RLDOCK: A new method for predicting RNA-ligand interaction

Journal:	*Journal of Chemical Theory and Computation*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Chen, Shi-Jie; University of Missouri Columbia, Physics and Biochemistry Sun, Lizhen; Zhejiang University of Technology

RLDOCK: A new method for predicting RNA-ligand interaction

Li-Zhen SUN[1,2], Shi-Jie CHEN[2,‡]

[1]Department of Applied Physics, Zhejiang University of Technology, Hangzhou 310023, China

[2]Department of Physics, Department of Biochemistry, and Informatics Institute, University of Missouri, Columbia, MO 65211

Abstract

The ability to accurately predict the binding site, binding pose, and binding affinity for a ligand bound to an RNA is important for RNA-targeted drug design. Here we describe a new computational method, RLDOCK, for predicting the binding site and binding pose for ligand-RNA binding. By developing an energy-based scoring function, we sample exhaustively all the possible binding sites for a ligand-RNA pair based on the geometric and energetic scores. The model distinguishes from other approaches in three notable features. First, the model enables exhaustive scanning of all the possible binding sites, including multiple alternative or coexisting binding sites, for a given ligand-RNA pair. Second, the model is based on a new energy-based scoring function developed here. Third, the model employs a novel multi-step screening algorithm to improve computational efficiency. Specifically, first, we rank order the different binding sites according to the minimum LJ potential energy for the different ligand poses at the given site. Second, for each highly ranked sites, we predict the ligand pose using a two-step algorithm. In the first step, we quickly identify the probable ligand poses using a coarse-grained simplified energy function. In the second step, for each of the probable ligand poses, we predict the ligand poses using a refined energy function. Tests of the RLDOCK for a set of 230 RNA-ligand bound structures indicate that RLDOCK can successful predict ligand poses for 34.8%, 61.3%, and 65.7% of all the test cases with the root-mean-square deviation within 1.0, 2.0, and 3.0 Å, respectively, for the top-three predicted docking poses. The computational method presented here may enable the development of a new, more comprehensive framework for the prediction of ligand-RNA binding with allosteric conformational changes and metal ions effects.

---

‡Author to whom correspondence should be addressed; E-mail: chenshi@missouri.edu

1 Introduction

RNA molecules can fold into complicated tertiary structures that contain various motifs such as pseudoknots, kissing loops, hairpins, and deep major groove.[1] Such a plethora of tertiary structural motifs can lead to a variety of ideal binding sites for small molecules and hence make RNAs potential drug targets.[2] The RNA druggability is particularly important as protein druggability is severely limited by the lack of proper binding pockets.[3,4] Furthermore, RNA is critical to gene expression,[5] thus inhibiting RNA function may lead to the termination of the production of dozens or more proteins.[6,7] A notable example for the effect of ligand-RNA binding is riboswitch,[8–12] where ligand binding results in the termination or go-through for transcription or translation.[12] Furthermore, because RNAs can directly participate in the protein syntheses[6] through the formation of specific active sites in ribosome,[13] or in viral gene regulation.[14] Ligand-RNA binding can directly impact gene expression and viral replication.

In recent years, computational docking/scoring methods, such as rDOCK,[15,16] LigandRNA,[17] DrugScore$^{\mathrm{RNA}}$,[18,19] Dock6,[20] and MORDOR,[21] have been developed to predict RNA-ligand interactions. These docking/scoring methods can be classified into two types:

1. **Knowledge-based methods.** DrugScore$^{\mathrm{RNA}}$[18,19] uses a set of distance-dependent knowledge-based potentials to score and rank the ligand poses for an RNA target. The potentials used in DrugScore$^{\mathrm{RNA}}$[18,19] are described by the distance between a pair of atoms in the ligand and the RNA, respectively. The potentials are derived from 670 crystallographic RNA-ligand and RNA-protein complexes. However, only 50 RNA-ligand complexes are contained among all the ligand-RNA complexes. LigandRNA[17] is another knowledge-based method. Its potentials consider not only the distances but also the angles between RNA atom pairs and ligand atoms. The potentials used in LigandRNA[17] are derived from 251 RNA-ligand complexes. rDOCK[16] is an open-source software that can predict ligand binding poses in nucleic acids and proteins. The (empirical) scoring functions in rDOCK[15] account for specific RNA-ligand interactions such as those involving aromatic groups.

2. **Physics-based methods.** MORDOR[21] applies the all-atom CHARMM27 force field[22] to the target and the general AMBER force field[23] to the ligand. The model allows both the ligand and the target to be flexible.

DOCK6[20] is another typical physics-based docking procedure. In DOCK6, AMBER force field[23,24] is applied to both ligand and RNA, and the scoring function contains the solvent and the (sodium) ion effects on docking. In comparison, MORDOR may provide a higher accuracy in predictions of a specific ligand database, while DOCK6 has a higher efficiency and be applicable to broader databases.[20] In addition to MORDOR and DOCK6, other physics-based scoring/docking methods focus mainly on specific types of ligands such as aminoglycoside antibiotics[25] or targets such as purine riboswitch.[26]

The applications of the existing scoring/docking methods have been successful at different levels.[27] The accuracy of the knowledge-based methods is dependent on the training database of RNA-ligand complexes. Compared with protein-ligand complexes,[28] we have much less experimentally determined structures available for RNA-ligand complexes. Nevertheless, with the development of rigorous physical models, we can realistically expect that with the expansion of the RNA-ligand complex structure family, the accuracy of the predictions can be continuously improved.[29] Moreover, given the limited dataset of available structures, physics-based docking/scoring methods are particularly needed. Despite the significant efforts devoted on the development of physics-based algorithms, limitations of the current approaches remain. For example, Dock6[20] requires expert knowledge to choose the various parameter options for reliable calculations.

Here we report a new RNA-ligand docking (RLDOCK) model. The model has two key novel ingredients. First, we develop a new scoring function based on ligand-RNA energetics,[27] including the van der Waals (VDW) interaction, electrostatic interaction, polar and nonpolar hydration effects, and hydrogen-bond effect. Second, we employ a novel global search algorithm to identify all the potential docking sites for a given ligand. To derive the scoring function, we use a total of 230 RNA-ligand complexes that have been experimentally determined (Table S1 in the Supplemental Information (SI)), among which 30 are the training set for the extraction of the statistical potential and 200 are the test set. Comparisons between other existing models suggest that RLDOCK can give better predictions for ligand docking pose.

2 Model and methods

2.1 RNA and ligand Preparation

We download all the (totally 230) experimentally determined RNA-ligand complex structures from the protein data bank (PDB).[30] For the NMR structures that contain multiple structure models, we select the first model. If an RNA-ligand complex contains multiple types of ligands, we select the ligand that has the largest number of heavy atoms. and ignore the other ligands. For example, the RNA-ligand complex of PDB identifier (ID) 3DIL[31] has five ligand molecules, a lysine with 8 heavy atoms, three fragments of pentaethylene glycol with 3, 3, and 5 heavy atoms, respectively, and an isopropyl alcohol with 4 atoms. We keep only the lysine for the docking prediction. Furthermore, if a ligand has multiple binding sites in the RNA, although the structures of this ligand at various sites may be slightly different due to the flexibility of the ligand, we use the ligand structure at the first binding site (site 1) in the PDB file and delete others. In such cases, if a predicted ligand pose finds the the binding in other sites (e. g. site 2), we use the corresponding experimental ligand pose (i. e., the pose at site 2) as the reference state for the heavy-atom root-mean-square deviation (RMSD) calculation. Here we use the RMSD to assess the accuracy of the theoretical predictions for the ligand poses. The detailed information about the RNA-ligand complexes, such as the experiment method (NMR or Xray) and the ligand name, are listed in Table S2 in the SI.

To calculate the electrostatic interaction energy, we assign partial charges using the "Dock prep" module in Chimera.[32] Specifically, for a given RNA or ligand, we implement the "Dock prep" module for alternate location deletion (keeping the highest occupancy), hydrogen addition, partial charges addition, and output with Mol2 format. For standard RNA residues, we use AMBER ff14SB (with Parm99) to assign charges.[33,34] For the ligand and non-standard residues in RNA, we first calculate the partial charges using ANTECHAMBER with the AM1-BCC (AM1 for short) method.[35,36] If AM1 fails to assign the partial charges, we then use Gasteiger (GAS for short)[37] to calculate the partial charges. Both AM1 (by default) and GAS are included in the "Prep Dock" module. See Table S2 in the SI for a summary of the methods that we use to assign partial charges.

2.2 RLDOCK model and the ligand-RNA scoring function

The procedure above for an RNA-ligand complex results in two Mol2 files for the RNA and the ligand, respectively, that contain the information about the atomic coordinates and partial charges. In our model, the different types of atoms are treated as spheres with specific VDW radii (listed in Table S3 in the SI). We describe a ligand pose using three variables (R, A, O), where R denotes the coordinate of the binding site, i.e., the center of the bound ligand (an effective anchor point for the ligand), A denotes the ligand (heavy) atoms involved at the binding site (in close vicinity of the RNA atoms), and O is the three-dimensional orientation of the ligand. We use Euler rotation angles (α, β, γ) about the anchor point R to represent the ligand orientation O.

To efficiently sample the ligand binding pose for a given RNA structure, we first sample the possible binding sites and the ligand atoms to be placed at the binding site, then generate all the possible ligand orientations about the atom at the binding site. We configure the RNA in a box such that the six boundaries of the box are 3 Å away from the outermost atoms of the RNA, and discretize the space using a simple cubic lattice of grid 0.5 Å. For each grid site R, we place a ligand (heavy) atom A such as the atom i of the ligand at each grid site R (potential binding site). Here $i = 1, 2, ..., N_{\rm L}$ and $N_{\rm L}$ is the number of the heavy atoms of the ligand. For each given pair of the binding site R and bound ligand atom A, we generate the ligand poses by rotating the ligand about the atom A positioned at coordinate R. The 3D rotation is produced by the Euler angles O with 5° increment in each step. In the next step, we score each binding pose (R, A, O) using a new energy-based scoring function described below.

The RLDOCK scoring function is based on the physical interaction energies between RNA and ligand, including the VDW interaction, electrostatic interaction, polar and nonpolar hydration interactions, and hydrogen-bond interaction. We use the generalized Born approximation with solvent-accessible surface area (GB/SA model)[38–43] to treat the polar and nonpolar hydration interactions. In the energy calculation, we neglect the hydrogen atoms in RNA and ligand and add their charges to the directly connected heavy atoms. For a given ligand pose (R, A, O), the energy score is given by:

$$\begin{aligned} S(R, A, O) =& c_{lj} \times \Delta U_{lj} + c_e \times \Delta U_e + c_h \times \Delta U_h + c_{sa} \times \Delta U_{sa} + \\ & c_{pol} \times \Delta U_{pol} + c^{\rm R}_{self} \times \Delta U^{\rm R}_{self} + c^{\rm L}_{self} \times \Delta U^{\rm L}_{self}, \end{aligned} \tag{1}$$

where we use the weight coefficients c[44] to account for the correlation (nonadditivity) effects for the different interactions. In what follows, we illustrate the calculation of each energy term in the scoring function above.

We use the Lennard-Jones (LJ) potential ΔU_{lj} to represent VDW interaction:

$$\Delta U_{lj} = \sum_r \sum_l [(\frac{\sigma_{rl}}{r_{rl}})^{12} - (\frac{\sigma_{rl}}{r_{rl}})^6]. \tag{2}$$

Here the subscripts r and l denote an atom r in the RNA and an atom l in the ligand, respectively, r_{rl} is the distance between the two atoms, and $\sigma_{rl} = 0.8(R_r + R_l)$ is the equilibrium distance, where R_r and R_l are radii of atoms r and l, respectively. We apply a cut-off distance $r_{\rm cut} = 2.5(R_r + R_l)$ in the LJ potential calculation.

We calculate the electrostatic interaction ΔU_e between the ligand and RNA using the following formula:

$$\Delta U_e = \sum_r \sum_l \frac{Z_r Z_l \, e^2}{\varepsilon_c r_{rl}}. \tag{3}$$

Here $Z_r \, e$ and $Z_l \, e$ are the electric charges of the atoms r in RNA and l in ligand, respectively, e is the electronic charge, ε_c (=20 in our calculation) is the dielectric constant of the RNA-ligand complex.

We evaluate the hydrogen-bond interaction energy ΔU_h between the RNA and ligand as:

$$\Delta U_h = \sum_r \sum_l u_h(r_{rl}), \tag{4}$$

where $u_h(r_{rl})$ is the hydrogen-bond energy of an RNA-ligand atom pair. We apply an empirical formula[15] to evaluate the hydrogen-bond energy:

$$u_h(r_{rl}) = \begin{cases} -1 & r_{rl} \le r_{\rm min} \\ -1 + \frac{r_{rl} - r_{\rm min}}{r_{\rm max} - r_{\rm min}} & r_{\rm min} < r_{rl} < r_{\rm max} \\ 0 & r_{rl} \ge r_{\rm max} \end{cases} \tag{5}$$

Here $r_{\rm min} = 0.8(R_r + R_l)$ and $r_{\rm max} = 1.3(R_r + R_l)$.

To account for the change of the hydration energy upon ligand-RNA binding, we consider a hydration layer of width 1.4 Å around the surface[45,46] of the RNA, ligand, and ligand-RNA complex structures. We evaluate the nonpolar hydration energy ΔU_{sa} according to the change in the solvent-accessible surface area (SASA):[47–49]

$$\Delta U_{sa} = \sigma \times \Delta SA \tag{6}$$

where ΔSA is the total SASA change before and after the ligand docking

$$\Delta SA = SA_{\text{complex}} - (SA_{\text{RNA}} + SA_{\text{ligand}}). \tag{7}$$

Here SA_{complex} denotes the SASA of the RNA-ligand complex with the ligand docked to the RNA with pose (R, A, O). SA_{RNA} and SA_{ligand} are the SASA of the RNA alone and ligand alone, respectively (see Fig. S1 in the SI). We choose $\sigma = 0.0054$ kcal/(mol·Å^2) for the empirical atomic solvation parameter σ in Eq. 6.[50]

We decompose the polar hydration energy into three parts:[51] the self-polarization energy changes $\Delta U^{\text{R}}_{self}$ for the RNA and $\Delta U^{\text{L}}_{self}$ for the ligand, and the mutual polarization energy change ΔU_{pol} induced by other atoms:

$$\Delta U_{pol} = U_{pol}^{\text{complex}} - (U_{pol}^{\text{RNA}} + U_{pol}^{\text{ligand}}), \tag{8}$$

where U_{pol}^{complex}, U_{pol}^{RNA}, and U_{pol}^{ligand} are the mutual polarization energy of the complex, the RNA alone, and the ligand alone, respectively. We estimate the three mutual polarization energies from the GB model:[38–43]

$$U_{pol} = \frac{1}{2}(\frac{1}{\varepsilon_w} - \frac{1}{\varepsilon_c}) \sum_{ij} \frac{Z_i Z_j e^2}{\sqrt{r_{ij}^2 + B_i B_j \exp(-\frac{r_{ij}^2}{4B_i B_j})}}, \tag{9}$$

where ε_w (=78) denotes the dielectric constant of water. We assume the same dielectric constant ε_c for the bound and the unbound RNA and ligand. The subscripts i and j ($i \neq j$) represent the atoms i and j of the respective molecule (complex, RNA alone, or ligand alone). r_{ij} denotes the distance between atoms i and j. B_i and B_j are the Born radii of atoms i and j (see Eqs. S1-S4 in the SI). We compute The self-polarization energies $\Delta U^{\text{R}}_{self}$ of the RNA and $\Delta U^{\text{L}}_{self}$ of the ligand as the following:

$$\begin{aligned} \Delta U^{\text{R}}_{self} &= (\frac{1}{\varepsilon_w} - \frac{1}{\varepsilon_c}) \sum_r (\frac{1}{B_r^a} - \frac{1}{B_r^b}) Z_r^2 e^2 \\ \Delta U^{\text{L}}_{self} &= (\frac{1}{\varepsilon_w} - \frac{1}{\varepsilon_c}) \sum_l (\frac{1}{B_l^a} - \frac{1}{B_l^b}) Z_l^2 e^2 \end{aligned} \tag{10}$$

Here $B^b_{r(\text{or } l)}$ and $B^a_{r(\text{or } l)}$ denote the Born radius of atom r (or l) in the RNA (or ligand) before and after the ligand-RNA docking, respectively.

The total number of the ligand poses generated in the sampling algorithm above is equal to (the number of grid points) $\times$ (the number of ligand atoms) $\times$ (the number of orientations). The complete enumeration of all these

possible ligand docking poses and the evaluation of their energy scores are computationally demanding. To enhance the computational efficiency, we develop a novel approach.

2.3 Global search for the binding sites R and the binding atoms A

To enhance the efficiency in the sampling of the ligand poses, we develop the following multi-step screening approach (see the flowchart in Fig. 1).

First, we remove all the grid points R that can cause steric clashes between the ligand and the RNA. We determine the steric clash using a spherical probe of radius $r_{\rm ball} = 2$Å. For a given site $R = (x, y, z)$, if the probe centered at R touches RNA heavy atoms, i.e., the distance between R and an RNA atom is within $r_{\rm ball} + r_{\rm atom}$, where $r_{\rm atom}$ is the radius of the atom, we exclude the grid site R from the sampling space of the binding site (anchor point for the ligand).

Second, we search for all the pocket regions on RNA surface. These regions are potential RNA binding sites. To identify the binding pockets, we move the probe sphere a distance (6 Å) along the positive and negative x, y, and z directions (a total of 6 directions). If this probe sphere in the above movements meets the heavy atom(s) of RNA in at least 5 directions, we keep the site R. Otherwise, the site is not in an RNA pocket and is removed from further sampling steps.

The above two-step screening results in a significant reduction in the available number of grid points R as candidate binding sites; See Fig. 2A for three examples (PDB IDs: 1KOC, 1AKX, and 2BEE) and Table S4 in the SI for the resultant number of candidate binding sites.

After generating the putative binding sites, we sample the different ligand configurations at each binding site. The sampling involves two steps. First, for each candidate binding site R (anchor points) sampled above, we place the different ligand atoms A at the site R. Second, for a given assignment of the ligand atom A, we sample the ligand orientation by rotating the ligand about the atom A (anchored at R). We use Euler angles for the 3D rotation about R to generate an ensemble of ligand orientations. This step results in an ensemble of ligand pose (and RNA-ligand complex structure) described by (R, A, O).

To further sieve the ensemble of ligand binding poses above, we employ a crude, computationally efficient scoring

function. For a given pair (R, A), we calculate the Lennard-Jones (LJ) potential $U_{lj}(R, A, O)$ (see Eq. 2) for each orientation O. Among all the different orientations, we find the one $O^*(R, A)$ with the lowest LJ potential as the "geometrically preferred orientation" $\mathrm{Pose}_{lj}(R, A)$ and the corresponding LJ potential as the geometric compatibility score $\mathrm{Score}_{lj}(R, A)$ for the given binding mode (R, A). Similarly, for a given binding site R, among all the different assignments of the ligand atom A, we find the one $A^*(R)$ with the best (minimum) $\mathrm{Score}_{lj}(R, A)$ as the ligand binding atom to be placed at the given binding site R.

We evaluate the success of the global search algorithm above for the binding site R and the binding atom A using the experimentally determined RNA-ligand complex structure. We rank the candidate binding sites R according to $\mathrm{Score}_{lj}(R, A^*)$ and consider a prediction for the binding site R to be successful if R is the closest to the atom $A^*(R)$ in the experimental structure and the distance is within 0.5Å (grid spacing). As shown in Fig. 2B for the success rate for the 30 RNA-ligand complexes as the training set (see also Table S4 in the SI), we find that the most poorly predicted case is (PDB ID) 3DIX, for which the most native-like ligand binding pose is ranked the 207th in list of the candidate poses. Among the 30 cases, 3DIX has the largest RNA (174-nt) and the ligand has a relatively simple structure with only 10 heavy atoms. Predicting the binding sites for large RNAs and small ligands is challenging because a larger RNA usually has a larger number of potential anchor points and a smaller ligand is less sensitive to the geometric compatibility scores at the different anchor points.

After generating the above ensemble of binding poses (R, A, O) on the basis of the LJ potential, we apply a more refined energy function to score the candidate poses. To reach the optimal balance between the computational efficiency and the accuracy, we keep the top 300 ranked binding sites R and for each R, we keep two binding atoms A^* in the ligand corresponding to the lowest and the second-lowest $\mathrm{Score}_{lj}(R, A^*)$, respectively.

Although the above procedure leads to a drastic reduction in the number of the candidate binding sites and the binding atoms in each site, for a given pair of binding site R and atom $A^*(R)$, there exist a huge number ($> 10^5$) of ligand orientations. Because a slight change in the ligand orientation may incur a notable change in the SASA of the RNA-ligand complex SA_{complex} (see Eq. 7) and the Born radii of the atoms in the complex (see Eqs. S1-S4 in the SI), we need to update the SASA and Born radii for each ligand orientation sampled. SASA and Born radii calculations

involve all the atoms in the RNA-ligand complex, thus, an exhaustive computation for all the possible orientations can be time-consuming (several seconds per orientation). To speed-up the overall computational efficiency, before applying the complete (high-resolution) scoring function (Eqs. 1-10; denoted as *SF-h* for the *high*-resolution scoring fucntion), we first use a computationally efficient, simplified (low-resolution) scoring function (denoted as *SF-l* for the *low*-resolution scoring function) to select the highly probable ligand orientations for a given (R, A).

2.4 Simplified scoring function

In the simplified scoring function (*SF-l*), we employ a fast method to estimate the GB/SA energy terms.[38–43]

1. **Simplified SASA.** When considering the SASA change for a pair of ligand-RNA atoms, we ignore the effect from other atoms (see Fig. S1 in SI), and estimate the SASA change ΔSA using the following approximation:

$$\Delta SA = \sum_r \sum_l -2\pi(R_r^* \times H_r^* + R_l^* \times H_l^*). \tag{11}$$

   Here R_r^* (=$R_r + 1.4$Å) and R_l^* ($R_l + 1.4$Å) denote the radii of the RNA atom r (with a hydration shell) and the ligand atom l (with a hydration shell). H_r^* and H_l^* are the heights of the overlapping spherical crown of the atoms (see Fig. S1 in SI).

2. **Simplified Born radii.** Ignoring the changes in the Born radii for the RNA upon ligand docking, we assume $\Delta U_{self}^{\mathrm{R}} = 0$ in Eq. 10 and neglect the changes in the RNA-RNA mutual polarization interactions $U_{pol}^{\mathrm{complex}}$ and U_{pol}^{RNA} in Eq. 9. For the ligand atoms before docking, we estimate the Born radii using the VDW radii ($B_l^b = r_{\mathrm{atom}}$). For a ligand atoms, we estimate the post-docking Born radii by placing the atom at the binding site R. For example, for a ligand consisting of three types of atoms N, C, and O, we position atoms N, C, and O one by one at the binding site R and calculate their Born radii separately. With the above approximations, we can assume atoms in a ligand bound at a given site with different orientations to have the same Born radii.

Compared to the original full scoring function, the above simplified scoring function (*SF-l*) can lead to thousands-fold reduction in computational time.

2.5 Refined scoring function

We first employ the *SF-l* above to perform a quick screening of ligand orientations in a binding site, then apply the rigorous *SF-h* (Eqs. 1-10) to refine the ranking of the ligand poses. We use the aforementioned 30 RNA-ligand complexes to train the weight coefficients in the scoring function. The training set covers a wide variety of ligands from the totally 230 RNA-ligand structures. In the training process, each coefficient varies from 0.02 to 5.00 with a step 0.02. We determine the optimal values of the weight coefficients by minimizing the heavy-atom RMSD between the predicted ligand pose and the native pose (in the PDB structure). The resultant weight coefficients can give the heavy-atom RMSD within 2 Å for the top-three ranked poses. In detail, the computation for the weight coefficients involves three steps.

First, we use the crude *SF-l* to calculate the seven interaction terms in Eq. 1 for a quick estimation of the coefficients; See Table 1. Because $\Delta U^{\mathrm{R}}_{self} = 0$ in *SF-l*, the corresponding coefficient is not included in the calculation.

Second, we apply the rigorous *SF-h* to refine the predicted weight coefficients. Specifically, we run the calculation for the top-five poses predicted in the first step above for each pair of binding site and atom (R, A). In this step, the coefficients of the VDW interaction, electrostatic interaction, and hydrogen-bond interaction remain the same as those obtained in the first step because these ligand-RNA interactions are the same in the *SF-l* and *SF-h*. The refined results for the coefficients are shown in Table 1. This step leads to the final scoring function.

Third, we start from the top-ranked pose and group the ligand poses into clusters using a heavy-atom RMSD cut-off 2 Å. In each cluster, the pose with the best score is chosen to represent the cluster. This step leads to a new list of ranked poses, each representing a cluster.

3 Results

To focus on the ligand-RNA binding site search algorithm and the ligand pose scoring function, we exclude the effect from the possible ligand and RNA conformational changes in the binding process. Therefore, we use RNA and ligand structures adopted from the experimentally determined native structure of the bound complex and assume the structures are rigid.

The prediction of a ligand-RNA bound structure involves five steps in the RLDOCK model: (1) search for the possible binding sites R using the spherical probe; (2) determine the binding atoms of ligand A through the geometric fit; (3) select the binding orientations O using the crude, simplified scoring function; (4) rank the ligand docking poses using the original, refined scoring function; and (5) generate the final rank list after the cluster calculation. Our test results are shown in Table S2 in the SI.

3.1 Success rate of RLDOCK

We measure the accuracy for the prediction of the ligand binding pose in terms of the heavy-atom RMSD to the ligand pose in the crystal structure. Fig. 3 shows the success rate with the increased RMSD cutoff from (A) 1 Å, (B) 2 Å, to (C) 3 Å for all the 230 RNA-ligand complexes. In Fig. S2 of the SI, we show the success rates for the training set (30 complexes) and the test set (200 complexes), respectively. If we use only the crude *SF-l* (without using the more accurate *SF-h*) in the RLDOCK model, the model can successfully predict 13.0%, 36.1%, and 42.2% of all the cases within RMSD thresholds 1 Å, 2 Å, and 3 Å, respectively. The use of *SF-h* (after the *SF-l*-based quick, initial screening) would increase the fractions to 16.5%, 46.5%, and 51.3%, respectively, for the top-ranked pose.

Because RLDOCK uses the top-ranked pose as the starting point for clustering, the success rates of the *SF-h*-based predictions with and without clustering are the same for the top-ranked pose. However, for the results that include the top-10 poses, the use of clusters can cause a notable increase in the success rate by 7.9% (from 47.8% to 55.7%), 9.6% (from 66.5% to 76.1%), and 10.9% (from 70.4% to 81.3%) for RMSD within 1 Å, 2 Å, and 3 Å, respectively. According to the results shown in Fig. 3, within top-50 ligand docking poses, the RLDOCK model based on *SF-h* (after applying *SF-l*) and the cluster of poses can give successful predictions for 73% (RMSD $<$ 1 Å), 88.2% (RMSD $<$ 2 Å), and 92.2% (RMSD $<$ 3 Å) of all the RNA-ligand complexes.

3.2 Predictions for ligand binding with multiple binding sites

Among all the 230 RNA-ligand complexes, there are 51 cases that contain multiple ligands, each bound to different binding sites (PDB IDs listed in Table S5 in the SI). The novel global search method for all the possible binding sites enables the RLDOCK model to find out the multi-ligand poses in these (51) cases (also see Table S5 in SI).

Our test results show that RLDOCK can successfully predict 45.1% of the 51 cases, where the top ranked pose is an experimental pose within 2 Å RMSD; See Fig. 4A. The success rate is increased to 68.7% and 72.5% if we consider the top-three and top-five poses, respectively. For 5 out of the 51 (9.8%) cases, the top-two ranked poses are the correspond exactly to the two experimental ligand poses. The apparent low success rate (9.8%) is due to our strict criteria that the top-two poses correspond exactly to the experimentally observed two alternative binding poses. In fact, the top few ranked poses often cluster around the same (experimental) pose before the second, distinct, binding pose/site emerges in the ranked list. Indeed, the success rate for the prediction of the multiple binding poses rises rapidly to 19.8% and 43.3% if we include the top-three and top-ten poses, respectively. Moreover, the success rate increases to 80% cases if we include the top-100 ranked poses. These 100 predicted poses form less than ten clusters, demonstrating the reliability of RLDOCK in predicting multiple docking poses.

3.3 Comparisons to other models

Here we compare our RLDOCK model to other models, such as rDOCK,[15] LigandRNA,[17,29] DrugScore$^{\mathrm{RNA}}$,[18,19] Dock6,[20] and MORDOR.[21] In the comparisons, we use a collection of 42 RNA-ligand complexes reported in Ref. 17 as the test set (see Table S6 and S7 in the SI). Among the 42 test cases, 1FJG, 1HNW, 1XPB, and 2OGN (PDB id) are ribosomal complexes. RLDOCK may not provide accurate predictions for these systems because the proteins in complex may influence ligand binding and the large system (> 2000 nts) makes the global scanning for the binding sites computationally infeasible.

Using the top-ranked pose with RMSD ≤ 2Å as the criteria for a successful prediction, out of the 42 test cases, LigandRNA, DrugScore$^{\mathrm{RNA}}$, and DOCK6 can successfully predict 15, 13, and 15 cases, respectively (see Table S6 in SI). The combination of the LigandRNA and DOCK6 can find 20 cases. Our RLDOCK model alone finds 16 out of 38 cases. Furthermore, the original version of rDOCK, the original DrugScore$^{\mathrm{RNA}}$, and MODDOR can predict 5 out of 10, 12 out of 21 cases, and 20 out of 32 cases, respectively. Considering the 38 cases that RLDOCK can treat, the combination of the LigandRNA and DOCK6 is the best method among the five models listed in Table 2 based on the top-ranked pose. Such a combination of models can provide 50% successful prediction for the 38 test cases, while the success rate of RLDOCK alone is 42.1%. However, if we extend the prediction results to top-three poses

(see the details in Table S7 in SI), RLDOCK can have 60.1% (23 out of 38 cases) success rate, which is slightly better than the combination of LigandRNA and DOCK6.

4 Discussions

RLDOCK is a robust physics-based model for ligand-RNA binding. The model has two key ingredients: (a) the global search algorithm for the potential binding sites and (b) the scoring function for ligand-RNA interactions. The model can identify the native ligand pose with a high success rate; See Fig. 5A-C for three selected examples where the top-scored poses have RMSDs less than 2.0 Å. Even for the case that a ligand has two binding sites in the RNA receptor, the RLDOCK model can find both ligand poses/sites, such as the case of PDB ID 2BE0 as shown in Fig. 5D. Here, in order to discriminate the predicted ligand poses in Fig. 5D, we use the native ligand pose in site 1 as the reference ligand pose to calculate the RMSD (same for the case of PDB ID 2BEE in Fig. 6A3). The first and the second top-ranked poses correspond to the native ligand pose in site 2 and site 1 respectively, as shown in Fig. 5E.

However, due to the fact that ligand pose prediction is based on the global search around the given RNA receptor, we find that the first or the first few high-ranked poses could be located in a false binding pocket in the RNA (see the four selected examples in Fig. 6). A close examination of such cases suggests other potentially important factors that need to be considered for ligand-RNA docking.

1. **Metal ion binding.** Metal ions play an essential role in RNA folding and stability.[52,53] The metal ions can be trapped at specific binding sites in RNAs.[27] The bound ions, especially multivalent ions, can exclude ligand binding in the same pocket region. For example, in the case of PDB ID 1NTA, the top-ranked pose is located at the pocket which is actually occupied by two Ba^{2+} ions (yellow balls) in the crystal structure, as shown in Fig. 6A1 and B1. Moreover, other ions such as Na^{+} ions (purple balls) can also influence the ligand docking pose. In the current version of RLDOCK, effects from such specifically bound ions are ignored. In fact, our test calculation shows that if we consider the bound ions (keep them in the RNA mol2 file), the RMSD of the top-ranked pose would be improved and become 1.13 Å.

2. **Competition between the different types of ligands.** In some cases, the crystal structure contains different

types of ligands bound to the RNA. For example, the structure of PDB ID 2EES contains two types of ligands (ACT and HPA), as shown in Fig. 6B2. In RLDOCK, only the ligand with the most heavy atoms (HPA in this example) is used for the prediction (see the subsection 2.1). The experimental structure shows that the RLDOCK-predicted top two poses are located in the RNA pocket occupied by the ACT ligand (Fig. 6-A2 and B2). The result suggests that ACT out-competes HPA in the binding to this pocket.

3. **Correlation between the same type of ligands.** For the cases of a ligand having two or more binding sites, the current RLDOCK model dock ligands one by one as if the different ligands bind to RNA independently. The model neglects the correlation between the different ligand binding events and may cause false predictions. For example, in the case of PDB ID 2BEE as shown in Fig. 6A3 and B3, the crystal structure shows two ligands bound at site 1 and 2, respectively. The top-ranked pose predicted from RLDOCK is located at the location in the middle of site 1 and site 2. However, if we consider the correlation between the two ligands by placing a ligand in site 1 (or site 2), the RLDOCK model can predict the top-ranked pose with RMSD = 0.805 Å at the site 2 (or RMSD = 0.861 Å at the site 1).

4. **Other RNA-ligand interactions neglected in the scoring function.** We find that in some cases, such as 3SUX (see Fig. 6A4 and B4) and 4LVX (see group 3 in Fig. 4B), a few top-ranked poses will be located at the "false" binding pockets. The result may be attributed to RNA-ligand interactions that are neglected in the scoring function, such as the attractive lipophilic interaction and aromatic stacking interaction.[15]

One of the limitations in the current RLDOCK model is the computational efficiency. The bottleneck is the sampling for all the possible binding sites; See Table S4 in the SI for the computer time of the global search for the training set (30 complexes). The computational time is dependent on not only the number of atoms in the RNA receptor and the ligand but also in the candidate binding sites after the detection of the ball probe. For example, for the case of PDB ID 5C45 with 2337 heavy atoms in RNA, the initial screening using the spherical probe for all the available sites in the RNA box results in 2287 candidate binding sites. At each site, we need to calculate the geometric compatibility for all the 27 heavy atoms in the ligand. The whole calculation is time demanding. Therefore, the current version of RLDOCK cannot be used to treat large RNA-ligand system.

5 Conclusions

We here present a newly developed model, RLDOCK, to predict the ligand docking pose in a ligand-RNA complex. In the prediction, the structures and partial charges for the RNA and the ligand are used as the input information, RLDOCK predicts the ligand binding poses in the following steps: (a) To select the candidate binding sites over the entire RNA using a sphere probe; (b) To determine the ligand atom to be placed at the selected binding sites according to the geometric compatibility between the RNA and ligand, (c) To choose the probable ligand orientations through quick computation by applying the simplified scoring function (*SF-l*); (d) To score the (hundreds to thousands) ligand poses using the rigorous scoring function (*SF-h*); (e) To rank the poses with cluster analysis.

Currently the database of the experimentally determined RNA-ligand complex structures is relatively limited compared with the protein-ligand complexes.[28] Therefore, a physics-based approach such as RLDOCK is highly needed. We note that the RLDOCK model developed here has several unique advantages.

1. The novel multi-step sieving algorithm for the global search of the binding sites and poses (section 2.3) can optimize the balance between efficiency and robustness. In particular, the global search method enables the RLDOCK model to predict multiple binding sites/poses.

2. The RLDOCK scoring function consists (see 1). of various RNA-ligand interactions, such as the VDW interaction, electrostatic interaction, polar and nonpolar solvent effects, and the hydrogen-bond interaction.

The model, which is trained using only 30 known complexes, can successfully predict other 200 test complexes (Table 3 and Fig. S2). The result suggests that the parameters are transferable and the model may be robust. In fact, even the "false" predicted poses by the RLDOCK may provide useful information (see the example in Fig. 6A3 and B3).

Future development of the RLDOCK model should address several significant issues such as the low computational efficiency because of the global search for the binding sites, the possible flexibility of ligand and RNA, the effects of metal ions and other ligands, and other possible neglected interactions in the scoring function.

Supplementary informations

Detailed formulas of Born radii (Eq. S1-Eq. S4), 2D-sketch of SASA in *SF-h* and *SF-l* (Fig. S1), success rate for training set and test set (Fig. S2), PDB IDs for the training set and the test set (Table S1), detailed information and prediction results for all the RNA-ligand complexes (Table S2), VDW radii and Born scaling factors for various atoms (Table S3), results of the global search for the training set (Table S4), predicted complexes containing multi-binding poses (Table S5), and the details of comparisons between RLDOCK and other models (Tables S6 and S7).

Acknowledgments

This research was supported by NIH grants R01-GM117059 and R01-GM063732 (to S.-J. C.) and National Natural Science Foundation of China (NSFC) under Grant No. 11704333 (to L.-Z.S.).

References

[1] Batey ,R. T.; Rambo, R. P.; Doudna, J. A.; Tertiary motifs in RNA structure and folding. *Angew. Chem. Int. Ed.* **1999**, *38*, 2326–2343.

[2] Cheng, A. C.; Calabro, V.; Frankel, A. D.; Design of RNA-binding proteins and ligands. *Curr. Opin. Struct. Biol.* **2001**, *11*, 478–484.

[3] Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L.; How many drug targets are there? *Nat. Rev. Drug Discov.* **2006**, *5*, 993–996.

[4] Shortridge, M. D.; Varani, G.; Structure based approaches for targeting non-coding RNAs with small molecules. *Curr. Opin. Struct. Biol.* **2015**, *30*, 79–88.

[5] Chapeville, F.; Lipmann, F.; Ehrenstein, G.; Weisblum, B.; Ray, W. J. Jr; Benzer, S.; On the role of soluble ribonucleic acid in coding for amino acids. *Proc. Natl. Acad. Sci. U.S.A.* **1962**, *48*, 1086–1092.

[6] Sucheck, S. J.; Wong, C. H.; RNA as a target for small molecules. *Curr. Opin. Chem. Biol.* **2000**, *4*, 678–686.

[7] Hermann, T.; Tor, Y.; RNA as a target for small-molecule therapeutics. *Expert Opin. Ther. Pat.* **2005**, *15*, 49–62.

[8] Mironov, A. S.; Gusarov, I.; Rafikov, R.; Lopez, L. E.; Shatalin, K.; Kreneva, R. A.; Perumov, D. A.; Nudler,

E.; Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **2002**, *111*, 747–756.

[9] Mandal, M.; Boese, B.; Barrick, J. E.; Winkler, W. C.; Breaker, R. R.; Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell* **2003**, *113*, 577–586.

[10] Blount, K. F.; Breaker, R. R.; Riboswitches as antibacterial drug targets. *Nat. Biotechnol.* **2006**, *24*, 1558–1564.

[11] Montange, R. K.; Batey, R. T.; Riboswitches: emerging themes in RNA structure and function. *Annu. Rev. Biophys.* **2008**, *37*, 117–133.

[12] Garst, A. D.; Edwards, A. L.; Batey, R. T.; Riboswitches: structures and mechanisms. *Cold Spring Harb Perspect Biol.* **2011**, *3*, a003533.

[13] Youngman, E. M.; Brunelle, J. L.; Kochaniak, A. B.; Green, R.; The active site of the ribosome is composed of two layers of conserved nucleotides with distinct roles in peptide bond formation and peptide release. *Cell* **2004**, *117*, 589–599.

[14] Bannwarth, S.; Gatignol, A.; HIV-1 TAR RNA: the target of molecular interactions between the virus and its host. *Curr. HIV Res.* **2005**, *3*, 61–71.

[15] Morley, S. D.; Afshar, M.; Validation of an empirical RNA-ligand scoring function for fast flexible docking using Ribodock. *J. Comput. Aided Mol. Des.* **2004**, *18*, 189–208.

[16] Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D.; rDock: a fast, versatile and open source code for docking ligands to proteins and nucleic acids. *PLoS Comput. Biol.* **2014**, *10*, e1003571.

[17] Philips, A.; Milanowska, K.; Łach, G.; Bujnicki, J. M.; LigandRNA: computational predictor of RNA-ligand interactions. *RNA* **2013**, *19*, 1605–1616.

[18] Pfeffer, P.; Gohlke, H.; DrugScore$^{\mathrm{RNA}}$ –knowledge-based scoring function to predict RNA-ligand interactions. *J. Chem. Inf. Model.* **2007**, *47*, 1868–1876.

[19] Krüger, D. M.; Bergs, J.; Kazemi, S.; Gohlke, H.; Target flexibility in RNA-ligand docking modeled by elastic potential grids. *ACS Med. Chem. Lett.* **2011**, *2*, 489–493.

[20] Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D.; DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* **2009**, *15*, 1219–1230.

[21] Guilbert, C.; James, T. L.; Docking to RNA via root-mean-square-deviation-driven energy minimization with flexible ligands and flexible targets. *J. Chem. Inf. Model* **2008**, *48*, 1257–1268.

[22] MacKerell, A. D.; Banavali, N.; Foloppe, N.; Development and current status of the charmm force field for nucleic acids. *Biopolymers* **2000**, *56*, 257–265.

[23] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A.; Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[24] Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.; The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

[25] Moitessier, N.; Westhof, E.; Hanessian, S.; Docking of aminoglycosides to hydrated and flexible RNA. *J. Med. Chem.* **2006**, *49*, 1023–1033.

[26] Daldrop, P.; Reyes, F. E.; Robinson, D. A.; Hammond, C. M.; Lilley, D. M.; Batey, R. T.; Brenk, R.; Novel ligands for a purine riboswitch discovered by RNA-ligand docking. *Chem. Biol.* **2011**, *18*, 324–335.

[27] Sun, L. Z.; Zhang, D.; Chen, S. J.; Theory and modeling of RNA structure and interactions with metal ions and small molecules. *Annu. Rev. Biophys.* **2017**, *46*, 227–246.

[28] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K.; BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

[29] Philips, A.; Łach, G.; Bujnicki, J. M.; Computational methods for prediction of RNA interactions with metal ions and small organic ligands. *Methods Enzymol.* **2015**, *553*, 261–285.

[30] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E.; The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

[31] Serganov, A.; Huang, L.; Patel, D. J.; Structural insights into amino acid binding and gene control by a lysine riboswitch. *Nature* **2008**, *455*, 1263–1267

[32] Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E.; UCSF Chimera- visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

[33] Cornell, W. D.; Cieplak, P.; Baily, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. C. Jr.; Fox, T.; Caldwell, J. W.; Kollman, P. A.; A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[34] Cheatham, T. E. III; Cieplak, P.; Kollman, P. A.; A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.

[35] Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I.; Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[36] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A.; Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model* **2006**, *25*, 247–260.

[37] Gasteiger, J.; Marsili, M.; Iterative partial equalization of orbital electronegativity rapid access to atomic charges. *Tetrahedron*, **1980**, *36*, 3219–3288.

[38] Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T.; Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.

[39] Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G.; Pairwise solute descreening of solute charges from a dielectric medium. *Chem. Phys. Lett.* **1995**, *246*, 122–129.

[40] Zou, X.; Sun, Y.; Kuntz, I. D.; Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. *J. Am. Chem. Soc.* **1999**, *121*, 8033–8043.

[41] Nymeyer, H.; Garcia, A. E.; Simulation of the folding equilibriumof a helical peptides: a comparison of the generalized Born approximation with explicit solvent. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13934–13939.

[42] Liu, H.-Y.; Kuntz, I. D.; Zou, X.; Pairwise GB/SA scoring function for structure-based drug design. *J. Phys. Chem. B* **2004**, *108*, 5453–5462.

[43] Liu, H. -Y.; Zou, X.; Electrostatics of ligand binding: parameterization of the generalized Born model and comparison with the Poisson-Boltzmann approach. *J. Phys. Chem. B* **2006**, *110*, 9304–9313.

[44] Kang, X.; Shafer, R. H.; Kuntz, I. D.; Calculation of ligand-nucleic acid binding free energies with the generalized-born model in DOCK. *Biopolymers* **2004**, *73*, 192–204.

[45] Lee, B.; Richards, F. M.; The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379–400

[46] Durham, E.; Dorr, B.; Woetzel, N.; Staritzbichler, R.; Meiler, J.; Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J. Mol. Model.* **2009**, *15*, 1093–1108.

[47] Simonson, T.; Brunger, A. T.; Solvation free energies estimated from macroscopic continuum theory: an accuracy assessment. *J. Phys. Chem.* **1994**, *98*, 4683–4694.

[48] Vallone, B.; Miele, A.; Vecchini, P.; Chiancone, E.; Brunori, M.; Free energy of burying hydrophobic residues in the interface between protein subunit. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 6103–6107.

[49] Raschke, T. M.; Tsai, J.; Levitt, M.; Quantification of the hydrophobic interaction by simulations of the aggregation of small hydrophobic solutes in water. *Proc. Natl. Acad. Sci. U.S.A.* **2001** *98*, 5965–5969.

[50] Treesuwan, W.; Wittayanarakul, K.; Anthony, N. G.; Huchet, G.; Alniss, H.; Hannongbua, S.; Khalaf, A. I.; Suckling, C. J.; Parkinson, J. A.; Mackay, S. P.; A detailed binding free energy study of 2:1 ligand-DNA complex formation by experiment and simulation. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10682–10693.

[51] Sun, L. Z.; Chen, S. J.; Monte carlo tightly bound ion model: predicting ion binding properties of RNA with ion correlations and fluctuations. *J. Chem. Theory Comput.* **2016**, *12*, 3370–3381.

[52] Brion, P.; Westhof, E.; Hierarchy and dynamics of RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **1997**, *26*, 113–137.

[53] Tinoco, I. Jr; Bustamante, C.; How RNA folds. *J. Mol. Biol.* **1999** *293*, 271–281.

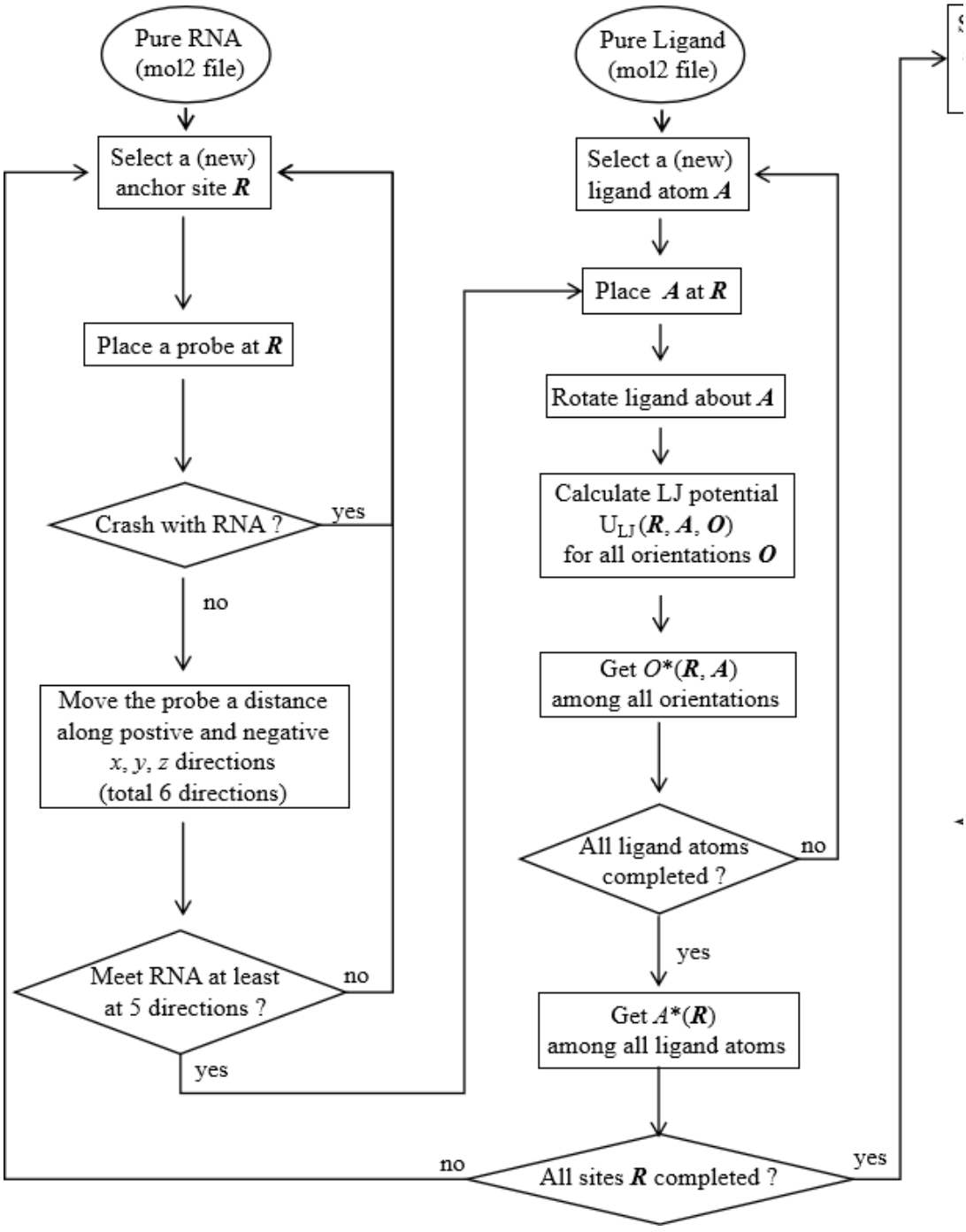

Figure 1: The flowchart of the RLDOCK model.

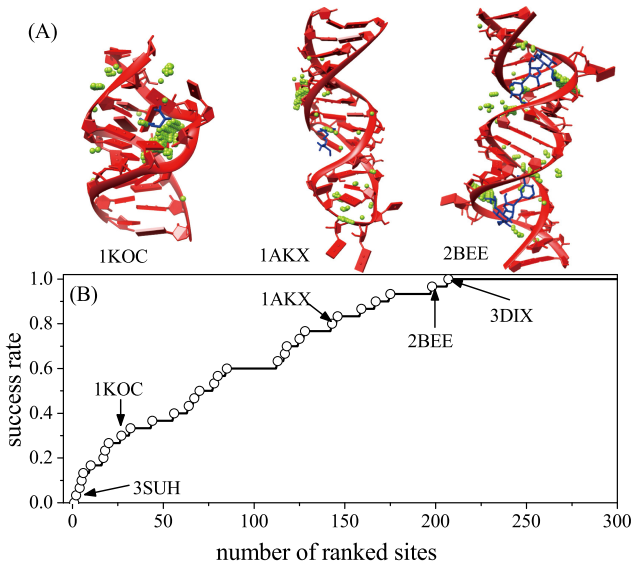

Figure 2: (A) The selected binding sites after the screening using the spherical probe. The RNAs and ligands are marked with red and blue colors, respectively. The green points represent the candidate binding sites. (B) The success rate of the global search for the binding sites as a function of the number of the ranked sites included for the 30 RNA-ligand cases in training set.

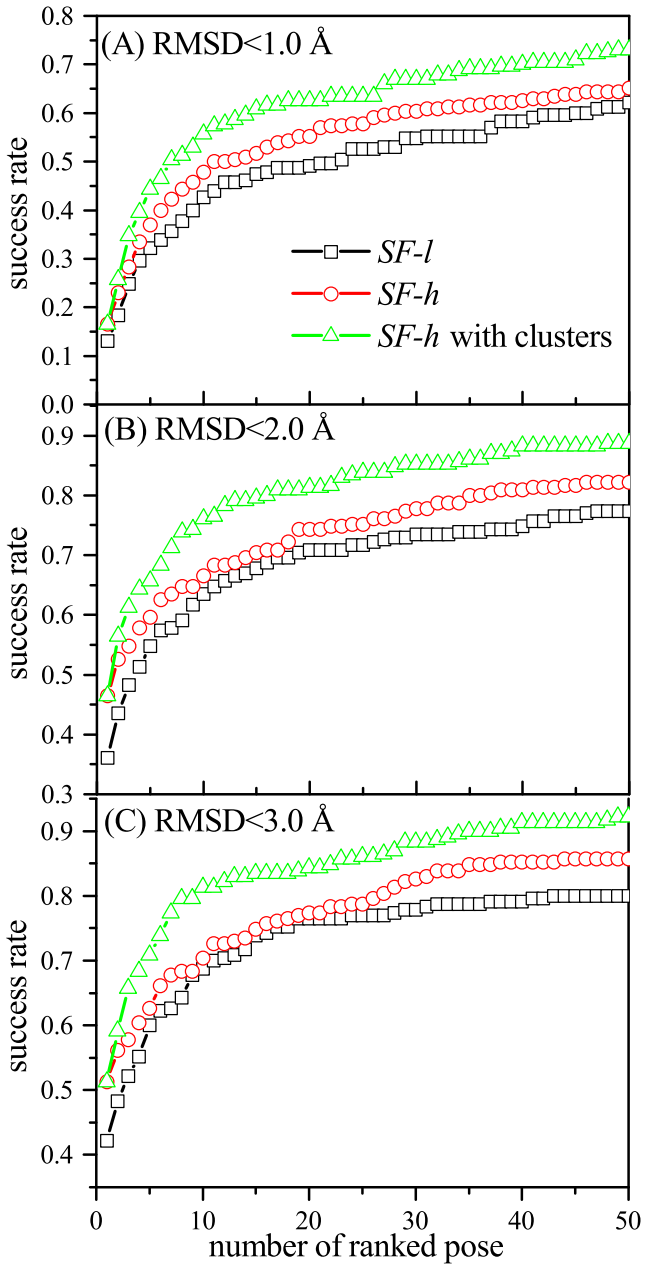

Figure 3: The success rate as a function of the number of the (top) ranked poses with RMSD within 1 Å (A), 2 Å (B), and 3 Å (C) for all the 230 RNA-ligand binding cases. In the evaluation for the success rate for RMSD within 1 Å, we use the cut-off RMSD 1Å to classify the different poses into clusters.

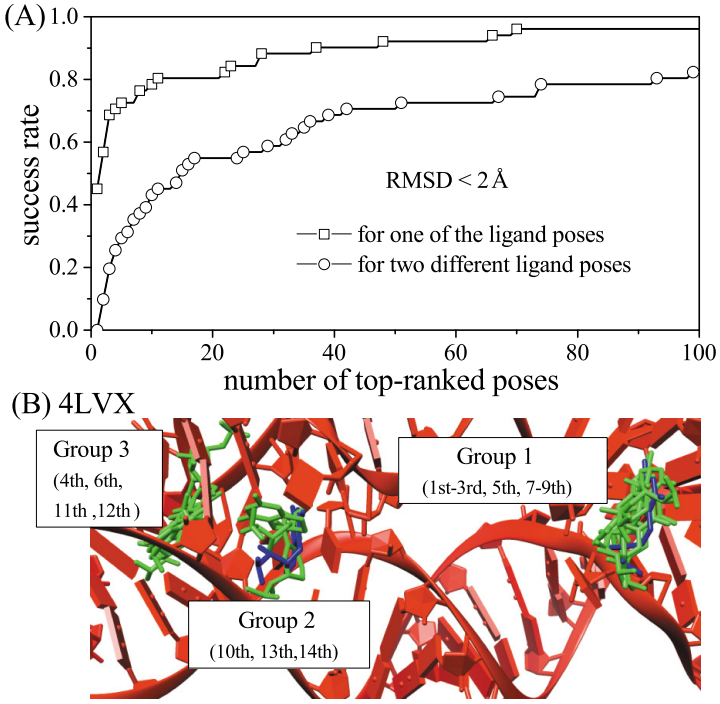

Figure 4: (A) The success rates as a function of the number of the (top) ranked poses (for one or two of the experimental ligand poses). Here a predicted pose with RMSD $\leq$ 2Å is considered to be successful. (B) The structure of RNA-ligand complex (PDB ID: 4LVX). The experimentally determined ligand pose and the predicted ligand pose are depicted in blue and green, respectively. Group 1 includes the 1st-3rd, 5th, and 7-9th ranked poses, group 2 includes the 10th, 13th, and 14th ranked poses, and group 3 includes the 4th, 6th, 11th, and 12th ranked poses.

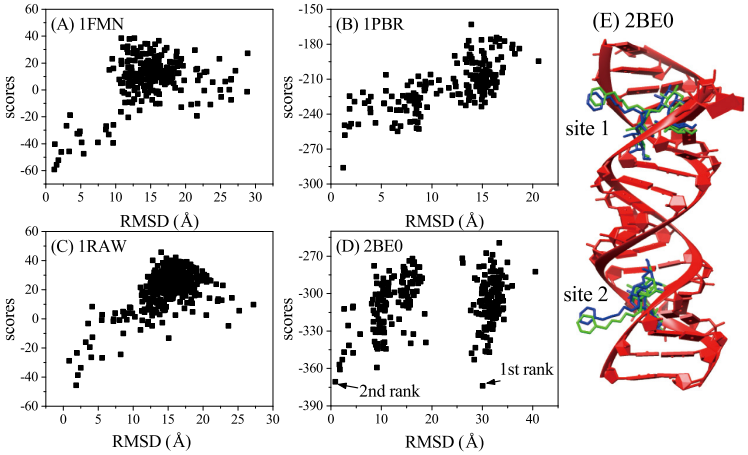

Figure 5: (A)-(D) Score-RMSD relationship for the four selected RNA-ligand complexes with PDB IDs 1FMN (A), 1PBR (B), 1RAW (C), and 2BE0 (D), respectively. (E) The structure of RNA-ligand complex with PDB ID 2BE0. The experimentally determined ligand pose and the predicted ligand pose are depicted in blue and green colors, respectively.

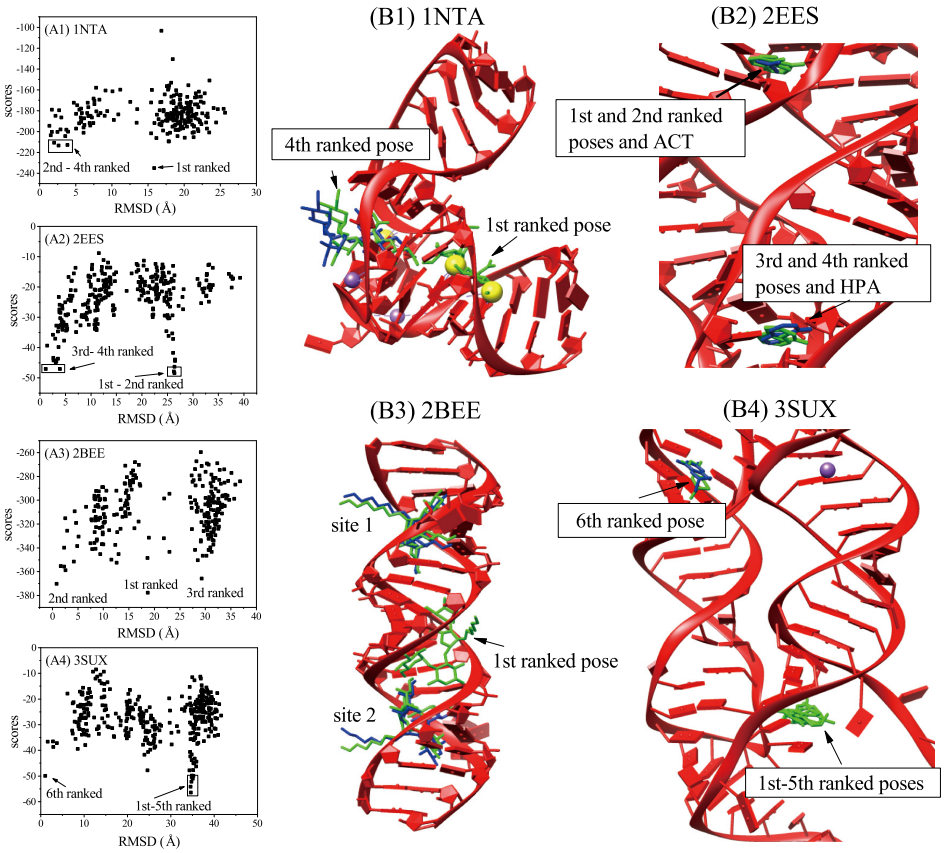

Figure 6: (A1)-(A4) Score-RMSD relationship for the four selected RNA-ligand complexes with PDB IDs 1NTA (A1), 2EES (A2), 2BEE (A3), and 3SUX (A4), respectively. (B1)-(B4) The structures of RNA-ligand complexes corresponding to the four cases in (A1)-(A4). The experimental ligand pose and the predicted ligand pose are depicted in blue and green, respectively. Yellow and purple balls represent the Ba^{2+} and Na^{+} ions observed in the crystal structures.

Table 1: The weight coefficients in the simplified (*SF-l*) and the rigorous (*SF-h*) scoring functions, respectively

coefficients	c_{lj}	c_e	c_h	c_{sa}	c_{pol}	c_{self}^{R}	c_{self}^{L}
SF-l	3.30	1.32	0.12	0.24	0.76	—	0.14
SF-h	3.30	1.32	0.12	0.92	0.78	4.96	2.52

Table 2: The success rate of various docking models[a]

	LigandRNA[17]	Drugscore[RNA 18,19]	DOCK6[20]	ligandRNA +DOCK6[17]	RLDOCK
top 1[b]	39.5%(35.7%)	28.9%(31.0%)	36.8%(35.7%)	**50%**(47.6%)	42.1%(46.5%)
top 3[c]	47.4%(45.2%)	39.5%(42.9%)	44.7%(42.9%)	57.9%(47.6%)	**60.1%** (61.3%)

[a] The precentage without parenthesis is calulated for the 38 RNA-ligand cases in Ref. (17) tested with the RLDOCK model. The value with parenthesis is calculated for the 42 test cases in Ref. (17) and the 230 complexes in the present study. [b] The success rate of the top-ranked poses. [c] The success rate (measured by the best RMSD) of the top-three poses. The best results are indicated in bold.

Table 3: Comparisons of the success rate between the training set and test set

	training set			test set		
RMSD	top 1	top 3	top 10	top 1	top 3	top 10
< 1Å	20.0%	36.7%	50.0%	16.0%	34.5%	56.0%
< 2Å	46.7%	66.7%	86.7%	46.5%	60.5%	74.5%
< 3Å	53.3%	66.7%	86.7%	51.0%	65.5%	80.0%