BCSJ Bulletin of the Chemical Society of Japan

A Structural Refinement Technique for Protein-RNA Complexes Using Combination of AI-based Modeling and Flexible Docking: A Study of Musashi-1 Protein

Nitchakan Darai, Kowit Hengphasatporn, Peter Wolschann, Michael T. Wolfinger, Yasuteru Shigeta, Thanyada Rungrotmongkol,* and Ryuhei Harada*

Advance Publication on the web June 9, 2023
doi:10.1246/bcsj.20230092

A Structural Refinement Technique for Protein-RNA Complexes Using Combination of AI-based Modeling and Flexible Docking: A Study of Musashi-1 Protein

Nitchakan Darai,[1‡] Kowit Hengphasatporn,[2‡] Peter Wolschann,[3] Michael T. Wolfinger,[3,4,5] Yasuteru Shigeta,[2] Thanyada Rungrotmongkol,[1,6*] Ryuhei Harada[2*]

‡ Equal contribution.

[1]Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok, 10330, Thailand.

[2]Center for Computational Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan.

[3]Department of Theoretical Chemistry, University of Vienna, Währinger Strasse 17, Vienna, 1090, Austria.

[4]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währinger Strasse 29, Vienna, 1090, Austria.

[5]RNA Forecast, 1100 Vienna, Austria.

[6]Center of Excellence in Biocatalyst and Sustainable Biotechnology, Faculty of Science, Chulalongkorn University, Bangkok, 10330, Thailand.

E-mail: Thanyada.r@chula.ac.th, ryuhei@ccs.tsukuba.ac.jp

Nitchakan Darai received her Ph.D. degree in 2022 from Program in Bioinformatics and Computational Biology, Chulalongkorn University. Currently, she is a postdoctoral research fellow at Chulalongkorn University. Her research interests are in the Molecular modeling of biological systems.

Dr. Thanyada Rungrotmongkol obtained her Ph.D. degree from Kasetsart University, Thailand, in 2006. She was recently promoted to the position of Associate Professor in Theoretical Chemistry at Chulalongkorn University in 2020. Her research interests involve the use of computational simulations to gain atomic-level insights into the molecular recognition, structural, and dynamic properties of proteins in biological processes.

Abstract

An efficient structural refinement technique for protein-RNA complexes is proposed based on a combination of AI-based modeling and flexible docking. Specifically, an enhanced sampling method called parallel cascade selection molecular dynamics (PaCS-MD) was extended to include flexible docking to construct protein-RNA complexes from those obtained by AI-based modeling (AlphaFold2). With the present technique, the conformational sampling of flexible RNA regions is accelerated by PaCS-MD, enabling one to construct plausible models for protein-RNA complexes. For demonstration, PaCS-MD constructed several protein-RNA complexes of the RNA-binding Musashi-1 (MSI1) family of proteins, which were validated by comparing a group of crucial residues for RNA-binding with experimental complexes. Our analyses suggest that PaCS-MD improves the quality of complex modeling compared to the standard protocol based on template-based modeling (Phyre2). Furthermore, PaCS-MD could also be a beneficial technique for constructing complexes of non-native RNA-binding to proteins.

Keywords: PaCS-MD, Musashi proteins, RNA-protein complex construction

1. Introduction

RNA-binding proteins (RBPs) are generally recognized as proteins that bind to RNA via one or more globular RNA-binding domains (RBDs).[1] In all eukaryotes, RBPs are essential for modifying the bound RNA function, such as RNA capping, RNA editing, or the fate of the bound RNA.[2] Additionally, RBPs have critical roles in the regulation of mRNA translation, mRNA transport, and splicing control in post-transcriptional processes.[2,3] As a representative RBP, Musashi is a member of the RBP family, which comprises two orthologs in humans, known as Musashi-1 (MSI1) and Musashi-2 (MSI2).[4] Specifically, MSI1 is a translational regulator that promotes stem cell maintenance and self-renewal and is linked to the enhancement of ZIKV replication.[5] In contrast, MSI2 controls the mRNA translation of many intracellular targets and affects a variety of biological functions,[6] including the preservation of stem cell identity,[7] stem cell self-renewal, and cancer growth.[8] In our previous study,[9] molecular dynamics (MD) simulations and binding free-energy calculations of the MSI1 RNA-binding protein were performed to characterize the interaction energies of two individual systems: RNA-binding domain 1 (RBD1) and RNA-binding domain 2 (RBD2) because experimental data on

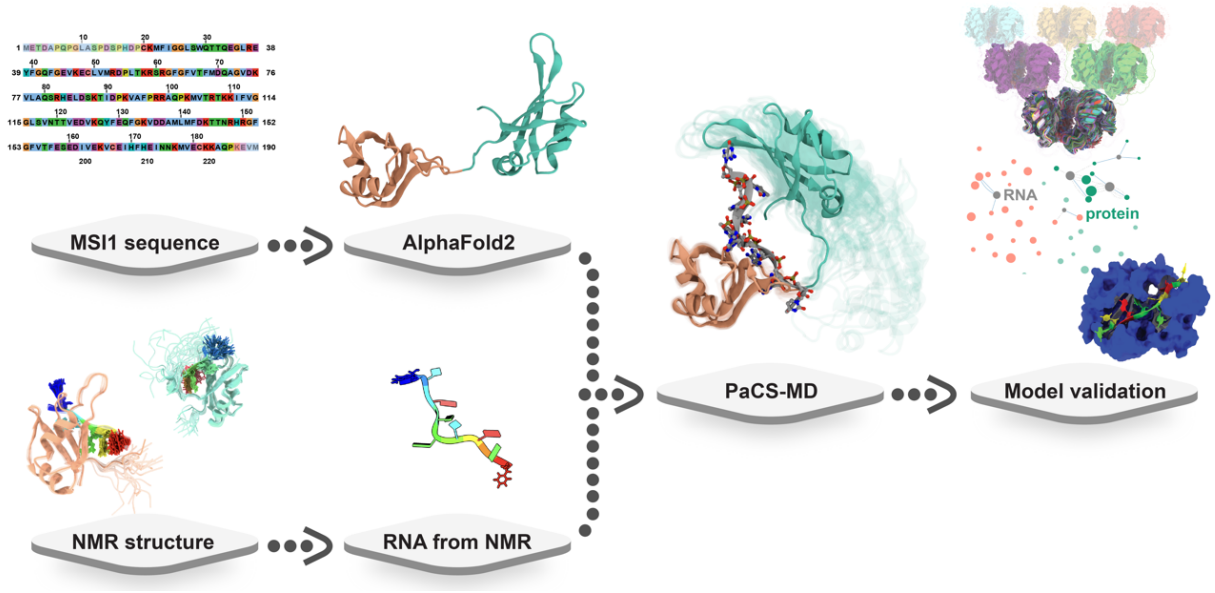

Figure 1. Workflow of an efficient structural refinement technique for protein-RNA complexes based on a combination of AI-based modeling (AlphaFold2) and flexible docking (PaCS-MD). Initially, AI-based modeling of protein-RNA structures is conducted with AlphaFold2. Subsequently, experimental data provides RNA/protein recognition site information, which is incorporated into the model. PaCS-MD is then employed to generate a variety of protein-RNA complexes based on flexible docking. Finally, the structural validation of protein-RNA complexes is performed through MD simulation, end-point binding free energy calculation, and interaction network analysis described by the orange and blue circles in the top right corner of the workflow.

the complexes of MSI1-RBD1/RBD2 is not available.

Our preceding results showed that RBD1 and RBD2 bind to the canonical RNA motif GUAGU better than GGAGU based on their binding affinities.[9] The RNA-binding is thought to occur through the numerous RBDs, which can create an expanded RNA-binding interface, allowing them to recognize cognate RNA sequences of various targets and lengths.[10] To gain structural insight into MSI1-RBD1/RBD2, the full-coverage protein structure in complex with RNA should be constructed when performing MD simulations. Here, a standard protocol to generate an RBP complex is to connect a set of partial structures and subsequently overlay the available RNA template to construct the complex.[11-13] However, as a limitation, the standard protocol is unsuitable for a system that interacts with a non-canonical RNA binding motif that is not available in the Protein Data Bank (PDB).

To overcome this issue, we propose an efficient structural refinement technique that searches for plausible protein-RNA complexes based on a combination of AI-based modeling and flexible docking. To perform the flexible docking, we focused on our conformational sampling methods.[14-15] Specifically, the parallel cascade selection molecular dynamics (PaCS-MD)[16,17] was extended to sample the RNA-binding process. PaCS-MD was originally designed to sample the conformational transitions of proteins induced in timescales inaccessible to traditional MD. In our previous studies, PaCS-MD was applied to several proteins and successfully sampled their large-amplitude domain motions.[18-22] To efficiently promote a conformational transition, PaCS-MD repeats multiple cycles of MD simulations from important protein structures with a higher potential to make transitions. In analogy to the conformational transitions of proteins, the RNA-binding process is regarded as having long timescale dynamics and is difficult to sample with traditional MD. Therefore, the conformational sampling of flexible RNA regions is accelerated by PaCS-MD, enabling one to construct models for plausible protein-RNA complexes based on flexible docking.

In this study, PaCS-MD was extended to include flexible docking to refine RNA-protein complexes after being predicted by AI-based structural modeling. Figure 1 shows an overview of protein-RNA modeling using MSI1-RBDs with RNAs. First, the AI-based structural modeling program called AlphaFold2 (AF2) predicted a model structure as a template for an entire MSI1-RBD1/RBD2. Next, the RNA 3D structure was placed in the MSI1-RBD1 using the information from the NMR structure. Then, the other side of RNA was inserted into MSI1-RBD2 using PaCS-MD, which is referred to as an efficient structural refinement of protein-RNA complexes predicted by AF2. In PaCS-MD, important conformations with a higher potential to form plausible protein-RNA complexes were selected from MD snapshots of multiple MD simulations using a measure. By repeating multiple MD simulations, PaCS-MD gradually sampled the RNA-binding process.

The benefit of the combination of AI-based modeling (AF2) and flexible docking (PaCS-MD) is as follows. Generally, AF2 fails to predict flexible regions that are undetermined by experimental studies owing to their large structural fluctuations, although AF2 predicts rigid regions such as RBDs with high accuracy. Therefore, PaCS-MD enables one to efficiently sample conformations of linker regions of RBDs predicted by AF2. Thus, PaCS-MD structurally refines flexible regions of protein-RNA complexes. In summary, the rigid regions of RBD bound with the RNA are predicted by AF2, and subsequentially PaCS-MD refines the remaining regions based on flexible docking.

Finally, as validation, protein-RNA complexes constructed by PaCS-MD were compared with those predicted by a standard protocol (homology modeling by Phyre2). Furthermore, we elucidated a group of crucial residues for maintaining protein-RNA interactions between RBDs and RNA by analyzing the trajectories of PaCS-MD. In conclusion, the combination of AI-based modeling (AF2) and flexible docking (PaCS-MD) could be a structural refinement technique for constructing plausible RNA-protein complexes.

2. Computational Details

2.1. Complexes of MSI1-RBDs bound with the RNA constructed by PaCS-MD

To sample MSI1-RNA complexes using PaCS-MD, a set of end-point structures (reactant and product) of the binding process was constructed, that is, a starting conformation (reactant before RNA binding) and a template conformation (product after RNA binding) of PaCS-MD. First, we constructed a reactant of PaCS-MD. In a previous NMR study, a set of the Musashi RNA binding domain 1 (MSI1-RBD1) and 2 (MSI1-RBD2) structures were deposited in PDB with PDB ID: 2RS2[23] and 5X3Z,[10] respectively. Because the entire RBD1-RBD2 structures of RNA binding to MSI1 are still lacking, we constructed the whole conformations as complexes. To build a complex, we used an amino-acid sequence of human RBD1-RBD2 from the UniProt database (UniProt accession: O43347). Indeed, the amino acids within both RBDs from residue numbers 20 to 186 were considered. Then, a set of complexes of RBD1-RBD2 was constructed based on the amino-acid sequence in humans. In the present study, AF2 developed by DeepMind[24] was adopted to predict the whole structure of RBD1-RBD2 of MSI1 using the ColabFold notebook.[25]

First, we constructed a reactant of PaCS-MD based on the structure of RBD1-RBD2 predicted by AF2. This structure of RBD1-RBD2 was superimposed with the NMR structure of RBD1 in the RNA-bound state to maintain the orientation of the RNA around RBD1. We then regarded the RBD1-RBD2-RNA complex as a reactant of PaCS-MD. Second, we constructed a product of PaCS-MD using Phyre2.[26] Practically, Phyre2 allows users to predict and analyze the changes in structures, functions, and mutations of proteins. In Phyre2, similar energy functions and template-based modeling are used to determine protein structures. In the present study, Phyre2 was used to construct a 3D model based on the human MSI1 protein sequence. For the options of Phyre2, we selected an intensive mode to combine multiple template modeling with a simplified ab initio folding simulation to produce an entire full-length model of the human MSI1 protein sequence. Sequentially, the two NMR structures of RBD1 and RBD2 were superimposed with the model of Phyre2 to preserve their orientation with the RNAs (GUAG and UAG for RBD1 and RBD2, respectively). Discovery Studio Visualizer, 2020[27] was employed to remove part of the duplicated RNA and connect the RNAs from the NMR structures. Then, the ligated RNA bound by RBD1 and RBD2 was regarded as the product of PaCS-MD. Finally, the MSI1-RBDs structures with doublecortin mRNA were constructed using PaCS-MD because the mRNA has been found to bind to a specific site of MSI1 (5'-GUAGGUAGU-3'). Finally, we used the tLEaP module of AMBER20 to construct a set of MD parameters for MSI1-RBDs containing the RNAs.[28]

2.2. PaCS-MD

PaCS-MD was used to construct protein-RNA complexes consisting of RBD1 and RBD2 bound with the target RNA from the unbound (reactant) to the bound (product) states. In PaCS-MD, as a measure, we partially specified the all-atom RMSD for RBD2 bound with the RNA predicted by Phyre2 because the reactant of PaCS-MD was already superimposed with RBD1 of the NMR structure. Therefore, PaCS-MD specified the partial RMSD for RBD2 bound with the RNA to select initial structures for short-time MD simulations in each cycle. At the begging of the current (*i*th) cycle, the MD snapshots of the short-time MD simulations in the previous (*i*-1th) cycle were ranked according to their RMSD values. Some of the MD snapshots with smaller RMSD values were highly ranked and regarded as important conformations with a high potential to bind to the RNA. Then, the important conformations were selected as the initial structures of the next (*i*+1th) cycle to restart the short-time MD simulations by resetting their initial velocities based on the Maxwell–Boltzmann distribution. Repeating a cycle of resampling from important conformations encourages frequent RNA binding to RBDs. In PaCS-MD, a variety of initial structures and velocities used in restarting the short-time MD simulations act as a perturbation, which intensively promotes RNA-binding to form protein-RNA complexes. Finally, PaCS-MD is used to construct RNA-protein complexes, enabling one to predict experimentally undetermined (highly fluctuated) regions.

To generate statistically reliable protein-RNA complexes, we independently performed five PaCS-MD trials by changing their initial conditions. In each cycle, the top-ranked MD snapshot was identified by the partial RMSD for RBD2 and specified as an initial structure of a short-time (100-ps) MD simulation in the next cycle, where MD snapshots were recorded every 1 ps for 100 ps each. Finally, we performed PaCS-MD until the 200th cycle, that is, the total simulation time was 20 ns per trial (one initial structure × 100-ps MD simulation × 200 cycles).

2.3. Interaction-based clustering of complexes

The trajectories of PaCS-MD were used to elucidate the intermolecular interactions between RBDs and the target RNA. The residue–residue interaction networks of the protein-RNA complexes were generated using the Residue Interaction Network Generator (RING) web server.[29] To generate the networks, we periodically selected 100 MD snapshots from the trajectories of the last 100 cycles of each PaCS-MD trial (5 trials, 500 MD snapshots). Finally, the RING web server generated the protein-RNA interactions using a set of default parameters: sequence separation: 3, nodes: closest, edges: multiple, hydrogen bond: 3.5 Å, van der Waals: 0.5 Å, ionic or salt bridge: 4 Å, π–π stacking: 6.5 Å, π-cation: 5 Å, and disulfide bond: 2.5 Å. Through interaction-based clustering, we selected a set of representative protein-RNA complexes from the MD snapshots generated by each PaCS-MD trial.

2.4. MD simulations from a set of representative complexes

After constructing the representative protein-RNA complexes, MD simulations were performed from them. First, the protonation states of the complexes generated by both protocols were determined using the PDB2PQR server[30] at pH 7.4. The AMBER ff14SB[31] and chiOL3 (OL3)[31] force fields were employed for protein and RNA, respectively. The LEaP module[32] added the missing hydrogen atoms of each system in accordance with the conventional approach. The additional hydrogen atoms were reduced for 1,000 steps using the steepest descent (SD) method and then for 3,000 steps using the conjugated gradient (CG) method. The TIP3P water molecules[33] were randomly arranged around each solute, amounting to approximately 12,436 atoms. Each system was solvated using a distance of 12 Å from each solute, producing each periodic box of both systems with dimensions 60 × 43 × 51 $Å^3$ (PaCS-MD) and 55 × 46 × 43 $Å^3$ (Phyre2). To neutralize the solvated systems, five Na^+ counter ions were added to each system. A time step of 2 fs for integrating Newton's equation of motion and a periodic boundary condition with an isothermal-isobaric (*NPT*) ensemble were used. The ions and water molecules were then minimized using the SD method for 1,000 steps, followed by the CG method for 3,000 steps. In the final step, a similar technique completely minimized each system. The SHAKE algorithm[34] was used to restrict all the bonds containing hydrogen atoms. The AMBER20 software[31] was used to perform the MD simulations. Each

system was gradually heated from 100 to 310 K during the first stages of the MD simulations. Subsequently, each system was equilibrated at a 310 K constant temperature. For each representative complex, a set of 300-ns MD simulations was conducted from five replicas under the *NPT* condition (1 atm and 310 K) to obtain statistically reliable trajectories. As a reference, MD simulations were also independently performed from five replicas of the complex predicted by Phyre2. In the present study, "replicas" refer to MD simulations from an identical structure with different initial velocities re-generated based on the Maxwell–Boltzmann distribution.[35]

2.5 Trajectory analyses

To analyze each protein-RNA complex structurally and energetically, only the final 100-ns trajectory of each 300-ns MD simulation was considered. We used two energetic analyses to evaluate the binding affinity of MSI1-RBDs with the RNA. First, the solvated interaction energy (SIE) was calculated using sietraj software.[36-38] Then, the generalized Born (GB) surface area (MM/GBSA) with the per-residue decomposition energies was calculated by the MM-PBSA.py module.[28] Based on the MM/GBSA method, the binding free energy of each complex was calculated using the 100 MD snapshots taken from the last 100-ns trajectory of each 300-ns MD simulation.[39,40] Moreover, the LigandScout 4.4 program[41-43] was used to determine pharmacophore models having 100% occurrence to elucidate the types of intermolecular interactions between RBDs and RNA.

3. Results and Discussion

3.1. Protein-RNA complexes constructed by PaCS-MD

First, we validated the structural quality of the AF2 models to specify a reactant of PaCS-MD. For the structural validation, we used the number of sequences per position and the per-residue confidence measure (pLDDT). Based on the pLDDT value, we evaluated the structural reliability of the top five AF2 models, for which their pLDDT values dropped to 40 or 50–70 in some residues due to the high flexibility of the position connection between RBD1 and RBD2 (Figure S1B). The top five AF2 models were superimposed **(**Figure S1C). From the superposition, some structural regions of the protein did not superimpose because of their different linking orientations. Finally, we specified the rank 1 model (Figure S1D) as a reactant for PaCS-MD simulation.

Next, we constructed protein-RNA complexes (MSI1-RBDs bound with the RNA) based on PaCS-MD. The rank 1 model was superimposed with the NMR structure of RBD1 (PDB ID: 2RS2) to preserve the orientation of the RNA in RBD1. Sequentially, to model the RNA to bind with RBD2 in a different orientation, we then launched PaCS-MD using the superimposed structure. Figure S2A shows a set of partial RMSD profiles for the five PaCS-MD trials (R1–R5), indicating that the protein-RNA complexes of each trial structurally converged after the 100th cycle. We confirmed how the RNA-protein complexes constructed by PaCS-MD was converged until the 200th cycle by referring to the partial RMSD values, which measured by comparing all structures obtained from PaCS-MD to the RNA-RBD2 domain of Phyre2 structure. Therefore, we selected 100 RBDs-RNA complexes from the last 100 cycles of each trial and regarded them as representative complexes of MSI1-RBDs bound with the RNA (Figure S2B–F). From the representative complexes, PaCS-MD constructed a variety of RBDs-RNA binding poses in all five trials. Next, we quantitatively evaluated which binding pose had plausible intermolecular interactions between RBDs and the RNA.

3.2. Protein-RNA intermolecular interaction analyses

After constructing the representative binding poses, their protein-RNA intermolecular interactions were quantitatively evaluated by several quantities. First, the intermolecular interactions were addressed using an interaction-based clustering method using the RING 3.0 web server. For each complex, the RING 3.0 web server created an intermolecular interaction network among all the residues, enabling one to identify highly conserved and dynamically variable heterogeneous intermolecular interactions. As the first validation, the probabilistic networks of RBDs in their NMR structures generated by the RING 3.0 web server were represented using a NUCPLOT diagram (Figure 2A). Based on the present interaction-based clustering, we found the same intermolecular interactions as those observed in the NMR structures of RBDs, that is, K21, W29, R61, and F96 in RBD1 and K110, T146, R150, K182, and Q185 in RBD2. This indicates that the RING 3.0 web server can identify plausible protein-RNA intermolecular interactions that match those observed in experimental structures.

After validating the interaction-based clustering using the NMR structures of RBDs, we addressed the protein-RNA intermolecular interactions in the complexes of each PaCS-MD trial. Figure 2B–F displays probabilistic networks in the complexes of each trial, which are represented as edges and nodes created by the RING 3.0 web server. Based on the probabilistic networks, we observed the formation of crucial hydrogen bonds between RBDs and the RNA, including the following set of pairs (Figure 2B–F): G1-K177, G1-M178, U2-G115, G4-R150, G5-W29, G5-K88, U6-G26, U6-K93, A7-F96, A7-G8, A7-K21, U9-R98, U9-R99, U9-K103, U9-K182, and U9-Q185. These crucial hydrogen bonds between RBDs and the RNA were identical to those observed in the NMR structures, indicating that PaCS-MD can construct plausible protein-RNA complexes.

3.3. Performance of constructing complexes with PaCS-MD

Finally, we compared the performance of constructing protein-RNA complexes between the present protocol (PaCS-MD) and the standard protocol (Phyre2). For this comparison, we performed 300-ns MD simulations from the most representative complex constructed by one of the PaCS-MD trials (the R3 complex with the smallest RMSD value in Figure S2D) and the complex predicted by Phyre2. First, we evaluated the intermolecular interactions between RBDs and the RNA by counting the number of intermolecular contact atoms and hydrogen bonds on the protein-RNA interface calculated using the trajectories of each protocol. Figure S3 shows the time series of the number of intermolecular contact atoms and hydrogen bonds computed using the last 100-ns trajectory of each 300-ns MD simulation. For the number of intermolecular contacts, the means and standard deviations of the five replicas were 434.00 ± 79.00 (PaCS-MD) and 301.00 ± 94.00 (Phyre2). Additionally, for the number of hydrogen bonds, their means and standard deviations of the five replicas were 23.00 ± 4.00 bonds (PaCS-MD) and 19.00 ± 5.00 bonds (Phyre2). Therefore, the R3 complex of PaCS-MD interacted with the RNA more tightly than the complex of Phyre2.

As a comparison with the conventional flexible docking, we independently performed five traditional MD simulations from the reactant for PaCS-MD by changing the initial velocities. Quantitatively, the accumulated computational costs of the five PaCS-MD trials and the five traditional MD simulations were 100 ns (100-ps MD × 200 cycles × 5 replicas) and 1 μs (200-ns MD × 5 replicas), respectively. Figure S4 shows a comparison of RMSD profiles between the present PaCS-MD and traditional MD, where RMSD was measured from the product.

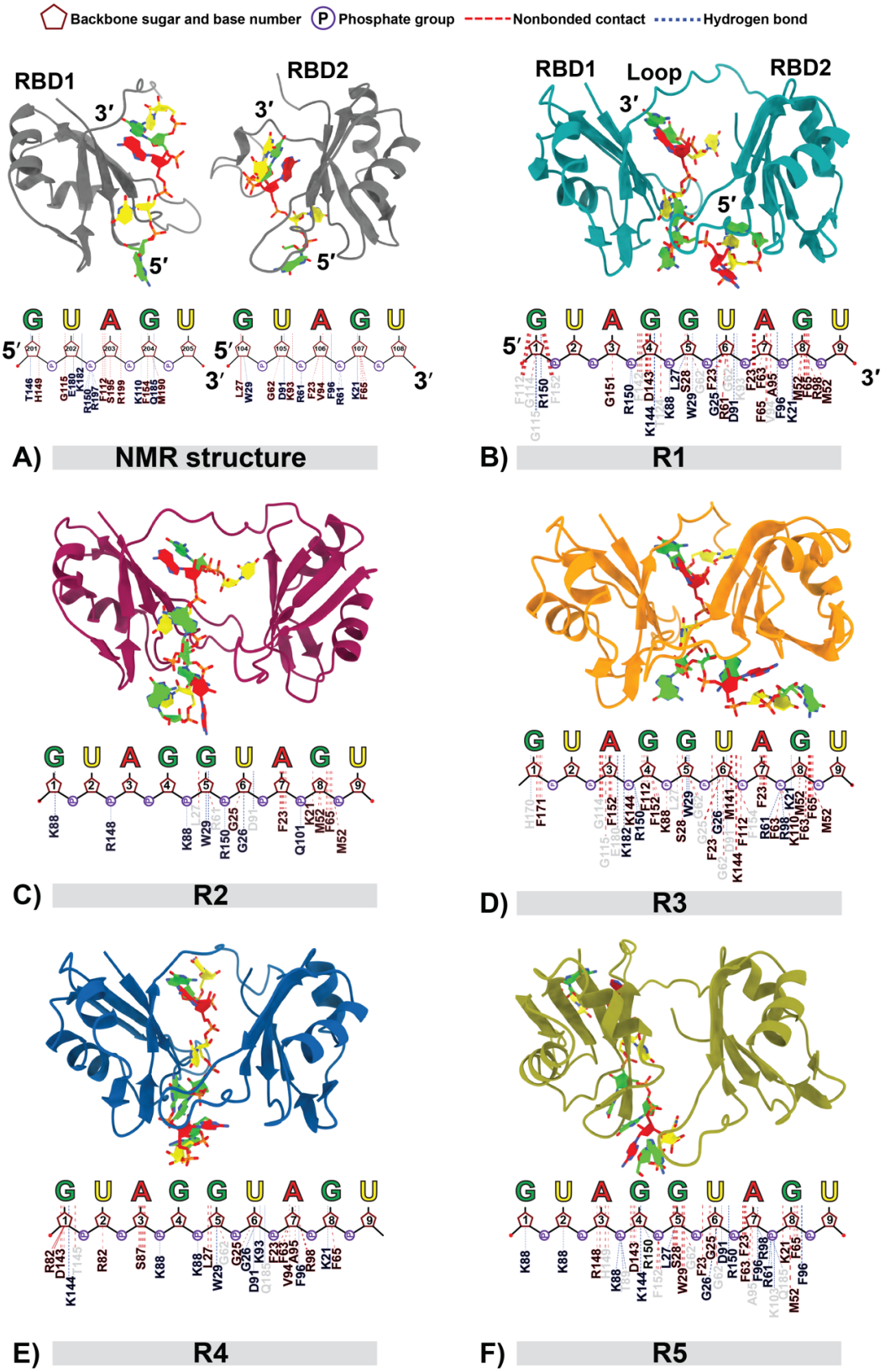

Figure 2. (A) Protein-RNA intermolecular interaction networks in the NMR structures (PDB IDs: 2RS2 for RBD1 and 5X3Z for RBD2). RNAs are represented as sticks: guanine is green, adenine is red, and uracil is yellow. (B–F) Intermolecular interaction networks in the representative protein-RNA complexes constructed by the five PaCS-MD trials (R1–R5) and the target sequences RNA (5'-GUAGGUAGU-3'). The bases are represented by their one-letter codes and colors: guanine is green, adenine is red, and uracil is yellow.

From the RMSD profiles, the traditional MD simulations failed to sample the plausible protein-RNA complex in all five trials because the RMSD values fluctuated largely over 10 Å. In contrast, the flexible docking based on PaCS-MD exhibited a higher conformational sampling efficiency than traditional MD because the RMSD values after the 100th cycle had converged well. In summary, PaCS-MD can construct plausible RNA-protein complexes with a lower computational cost compared with traditional MD judging from the accumulated computational costs in the present comparison, that is, 100 ns (PaCS-MD) versus 1 μs (traditional MD).

We also quantified the intermolecular interactions between RBDs and the RNA based on their binding free energy (ΔG_{bind}). The ΔG_{bind} values were calculated using the last 100-ns trajectories of each 300-ns MD simulation (Tables S1 and S2). Figure 3A shows the ΔG_{bind} values averaged over the five replicas in each protocol. From the means of ΔG_{bind}, the protein-RNA intermolecular interaction of the R3 complex of PaCS-MD was stronger than the complex of Phyre2. For a more detailed analysis, ΔG_{bind} was decomposed into each residue, that is, the

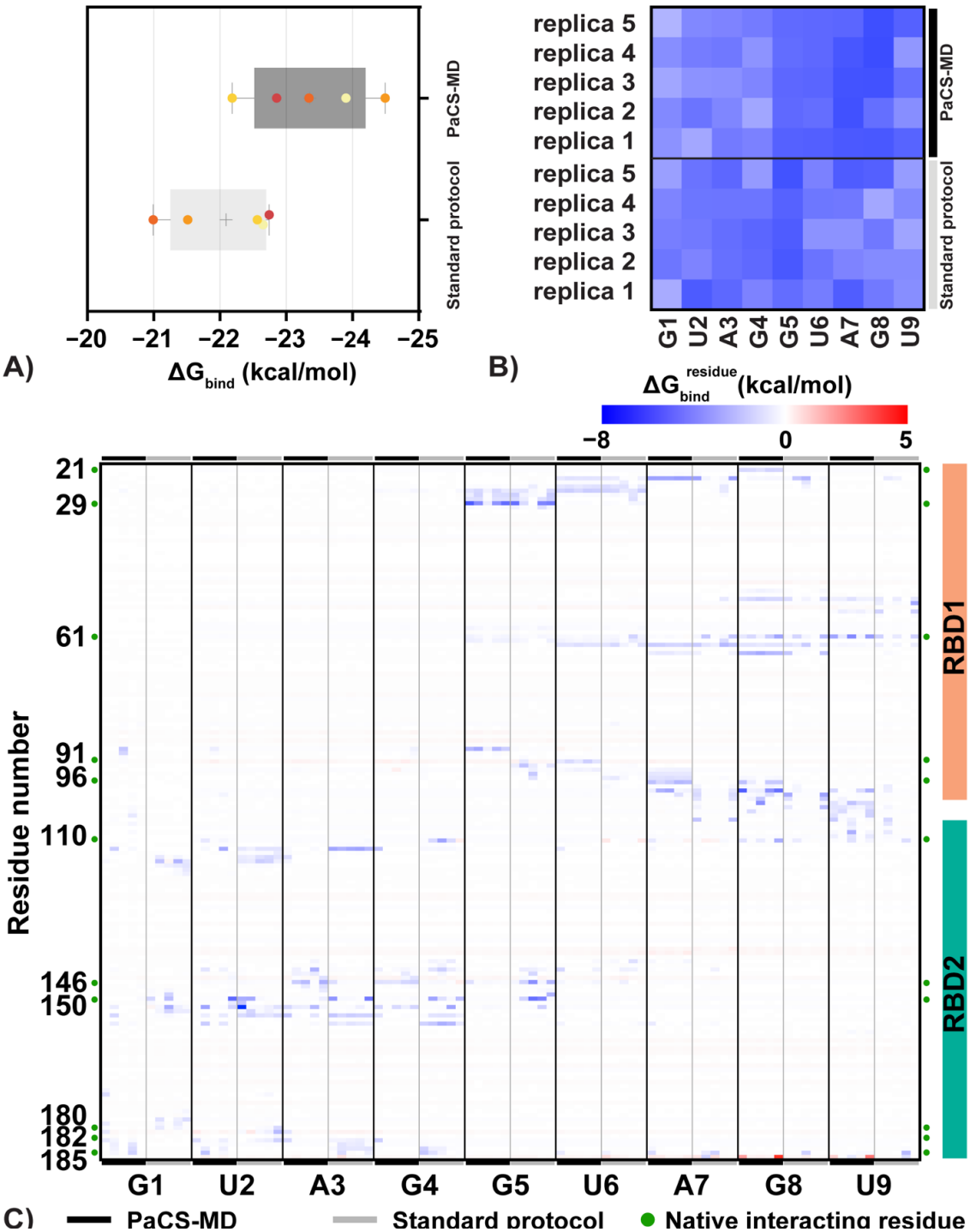

Figure 3. (A) ΔG_{bind} averaged over the five replicas of both PaCS-MD and Phyre2. (B) $\Delta G_{\mathrm{bind}}^{\mathrm{residue}}$ of the five replicas of both protocols versus the target sequence (5'-GUAGGUAGU-3'). (C) The decomposition of $\Delta G_{\mathrm{bind}}^{\mathrm{residue}}$ into amino-acid residues and RNA units. The native interacting residues in the NMR structures of RBD1 and RBD2 are highlighted by green dots. The same spectrum bar is used for both (B) and (C).

binding free-energy contribution from each residue was calculated as $\Delta G_{\text{bind}}^{\text{residue}}$. Figure 3B shows $\Delta G_{\text{bind}}^{\text{residue}}$ calculated using the last 100-ns trajectories of the 300-ns MD simulations from the five replicas of the R3 complex (PaCS-MD) and the protein-RNA complex (Phyre2). After the decomposition, there was no significant difference in $\Delta G_{\text{bind}}^{\text{residue}}$ among the five replicas of each protocol, indicating that PaCS-MD is sufficient to estimate the protein-RNA intermolecular interaction using the binding free-energy decomposition, $\Delta G_{\text{bind}}^{\text{residue}}$.

For additional validation, we compared the protein-RNA intermolecular interactions derived from each protocol with those derived from the NMR structures of RBDs. Figure 3C shows the further decomposition of $\Delta G_{\text{bind}}^{\text{residue}}$ into amino-acid residues and RNA units. From further decomposition, the native interacted residues in the NMR structures showed good correspondence with the high binding affinity sites in RBDs identified by $\Delta G_{\text{bind}}^{\text{residue}}$. This indicates that both PaCS-MD and Phyre2 can reproduce the crucial protein-RNA intermolecular interactions between RBDs and RNA. In conclusion, the flexible docking based on PaCS-MD can construct plausible protein-RNA complexes by considering their crucial intermolecular interactions.

4. Conclusion

In the present study, we proposed an efficient structural refinement technique for constructing plausible protein-RNA complexes based on AI-based modeling (AF2) and flexible docking (PaCS-MD). As a demonstration, we used MSI1-RBDs with the RNA. First, a complete complex (MSI1-RBDs) was predicted based on the amino-acid sequences in humans using AF2. Sequentially, PaCS-MD was used to construct the protein-RNA complexes consisting of RBDs and the RNA. To validate the structural stability of the protein-RNA complexes, we evaluated their intermolecular interactions based on the SIE method. Our findings corroborate previous studies showing that core trinucleotide-based MSI1-RBDs play a significant role in the intermolecular interaction of MSI1-RBDs with the target RNAs.[9,23] Additionally, a group of crucial residues for RNA-binding of the complexes constructed by PaCS-MD related well to the previous findings.

Our technique has an advantage over template-based modeling, which is limited by the need to connect RNAs with RBDs manually after constructing their overall conformation. In the present study, for the MSI1 protein, we used the tool to remove the duplicated RNA region and connect the RNAs from the NMR structures. However, this does not ensure that the manually ligated RNA bounded with RBDs will be a natural complex. In contrast, PaCS-MD enables one to automatically construct protein-RNA complexes from those predicted using AF2 based on flexible docking without ligating the RNAs, which might be an advantage of our technique.

The concept of our study is to construct reasonable protein-RNA complex structures using the PaCS-MD simulation technique, which has demonstrated superiority over the conventional method (Phyre2), highlighting the advantages and improvements achieved by our technique. Unlike conventional homology modeling methods, our technique promotes essential transitions for forming protein-RNA complexes by capturing the binding interactions between RNA and protein. However, our technique utilizes the NMR coordinates of RNA in complex with RBDs as a reference for PaCS-MD. Therefore, in the absence of any structural information, our method, similar to previous studies employing PDB data for the conventional docking methods based on structural homology, would face limitations. It would indeed be challenging to achieve accurate flexible docking without any initial structural insights, which might be achieved by an extended (non-targeted) PaCS-MD[44-46] that samples transition pathways of proteins without any reference. In our future study, we will extend our method as more versatile flexible docking based on non-targeted PaCS-MD.

Finally, in future studies, we will use flexible docking (PaCS-MD) to construct the binding of a longer RNA chain that includes both binding motifs of the two RBDs of MSI proteins. For this issue, PaCS-MD could be a beneficial technique for constructing complexes of non-native RNA that bind to RBDs. From the viewpoint of medical applications, a deeper understanding of how the MSI1 proteins interact with RNAs may be helpful for several therapeutic approaches, such as for the Zika virus, cancer therapy, and targeted therapy.

Acknowledgement

This research project was funded by the Second Century Fund (C2F), Chulalongkorn University, the National Research Council of Thailand (NRCT), NRCT5–RGJ63001–009, and the 90th anniversary of CU Fund (Ratchadaphiseksomphot Endowment Fund), GCUGR1125651012D. We also thank the ASEAN–European Academic University Network (ASEA–UNINET) for a short visit grant and the Vienna Scientific Cluster (VSC) for facilities and computing resources. This research was partly funded by the Austrian Science Fund FWF (Grant I 6440-N) to MTW. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission. This research was also funded by the Mishima Kaiun Memorial Foundation and the Sumitomo Foundation, Japan.

References

[1] M. W. Hentze, A. Castello, T. Schwarzl, T.Preiss, *Mol. Cell Biol*. **2018**, *19*, 327.

[2] C. A. O. Shotwell, J. D. Cleary, J. A. O. Berglund, *WIREs* **2020**, *11*, e1573.

[3] C. Oliveira, H. Faoro, L. R. Alves, S. Goldenberg, *Genet. Mol. Biol.* **2017**, *40*, 22.

[4] A. E. Kudinov, J. Karanicolas, E. A. Golemis , Y. Boumber, *Clin. Cancer Res.*, **2017**, *23*, 2143.

[5] A. D. B. Schneider, M. T. olfinger, *Sci. Rep.* **2019**, *9*, 6911.

[6] A. E. Kudinov, A. O. Deneka, A. S. Nikonova, T. N. Beck, Y. H. Ahn, X. Liu, C. F. Martinez, F. A. Schultz, S. Reynolds, D. H. Yang, K Q. Cai, K. M. Yaghmour, K. A. Baker, B. L. Egleston, E. Nicolas, A. Chikwem, G. Andrianov, S. Singh, H. Borghaei, I. G. Serebriiskii, D. L. Gibbons, J. M. Kurie, E. A. Golemis, Y. Boumber, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 6955.

[7] A. Deneka, L. Kharin, N. S. Karnaukhov, M. Voloshin, T. G. Ayrapetova, E. Ratner, A. Sabirov, Y. Topchu, A. Mazitova, Z. Abramova, V. Yugai, E. M. Frantsiyants, O. I. Kit, Y. Boumber, *J. Clin. Oncol.* **2020**, *38*, e21583.

[8] G. B. Caldas-Garcia, L. Schüler-Faccini, A. Pissinatti, V. R. Paixão-Côrtes, M. C. Bortolini, *Infect. Genet. Evol.* **2020**, *84*, 104364.

[9] N. Darai, P. Mahalapbutr, P. Wolschann, V. S. Lee, M. T. Wolfinger, T. Rungrotmongkol, *Sci. Rep.* **2022**, *12*, 12137.

[10] R. Iwaoka, T. Nagata, K. Tsuda, T. Imai, H. Okano, N. Kobayashi, M. Katahira, *Molecules* **2017**, *22*, 1207.

[11] J. F. Zheng, X. Hong, J. Xie, X. X. Tong, S. Y. Liu, Bioinformatics, **2020**, *36*, 96.

[12] K. Kappel, R. Das, *Structure*, **2019**, *27*, 140.
[13] H. Q. Meng, C. Q. Li, Y. Wang, G. J. Chen, *PLoS One*, **2014**, *9*, e86104.
[14] R. Harada, Y. Takano, T. Baba, Y. Shigeta, *Phys. Chem. Chem. Phys.* **2015**, *17*, 6155.
[15] R. Harada, *Bull. Chem. Soc. Jpn.* **2018**, *91*, 1436.
[16] R. Harada, A. Kitao, *J. Chem. Phys.* **2013**, *139*, 035103.
[17] A. Kitao, R. Harada, Y. Nishihara, D. P. Tran, *AIP Conf. Proc.* **2016**, *1790*, 020013.
[18] R. Harada, T. Nakamura, Y. Shigeta, *Chem. Phys. Lett.* **2015**, *639*, 269.
[19] R. Harada, T. Nakamura, Y. Shigeta, *Bull. Chem. Soc. Jpn.* **2016**, *89*, 1361.
[20] R. Harada, Y. Shigeta, *J. Comput. Chem.* **2017**, *38*, 2671.
[21] T. Baba, R. Harada, M. Nakano, Y. Shigeta, *J. Comput. Chem.* **2014**, *35*, 1240.
[22] J. Fujita, R. Harada, Y. Maeda, Y. Saito, E. Mizohata, T. Inoue, Y. Shigeta, H. Matsumura, *J. Struct. Biol.* **2017**, *198*, 65.
[23] T. Ohyama, T. Nagata, K. Tsuda, N. Kobayashi, T. Imai, H. Okano, T. Yamazaki, M. Katahira, *Nucleic Acids Res.* **2012**, *40*, 3218.
[24] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583.
[25] M. Mirdita, S. Ovchinnikov, M. Steinegger, *Nat. Methods*, **2022**, *19*, 679.
[26] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, *Nat. Protoc.* **2015**, *10*, 845.
[27] BIOVIA, D. S., Discovery Studio Visualizer, 2020, San Diego: Dassault Systèmes **2020**.
[28] D.A. Case, H.M. Aktulga, K. Belfon, I.Y. Ben-Shalom, J.T. Berryman, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, G.A. Cisneros, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, G. Giambasu, M.K. Gilson, H. Gohlke, A.W. Goetz, R. Harris, S. Izadi, S.A. Izmailov, K. Kasavajhala, M.C. Kaymak, E. King, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, M. Machado, V. Man, M. Manathunga, K.M. Merz, Y. Miao, O. Mikhailovskii, G. Monard, H. Nguyen, K.A. O'Hearn, A. Onufriev, F. Pan, S. Pantano, R. Qi, A. Rahnamoun, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, A. Shajan, J. Shen, C.L. Simmerling, N.R. Skrynnikov, J. Smith, J. Swails, R.C. Walker, J. Wang, J. Wang, H. Wei, R.M. Wolf, X. Wu, Y. Xiong, Y. Xue, D.M. York, S. Zhao, and P.A. Kollman (2022), Amber 2022, University of California, San Francisco.
[29] D. Clementel, A. Del Conte, A. M. Monzon, Camagni, F. Giorgia, G. Minervini, D. Piovesan, S. C. E. Tosatto, *Nucleic Acids Res.* **2022**, *50*, 651.
[30] E. D. Jurrus, K. Star, K. Monson , J. Brandi, L. E. Felberg, D. H. Brookes, L Wilson, J. Chen, K. Liles, M. Chun, D. W. Gohara, T. Dolinsky, R. Konecny, D. R. Koes, J. E. Nielsen, T. Head-Gordon, W. Geng, R. Krasny , G. W. Wei, M. J. Holst, J. A. McCammon, N. A. Baker, *Protein Sci.* **2018**, *27*, 112.
[31] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, C. Simmerling, *J. Chem. Theory Comput.* **2015**, *11*, 3696.
[32] D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman (2014), AMBER 14, University of California, San Francisco.
[33] P. Mark, L. Nilsson, *J. Phys. Chem. A* **2001**, *105*, 9954.
[34] V. Krautler, W. F. Van Gunsteren, P. H. Hunenberger, *J. Comput. Chem.* **2001**, *22*, 501.
[35] B. Knapp, L. Ospina, C. M. Deane, *J. Chem. Theory Comput.* **2018**, *14*, 6127.
[36] R. Baron, *Computational Drug Discovery and Design*. Springer, New York, Dordrecht, Heidelberg, and London, 2012.
[37] M. Naim, S. Bhat, K. N. Rankin, S. Dennis, S. F. Chowdhury, I. Siddiqi, P. Drabik, T. Sulea, C. I. Bayly, A. Jakalian, E. O. Purisima, *J. Chem. Inf. Model.* **2007**, *47*, 122.
[38] K. Hengphasatporn, N. Kungwan, T. Rungrotmongkol, *J. Mol. Liq.* **2019**, *274*, 140.
[39] K. Hengphasatporn, P. Wilasluck, P. Deetanya, K. Wangkanont, W. Chavasiri, P. Visitchanakun, A. Leelahavanichkul, W. Paunrat, S. Boonyasuppayakorn, T. Rungrotmongkol, S. Hannongbua, Y. Shigeta, *J. Chem. Inf. Mod.* **2022**, *62*, 1498.
[40] K. Hengphasatporn, R. Harada, P. Wilasluck, P. Deetanya, E. R. Sukandar, W. Chavasiri, A. Suroengrit, S. Boonyasuppayakorn, T. Rungrotmongkol, K. Wangkanont, Y. Shigeta, *Sci. Rep.* **2022**, *12*, 17984.
[41] G. Wolber, T. Langer, *J. Chem. Inf. Model.* **2005**, *45*, 160.
[42] K. Hengphasatporn, A. Garon, P. Wolschann, T. Langer, Y. Shigeta, T. N. Huynh, W. Chavasiri, T. Saelee, S. Boonyasuppayakorn, T. Rungrotmongkol, *Sci. Pharm.* **2020**, *88*.
[43] K. Sanachai, P. Mahalapbutr, K. Hengphasatporn, Y. Shigeta, S. Seetaha, L. Tabtimmai, T. Langer, P. Wolschann, T. Kittikool, S. Yotphan, *ACS Omega* **2022**, *7*, 33548.
[44] R. Harada, A. Kitao, *J. Chem. Theory Comput.* **2015**, *11*, 5493.
[45] R. Harada, V. Sladek, Y. Shigeta, *J. Chem. Theory Comput.* **2019**, *15*, 5144.
[46] R. Harada, K. Yamaguchi, Y. Shigeta, *J. Chem. Theory Comput.* **2020**, *16*, 6716.