (WebServer issue), W233−W238.
(18) Fernandez-Recio, J.; Totrov, M.; Abagyan, R. Soft protein-protein docking in internal coordinates. *Protein Sci.* **2002**, *11*, 280−91.
(19) Cheng, T. M.; Blundell, T. L.; Fernandez-Recio, J. pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins: Struct., Funct., Genet.* **2007**, *68*, 503−15.
(20) Pons, C.; Talavera, D.; de la Cruz, X.; Orozco, M.; Fernandez-Recio, J. Scoring by intermolecular pairwise propensities of exposed residues (SIPPER): a new efficient potential for protein-protein docking. *J. Chem. Inf. Model.* **2011**, *51*, 370−7.
(21) Ravikant, D. V.; Elber, R. PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 400−19.
(22) Lensink, M. F.; Wodak, S. J. Docking, scoring, and affinity prediction in CAPRI. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 2082−95.
(23) Pons, C.; Grosdidier, S.; Solernou, A.; Perez-Cano, L.; Fernandez-Recio, J. Present and future challenges and limitations in protein-protein docking. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 95−108.
(24) Pallara, C.; Jimenez-Garcia, B.; Perez-Cano, L.; Romero-Durana, M.; Solernou, A.; Grosdidier, S.; Pons, C.; Moal, I. H.; Fernandez-Recio, J. Expanding the frontiers of protein-protein modeling: from docking and scoring to binding affinity predictions and other challenges. *Proteins: Struct., Funct., Genet.* **2013**, *81*, 2192−200.
(25) Fernandez-Recio, J.; Totrov, M.; Abagyan, R. ICM-DISCO docking by global energy optimization with fully flexible side-chains. *Proteins: Struct., Funct., Genet.* **2003**, *52*, 113−7.
(26) Mashiach, E.; Nussinov, R.; Wolfson, H. J. FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1503−19.
(27) Zacharias, M. Protein-protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci.* **2003**, *12*, 1271−82.
(28) Zacharias, M. Rapid protein-ligand docking using soft modes from molecular dynamics simulations to account for protein deformability: binding of FK506 to FKBP. *Proteins: Struct., Funct., Genet.* **2004**, *54*, 759−67.
(29) Moal, I. H.; Bates, P. A. SwarmDock and the use of normal modes in protein-protein docking. *Int. J. Mol. Sci.* **2010**, *11*, 3623−48.
(30) Straub, F. B.; Szabolcsi, G. O dinamicseszkij aszpektah sztukturü fermentov (On the dynamic aspects of protein structure). In *Molecular Biology, Problems and Perspectives*; Braunstein, A. E., Ed.; Izdat. Nauka: Moscow, 1964; pp 182−187.
(31) Závodszky, P.; Abaturov, L. V.; Varshavsky, Y. M. Structure of glyceraldehyde-3-phosphate dehydrogenase and its alteration by coenzyme binding. *Acta Biochim. Biophys. Acad. Sci. Hung.* **1966**, *1*, 389−403.
(32) Foote, J.; Milstein, C. Conformational isomerism and the diversity of antibodies. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 10370−4.
(33) Tsai, C. J.; Ma, B.; Nussinov, R. Folding and binding cascades: shifts in energy landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9970−2.
(34) Ma, B.; Kumar, S.; Tsai, C. J.; Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng., Des. Sel.* **1999**, *12*, 713−20.
(35) 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), 539−46.
(36) Stein, A.; Rueda, M.; Panjkovich, A.; Orozco, M.; Aloy, P. A systematic study of the energetics involved in structural changes upon association and connectivity in protein interaction networks. *Structure* **2011**, *19*, 881−9.
(37) Boehr, D. D.; Nussinov, R.; Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5*, 789−96.
(38) Fraser, J. S.; Clarkson, M. W.; Degnan, S. C.; Erion, R.; Kern, D.; Alber, T. Hidden alternative structures of proline isomerase essential for catalysis. *Nature* **2009**, *462*, 669−73.
(39) Volkman, B. F.; Lipson, D.; Wemmer, D. E.; Kern, D. Two-state allosteric behavior in a single-domain signaling protein. *Science* **2001**, *291*, 2429−2433.
(40) Moal, I. H.; Bates, P. A. Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. *PLoS Comput. Biol.* **2012**, *8*, e1002351.
(41) Bastard, K.; Thureau, A.; Lavery, R.; Prevost, C. Docking macromolecules with flexible segments. *J. Comput. Chem.* **2003**, *24*, 1910−20.
(42) Bastard, K.; Prevost, C.; Zacharias, M. Accounting for loop flexibility during protein-protein docking. *Proteins: Struct., Funct., Genet.* **2006**, *62*, 956−69.
(43) van Dijk, M.; van Dijk, A. D.; Hsu, V.; Boelens, R.; Bonvin, A. M. Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res.* **2006**, *34*, 3317−3325.

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(44) Grunberg, R.; Leckner, J.; Nilges, M. Complementarity of structure ensembles in protein-protein binding. *Structure* **2004**, *12*, 2125−36.
(45) Smith, G. R.; Sternberg, M. J.; Bates, P. A. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *J. Mol. Biol.* **2005**, *347*, 1077−101.
(46) Chaudhury, S.; Gray, J. J. Conformer selection and induced fit in flexible backbone protein-protein docking using computational and NMR ensembles. *J. Mol. Biol.* **2008**, *381*, 1068−87.
(47) Pons, C.; Fenwick, R. B.; Esteban-Martín, S.; Salvatella, X.; Fernandez-Recio, J. Validated Conformational Ensembles Are Key for the Successful Prediction of Protein Complexes. *J. Chem. Theory Comput.* **2013**, *9*, 1830−1837.
(48) Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M. S.; Eramian, D.; Shen, M. Y.; Pieper, U.; Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc Bioinformatics*; 2006; Chapter 5, Unit 5.6.10.1002/0471250953.bi0506s15
(49) Cheatham, T. E., 3rd; Cieplak, P.; Kollman, P. A. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845−862.
(50) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668−88.
(51) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Perez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpi, J. L.; Orozco, M. MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure* **2010**, *18*, 1399−409.
(52) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247−260.
(53) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905−1908.
(54) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* **2001**, *80*, 505−15.
(55) Kovacs, J. A.; Chacon, P.; Abagyan, R. Predictions of protein flexibility: first-order measures. *Proteins: Struct., Funct., Genet.* **2004**, *56*, 661−8.
(56) Hinsen, K. Analysis of domain motions by approximate normal mode calculations. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 417−429.
(57) Rueda, M.; Chacon, P.; Orozco, M. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure* **2007**, *15*, 565−75.
(58) Sali, A.; Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **1993**, *234*, 779−815.
(59) Hwang, H.; Pierce, B.; Mintseris, J.; Janin, J.; Weng, Z. Protein-protein docking benchmark version 3.0. *Proteins: Struct., Funct., Genet.* **2008**, *73*, 705−9.
(60) Abagyan, R.; Lee, W. H.; Raush, E.; Budagyan, L.; Totrov, M.; Sundstrom, M.; Marsden, B. D. Disseminating structural genomics data to the public: from a data dump to an animated story. *Trends Biochem. Sci.* **2006**, *31*, 76−8.
(61) Garbuzynskiy, S. O.; Melnik, B. S.; Lobanov, M. Y.; Finkelstein, A. V.; Galzitskaya, O. V. Comparison of X-ray and NMR structures: is there a systematic difference in residue contacts between X-ray and NMR-resolved protein structures? *Proteins: Struct., Funct., Genet.* **2005**, *60*, 139−147.
(62) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Genet.* **2010**, *78*, 1950−1958.
(63) Lyman, E.; Zuckerman, D. M. On the structural convergence of biomolecular simulations by determination of the effective sample size. *J. Phys. Chem. B* **2007**, *111*, 12876−82.
(64) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.
(65) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141−151.
(66) Costa, M. G. S.; Batista, P. R.; Bisch, P. M.; Perahia, D. Exploring free energy landscapes of large conformational changes: molecular dynamics with excited normal modes. *J. Chem. Theory Comput.* **2015**, *11*, 2755−2767.
(67) Zhou, H. X. From induced fit to conformational selection: a continuum of binding mechanism controlled by the timescale of conformational transitions. *Biophys. J.* **2010**, *98* (6), L15−7.

DOI: 10.1021/acs.jctc.6b00204
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX