Wang *et al. BMC Bioinformatics* (2017) 18:327
DOI 10.1186/s12859-017-1733-6

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Using the multi-objective optimization replica exchange Monte Carlo enhanced sampling method for protein–small molecule docking

Hongrui Wang[1*], Hongwei Liu[1], Leixin Cai[1], Caixia Wang[1] and Qiang Lv[1,2]

Abstract

Background: In this study, we extended the replica exchange Monte Carlo (REMC) sampling method to protein–small molecule docking conformational prediction using RosettaLigand. In contrast to the traditional Monte Carlo (MC) and REMC sampling methods, these methods use multi-objective optimization Pareto front information to facilitate the selection of replicas for exchange.

Results: The Pareto front information generated to select lower energy conformations as representative conformation structure replicas can facilitate the convergence of the available conformational space, including available near-native structures. Furthermore, our approach directly provides min-min scenario Pareto optimal solutions, as well as a hybrid of the min-min and max-min scenario Pareto optimal solutions with lower energy conformations for use as structure templates in the REMC sampling method. These methods were validated based on a thorough analysis of a benchmark data set containing 16 benchmark test cases. An in-depth comparison between MC, REMC, multi-objective optimization-REMC (MO-REMC), and hybrid MO-REMC (HMO-REMC) sampling methods was performed to illustrate the differences between the four conformational search strategies.

Conclusions: Our findings demonstrate that the MO-REMC and HMO-REMC conformational sampling methods are powerful approaches for obtaining protein–small molecule docking conformational predictions based on the binding energy of complexes in RosettaLigand.

Keywords: Monte Carlo, Enhanced sampling method, Multi-objective optimization, Protein–small molecule docking, Complex structure prediction

Background

Simulating the interactions between a macromolecule and small molecule (ligand) is important for understanding the molecular basis of the mechanisms found in healthy and diseased cells [1]. The complex conformational search problem has been investigated in recent decades in order to predict the conformations of protein–small ligand docking [2]. Given the importance of conformational search, several software systems have been developed over the past 20 years, including Dock [3], FlexX [4, 5], GOLD [6, 7], Autodock [8–10], Glide [11] and others [12–14]. These software systems and sampling methods can efficiently predict realistic complex protein–ligand docking structures according to predefined sets of criteria [15]. In general, a protein–ligand docking conformational search method uses either Monte Carlo (MC) [16] search strategies or genetic algorithms [17]. However, in order to improve the sampling procedure, various advanced sampling approaches have been developed in recent years [18–20].

The MC method comprises a class of numerical methods based on random sampling and estimating the

*Correspondence: riihon@yeah.net
[1]School of Computer Science and Technology, Soochow University, 1 Shizi Street, 215006 Suzhou, People's Republic of China
Full list of author information is available at the end of the article

desired outputs using this sample. Integration by MC simulation evaluates $E[f(x)]$ by drawing samples $\{X_t, t = 1, \ldots, n\}$ from the state space Ω and then approximating

$$E\left[f(x)\right] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t). \quad (1)$$

Thus, the function mean of $f(X)$ is estimated based on a sample mean. When the samples $\{X_t\}$ are independent, the law of large numbers ensures that the approximation can be as accurate as required by increasing the sample size n.

The replica exchange MC (REMC) method [21] implemented using independent Markov chains $X_n^i (n \geq 0)$ is defined on the same state space Ω and it can be used to test several replicas in parallel in order to explore the same stationary normalized distributions $\rho_i(x)(x \in \Omega, 1 \leq i \leq N)$ (due to the central limit theorem) at different "temperatures" [22, 23]. Replicas at sufficiently high temperatures are sampled broadly so the barriers will be crossed, whereas low temperature replicas can used to deeply explore the local energy minima. In the REMC method, frequent exchanges are attempted between states X_n^i and X_n^j of two "neighboring" Markov chains with indices i and j, which belong to different thermodynamic states, and the configurations can be identified that cross the local energy barriers more easily.

Many versions of the REMC sampling method have been used in studies related to simulation [24–26]. These search methods provide significant improvements in terms of computational efficiency compared with the traditional MC search methods. Hamiltonian [27–29] and well-tempered ensemble [30, 31] methods are used widely as MC search methods. Hamiltonian MC is a Markov chains MC method that uses the physical system dynamics rather than a probability distribution to estimate future states in the Markov chain. This allows the Markov chain to explore the target distribution much more efficiently, thereby resulting in faster convergence in Ω. The well-tempered ensemble can be designed to have approximately the same average energy as the canonical ensemble but much larger fluctuations. An even greater advantage is obtained when a well-tempered ensemble is combined with parallel tempering [32]. Using a well-tempered ensemble, it is possible to observe transitions between states, which would be impossible to study using the standard MC method [33]. In this study, we present novel multi-objective optimization (MO)-REMC sampling methods.

A multi-objective optimization problem (MOP) comprises several conflicting objectives that need to be optimized. In general, a MOP is defined mathematically as presented in [34].

Definition 1 *(General MOP): A MOP minimizes* $F(\vec{x}) = (f_1(\vec{x}), \ldots, f_k(\vec{x}))$ *subject to* $g_i(\vec{x}) \leq 0, i = 1, \ldots, m, \vec{x} \in \Omega$. *A MOP solution minimizes the component functions of a vector function* $F(\vec{x})$, *where* $\vec{x}$ *is an n-dimensional decision variable vector* $(\vec{x} = x_1, \ldots, x_n)$ *from some space* Ω, *the vector* $\vec{x}$ *minimizes every component of* $F(\vec{x})$, *or at least one, and the component functions of the vector function* $F(\vec{x})$ *should be computable for every* $\vec{x}$.

The objectives of *DEFINITION* **1** contradict each other because no point in Ω maximizes all of the objectives simultaneously. Thus, in order to balance them, the best tradeoffs among the objectives can be defined in terms of Pareto optimality. Using the MOP presented in *DEFINITION* **1**, the key Pareto concepts of Pareto dominance, Pareto optimality, Pareto optimal set, and the Pareto front (non-dominated solutions set) are defined mathematically as presented in [34, 35]. The multi-objective optimization approach finds the Pareto optimal set of the population, which comprises a set of solutions that are non-dominated with respect to each other. In the objective space, the set of non-dominated solutions lie on a surface known as the Pareto front. Non-dominated solution sets are those in which no other solutions are superior in terms of all attributes (objectives). Pareto optimality is effective for facilitating the convergence of the population in a low-dimensional search space [36]. By comparing every solution in the Pareto optimal set, it is always possible to improve one attribute to achieve a better gain without another becoming worse. However, each objective can be minimized or maximized when considering optimization problems with two objectives. The Pareto front approach offers a method based on attributes for finding the subset of promising solutions. This method also considers the solution attributes directly without converting them into a standard form initially. Figure 1 illustrates the case of a Pareto front with two objectives (colored points), where there is a tradeoff between minimizing and maximizing the Pareto optimal points of both the x and y coordinate values in min-max, max-max, min-min, and max-min scenarios. The scatter plots indicate the Pareto optimal set with discrete points for four different scenarios and two objectives. In each case, the Pareto optimal set always comprises solutions from a particular edge of the feasible search space for discrete points [37].

In recent studies, protein–small ligand docking prediction has focused on improving the convergence speed using sampling methods. A form of solution is used as an important component of evolutionary multi-objective optimization algorithms. It has been shown that using an elitist solution improved the convergence speed for various sampling algorithms. Therefore, in this study, we

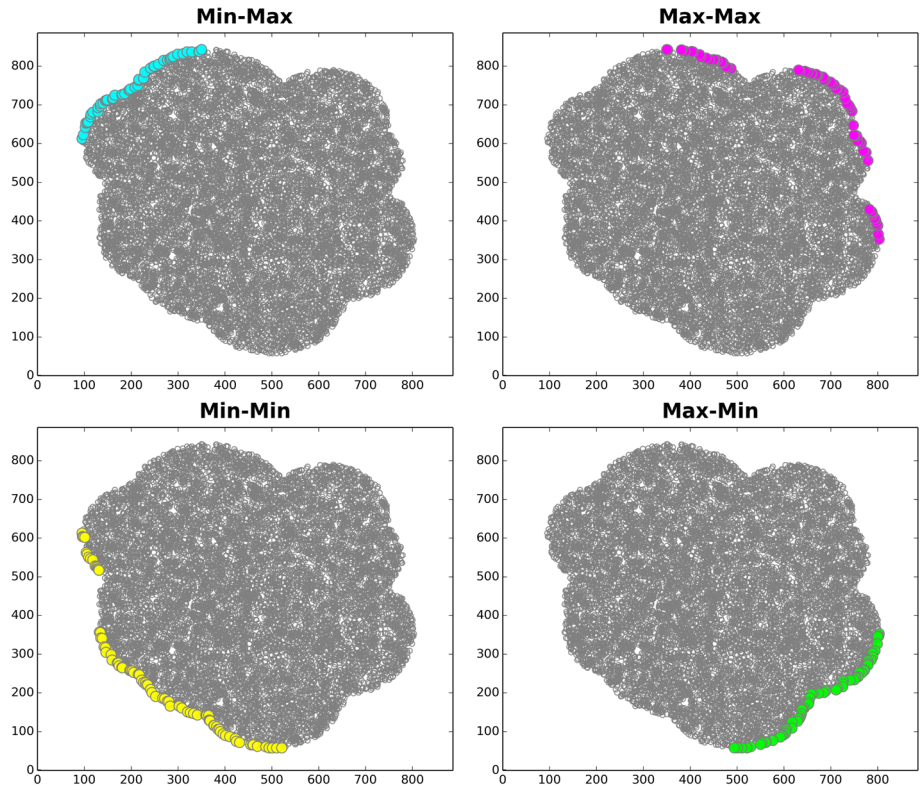

Fig. 1 Pareto optimal solutions used to search four combinations of two objective types with discrete points

developed MO-REMC methods by using multiple non-dominated solutions as replicas for exchange during optimization at different temperatures, thereby improving the REMC sampling algorithm convergence speed associated with replica selection. We also developed methods for choosing replicas to enhance search and to improve exploration of the state space by using the Pareto front energy information. We demonstrated that the MO-REMC methods could enhance the performance of sampling methods based on a suite of benchmark test sets using the RosettaLigand protocol [38, 39]. We also performed an extensive comparative study of the proposed methods with traditional MC (detailed implementation is presented in the "Sampling methods" section in reference Algorithm 1) and REMC (see Algorithms 3 and 2) sampling algorithms based on 16 benchmark test cases. As part of this investigation, the RosettaLigand energy function total score (TScore), binding energy interface delta (IFDelta), and ligand of RMSD(Lrmsd) obtained with the proposed MO-REMC algorithms were compared with those produced by MC and REMC sampling methods, which showed that the proposed methods generally performed better than MC and REMC. The MO-REMC (see Algorithms 3, 4 and 5) and hybrid MO-REMC(HMO-REMC, see Algorithms 3, 4 and 6) methods were found to enhance the convergence to solutions compared with the MC and REMC sampling methods.

Methods

Test data set

The RosettaLigand protocol yielded better results with the classic MC sampling method when using a data set of 100 native protein-ligand complexes. In 71/100 cases, the lowest energy model had an Lrmsd less than 2Å [39]. We suggest that the RosettaLigand protocol cannot obtain satisfactory results in the remaining cases mainly because the MC sampling technique employed in docking is not sufficiently efficient for sampling or optimization in challenging cases. In the present study, we considered cases where satisfactory result could not be obtained with the MC approach. In all of these cases, the native complex was not recognized as a particularly low energy pose even after minimization. The 16 complexes used in this study are summarized in the "Summary of the docking results obtained using different sampling methods and scales" section.

Preparation of the protein and ligand

A validated receptor is crucial for the successful prediction of targets. In this study, we performed repacking of

the side-chain of the receptor near the initial ligand position in a similar manner to the RosettaLigand protocol [38]. Placing a ligand near clashing residues allowed the side-chains to be repacked stochastically. We generated 10 structures per receptor and the receptor structure was directly derived based on the RosettaLigand TScore to select the protein conformation with top minor TScore value. This selection process used the RosettaLigand protocol to generate 10 structures per receptor and we only selected that with the lowest energy. This procedure can resolve any pre-existing clashes between the protein side-chains and ligand, thereby gaining a large energy increase [39].

Alternatively, we treated ligand conformations as "rotamers," which were sampled at the same time as the protein side-chains were repacked. Ligands were represented as a set of discrete conformations. To generate these conformations, all the torsional degrees of freedom in the ligand were identified and each of the torsion angles with probable conformations was compiled based on the atom type and hybridization state of the linked atoms. Next, each torsion angle was placed in one of the states considered, but conformations with internal clashes in ligand atoms were not considered, especially the conformations where the closed ring systems were not altered. Finally, we evaluated the internal ligand energy and energy minimization was applied [40]. At present, ligand conformers are generated externally in the RosettaLigand protocols. Thus, we used the Omega program (v2.3.2, OpenEye) [41] with its default settings and restrained the ligand torsions with a harmonic potential during minimization.

Scoring function for docking

In the coarse-grained sampling stage, the coarse-grained complementary score S_{cg} is defined as

$$S_{cg} = R - min(A/N, 0.85), \tag{2}$$

where R denotes ligand atoms within 2.25Å of the receptor backbone or C^{β}s (repulsive clashes), A denotes ligand atoms between 2.25Å and 4.75Å of any protein atom (attractive contacts), and N denotes the total ligand atoms. The best-scoring poses were filtered by stochastic elimination of near duplicates with a threshold of $0.65\sqrt{N}$Å, where N is the number of non-hydrogen ligand atoms [39].

In the high-resolution refinement stage, the full-atom score is a linear combination of the different scoring items. These scoring items include the attractive Lennard-Jones score, repulsive Lennard-Jones score, implicit Lazaridis-Jarplus solvation score, reference energy for each amino acid, proline ring closure energy score, backbone-backbone H-bonds distant and close scores in the primary sequence, hydrogen bond energy score, probability of an amino acid at *phi* and *psi* angles, residue–residue pair probability score, and *omega* dihedral in the backbone. The high-resolution refinement scoring function S_{fa} is defined as

$$S_{fa} = \sum_{t=1}^{n} w_i s_i, \tag{3}$$

where s_i denotes different scoring items and w_i denotes alternative weights. The full details are described in Table 1, reference [42]. In this research, we are simply using coarse-grained sampling stage and high-resolution refinement stage scoring functions for docking, including TScore and IFDelta functions, as implemented in RosettaLigand [39].

Sampling methods

Our docking methods are based on the Rosetta Ligand(v3.4) protocol, where we use the repacking side-chain method in ROSETTA suites to generate the receptor and represent ligands as a set of discrete conformations generated by the Omega program. Finally, we examined the capability of the RosettaLigand docking protocol based on MC, REMC, MO-REMC, and HMO-REMC sampling methods.

MC sampling method

The MC method approximates an expectation based on the sample mean of a function of simulated random variables. The term MC generally applies to all simulations

Table 1 Scoring function weights used in the four sampling methods

Score items	Weight (Hard)	Weight (Soft)
Proling ring closure energy	1.00	1.00
Lennard-Jones attractive	0.80	0.80
Lennard-Jones repulsive	0.40	0.60
Lazaridis-Jarplus solvation energy	0.60	0.50
Pair energy	0.80	0.50
Reference energy for each amino acid	1.00	1.00
In primary sequence		
Backbone-backbone hbonds distant	2.00	1.20
Backbone-backbone hbonds close	2.00	1.20
Hydrogen bond energy		
Sidechain-backbone	2.00	1.20
Sidechain-sidechain	2.00	1.20
Probability of amino acid at *phi* and *psi*	0.50	0.32
Omega dihedral in the backbone	0.50	0.50

(Hard) indicates weights used during side-chain repacking
(Soft) indicates weights used during rigid-body minimization

that utilize random sampling to obtain numerical solutions for a system of interest. In the general RosettaLigand protocol, MC refers to Metropolis-Hastings sampling, which samples from the Boltzmann distribution, and it was developed by Metropolis et al. in the Los Alamos team [43]. In the present study, MC simulations were performed as follows. Starting from an initial conformation of the protein–ligand interaction, a perturbation of *rotamerTrialMover*() or *packRotamersMover*() was attempted that changed the conformation of the complex. This trail *Mover*() from state last accepted (old) to state perturbed (new) is accepted based on an acceptance probability such that [39]

$$prob\,[old \rightarrow new] := e^{min(40.0, max(-40.0,\ boltz\_factor))}, \tag{4}$$

where the $boltz\_factor = (last\_accepted\_score - score)/k_BT$, $last\_accepted\_score$ denotes the energy value of the last accepted structure of the complex, *score* denotes the energy value of the perturbed structure of the complex, T denotes the current temperature, and k_B denotes the Boltzmann constant, which is considered to be one. In order to decide whether to accept or reject the trail *Mover*(), we generate a random number, denoted by *mc_RG_uniform*, from a uniform distribution in the interval[0, 1].

Clearly, the probability that *mc_RG_uniform*[0, 1] is less than *prob*[*old* → *new*] is equal to *prob*[*old* → *new*]. We now accept the trail *Mover*() if *mc_RG_uniform*[0, 1] < *prob*[*old* → *new*] or *prob*[*old* → *new*] ≥ 1 and reject it otherwise. The transition probability for the MC sampling method from conformation p to a perturbed conformation p' depends on the difference in $last\_accepted\_score - score$ between the last accepted (old) conformation and the perturbed (new) conformation, which is determined such that

$$P[p \rightarrow p']:=\begin{cases}0,\ if\ prob[old \rightarrow new] \leq mc\_RG\_uniform[0,1],\\ 1,\ if\ prob[old \rightarrow new] > mc\_RG\_uniform[0,1],\\ 1,\ if\ prob[old \rightarrow new] \geq 1.\end{cases} \tag{5}$$

where *prob*[*old* → *new*] is the acceptance probability between conformations p' and p. This rule guarantees that the probability to accept a trail *Mover*() from the last accepted conformation to perturbed conformation is indeed equal to *prob*[*old* → *new*] [44]. If the current conformation structure is rejected, MC can retain an additional duplicate of the previous sampling structure as the sample accepted by the system. Figure 2 (left and upper panel) shows that the last sampling structure (red point) is accepted by the MC method as the exclusive solution. After many iterations, an accurate average energy value can be obtained for a complex structure. Algorithm 1 shows the pseudo-code for the RosettaLigand MC Boltzmann sampling method implementation.

Algorithm 1: MCBOLTZMANN(p, T)

Input: p – current structure of the complex, T – temperature of the current system, $E()$ – donated energy function

Output: *mc_accepted* – true or false, donated acceptance or rejection of the current structure

```
1  mc_accepted ← 0;
2  score ← E(p);
3  boltz_factor ← (last_accepted_score − score)/T;
4  prob ← e^min(40.0,max(−40.0,boltz_factor));
5  if prob < 1 then
6  |  mc_RG_uniform ← U(0, 1);
7  |  if mc_RG_uniform ≥ prob then
   |  mc_accepted ← 0;
8  |  else mc_accepted ← 1;
9  else
10 |  mc_accepted ← 1;
11 end
12 if mc_accepted then
13 |  last_accepted_score ← score;
14 end
15 return mc_accepted
```

In RosettaLigand, the efficiency of the MC Boltzmann sampling method can be improved by avoiding the computation of the exponential function (line 4, Algorithm 1). A more detailed interpretation is given in reference [44].

REMC sampling method

In current protocols, replica exchange is the most widely used method for enhancing sampling in bio-molecular simulations, where it can be viewed as a parallel version of simulation tempering, and it is also known as parallel tempering or multiple Markov chains. In the proposed method, REMC search maintains M identical copies of replicas as M sampled canonical ensembles at different temperatures. Each temperature value is unique and each of the M replicas has an associated temperature value (T_1, T_2, ..., T_M). Each of the M replicas independently performs a simple *MCBoltzmann*(p, T) search at the respective temperature setting. In addition, in our REMC algorithm, each replica p'_i is perturbed and the associated energy value $E(p'_i)$ is archived in ensembles P' and E'. The elite replicas in the archives are selected using a procedure called *select_REMC_Replicas*(E', P'). In this procedure, we select the last "*numR*" conformations that have been pushed into the queue in the archives as replicas

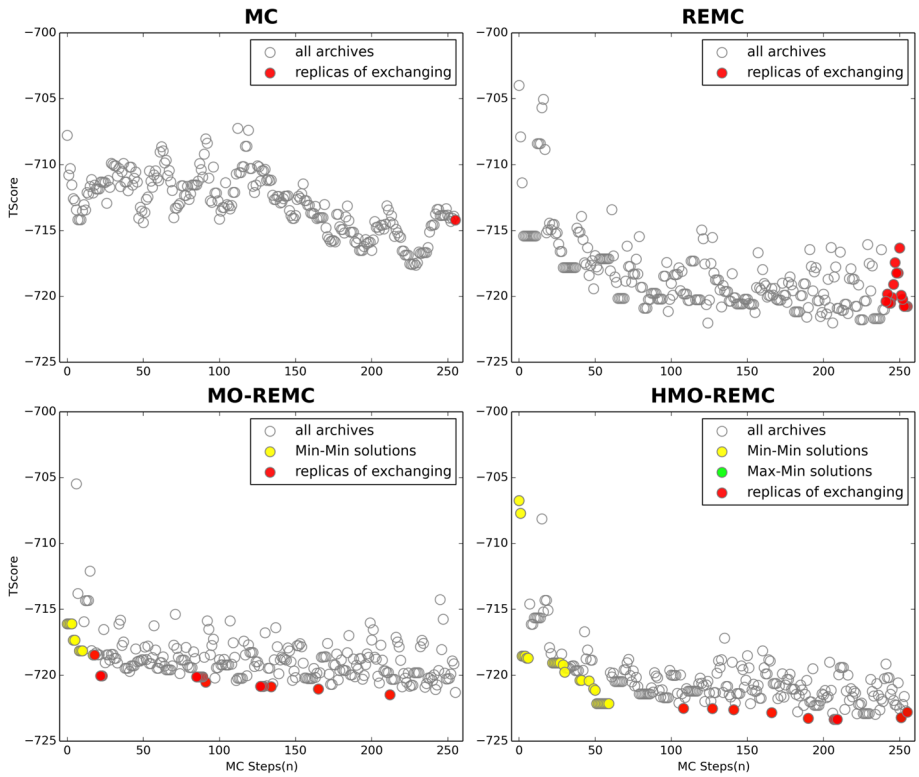

Fig. 2 Target 2PRG replicas selected by the MC, REMC, MO-REMC, and HMO-REMC sampling methods in one iteration

for exchange, as shown in Fig. 2 (right and upper panels), where the last "*numR*" sampling structures are used as replicas(red points) for exchange in the REMC method. Algorithm 2 presents the pseudo-code for the selection of replicas from the archives in the implementation of the REMC sampling method.

Algorithm 2: SELECT_REMC_REPLICAS(E', P')

Input: E' – energy score in the archives, P' – conformation ensemble in the archives

Output: pe' – protein conformation ensemble of the selected elite

```
i ← 0;
while i < numR do
    pe' ← P'_{|E'|−i};
    i ← i + 1;
end
return pe'
```

We can represent the current state of the "*numR*" replicas selected from the archives as a protein conformation ensemble pe': = (pe'_1, ..., pe'_{numR}), as follows, where pe'_j is the conformation of replica j, which (as stated previously) runs at temperature T_j. During replica exchange, the temperature values of neighboring replicas are exchanged at a probability proportional to their energy value and difference in temperature. The transition probability from some current conformation pe'_i to a perturbed (trail *Mover*()) conformation pe''_i is determined using the so-called Metropolis criterion, as shown in the MC sampling method section.

Exchanges are performed between neighboring temperatures, T_i and T_j. The probability of an exchange depends on the energy values, $E(pe'_i)$ and $E(pe'_j)$, and the inverse temperatures, β_i and β_j. An exchange of temperatures, and thus the relabeling of replicas, affects the state of the replica ensemble pe'. Therefore, we define an exchange between two replicas i and j more generally as a transition from the current ensemble state pe' to an exchanged state pe''. We define $l(pe'_i) = i$, the current label or replica number, for all pe'_i. The probability of a transition from the current ensemble state pe' to an altered state pe'' by exchanging replicas i and j is defined as [45]:

$$P\left[pe' \rightarrow pe''\right] := P\left[l(pe'_i) \leftrightarrow l(pe'_j)\right] := \begin{cases} 1, & \Delta \leq 0, \\ e^{-\Delta}, & \text{otherwise.} \end{cases} \quad (6)$$

The value Δ is the product of the energy difference and inverse temperature difference:

$$\Delta := \left(\beta_j - \beta_i\right)\left(E(pe_i) - E(pe_j)\right), \tag{7}$$

where $\beta_i = 1/T_i$ is the inverse of the temperature of replica i. Potential replica exchanges are only performed between neighboring temperatures because the acceptance probability of the exchange decreases exponentially as the temperature difference between replicas increases.

The pseudo-code for Algorithm 3 illustrates the details of our REMC search procedure performed for "*numR*" replicas and a predetermined temperature range between *minT* and *maxT*. In the "*while* $i + 1 < numR$ do" loop, which runs over the pairs of replicas to be swapped, it can be seen that the swaps being attempted include pairs (0,1), (2,3), (4,5), etc., but never pairs (1,2), (3,4), (5,6), etc. This scheme will not satisfy the "detailed balance condition"(transition probabilities $i \to j \neq j \to i$). Moreover, in the condition structure for Δ, it is obvious that the swap is rejected if Δ is larger than some threshold number (often 75, but also depends on the computer architecture), then the swap is rejected because $e^{-\Delta}$ can never be larger than any random number *mc_RG_uniform*$[0, 1]$, and hence one call of the random number generator is saved, making the algorithm computationally more efficient.

Algorithm 3: REMC(*numR, numC, repackNth, minT, maxT*)

Input: p_0 – ensemble of initial conformations, *numR* – number of conformation replicas, *numC* – number of cycle steps, *repackNth* – repack receptor side-chain of interface padding every N cycle steps, *minT* – minimum temperature, *maxT* – maximum temperature

Output: p' – ensemble of modified state perturbed conformations

```
E' ← 0; P' ← 0;
TStep ← (maxT − minT)/numR;
foreach temperature i in numR do
    T_i ← minT + TStep;
end
foreach cycle k in numC do
    foreach replica i in numR do
        p_i ← p_0;
        if i%repackNth = 1 then
            p'_i ← packRotamersMover(p_i);
        else
            p'_i ← rotamerTrialsMover(p_i);
        end
        MCBoltzmann(p'_i, T_i);
        E' ← E(p'_i);
        P' ← p'_i;
    end
    pe' ← select_REMC_Replicas(E', P');
    i ← 0; j ← 0;
    while i + 1 < numR do
        j ← i + 1;
        Δ ← (β_j − β_i)(E(pe'_i) − E(pe'_j));
        if Δ ≤ 0 then
            swapLabels(pe'_i, pe'_j)
        else
            remc_RG_uniform ← U(0, 1);
            if remc_RG_uniform ≤ e^{−Δ} then
                swapLabels(pe'_i, pe'_j);
            end
        end
        i ← i + 2;
    end
    p_0 ← 0; p_0 ← pe';
end
```

MO-REMC sampling method

The REMC method involves a group of MC moves that generate a Markov chain of states. This Markov process has no dependence on history in the sense that new configurations are generated with a probability that depends only on the current configuration and not on any previous configurations. In this study, we developed the MO-REMC sampling method where the random configuration process is not Markovian so the "detailed balance criterion" is not satisfied. In contrast to the traditional REMC algorithm, which typically samples a canonical ensemble of states, we introduce a dependence on history into the REMC method and use historic multi-objective optimal Pareto front information to facilitate the selection of critical replicas of current states, which comprise a set of replicas that are similar to lower energy states but also as diverse possible. Using the generated Pareto front as representative conformation structure templates can improve the convergence of the available conformational space including possible near-native structures.

The aim of the MO-REMC sampling method is to enhance the speed of convergence for the available conformational space. The MO-REMC method employs a history-dependent Pareto frontier list to explicitly maintain a limited number of non-dominated conformations found by the REMC sampling method. Each individual in the archives generated by the REMC sampling method is evaluated using binary objectives: the sampling search steps (MC steps) and the TScore values of the perturbed conformations. The objective MC steps denote the time series for the search process and the TScore values for the perturbed conformations in RosettaLigand denote a history-dependent information map of the available conformational space. The MO-REMC sampling method is

inspired by evolutionary, population-based algorithms. In the traditional REMC method, replicas at sufficiently high temperatures are sampled broadly so the barriers will be crossed, whereas low-temperature replicas can used to deeply explore the local energy minima principle. Included in multi-objective optimal method critical replicas of current states are similar greedy states, dominated non-Pareto frontier list replicas, and diverse possible characteristics. This method is effectively a combination of the REMC sampling method and historic multi-objective optimal Pareto front critical conformation structures. The experimental results show that the elite replicas generated by the historic multi-objective optimal Pareto front can enhance the speed of convergence of the available conformational space.

Algorithm 4 presents the pseudo-code for calculating the binary objectives based on the Pareto front of archives in the implementation of the MO-REMC sampling method. Each objective can be minimized or maximized according to the values of Boolean variables *maxX* and *maxY*. In this procedure, in the first step (lines 1–6), all of the solutions $x_0, \ldots, x_{n-1}$ in the archives are the alternatives sorted in order of increasing/decreasing objective X, which can be minimized or maximized. Let pf':=$\{x_0, y_0\}$ and i:=1, where $\{x_0, y_0\}$ denotes the combination containing the first non-dominated front. In the second step (lines 8–17), for each combination in the archives $\{x_i, y_i\} \in \{X, Y\}$, let pf':=$pf' \cup \{x_i, y_i\}$, If $\{x_i, y_i\}$ is not dominated by any combination according to objective Y that has been be minimized or maximized already in pf', then add $\{x_i, y_i\}$ to pf'. In the third step (lines 7–18), repeat from the step second until no more combinations can be added to pf'. In the last step, iteration stops when i=N, where N denotes the number of combinations in the archives.

In addition, in the middle of each iteration of the MO-REMC sampling method, a set of conformations is provided instead of the last set of conformations using the $select\_MO - REMC\_Replicas(E', P')$ procedure, whereas the REMC sampling method uses $select\_REMC\_Replicas(E', P')$. The $select\_MO-REMC\_Replicas$ function is obviously designed to select the conformations from the archived and the last "*numR*" min-min scenario Pareto optimal solutions set that are non-dominated relative to the other conformations, as shown in Fig. 2 (left and lower panel), where in the last circle, the last "*numR*" sampling structures are used as replicas(red points) for exchange in the MO-REMC method, and the min-min scenario Pareto optimal solutions set is denoted by yellow points (partial points are covered by red points in Fig. 2). These min-min scenario Pareto optimal solutions from the archives provide a natural and rapid convergence source, which is used to obtain alternative comparison sets from the archives. The pseudo-code in Algorithm 5 describes the procedure for determining whether to accept or reject the Pareto front, as well as for deciding whether to select replicas for exchange or not.

Algorithm 4: PARETOFRONTIER(X,Y,$maxX$,$maxY$)

Input: X – objective X, Y – objective Y, $maxX$ – Boolean value of the maximized objective X, $maxY$ – Boolean value of the maximized objective Y

Output: pf' – conformation ensemble of Pareto optimal solutions

```
1  if maxX = 1 then
2  |  inverse_sorted({X, Y});
3  else
4  |  sorted({X, Y});
5  end
6  pf' ← {x_0, y_0}; i ← 1;
7  foreach {x_i, y_i} in {X, Y} do
8  |  {pair_x−previous, pair_y−previous} ← pf'_|pf'|−1;
9  |  if maxY = 1 then
10 |  |  if pair_y−previous ≥ y_i then
11 |  |  |  pf' ← {x_i, y_i};
12 |  |  end
13 |  else
14 |  |  if pair_y−previous ≤ y_i then
15 |  |  |  pf' ← {x_i, y_i};
16 |  |  end
17 |  end
18 end
19 return pf'
```

HMO-REMC sampling method

The pseudo-code of our implemented method for selecting HMO-REMC replicas is presented in Algorithm 6. We experimented using this variant of the MO-REMC

Algorithm 5: SELECT_MO-REMC_REPLICAS(E', P')

Input: E' – energy score in the archives, P' – conformation ensemble in the archives

Output: pe' – conformation ensemble of the last selected "*numR*" min-min scenario Pareto optimal solutions

```
1  PF ← paretoFrontier(E'_id, E', false, false);
2  i ← 0;
3  while i < numR do
4  |  pe' ← PF_|PF|−i;
5  |  i ← i + 1;
6  end
7  return pe'
```

algorithm with 16 protein–small ligand docking cases, which differed only in terms of the procedure used for selecting elite solutions in the MO-REMC sampling method. Updating of the replicas occurs in the MO-REMC method, which ensures that it only contains non-dominated solutions where both the objective MC steps and TScore can be minimized. Thus, the replicas for exchange cover a diverse range of individuals so the min-min scenario non-dominated solutions assigned to replicas truly reflect the quality of the MO-REMC sampling method. The MO-REMC sampling method exclusively uses replicas from the archives where both the objective MC steps and TScore are minimized.

Algorithm 6: SELECT_HMO-REMC_REPLICAS(E',P')

Input: E' – energy score from the archives, P' – conformation ensemble from the archives
Output: pe' – conformation ensemble of selected elite replicas

$PF_{ff} \leftarrow paretoFrontier(E'_{id}, E', false, false)$;
$PF_{tf} \leftarrow paretoFrontier(E'_{id}, E', true, false)$;
$i \leftarrow 0$; $j \leftarrow 0$; $k \leftarrow 0$;
while $(i < |PF_{ff}|)$ && $(j < |PF_{tf}|)$ **do**
  if $E(PF_{ff(i)}) <= E(PF_{tf(i)})$ **then**
    $PF \leftarrow PF_{ff(i)}$; $i \leftarrow i + 1$;
  else
    $PF \leftarrow PF_{tf(j)}$; $j \leftarrow j + 1$;
  end
end
while $i < |PF_{ff}|$ **do**
  $PF \leftarrow PF_{ff(i)}$; $i \leftarrow i + 1$;
end
while $j < |PF_{tf}|$ **do**
  $PF \leftarrow PF_{tf(j)}$; $j \leftarrow j + 1$;
end
while $k < numR$ **do**
  $pe' \leftarrow PF_k$; $k \leftarrow k + 1$;
end
return pe'

Similarly, in the HMO-REMC sampling method, the replica selection method is based on the solutions in the archives where the non-dominated solutions from both the objective MC steps and TScore are minimized, as well as the maximized objective MC steps and minimized objective TScore values. Figure 2 (right and lower panel) shows that lower energy non-dominated solutions are used in min-min and max-min scenarios Pareto optimal solutions as replicas(red points) for exchanging in the HMO-REMC method. The min-min scenario Pareto optimal solutions set is denoted by yellow points and the max-min scenario Pareto optimal solutions set by green points. Obviously, the replicas do not include all of the lower energy non-dominated solutions in the MO-REMC sampling method. Our MO-REMC variant, the HMO-REMC sampling method, uses hybrid non-dominated solutions to select the solutions where both the objective MC steps and TScore are minimized, as well as the maximized objective MC steps and minimized objective TScore non-dominated solutions. In particular, in each replica selection step, all the lower energy non-dominated solutions in both the min-min and max-min scenarios will be used preferentially as replicas for exchange. If the number of solutions is less than *numR*, which is the number of replicas used for exchanging, the non-dominated solutions set is hybridized, where both the min-min and max-min scenarios non-dominated solutions are used iteratively to fill the replica set in order of the TScore value sequence. Replica selection in the MC, REMC, MO-REMC, and HMO-REMC sampling methods is illustrated in Fig. 2.

Implementation in Rosetta

All versions of our MC protein–ligand docking sampling methods were coded in C++ and compiled using g++ (GCC v4.4.7). Algorithm 1 presents the pseudo-code to illustrate the details of our MC search procedure for a single replica with N times MC runs ($N = numR \times numC$) and a predetermined number of temperatures ($T = 2.0$). Algorithm 3, presents the pseudo-code for the implementations of our REMC sampling methods. In order to demonstrate the effectiveness of the REMC algorithms, including REMC, MO-REMC, and HMO-REMC, and without prior knowledge of the problem instances, we fixed the parameter configuration in all of the experimental cases ($numR, numC, repackNth, minT, maxT$) : = (16, 16, 5, 2, 4), where *numR* is the number of replicas simulated, *numC* is the number of local circle steps in REMC search, *repackNth* is the number of iterative steps performed by a *packRotamersMover*() mover, and *minT* and *maxT* are the minimum and maximum temperature values, respectively. All versions of our REMC algorithms were run on 16 processors and they were parallelized.

Multiple independent trajectories were used to generate an ensemble of docking models near the native complex using the MC, REMC, MO-REMC, and HMO-REMC sampling methods. In all of the tests in this study, we performed 5000 docking trajectories (runs), 16 × 16 × 5000 MC steps, for each receptor–ligand pair in the predictive structures, which required 30–50 processor-hours on a 1.9 GHz CPU and 2 GB memory per core Linux cluster. The results of these docking calculations were typically evaluated based on the "energy *versus* rmsd" plot where IFDelta scores were plotted *versus* Lrmsd values, and the effectiveness of each sampling method was judged according to the "funnel-like" character of the plot. In this procedure, we first discarded any

structures where the ligand was not touching the protein (scoring function item ligand_is_touching=0). Second, we took the top 5% of structures based on the total energy. Finally, we ranked the remaining decoys based on the RosettaLigand IFDelta between the protein and ligand. We obtained better results with these ranking scheme and parameters.

Results and discussion

Comparison of different sampling methods

In the procedure using different sampling algorithms, for each crystal structure target in the test data set, the ligand was extracted from the native complex and re-docked into the binding pocket. The Lrmsd value was calculated between the predicted positions C^{α} of the ligand and the ligand C^{α} in the experimental crystal structure, and Lrmsd$\leq$2Å was used as the criterion for success. Using the classic MC sampling method, the protein included backbone translation and rotation as well as repacking of the side-chain of the receptor, and we only selected the lowest pose in terms of energy with the traditional RosettaLigand docking protocol. As shown in Fig. 3, for the 1K3U, and 1OWE targets, the MC sampling method could not produce better experimental binding poses for the ligand in these complexes compared with those reported previously [39] even after 1.28×10^6 MC steps. For 1K3U, and 1OWE, the docking results did not satisfy the requirement in terms of Lrmsd$\leq$2Å, but they converged based on "IFDelta *versus* Lrmsd," as shown by the "funnel-like" character of the plot at the lower left. Successful predictions were made for the 1AQ1 and 2PRG targets using the MC sampling method, but the predictions were not sufficiently good for all of the target protein structures using the four sampling methods (see the docking results obtained using the REMC, MO-REMC, and HMO-REMC sampling methods in the figure).

The aim of REMC sampling methods is to increase the scope and depth of sampling by exchanging configurations between replicas characterized by slightly different temperature parameters. The REMC sampling method has been employed widely to enhance sampling methods by crossing energy barriers and accelerating the convergence of MC simulations. For a specific target, the MC sampling method may not be sufficient to cover some important regions of the conformational space that can be recognized by a number of ligands. However, enhanced sampling methods such as REMC, MO-REMC, and HMO-REMC can be used to generate a large number of receptor conformations for protein–ligand docking. Thus, in this study, in order to sample more of the receptor backbone and side-chain flexibility in each case, we tested 5000 decoys with each enhanced sampling method and only selected the lowest energy pose from these trajectories based on the IFDelta function as implemented in RosettaLigand [38, 39]. As shown in Fig. 3, the RosettaLigand protocol based on the REMC method obtained the lower energy pose (1OWE), faster convergence of the lower energy pose (2PRG), cross-local energy minima (1K3U), and the binding poses of the alternative ligand for the first pose within 2Å Lrmsd. By contrast, for 2PRG, the MO-REMC and HMO-REMC sampling algorithms obtained nearly perfect results within 1Å Lrmsd as well as faster convergence for more of the predicted structures with the lowest IFDelta scores.

Comparison of different sampling scales

The evolution of sampling in terms of the IFDelta and Lrmsd scores with different sampling scales is shown for one representative target (2PRG) in Fig. 4. For 2PRG, the four sampling methods could progressively sample lower (more favorable) IFDelta values as the number of MC steps increased from 2.56×10^5 to 1.28×10^6. However, the enhanced sampling methods obtained faster convergence in terms of IFDelta, as well as the HMO-REMC method compared with the MO-REMC method for Lrmsd<=2Å. The MC sampling method successfully sampled solutions with Lrmsd<=2Å after 1.28×10^6 steps, whereas the REMC, MO-REMC, and HMO-REMC sampling methods could reach near-native solutions, particularly the MO-REMC method, which obtained Lrmsd<1Å solutions after only 7.68×10^5 MC steps. In terms of the IFDelta scores, after 1.28×10^6 MC steps, the MC sampling algorithm successfully sampled near-native solutions with Lrmsd of 1.42Å and the IFDelta score value was −18.8. By contrast, after only 2.56×10^5 MC steps, the REMC, MO-REMC, and HMO-REMC methods obtained Lrmsd scores within 1.20Å, 1.14Å, and 1.33Å, respectively, and the IFDelta scores were −18.4, −18.9, and −17.2, respectively. Furthermore, after 1.28×10^6 MC steps, the three enhanced sampling algorithms sampled near-native solutions with Lrmsd scores of 1.20Å, 0.79Å, and 0.69Å, respectively. In addition, the IFDelta scores converged around -18.6 ± 0.3. Similar trends were also observed in all the other test cases.

Summary of the docking results obtained using different sampling methods and scales

In general, better docking results are achieved by sampling lower docking score value conformations. So, the first parameter that we evaluated was the global performance of the docking results in terms of the IFDelta score. For all 16 cases, the evolution in terms of IFDelta using different sampling scales in the four sampling methods is shown in Fig. 5. As shown by the histogram of IFDelta values for the 16 individual targets, the four sampling methods could sample near-native docking solutions with more negative IFDelta scores at three sampling scales in 2.56×10^5, 7.68×10^5, and 1.28×10^6 MC steps.

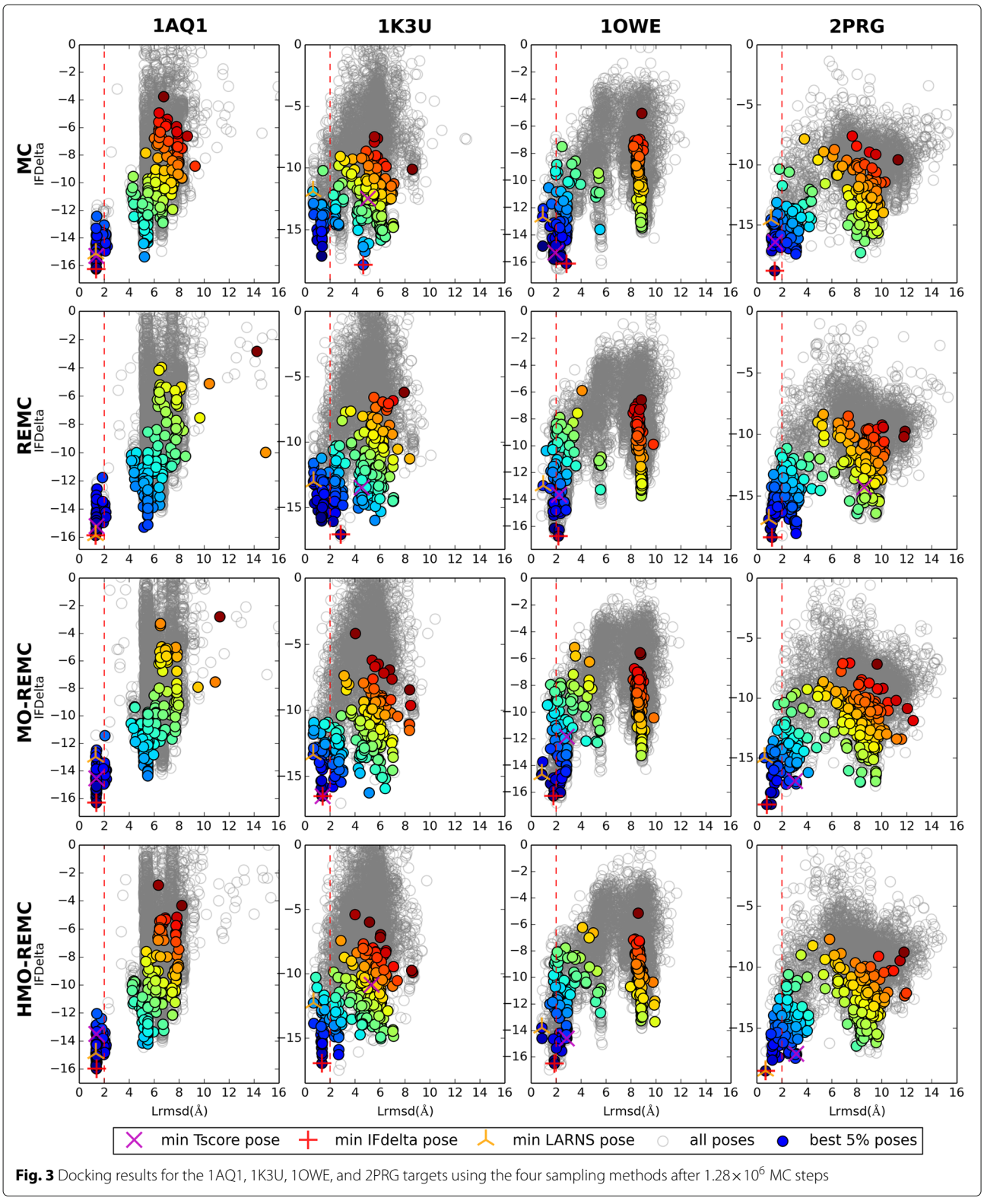

Fig. 3 Docking results for the 1AQ1, 1K3U, 1OWE, and 2PRG targets using the four sampling methods after 1.28×10^6 MC steps

However, using the same number of MC steps (2.56×10^5, 7.68×10^5, or 1.28×10^6), the enhanced sampling methods could sample solutions with lower IFDelta scores than the classic MC sampling method. The MO-REMC, and HMO-REMC enhanced sampling methods obtained better docking results in 9/16 cases (1AQ1, 1DBJ, 1JJE,

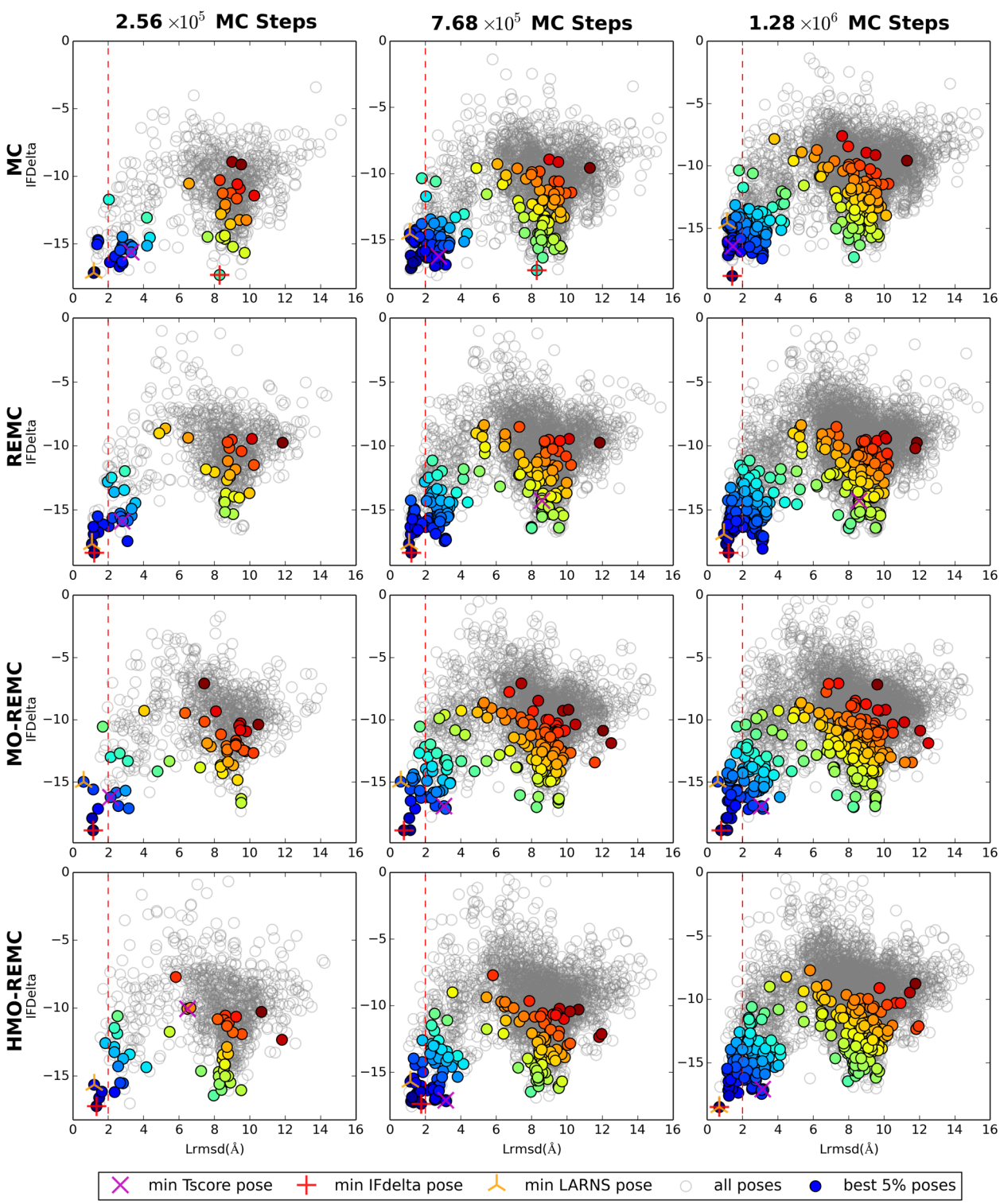

Fig. 4 Docking results for the 2PRG target using the four sampling methods after 2.56×10^5, 7.68×10^5, or 1.28×10^6 MC steps

1TOW, 1V48, 1Y6B, 2DBL, 2PRG, and 7CPA) with lower final IFDelta scores compared with the standard MC and REMC sampling methods after 1.28×10^6 MC steps. It was interesting that the MO-REMC sampling method obtained better docking results in 7/16 cases, with lower IFDelta scores compared with the HMO-REMC sampling method. However, in 3/16 cases (1OWE, 1PQ6, and 4TIM), the REMC method obtained configurations,

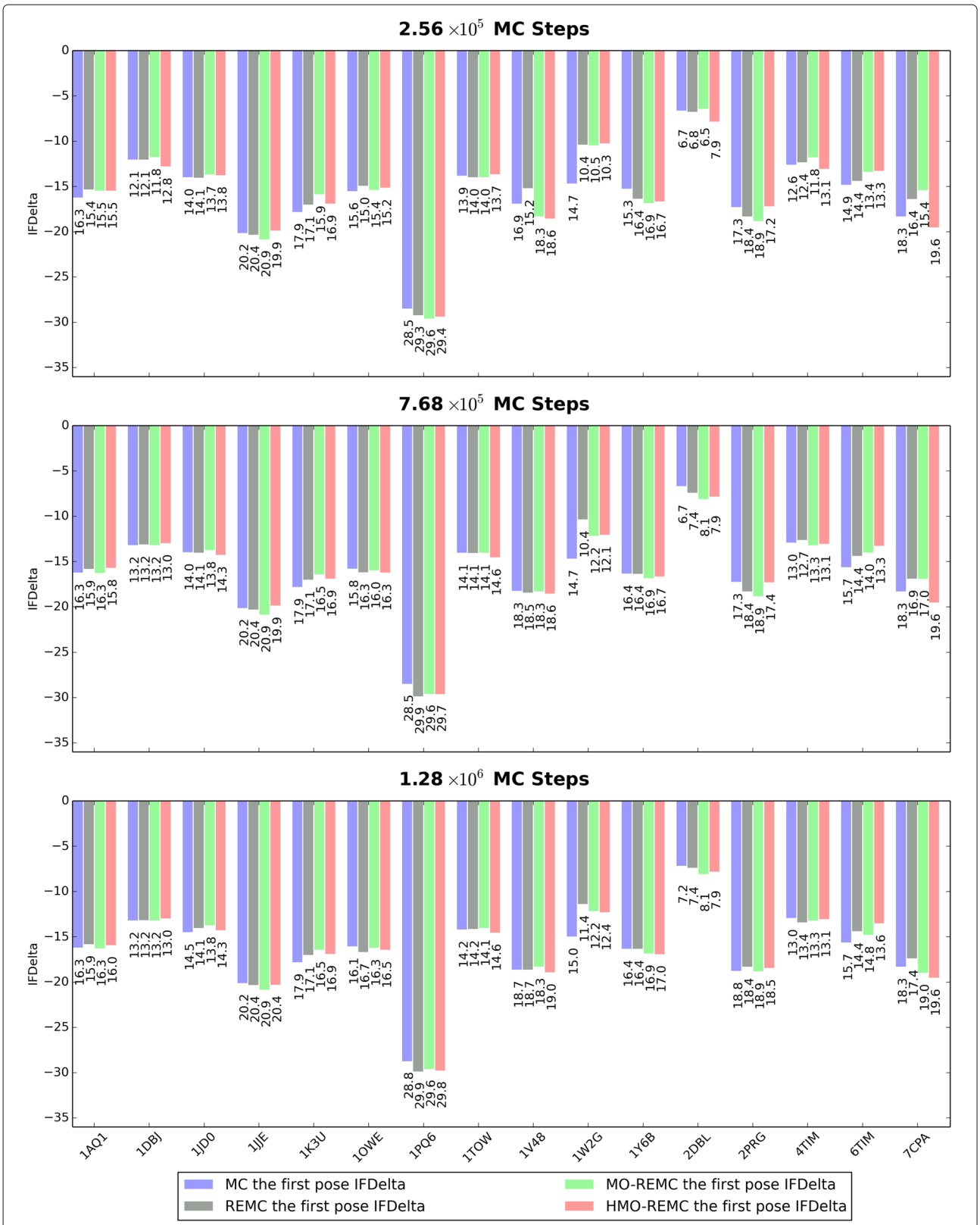

Fig. 5 Summaries of the IFDelta scores with the MC, REMC, MO-REMC, and HMO-REMC sampling methods at three sampling scales

which were closer to the lower binding energy form compared with the MO-REMC methods. By contrast, the MC sampling algorithm succeeded also in the cases of 1JD0, 1K3U, 1W2G, and 6TIM after 1.28×10^6 MC steps. The results based on the 16 test cases indicate that the MO-REMC and HMO-REMC enhanced sampling methods performed better than the MC and REMC sampling methods. The results also showed that the IFDelta values could vary dramatically for different targets and sampling methods, whereas the IFDelta scores obtained for the same target with the REMC, MO-REMC, and HMO-REMC enhanced sampling methods varied only slightly. For example, for 1PQ6, the MC method achieved an IFDelta value of around −28.8, whereas the REMC, MO-REMC, and HMO-REMC sampling methods obtained a binding energy value of around -29.8 ± 0.2. In addition, for 1DBJ, the four sampling methods achieved similar IFDelta scores of around 13.2 ± 0.2 after 1.28×10^6 MC steps.

The second parameter that we analyzed was the overall performance of the docking results in terms of the Lrmsd value. An overview of the Lrmsd values obtained for the individual targets is shown in Fig. 6 at three sampling scales of 2.56×10^5, 7.68×10^5, or 1.28×10^6 MC steps. For each target, the Lrmsd values are presented after docking the ligand into alternative receptor structures using the MC, REMC, MO-REMC, and HMO-REMC sampling methods. For each of the 16 targets, the bars from left to right correspond to the results for the protein based on the MC, REMC, MO-REMC, and HMO-REMC sampling methods, respectively. For 1OWE, and 1W2G, the MC sampling method achieved better (slightly) solutions within 2Å Lrmsd compared with the enhanced sampling methods after 2.56×10^5 MC steps. However, for 14/16 cases, excluding 1JD0, and 1TOW, using the enhanced sampling methods achieved better minimum Lrmsd values for docking with the protein than the MC sampling method after 1.28×10^6 MC steps. In particular, for 1W2G, and 2PRG, the MO-REMC enhanced sampling method obtained Lrmsd values that were close to the perfect results within 1Å Lrmsd. These results have never been obtained before using MC sampling methods, and we showed that the MC sampling method could not obtain satisfactory samples of complicated protein flexibility after 1.28×10^6 MC steps. Finally, the Lrmsd values for individual protein structures could vary dramatically using different sampling methods and they changed greatly after 1.28×10^6 MC steps, thereby suggesting that in structure-based protein–ligand docking experiments, different sampling methods can significantly affect the docking results in terms of both depth and breadth. For example, for 2PRG, the best performing target obtained using the MC sampling method only achieved an Lrmsd value of around 8.29Å after 7.68×10^5 MC steps, whereas the REMC enhanced sampling method gave an Lrmsd value within 1.20Å, but the MO-REMC and HMO-REMC sampling methods obtained the best docking results within 0.79Å and 1.76Å Lrmsd, respectively.

Convergence with different sampling methods

Next, we briefly discuss how different sampling methods can affect the rate of convergence. Firstly, in order to demonstrate that the MO-REMC and HMO-REMC sampling methods proposed here provide an efficient sampling technique in temperature space, we calculate the probability of finding each replica at different temperatures. For the RosettaLigand docking protocol, the probability value with energy *score* in heat capacity and temperature T is described by Eq. (4), but no exponential calculation. In Fig. 7, for target 2PRG, we show that using the MC, REMC, MO-REMC, and HMO-REMC sampling methods, the probability of finding each replica at different temperatures progressively flattened over $numR = 16$ replicas simulated through $numC = 16$ local circles($numR\times numC$ MC steps). On each subfigure, the red circle points correspond to the probability average values of $numR = 16$ replicas simulated through $numC = 16$ local circles. The results obtained by the MC sampling method show that after $numR\times numC$ MC steps, the probability average values converged slowly to 1.56 with a wider fluctuation variance value of 8.78. However, using the enhanced sampling methods, REMC, MO-REMC, and HMO-REMC, the results show that the probability average values converged faster to 0.78, 0.73, and 0.99, with a narrow margin fluctuation variance value of 0.22, 0.18, and 0.07, respectively. Especially, for the HMO-REMC sampling method, the probability values of finding each replica at different temperatures show a fairly flat probability distribution. The probability results show that a strong temperature dependence of energy for complex protein–ligand docking systems.

Secondly, in Fig. 8, for the 2PRG and 4TIM targets, we show how the estimated TScore and IFDelta scores obtained using the MC, REMC, MO-REMC, and HMO-REMC sampling methods converged over a simulation of 1.28×10^6 MC steps. For comparison, we also show the same Lrmsd values calculated after 1.28×10^6 MC steps. For 2PRG, the results obtained by the enhanced sampling methods are shown that after 7.68×10^5 MC steps, which demonstrate that the IFDelta score converged almost exactly to −18.2, with small fluctuations in the order of $\approx$0.8. However, using the classic MC sampling method, almost all of the 7.68×10^5 MC steps were required to obtain a converged result with an IFDelta value in the order of −17.3, as shown in Fig. 8 (left and middle panels). After 1.28×10^6 MC steps, however, four sampling methods could obtain better convergence in terms of IFDelta,

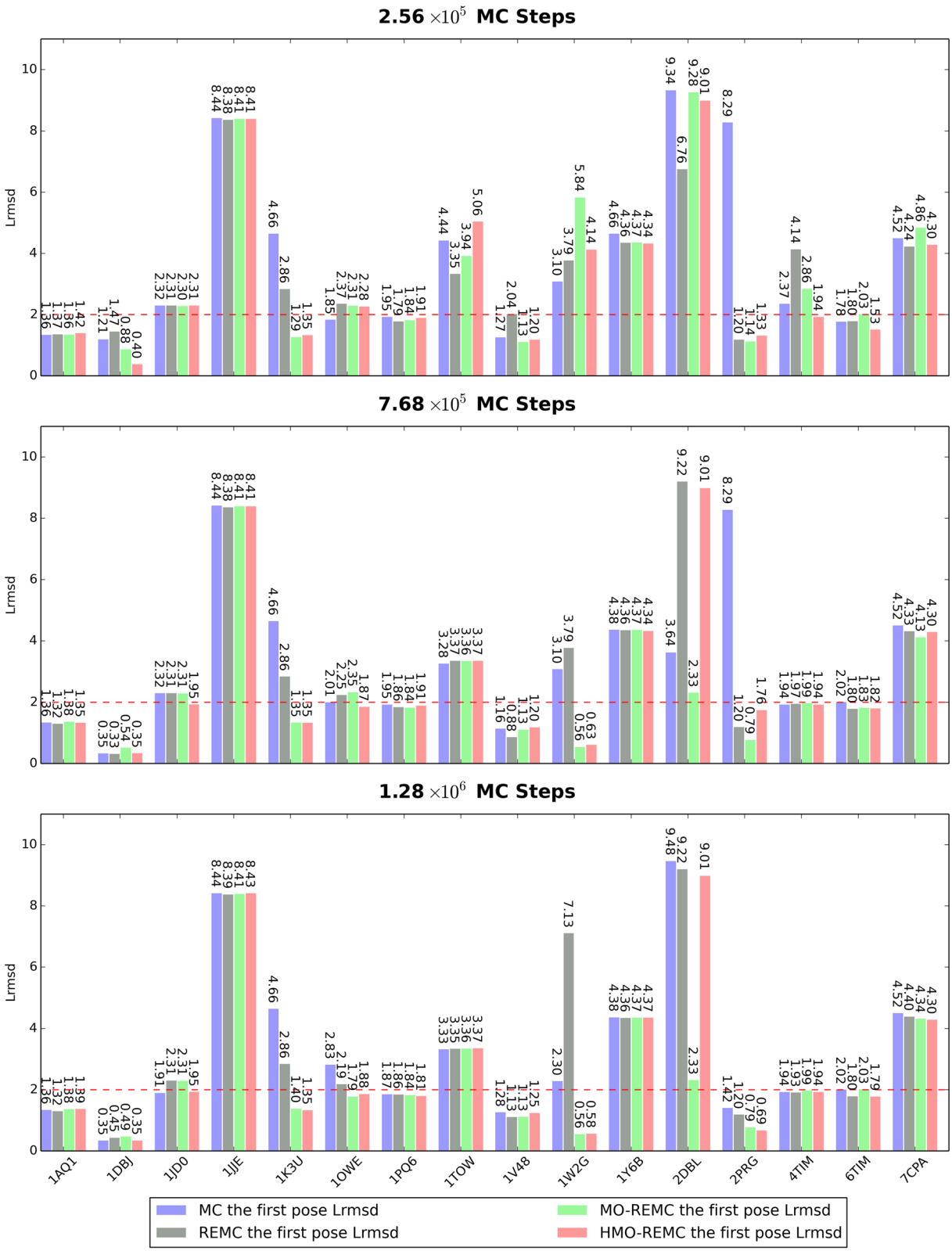

Fig. 6 Summaries of the Lrmsd values with the four sampling methods after 2.56×10^5, 7.68×10^5, or 1.28×10^6 MC steps

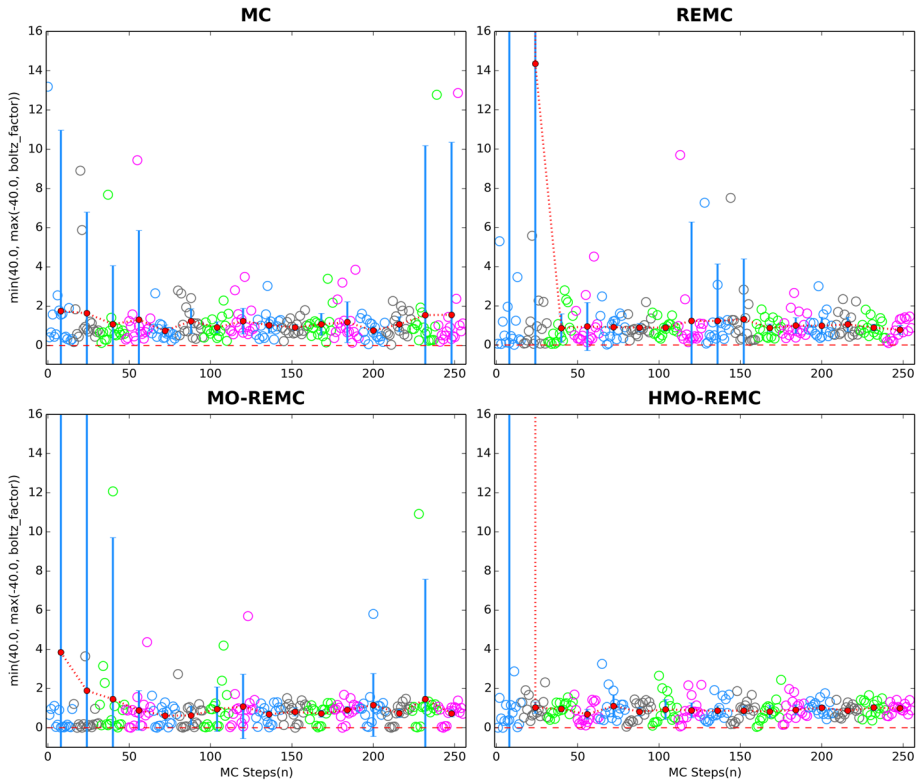

Fig. 7 For 2PRG, the probability of finding each replica at different temperatures

TScore, and Lrmsd. This represents an improvement in the sampling efficiency by one order of magnitude and it is very likely that this could be improved further, such as by incorporating information from the Hamiltonian. One of the most important tests of convergence for a protein–ligand interaction when sampling a complex transformation is the sensitivity of the results to different sampling methods. Thus, to exclude any dependence on different sampling methods, we also calculated the Lrmsd values for the four sampling methods after 1.28×10^6 MC steps and found that the estimated Lrmsd values agreed very well with the results based on the IFDelta values. These results are presented in Fig. 8 (left and lower panel). To facilitate a comparison with other targets, we also performed sampling for 4TIM using the four sampling methods, as shown in Fig. 8 (right panels), which clearly demonstrate that running 1.28×10^6 MC steps for 4TIM was sufficient to obtain a converged estimate of the IFDelta score. In particular, using the MO-REMC sampling method, the estimate of the IFDelta score converged rapidly. However, the MC sampling method might obtain better convergence in terms of IFDelta and TScore as well as Lrmsd, but the rate of convergence was slower. The REMC sampling method achieved better convergence after 1.28×10^6 MC steps, but the results indicated that the convergence rate was slower than that using the MO-REMC and HMO-REMC sampling methods in terms of speed and depth. In addition, the HMO-REMC sampling method performed slightly better than the MO-REMC sampling method in 9/16 cases after 1.28×10^6 MC steps, as shown in Fig. 6 (lower panel).

Finally, we present further evidence that the MO-REMC and HMO-REMC sampling methods are effective sampling techniques in temperature space. For 2PRG, we conduct more than one simulation run using different initial parameters($(numR, numC, repackNth, minT, maxT)$: = (8, 8, 3, 2, 4), (8, 8, 3, 2, 6), (8, 16, 5, 2, 4), (8, 16, 5, 2, 6), (16, 8, 3, 2, 4), (16, 8, 3, 2, 6), (16, 16, 5, 2, 4), and (16, 16, 5, 2, 6), respectively.). In this way, the summaries of IFDelta and Lrmsd through creating different trajectories(decoys) of configurations for each initial parameter are compared in Fig. 9. We show that using different initial parameters, the new methods (including MO-REMC and HMO-REMC) proposed in this research can efficiently converge to possible near-native solutions. In addition, we also compare the necessary number of sampling decoys to reach convergence in simulation runs using different initial parameters. The results show that when using different initial parameters, better near-native

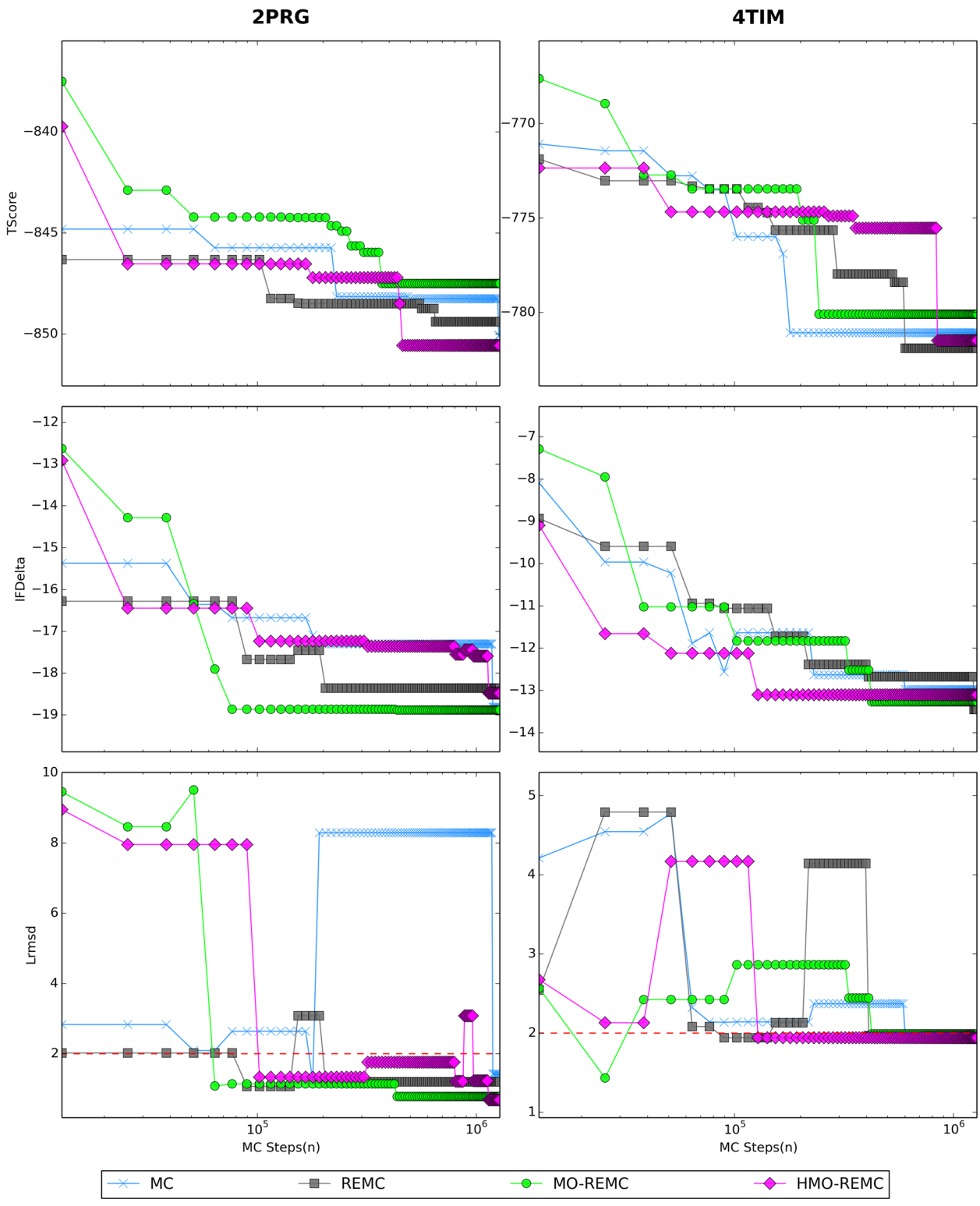

Fig. 8 Convergence of TScore, IFDelta, and Lrmsd *vs.* MC steps for the 2PRG and 4TIM targets

conformations can be achieved after sampling 5×10^3 decoys($numR \times numC \times N$ MC steps, the N is the number of decoys). However, for different initial parameters, the necessary number of sampling decoys may be conspicuously different when using different sampling methods.

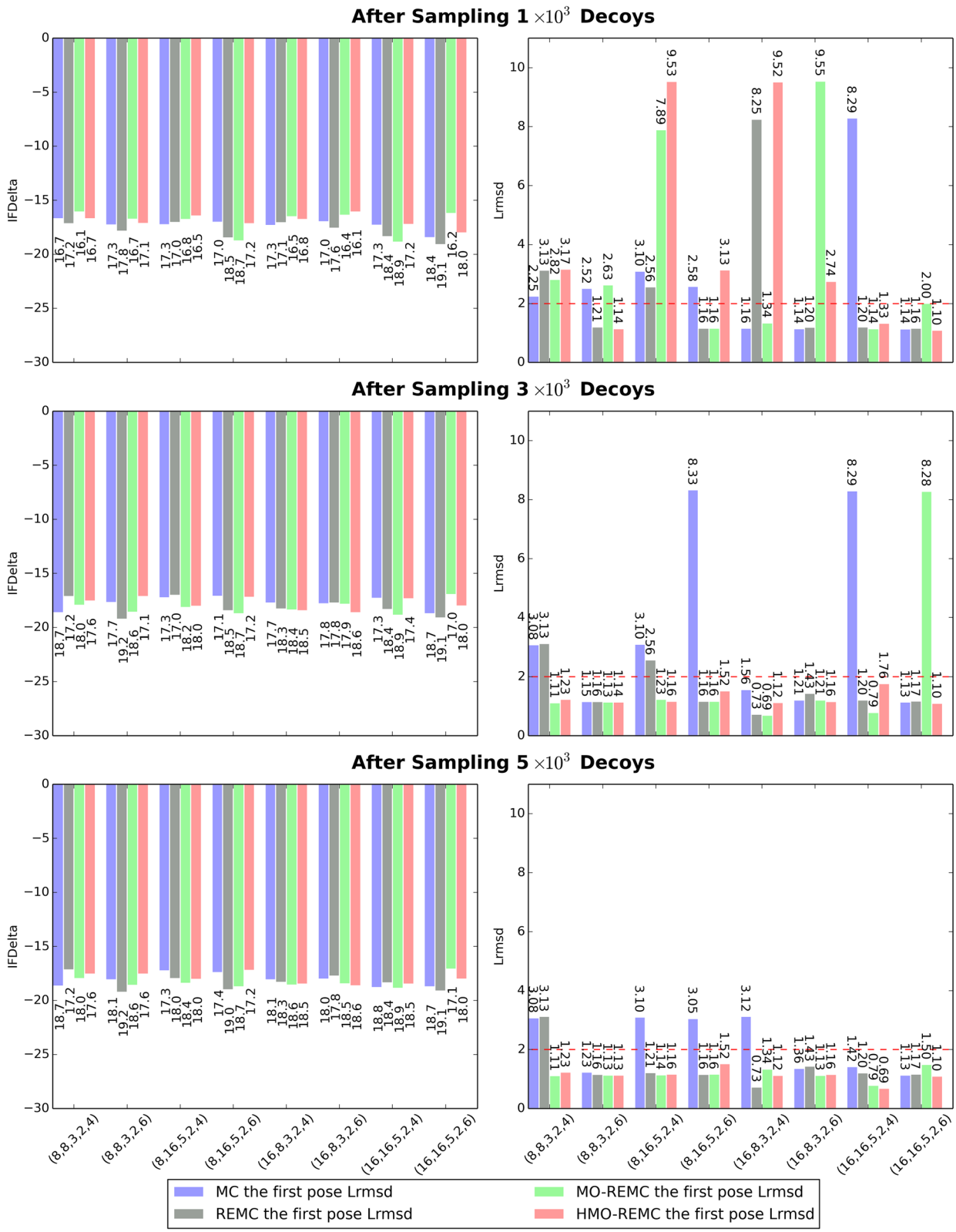

Fig. 9 For 2PRG, summaries of the IFDelta and Lrmsd using different initial parameters, sampling methods, and scales

Conclusions

In this study, we developed REMC sampling methods based on multi-objective optimization for predicting conformations in protein–small ligand docking with RosettaLigand. We used temperature replica exchange to enhance conformational sampling between Pareto optimal solutions, and the concept of non-dominated solutions was applied to solve the replica selection problem in our REMC enhanced sampling methods. In contrast to most other MC and REMC methods, the MO-REMC method selects non-dominated solutions, which depend on archived solutions measured in terms of the objective MC steps and TScore values, in order to find a set of similar replicas with lower energy conformations but that are also as diverse as possible. The MO-REMC and HMO-REMC methods achieve better integration of the REMC sampling method and critical conformation structures of the current sampling state. Using a benchmark data set of 16 protein-ligand test cases with different chain lengths in terms of amino acids, we assessed various comparison measures, i.e., TScore, IFDelta, and Lrmsd. We also considered the funnel-like character of the energy landscape, the probability of finding each replica at different temperatures, and the rate of convergence in the TScore, IFDelta, and Lrmsd scores.

For the targets tested in our benchmark data set, we found that the ligand pose was correctly positioned within 2Å Lrmsd for 11/16 of these targets using the HMO-REMC sampling method after 1.28×10^6 MC steps. The performance of the proposed MO-REMC sampling method was better than that of the MC and REMC methods in most cases, whereas MC generally performed better but converged slowly. The MO-REMC sampling method achieved significantly faster convergence of the lower energy poses and identified more correct docked complexes with near-native decoys than the MC and REMC sampling methods. Moreover, the results showed that HMO-REMC obtained faster convergence and more distinct solutions than MO-REMC in each run for most of the targets. The MO-REMC and HMO-REMC methods required the same or slightly more time than MC and REMC for the same number of sampling steps. Moreover, for the 1DBJ, 1JD0, 1K3U, 1PQ6, 1Y6B, 2PRG, 4TIM, 6TIM, and 7CPA targets, the performance of HMO-REMC was much better than that of MO-REMC. The HMO-REMC sampling method captured much of the possible variation in the conformations for most of the test cases, and it also sampled lower binding energy conformations within an Lrmsd of 2Å for the conformations of these targets. The HMO-REMC hybridizes two scenario combinations for the Pareto optimal solutions with the ranking-based MO-REMC method and it worked well for many targets. An interesting feature of the MO-REMC method compared with other REMC algorithms is that many non-dominated solutions are chosen as the current replicas for exchange. Thus, sampling at a lower energy is a much more greedy process, which leads to better protein–ligand conformational sampling performance. Clearly, this feature can also be incorporated in addition to the concept of Pareto front solutions in other ensemble-based sampling methods in order to improve their performance. In addition, experimental results showed that faster convergence to the global optimal solution does not necessarily provide an efficient algorithm for enhancing conformational sampling of the phase space. Use of temperature replica exchange to enhance conformational sampling between non-dominated solutions can also provide good convergence of the available conformational space including available near-native structures.

In the future, the proposed MO-REMC method may be extended in several ways. Even though detailed balance is not satisfied in the MO-REMC and HMO-REMC sampling methods, some balance condition may still be efficient if it is proved that it provides a good sampling method. We can still generate an algorithm that may satisfy the balance criteria, for example, instead of selecting the conformation ensemble of Pareto optimal solutions, the configurations to be swapped from the history archives can be randomly selected instead. Moreover, in order to obtain performance improvements, several enhanced sampling techniques, including Hamiltonian replica exchange and well-tempered ensemble approaches, or even a dynamic temperature selection strategy, can be incorporated in the MO-REMC and HMO-REMC methods.

Abbreviations

HMO-REMC: Hybrid MO-REMC; IFDelta: Binding energy interface delta; Lrmsd: Ligand of RMSD; MC: Monte Carlo; MO-REMC: Multi-objective optimization-REMC; MOP: Multi-objective optimization problem; REMC: Replica exchange Monte Carlo; TScore: The RosettaLigand energy function total score

Acknowledgements

The authors thank W. Yang and C.Y. Cao for advice, valuable ideas, and comments. The authors thank Y. Su for help with preparing the paper and H.Zhou for research support. The authors would like to thank Soochow University for support to complete this study.

Funding

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed in the current study are available in the RCSB PDB repository, http://www.rcsb.org.

Authors' contributions

HRW conceived the study. HRW and HWL developed the MO-REMC approach and wrote the statistical software. LXC was involved with data analysis. CXW ran analyses and prepared the text, data, and figures. HRW and QL drafted the manuscript. All of the authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Author details
[1]School of Computer Science and Technology, Soochow University, 1 Shizi Street, 215006 Suzhou, People's Republic of China. [2]Jiangsu Provincial Key Lab for Information Processing Technologies, 1 Shizi Street, 215006 Suzhou, People's Republic of China.

Received: 2 November 2016 Accepted: 15 June 2017
Published online: 10 July 2017

References

1. Maximova T, Moffatt R, Ma BY, Nussinov R, Shehu A. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. Plos Comput Biol. 2016;12(4):e1004619. doi:10.1371/journal.pcbi.1004619.
2. Dror RO, Dirks RM, Grossman JP, Xu HF, Shaw DE. Biomolecular simulation: A computational microscope for molecular biology. Annu Rev Biophys, Vol 41. 2012;41:429–52. doi:10.1146/annurev-biophys-042910-155245.
3. Ewing TJA, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. J Comput Aid Mol Des. 2001;15(5):411–28. doi:10.1023/A:1011115820450.
4. Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins. 1999;37(2): 228–41. doi:10.1002/(SICI)1097-0134199911013 7:2<228::AID-PROT8>3.0.CO;2-8.
5. Wagener M, de Vlieg J, Nabuurs SB. Flexible protein-ligand docking using the Fleksy protocol. J Comput Chem. 2012;33(12):1215–7. doi:10.1002/jcc.22948.
6. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins-Structure Funct Genet. 2003;52(4):609–23. doi:10.1002/prot.10465.
7. Verdonk ML, Chessari G, Cole JC, Hartshorn MJ, Murray CW, Nissink JWM, Taylor RD, Taylor- R. Modeling water molecules in protein-ligand docking using GOLD. J Med Chem. 2005;48(20):6504–15. doi:10.1021/jm050543p.
8. Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of AutoDock. J Mol Recognit. 1996;9(1):1–5. doi:10.1002/(SICI)1099-13521996019:1<1::AID-JMR241>3.0.CO;2-6.
9. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785–91. doi:10.1002/jcc.21256.
10. Trott O, Olson AJ. Software news and update AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61. doi:10.1002/jcc.21334.
11. Vass M, Tarcsay A, Keseru GM. Multiple ligand docking by Glide: implications for virtual second-site screening. J Comput Aid Mol Des. 2012;26(7):821–34. doi:10.1007/s10822-012-9578-6.
12. Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res. 2011;39: W270–W277. doi:10.1093/nar/gkr366.
13. Rao L, Chi B, Ren YL, Li YJ, Xu X, Wan J. DOX: A new computational protocol for accurate prediction of the protein-ligand binding structures. J Comput Chem. 2016;37(3):336–44. doi:10.1002/jcc.24217.
14. Huang SY, Li M, Wang JX, Pan Y. HybridDock: A hybrid protein-ligand docking protocol integrating protein- and ligand-based approaches. J Chem Inf Model. 2016;56(6):1078–87. doi:10.1021/acs.jcim.5b00275.
15. Pan L-L, Zheng Z, Wang T, Merz KM. Free energy-based conformational search algorithm using the movable type sampling method. J Chem Theory Comput. 2015;11(12):5853–64. doi:10.1021/acs.jctc.5b00930.
16. Kirkpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science. 1983;220(4598):671–80. doi:10.1126/science.220.4598.671.
17. Goldberg D. Genetic algorithms in search, optimization and machine learning. New York: Addison-Wesley Publishing Company, Inc.; 1989.
18. Luitz M, Bomblies R, Ostermeir K, Zacharias M. Exploring biomolecular dynamics and interactions using advanced sampling methods. J Phys-Condens Mat. 2015;27(32):323101. doi:10.1088/0953-8984/27/32/323101.
19. Valsson O, Parrinello M. Variational approach to enhanced sampling and free energy calculations. Phys Rev Lett. 2014;113(9):090601. doi:10.1103/Physrevlett.113.090601.
20. Bernardi RC, Melo MC, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. Biochim Biophys Acta. 2015;1850(5):872–7. doi:10.1016/j.bbagen.2014.10.019.
21. Swendsen RH, Wang J-S. Replica Monte Carlo simulation of spin-glasses. Phys Rev Lett. 1986;57(21):2607–9.
22. Geyer CJ. Markov chain Monte Carlo maximum likelihood. In: Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface. Fairfax Station: Interface Foundation of North America; 1991. p. 156–63.
23. Earl DJ, Deem MW. Parallel tempering: Theory, applications, and new perspectives. Phys Chem Chem Phys. 2005;7(23):3910–6. doi:10.1039/b509983h.
24. Zhang Z, Lange OF. Replica exchange improves sampling in low-resolution docking stage of RosettaDock. Plos One. 2013;8(8):e72096. doi:10.1371/journal.pone.0072096.
25. Sambridge M. A parallel tempering algorithm for probabilistic sampling and multimodal optimization. Geophys J Int. 2014;196(1):357–74. doi:10.1093/gji/ggt342.
26. Russo A, Scognamiglio PL, Enriquez RPH, Santambrogio C, Grandori R, Marasco D, Giordano A, Scoles G, Fortuna S. In silico generation of peptides by replica exchange Monte Carlo: docking-based optimization of Maltose-binding-protein ligands. Plos One. 2015;10(8):e0133571. doi:10.1371/journal.pone.0133571.
27. Fukunishi H, Watanabe O, Takada S. On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. J Chem Phys. 2002;116(20): 9058–67. doi:10.1063/1.1472510.
28. Luitz MP, Zacharias M. Protein-ligand docking using Hamiltonian replica exchange simulations with soft core potentials. J Chem Inf Model. 2014;54(6):1669–75. doi:10.1021/ci500296f.
29. Ostermeir K, Zacharias M. Hamiltonian replica-exchange simulations with adaptive biasing of peptide backbone and side chain dihedral angles. J Comput Chem. 2014;35(2):150–8. doi:10.1002/jcc.23476.
30. Barducci A, Bussi G, Parrinello M. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. Phys Rev Lett. 2008;100(2):020603. doi:10.1103/Physrevlett.100.020603.
31. Bonomi M, Parrinello M. Enhanced sampling in the well-tempered ensemble. Phys Rev Lett. 2010;104(19):190601. doi:10.1103/Physrevlett.104.190601.
32. Valsson O, Parrinello M. Well-tempered variational approach to enhanced sampling. J Chem Theory Comput. 2015;11(5):1996–2002. doi:10.1021/acs.jctc.5b00076.
33. Zhang Z, Schindler CEM, Lange OF, Zacharias M. Application of enhanced sampling Monte Carlo methods for high-resolution protein-protein docking in Rosetta. Plos One. 2015;10(6):e0125941. doi:10.1371/journal.pone.0125941.
34. Li BD, Li JL, Tang K, Yao X. Many-objective evolutionary algorithms: A survey. Acm Comput Surv. 2015;48(1):13. doi:10.1145/2792984.
35. von Lücken C, Barán B, Brizuela C. A survey on multi-objective evolutionary algorithms for many-objective problems. Comput Optim Appl. 2014;58(3):707–56. doi:10.1007/s10589-014-9644-1.
36. Deb K. Multi-objective optimization In: Burke KE, Kendall G, editors. Search Methodologies: Introductory Tutorials in Optimization and Decision. Support Techniques. Boston: Springer; 2014. p. 403-49.
37. Deb K, Kalyanmoy D. Multi-objective optimization using evolutionary algorithms. Chichester: John Wiley & Sons, Inc; 2001. pp. 389–400.
38. Meiler J, Baker D. ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility. Proteins. 2006;65(3):538–48. doi:10.1002/prot.21086.

39. Davis IW, Baker D. ROSETTALIGAND docking with full ligand and receptor flexibility. J Mol Biol. 2009;385(2):381–92. doi:10.1016/j.jmb.2008.11.010.
40. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and Cambridge structural database. J Chem Inf Model. 2010;50(4):572–84. doi:10.1021/ci100031x.
41. Hawkins PCD, Nicholls A. Conformer generation with OMEGA: Learning from the data set and the analysis of failures. J Chem Inf Model. 2012;52(11):2919–36. doi:10.1021/ci300314k.
42. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol. 2003;331(1):281–99. doi:10.1016/S0022-28360300670-3.
43. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J Chem Phys. 1953;21(6):1087–92. doi:10.1063/1.1699114.
44. Daan F, Berend S. Understanding molecular simulation from algorithms to applications. San Diego: Academic Press; 2002, pp. 111–38.
45. Thachuk C, Shmygelska A, Hoos HH. A replica exchange Monte Carlo algorithm for protein folding in the HP model. BMC Bioinforma. 2007;8:342. doi:10.1186/1471-2105-8-342.