Received: 10 December 2019 | Revised: 4 February 2020 | Accepted: 22 February 2020

DOI: 10.1002/jcc.26187

FULL PAPER

Journal of COMPUTATIONAL CHEMISTRY WILEY

Prediction of protein–protein complexes using replica exchange with repulsive scaling

Till Siebenmorgen | Michael Engelhard | Martin Zacharias

Physik-Department T38, Technische Universität München, Garching, Germany

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

Funding information
DFG (German Research Foundation) grant SFB (Collaborative Research Center), Grant/Award Number: 863 (project A10)

Abstract

The realistic prediction of protein–protein complex structures is import to ultimately model the interaction of all proteins in a cell and for the design of new protein–protein interactions. In principle, molecular dynamics (MD) simulations allow one to follow the association process under realistic conditions including full partner flexibility and surrounding solvent. However, due to the many local binding energy minima at the surface of protein partners, MD simulations are frequently trapped for long times in transient association states. We have designed a replica-exchange based scheme employing different levels of a repulsive biasing between partners in each replica simulation. The bias acts only on intermolecular interactions based on an increase in effective pairwise van der Waals radii (repulsive scaling (RS)-REMD) without affecting interactions within each protein or with the solvent. For a set of five protein test cases (out of six) the RS-REMD technique allowed the sampling of near-native complex structures even when starting from the opposide site with respect to the native binding site for one partner. Using the same start structures and same computational demand regular MD simulations sampled near native complex structures only for one case. The method showed also improved results for the refinement of docked structures in the vicinity of the native binding geometry compared to regular MD refinement.

KEYWORDS

binding free energy, docking refinement, free energy simulation, implicit solvent modeling, molecular dynamics simulation

1 | INTRODUCTION

Biomolecular binding and in particular protein–protein binding processes to form functional complexes are key elements of almost all biological processes. Knowledge of the three-dimensional (3D) structure of protein–protein complexes is a prerequisite for understanding its function. Experimental structure determination as well as prediction of protein–protein complex structures are also of significant interest for the rational design of drug molecules to influence biological processes. Computationally efficient docking algorithms are frequently applied to identify putative protein–protein binding geometries based on surface complementarity or simple pairwise interaction potentials.[1–4] Molecular docking, however, often largely neglects or only approximately accounts for the flexibility of the binding partners and interactions with the solvent.[3,5–7] It is possible to include a moderate degree of flexibility using for example deformations in soft normal modes at reasonable computational costs.[8–12] In some approaches a refinement stage with side chain flexibility is performed, mainly focusing on interfacial rearrangements.[13,14] In addition, the evaluation of identified binding geometries is largely based on empirical scoring functions applied to single complex conformations neglecting conformational and orientational entropic contributions to binding. Ideally, molecular association should be simulated including full flexibility of both partners and accounting for surrounding water molecules and ions.[15] Molecular dynamics (MD) simulations are in principle well suited for investigating biomolecular association processes including full atomic flexibility. The methodology has already been used to refine potential binding geometries identified in the docking efforts.[16–18] In selected cases it is even possible to skip any initial docking but to use ultra-long atomistic MD simulations and directly mimic the physical binding process.[19,20] This is, however, computationally very demanding and only possible up to timescales on the order of microseconds to milliseconds for individual examples with current computational resources. The search for putative binding regions on the surface of proteins is associated with a rough energy landscape. Hence, the binding partners often get kinetically trapped in local energy minimum for long time intervals resulting in a waste of computational resources. Several efforts have been undertaken to accelerate the search for binding sites. It is possible to employ temperature replica exchange molecular dynamics (TREMD) with multiple parallel MD simulations and periodic exchanges. It can improve the sampling by exploring the surface of the receptor more rapidly at higher temperatures and extracting relevant states at lower temperatures. However, TREMD does not scale well with system size and another method, Hamiltonian REMD (H-REMD) might be more suitable[21] because one can specifically scale force field parameters affecting receptor–ligand interactions. One possibility is to linearly scale the Lennard-Jones (LJ) and electrostatic potential across replicas[22] or reduce the ruggedness of the energy landscape by introducing soft core potentials.[23] The latter method has shown promising results to refine complex geometries close to the native binding mode but do not effectively reduce the problem of trapped binding sites on the receptor surface.[23] Transient binding states in agreement with experiment could be recognized using replica exchange Monte Carlo simulations for three protein–protein complexes using a coarse grained representation of the molecules.[24]

It is also possible to use metadynamics methods to reconstruct the free energy surface of association and dissociation of protein–ligand systems by gradually adding biasing potentials that destabilize already sampled protein surface regions.[25] In the latter study the choice of only two collective variables (CVs) was enough to identify the binding site of four protein–ligand systems. In general, a higher number of CVs is necessary to completely describe the relative ligand–receptor position and orientation.[26] A larger set of CVs can be chosen using reconnaissance metadynamics that incorporates a self-learning algorithm that gradually pushes linear combinations of the CVs.[27] For a protein–ligand system this method was able to identify multiple binding sites of the protein.

In a recent study by Pan et al., reversible association and dissociation of five protein–protein complexes was observed using tempered binding MD simulations.[28] However, the binding and unbinding events were captured in still expensive computer simulations (simulation times of several hundred microseconds) on the special purpose machine Anton.[29] In tempered binding the interaction strength between the two solutes (but not within each protein partner) is used as tempering coordinate (instead of the total temperature as used in standard simulated tempering). As another alternative it is possible to add an explicit repulsive biasing potential between partners in a series of replicas (BP-REMD) that keeps the ligand and receptor at various distance intervals apart in higher replicas.[30] The higher replicas allow to keep some space between partner molecules and therefore result in fast diffusion. Upon exchange with lower replicas favorable binding sites can be rapidly sampled also in the reference replica. This method showed promising performance to specifically accelerate the search process for identifying ligand binding sites on protein surfaces under realistic conditions.[30] However, the method requires to calculate an ambiguity distance between all pairs of surface atoms of both partners which is computationally demanding and not well suited to run in parallel on many cores such as graphical processing units (GPUs).

In the present study, the possibility of increasing the repulsion between ligand and receptor by specifically increasing the pairwise effective van der Waals (vdW) radii and reducing the vdW attraction along the replicas in an H-REMD simulation is explored. It weakens not only the LJ contribution to binding interactions but also reduces the number of hydrogen bonds and electrostatic interactions due to an increased average distance between ligand and receptor. Hence, the biasing potential in the replicas allows the partners to rapidly dissociate from possible suboptimal binding sites to effectively search the protein surface. The method is promising for its simplicity of implementation and only requires adjusting parameters and can therefore be used with existing simulation software that runs on GPUs. The approach was tested on several protein–protein complexes of different sizes and types. In contrast to regular MD simulations it allowed the identification of near-native complexes even when starting far from the native binding region. In addition, we tested the approach for refinement of complexes starting from geometries in the vicinity of the native binding arrangement. In this case also a slightly better performance than regular MD simulations at the same computational effort was achieved.

2 | MATERIALS AND METHODS

For all atomistic simulations the Amber16 or Amber18 software packages,[31,32] were used employing the *pmemd.cuda* module for efficient calculations on GPUs. The ff14SB[33] force field was used together with an implicit water representation using the OBC Generalized Born (GB) model[34] (igb = 8 option in amber) involving an infinite cutoff.

2.1 | Simulations of protein–protein complexes starting far from the binding geometry

MD simulations on protein–protein complexes in order to identify the native binding arrangement were started from an initial placement of one partner (termed ligand) at the opposite side of the second (receptor) protein with respect to the native binding site. Six complexes were considered for these simulations (pdb-id of complexes: 2oo9, 2cfh, 7cei, 2sni, 1gcq and 1syx, see also Supporting

Information Table S1). In all cases the unbound protein structures were used for the simulations. The six complexes were selected due to the relatively small size from the docking benchmark 3.0.[35] We distinguish between receptor and ligand protein according to the assignment in the benchmark 3.0[35] (typically the large protein partner is the receptor and the smaller partner is the ligand). The root mean square deviation ($\text{Rmsd}_{\text{ligand}}$: Rmsd of the ligand after best superposition of the receptor with respect to the native complex structure) of the initial relative placement of the partner proteins from the native complex geometry was between 30 and 61 Å.

MD simulations were performed using the OBC (Onufriev, Bashford, Case) Generalized Born (GB) implicit solvent model[34] (igb = 8 option) and using an infinite cutoff radius for both the GB radii and nonbonded interactions. A Langevin thermostat with a collision frequency of $\gamma = 5\ \text{ps}^{-1}$ was used to control the temperature. The collision frequency is reduced relative to a more physical value of $50\ \text{ps}^{-1}$ to reduce the apparent viscosity of the solvent and speed up sampling.[36] Equilibration of the start geometry was achieved after energy minimization (50 steps steepest descent followed by 1,500 steps conjugate gradient) and heating in three steps (each 12 ps) to 300 K with positional restraints of $0.05\ \text{kcal mol}^{-1}\ \text{Å}^{-2}$ applied on the heavy atoms relative to the starting structure. Since we observed in long MD simulations for some proteins a partial unfolding positional restraints on the receptor C_α atoms (force constant $0.05\ \text{kcal mol}^{-1}\ \text{Å}^{-2}$) were also included during production simulations. Note, that such weak restraints allow still considerable backbone fluctuations and full side chain flexibility but prevent unfolding or large domain motions in the proteins. To prevent the ligand from diffusing too far away from the receptor, restraints between the center of masses (COM) of the C_α atoms of the proteins were employed. The restraining energy was zero for COM distances below a certain threshold and increased quadratically beyond the threshold (force constant $1.0\ \text{kcal mol}^{-1}\ \text{Å}^{-2}$) so that it prevents large receptor–ligand separation but still allows the ligand to dissociate from the receptors up to a certain distance. The COM distance threshold ranged from 27 to 50 Å for the different protein–protein complexes and was larger than the sum of the largest center to surface distances for the two partner proteins (given for each protein pair in Table 1). It was chosen such that the ligand protein can reach every position on the surface of the receptor protein without violating the COM distance threshold. The mean difference of the applied COM distance restraint threshold and the native COM distances was 11 Å. For avoiding unfolding of the ligand protein additional intramolecular pairwise harmonic distance restraints between the C_α atoms of the ligand protein (only distances between 5 and 10 Å) were applied, that prevented the ligand backbone from unfolding (force constant $2.0\ \text{kcal mol}^{-1}\ \text{Å}^{-2}$) but allow full side chain flexibility. Note, that the above harmonic backbone distance restraints do not restrict the orientational and translation freedom of the ligand protein relative to the receptor protein.

In order to perform Hamiltonian replica exchange simulations (H-REMD), 16 replicas for each protein were generated with different Lennard-Jones (LJ) parameters for atom pairs involving atoms from different protein molecules (all intramolecular nonbonded parameters within were preserved). The intermolecular LJ potentials were scaled by a parameter d that adjusts the effective van der Waals radius and a factor e that changes the potential well depths (see next section for a detailed description). The following parameter set for d and e, with a smaller step size between the parameters close to the reference replica that increases in the higher replicas, gave the best results for protein–protein test simulations (see Table 2). For each replica, a short equilibration was performed for 32 ps with no exchange attempts. In the production run every 1,000

TABLE 1 Simulation setups for each complex indicated by the PDB-id for the repulsive scaling H-REMD (RS-REMD) approach and the regular MD simulations

	Simulation time		COM d[a] (Å)
PDB	RS-REMD (ns/replica)	Regular MD (ns/simulation)	
7cei	772	400	40
2oo9	730	684	27
2cfh	845	899	50
1syx	438	380	35
2sni	340	308	35
1gcq	640	640	30

[a]Distance was chosen larger than the sum of the largest center to surface distances for the two partner proteins such that every surface position can be reached without violating the COM restraint.

TABLE 2 Lennard-Jones scaling parameters for the different RS-REMD simulation setups

Replica number	16 replicas		8 replicas	
	d (Å)	*e*	*d* (Å)	*e*
1	0.0	1.0	0.0	1.0
2	0.01	0.99	0.015	0.99
3	0.02	0.98	0.03	0.985
4	0.04	0.97	0.045	0.98
5	0.08	0.96	0.06	0.97
6	0.12	0.94	0.075	0.96
7	0.16	0.92	0.09	0.95
8	0.2	0.9	0.12	0.935
9	0.24	0.88		
10	0.28	0.86		
11	0.32	0.84		
12	0.38	0.82		
13	0.44	0.8		
14	0.5	0.78		
15	0.58	0.76		
16	0.68	0.74		

Note: For the repulsive scaling simulations starting far from the binding geometry and for the refinement of a docking ensemble the 16 replica scheme was used (columns 2 and 3). In the refinement simulations of individual docking poses the eight replica setup was used (columns 4 and 5).

MD steps an exchange between neighboring replicas was attempted, yielding a total simulation time (per replica) ranging from 340 to 845 ns (see Table 1).

Finally, starting from the same equilibration runs, 16 regular MD simulations with no H-REMD but different initial velocities (using the same restraints) were performed for comparable timescales as the H-REMD simulations (see Table 1).

2.2 | Refinement of individual protein–protein docking poses in implicit solvent

In addition to simulations starting far away from the native binding geometry, H-REMD and regular MD simulations were also performed for arrangements in the vicinity of the native complex structure obtained by an initial protein–protein docking run using the program ATTRACT.[37,38] The same set of structures and docking procedure as used in a previous study[39] were employed (see Supporting Information Table S2). Since the H-REMD method for refinement of docked complexes is computationally demanding the number of test complexes was limited to 20 complexes from the docking benchmark 3.0.[35] The docking was performed using a standard docking protocol on the unbound partner structures with the program ATTRACT.[37,38] The 300 top-ranked complexes were considered. Out of this set, the 50 models with lowest Rmsd to the native complex structure were used for further refinement using the RS-REMD or regular MD simulations. In order to refine the docking solutions atomistic replica exchange simulations in implicit solvent were performed (OBC model,[34] using the same conditions as described above) starting from the 50 docking poses of 20 protein–protein complexes. Energy minimization consisted of 2,500 minimization steps (400 steps of steepest descent, 2,100 steps of conjugate gradient). The systems were heated gradually in three steps of 15 ps to 300 K using a Langevin thermostat for temperature scaling. For each equilibrated pose eight replicas were generated with increasing bias for higher replica numbers of the intermolecular LJ parameters. As described above for the simulations starting from the opposite side of the receptor protein a parameter *d* adjusting the effective van der Waals radius and a factor *e* that changes the potential well depths was varied between the eight replicas (see Table 2).

Each replica was simulated for 0.5 ns with an exchange attempt every 125 steps amounting to 4 ns simulation time per pose. Intramolecular pairwise harmonic distance restraints between the C_α atoms of each individual protein were applied (force constant 0.5 kcal mol^{-1} Å^{-2}) together with a COM distance restraint of interfacial C_α atoms between the ligand and receptor (atoms with distances between 10 and 15 Å were considered) with a half parabolic shape (force constant 1.5 kcal mol^{-1} Å^{-2}) that prevents full dissociation in the high replicas and shrinks the possible sampling space for these short simulations. The same simulation conditions and restraints were applied for regular MD simulations of each pose (no replica exchange and bias involved) of the same simulation time (4 ns) following a standard refinement protocol developed previously.[39,40] For evaluating the interaction energy a short MD simulation (30 ps) was applied on the reference replica of the REMD simulations followed by a minimization (500 steps of steepest descent, 2000 steps of conjugate gradient), which was also applied to evaluate the final structures from regular MD simulations. Finally, the minimized structures were scored by subtracting the potential energy of the partners from the energy of the complex.[39] To access the deviation of the refined structures from the native binding site the Rmsd$_{\text{ligand}}$ was calculated, the root mean square deviation of the ligand to the native ligand after superpositioning the receptor on the native receptor (only heavy atoms were considered).

2.3 | Refinement of a protein–protein docking ensemble in implicit solvent

Multiple docking poses were considered in a single REMD run to perform refinement simulations for each of the 20 protein–protein complexes. Only docking poses with a rmsd$_{\text{ligand}}$ above 10 Å (for the complex 7cei 8 Å was chosen, due to a lack of poses with large Rmsd) were considered as starting structures for the subsequent RS-REMD refinement. Each of these poses was first scored based on the potential energy difference of the complex to the individual partners. The 16 highest ranked poses were considered for the subsequent REMD simulations. The poses formed the start structures in the 16 replicas and were distributed based on the ranking (best ranked pose in replica 1, second best in replica 2, etc.). Thus, the (initially) best ranked pose started in the reference replica. Each replica was simulated for 30 ns amounting to a total simulation time of 480 ns per RS-REMD run with an exchange attempt every 250 steps between neighboring replicas. Finally, for comparison, regular MD simulations (no biases and no replica exchange involved) were performed starting from the same poses and simulating the same time as in the RS-REMD case.

2.4 | LJ parameter scaling between partner molecules

The LJ interaction consists of an attractive part proportional to $1/r^6$ and a repulsive contribution typically modeled by a term proportional to $1/r^{12}$. The parameters ϵ_{ij} and R_{ij} in the LJ potential determine the magnitude of attractive interaction and the effective (pairwise) van der Waals radius of the interaction between a pair of atoms of type *i* and *j*.

Typically, only the parameters between atoms of the same type, ϵ_{ii} and R_{ii} are used and one obtains parameters for pairs of different atom types using the Lorentz-Berthelot rules[41]:

$$R_{ij} = \frac{R_{ii} + R_{jj}}{2} \tag{1}$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}}\ . \tag{2}$$

By defining new atom types, it is possible to specifically modify the LJ potential for interactions between the ligand and receptor without affecting the LJ interaction within one partner molecule or with the solvent. We used this possibility by adding an adjustable parameter d to an effective pair-wise van der Waals interaction between pairs of ligand and receptor atoms,

$$R'_{ij} = R_{ij} + d. \tag{3}$$

One might also scale R_{ij} by multiplying with a factor, but this would increase the effective radius of pairs of atoms by different amounts and may strongly distort an interaction interface. A change in the potential well depth ϵ_{ii} is described by a factor e:

$$\epsilon'_{ij} = e \cdot \epsilon_{ij}. \tag{4}$$

As this factor enters multiplicative instead of additive the same relative scaling of attractive interactions of different pairs of atom types is possible. One subtle problem with increasing d, however, is that the number of atoms that can interact increases (illustrated in Figure 1).

This increases the total binding strength, even though any individual interaction might be weaker due to an $e < 1$. ϵ_{ij} has to be decreased further to compensate for that effect. The number of atoms that can exactly fit into the energy minimum around one atom is proportional to the surface area of a sphere with van der Waals radius, which would suggest the following quadratic correction:

$$\epsilon''_{ij} = \left(\frac{R}{R'}\right)^2 \epsilon'_{ij} = \left(\frac{R}{R+d}\right)^2 \epsilon'_{ij}. \tag{5}$$

The LJ potential minimum also gets wider linearly as R_{ij} increases and more atoms can fit into the minimum along the radial direction, leading to a cubic correction:

$$\epsilon''_{ij} = \left(\frac{R}{R+d}\right)^3 \epsilon'_{ij}. \tag{6}$$

In the cubic case, the binding energy stays approximately constant so that the correction with Equation (6) compensates well for the additional possible interactions and a lowering of $e < 1$ indeed weakens the attractive LJ interaction between the partner molecules. Hence, in all cases a cubic correction of ϵ_{ij} was used.

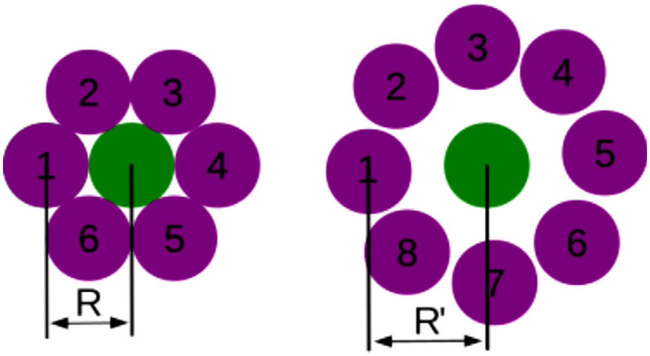

FIGURE 1 Effect of increasing the van der Waals radius on the number of possible interactions [Color figure can be viewed at wileyonlinelibrary.com]

3 | RESULTS AND DISCUSSION

3.1 | Simulations of near-native protein–protein complex formation

For six protein–protein complexes regular MD simulations and the RS-REMD (repulsive scaling replica exchange molecular dynamics) method were compared to identify the native complex geometry after starting from distant initial location of partner proteins. In the starting arrangement the ligand protein was located on the opposite side of the receptor partner with respect to the native binding site (worst case scenario of the initial guess). In each case, 16 MD simulations with different initial velocities were performed using an implicit Generalized Born (GB) solvent model (see Methods for details). The use of an implicit solvent model reduces the computational demand and allows for faster free diffusion of the proteins due to appropriate reduction of the viscosity compared to an explicit solvent model. The simulations started from the unbound ligand and receptor conformations. Only in two of the test cases (2oo9, 1syx) individual regular MD simulations reached locations near to the native binding site and sampled it for longer than a few ns with the 16 × (300 to 900) ns (see Table 1). In the smallest test case (2oo9) the native binding site ($\text{Rmsd}_{\text{ligand}}$ < 10 Å) (root mean square deviation of the ligand to the native ligand after best superposition of native and simulated receptors) was identified in 10 runs after an average simulation time of 186 ns (relative occupancy of binding site 47% in second half of the simulations, see Supporting Information Table S1). For 1syx one simulation reached placements near the binding site ($\text{Rmsd}_{\text{ligand}}$ < 10 Å) after a long simulation time of 236 ns where it stayed for a short time span (approximately 44 ns) until the $\text{Rmsd}_{\text{ligand}}$ grew again beyond 10 Å. In the other simulations including all 16 regular MD simulations of the other 4 protein cases (2cfh, 2sni, 1gcq, 7cei) trapping at locally stable sites but no approach of the native binding site was observed (see Figure 2).

Next, we employed the repulsive scaling (RS)-REMD technique using 16 replicas and started them from the same initial placement as the regular MD simulations. In this technique a repulsive biasing potential acting only between the proteins is added in the replica runs (see Methods for details) that destabilizes protein–protein binding for incorrect trapped states (as well as the native binding state). However, no biasing is employed in the reference replica (under control of the original force field) that was used for further evaluation and analysis. In all but one case sampling of near native arrangements was observed in the reference replica after ~20–400 ns (see Figure 2 and Table 3). For the 2oo9 system, it took 55 ns and thus a bit longer than in case of the regular MD simulations (28 ns). Here, the interacting proteins are very small (contain fewer than 70 residues) with an apparently small number of alternative locally stable binding geometries. However, using RS-REMD the ligand of 2oo9 reached the native site on average faster than in case of regular MD simulations

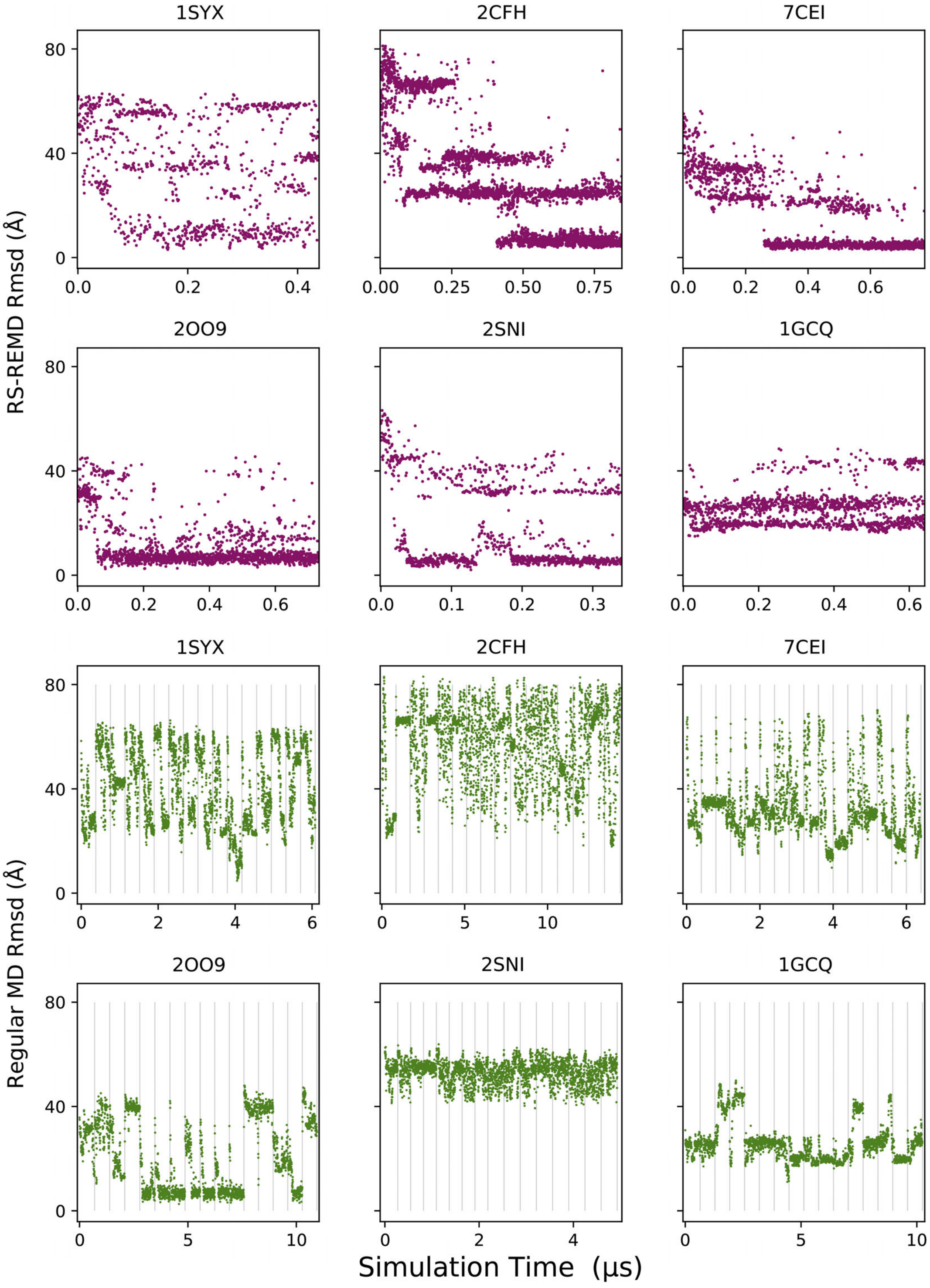

FIGURE 2 $\text{Rmsd}_{\text{ligand}}$ from the native structure for the reference replica of the RS-REMD simulations (magenta dots; first and second row) and for the regular MD simulations (green dots; third and fourth row) of the six protein–protein test cases. The results of the 16 individual simulations (separated by vertical lines) were concatenated in the regular MD cases [Color figure can be viewed at wileyonlinelibrary.com]

(by ~100 ns, see Table 3) and the relative occupancy of the binding site was 78% in the reference replica, higher than the 47% observed when combining all regular MD results (see Supporting Information - Table S1). The RS-REMD simulation on 1syx explored near-native binding arrangements ($\text{Rmsd}_{\text{ligand}}$ < 10 Å) after 68 ns and thus more than 150 ns faster than the free simulations. For the other three proteins the RS-REMD simulations captured the native binding site after 21 ns (2sni), 258 ns (7cei) and 405 ns (2cfh). For these complexes one can observe that the ligand continuously approaches the native binding site through several intermediate states (see Figure 2). It illustrates

TABLE 3 Simulation details for each complex indicated by the PDB-id

	Time to encounter native site		Lowest Rmsd	
PDB	RS-REMD (ns/replica)	Regular MD (ns/simulation)	RS-REMD (Å)	Regular MD (Å)
7cei	258	363[a]	2.7	7.6[a]
2oo9	55	186[b]	2.4	2.4
2cfh	405		3.0	14.9
1syx	68	236	3.1	4.5
2sni	21		2.1	37.4
1gcq			11.6	10.2

[a]The ligand was not stable at the binding site and stayed only for 1.4 ns.
[b]The mean value of all encounter times was taken.

the advantage of RS-REMD compared to regular MD simulations: while regular MD simulations can get easily trapped in intermediate binding states for significant simulation times the RS-REMD allows the system to more rapidly dissociate from such states and reach near-native geometries. The process of approaching the binding site is illustrated for the 7cei case in Figure 3. The initial population of the centers of mass of the ligand protein is located on the opposing side of the receptor in the first third of the simulation (in the reference replica). The sampled distribution eventually shifts toward the binding site on the receptor protein and there increases continuously until the ligand is mostly populated at the binding site or in the vicinity of the binding site in the reference replica. A similar representation for the highest replica shows a quite uniform spherical population of the ligand around the receptor (see Supporting Information Figure S1).

In all the cases for which the binding site was identified RS-REMD led occasionally to a very close agreement to the native structure with a lowest $\text{Rmsd}_{\text{ligand}}$ of ~3 Å (see Table 3). In particular, the lowest $\text{Rmsd}_{\text{ligand}}$ using RS-REMD was closer than for regular MD also in those cases were both methods identified the correct binding site. Thus, in these cases RS-REMD not only performs better than regular MD in the global searching process for the binding site, but also for local rearrangements at the binding site.

In only one case, 1gcq, the near-native binding geometry was not detected after the upper limit of 640 ns of simulation time in the RS-REMD and also not during any of the 16 regular MD simulations (Figure 2). In the 1gcq case the correct position of the ligand at the protein-interface site of the receptor was captured but the orientation of the ligand was incorrect (see Supporting Information Figure S2). It is possible that the force field and implicit solvent representation favor in this case the non-native binding geometry. Stabilization of alternative (non-native) binding geometries (in the current force field setup) is also observed for some of the other test cases. For example, in the 1syx case, complexes with an $\text{Rmsd}_{\text{ligand}} < 5$ Å are occasionally visited in the reference replica but alternative states with larger $\text{Rmsd}_{\text{ligand}} \sim 8$ Å are more frequently sampled. Besides of force field artifacts, such deviation can also be due to the conformational restraining with respect to the unbound (backbone) protein conformation that we include during all simulations. Indeed, the protein association in the case of 1syx involves some backbone changes toward the bound structure at the protein interface (1syx corresponds to a target of medium difficulty, Supporting Information Table S2). In particular, a loop conformation at the interface of the receptor protein differed in the sampled near-native complexes from the structure in the bound form (see Supporting Information Figure S3). Also, states with larger $\text{Rmsd}_{\text{ligand}}$ are still populated in the reference replica in the final stage of the RS-REMD simulation (Figure 2). In the 2cfh case a near-native geometry ($\text{Rmsd}_{\text{ligand}} < 5$ Å) forms the dominant sampled state in the final simulation stage but an alternative binding geometry with $\text{Rmsd}_{\text{ligand}} \sim 25$ Å remains also highly populated. The result indicates that the force field setup stabilizes in many cases not only the native binding geometry but also alternative states in the vicinity of the native structure but also some binding modes quite far from the experimentally observed complex structure. All protein simulations for which the binding site was captured were extended for more than 300 ns after having encountered the native binding site. The relative population in the reference replica of near-native states at the binding site grows in several cases with ongoing simulation time (reaching >50%) (see Supporting Information Table S1). This is not the case for 1syx with a population of the near-native complex of ~ 25% (still forming the largest populated cluster; Supporting Information Figure S4) but some alternative binding modes reaching similar population indicating similar binding affinity. The population of ligand placements at the native binding site is highest in the reference replica and decreases for the higher (larger bias) replicas (see Supporting

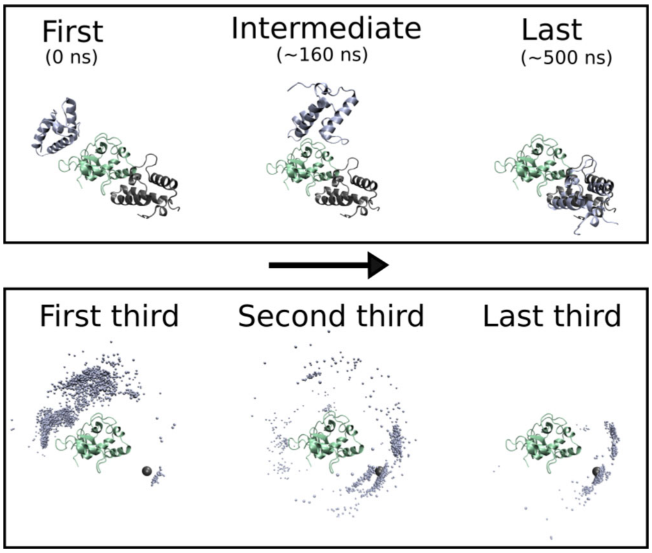

FIGURE 3 (Upper panel) Three snapshots from the reference replica trajectory of the RS-REMD of the 7cei complex example (green cartoon: receptor protein, blue cartoon: ligand protein, black cartoon: native ligand protein placement). (Lower panel) The population of the sampled ligand (center-of-mass) placements during RS-REMD is indicated as blue spheres around the receptor (green cartoon). The ligand protein placement in the native complex is shown by an enlarged black sphere [Color figure can be viewed at wileyonlinelibrary.com]

Information Figure S5 for the example case 7cei) due to the higher repulsive bias. Hence, an advantage of the RS-REMD technique relative to regular MD is that the near-native binding site can be identified by just looking at the population in the different replicas.

3.2 | Refinement of individual protein–protein docking poses in implicit solvent

Significant computational demand and simulation times are still necessary to reach near-native binding geometries using RS-REMD from distant initial placements. However, this corresponds to a worst-case scenario. Instead, it is also possible to first perform a rapid protein–protein docking (not including solvent or partner flexibility) in order to first identify potential binding sites possibly not too far from the native binding geometry. In a second step short RS-REMD simulations are used as a refinement procedure to further improve the docking results. We first performed a docking run on a subset of 20 protein–protein complexes of the protein–protein benchmark 3.0[35] using the docking program ATTRACT and obtained 50 top ranked poses with different ligand deviations from the bound complex ($\text{Rmsd}_{\text{ligand}} < 25$ Å) around the receptor protein (same as used in a recent study on protein–protein docking scoring[39]). For each of these poses a short RS-REMD refinement (4 ns) was performed to refine the docking results. In order to limit the computational demand for this procedure an RS-REMD with eight replicas was performed.

Overall, upon RS-REMD refinement of all 50 decoys for the 20 test cases a slightly larger number of models (65%) with higher $\text{Rmsd}_{\text{ligand}}$ was observed compared to the starting structures (35% of poses had higher $\text{Rmsd}_{\text{ligand}}$ before refinement) (Figure 4). Likely, because of the short simulation time, no $\text{Rmsd}_{\text{ligand}}$ improvements better than 13 Å were sampled. More important than the $\text{Rmsd}_{\text{ligand}}$ improvement of poses with a high deviation from the native binding mode is the refinement performance of the near-native models. For the docking model closest to the native binding site a smaller $\text{Rmsd}_{\text{ligand}}$ after RS-REMD refinement was observed for 13 complexes. Also, the improvement in $\text{Rmsd}_{\text{ligand}}$ was higher, so an overall improvement of 0.49 Å was found considering the mean difference in $\text{Rmsd}_{\text{ligand}}$ before and after RS-REMD refinement of the closest to native pose (see Figure 4, left panel).

The RS-REMD scaling results are compared to a well-established atomistic refinement procedure[39,40] using regular MD simulations with 8 times longer simulation time per decoy and final energy minimization (same force field setup and positional and distance restraints as in the RS-REMD, see Methods). In Figure 4 (right panel), the $\text{Rmsd}_{\text{ligand}}$ after RS-REMD refinement is plotted vs. $\text{Rmsd}_{\text{ligand}}$ after regular MD refinement. A slightly higher amount of structures had a lower Rmsd with RS-REMD refinement (57%, magenta dots) than with regular MD refinement (43%, green dots). RS-REMD refinement also performed better than the regular MD refinement considering the model with the lowest $\text{Rmsd}_{\text{ligand}}$ for each protein–protein complex. For 13 complexes the near-native pose was closer for RS-REMD refinement than regular MD refinement and the mean $\text{Rmsd}_{\text{ligand}}$ of the near-native poses of all structures was slightly lower (0.33 Å) for RS-REMD refinement.

Finally, the refined poses were scored based on the interaction energy of ligand and receptor, the total energy of the complex was subtracted from the total energy of the individual ligand and receptor. The selectivity of the resulting funnel plots ($\text{Rmsd}_{\text{ligand}}$ versus scoring) was compared (see Figure 5 and Supporting Information Figures S6 and S7), measuring the ability of the refinement procedures to distinguish near-native from other decoys. The selectivity was calculated based on an approach introduced recently,[39] by calculating the normalized difference in binding energy of the highest scored pose at the binding site ($\text{Rmsd}_{\text{ligand}} < 10$ Å from the pose of minimal $\text{Rmsd}_{\text{ligand}}$) S'_T and not at the binding site S'_F:

$$\text{Selectivity} = S'_T - S'_F. \tag{7}$$

The two key poses were shifted by the mean scoring value of all poses and divided by the minimum scoring value in order to obtain comparable results for each protein.

$$S'_i = \frac{S_i - \bar{S}}{S_{\text{min}}}. \tag{8}$$

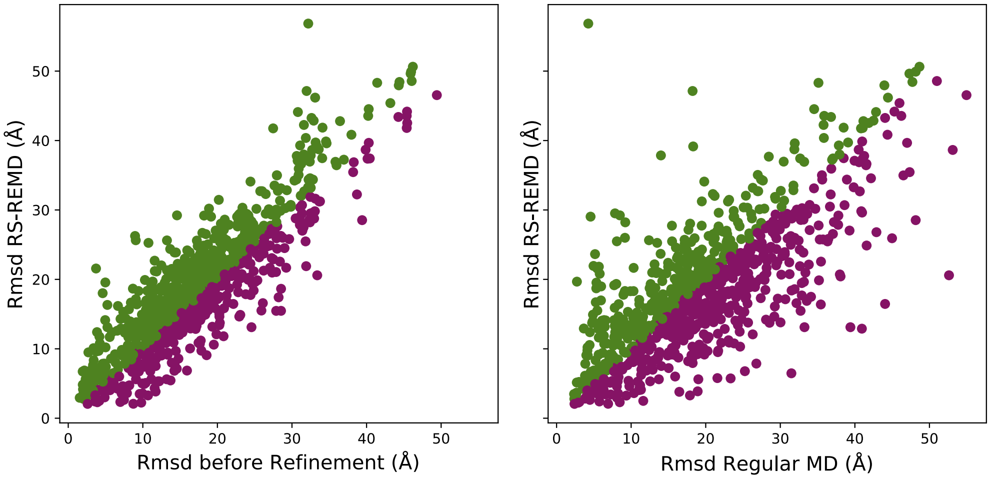

FIGURE 4 The $\text{Rmsd}_{\text{ligand}}$ of all refined complexes after RS-REMD refinement is plotted against the $\text{Rmsd}_{\text{ligand}}$ before refinement (left panel) and after regular MD refinement (right panel). Magenta dots mark poses where the $\text{Rmsd}_{\text{ligand}}$ decreased due to RS-REMD refinement (35% for the left panel and 56% for the right panel) and green dots depict the poses for which the $\text{Rmsd}_{\text{ligand}}$ increased after RS-REMD refinement [Color figure can be viewed at wileyonlinelibrary.com]

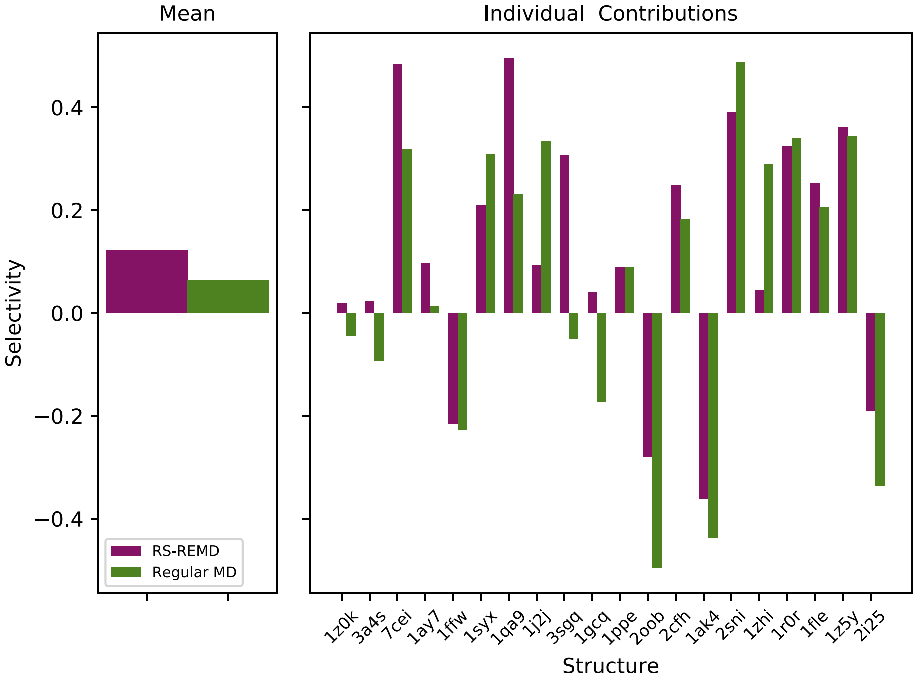

FIGURE 5 Mean selectivity of all structures and the selectivity of each structure for the different refinement procedures. The selectivity was calculated as described in the main text (see Equations 7 and 8) [Color figure can be viewed at wileyonlinelibrary.com]

A selectivity of 1 means a perfectly selective funnel plot for the best bound pose at the binding site and −1 means that the funnel plot is very selective for the highest scored decoy not at the binding site. A value of 0 means that the highest scored near-native pose and the highest scored pose not at the binding site have the same binding affinity.

In 16 cases RS-REMD was able to identify the near-native binding placement (positive selectivity) and only in four cases the refinement approach resulted in a clearly negative selectivity (1ffw, 2oob, 1ak4, 2i25), identifying an incorrect binding site.

The selectivity was slightly higher for 14 structures in RS-REMD (1z0k, 3a4s, 7cei, 1ay7, 1ffw, 1qa9, 3sgq, 1gcq, 2oob, 2cfh, 1ak4, 1fle, 1z5y, 2i25) compared to using regular MD refinement. The mean selectivity of all structures was also higher after the RS-REMD refinement procedure (0.12) in contrast to regular MD refinement (0.06).

In summary, the RS-REMD refinement was able to improve in many cases the placements of the near-native poses. It overall performed slightly better than an established regular MD refinement procedure (at same computational costs) in terms of the selectivity in identifying the native binding site.

3.3 | Refinement of a protein–protein docking ensemble in one RS-REMD

Instead of refining every docked pose separately, it is also possible to start from a different docking decoy in each replica leading to a higher diversity in the starting conditions such that multiple possible binding sites are represented in different replicas. In the case of a sufficient number of replicas, it is then possible to perform only one RS-REMD simulation per complex in contrast to 50 separate simulations for individual refinement of decoys (see above). To increase the challenge, the refinement was initialized exclusively from starting placements that were not located at the binding site. Only poses with an $\text{Rmsd}_{\text{ligand}}$ above 10 Å (for the complex 7cei 8 Å, due to a lack of poses with large Rmsd) were considered as starting structures. An increasing replica number was linked to a lower ranking after docking for the selected starting poses. The results of the RS-REMD (with 16 replicas) are again compared to 16 regular MD simulations of the same length starting from the same initial placements.

The resulting total population of sampled near-native states close to the binding site ($\text{Rmsd}_{\text{ligand}}$ < 10 Å) increased significantly (30%) in

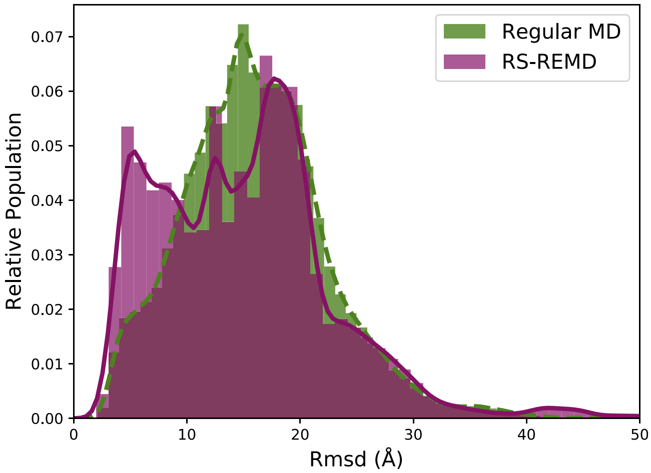

FIGURE 6 Histograms of the $\text{Rmsd}_{\text{ligand}}$ from the native structure for the reference replica of the RS-REMD refinement (magenta) (for all 20 protein–protein test cases) is compared to the $\text{Rmsd}_{\text{ligand}}$ histograms of the regular MD simulations (green). The refinement was performed starting from initial placements not at the binding site [Color figure can be viewed at wileyonlinelibrary.com]

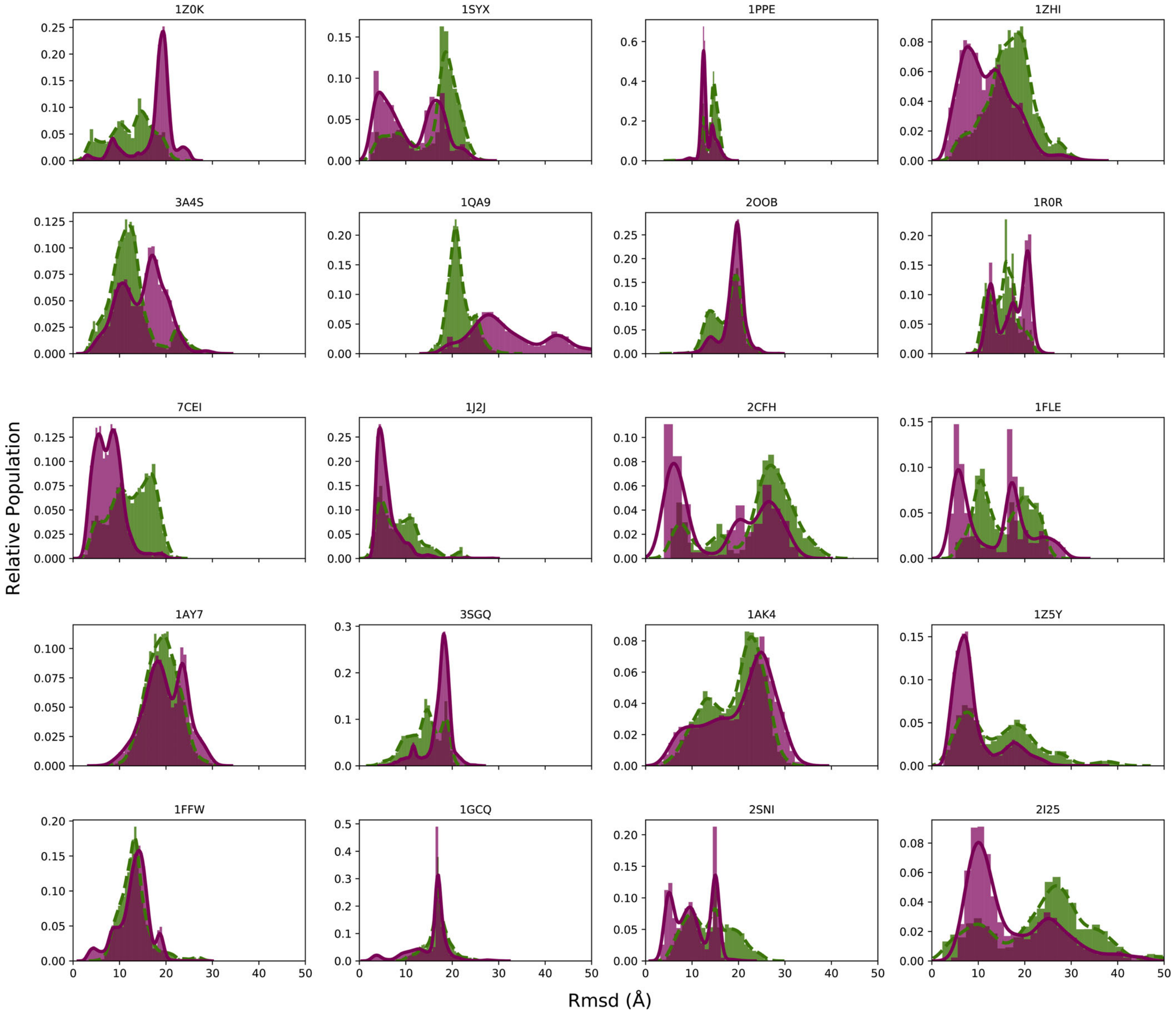

FIGURE 7 Histograms of the $\text{Rmsd}_{\text{ligand}}$ from the native structure for the reference replica of the RS-REMD refinement (magenta) (for all 20 protein–protein test cases in an individual figure) is compared to the $\text{Rmsd}_{\text{ligand}}$ histograms of the regular MD simulations (green). Both refinement procedures were initialized from poses that were not at the binding site [Color figure can be viewed at wileyonlinelibrary.com]

comparison to regular MD refinement (17%) (see Figure 6). For 14 of the 20 structures RS-REMD was able to capture the binding site, in some cases the population was highest at the binding site (see Figure 7). It points out, that even relatively short RS-REMD simulations can capture near-native binding geometries that were not already found in the initial docking search.

Performing only one refinement simulation starting from an ensemble of promising docking solutions and not refining every single pose individually can significantly reduce the computational demand. Hence, the refinement starting from an ensemble of promising docking solutions can be considered as overall more efficient than refinement of every single pose.

Comparing the $\text{Rmsd}_{\text{ligand}}$ histograms of the refinement procedure to the histograms of the pure RS-REMD simulations (see Figure 2 and Supporting Information Figure S4) the dominant states are consistent in most cases, especially for $\text{Rmsd}_{\text{ligand}}$ values under 10 Å (see 2cfh, 7cei, 2sni, 1syx). In case of 1gcq the binding site was not captured in the long repulsive scaling simulation, still the two populated spikes around 20 Å are also present in the refinement simulations.

4 | CONCLUSION AND OUTLOOK

A new H-REMD scheme is presented that includes a repulsive scaling potential (RS-REMD) between different protein molecules based on modification of the intermolecular LJ parameters. The bias requires a modification of the simulation parameter file but no changes in the underlying MD program are involved and full GPU support is possible.

The replica exchange scheme was applied and tested on three tasks that seek to identify the native binding geometry of protein–protein complexes using an implicit solvent model. First, RS-REMD allowed to sample near-native binding placements in five out of six example complexes, starting from a random placement far away from the native binding site. In contrast to multiple regular MD simulations, which were stuck mostly at locally stable sticky sites, these sticky sites were overcome through several intermediate steps in the RS-REMD. While the higher replicas sampled the whole receptor surface, the reference replica sampled locally favorable sites quickly until the native binding site was captured but depending on the case alternative binding modes were also still sampled. Although much less demanding than regular continuous (c)MD simulations still quite extensive sampling is needed for this approach that may limit its applicability. Our RS-REMD approach is designed to identify native binding geometries given the knowledge that the protein partners form a complex. For two proteins that do not bind in reality the approach will likely also suggest putative binding geometries. It might, however, be possible to predict the likelihood of complex formation for two protein partners by studying the population of associated states in the higher replicas (with larger repulsive bias) or from the lifetime and accumulation of complexed states in the reference replica. This will be subject of future studies.

In addition to starting from a worst-case scenario, we also used the approach for refining pre-docked poses. By applying a short RS-REMD run for each of the 50 poses of a benchmark set of 20 protein–protein complexes, it was possible to decrease the mean deviation from the native binding site. Moreover, the mean selectivity of identifying the native binding site according to a simple scoring function (based on the interaction energy) was increased in comparison to a regular MD refinement.

The simulation effort could be further reduced using a refinement scheme that associates each replica in the RS-REMD run with a different docking pose as starting structure. In contrast to the first refinement procedure, only one refining simulation had to be performed for each protein–protein complex. The population of the ligand protein partners near the binding site was clearly increased with RS-REMD beyond the result achieved by regular MD simulations. The benchmark set also contained difficult test cases (also reported in[39]), where the identification of the native binding site was not possible in both refinement procedures. Possible reasons are inaccuracies of the implicit solvent model that may not always favor correct complex structure relative to alternative arrangements. Explicit solvent simulations may help to solve this issue and will be tested in future studies. However, the probably slower diffusion and increase in number of particles will likely demand higher computational efforts. Another limitation of our setup is the inclusion of conformational restraints of the partner molecules with respect to the unbound conformations. This avoids any large-scale conformational change or unfolding of partners but in some cases may prevent conformational adaptations necessary for productive protein–protein complex formation. More global restraining methods like inclusion of backbone Rmsd restraints can help to overcome this issue in future efforts.

In principle, the RS-REMD biasing scheme can also be helpful to study folding/unfolding events or dissociation/association of parts of a protein structure. In such a case only the interactions of the selected part of the protein with other protein segments are scaled in the replica simulations.

ACKNOWLEDGMENTS

We thank the DFG (German Research Foundation) grant SFB (Collaborative Research Center) 863 (project A10) for financial support. Computer resources for this project have been provided by the Gauss Centre for Supercomputing/Leibniz Supercomputing Centre under grant pr27za.

ORCID

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

REFERENCES

[1] J. J. Gray, S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl, D. Baker, *J. Mol. Biol.* **2003**, *331*, 281.

[2] D. W. Ritchie, V. Venkatraman, *Bioinformatics* **2010**, *26*, 2398.

[3] M. Zacharias, *Curr. Opin. Struct. Biol.* **2010**, *20*, 180.

[4] M. M. Gromiha, K. Yugandhar, S. Jemimah, *Curr. Opin. Struct. Biol.* **2017**, *44*, 31.

[5] M. Totrov, R. Abagyan, *Curr. Opin. Struct. Biol.* **2008**, *18*, 178.

[6] C. Wang, P. Bradley, D. Baker, *J. Mol. Biol.* **2007**, *373*, 503.

[7] A. D. J. van Dijk, A. M. J. J. Bonvin, *Bioinformatics* **2006**, *22*, 2340.

[8] A. May, M. Zacharias, *Biochim. Biophys. Acta* **1754**, *2005*, 225.

[9] A. May, M. Zacharias, *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 794.

[10] E. Mashiach, R. Nussinov, H. J. Wolfson, *Nucleic Acids Res.* **2010**, *38*, W457.

[11] V. Venkatraman, D. W. Ritchie, *Proteins: Struct., Funct., Bioinf.* **2012**, *80*, 2262.

[12] I. H. Moal, P. A. Bates, *Int. J. Mol. Sci.* **2010**, *11*, 3623.

[13] J. Fernández-Recio, M. Totrov, R. Abagyan, *Proteins: Struct., Funct., Bioinf.* **2003**, *52*, 113.

[14] S. J. de Vries, A. D. J. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, A. M. J. J. Bonvin, *Proteins: Struct., Funct., Bioinf.* **2007**, *69*, 726.

[15] T. Siebenmorgen, M. Zacharias, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2019**, *1448*, 18.

[16] G. Rastelli, G. Degliesposti, A. Del Rio, M. Sgobba, *Chem. Biol. Drug Des.* **2009**, *73*, 283.

[17] K. Takemura, N. Matubayasi, A. Kitao, *J. Chem. Phys.* **2018**, *148*, 105101.

[18] A. Shinobu, K. Takemura, N. Matubayasi, A. Kitao, *J. Chem. Phys.* **2018**, *149*, 195101.

[19] I. Buch, T. Giorgino, G. De Fabritiis, *Proc. Natl. Acad. Sci.* **2011**, *108*, 10184.

[20] Y. Shan, E. T. Kim, M. P. Eastwood, R. O. Dror, M. A. Seeliger, D. E. Shaw, *J. Am. Chem. Soc.* **2011**, *133*, 9181.

[21] H. Fukunishi, O. Watanabe, S. Takada, *J. Chem. Phys.* **2002**, *116*, 9058.

[22] K. Wang, J. D. Chodera, Y. Yang, M. R. Shirts, *J. Comput. Aided Mol. Des.* **2013**, *27*, 989.

[23] M. P. Luitz, M. Zacharias, *J. Chem. Inf. Model.* **2014**, *54*, 1669.

[24] Y. C. Kim, C. Tang, G. M. Clore, G. Hummer, *Proc. Natl. Acad. Sci.* **2008**, *105*, 12855.

[25] F. L. Gervasio, A. Laio, M. Parrinello, *J. Am. Chem. Soc.* **2005**, *127*, 2600.

[26] S. Boresch, F. Tettinger, M. Leitgeb, M. Karplus, *J. Phys. Chem. B* **2003**, *107*, 9535.

[27] P. Söderhjelm, G. A. Tribello, M. Parrinello, *Proc. Natl. Acad. Sci.* **2012**, *109*, 5170.

[28] A. C. Pan, D. Jacobson, K. Yatsenko, D. Sritharan, T. M. Weinreich, D. E. Shaw, *Proc. Natl. Acad. Sci.* **2019**, *116*, 4244.

[29] D. E. Shaw, J. P. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Denero, R. O. Dror, A. Even, C. H. Fenton, A. Forte, J. Gagliardo, G. Gill, B. Greskamp, C. R. Ho, D. J. Ierardi, L. Iserovich, J. S. Kuskin, R. H. Larson, T. Layman, L. Lee, A. K. Lerer, C. Li, D. Killebrew, K. M. Mackenzie, S. Y. Mok, M. A. Moraes, R. Mueller, L. J. Nociolo, J. L. Peticolas, T. Quan, D. Ramot, J. K. Salmon, D. P. Scarpazza, U. B. Schafer, N. Siddique, C. W. Snyder, J. Spengler, P. T. P. Tang, M. Theobald, H. Toma, B. Towles, B. Vitale, S. C. Wang, C. Young, *SC'14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* **2014**, *9*, 41.

[30] K. Ostermeir, M. Zacharias, *PloS One* **2017**, *12*, e0172072.

[31] D. A. Case, R. M. Betz, D.S. Cerutti, T. E. Cheatham, III, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, C. Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K. M. Merz, G. Monard, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, C. Sagui, C. L. Simmerling, W. M. Botello-Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, L. Xiao and P. A. Kollman (**2016**), AMBER 2016, University of California, San Francisco.

[32] D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham, III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R Harris, N. Homeyer, S. Izadi, A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York and P. A. Kollman (**2018**), *AMBER 2018*, University of California, San Francisco.

[33] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712.

[34] A. Onufriev, D. Bashford, D. A. Case, *J. Phys. Chem. B*, **2000**, *104*, 3712.

[35] H. Hwang, B. Pierce, J. Mintseris, J. Janin, Z. Weng, *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 705.

[36] R. Anandakrishnan, A. Drozdetski, R. Walker, A. Onufriev, *Biophys. J.* **2015**, *108*, 1153.

[37] M. Zacharias, *Protein Sci.* **2003**, *12*, 1271.

[38] S. de Vries, M. Zacharias, *Proteins: Struct., Funct., Bioinf.* **2013**, *81*, 2167.

[39] T. Siebenmorgen, M. Zacharias, *J. Chem. Theory Comput.* **2019**, *15*, 2071.

[40] C. E. Schindler, S. J. de Vries, M. Zacharias, *Proteins: Struct., Funct., Bioinf.* **2015**, *83*, 248.

[41] H. Lorentz, *Ann. Phys.* **1881**, *248*, 127.

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: Siebenmorgen T, Engelhard M, Zacharias M. Prediction of protein–protein complexes using replica exchange with repulsive scaling. *J Comput Chem*. 2020; 41:1436–1447. https://doi.org/10.1002/jcc.26187