Received: 15 September 2019 | Revised: 20 November 2019 | Accepted: 27 November 2019

DOI: 10.1002/prot.25860

PROTEINS WILEY

RESEARCH ARTICLE

Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach

Glenn Glashagen[1] | Sjoerd de Vries[2,3] | Urszula Uciechowska-Kaczmarzyk[4] | Sergey A. Samsonov[4] | Samuel Murail[2] | Pierre Tuffery[2,3] | Martin Zacharias[1]

[1]Physik-Department T38, Technische Universität München, Garching, Germany

[2]Université de Paris, CNRS UMR 8251, INSERM ERL U1133, Paris, France

[3]Ressource Parisienne en Bioinformatique Structurale (RPBS), Paris, France

[4]Faculty of Chemistry, University of Gdańsk, Gdańsk, Poland

Correspondence
Martin Zacharias, Physik-Department T38, Technische Universität München, James Franck Str. 1, 85748 Garching, Germany.
Email: martin.zacharias@ph.tum.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: SFB1035/B02; National Science Center of Poland (Narodowe Centrum Nauki), Grant/Award Number: UMO-2018/30/E/ST4/00037

Abstract

The ATTRACT protein-protein docking program has been employed to predict protein-protein complex structures in CAPRI rounds 38-45. For 11 out of 16 targets acceptable or better quality solutions have been submitted (~70%). It includes also several cases of peptide-protein docking and the successful prediction of the geometry of carbohydrate-protein interactions. The option of combining rigid body minimization and simultaneous optimization in collective degrees of freedom based on elastic network modes was employed and systematically evaluated. Application to a large benchmark set indicates a modest improvement in docking performance compared to rigid docking. Possible further improvements of the docking approach in particular at the scoring and the flexible refinement steps are discussed.

KEYWORDS

docking minimization, elastic network model, induced fit, protein-protein complex formation, protein-protein interaction

1 | INTRODUCTION

The interaction of proteins to form functional complexes is a fundamental property of all living systems. A major goal of structural biology is to obtain structural insight into all relevant protein-protein complexes of living systems. Although the structures of many protein-protein complexes have been determined experimentally there is still a demand for docking methods to predict entirely new complex structures. For many cases, it is possible to find a homologous complex in the database of known complexes to generate a template-based model.[1] However, in particular for transient protein-protein interactions or low-affinity complexes it is still difficult to determine the complex structure experimentally or based on a homologous template. Within the blind evaluation of different methods and protocols for protein-protein docking by the CAPRI (Critical Assessment of Predicted Interaction) challenge,[2-5] we employed the ATTRACT protein-protein docking approach.[6-11] In ATTRACT, an atomistic or coarse-grained (CG) protein model can be employed for predicting the targets of CAPRI rounds 38-45. The ATTRACT CG protein model represents each amino acid of a protein by up to four pseudo centers and is intermediate between a residue-level and full atomistic description.[11,12] In contrast to most protein-protein docking methods, the ATTRACT approach includes conformational flexibility of binding partners already approximately during the early systematic stage of protein-protein docking. To speed up the docking calculations, the potential energy can be precalculated on a grid and interactions are then calculated by interpolation from nearest grid points.[6,7] A docking search consists of several Monte Carlo simulations or energy minimizations starting from thousands of start configurations. It is possible to perform docking with an arbitrary number of partner molecules. The standard reduced CG representation provides a smooth protein surface representation containing a limited number of docking energy minima that allows for more rapid and fully converged energy minimization compared to an atomistic model. Besides side chain flexibility

and representation of partners by an ensemble of conformations, it is possible to include global flexibility (eg, domain-domain motion) by energy minimization along the directions of precalculated soft normal modes (simultaneous to minimization in rotational and translational degrees of freedom) for each partner structure.[10,13,14] In recent years, the possibility for global peptide-protein docking,[9,15,16] protein-DNA,[17] or protein-RNA docking[18,19] was implemented. In addition, experimental or bioinformatics data can be included directly during the search as restraints or to guide/bias the search. Thus, it is possible to account for low-resolution CryoEM density[20] or small angle Xray-scattering data.[21]

In CAPRI rounds 38-45, we provided docking predictions for all targets and tested in several cases the inclusion of global flexibility during the docking searches. The possible benefit of a systematic docking in translational and rotational degrees of freedom and simultaneous optimization in precalculated collective degrees of freedom was evaluated. Furthermore, refinement efforts at atomic resolution to improve CAPRI protein-protein docking target predictions will also be reported.

2 | MATERIALS AND METHODS

2.1 | Protein-protein docking approach

The ATTRACT coarse-grained protein model represents the main chain by two pseudo atoms per residue (located at the backbone nitrogen and backbone oxygen atoms, respectively). The side chains of small amino acids (Ala, Asp, Asn, Cys, Ile, Leu, Pro, Ser, Thr, and Val) are represented by one pseudo atom corresponding to the geometric mean of side chain heavy atoms.[11,12] Larger and more flexible side chains are represented by two pseudo atoms to better account for the shape and dual chemical character of some side chains. Effective interactions between pseudo-atoms are described by soft distance (r_{ij})-dependent Lennard-Jones (LJ)-type potentials of the following form,[12]

$$V = \varepsilon_{AB}\left[\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right] + \frac{q_i q_j}{\varepsilon(r_{ij})\, r_{ij}} \text{ in case of attractive pair}$$

repulsive pair :

$$V = -\varepsilon_{AB}\left[\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right] + \frac{q_i q_j}{\varepsilon(r_{ij})\, r_{ij}} \text{ if } r_{ij} > r_{min}$$

$$V = 2\, e_{min} + \varepsilon_{AB}\left[\left(\frac{R_{AB}}{r_{ij}}\right)^8 - \left(\frac{R_{AB}}{r_{ij}}\right)^6\right] + \frac{q_i q_j}{\varepsilon(r_{ij})\, r_{ij}} \text{ if } r_{ij} \le r_{min}$$

where R_{AB} and ε_{AB} are effective pairwise radii and attractive or repulsive Lennard-Jones parameters. At the distance r_{min} between two pseudo atoms the standard LJ-potential has the energy e_{min}. However, in case of the repulsive pair (LJ-type) potential (lower two lines in the above equation), the energy minimum is replaced by a saddle point and the potential is positive for all distances (always repulsive). Which pseudo atom pairs are attractive or repulsive has been determined by an optimization procedure based on native complexes and many decoy structures (details given in Reference 12. A Coulomb type term accounts for electrostatic interactions between real charges (Lys, Arg, Glu, and Asp) damped by a distance-dependent dielectric constant (ε = 15r, with r measured in Å). For protein-peptide docking, we used the pepATTRACT approach.[15] The pepATTRACT approach uses the same CG model for an initial systematic search for putative peptide binding regions and the iATTRACT approach for refinement (see below).

2.2 | Inclusion of global flexibility during docking using normal modes

In the ATTRACT docking program, it is possible to perform energy minimization not only in rigid body degrees of freedom but simultaneously in any combination of (Cartesian) collective degrees of freedom.[14] Collective degrees of freedom representing normal modes with small eigenvalues obtained as eigenvectors of the Hessian matrix of the proteins using an elastic network model (ENM) have been found most useful to describe global conformational changes in partner structures associated with complex formation.[22,23] Hence, the partner structures can relax (deform) along precalculated soft collective degrees of freedom during the docking search. Energy minimization is performed along the directions of a subset of normal modes simultaneously with the rigid body movement of protein partners. The standard soft degrees of freedom corresponded to eigenvectors of the proteins calculated using an ENM developed by Hinsen[22] related to Anisotropic ENMs.[23] In our implementation, the normal modes were calculated with respect to only the protein backbone (Cα atoms). This results in collective degrees of freedom that allow backbone loop motions and/or domain-domain motions. Each side chain (bound to a Cα atom) performs exactly the same (translation) movement as the corresponding Cα atom. Hence, the complete side chain moves (in translation) as a rigid body following the motion of the backbone. The method was implemented in a GPU-version of the ATTRACT program in order to achieve high computational efficiency. Note, that this includes only the search (energy minimization) along rigid degrees of freedom and along the normal mode directions (starting from thousands of initial placements) but not the Hessian matrix diagonalization to calculate the normal modes. This step needs to be performed only once for each partner prior to the docking search and usually takes only a few seconds or minutes on a standard CPU-based PC.

2.3 | Refinement of docked protein-protein or peptide-protein complexes at atomic resolution

The iATTRACT-flexible interface refinement is one standard refinement technique of docked complexes in ATTRACT.[24] It combines rigid body motions of partners directly with full atomic resolution flexibility of the predicted interface regions. Briefly, the docked protein (or peptide) partners are converted into an all-atom model with the

ATTRACT tool *aareduce* based on the OPLS force field.[25] Missing hydrogens are generated by PDB2PQR[26] and protonation states were determined by PropKa.[27] The atomistic refinement uses a physical force field based on the OPLS parameters[25] to calculate intermolecular nonbonded and electrostatic interactions between the protein partners. Contacts from the input structure are treated as flexible during a simultaneous potential energy minimization in rigid body degrees of freedom and interface flexibility. A structure-based force field is determined on-the-fly (for each complex) to evaluate intraprotein interactions for the flexible interface part.[24]

The AMBER16 molecular dynamics (MD) package[28] was used for atomistic refinement MD simulations using an implicit solvent description. The structures were converted to the AMBER atom type description using the pdb4amber tool. A Generalized-Born (GB) implicit solvent model (igb = 8) was used with the newest version of the AMBER force field ff14SB.[29] The structures were first minimized with the sander program (1000 steepest descent steps) with a short cutoff (cut = 9.0) and a small initial step size (dx0 = 0.0000001) to relax possible atom overlap and deformations. Then two short MD simulations were run with the pmemd.cuda program of the AMBER package for 5000 and 2500 time steps (2 fs) at temperatures T = 400 K and T = 350 K, respectively. During the MD simulations, intramolecular distances and intermolecular distances between backbone atoms were restrained to prevent large deformations and dissociation of the binding partners. Finally, the structures were minimized for 5000 steps with a large cutoff (cut-off-radius = 9999.0 Å) to include all pairwise interactions using the pmemd.cuda program without restraints. The force field energy was evaluated for the complex and the individual protein partners by the sander program. The binding interaction energy score was calculated by subtracting the energy of the free protein partners from the energy of the complex. The final models were ranked by their binding interaction energy score without using any reweighting scheme of the energy contributions. In case of peptide-protein complexes of round 44, we also performed in some cases a refinement protocol that involved explicit solvent simulations by all-atomic MD simulations using CHARMM36 force field[30] for protein and TIP3P model[31] for water. Models were solvated using ~8000 water molecules in a rhombic dodecahedron box at salt concentration of ~150 mM. The Gromacs 5.141 software[32] was used to run the simulations using the virtual interaction sites approach, allowing a 5 fs integration time step. Neighbor searching was performed every five steps. All bonds were constrained using the LINCS algorithm. The particle mesh Ewald algorithm was used to handle electrostatics with a 10 Å cutoff for the short-range part and a grid spacing of 1.2 Å for the long-range contribution in reciprocal space. The Verlet buffer scheme was used for nonbonded interactions, the neighbor list was updated every 10 steps. Two independent baths for protein and solvent were coupled to a temperature of 310 K using the Bussi velocity rescaling thermostat with a time constant of τ = 0.1 ps. Pressure coupling was scaled isotropically to a reference pressure of 1 bar, τ = 5.0 ps and compressibility of 4.5 10^{-5} bar^{-1}. Systems were minimized for 20 000 steps with the steepest descent algorithm, equilibrated for 500 ps with a 2 fs integration time step (and 5 fs for the following steps), using position restraints of 10 kJ mol^{-1} $Å^{-2}$ on heavy atoms with the initial model structure as a reference, followed by an equilibration of 1 ns using position restraints with a force constant of 10 kJ mol^{-1} $Å^{-2}$ on Cα atoms, and finally 2.5 ns using position restraints with a force constant of 0.1 kJ mol^{-1} $Å^{-2}$ on Cα atoms. Two production runs were computed for 100 ns without any position restraints for round 134, two production runs of 790 and 490 ns were computed for round 135. Submitted structures correspond to the structures the closest to the average structure of the peptide.

2.4 | Molecular docking of carbohydrate-protein complexes

For the docking and evaluation of carbohydrate-protein complexes (Capri round 41), we applied Autodock 3[33] to perform molecular docking calculations because at the time of the Capri challenge no carbohydrate specific force field was available in ATTRACT. Receptor structures were prepared by the Autodock Tools[34] after parameterization with the Xleap module of AMBER 16 using ff99SB force field parameters.[35] The internal charges from Autodock 3 were assigned, while nonpolar hydrogen atoms were merged on corresponding carbon atoms using a default procedure. Ligands were prepared using the Xleap module of AMBER 16 with GLYCAM06 parameters.[36] For oligosaccharide ligands, the charges assigned in this procedure were kept unaltered in the docking calculation. The atomic potential grid of the protein receptor was calculated with a spacing of 0.375 Å and consisted of 90 x 90 x 90 grid points. The gridbox center was manually placed in the approximate center of the ligand binding site for each of the proteins to allow for substantial rotational and translational degrees of freedom for the docked oligosaccharide ligands within the binding site. The Lamarckian genetic algorithm was used as the search algorithm, with an initial population size of 300 and a termination condition of 10^5 generations or 9995 x 10^5 energy evaluations. These parameters previously proved to be successful when applied for linear periodic polysaccharide ligands.[37] All ligands were kept fully flexible in docking simulations, while the receptor structures were kept rigid. A total of 1000 independent docking runs were performed. For further analysis, 50 top scored docking solutions were taken into account. Spatial clustering of the docking solutions was performed using the DBSCAN algorithm[38] with parameters manually chosen for each protein-oligosaccharide pair in a way to obtain up to 5-7 representative clusters. Solutions for AbnE were further filtered so that longer oligo-arabinose molecule poses represented elongation of the shorter ones. The distance metric between two structures used in the clustering procedure was defined as the root-mean-square deviation of atomic distances, while pairing up nearest atoms of the same type, which is different from the classical RMSD of atomic distances that pairs up atoms with the same atomic ID. The use of such a metric is beneficial for periodic linear polysaccharides.[39] From each cluster, one representative docking solution was chosen for the subsequent analysis.

2.5 | Molecular dynamics refinement of docked protein-carbohydrate complexes

The structures obtained by molecular docking were further analyzed by all-atomic MD simulations using Amber16. Periodic boundary conditions in a truncated octahedron TIP3P water box with at least 6 Å distance from the solute to the periodic box border were used. Counter-ions were used to neutralize the system. ff99SB force field parameters for protein and the GLYCAM06 for oligosaccharides were applied, respectively. Prior to an MD production run, two energy-minimization steps were performed: first, 0.5×10^3 steepest descent cycles and 10^3 conjugate gradient cycles with harmonic force restraints on solute (10 kcal·mol^{-1} Å^{-2}), then, 3×10^3 steepest descent cycles and 3×10^3 conjugate gradient cycles without constraints. After the minimization, a system was heated up to 300 K for 10 ps with harmonic force restraints on solute (10 kcal·mol^{-1} Å^{-2}), equilibrated for 100 ps at 300 K and 10^5 Pa in isothermal isobaric ensemble (NTP). This was followed by a 10 ns MD production run in the same NTP ensemble. The SHAKE algorithm, 2 fs time integration, 8 Å cutoff for nonbonded interactions and the Particle Mesh Ewald method were used. The trajectories were analyzed using the cpptraj module of AMBER Tools.

2.6 | Scoring of protein-carbohydrate complexes

In order to score the docking poses, MD trajectory postprocessing calculations were performed. This accounts for the often observed significant conformational flexibility of protein-carbohydrate complexes. For this, free energy calculations were carried out using Molecular Mechanics-Generalized Born Surface Area model igb = 2[40] for protein-oligosaccharide complexes. The obtained energy values (often also termed "free energy") account explicitly for the enthalpy and implicitly for the solvent entropy and were used to score and rank the poses.

3 | RESULTS AND DISCUSSION

3.1 | Inclusion of ENM-based normal modes during protein-protein docking

The idea of using conformational relaxation of partner structures in soft normal modes to account for partner flexibility during docking dates back to Zacharias and Sklenar.[41] In particular, in the field of protein-protein docking, the methodology has been further developed in several studies[13,14,42] and implemented in docking approaches such as ATTRACT[14] and SwarmDock.[42] In the ATTRACT approach, it is possible to simultaneously energy-minimize partner structures in rotational, translational, and collective degrees of freedom during docking. The collective degrees of freedom are typically precalculated using an ENM for the partner structures. This can be performed very rapidly and often the softest modes (those with smallest eigenvalue) of the ENM show good overlap with the conformational (backbone) structural difference between unbound and bound partner structures. Indeed, in previous studies, it was demonstrated that the inclusion of soft collective modes during docking of proteins can improve the docking performance. However, this concerned mostly preselected small sets of cases where it is known in advance that protein-protein association involved considerable global protein backbone changes.[14] With our recent GPU-based implementation protein-protein docking based on energy minimization starting from tens of thousands of start configurations can be performed within a few minutes of computer time for a medium sized complex including also energy minimization in normal modes. Hence, it is possible to test the approach systematically on whole benchmarks of complexes before applying it on Capri targets.

For a benchmark set of 207 complexes of the benchmark 5 set[43] with known unbound and bound structures (excluding complexes with >1000 residues, see Table S1), we first checked how well a subset of calculated softest ENM modes overlaps with the observed difference in protein backbone of unbound vs bound structures (Figure S1). Using just the first five modes results in many cases in overlaps of 20% to 75%. Overlap means, here, how much of the conformational difference between bound and unbound structure can be presented by a (best possible) linear combination of deformations in the first five modes (it corresponds to the projection of the conformational difference vector onto the subspace of the first five orthogonal modes). It agrees qualitatively well with previous studies on other protein structures and using similar ENMs[44-46] and indicates that in principle a subset of ENM-derived modes describes global conformational differences between unbound and bound structures often reasonably well.

For evaluating the benefit of including normal mode flexibility during docking, we compare the number of acceptable, medium and high quality docking solutions (Capri criteria) using rigid unbound docking (with the same start structures) as reference. In each case, systematic docking by energy minimization is started from tenth of thousand starting structures (see section 2). In ATTRACT, the motion in collective degrees of freedom is controlled on the one hand by the intermolecular sterical interactions (the proteins can move in the mode directions driven by intermolecular forces). In addition, the deformation in each mode is penalized by a fourth order function in the deformation magnitude multiplied by the eigenvalue (force constant) of the selected mode.[14] We found in previous studies[10,14] that such a penalty avoids unrealistically large deformations during docking and at the same time allows for small deformations with very little penalty. The variation of the force constant for this penalty indicates a scaling of 0.7 to give optimal results (Figure 1A). Also considering the whole benchmark inclusion of only the softest (first) mode for both protein partners gave an overall improvement of docking performance by ~12% (Figure 1B) whereas inclusion of more modes resulted in overall poorer performance compared to rigid docking. A strict criterion was used to measure the performance: only those cases were counted as success (of flexible docking) for which rigid docking gave no acceptable (or better) solution but inclusion of normal mode flexibility gave at least one acceptable (or medium, or high quality)

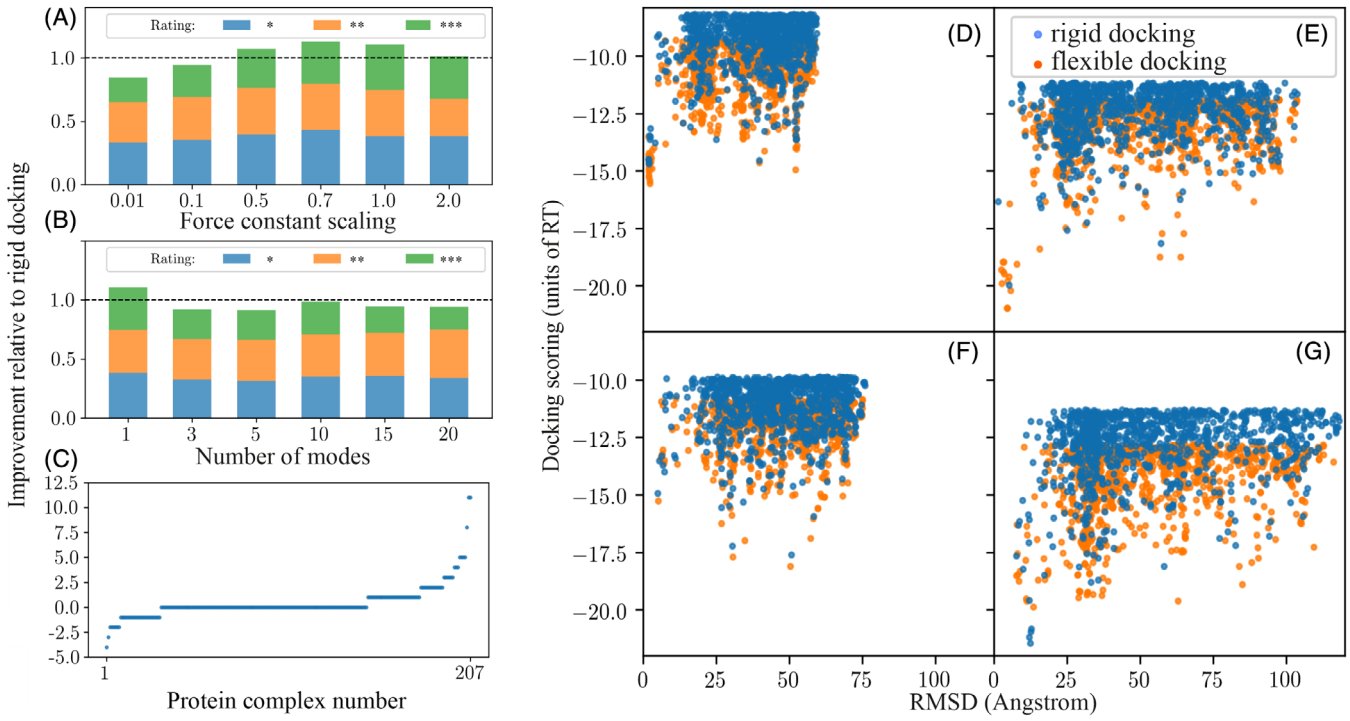

FIGURE 1 Testing inclusion of harmonic mode flexibility on docking performance for a docking benchmark set of 207 complexes using unbound partner structures. A, Average docking performance relative to rigid docking upon variation of the force constant to control the deformation magnitude in normal mode directions. A very strict criterion for improved docking performance was used: a success of flexible vs rigid docking was only counted if rigid docking gave no acceptable (or better) solution and flexible docking gave at least one acceptable docking solution. B, Effect of inclusion of increasing number of modes on average docking performance (average over all 207 cases). C, Increase or decrease (negative numbers) of successful docking cases in the top 50 solutions for each of the 207 complexes relative to rigid docking (0 means no improvement compared to rigid docking). The cases are ordered with respect to increasing performance of the flexible docking approach (performance means here the increase/decrease of the absolute number of acceptable, medium or high quality docking solutions in the top 50 relative to rigid docking). Results are given for inclusion of only the softest mode for both partner proteins because only for this case an average improvement relative to rigid docking is observed (see B). D-G, Ranking of top1000 docking solutions for pdb3pc8 (D), 3mxw (E), 1zhi (F), and 2fd6 (G) vs L-rmsd (RMSD of smaller protein after best superposition on larger partner protein) upon rigid docking (blue dots) and including normal mode flexibility (1 mode for each partner, orange dots). In cases D and E, the ranking of near native solutions improves relative to incorrect solutions whereas in case F and G incorrect docking solutions (large RMSD) improve more than near-native solutions

prediction according to Capri criteria in the top 50 scored solutions. For example, for rigid docking, we found at least acceptable docking solutions in the top 50 for 107 cases compared to 120 successful cases if the softest normal mode for both receptor and ligand were included (translates to 12% improvement). Analysis of each individual docking case in the benchmark indicates that for the majority of cases little or no visible improvement compared to rigid docking was observed (Figure 1A,B). Hence, the number of cases with successful docking (*, **, or ***) among the top 50 scoring solutions does not change significantly (even slightly drops with increasing number of modes, Figure 1B). It is interesting to note that the total number of at least acceptable solutions among all top 50 solutions increases significantly upon inclusion of normal mode flexibility (Figure S2). Hence, in those cases for which inclusion of normal mode flexibility is really helpful the number of successful solutions in the top 50 can increase significantly. It demonstrates that for a subset of complexes a significant improvement but for another set a significant drop in docking performance was observed (shown for the case of including just the softest mode in Figure 1C). The simple reason for this result is that both correctly docked but also incorrect complexes can benefit from the inclusion of normal mode flexibility.

For a few examples, the effect of including normal mode flexibility on the ranking vs L_rmsd is illustrated (Figure 1D-G). In the example case of pdb3pc8 and pdb3mxw, an improvement of the ranking of near-native solutions is observed but for pdb1zhi and 2fd6, the ranking of several incorrect solutions improves more than for the near-native binding modes. Although the average improvement of the docking results (without including any prior information on the target) is only modest it was included in several of the Capri targets.

3.2 | Capri targets and predictions

In the Capri rounds 38-45, predictions were submitted for targets 121-136. The targets included protein-protein but also protein-peptide and protein-carbohydrate complexes. Our prediction results

are summarized in Table 1. In each case, different docking strategies were used because the difficulties for each target varied considerably.

3.3 | Round 38 and 39 (targets 121-124)

A number of Capri targets were protein peptide complexes. This includes Capri round 38/target 122 that consisted of the TolAIII domain, a subdomain of the periplasmic protein TolA from *Pseudomonas aeruginosa*, in complex with a peptide (sequence: ADPLVISSGNDRA). Only the sequence of both partners was provided and template-based modeling of the TolAIII domain was required prior to docking following the pepATTRACT method.[15] Our best model showed an interface RMSD of ~5 Å and a fnat ~0.16 and was counted as incorrect result (Table 1). A direct comparison to the experimental structure is not yet possible but we assume the failure is likely due to an inaccurate modeled TolAIII receptor structure.

The first target (122) of round 39 corresponded to a trimeric interleukin (IL) 23 complex consisting of IL23α, IL12β, and IL23R. Based on known structures of related interleukins such as IL12 and IL6, it was possible for this target to limit the number of docking starting geometries. We employed the option in ATTRACT to perform multibody minimization. The final complexes were subject to atomistic refinement simulations as described in the section 2. For this target, our docking efforts provided several acceptable solutions with only limited fraction of native contacts (fnat ~0.11) but among the best predictions for the partner placement (L_rmsd~12 Å for model 4, Figure 2) compared to the experimental structure.[47] Most other predictions for this target including the medium quality solutions indicated a larger L_rmsd.

Targets 123 and 124 corresponded to subdomains of PorM that is part of a bacterial secretion system, in complex with single chain camelid nanobodies. For target 123, this was the N-terminal PorM domain and for target 124 the C-terminal domain as a dimer. In both cases, structural models for the PorM domain and the nanobodies had to be generated by homology modeling using MODELLER.[48] None of our submissions were acceptable. Comparison of the homology-modeled partner structures especially of the PorM domains with the published complex structure indicated significant differences (Rmsd >10 Å) that caused or contributed to the failure for these targets.

3.4 | Round 40 (target 125)

Round 40 included a single target (125) that consisted of a hetero-hexameric complex formed by the extracellular domain of the LLT1

TABLE 1 Results of CAPRI predictions

Target	Type	Best model	fnat	IRMSD (Å)	Classification
121	Peptide-protein	2	0.16	5.3	incorrect
122	Protein-protein	4	0.11	0.98	acceptable quality (*)
123	Protein-protein	9	60	16.7	incorrect
124	Protein-protein	6	0	13.1	incorrect
125.1	Protein-protein	4(70)	0.44(0.69)	1.8(0.99)	medium (**)(high(***))
125.2	Protein-protein	4	0.89	0.26	high (***)
125.3/4	Protein-protein	3	0	10.1	incorrect
126	Protein-carbohydrate	1	0.22	4.4	acceptable (*)
127	Protein-carbohydrate	1	0.29	4.3	acceptable (*)
128	Protein-carbohydrate	4	0.3	2.3	medium (**)
129	Protein-carbohydrate	7	0.24	3.6	acceptable (*)
130	Protein-carbohydrate	5	0.23	2.3	acceptable (*)
131	Protein-protein	(31)	(0.59)	(3.9)	incorrect (but ** in 90-100)
132	Protein-protein	(69)	(0.43)	(2.6)	incorrect (but ** in 90-100)
133	Protein-peptide	5	0.55	1.5	medium (**)
134	Protein-peptide	1	0.87	0.53	high (***)
135	Protein-peptide	4	0.90	0.66	high (***)
136.1	Protein-protein	1	0.11	4.2	acceptable (*)
136.2	Protein-protein	4	0.14	4.1	acceptable (*)
136.3	Protein-protein	5	0	4.0	incorrect

Notes: IRMSD is the root mean square deviation between prediction and experiment of protein backbone atoms within 10 Å of the protein-protein interface. Models not within the top 10 submitted complexes are given in parenthesis. For a definition of acceptable (*), medium (**), and high (***) quality solutions see Reference 5.

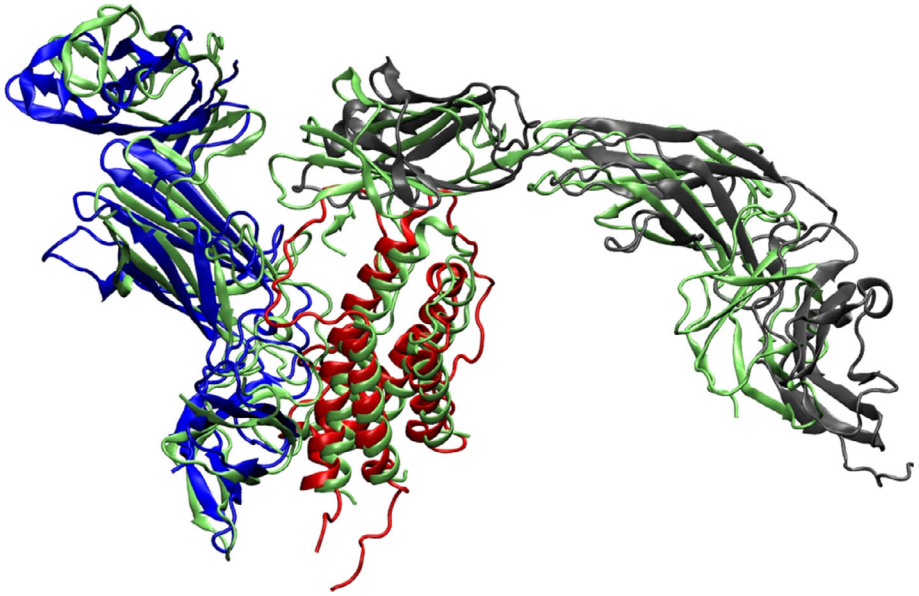

FIGURE 2 Superposition of the best predicted model (4) for target 122 (complex of IL23α [red cartoon], IL12β [blue], and IL23R [gray]) onto the experimental complex (pdb5mzv,[47] narrow green cartoon)

protein with the extracellular domain of its inhibitor NKR-P1. Initial complex structures for this target were generated based on a template and based on symmetry. The generated initial models were energy minimized using ATTRACT and subsequently further refined at atomic resolution following the protocol described in section 2. We were able to predict two of the interacting protein surfaces in the complex (out of four possible) with several medium (**) and high quality (***) solutions, respectively (Table 1). Similar to all other predictors, we were unable to correctly predict the other two interfaces correctly.

3.5 | Round 41 (targets 126-130)

The targets of round 41 consisted entirely of protein domains in complex with carbohydrate chains. Although we designed recently a coarse-grained ATTRACT force field for protein-carbohydrate docking at the time of the round 41 challenge no such force field was available. In case of the AbnB protein (1,5-alpha-arabinanase catalytic domain), the structure of the 1,5-alpha-arabinanase catalytic mutant (AbnBE201A) complexed to arabinotriose (pdb3d5z, res. 1.9 Å) served as protein receptor template. The structure of the protein was energy-minimized in the absence and in the presence of arabinotriose using the Amber16 package (see section 2). All-atom RMSD between the resulting two structures was 1.07 Å, which predominantly accounted for the residues side chains involved in the arabinotriose binding. Both structures were considered for molecular docking simulation allowing for conformational variance in the oligosaccharide binding site following the approach described in section 2. A Ca2+ ion was included as a part of the receptor structure. For the second receptor protein, AbnE, no experimental structure was available, and therefore we modeled its structure based on three homologous structures (pdb1eu8, res. 1.9 Å; pdb5ci5, 1.61 Å; pdb5f7v, 1.4 Å) using MODELLER.[48] Percent identity between all the sequences including the target protein was in the range of 20% to 30%. Similar to the AbnB protein target, two structures were selected for molecular docking simulation based on the best MODELLER scores. With our docking protocol, we achieved acceptable solutions for all targets and even medium quality for the target 128 case. Since the approximate binding geometry was in all cases correctly identified, further improvement is possible by an optimization of the force field and refinement procedure.

3.6 | Round 42 (targets 131, 132)

Targets 131 and 132 corresponded to the cell adhesion proteins HopQ type I and HopQ type II in complex with the human cell adhesion protein CEACAM1.[49] For the CEACAM1 and the HopQ type I unbound structures, pdb2gk2 and pdb5lp2, respectively, were available. However, to create a suitable structure for docking missing segments in the unbound HopQ I were generated using template based modeling with MODELLER.[48] The HopQ type II partner structure was generated using template-based modeling with pdb5lp2 as template. Unfortunately, the template-based modeling of HopQ II and of missing loop segments in case HopQ I generated model structures that deviated from the bound structures especially at the protein interface region (Figure 3). During systematic docking, the softest normal mode of each partner was included as flexible degree of freedom. In both target 131 and target 132 cases, no acceptable (or better) docking solution were among the top 10 models that we submitted. However, inspection of all 100 submitted models indicated several very good (medium quality) predictions (Table 1). These docking solutions were found as putative geometries during docking including normal mode flexibility with reasonable score because the normal mode flexibility allowed structural relaxation at the interface (illustrated in Figure 3B). Upon atomic resolution refinement, the solutions still showed some sterical clashes and were therefore not included in the top 10 submitted models (Figure 3C). As described in section 2, our standard atomistic refinement allowed only limited protein backbone motions and hence in some case no full relaxation of residual strain.

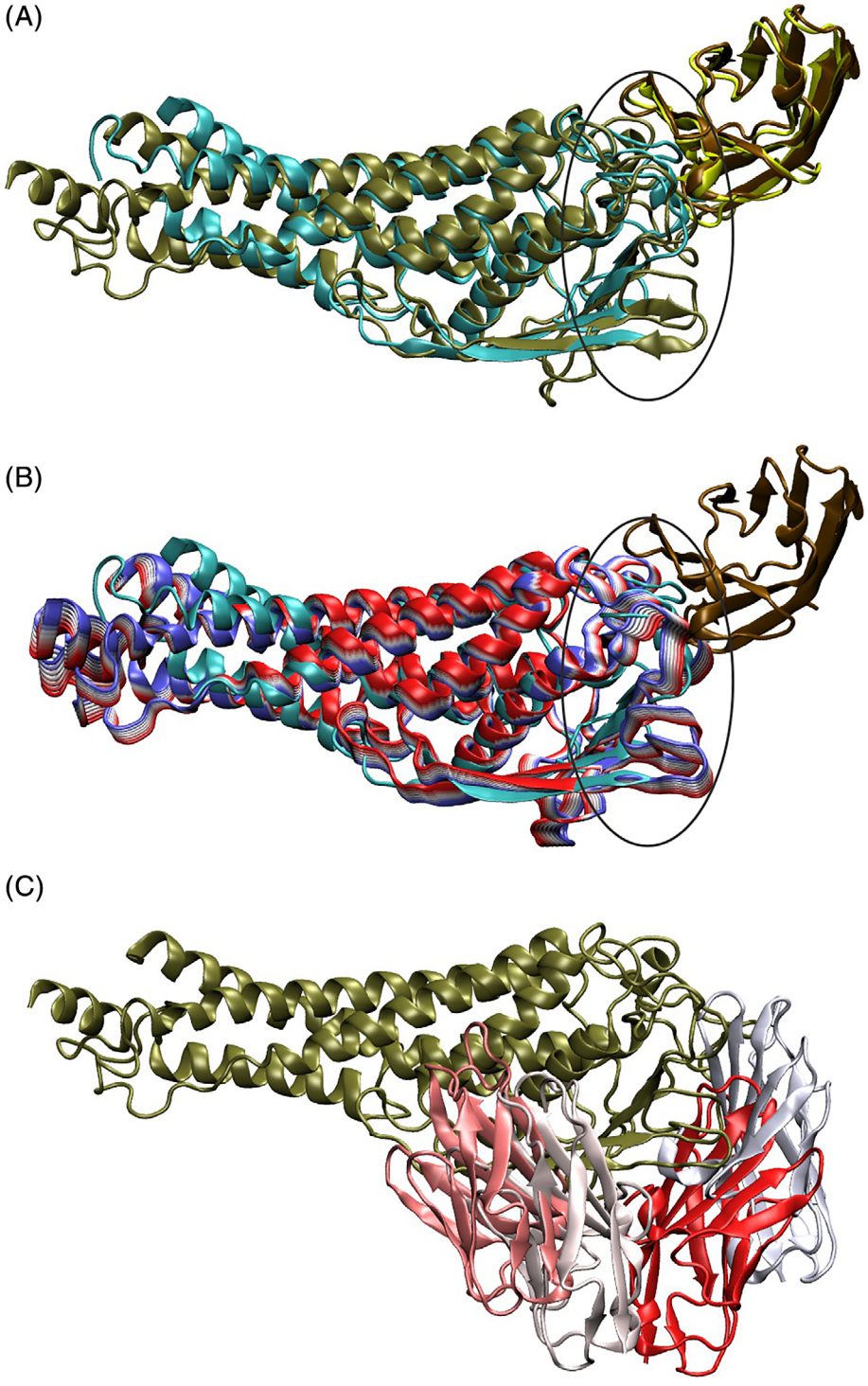

FIGURE 3 A, Superposition of predicted model 69 for docking target 132 (HopQ type II in complex with CECAM1) onto the experimental complex structure (pdb6gbh[49]). The HopQ type II is in cyan (experiment) and tan (prediction) whereas the CECAM1 partner is shown in gray (experiment) and yellow (prediction). The structural difference especially at the interface is indicated (black line). B, Superposition of several unbound HopQ type II structures deformed along the softest normal mode (colored red to blue) onto the experimental complex structure (same color as in [A]). The large deformability at the interface in the softest mode is encircled (black line). C, Superposition of example docking solutions among our top five predictions for target 132 (different color from red to white for the docked CECAM1 molecule, HopQ type II in tan color)

3.7 | Round 43 (target 133)

The single target of round 43 corresponded to a redesigned version of the Colicin E2 DNase-Im2 complex[50] where both partners have been redesigned at the level of the sequence relative to the wild-type complex. For this system, we used the partner structures in the wild-type complex (pdb3u43[50]) as templates to generate docking partner structures. The systematic search was restricted to the regions near the mutated residues on both partners. After atomistic refinement, we were able to provide a medium (**) quality solution among the top 5 docking solutions (Table 1).

3.8 | Round 44 (targets 134-135)

The targets of round 44 were complexes of the Dynein Light Chain subunit 8 (DLC8; a homodimer) with peptide segments from a cognate protein. The DCL8 was crystallized starting from both, a 50-residue

segment of the interacting protein and from a 12-residue peptide from the same segment. The resulting complexes adopt very similar structures. Hence for the target 134, the task was first to predict which 12 mer subsegment of the 50 residue segment binds to DCL8 and to predict the structure of the resulting complex. For the second target 135, the sequence of the 12 residues binding peptide was given and the task was to predict just the complex of the 12 mer with DCL8.

There are several structures of DCL8 with bound peptides available in the protein data base and all of them show the same binding mode with the peptide bound as an extended beta-sheet in groove between the subunits of the homodimer (stoichiometry two peptides per DCL8 dimer). It was reasonable to assume that also for targets 134 and 135, the binding mode will be conserved and we used a template-based protocol instead of a docking approach for this target. Based on the program Pep-Fold,[51] extended beta-sheet regions in the 50-residue sequence were predicted. Using an alignment of the known DCL8-peptide complexes, conserved residue positions in the bound peptides were identified and this allowed us to predict the most likely bound peptide segment and the register of binding. For refinement, we used both implicit and explicit solvent approaches (see section 2). For target 134, our top 1 model was a high (***) quality model (second: medium quality model). For target 135, we got three high quality models and seven medium quality models in the top 10 predictions.

3.9 | Round 45 (target 136)

Round 45 consisted of a single target, a bacterial lysine decarboxylase LdcA,[52] that forms a homo-decamer with D5 symmetry determined by Gutsche and coworkers using CryoEM. For this target, several homologous structures were available that could serve as template.[53] We used pdb2vyc (an arginine decarboxylase)[53] as template to generate starting homology models. These were used as start structures for rigid body minimization using ATTRACT and subsequent atomistic refinement using our standard atomistic refinement protocol (see section 2). At least for the interfaces 1 and 2 but not for interface 2 (Table 1), we obtained acceptable quality for all top 10 models.

4 | CONCLUSION AND OUTLOOK

In the course of Capri rounds 38-45 with 16 targets, we were able to provide docking solutions of acceptable or better quality for 11 cases within the top 10 solutions. For three targets, our predictions were of high quality. It indicates the high quality of our docking approaches and proves its versatility in dealing with a variety of different targets ranging from large protein-protein complexes, peptide-protein complexes to association of carbohydrates with protein receptors. The failure in some CAPRI docking cases can be likely attributed to inaccuracies in homology modeling of proteins partners used for docking and to significant conformational changes upon association but is also due to limitations of the scoring scheme. The ATTRACT docking approach includes the standard option to account for conformational changes during docking using energy minimization in precalculated soft harmonic modes of each partner based on an ENM. The CAPRI challenge has also triggered our efforts to systematically test the benefit of including such normal harmonic modes as additional variables on top of the rigid degrees of freedom. For one Capri round, the inclusion allowed us to identify medium quality solutions among the top 100 (but not among top 10). The systematic application to a large benchmark set yielded an overall modest improvement in terms of cases where at least an acceptable (or better) solution was found compared to rigid docking (~12% more cases upon inclusion of just the softest mode for each partner). It indicates that indeed in several cases near-native docking solutions significantly benefit but in several other cases incorrect docking solutions benefit even more than near-native solutions from inclusion of global harmonic modes. Application of more advanced normal mode methods such as the nonlinear rigid block normal mode method[54] might be helpful to improve the performance of the flexible docking approach. Furthermore, improvements of the scoring function or the combination with alternative scoring schemes may also enhance the flexible docking approach and will be the subject of future studies.

ACKNOWLEDGMENTS

We thank the organizers of the CAPRI challenge for this opportunity and the assessors for the hard work to evaluate the predictions. We thank all structural biologists who contributed target structures for the CAPRI experiment. The work was supported in part by a grant (SFB1035/B02) to M.Z. from the DFG (Deutsche Forschungsgemeinschaft) and by the CIPSM (Center of Integrated Protein Science Munich) excellence cluster. This work was supported to S.A.S. by the grant UMO-2018/30/E/ST4/00037 from the National Science Center of Poland (Narodowe Centrum Nauki). Computational resources were provided by the Polish Grid Infrastructure (PL-GRID, grants protgags, gagstr), ZIH at TU Dresden (grant p_gag).

ORCID

Martin Zacharias https://orcid.org/0000-0001-5163-2663

REFERENCES

1. Kundrotas PJ, Anishchenko I, Dauzhenka T, et al. Dockground: a comprehensive data resource for modeling of protein complexes. *Protein Sci*. 2018;27:172-181.
2. Lensink MF, Velankar S, Wodak SJ. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins Struct Funct Bioinforma*. 2017;85:359-377.
3. Lensink MF, Wodak SJ. Docking and scoring protein interactions: CAPRI 2009. *Proteins*. 2010;78:3073-3084.
4. Lensink MF, Méndez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins*. 2007;69:704-718.
5. Janin J. Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein Sci*. 2005;14:278-283.
6. de Vries SJ, Schindler CEM, Chauvot de Beauchêne I, Zacharias M. A web Interface for easy flexible protein-protein docking with ATTRACT. *Biophys J*. 2015;108:462-465.
7. Vries S, de Zacharias M. Flexible docking and refinement with a coarse-grained protein model using ATTRACT. *Proteins*. 2013;81:2167-2174.

8. Schneider S, Saladin A, Fiorucci S, Prévost C, Zacharias M. ATTRACT and PTOOLS: open source programs for protein–protein docking. In: Baron R, ed. *Computational Drug Discovery and Design. Methods in Molecular Biology*. New York, NY: Springer New York; 2012:221-232.
9. Schindler CEM, Beauchêne IC, de Vries SJ, de Zacharias M. Protein-protein and peptide-protein docking and refinement using ATTRACT in CAPRI. *Proteins*. 2017;85:391-398.
10. May A, Zacharias M. Protein–protein docking in CAPRI using ATTRACT to account for global and local flexibility. *Proteins*. 2007;69:774-780.
11. Zacharias M. Protein–protein docking with a reduced protein model accounting for side-chain flexibility. *Protein Sci*. 2003;12:1271-1282.
12. Fiorucci S, Zacharias M. Binding site prediction and improved scoring during flexible protein–protein docking with ATTRACT. *Proteins*. 2010;78:3131-3139.
13. May A, Zacharias M. Accounting for global protein deformability during protein–protein and protein–ligand docking. *Biochim Biophys Acta BBA Proteins Proteom*. 2005;1754:225-231.
14. May A, Zacharias M. Energy minimization in low-frequency normal modes to efficiently allow for global flexibility during systematic protein–protein docking. *Proteins*. 2008;70:794-809.
15. Schindler CEM, de Vries SJ, Zacharias M. Fully blind peptide-protein docking with pepATTRACT. *Structure*. 2015;23:1507-1515.
16. de Vries SJ, Rey J, Schindler CEM, Zacharias M, Tuffery P. The pepATTRACT web server for blind, large-scale peptide–protein docking. *Nucleic Acids Res*. 2017;45:W361-W364.
17. Setny P, Bahadur RP, Zacharias M. Protein-DNA docking with a coarse-grained force field. *BMC Bioinformatics*. 2012;13:228.
18. Setny P, Zacharias M. A coarse-grained force field for protein–RNA docking. *Nucleic Acids Res*. 2011;39:9118-9129.
19. Setny P, Zacharias M. Elastic network models of nucleic acids flexibility. *J. Chem. Theory Comput*. 2013;9:5460-5470.
20. de VSJ, Zacharias M. ATTRACT-EM: a new method for the computational assembly of large molecular machines using Cryo-EM maps. *PLOS One*. 2012;7:e49733.
21. Schindler CEM, de Vries SJ, Sasse A, Zacharias M. SAXS data alone can generate high-quality models of protein-protein complexes. *Structure*. 2016;24:1387-1397.
22. Hinsen K. Analysis of domain motions by approximate normal mode calculations. *Proteins*. 1998;33:417-429.
23. Bahar I, Lezon TR, Yang L-W, Eyal E. Global dynamics of proteins: bridging between structure and function. *Annu Rev Biophys*. 2010;39:23-42.
24. Schindler CEM, Vries SJ, de Zacharias M. iATTRACT: simultaneous global and local interface optimization for protein–protein docking refinement. *Proteins*. 2015;83:248-258.
25. Jorgensen WL, Maxwell DS, Tirado-Rives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc*. 1996;118:11225-11236.
26. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res*. 2004;32:W665-W667.
27. Li H, Robertson AD, Jensen JH. Very fast empirical prediction and rationalization of protein pKa values. *Proteins*. 2005;61:704-721.
28. Case DA, Cheatham TE, Darden T, et al. The Amber biomolecular simulation programs. *J Comput Chem*. 2005;26:1668-1688.
29. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput*. 2015;11:3696-3713.
30. Huang J, MacKerell AD. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem*. 2013;34:2135-2145.
31. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys*. 1983;79:926-935.
32. Berendsen HJC, van der SD, van DR. GROMACS—a message-passing parallel molecular-dynamics implementation. *Comput Phys Commun*. 1995;91:43-56.
33. Morris GM, Goodsell DS, Halliday RS, et al. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*. 1998;19:1639-1662.
34. Morris GM, Huey R, Lindstrom W, et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem*. 2009;30:2785-2791.
35. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*. 2006;65:712-725.
36. Kirschner KN, Yongye AB, Tschampel SM, et al. GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J Comput Chem*. 2008;29:622-655.
37. Sage J, Mallèvre F, Barbarin-Costes F, et al. Binding of chondroitin 4-sulfate to Cathepsin S regulates its enzymatic activity. *Biochemistry*. 2013;52:6487-6498.
38. Ester M, Kriegel H-P, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. 6.
39. Samsonov SA, Gehrcke J-P, Pisabarro MT. Flexibility and explicit solvent in molecular-dynamics-based docking of protein–glycosaminoglycan systems. *J Chem Inf Model*. 2014;54:582-592.
40. Onufriev A, Case DA, Bashford D. Effective born radii in the generalized born approximation: the importance of being perfect. *J Comput Chem*. 2002;23:1297-1304.
41. Zacharias M, Sklenar H. Harmonic modes as variables to approximately account for receptor flexibility in ligand–receptor docking simulations: application to DNA minor groove ligand complex. *J Comput Chem*. 1999;20:287-300.
42. Moal IH, Bates PA. SwarmDock and the use of normal modes in protein-protein docking. *Int J Mol Sci*. 2010;11:3623-3648.
43. Vreven T, Moal IH, Vangone A, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J Mol Biol*. 2015;427:3031-3041.
44. Tobi D, Bahar I. Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. *Proc Natl Acad Sci USA*. 2005;102:18908-18913.
45. Dobbins SE, Lesk VI, Sternberg MJE. Insights into protein flexibility: the relationship between normal modes and conformational change upon protein–protein docking. *Proc Natl Acad Sci USA*. 2008;105:10390-10395.
46. Grudinin S, Laine E, Hoffmann A. Predicting protein functional motions: an old recipe with a new twist. 2019. Available from https://hal.archives-ouvertes.fr/hal-02291552
47. Bloch Y, Bouchareychas L, Merceron R, et al. Structural activation of pro-inflammatory human cytokine IL-23 by cognate IL-23 receptor enables recruitment of the shared receptor IL-12Rβ1. *Immunity*. 2018;48:45-58.e6.
48. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234:779-815.
49. Moonens K, Hamway Y, Neddermann M, et al. Helicobacter pylori adhesin HopQ disrupts trans dimerization in human CEACAMs. *EMBO J*. 2018;37:e98665.
50. Wojdyla JA, Fleishman SJ, Baker D, Kleanthous C. Structure of the ultra-high-affinity Colicin E2 DNase–Im2 complex. *J Mol Biol*. 2012;417:79-94.
51. Maupetit J, Derreumaux P, Tuffery P. PEP-FOLD: an online resource for de novo peptide structure prediction. *Nucleic Acids Res*. 2009;37:W498-W503.

52. Kandiah E, Carriel D, Perard J, et al. Structural insights into the *Escherichia coli* lysine decarboxylases and molecular determinants of interaction with the AAA+ ATPase RavA. *Sci Rep*. 2016; 6:24601–24609.
53. Andréll J, Hicks MG, Palmer T, Carpenter EP, Iwata S, Maher MJ. Crystal structure of the acid-induced arginine decarboxylase from *Escherichia coli*: reversible Decamer assembly controls enzyme activity. *Biochemistry*. 2009;48:3915-3927.
54. Hoffmann A, Grudinin S. NOLB: nonlinear rigid block Normal-mode analysis method. *J Chem Theory Comput*. 2017;13:2123-2134.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Glashagen G, de Vries S, Uciechowska-Kaczmarzyk U, et al. Coarse-grained and atomic resolution biomolecular docking with the ATTRACT approach. *Proteins*. 2019;1–11. https://doi.org/10.1002/prot.25860