Attracting Cavities 2.0: Improving the Flexibility and Robustness for Small-Molecule Docking

Ute F. Röhrig,* Mathilde Goullieux, Marine Bugnon, and Vincent Zoete*

Cite This: *J. Chem. Inf. Model.* 2023, 63, 3925−3940

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Molecular docking is a computational approach for predicting the most probable position of a ligand in the binding site of a target macromolecule. Our docking algorithm Attracting Cavities (AC) has been shown to compare favorably to other widely used docking algorithms [Zoete, V.; et al. *J. Comput. Chem.* **2016**, *37*, 437]. Here we describe several improvements of AC, making the sampling more robust and providing more flexibility for either fast or high-accuracy docking. We benchmark the performance of AC 2.0 using the 285 complexes of the PDBbind Core set, version 2016. For redocking from randomized ligand conformations, AC 2.0 reaches a success rate of 73.3%, compared to 63.9% for GOLD and 58.0% for AutoDock Vina. Due to its force-field-based scoring function and its thorough sampling procedure, AC 2.0 also performs well for blind docking on the entire receptor surface. The accuracy of its scoring function allows for the detection of problematic experimental structures in the benchmark set. For cross-docking, the AC 2.0 success rate is about 30% lower than for redocking (42.5%), similar to GOLD (42.8%) and better than AutoDock Vina (33.1%), and it can be improved by an informed choice of flexible protein residues. For selected targets with a high success rate in cross-docking, AC 2.0 also achieves good enrichment factors in virtual screening.
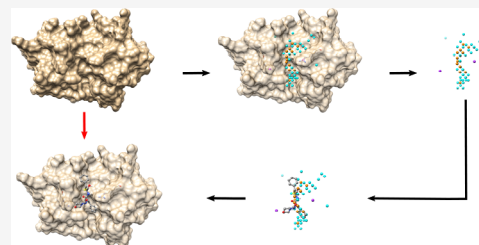
INTRODUCTION

Molecular docking is a computational approach for predicting the most probable binding mode of a small molecule to a macromolecular target, most commonly a protein but possibly also a DNA or RNA. Docking algorithms, which predict possible structures for ligand−target complexes and usually also estimate the corresponding binding affinities, constitute the cornerstone of structure-based computer-aided drug design. A docking algorithm generally consists of a sampling algorithm, which generates putative ligand binding modes, and a scoring function, which evaluates and ranks them.

The idea behind Attracting Cavities (AC)[1] is to replace the rough energy landscape of the macromolecule by a smooth attractive energy landscape generated by virtual attracting points surrounding the macromolecular surface (Figure 1). We demonstrated that simple rotations, translations, and geometry optimizations of the ligand in this smooth landscape are an efficient sampling algorithm for docking. These initial optimizations in the "mold" of the protein are followed by optimizations in the actual protein energy landscape and an implicit solvation treatment. The scoring function of AC is composed of the CHARMM force field[2−5] terms and the fast analytical continuum treatment of solvation (FACTS) model.[6] The use of this universal scoring ensures the applicability of the docking algorithm for diverse types of macromolecular targets (e.g., proteins, RNA, and DNA) and for diverse types of ligands (e.g., drug-like molecules, molecular fragments, and peptides).

The original AC algorithm, a Python code piloting the CHARMM[7] molecular mechanics program, was benchmarked on the Astex Diverse Set[8] of 85 non-covalent ligand−protein complexes. It reached a success rate of up to 84% for reproducing the native binding conformation of the ligand on its target within 2 Å root-mean-square deviation (RMSD) starting from a randomized conformer.[1]

Here we describe an update of the algorithm and the implementation of new features. The update comprises (1) the possibility to choose between the CHARMM22/27[2,3] or CHARMM36[4,5] force field, (2) the shared-memory parallelization of the code, (3) the free choice of initial ligand rotation angle values and removal of duplicate poses, (4) a modified definition of electrostatic cloud points, (5) a slightly modified definition of attractive cloud points using a switching function instead of cutoffs, (6) the additional definition of placement cloud points for initial ligand sampling, (7) the randomization of initial conditions to improve the sampling and the robustness of docking results, (8) the possibility to define parts of the receptor as flexible, and (9) the choice of the scoring function.
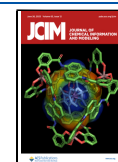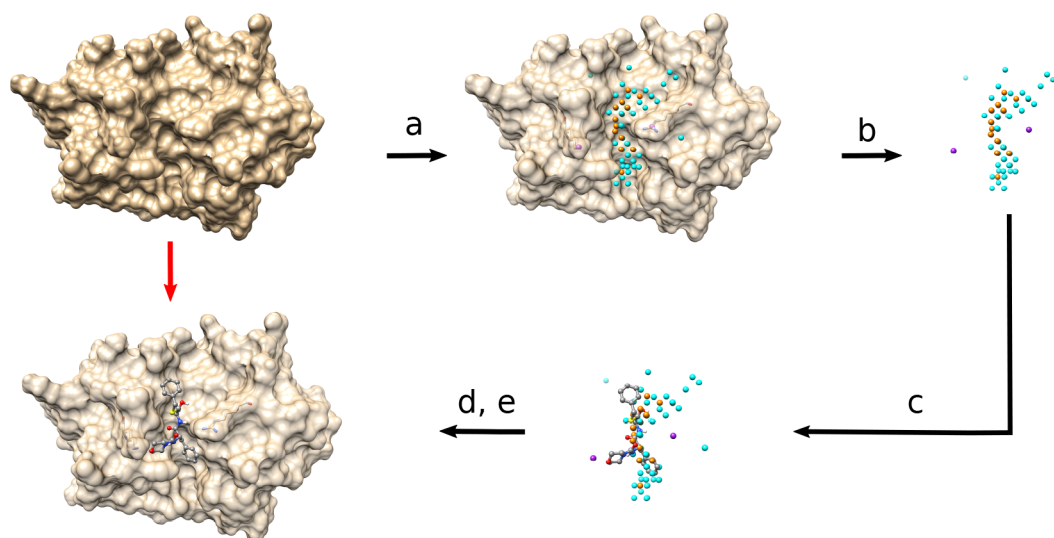
Figure 1. New Attracting Cavities algorithm. (a) Calculation of attracting (cyan), placement (orange), and electrostatic (purple) cloud points. (b) Removal of protein atoms. (c) Docking of the ligand in the "mold" of points (sampling). (d) Removal of cloud points and reintroduction of the protein. (e) Optimization of the ligand in the protein environment (refinement) and scoring including solvation terms.

To benchmark AC 2.0, we use the Core set from the PDBbind resource (http://www.pdbbind.org.cn),[9,10] version 2016, which aims at providing a relatively small set of 285 high-quality protein−ligand complexes for validating docking/scoring methods. The set consists of 57 target clusters with five complex structures each and has served as the primary test set in the Comparative Assessment of Scoring Functions (CASF) benchmark.[10] We chose this benchmark set because of its advertised quality and the availability of the curated starting structures including hydrogen atoms for proteins and ligands. Association constants (K_a values), binding pose decoys both for redocking and for cross-docking/screening, results of different scoring functions, and analysis scripts are also provided. For comparison to the previous version of AC, we also carried out some docking calculations on the Astex Diverse Set of 85 complexes. The popular state-of-the art docking codes Genetic Optimisation for Ligand Docking (GOLD)[11,12] and AutoDock Vina[13,14] were used for comparison with AC. We primarily benchmark the redocking performance of AC, but we also provide data on its scoring, ranking, cross-docking, and virtual screening performance.

METHODS

Update of Sampling Procedure. Using the original AC algorithm[1] (1.0, described in detail in the Supporting Information), we noted several potential points of improvement. For example, in the initial ligand sampling, the ligand was rotated by 90°, 60°, and 45° in the x, y, and z directions, resulting in 64, 216, and 512 poses, respectively. In the updated code, the rotational angle can be freely chosen, and duplicate poses due to symmetry (Table S2) are removed before sampling, so that the above-mentioned angles lead to 24, 108, and 208 poses, respectively, decreasing CPU time. We also noted that the definition of electrostatic cloud points, located in a thin layer of 0.3 Å around the target and calculated on a cubic grid with a step size of 1.5 Å, had two disadvantages: (1) a very high dependence of the exact location of the box center and (2) the potential accumulation of many electrostatic cloud points with similar charges in same regions of the target. We therefore replaced this definition of electrostatic

cloud points by looping over all residues with a net nonzero charge and placing a single bead carrying the same charge in the corresponding region (for details, see the Supporting Information). To reduce the dependence of the exact location of the box center for the attracting cloud points, we introduced smooth switching functions instead of cutoff values for determining the number of protein atoms around a point. We preserved the meaning of the threshold value N_{Thr} for cavity detection as closely as possible, so that a value of 70 detects mainly deep binding clefts, while a value of 50 also places attractive points in shallower protein cavities.[1] In AC 1.0, the attractive cloud points served a twofold purpose: (1) to provide a mold of the target active site and (2) to place the ligand in the initial sampling step. Here we provide the flexibility to separate the two functions and to use placement cloud points for the initial ligand sampling. This modification is based on the observation that attractive cloud points located close to the target surface are likely to cause clashes between the ligand and the target and therefore unlikely to lead to viable ligand poses. Therefore, the placement cloud points are defined analogously to the attractive cloud points, but removing points very close to the protein surface (see the Supporting Information for details). All cloud points (attractive, electrostatic, placement) can now also be predefined following the users' own rules and provided at the start of the docking calculation, so as to allow focusing on specific regions of the target. The use of electrostatic and placement cloud points is optional.

Besides the fixed protein/flexible ligand approach, we have now also implemented the option to leave parts of the receptor flexible, either by defining a radius around the ligand where all protein residues are flexible or by selecting specific residues.

Randomization of Initial Conditions. We observed that docking runs starting from ligand and target coordinates rotated in 3D space but otherwise identical yielded different results in numerous cases (data not shown), implying that the sampling in AC 1.0 was strongly dependent on the exact starting conditions. We therefore implemented a new feature, allowing random modification of the initial ligand rotation and the box center within a distance of −0.5 to 0.5 Å in each

Cartesian direction for the first sampling step. The seed for the random number generator can either be specified in the input file or is automatically generated and printed in the output file, so as to preserve the deterministic docking procedure, which is important for debugging and the testing of parameters. The number of random initial conditions (RIC) to be generated can be defined in the input file.

Parallelization and Acceleration. We implemented a parallel version of AC 2.0 using the `ProcessPoolExecutor` tool of the `concurrent.futures` module of Python, where the CPUs communicate via shared memory. The parallelization procedure can be switched on or off through the input file and necessitates the specification of a number of processors. At each minimization step in the AC algorithm, the total number of poses is then distributed over the number of requested processors. Each processor runs one CHARMM process treating the attributed group of poses. The remaining linear parts of the code were accelerated by reorganization and removal of inefficient I/O operations.

Scoring Function. In AC 2.0, it is possible to choose between CHARMM force field versions 22/27[2,3] and 36.[4,5] To allow for the consistent setup of ligand−target complexes with both force fields, the SwissParam (SP) approach,[15] which is based on the Merck Molecular Force Field,[16] was updated to be compatible also with the CHARMM36 force field. More details on these changes will be given in a separate communication. With the current force field, the treatment of cofactors and post-translation modifications was simplified. Topologies and parameters for the PDBbind Core set were generated with both force field versions, while for the Astex set only the CHARMM22/27 versions were generated. By default, the algorithm uses the AC score, which corresponds to the total energy of the complex (CHARMM force field energy plus FACTS solvation terms). However, it is now also possible to use the SwissParam score,[15] which consists of a weighted sum of polar and nonpolar terms fitted to reproduce the experimental binding free energies of 214 ligand−protein complexes of the Ligand Protein Database (LPDB).[17] The SwissParam score ΔG_{SP} is defined as

$$\Delta G_{SP} = -3.99 + 0.0946 \times (E_{vdW} + \Delta G_{desolv,np})$$
$$+ 0.0612 \times (E_{elec} + \Delta G_{desolv,elec}) \qquad (1)$$

with the van der Waals (E_{vdW}) and Coulomb (E_{elec}) electrostatic interaction energies between the ligand and the protein, and the nonpolar ($\Delta G_{desolv,np}$) and polar ($\Delta G_{desolv,elec}$) desolvation energies of the ligand and protein upon complexation calculated with the FACTS approach.[6]

The AC score has the advantage of being faster to calculate and applicable for a wider range of systems, such as ligand−nucleic acid or ligand−heme protein complexes. The SwissParam score, which can alternatively be calculated during postprocessing and clustering (see below), allows comparison of the scores of different ligands for the same target and therefore opens the possibility for scoring, ranking, and virtual screening. Besides performing a docking run, it is now also possible to provide a collection of poses in pdb or dock4 format and to calculate their scores, either with or without relaxation and with either a fixed or a flexible receptor.

Morse-like Metal Binding Potentials. We previously developed a procedure for treating the attractive interaction between heme cofactors and ligands with potential iron-binding functional groups. By introducing Morse-like metal binding potentials (MMBPs), which were fitted to reproduce density functional theory calculations, we were able to more than double the docking success rates for heme protein complexes.[18] This procedure is now integrated into AC 2.0.

Data Sets. *Astex Diverse Set.* The Astex Diverse set[8] is a set of 85 structures developed for the validation of protein−ligand docking performance with an emphasis on including diverse enzyme classes and diverse and drug-like ligands. Here we used the manually curated structures from our previous studies.[1,19] The overlap between the PDBbind Core set and the Astex Diverse set consists of four complexes (PDB IDs 1gpk, 1oyt, 1z94, and 2br1).

PDBbind Core Set. We downloaded the PDBbind v2016 Core set[10] from the PDBbind website (http://www.pdbbind.org.cn), including the ligand and target structure files in mol2 format, containing all hydrogen atoms and defining the ligand of interest and the protein structure with included cofactors. Regarding the provided target structures, we noted (1) that the hydrogen positions were not optimized, e.g., the neutral Nδ−H tautomer was used for all histidine residues regardless of their environment; (2) that some amino acids presented missing side-chain atoms, which poses a problem for the CHARMM setup; and (3) a few other structural errors, which we corrected (e.g., misinterpretation of *S*-(dimethylarsenic)-cysteine as a calcium ion in the five structures of elongin-B, wrong protonation states, wrong copy of ligand retained). In the provided files, only one copy of the ligand is kept by default, which can cause problems when the docking algorithm finds another copy to be more favorable or when the chosen copy extensively interacts with an absent copy. In the five structures of HIV-1 protease, we systematically protonated the catalytic Asp25 of chain A to be in its neutral form.[20] In order to generate coordinate and topology files for AC, we corrected the detected errors in the target structure, determined all histidine protonation states and tautomeric forms based on their potential hydrogen-bonding partners, reconstructed missing amino acid side chains with the UCSF Chimera program,[21] and reconstructed all hydrogens with the HBUILD command[22] of CHARMM.[7] A short minimization was carried out to remove potential clashes arising from the crystal structure and hydrogen atom placement. During this minimization, 100 steps of steepest descent and 200 steps of adopted basis Newton−Raphson (ABNR),[23] the FACTS solvation model was applied, and all heavy atoms were restrained with a force constant of 5 kcal mol^{-1} Å^{-2} to their original position. Randomized ligand conformations for initializing the docking calculations were generated with Open Babel.[24] Our characterization of the 285 complexes of this benchmark set is given in the Table S3. All files are available on Zenodo (DOI: 10.5281/zenodo.7940100).

X-ray Structure Quality Assessment. As a measure of X-ray structure quality, we calculated the diffraction-component precision index (DPI)[25,26] using the DPICalc algorithm from Mikko J. Vainio (http://users.abo.fi/mivainio/shaep/download.php). The calculations failed for two complexes of the Astex set and for six complexes of the PDBbind Core set.

We also calculated the molecular electron density support for individual atoms (EDIAm) proposed by Meyder et al.[27] to evaluate the structures. It was suggested by the authors that high-quality density maps should yield a ligand EDIAm value of 0.8 or higher. The calculations failed for one complex of the Astex set and six complexes of the PDBbind Core set.

In addition, we assessed for all structures whether the ligand of interest forms any van der Waals or electrostatic contact with a symmetry-related copy of the complex, which would be present in the experimental structure but absent in the docking calculations. To this end, we used the Crystal Contacts tool of UCSF Chimera with a cutoff length of 4.5 Å.

The portion of buried surface area of the ligands was calculated with CHARMM using a probe radius of 1.4 Å.

Docking with AC. We tested and varied many parameters during the docking calculations with AC. However, the orthorhombic search box center was always defined as the center of mass of the ligand in the corresponding X-ray structure. All solvent molecules were removed during setup. For local docking in the active site, the cubic box had an edge length of 20 Å. By default and unless otherwise stated, all docking calculations started from a randomized ligand conformation and were carried out with the CHARMM36 force field[4,5] and the CHARMM program,[7] version 47b1, using an attractive threshold value N_{Thr} of 70, a grid step length of 1 Å, and a rigid protein. The ligand topologies and parameters were generated with the SwissParam approach.[15] By default, we carried out all dockings with the AC scoring function (total energy with FACTS solvation). The number of saved poses depends on their diversity and was maximally 400 (50 clusters of 8 poses each) but about 200 on average. All calculations were performed on an AMD EPYC 7443 3.34 GHz CPU.

Definition of Scoring Failures. For benchmarking the redocking performance of AC with different docking parameters, we removed the complexes presenting scoring failures from the analysis of success rates because for these cases bad sampling might be rewarded with a better success rate. To detect scoring failures, we calculated the AC score for all experimental structures, relaxing the native pose with 200 steps of ABNR and keeping everything fixed besides the ligand. We then defined a scoring failure if the pose with the lowest score found across all docking runs was a failure (RMSD > 3.0 Å) and had a better score than the native pose (relative score < −1.0 kcal/mol). The success criterion was thus slightly relaxed in order to include fewer borderline cases where the best pose is a failure but another pose from the best cluster could be a success (RMSD ≤ 2.0 Å). This definition of a scoring failure has the disadvantage of depending on the number of docking runs but was observed to be relatively stable with respect to the addition of new docking results.

For assessing the overall redocking performance of AC and for comparison with other docking codes, we used the full benchmark sets.

Docking with GOLD. The GOLD docking program[11,12] from the Cambridge Crystallographic Data Centre, version 2022.2.0, was used for comparison with AC. The center of the search space was defined as the center of mass of the ligand, and a radius of 12.4 Å was chosen for the spherical search space in order to give approximately the same volume as a cubic search space with an edge length of 20 Å. Structure files for the protein and the ligand in mol2 format were generated from the respective CHARMM files in pdb format using the GOLD tools `gold_utils`, `conformer_generator`, and `check_mol2`. All files were manually checked for errors and corrected if necessary. The GoldScore (GS),[11] ChemScore (CS),[12] and Piecewise Linear Potential (PLP)[28] fitness functions were used for the benchmark calculations. By default, for each ligand 100 genetic algorithm (GA) runs were carried out without the "early termination" option. With

the PLP function, we also carried out dockings with 10 and 1000 GA runs, terminating each docking when the 16 top solutions had an RMSD below 1.0 Å. All calculations were performed on a single Intel Core i7-11700 2.5 GHz CPU, which in our experience is about 15% slower than the AMD processor on which AC and AutoDock Vina were run.

Docking with AutoDock Vina. The free and open-source docking program AutoDock Vina, version 1.2.3,[13,14] was used for comparison to AC. AutoDock Tools (ADT)[29] was used to generate the structure input files for proteins and ligands in pdbqt format starting from the respective CHARMM input files. The center of the search space was defined as the center of mass of the ligand, and a cubic search space with an edge length of 20 Å was used. For each docking, a maximum of 100 poses were saved (num_modes), setting the energy_range to 100. Different exhaustiveness parameters were tested (8, 100, 1000). Only the Vina score was used, as it has been shown to be faster and to perform substantially better than the AutoDock4 scoring function also available in Vina.[14] All calculations were performed on a single AMD EPYC 7443 3.34 GHz CPU.

Cross-Docking and Screening. The PDBbind Core set consists of 57 protein targets with five ligands binding to each. To evaluate the cross-docking performance of AC, GOLD, and AutoDock Vina, we docked all five ligands of each protein into the target structure of the highest-affinity ligand, as it was done for the decoy poses provided by the developers.[10] We superimposed the other four complexes with this complex structure in order to obtain the reference pose of each ligand, which was then used for RMSD calculation and success rate determination. It should be noted that among the provided 100 decoy poses of the five ligands for each target, 20% of cases do not include a pose within 2 Å of the native pose.

For screening, we docked all 285 molecules against five selected targets, namely β-trypsin (PDB ID 1o3f), U-plasminogen activator (1sqa), 3-dehydroquinate dehydratase (2xb8), catechol o-methyltransferase (3nw9), and MTA nucleosidase (4f3c). Enrichment factors were calculated according to ref 10.

We also performed cross-docking and screening with AC to the heme-free apo form of the anticancer target indoleamine 2,3-dioxygenase 1 (IDO1), based on the 13 complexes available in the PDB (PDB IDs 6azv, 6azw, 6e43, 6v52, 6wjy, 6wpe, 6x5y, 7b1o, 7m63, 7rrb, 7rrc, 7rrd, and 8i7l) and using the protein structure of 6wjy as a target. For all screenings with AC 2.0, we used the AC scoring function for docking and rescored the results with the SwissParam scoring function (eq 1).

Docking Success Criteria. In a postprocessing step, we calculated the symmetry-corrected RMSD values between all ligand poses and between the poses and the X-ray structure using the free and open-source DockRMSD approach.[30] As demonstrated by the authors, this algorithm outperforms similar ones especially for ligand molecules with complicated structural symmetry. All poses were ordered by their scores, which can be the same as used during the docking or different, depending on the application. For redocking and cross-docking, we used the AC score for both docking and postprocessing, while for screening we used the SwissParam score in the postprocessing step. The top-ranking pose was chosen as the center of the first cluster, and poses with an RMSD below 2 Å relative to this pose were assigned to the same cluster. The next unassigned pose with the best score was
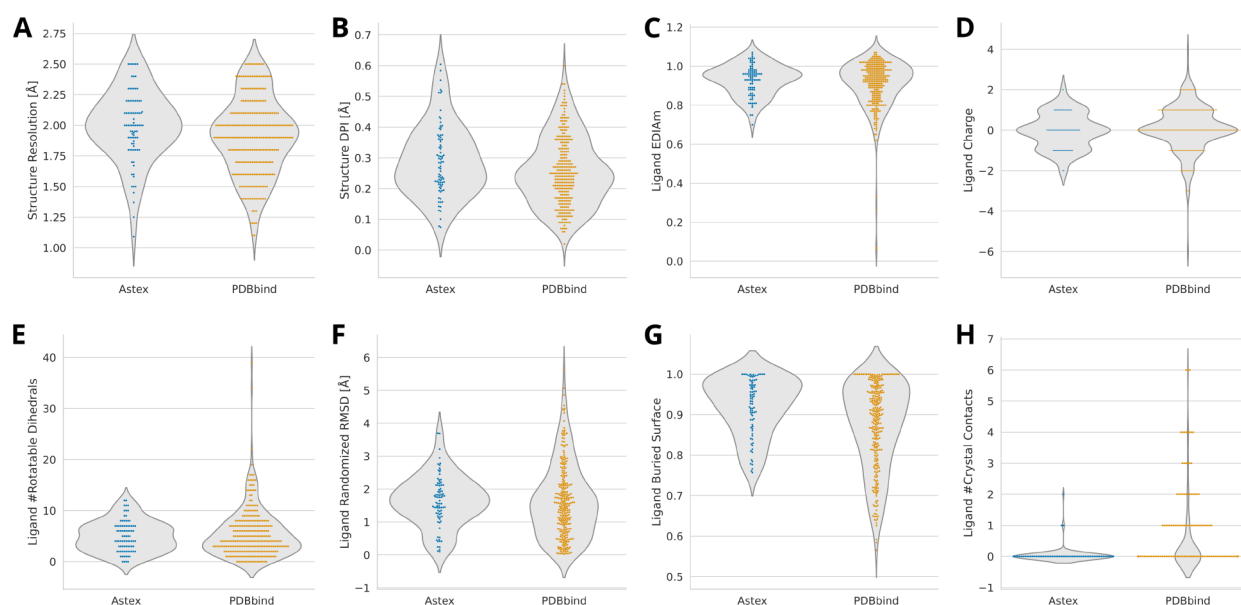
Figure 2. Properties of the Astex Diverse set (blue) and the PDBbind Core set (orange). (A) X-ray resolution of complex structure. (B) Diffraction-component precision index (DPI) of complex structure. (C) Ligand EDIAm value. (D) Ligand charge. (E) Ligand number of rotatable dihedrals. (F) Ligand RMSD between randomized and native pose. (G) Portion of buried surface area of the ligand. (H) Number of crystal contacts of the ligand.

chosen as the center of the second cluster, and its neighbors were assigned to this cluster. This procedure was continued until all poses were assigned to a cluster. A maximum of eight members were kept in each cluster, and the remaining similar poses of worse score were discarded. A maximum of 50 clusters were kept.

We used six different docking success criteria to assess the redocking and cross-docking performance of different parameters and programs. We analyzed whether the experimentally determined (native) pose was (i) within 1.0 Å of the best docking pose, (ii) within 1.5 Å of the best docking pose, (iii) within 2.0 Å of the best docking pose, (iv) within 2.0 Å of one pose of the best docking cluster, (v) within 2.0 Å of one pose within the best five docking clusters, and (vi) within 2.0 Å of one of all final docking poses. When only one criterion was analyzed, we chose definition (iii) as the most universally used definition.

RESULTS AND DISCUSSION

Characterization of the Benchmark Sets. We briefly investigated and compared the properties of the employed benchmark sets. As shown in Figure 2, both sets contain high-quality structures, but there are significant differences between the sets, some of which might be attributed to the fact that the PDBbind Core set is newer. The resolution is on average better in the PDBbind Core set (Figure 2A), leading to a lower DPI[26] (Figure 2B). Seven complexes of the Astex set and five of the PDBbind Core set have a DPI higher than 0.5 Å, suggesting lower precision on their coordinates. Another measure of local structure quality of the ligand is its EDIAm value.[27] Only four ligands of the Astex set (5%) but 35 ligands of the PDBbind Core set (12%) have an EDIAm value below 0.8 (Figure 2C), suggesting a significant coordinate uncertainty. The net charges of the ligands in the Astex set are balanced and range from −2 to +2, while the charges in the PDBbind Core set are skewed toward positive values and range from −6 to +4 (Figure 2D). The average ligand flexibility is similar in the two

benchmark sets (Figure 2E). However, the PDBbind Core set contains 23 ligands with more than 12 rotatable bonds and two with more than 30 rotatable bonds, whereas in the Astex set the maximum number of rotatable bonds is 12. This leads to a wider distribution of the RMSD between native and randomized ligand poses in the PDBbind Core set than in the Astex set (Figure 2F, RMSD calculated after superimposition of ligand poses). Conformations of ligands with a small number of rotatable dihedrals cannot efficiently be randomized. However, the bond lengths and bond angles as well as the 3D orientations of the randomized poses differ from the native poses.

The ligands of the PDBbind Core set are on average significantly more solvent-exposed than the ligands of the Astex set (Figure 2G). In the Astex set, three ligands (4% of the set) are in van der Waals contact with one or two residues from another symmetry-related copy of the complex in the crystal (Figure 2H). In the PDBbind Core set, 60 ligands (21%) form contacts with up to six symmetry-related residues. Since the symmetry-related copies of the complexes are absent in the docking calculations, these ligands are prone to display a lower buried surface than ligands not making any crystal contacts (Figure S1).

In 19 complexes of the Astex set and 20 complexes of the PDBbind Core set, the ligand is in direct contact with a zinc ion. In addition, in one complex of the Astex set, the ligand makes a bond to an iron ion in a heme cofactor, and in five complexes of the PDBbind Core set, the ligand is in direct contact with a magnesium ion.

In conclusion, the analysis suggests that the PDBbind Core set is more challenging for docking than the Astex Diverse set, as it contains more flexible ligands making fewer protein contacts. At the same time, it contains more potentially problematic experimental structures.

Scoring Failures. Based on our definition, six cases of the Astex set (7%) and 45 cases of the PDBbind Core set (16%) qualified as scoring failures, meaning that the AC scoring
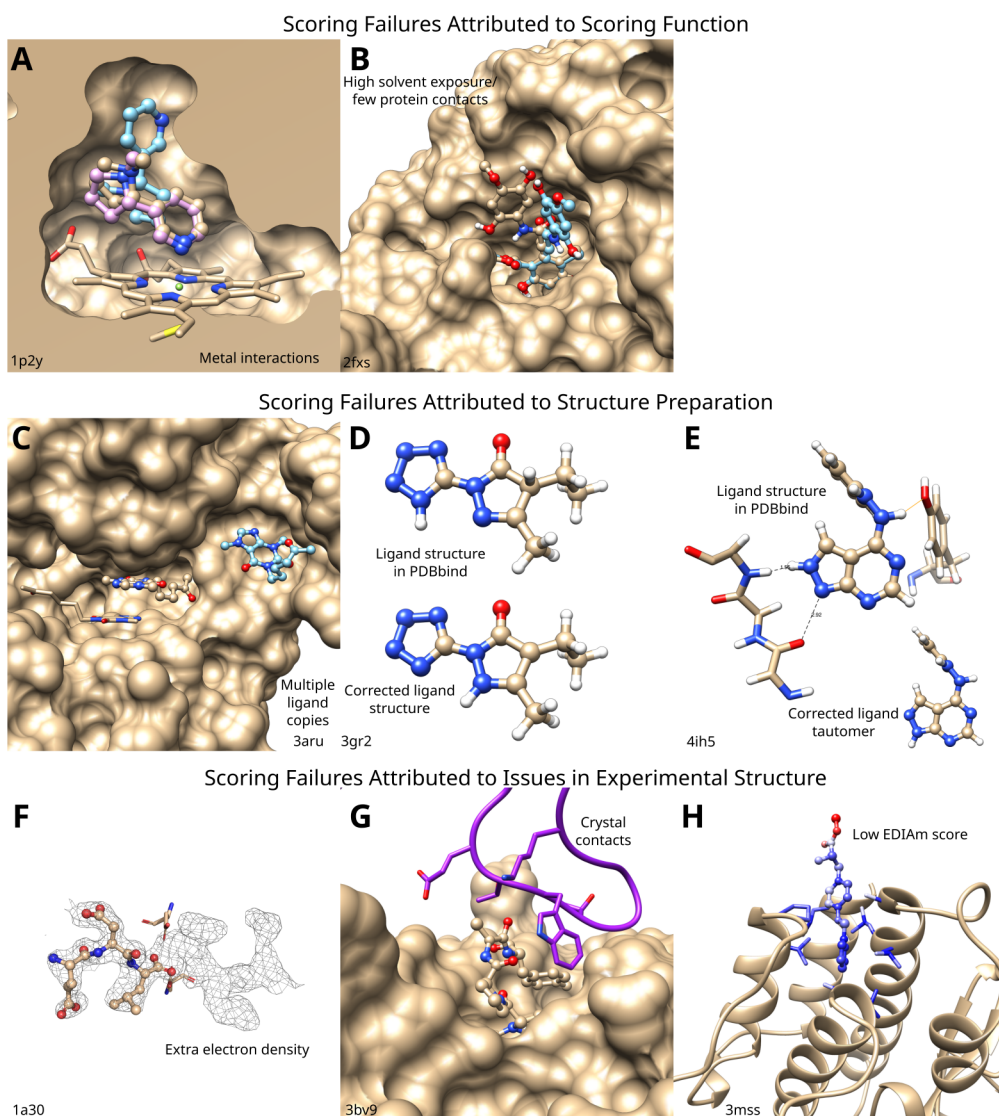
Figure 3. AC scoring failures. The ligand of interest is shown in ball and stick representation and other residues in stick representation. Experimental data are shown in tan color. (A) Heme-iron-bound ligand. Native structure shown in tan, best pose from standard AC docking in light blue, best pose from docking with MMBP in pink. (B) Very solvent-exposed ligand (proportion of buried surface area 0.77) making few protein interactions. The best docking pose (light blue) reproduces the protein-bound part of the native pose (tan) faithfully but fails for the solvent-exposed portion. (C) Ligand binding to protein and cofactor. Native structure shown in tan, best docking pose in light blue. (D) Wrong chemical structure of ligand. (E) Wrong ligand tautomer. (F) Unassigned electron density in active site. (G) Missing crystal contacts. Residues from a symmetry-related copy of the protein interacting with the ligand are shown in purple. (H) Limited electron density support for ligand coordinates. The ligand atoms are colored by their respective EDIA values, going from red (minimum) to blue (maximum).

function did not rank the X-ray pose as the most favorable binding pose. These failures can be divided into three classes according to their underlying cause: an issue in (1) the scoring function, (2) the structure preparation, or (3) the experimental structure. Approximately one-third of the AC scoring failures of the PDBbind Core set can be attributed to each class (Table S4).

Issues in the scoring function mainly concern (a) the presence of metal ions such as zinc or heme-bound iron in the active site (Figure 3A) and (b) some very solvent-exposed ligands, which make little interactions with the protein (Figure 3B). The first issue requires a better description of the polarization and charge transfer in metalloproteins, which can be achieved, for example, by a hybrid quantum/classical (QM/MM) description.[19,31] In the case of heme-iron-bound ligands, the scoring function can also be improved by using an

MMBP.[18] The Astex set contains one such complex, namely, P450cam bound to nicotine (PDB ID 1p2y). In standard docking with AC, this case is a scoring failure (Figure 3A, pose in light blue). However, use of the MMBP for the nicotine ligand leads to a docking success, with the best pose displaying a RMSD of 0.6 Å from the experimental pose (Figure 3A, pose in pink).

Issues attributed to structure preparation include (a) missing cofactors (Figure 3C), (b) errors in the chemical structure of the ligand (Figure 3D), (c) wrong protonation states of the ligand or the protein, and (d) ligand tautomers that cannot form the hydrogen bonds present in the X-ray structure (Figure 3E). These issues are in principal easy to fix but time-consuming, as they need to be detected and manually treated. This was not done in the present work.
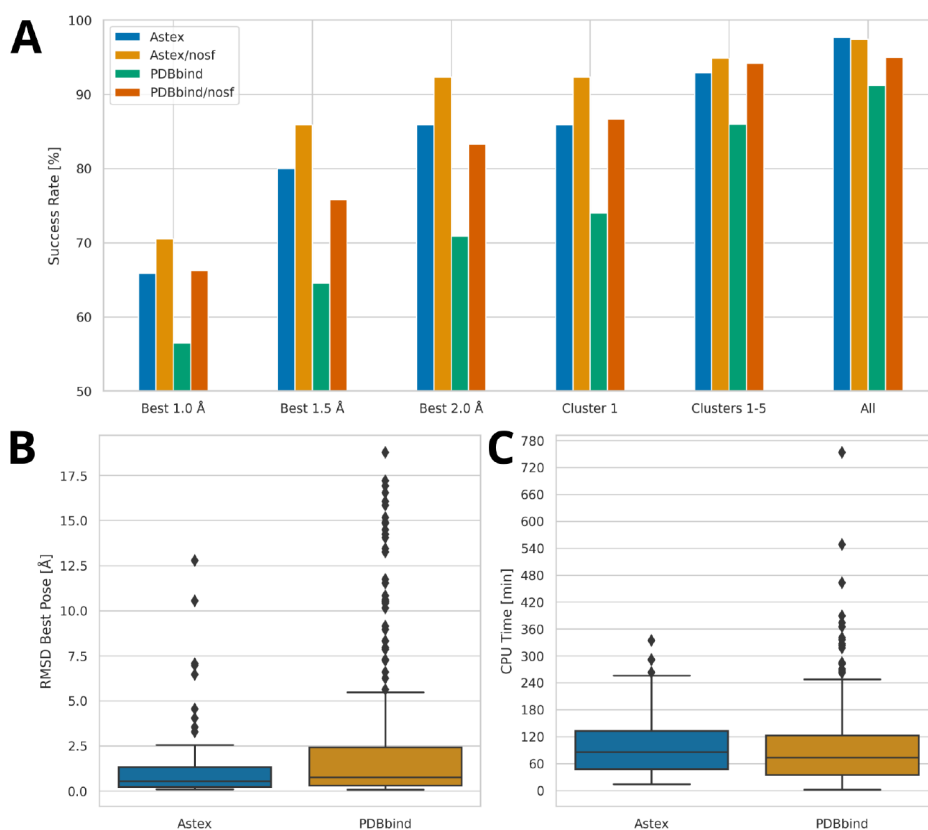
Figure 4. Docking results of AC on the Astex Diverse set and PDBbind Core set with same parameters. (A) Success rates for Astex and PDBbind set for all cases and after removal of scoring failures (nosf). (B) RMSD values of best poses. (C) CPU timings.

Issues in the experimental structures include (a) unassigned extra density in the active site, which could correspond to an an alternative ligand binding site (Figure 3F), (b) extensive contacts of the ligand with another copy of the protein (crystal contacts), which stabilize the experimentally observed ligand binding pose but which are not present in the docking calculations (Figure 3G), and (c) parts of the ligand with higher flexibility/less electron density support (Figure 3H). Supporting this assessment, eight complexes of the PDBbind Core set with a ligand EDIAm value below 0.8 are part of the scoring failures. These issues are not due to the scoring function but to the quality of the employed experimental structure. As a matter of fact, the scoring function is good enough to highlight these problematic cases.

For benchmarking the sampling algorithm of AC 2.0 and the different parameters influencing it, we removed scoring failures in some analyses as indicated because otherwise bad sampling could be rewarded with a higher success rate.

Redocking of Astex Diverse and PDBbind Core Sets. After assuring that AC 2.0 yields similar results as AC 1.0 (see the Supporting Information), we compared the performance of AC 2.0 on the two benchmark sets, using the same parameters (CHARMM27 force field, 90° rotation, eight RIC). The results confirm that the PDBbind Core set is more challenging, leading to lower success rates (Figure 4A) as expected due to the presence of more flexible and less buried ligands. The difference is less pronounced but still evident after the removal of scoring failures and can also be seen from the distribution of the RMSD values of the best poses (Figure 4B). The CPU timings are similar for the two sets (Figure 4C), although in the PDBbind Core set there are more outliers with very long

CPU times. In the following we consider only the PDBbind Core set and remove scoring failures for the analysis, unless otherwise stated.

Evaluation of Sampling Robustness. In order to evaluate how randomization of the initial conditions (RIC) influences the docking results, we carried out eight replicates of a docking run with standard parameters and a 90° ligand rotation angle. The results show some variability in the success rates (Figure 5A and Table S5), which is caused by a high variability in the docking results for each complex, as shown by the RMSD of the best pose per complex found in each docking run (Figure 5B). Of course, the exact variability and benefit depend on the chosen sampling parameters and test cases. Here we observe that for the best pose RMSD being below 2 Å with respect to the native pose, 18 cases (7.5%) are failures in all eight replicas, 131 cases (54.6%) are successes in all eight replica, and 91 cases (37.9%) change between failure and success depending on the initial conditions. It is also evident (Figure 5A, yellow bar, and Table 1) that merging the results of the eight replica displays significantly better success rates than each individual replica. These results support the use of multiple RIC in order to achieve more robust docking results.

We also carried out docking runs with different numbers of RIC (Figure 5C,D). The success rates (Figure 5C) demonstrate that with the chosen conditions, the use of six or more initial conditions leads to converged results. Taking into account only the runs with six or more replicas, the RMSD of the best pose per complex found in each docking run (Figure 5D) varies much less than in case of a single replica (Figure 5B). Here we observe that 22 cases (9.2%) are always failures, 186 cases (77.5%) are always successes, and only 32
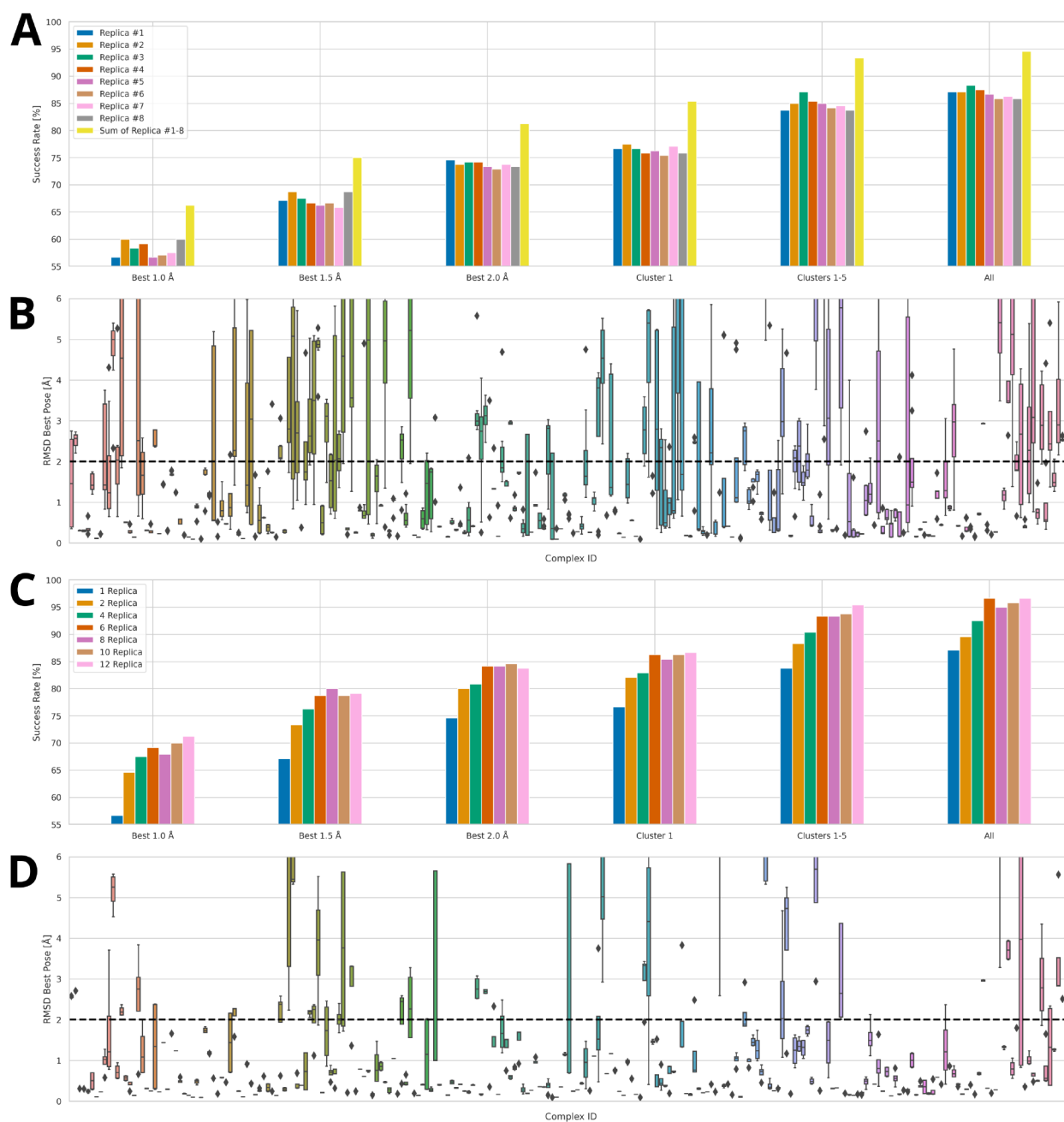
Figure 5. AC sampling depending on initial conditions (90° rotation of ligands). (A, B) Docking results using eight different RIC: (A) success rates [%] for each replica and the merged results of the eight replicas; (B) RMSD value [Å] of the best pose per complex (in alphanumerical order) for each of the replicas. (C, D) Docking results using different numbers of RIC (1, 2, 4, 6, 8, 10, and 12): (C) success rates [%] for each run; (D) RMSD value [Å] of the best pose per complex (in alphanumerical order) for the runs with 6−12 replicas only.

cases (13.3%) change between failure and success. In summary, the robustness of the docking results is greatly increased by increasing the sampling through randomization of the initial conditions.

Influence of the Placement Cloud Points. In AC 1.0, the attractive cloud points served a twofold purpose, defining the attractive grid and placing the ligand for the initial sampling procedure. Here we separated the second function from the first, creating separate cloud points similar to the attractive cloud points but excluding points close to the protein surface (see the Supporting Information for details). The results demonstrate that with a very coarse sampling (rotational angle of 180°, 4 RIC), use of placement points

decreases the success rates and does not speed up the docking runs (Figure 6). However, at normal sampling (rotational angles of 90° and 60°, 4 RIC), the use of placement points has a negligible influence on the success rate (Figure 6A) and the RMSD value of the best pose (Figure 6B) but greatly reduces CPU times for these runs. The mean CPU times with placement points are 0.9 h (90°) and 3.7 h (60°) versus 1.4 h (90°) and 6.2 h (60°) without placement points.

In summary, placement points can significantly reduce CPU times with only a minor influence on the docking success rates.

Influence of Other Docking Parameters. We tested different rotational angles for the initial ligand sampling. Because of symmetry relations, the number of initial poses per

Table 1. Selected AC Redocking Results after Removal of Scoring Failures[a]

parameters	Best	Cl1	Cl1−5	All	#Poses	#AP	#PP	Time [min]
90°, 1 RIC, Rep 1	74.6	76.7	83.8	87.1	407.5	38.1	23.2	13.3
90°, 1 RIC, Rep 1−8	81.2	85.4	93.8	95.0	409.4	38.3	23.3	103.2
90°, 2 RIC	80.0	82.1	88.3	89.6	765.4	38.3	23.1	27.6
90°, 4 RIC	80.8	82.9	90.4	92.5	1414.5	38.5	23.2	53.0
90°, 8 RIC	84.2	85.4	93.3	95.0	2575.4	38.4	23.1	100.3
90°, 12 RIC	83.8	86.7	95.4	96.7	3613.6	38.3	23.0	147.7
90°, 4 RIC, 4 CPU	81.7	85.0	91.2	92.1	1419.2	38.5	23.2	13.4
90°, 4 RIC, 16 CPU	81.2	85.0	92.9	94.6	1422.6	38.5	23.3	4.3
90°, 4 RIC, noEP	77.5	80.8	87.9	92.1	1203.7	38.2	23.1	48.8
90°, 8 RIC, c27	83.3	86.7	94.2	95.0	2557.4	38.8	23.4	100.5
90°, no RIC, noPP, c27	77.1	79.6	87.9	89.6	655.1	38.6	38.6	21.5
90°, no RIC, noPP	76.2	78.3	88.8	90.0	662.5	38.5	38.5	21.3
90°, 4 RIC, noPP	81.7	85.0	91.2	93.8	2297.0	38.4	38.4	81.8
90°, 4 RIC, N_{Thr} 50	84.6	87.1	92.9	95.4	3843.8	95.6	65.3	149.8
90°, 4 RIC, N_{Thr} 60	82.1	85.4	91.7	93.8	2432.1	63.5	40.8	97.5
90°, 4 RIC, blind	79.2	80.8	89.2	90.8	6119.7	166.8	88.2	298.1
90°, 4 RIC, native	91.2	93.3	97.1	97.1	1406.7	38.4	23.1	51.4
45°, 4 RIC	86.7	90.0	95.8	96.7	8935.4	38.5	23.2	394.3
60°, 4 RIC	86.2	87.5	96.2	97.5	5115.8	38.3	23.0	222.0
60°, 1 RIC, noPP	81.7	85.0	93.3	94.6	2519.9	38.2	38.2	92.2
60°, 4 RIC, noPP	86.7	88.8	95.4	97.1	8287.2	38.4	38.4	370.6
72°, 4 RIC	85.8	88.3	95.0	96.2	5744.8	38.4	23.0	250.7
120°, 4 RIC	80.0	83.8	91.7	92.5	1533.1	38.3	23.1	61.4
180°, 4 RIC	68.8	70.8	76.7	78.3	265.8	38.4	23.0	9.4

[a]Best: success rate (RMSD ≤ 2 Å) of best pose. Cl1: success rate of best cluster. Cl1−5: success rate of best five clusters. All: success rate in all final poses. #Poses: mean number of poses in final scoring. #AP: mean number of attractive cloud points. #PP: mean number of placement points. Time: mean CPU time. Abbreviations used for parameters: initial ligand rotation (deg), number of randomized inititial conditions (RIC), no use of placement points (noPP), no use of electrostatic points (noEP), CHARMM27 force field (c27), replica number (Rep), number of CPUs for parallel runs (CPU), docking on full protein surface (blind), docking starting from native pose (native). The complete table for all tested docking parameters is shown as Table S5.

placement point increases from 180° to 90°, 120°, 60°, 72°, and 45°, which is apparent from the CPU times (Figure 7A and Table 1). The success rates do not significantly increase from 90° to 120° and from 60° to 72° and 45°, respectively, suggesting the choices of 90° for fast docking and 60° for high-precision docking.

As we changed the definition of electrostatic cloud points, we tested their influence on the docking results. The use of the new electrostatic cloud points slightly increased CPU time but also improved the docking success rates and the RMSD value of the best poses (Figure 7B and Table 1). Based on these results, we recommend the use of electrostatic cloud points for the initial sampling.

As can be appreciated from Figure 7C, using a lower value for N_{Thr} and therefore more attractive cloud points located also in shallow protein pockets leads to a slight increase in the success rates. However, the computational cost for this improvement is important (Figure 7D and Table 1). For selected targets with known shallow binding pockets, the adaptation of this parameter may be helpful, but in general a value of 70 provides a good compromise between speed and accuracy.

We also tested the influence of defining the search space around the active site with "blind" docking runs, where the whole protein surface is included in the search space (Figure 7D and Table 1). Whereas the CPU times are considerably higher in the second case, the success rates remain almost constant (best pose RMSD ≤ 2 Å: local, 80.8%; blind, 79.2%),

suggesting that AC also provides good results when the active site of a target is unknown.

Another test investigated the influence of the randomization of the starting structure of the ligand. The results (Figure 7E and Table 1) showed that starting from a randomized ligand structure, it is possible to reach almost the same high success rates as when starting from the native ligand structure. For example, the success rate for the best pose being within 2 Å of the native pose is 91.2% when starting from the native structure and 86.7% when starting from the randomized structure. Of course, this necessitates a better sampling (shown here: rotational angle of 45° for the random pose vs 90° for the native pose) and therefore substantially longer CPU times.

We tested the parallelization of the program by carrying out docking runs with standard parameters (90° rotation, 4 RIC) using 1, 2, 4, 8, and 16 CPUs on the same computing node. The speedup is generally good but depends on the serial CPU time of the docking runs due to the overhead of serial parts of the program. For the fastest 15% of cases, the speedup on 16 CPU amounts only to 7.0 on average, but for the slowest 15% it amounts to 12.2. For the 70% of docking cases with a medium serial CPU timing of 12−80 min, the speedups on 2, 4, 8, and 16 CPU amount to 1.9, 3.5, 6.5, and 10.5, respectively (Figure 7F and Table S5).

Redocking Success Analysis by Complex. In the following, we do not exclude scoring failures in the analysis in order to be able to compare the performance of AC to those of GOLD and AutoDock Vina later.
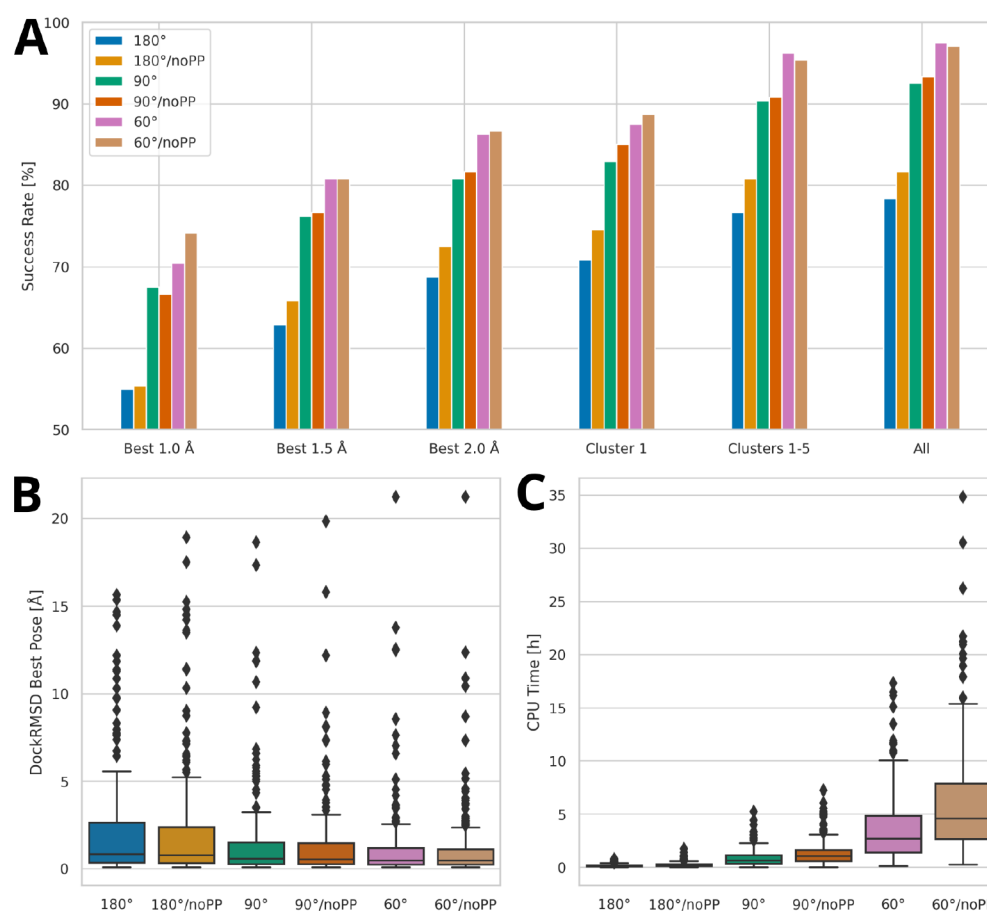
Figure 6. Placement cloud points. (A) AC redocking success rates with and without placement cloud points (PP) for three different rotational angles (180°, 90°, 60°) using 4 RIC. (B) RMSD value [Å] of the best poses. (C) CPU times [h].

We analyzed the AC docking success rate (best pose within 2 Å of native pose) by complex, taking into account all local docking runs of the PDBbind Core set using the CHARMM36 force field (40 runs). The success rate varies strongly depending on the target, with carbonic anhydrase 2 displaying the lowest success rates, partially attributable to the presence of zinc in the active site, and 16 targets displaying a median success rate of 100% (Table S7). Very small or shallow active sites, characterized by a low number of attractive cloud points, display on average a lower success rate than medium or large active sites (data not shown). In those cases, use of a lower N_{Thr} value would be recommended. Regarding the ligand properties, it is evident that rigid ligands display better success rates than flexible ligands (Figure 8A, light-gray violin/blue points). In particular, ligands with more than 10 rotatable bonds display significantly lower success rates. The portion of buried ligand surface also exerts a strong influence on the docking success (Figure 8B): the more buried a ligand, the higher the success rate. The AC redocking success rate is higher for ligands not making any crystal contacts (Figure 8C). Regarding the quality of the experimental data supporting the native binding mode, the electron density score for multiple atoms (EDIAm)[27] value of the ligands shows some correlation to the success rates (Figure 8D), suggesting that the docking works better for well-resolved ligands than for less-well-resolved ligands. The net charge on the ligand has no strong effect on the success rates (Figure 8E). As described earlier,[1,31] the AC scoring function performs less well for complexes having a zinc ion in the active site (Figure 8F). However, when

scoring failures are excluded (10 out of 20 zinc metalloprotein complexes), the AC success rates are similar for zinc-free and zinc-containing targets (Figure S3F). The performance of AC does not correlate with the binding strength of the ligand (data not shown), which makes it well-suited for fragment docking.

Redocking Comparison of AC, GOLD and AutoDock Vina. Redocking with GOLD was carried out with the three scoring functions GS, CS, and PLP. The docking program is extremely fast, especially with the CS and PLP scoring functions, requiring on average less than 2 min per docking. As reported in Table 2 and Figure 9, CS yields the lowest success rates, while GS and PLP show approximately the same performance, but PLP at lower computational cost. Varying the number of genetic algorithm runs from 10 to 100 and 1000 only marginally influences the success rate of the best pose but increases the success rates for cluster 1, clusters 1−5, and the full population. When starting from the native pose instead of a randomized ligand pose, a large increase in success rates is observed, for example, 10% for the best pose (RMSD 2 Å), suggesting a limited conformational sampling power. Based on our results, we estimate that CS fails in 80 cases (28%) to correctly rank the native pose, GS in 69 cases (24%), and PLP in 66 cases (23%). More than half of these scoring failures concern very solvent-exposed ligands (buried surface portion < 0.85).

When analyzing the docking success by ligand characteristics (Figure 8), it is noticeable that the GOLD success rates show a similar dependence on ligand flexibility and solvent-accessible surface as AC (Figure 8A,B). GOLD is more sensitive to
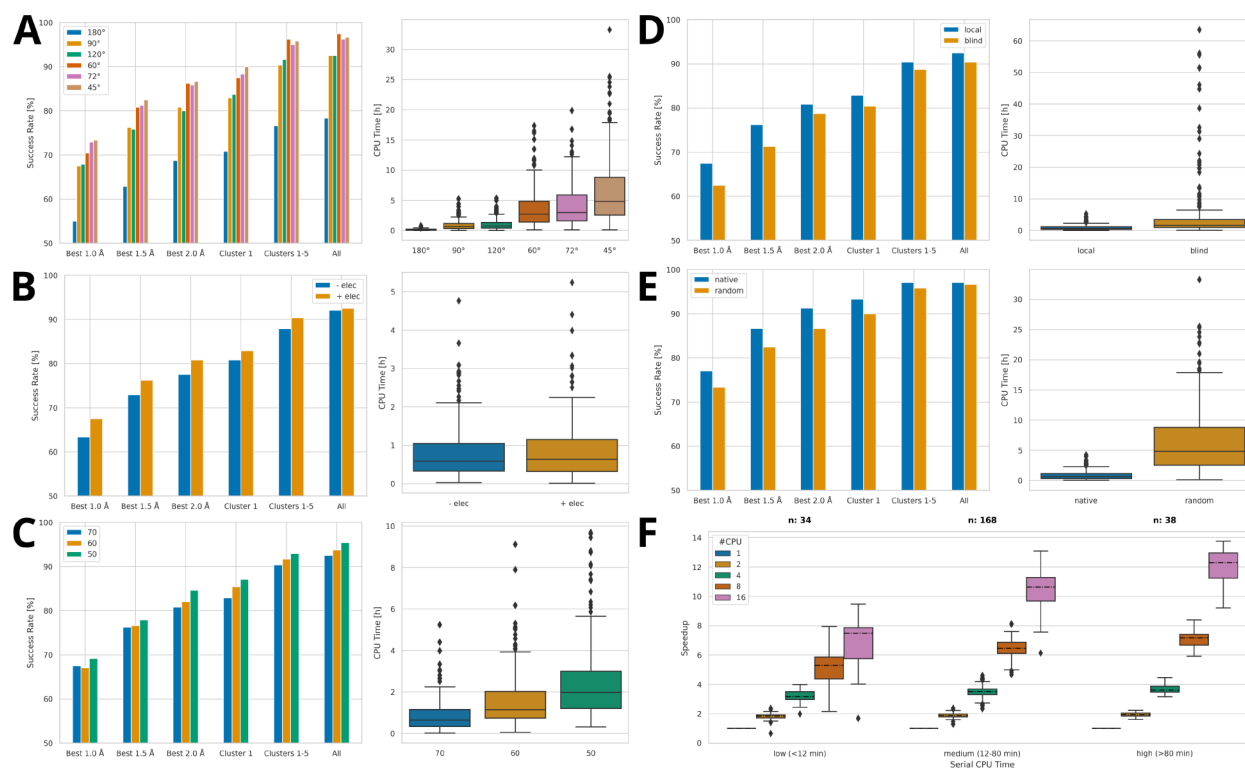
Figure 7. Influence of other docking parameters on the AC redocking success rates (left) and the CPU times (right). Unless otherwise stated, dockings were carried out with a rotational angle of 90° and an N_{Thr} value of 70 using placement points, electrostatic points, and 4 RIC. (A) Rotational angle. (B) Electrostatic cloud points. (C) N_{Thr} value. (D) Local/blind docking over the complete target surface. (E) Initial ligand conformation (native/randomized). Here a rotational angle of 90° was used for the runs starting from the native conformation and 45° for the runs starting from the randomized conformation. (F) Acceleration due to parallelization.

crystal contacts and the quality of the ligand structures, as reflected by the EDIAm values compared to AC (Figure 8C,D). The performance of GOLD for zinc metalloproteins is only slightly lower than for other targets (Figure 8F).

Redocking with AutoDock Vina was carried out only with the original Vina score, as it has been shown to clearly outperform the AutoDock4 score.[14] The main parameter which can be provided to determine docking speed and accuracy is the "exhaustiveness", with a default value of 8. We tested values of 8, 16, 32, 100, and 1000 (Table 2). These values do not influence much the docking results. AutoDock Vina uses a random seed for sampling, but its results seem robust with respect to this random seed (Table 2). Comparing the current version of AutoDock Vina, version 1.2.3, to an older version, 1.1.2, we noted that the new code yields somewhat better results at about half the CPU time.

The performance of AutoDock Vina depends strongly on ligand flexibility. Vina performs well for ligands with up to nine rotatable bonds but worse for more flexible ligands (Figure 8A). The docking program performs very badly for solvent-exposed ligands (Figure 8B) and better for neutral than for charged ligands (Figure 8E). AutoDock Vina shows some sensitivity to crystal contacts (Figure 8C) but little sensitivity to quality of the ligand structures (Figure 8D) and the presence of zinc in the active site (Figure 8F).

Comparing the three docking codes, AC yielded the highest success rates, followed by GOLD and AutoDock Vina (Figure 9). The best success rate (best pose RMSD ≤ 2 Å) was 73.3% for AC, 63.9% for GOLD, and 58.2% for AutoDock Vina. When starting from the native ligand conformation, the best success rates were 78.2% for AC (+4.9%), 72.3% for GOLD/

PLP (+8.4%), and 70.2% (+12.0%) for AutoDock Vina. These results suggest that the AC scoring function is better at correctly ranking the native pose and that its sampling algorithm is better at generating a native-like ligand conformation from a random conformation. Another distinguished feature of AC is its systematic potential for improvement. Better sampling, as determined by the input parameters such as the rotational angle, the number of RIC, and the N_{Thr} value, leads to better success rates (Figure 9).

Table S7 reports the median success rates per target and per docking code. The 12 targets with a median success rate of 100% over the three docking programs are all characterized by very buried and rather rigid ligands, while the targets with very low success rates (Chitinase A, β-lactamase, endothiapepsin, ITK/TSK, and RNase) on the other hand display highly solvent-exposed ligands, which are in the case of endothiapepsin also highly flexible. GOLD and AutoDock Vina outperform AC on the three zinc-containing targets (MMP-12, thermolysin, and carbonic anhydrase 2) and for unknown reasons on HSP82 and Factor X heavy chain. AC, on the other hand, performs much better on targets with relatively flexible ligands, such as β-lactoglobulin bound to saturated fatty acids of different lengths, DAP synthase, acetylcholinesterase, cellular tumor antigen p53, PKA C-α, and Chk1.

Scoring and Ranking. Scoring power refers to the ability of a scoring function to produce binding scores in a linear correlation with experimental binding data, while ranking power is the ability to correctly rank the known ligands of a target by their affinities.[10] Although not the main topic of the present article, we computed the scores for the experimental protein−ligand complexes in the PDBbind Core set in order to
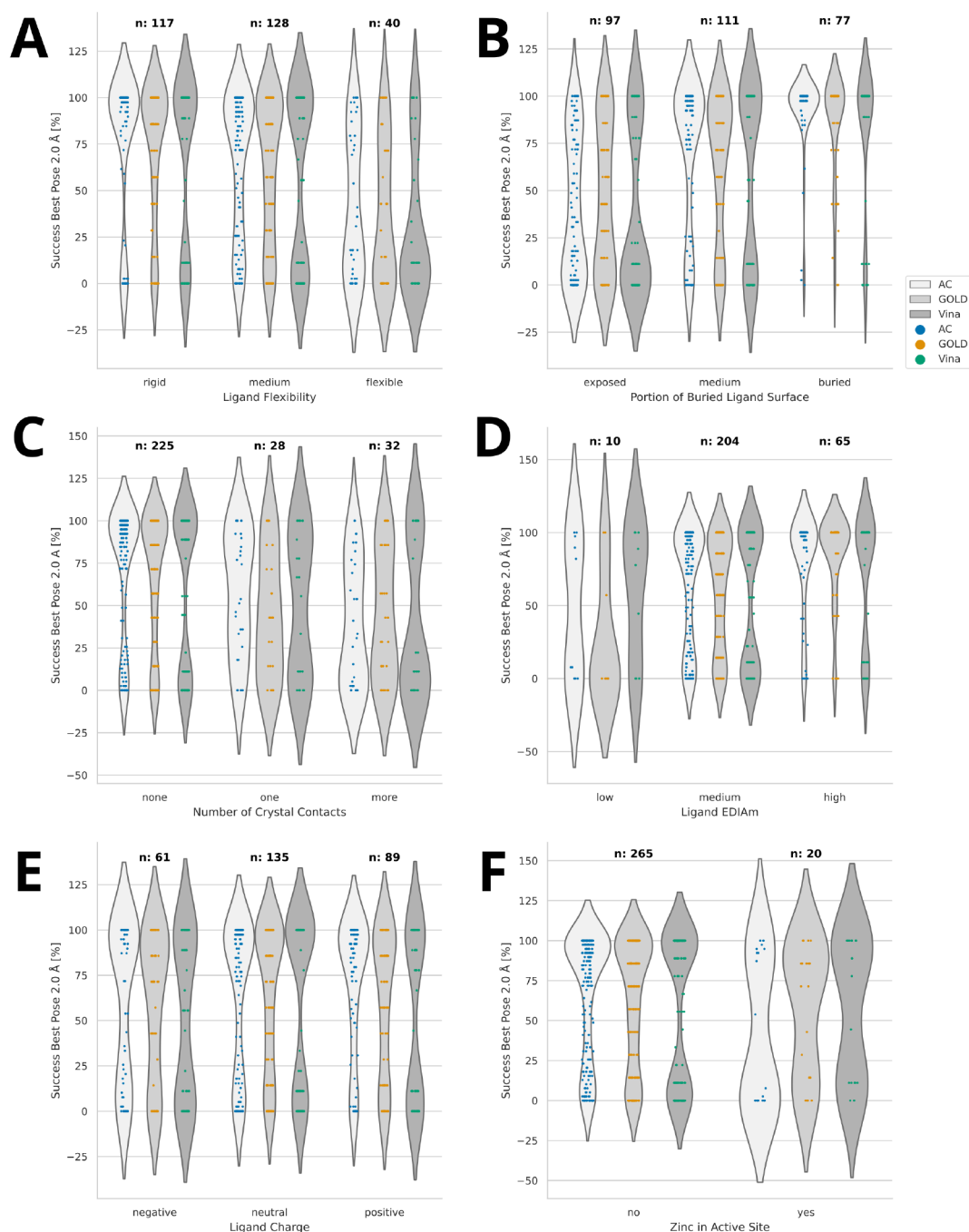
Figure 8. AC, GOLD, and AutoDock Vina redocking success rate analysis by complex merged over different docking conditions (AC, 40 docking conditions; GOLD, 6 docking conditions; AutoDock Vina, 8 docking conditions). Shown are success rates (best pose within 2 Å) by (A) number of rotatable bonds (rigid, 0−4 bonds; medium, 5−9; flexible, 10−32), (B) portion of buried ligand surface (exposed, <0.85; medium, 0.85−0.95; buried, >0.95), (C) number of crystal contacts of the ligand, (D) ligand EDIAm value (low, <0.7; medium, 0.7−1.0; high, 1.0−1.2), (E) net ligand charge, and (F) presence of zinc in the active site.

evaluate and compare our scoring function. Over the whole data set of 285 complexes, the Pearson correlation coefficient between calculated and experimentally determined binding free energy ΔG is 0.611, and the standard deviation is 1.72. These values are very similar to the ones obtained with ChemPLP/GOLD (0.614, 1.72) and with AutodockVina (0.604/1.73).[10] For ranking, the SwissParam scoring function shows a Spearman's rank correlation coefficient of 0.596 and a predictive index of 0.624, between the values for ChemPLP/GOLD (0.633/0.657) and AutoDock Vina (0.528/0.557;

definitions of the metrics and results from ref 10). The performance for scoring and ranking per protein target varies widely, as shown in the Figure S4.

Cross-Docking. We cross-docked the 5 ligands of each of the 57 protein targets of the PDBbind Core set to the target structure with the highest affinity ligand. Using the standard AC docking parameters with 6 RIC and a fixed protein, the success rate for the best pose at a RMSD below 2 Å was 42.5% (Table S8). The low success rate is due to clashes between the ligand and the protein when docking to a non-native protein

Table 2. Redocking Results of AC, GOLD, and AutoDock Vina for Complete PDBbind Core Set (For Abbreviations, See Table 1)

parameters	Best	Cl1	Cl1−5	All	CPU [min]
Attracting Cavities					
180°, 4 RIC	58.6	60.4	68.8	74.0	9.2
90°, 1 RIC	63.2	65.3	77.2	83.5	12.8
90°, 2 RIC	68.1	69.8	79.6	85.3	26.7
90°, 4 RIC	69.1	70.9	82.8	88.8	51.3
90°, 8 RIC	71.9	73.0	83.9	90.2	97.2
60°, 4 RIC	72.6	73.7	86.7	93.3	214.8
45°, 4 RIC	73.3	76.1	87.0	92.6	381.9
90°, 4 RIC, blind	67.0	68.4	79.6	86.3	272.0
90°, 4 RIC, native	78.2	80.0	88.1	93.0	49.8
GOLD					
ChemScore	57.2	63.9	79.6	83.9	1.6
GoldScore	62.8	68.1	83.5	85.6	9.3
PLP, 10 GA	61.1	63.9	74.0	74.0	0.2
PLP	62.1	68.8	85.6	88.1	1.3
PLP, 1000 GA	63.9	69.8	85.6	90.9	7.3
PLP, native	72.3	76.5	88.8	90.5	1.3
AutoDock Vina					
Exh. 8	56.8	58.6	77.9	86.3	0.9
Exh. 16	58.2	59.6	80.7	88.4	2.0
Exh. 32	56.8	58.9	79.3	90.5	3.9
Exh. 100, Rep 1	57.5	61.4	81.1	93.0	11.4
Exh. 100, Rep 2	57.9	61.1	81.4	91.9	11.4
Exh. 100 v1.1.2	57.2	57.9	79.3	83.5	21.1
Exh. 1000	58.0	61.8	83.0	91.5	79.6
Exh. 100, native	70.2	74.4	94.4	98.6	11.0

structure. Using a flexible protein within a radius of 1, 2, or 3 Å around the ligand during sampling and scoring to alleviate clashes did not significantly change the success rates. GOLD with the PLP scoring function reaches about the same cross-docking success rate (42.8%), while AutoDock Vina perfoms worse (33.1%).

For some protein targets, we visually inspected the superimposition of the five complexes and manually selected amino acids that clashed with native poses. For four targets, namely, acetylcholinesterase, coagulation factor X heavy chain, muscle glycogen phosphorylase (myophosphorylase), and the androgen receptor, this increased the success rate (best pose RMSD ≤ 2 Å) from 25% with a rigid protein to 30/50/55% with a flexible radius of 1/2/3 Å respectively, to 65% with manually selected flexible residues (Table S8). This demonstrates that manual curation can increase the success rate in cross-docking. The informed choice of a receptor structure, for example, one showing the lowest number of clashes with known ligands, might further increase success rates but was not tested here.

We also performed cross-docking with AC to the heme-free apo form of the anticancer target indoleamine 2,3-dioxygenase 1 (IDO1), an enzyme that we have studied thoroughly in the past.[32,33] There are 13 heme-free X-ray structures of IDO1 available. The active site of apo-IDO1 is characterized by a large hydrophobic pocket with few hydrogen bond donors and acceptors. Cross-docking of these 13 ligands to chain A of X-ray structure 6wjy, we obtained a cross-docking success rate of 46% with a rigid receptor and 62% with eight manually selected flexible residues.

Screening. A prerequisite for virtual screening is the capability of a docking code to generate near-native docking poses for known ligands of a protein target. To benchmark AC 2.0 for screening applications, we chose five protein targets from the PDBbind Core set, for which all three docking codes were able to correctly dock most of the native ligands, namely, β-trypsin, U-plasminogen activator, 3-dehydroquinate dehydratase, catechol o-methyltransferase, and MTA nucleosidase. The cross-docking success rates for these 25 cases were 92% for AC with a rigid protein, 92% for GOLD, and 88% for AutoDock Vina. Screening all 285 molecules against these five targets, we obtained an average enrichment factor among the top 1% of 4.00, top 5% of 6.27, and top 10% of 3.93 for AC using the SwissParam score (Table S9). Using different weighting factors for the nonpolar and polar terms of the CHARMM/FACTS energy (eq 1), much higher enrichment factors of 19.33/9.98/6.93 (top 1/5/10%) were obtained, namely, with a weight of 1 for the nonpolar terms and 1.5 for the polar terms. These results highlight known shortcomings in traditional scoring functions, which have been solely parametrized to reproduce experimental protein−ligand binding data with known crystal structures, excluding decoy data.[34] The corresponding enrichment factors calculated for GOLD/PLP (4.00/7.07/4.67) and for AutoDock Vina (8.00/6.93/3.47) were in a similar range as the values obtained with AC/SwissParam score.

Due to CPU time constraints, we did not test the screening performance of AC 2.0 on other targets of the PDBbind Core set, as it is clear that scores calculated on erroneous binding poses will be meaningless. However, we also screened the 285 molecules of the PDBbind Core set against apo-IDO1, in addition to its 13 native ligands. The reweighted CHARMM/FACTS energy (nonpolar terms + 1.5 × polar terms) performed again very well, resulting in enrichment factors among the top 1/5/10% of 23.08/7.69/5.38 when docking to a fixed protein structure (0.00/4.62/4.62 for the standard SwissParam score). Curiously, the enrichment factors for IDO1 among the top 1/5/10% were much better when docking to a fixed protein than when docking to a flexible protein. In fact, with a flexible protein environment, several decoy ligands, mainly large peptide-like ligands such as the HIV-1 protease inhibitor ritonavir, were predicted to bind very tightly to IDO1. The 13 known IDO1 inhibitors were nevertheless all found among the 20% top ranked compounds.

In summary, the screening performance of AC 2.0 is system-dependent. Successful cross-docking of known ligands to a target is a necessary but not sufficient condition. Allowing for protein flexibility can improve docking performance but also allow for a higher number of false positives. To obtain a better screening performance, the SwissParam score may need to be retrained on a balanced test set of native binding poses and decoys. Our limited assessment on six targets suggests that attributing a higher weight to polar force-field terms may be useful for the detection of true binders among decoy compounds.

■ SUMMARY AND CONCLUSIONS

In summary, AC 2.0 is a versatile high-accuracy docking program that performs equally well for blind docking as for local docking, polar and nonpolar active sites, and charged and neutral ligands. It performs best for ligands with up to 10 rotatable dihedrals, for which the solvent exposure does not exceed 15%. Its sampling procedure can easily be adapted to
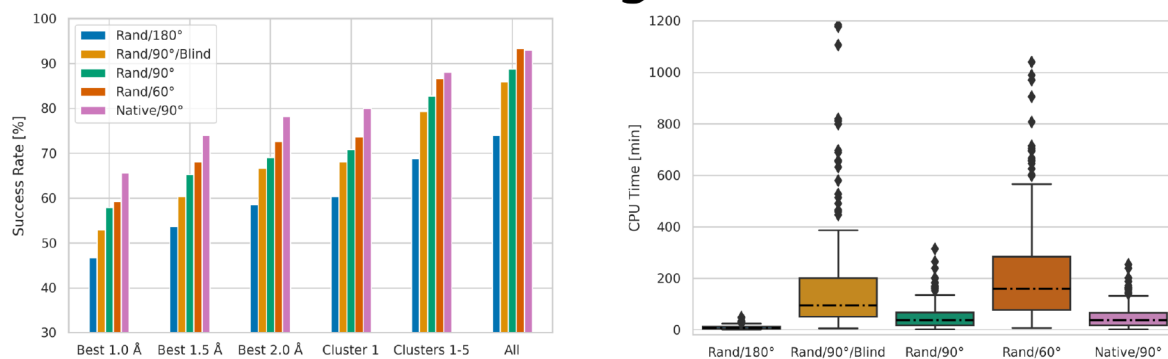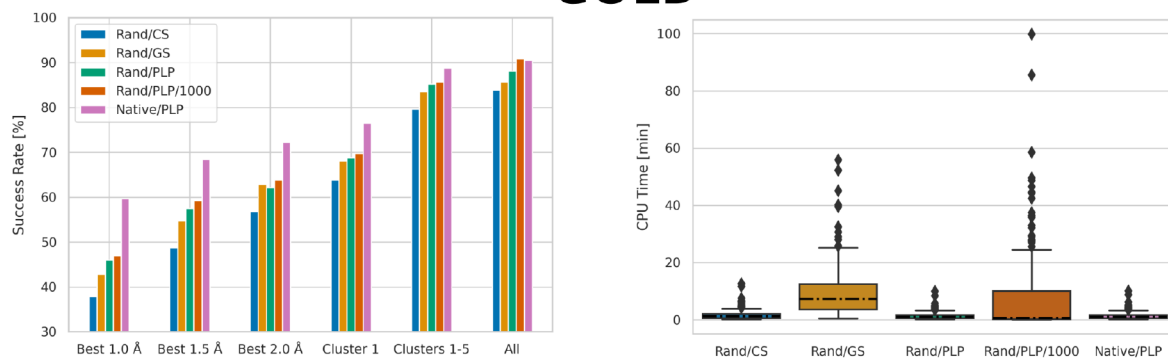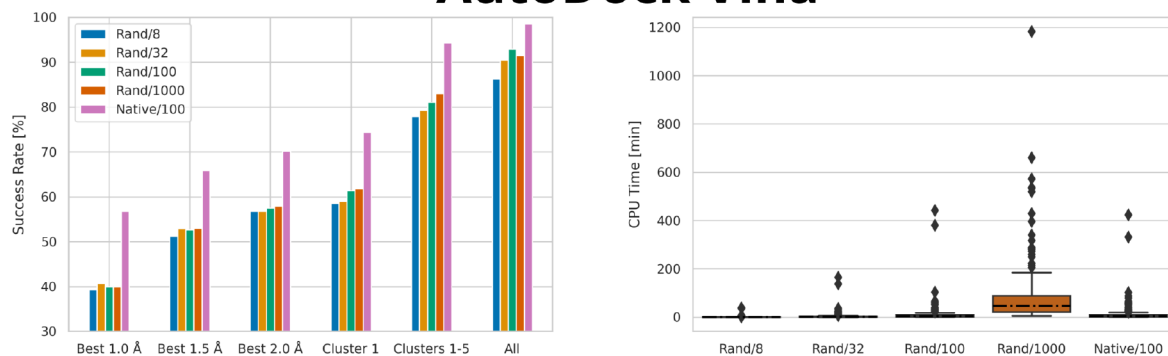
Figure 9. Redocking success rate (left) and CPU time (right) comparison of AC, GOLD, and AutoDock Vina. For each program, results from representative docking runs starting from a randomized ligand conformation (Rand) and the native ligand conformation (Native, pink) are shown. Standard parameters were used unless otherwise indicated. For GOLD, 100 GA runs were done, except for run Rand/PLP/1000, where 1000 GA runs were done. For AutoDock Vina, the respective exhaustiveness parameter is indicated. For the CPU time of AC/Rand/90°/Blind, not all outliers are shown.

fast docking as well as to high-precision docking. Its scoring function, composed of the CHARMM force field energy in combination with the FACTS implicit solvation model, has a known weakness in treating zinc metalloproteins but performs so well on other complexes that it is useful for the detection of problematic experimental data. A requirement for its good performance though is a good quality of the experimental structural data that it is supposed to reproduce. The design and preparation of a high-quality benchmark set from data of the Protein Data Bank is nontrivial, as our in-depth analysis of the PDBbind Core set demonstrates.

Compared to GOLD and AutoDock Vina, AC 2.0 achieves higher success rates for redocking of the PDBbind Core set than these two popular docking tools, albeit at higher computational cost. Starting from a randomized ligand conformation, AC reaches a success rate of 73.3% on the full PDBbind Core set (best pose RMSD $\leq$ 2 Å), GOLD/PLP a success rate of 63.9%, and AutoDock Vina a success rate of 58.2%. Both the sampling algorithm and the scoring function of AC contribute to its better performance. At variance with GOLD and AutoDock Vina, the sampling performance of AC can easily be adapted to available computational resources and systematically be improved. An explicit treatment of structural water molecules might further increase success rates.

For cross-docking, the AC 2.0 success rate on the full PDBbind Core set (42.5%) drops by about 30% due to clashes between the ligand and the protein when docking to a non-native protein structure. It is similar to the one of GOLD/PLP

(42.8%) and higher than the one of AutoDock Vina (33.1%). Our data demonstrate that the informed choice of a flexible protein environment can substantially increase the success rates. For screening applications, the performance of AC is case-dependent. It is similar to GOLD/PLP and AutoDock Vina for cases where all three codes can dock the native ligands with high confidence. A reparametrization of the weighting factors of the nonpolar and polar terms in the SwissParam score, taking decoy poses into account, may lead to an improved screening performance.

The AC 2.0 docking code will be made available to the scientific community free of charge through the SwissDock Web server (www.swissdock.ch),[35] which is currently based on the EADock DSS docking program[36] and which has been developed and maintained by the Molecular Modeling Group of the SIB Swiss Institute of Bioinformatics since 2011.

■ DATA AND SOFTWARE AVAILABILITY

The data used to generate the results of this article (coordinates, parameters, topologies, input files, docking results, analysis scripts) are available on Zenodo (DOI: 10.5281/zenodo.7940100). The AC 2.0 docking code will be made available through the SwissDock web server (www. swissdock.ch[35]) in the near future. Software and procedures mentioned can be accessed on the following websites: AutoDock tools for generating Vina input files (https:// autodocksuite.scripps.edu/adt/), AutoDock Vina docking program (https://vina.scripps.edu), CHARMM molecular simulation program (https://www.charmm.org), DockRMSD for RMSD calculation of symmetric molecules (https:// zhanggroup.org/DockRMSD/), DPI calculator (http://users. abo.fi/mivainio/shaep/download.php), server to calculate EDIAm values (https://proteins.plus/), GOLD docking program (https://www.ccdc.cam.ac.uk/solutions/software/ gold), Open Babel version 2.4.1 for ligand coordinate manipulation (https://openbabel.org/), SwissParam server for generating ligand force fields for AC (https://www. swissparam.ch/), and UCSF Chimera software for analysis and visualization (https://www.cgl.ucsf.edu/chimera).

■ ASSOCIATED CONTENT

ⓈⒾ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.3c00054.

> Technical details of AC implementation; comparison between AC 1.0 and 2.0; number of unique ligand poses per rotational angle; characterization of PDBbind Core set; PDBbind Core set scoring failures in AC; AC redocking results after removal of scoring failures; AC redocking results without removal of scoring failures; redocking results by target; cross-docking results; virtual screening results; buried ligand surface as a function of crystal contacts; comparison of AC redocking results with CHARMM22/27 and CHARMM36 force fields; AC redocking success rate by complex after removal of scoring failures; SwissParam score by target (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Ute F. Röhrig − *Molecular Modeling Group, SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland;* ⓞ orcid.org/0000-0002-4676-4087; Email: ute.roehrig@ sib.swiss

Vincent Zoete − *Molecular Modeling Group, SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland; Department of Oncology UNIL-CHUV, Lausanne University, Ludwig Institute for Cancer Research Lausanne Branch, CH-1066 Epalinges, Switzerland;* ⓞ orcid.org/0000-0002-2336-6537; Email: vincent.zoete@unil.ch

Authors

Mathilde Goullieux − *Molecular Modeling Group, SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland;* ⓞ orcid.org/0000-0001-9330-5353

Marine Bugnon − *Molecular Modeling Group, SIB Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland;* ⓞ orcid.org/0000-0002-4806-6784

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00054

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Zoete, V.; Schuepbach, T.; Bovigny, C.; Chaskar, P.; Daina, A.; Röhrig, U. F.; Michielin, O. Attracting cavities for docking. Replacing the rough energy landscape of the protein by a smooth attracting landscape. *J. Comput. Chem.* **2016**, *37*, 437−447.

(2) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(3) MacKerell, A. D.; Feig, M.; Brooks, C. L. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J. Comput. Chem.* **2004**, *25*, 1400−1415.

(4) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the Additive CHARMM All-Atom Protein Force Field Targeting Improved Sampling of the Backbone ϕ, ψ and Side-Chain χ_1 and χ_2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.

(5) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71−73.

(6) Haberthür, U.; Caflisch, A. FACTS: Fast analytical continuum treatment of solvation. *J. Comput. Chem.* **2008**, *29*, 701−715.

(7) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.;

Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(8) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726−741.

(9) Liu, Z.; Su, M.; Han, L.; Liu, J.; Yang, Q.; Li, Y.; Wang, R. Forging the Basis for Developing Protein−Ligand Interaction Scoring Functions. *Acc. Chem. Res.* **2017**, *50*, 302−309.

(10) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R. Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, *59*, 895−913.

(11) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(12) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins: Struct. Funct. Genet.* **2003**, *52*, 609−623.

(13) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.

(14) Eberhardt, J.; Santos-Martins, D.; Tillack, A. F.; Forli, S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **2021**, *61*, 3891−3898.

(15) Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: A fast force field generation tool for small organic molecules. *J. Comput. Chem.* **2011**, *32*, 2359−2368.

(16) Halgren, T. A. Merck molecular force field. III. Molecular geometries and vibrational frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553−586.

(17) Roche, O.; Kiyama, R.; Brooks, C. L. Ligand-protein database: linking protein-ligand complex structures to binding data. *J. Med. Chem.* **2001**, *44*, 3592−3598.

(18) Röhrig, U. F.; Grosdidier, A.; Zoete, V.; Michielin, O. Docking to heme proteins. *J. Comput. Chem.* **2009**, *30*, 2305−2315.

(19) Chaskar, P.; Zoete, V.; Röhrig, U. F. On-the-Fly QM/MM Docking with Attracting Cavities. *J. Chem. Inf. Model.* **2017**, *57*, 73−84.

(20) Smith, R.; Brereton, I.; Chai, R.; Kent, S. Ionization states of the catalytic residues in HIV-1 protease. *Nat. Struct. Biol.* **1996**, *3*, 946−950.

(21) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera − A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.

(22) Brünger, A. T.; Karplus, M. Polar hydrogen positions in proteins: Empirical energy placement and neutron diffraction comparison. *Proteins: Struct., Funct., Bioinf.* **1988**, *4*, 148−156.

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

(24) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.

(25) Blow, D. M. Rearrangement of Cruickshank's formulae for the diffraction-component precision index. *Acta Crystallogr. D. Biol. Crystallogr.* **2002**, *58*, 792−797.

(26) Goto, J.; Kataoka, R.; Hirayama, N. Ph4Dock: Pharmacophor−Based Protein−Ligand Docking. *J. Med. Chem.* **2004**, *47*, 6804−6811.

(27) Meyder, A.; Nittinger, E.; Lange, G.; Klein, R.; Rarey, M. Estimating Electron Density Support for Individual Atoms and Molecular Fragments in X-ray Structures. *J. Chem. Inf. Model.* **2017**, *57*, 2437−2447.

(28) Korb, O.; Stützle, T.; Exner, T. E. Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.

(29) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785−2791.

(30) Bell, E. W.; Zhang, Y. DockRMSD: An open-source tool for atom mapping and RMSD calculation of symmetric molecules through graph isomorphism. *J. Cheminf.* **2019**, *11*, 40.

(31) Chaskar, P.; Zoete, V.; Röhrig, U. F. Toward On-The-Fly Quantum Mechanical/Molecular Mechanical (QM/MM) Docking: Development and Benchmark of a Scoring Function. *J. Chem. Inf. Model.* **2014**, *54*, 3137−3152.

(32) Röhrig, U. F.; Awad, L.; Grosdidier, A.; Larrieu, P.; Stroobant, V.; Colau, D.; Cerundolo, V.; Simpson, A. J. G.; Vogel, P.; Van den Eynde, B. J.; Zoete, V.; Michielin, O. Rational Design of Indoleamine 2,3-Dioxygenase Inhibitors. *J. Med. Chem.* **2010**, *53*, 1172−1189.

(33) Röhrig, U. F.; Majjigapu, S. R.; Reynaud, A.; Pojer, F.; Dilek, N.; Reichenbach, P.; Ascencao, K.; Irving, M.; Coukos, G.; Vogel, P.; Michielin, O.; Zoete, V. Azole-Based Indoleamine 2,3-Dioxygenase 1 (IDO1) Inhibitors. *J. Med. Chem.* **2021**, *64*, 2205−2227.

(34) Wang, C.; Zhang, Y. Improving scoring-docking-screening powers of protein−ligand scoring functions using random forest. *J. Comput. Chem.* **2017**, *38*, 169−177.

(35) Grosdidier, A.; Zoete, V.; Michielin, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* **2011**, *39*, W270−277.

(36) Grosdidier, A.; Zoete, V.; Michielin, O. Fast docking using the CHARMM force field with EADock DSS. *J. Comput. Chem.* **2011**, *32*, 2149−2159.

▥ Recommended by ACS