Stony Brook University

MELD Application on refining DOCK results

A Thesis Presented

by

Cong Liu

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Master of Science

in

Chemistry

Stony Brook University

May 2016

Stony Brook University

The Graduate School

Cong Liu

We, the thesis committee for the above candidate for the

Master of Science degree, hereby recommend

acceptance of this thesis.

Professor. Ken A. Dill

Advisor

Director, Laufer Center for Physical and Quantitative Biology

Professor, Physics and Chemistry, Stony Brook University

Prof. Carlos Simmerling

Associate Director, Laufer Center for Physical & Quantitative Biology

Professor of Chemistry ,Stony Brook, NY

&

Prof. David F. Green

Associate Professor and Graduate Program Director

Department of Applied Mathematics and Statistics

Stony Brook University

This thesis is accepted by the Graduate School

Charles Taber

Dean of the Graduate School

Abstract of the Thesis

MELD application on refining DOCK results

by

Cong Liu

Master of Science

in

Chemistry

Stony Brook University

2016

In order to bridge the computational gap between DOCK and MD-based methods, a pipeline was designed to combine external information derived from DOCK with replica exchange molecular dynamics using MELD, an algorithm recently developed by our lab. Our goal was to predict the lowest free energy binding pose within the DOCK ensemble, therefore enabling us to refine DOCK results in cases where the correct structure is incorrectly scored. Our method was tested on 8 such 'scoring failures' found in the DOCK database SB2012. Three protocols of extracting structural information from DOCK ensembles were designed and tested. Crystal-like poses were correctly identified from the DOCK ensemble using all three protocols. In particular, the prediction within 3Å of the crystal pose using information derived from "Automatically cluster and choose based on KGS penalty function" protocol demonstrated the ability of this pipeline to refine DOCK results.

List of Contents

List of Figures

List of Tables

List of Abbreviations

MELD - Modeling Employing Limited Data
MD - Molecular Dynamic simulation
KGS - Kelley-Gardner-Sutcliffe penalty function
GAFF - General AMBER force field
T-REMD - Temperature replica exchange simulation
RMSD - Root mean square deviation

Acknowledgement

First, I would like to express my sincere and deepest gratitude to my dissertation advisor Professor Ken A. Dill for his full support, expert guidance and encouragement during my graduate study. Without his incredible patience and wisdom and instruction, I couldn't complete my research work. And I also would like to thank you and Mrs. Jolanda Dill for your generous hospitality during annual Thanksgiving Eve parties at your home.

I would also like to thank my committee members, Professor Carlos Simmerling and Professor David F. Green for providing advices and schedule time to attend my defense.

I would also like to thank Professor Robert C. Rizzo for providing the SB 2012 DOCK database and DOCK outcomes summary file. Thanks Brian Fochtman for technical details about clustering DOCK outcomes.

I also want to express my appreciation to my mentor, Dr. Joseph Morrone, for the encouragement, guidance and valuable advice I received from him. Your inspirational advice to my research really helps me go through a lot of obstacles.

I would like to thanks my Karate Sensei, Dan Hayes. You taught me how to individually develop spiritual and physical personalities. Every Tuesday and Thursday night's physical training at your Dojo really help me release the pressure from academic research and make me better understand my body.

I would like to acknowledge members in Ken research group: Dr. Alberto Perez, Dr. Emiliano Brini, Dr. Jason Wagoner, Dr. Lane Votapka, Mariola Szenk, Elizaveta Guseva, Michael Hazoglou, Mantu Santra. It is a great pleasure working with you and I wish you all the best of luck in your future. I would also like to give a special thanks to Nancy Rohring, Feng Zhang and Eileen Dowd for your kindness and support.

I am most grateful to my family members and friends for their love and support. To both of my parents, JinMin and Liping, I am so lucky to have you always with me and support me. I also want to express my appreciation to my friends. Thank you for always be there for me.

MELD application on refining DOCK results

1.Introduction

The methodology of Computer-aided drug design has been widely used in all stages of the 'hit to lead' optimization protocol of the modern drug discovery industry[1]. During the 'hit' process, the primary goal is to quickly screen millions of molecules and find the most promising ones. Compromising some accuracy for speed is acceptable. However, for the lead optimization process, accurately predicting the binding free energy, therefore distinguishing the 'best' ligand from the 'better' ones is the main goal[2]. Based on different needs, various methods have been used in this field and they could be roughly separated into two groups: the structure-based DOCK method and more computationally expensive MD-based method[3] which sample the free energy landscape. Here we bridge these two approaches using MELD[4,5] a physics-based, enhanced sampling technique that can utilize information provided by approaches such as DOCK to rigorously sample ligand binding poses and find the most stable structure.

The DOCK method, which was first implemented based on shape complimentary and later included more sophisticated force field scoring functions and sampling strategies, is noted for its speed and reasonable accuracy[6]. It enables the user to quickly evaluate the fitness of millions of molecules to a known protein structure and exhibits a quite high success rate to certain protein families[7]. DOCK uses the genetic algorithm quickly sample ligand conformation space and a scoring function to rank all possible conformations based on shape and energetic compatibility[6]. It has achieved a remarkable success in drug discovery. The discovery of DNA gyros inhibitors is a well-known case[8].

As implemented, DOCK has limitations. In most cases, the DOCK method uses a rigid protein structure. This might cause insufficient sampling of the protein conformational space during the binding process[3]. Moreover, a rigid protein binding cavity may also cause a insufficient sampling of ligand conformation space. To address this, various methods had been implemented, including rotamer libraries, soft docking and use of multiple receptor conformation of the binding pocket[3]. However, these simple methods still have difficulty handling the protein backbone level rearrangement[9]. Another limitation of DOCK is incapable of truly evaluating binding affinities. Due the intrinsic implementation of DOCK algorithm, the entropy effect is not included in the binding process.

These kind of limitations of DOCK might contribute to failures in DOCK calculations. Recently a test of DOCK was carried by Allen, et al[10]. They test the pose reproducibility of DOCK with respect to a DOCK validate dataset SB2012 which consists of 1042 systems . It was found that in 73.3%(765/1043) of the cases DOCK was to be correctly predict the binding poses. However, for the other 26.9% failure cases, 66%(183/278) of the case was due to a scoring failure, that is the DOCK program could successfully sample the crystal structure but the scoring function didn't correct rank it on the top[6]. They also argued that the lacking of an entropic term

in the scoring function or limited protein flexibility in the DOCK process might contribute to scoring failures.

Molecular dynamics(MD) can be used to refine docking results, especially in the case of scoring failures. The underlying assumption is that the correct, experimentally observable docking pose is a minimum on the free energy landscape and therefore will be frequently sampled by the MD simulation[3]. A bad docking pose is less stable minimum and therefore the simulation would sample around it infrequently[3]. Moreover, unlike DOCK, the MD simulation could capture the so-called "induced-fit effect" where the binding cavity of the protein rearranged to establish more specific and stronger interaction with ligand molecule. Also, MD can better capture solvent effect.

However MD also has limitations. Due to the rugged nature of the free energy landscape[11], the system might be trapped in a local minimum region and unable to get out and sample other regions. Although advanced computation tools like T-REMD[3,12,13,14] were invented to address this problems, the efficiency still wasn't enough to allow fast sampling in the binding process. Problems like repeated sampling at certain region and longer equilibration time[11] hinder its application in the lead process of drug discovery.

MELD was introduced by MacCallum,Perez, and Dill. to address this sampling efficiency problem by utilizing sparse, ambiguous information[4]. In our implementation of MELD, the information constraints which derived from DOCK ensembles were combined with T-REMD simulation to facilitate the sampling of ligand binding poses. Information uncertainties were handled by group and collection strategies to make sure that the sampling results were consistent with those from sampling the unbiased free energy landscape. MELD can shorten the sampling time without losing the physical reality of the system as described by standard force fields.

In this project, we designed a pipeline to refine DOCK results using the physics-based MD simulation method MELD. The structural information of DOCK ensemble was converted to MELD constraints through three successful constraint generating protocols. Then those constraints were used in MELD. After applying this pipeline to several protein systems, we found eight systems where DOCK experienced a scoring failure and could predict some of these complex systems' crystal conformation within 3Å. The high level of accuracy obtained in this study demonstrated the power of MELD and potential application to structure-based drug design.

2.DOCK

DOCK has became a powerful tools in the structure-based "hit to lead" drug design process[7,10]. It relies on the generic algorithm to sample ligand conformation space and generate DOCK ensemble structure[1]. Then scoring functions are used to rank poses in the DOCK ensemble based on their shape and energetic compatibility with the receptor proteins. This kind of process can be quite fast and efficient in generating preliminary hit structure and has achieved lots of success in the past. Recent studies by He et al. [14]and Khare et al. [15] were two of these examples. But the limitations of it, including lack of protein flexibility and entropic effect, also cause failure cases.

MD based methods were powerful tools to refine DOCK failure cases[16]. The protein flexibility and entropic effect could be captured by MD simulation.

2.1. DOCK Database

The DOCK database SB2012, which were constructed by Allen et al[10], consisted of 1048 ligand-receptor complex systems was the ideal DOCK database to search for refinement targets. The SB 2012 consisted of 25 protein families and various ligand types with different amount of rotation bonds. SB2012 was available on Rizzo's lab webpage and free to download.

2.2. Refinement Target Systems.

DOCK scoring failures are the refinement targets. The DOCK outcomes could be classified into three groups: success, scoring failure and sampling failure[7]. By definition, the DOCK scoring failure means that the crystal conformation is successfully sampled by DOCK but the scoring functions don't rank it at the top. Sampling failure means that the crystal conformation aren't sampled by DOCK.

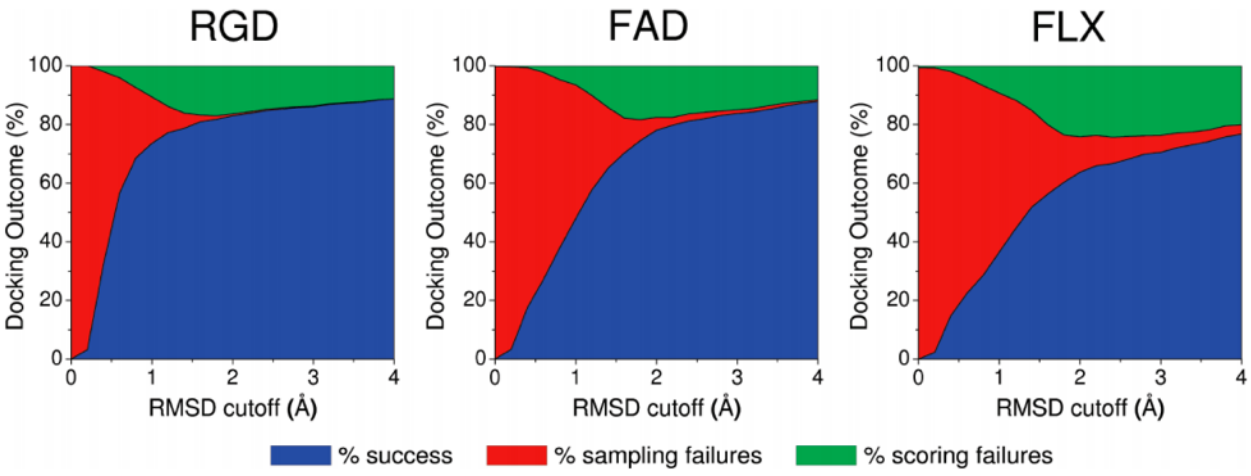

Figure 1. DOCK outcomes vs RMSD cutoff distance.

In order to define whether a pose in DOCK ensemble is accounted as crystal pose, a RMSD cutoff distance is chose. Based on how big this RMSD cutoff, the DOCK outcomes could varies. Figure 1. shows that the composition of DOCK outcomes varies depends on the cutoff RMSD distance. In practice, balancing the DOCK success rate and the accuracy of DOCK ensemble poses, 2Å RMSD cutoff distance was used to define whether a DOCK pose is accounted as crystal pose.

Recently pose reproducible test by Allen[10] shown that the scoring failure consisted the main part of DOCK failure cases. 66.7% of all failure cases were due to scoring failure. Therefore, DOCK scoring failure became our refinement targets.

Based on the computing power of our refinement tools and the complexity of DOCK scoring failure systems, we decided to focus on the DOCK scoring failure cases with receptor protein length shorter than 250 amino acid and single ligand systems.

After searching the DOCK database SB2012 based on these criterions, we found out twenty systems that meet our request. These systems are listed below in **Table I**

Table I: Refinement Target systems:

PDBID	Protein Length	Charge	Protein Type
1SV9	121	-2	Phospholipase A2
2QVD	121	-1	Lysozyme
1Z6S	124	-2	Ribonuclease pancreatic
1BB7	129	0	Lysozyme
1HEW	129	0	Lysozyme
1LMO	129	0	Lysozyme
1LZB	129	0	Lysozyme
1LZC	129	0	Lysozyme
1LZE	129	0	Lysozyme
1BB5	130	0	Lysozyme
1TOW	131	-1	Fatty acid-binding protein
9HVP	198	0	HIV-I protease
1JLD	198	0	HIV-I protease
1HPX	198	0	HIV-I protease
1HXB	198	1	HIV-I protease
1UY8	208	0	HEAT SHOCK PROTEIN
1UYC	209	0	HEAT SHOCK PROTEIN
1UYH	209	0	HEAT SHOCK PROTEIN
1UYK	209	0	HEAT SHOCK PROTEIN
1UYF	209	0	HEAT SHOCK PROTEIN

3.MELD

MELD is a novel, physics-based computational tools designed by Maccallum, Perez, Dill[4,5] to utilize information constraints to accelerate temperature replica exchange(T-REMD) simulation.

MELD consists of two parts: T-REMD simulation and information constraints. Information constraints defines specific regions on free energy landscape where the most likely the global minimum would be found. The T-REMD simulation provides a efficient sampling tools to sample around those regions and find out the most global likely conformation.Moreover, MELD also contains fancy strategies: group and collection to handle the uncertainty of the information constraints to make sure the sampling results are consistent with force field.

In this section, we will introduce these three main feature of MELD:

- Information constraints;
- T-REMD;
- Group and collection;

3.1. Information constraints:

There are information constraints in MELD: distance constraints and angular constraints. The distance constraints were assigned to among atoms to restraint the distance among atoms. The angular constraints were used to restraint torsion angle in the protein structure. Here we use constraints to guide the ligand to binding cavity on the protein surface, therefore the distance constraints are the most useful.

The distance constraints has the so called “Flat-bottom shape”.

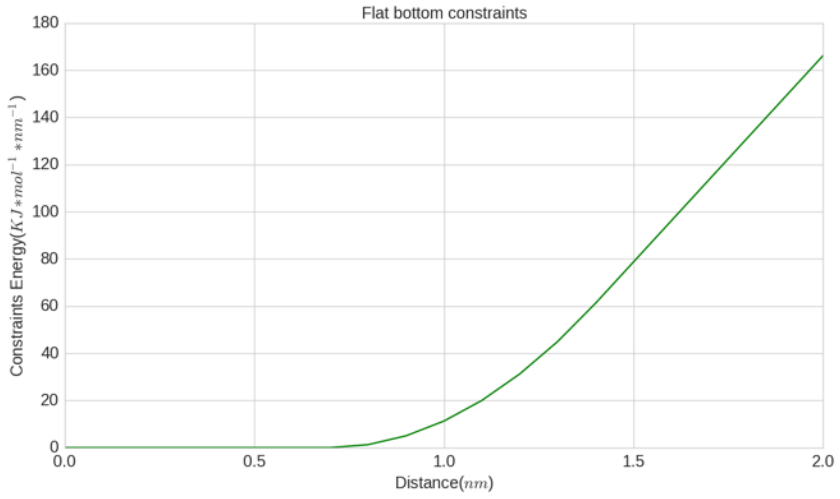

Figure 2. Flat bottom constraints

As the Figure 2. showed here, the distance constraints in MELD has the “flat bottom shape”.

- Flat bottom region: Distance <= cut_off_1, $E_{constraints} = 0$;
- Quartic region: cut_off_1 < Distance <= cut_off_2, $E_{constraints} > 0$;
- Linear region: Distance > cut_off_2, $E_{constraints} > 0$;

The information constraints accelerate the T-REMD simulation. When ligand samples at region which doesn’t consistent with constraints(Quartic region and Linear region), the information constraints are activated and external forces are applied to constrained atom pairs. These external forces drive the ligand towards the region which consistent with the information constraints(Flat bottom region). Therefore it accelerates the sampling in the T-REMD simulation.

Moreover, MELD preserves the relative binding affinity.

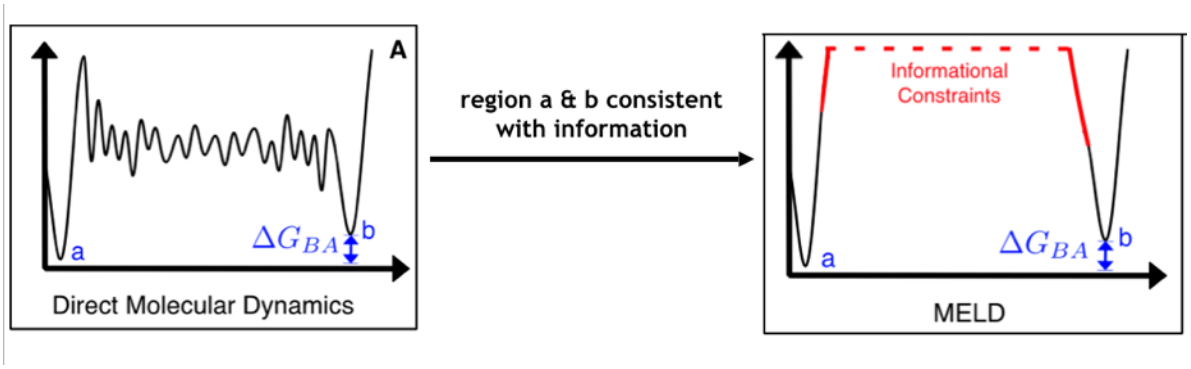

Figure 3. MELD preserves the relative binding affinity.

The Figure 3. shows how MELD information constraints help preserve the relative free energy. The left graph represents the original free energy landscape which described by force field and solvent model. Two minimum a and b represents the region which consistent with MELD information constraints. The original relative free energy difference is represented by ΔG_{BA}. When ligand sample around region a and b which are consistent with information constraints in MELD, no MELD constraint is activated due to the flat bottom shape of constraint energy function. Therefore no MELD energy is introduced into simulation system and the shape of the free energy landscape is unaltered around those two region. Thus the relative free energy preserves.

In this project, we used a distance constraints function with details listed below[4].

Table II: Information constraint function details:

cut_off_1(nm)	0.7
cut_off_2(nm)	0.9
Force Constant($KJ*mol^{-1}*nm^{-2}$)	250

Moreover, in order to cooperate with T-REMD simulation to maximize the sampling efficiency around the binding free energy landscape, the constraint force constant is scaled along temperature replica ladder. Each replica in T-REMD simulation is assigned a value alpha to determine the exact force constant on that replica. The exact math formula is given below[4]:

$$\mathrm{k} = \begin{cases} 1 & \text{if } \alpha < \alpha_{\min} \\ \dfrac{e^{\text{factor}} * e^{\frac{\alpha_{\max} - \alpha}{\alpha_{\max} - \alpha_{\min}}} - 1}{e^{\text{factor}} - 1} & \text{if } \alpha_{\min} \leq \alpha < \alpha_{\max} \\ 0 & \text{if } \alpha \geq \alpha_{\max} \end{cases} \quad \textbf{(Eq.1)}$$

α_{min} α_{max} are the starting replica and ending replica of constraint scaling. factor controls the steepness of constraints scaling changing.

For replicas which had alpha value lower than the setting alpha mim, no scaling was applied to the constraints' force constant.

For replicas within the range of alpha in and alpha max, the constraints' force constant was generally weaken to relax the interaction between ligand and receptor. This process coupled with temperature increasing, therefore the ligand had probability to escape trap and sample wider region.

At the top several replicas, which the alpha greater than the setting alpha max, the constraints' force constant equaled to zero and no external force applied between ligand and receptor. The ligand had the full ability to "walk around" the free energy landscape and sampled around the widest region.

In this project, the details about the force constant scaling parameters are listed below:

Table III: Force constant scaling parameters

Alpha min	0.0
Alpha max	0.7
Factor	2.5

3.2. T-REMD

In order to avoid kinetic traps, T-REMD simulation is introduced. The modeling system has "multiple copies"(replicas) which running at different temperatures. At the top replica, where the system is running at the highest temperature. It has the largest probability to escape kinetic traps. When the temperature cools down, the system could sample more diverse local minimums[14].

In order to strictly determine the temperature of each replica, a replica index alpha was given.

The temperature(T) and replica index(α) was linked through Eq.13 [4]:

$$\mathrm{T} = \begin{cases} \mathrm{T}_{\min} & \text{if } \alpha < \alpha_{\min} \\ \mathrm{T}_{\max} * \left[e^{\text{factor}} * e^{\frac{\alpha - \alpha_{\min}}{\alpha_{\max} - \alpha_{\min}}} \right] & \text{if } \alpha_{\min} \leq \alpha < \alpha_{\max} \\ \mathrm{T}_{\max} & \text{if } \alpha \geq \alpha_{\max} \end{cases} \quad \textbf{(Eq.2)}$$

T**min*,T**max* were the corresponding maximum T and minimum T.

For this project, the temperature scaling parameters are listed in **Table III**:

Table IV: Temperature scaling parameters

T min	300K
T max	450K
Alpha min	0.0
Alpha max	0.7

3.3. Group and Collection

The group and collection were designed in a hierarchical way to enable MELD to handle the uncertainty of constraints by using the enforced factor.

"Group aggregated multiple constraints in collective, multi-body constraints such that only a specific fraction of constraints are activated."[4]

$$\mathrm{E}_{\mathrm{i}}^{\mathrm{grp}} = \sum_{\mathrm{j}=1}^{n_{\mathrm{active,i}}^{\mathrm{grp}}} \mathrm{E}_{\mathrm{i,j}}^{\mathrm{rest}} \quad \textbf{(Eq.3)}$$

where the component constraints are sorted by energy[4]:

$$\mathrm{E}_{\mathrm{i},1}^{\mathrm{rest}} \leq \mathrm{E}_{\mathrm{i},2}^{\mathrm{rest}} \leq \cdots \leq \mathrm{E}_{\mathrm{i},\mathrm{N}_i}^{\mathrm{rest}} \quad \textbf{(Eq.4)}$$

The $n_{active,i}$ was the enforced factor which represented how many constraints we could trust.

The collection was the upper layer of group providing additional sorting and activation to handle ambiguity of grouped constraints:

$$\mathrm{E}_{\mathrm{i}}^{\mathrm{coll}} = \sum_{\mathrm{j}=1}^{n_{\mathrm{active,i}}^{\mathrm{coll}}} \mathrm{E}_{\mathrm{i,j}}^{\mathrm{grp}} \qquad \textbf{(Eq.5)}$$

where different groups in collection are sorted by energy[4]:

$$\mathrm{E}_{\mathrm{i},1}^{\mathrm{grp}} \leq \mathrm{E}_{\mathrm{i},2}^{\mathrm{grp}} \leq \cdots \leq \mathrm{E}_{\mathrm{i},\mathrm{N}_i}^{\mathrm{grp}} \qquad \textbf{(Eq.6)}$$

In this project, we used different group and collection strategies to handle the uncertainty. Details will be listed in the next section "Constraint generating protocols".

4. Testing the stability of crystal pose in force field.

The motivation of stability test came from the studies[3,17] which had shown that force field and solvent model might prefer certain conformations. If that were the case, we must limit our refinement to systems which were consistent with the current best force field and solvent model. Based on this understanding, we designed a standard MD simulation protocol to test whether the crystal structures of our target systems were stable. We the current state-of-art protein force field ff12SB[18] and implicit solvent model gbNeck2[19]. The technical details of stability test were separated into four parts:

- Prepare ligand library file;
- Energy minimization of complex structure;
- MD simulation;
- Results;
- Discussion;

4.1. Prepare ligand library file

First, the DOCK pose which had the lowest root-mean-square deviation(rmsd) to crystal pose was chose as the crystal conformation. Visual checking was performed to make sure that this crystal conformation was consisted with its chemical structure.

Second, the charges of ligand atoms were calculated using AM1-BCC[20] charge model in Chimera[21]'s built-in function "Add Charge". Then a mol2 file of the charged ligand was saved relative to protein structure using Amber/GAFF atom types.

Third, this ligand mol2 file was used as template to build a ligand library file using Amber package tLeap[22,23]. The details are listed in Amber's "Antechamber Tutorial"[24]. Once the library was created, the topology(.top) and coordinate file(.inpcrd) of the complex systems were then generated using this ligand library file and ff12SB protein force field.

4.2.Energy minimization of complex structure

The purpose of energy minimization was to relax the system and avoid potential atom clashes

The topology and coordinate file were used as input file for the energy minimization process. The initial crystal structure was energy minimized using Amber's Pemed function[22]. The key parameters were listed in **Table V**:

Table V: Energy minimization parameters

Max number of iteration(**maxcyc**)	500
Gradient Descend steps(**ncyc**)	150
Implicit Solvent Model(**igb**)	GB-Neck2[19]
Cut-off distance for electrostatic interaction(**cut**)	999999
Force Field	ff12SB[18]

After energy minimization, a .rst file which stores the minimized structure was saved. Then a file format converting from rst to pdb was implemented using Amber package program ambpdb. This pdb file was used as input file to run the MD simulation.

4.3.MD simulation

The purpose of the MD simulation at this stage was to find out systems which were consistent with the force field and solvent model and therefore could be stabilized at their crystal conformation.

Through the previous two steps, we got the ligand library file and the energy minimized crystal structure file. Now we could run MD simulation to check which systems were stable. The simulation details are listed below:

Table VI: Stability Test Run Parameters:

Protein Force Field	FF12SB[18]	Software	MELD[4,5]
Ligand Force Field	GAFF[25]	T_scaling	0.0,0.7,300.300
Charge Model	AM1-BCC[20]	Number of Replicas	2
Temperature	300K	Cartesian Constraints	0.35 250
Solvent Model	GB-Neck2[19]		

Note that instead of running a normal MD simulation at 300K, we used a trick here using 2 replicas in MELD to run the stability test. Since the MELD requires minimum two replicas, therefore we set up a temperature scaling with T_{min} = 300k, T_{max} = 300k, $alpha_{min}$ = 0.0, $alpha_{max}$ = 0.7 to make sure that the two replicas were running at the same temperature.

In order to keep consistent with the normal MD simulation with only one replica, the "extract_trajectory follow_dcd —replica 0 " command was used to follow one structure.

Also a harmonic cartesian constraints was applied to protein carbon alpha atoms with a displacement 0.35nm and force constant 250 KJ*mol^{-1}*nm^{-2}.

4.4. Results:

As we mentioned before, the purpose of stability test was to find out complex systems which their crystal conformations suffered least from the imperfectness of force field and solvent model during MD simulation and therefore could be stabilized at conformations that were similar to their crystal conformation.

Two typical results(1HEW and 1Z6S) are listed below. For simulation of all other systems, please refer to Appendix 8.1 "Testing the stability of crystal pose in force field".

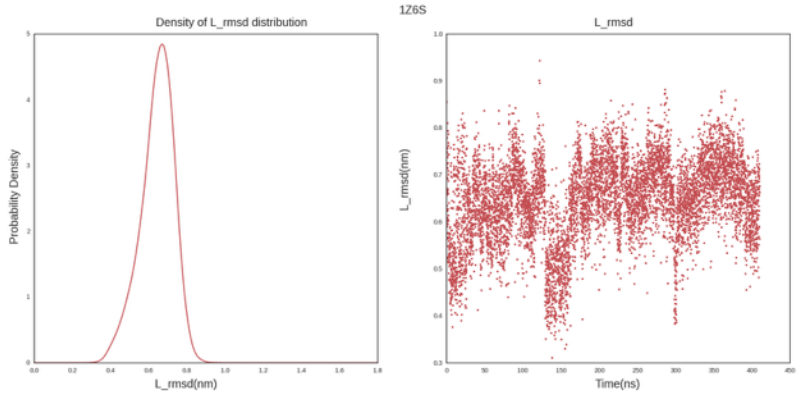

Figure. 4 Stability Test results of 1Z6S, simulation parameters are listed in Table VI, the left is the probability density plot of the L_rmsd distribution, the right one is ligand rmsd plot.

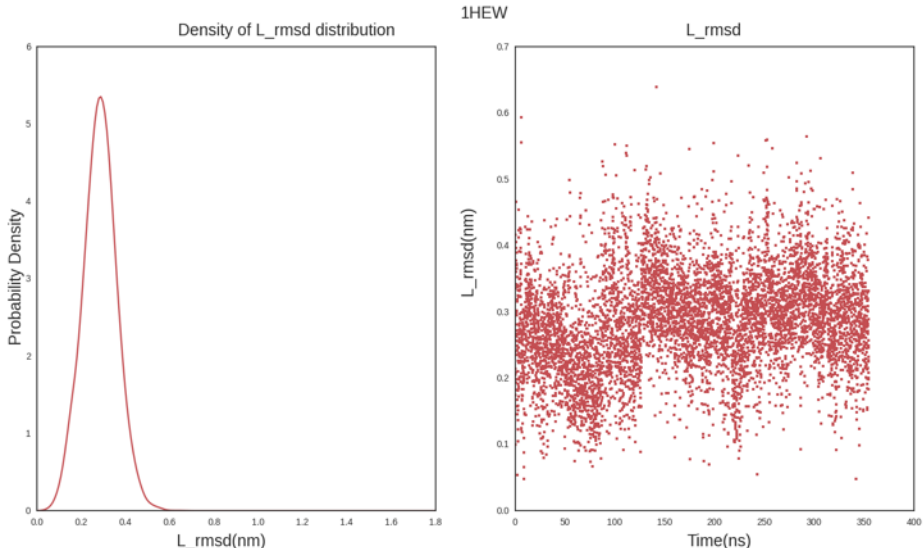

Figure 5. Stability Test results of 1HEW, simulation parameters are listed in **Table VI**, the left is the probability density plot of the L_rmsd distribution, the right one is ligand rmsd plot.

As we could see that, for systems like 1HEW, the ligand could stable at conformation closer than 4Å to crystal pose. These kind of systems are defined as stable system and our refinement will focus on those systems.

For systems like 1Z6S, the ligand stable at confirmation at about 7 Å away from the crystal pose. Virtually checking found out that the ligand was completely misplaced at the binding cavity. For those systems which couldn't stable at crystal conformation, we defined them as unstable systems. We didn't include them in our refinement targets.

After running the stability test and calculate the ligand RMSD with respect to crystal pose, we found out 8 stable systems which their ligand could stable at conformation closer than 4Å to crystal pose. All those systems were listed below:

Table VII: Stable systems

PDB Code	Protein
1LMO	Lysozyme
1LZB	Lysozyme
1HEW	Lysozyme
1TOW	Fatty acid-binding protein
1BB7	Lysozyme
1BB5	Lysozyme
1LZC	Lysozyme
1LZE	Lysozyme

4.5.Discussion

As we mentioned in previous section, the results was divided into two groups: Stable systems and unstable systems.

- **Unstable Systems:**

Those systems were 1HPX(**Fig.33**), 1HXB(**Fig.34**), 1UY8(**Fig.35**), 1UYC(**Fig.36**), 1UYF(**Fig.37**), 1UYH(**Fig.38**), 1Z6S(**Fig.4)**, 2QVD(**Fig.39**), 1SV9(**Fig.40)**, 1UYK**(Fig.41**).

All members of this group except 1UYC, 1UY8, 1UYK, 1UYH used ff12SB protein force field and GB-neck2 implicit solvent model. 1UYC, 1UY8, 1UYK, 1UYH used ff14SB protein force field[26] and GB-neck2 implicit solvent model[19]. The reason for using different protein force field was due to the poor performance of 1UYF under ff12SB and GB-neck2 implicit solvent model. All systems were using standard cartesian constraints method.

For those failed systems, 2 out of 10, 1Z6S and 1SV9, were highly charged. 2QVD and 1HXB were slightly negatively/positively charged. 6 out of 10, 1HPX, 1UY8, 1UYC, 1UYH, 1UYK, 1UYF, were neutral.

The diverse character of those failed system, which range from system with highly charged ligand to system with neutral ligand or protein family from lysozyme to HIV-I protease to Heat shock proteins, indicated that these failure might be cause by different reasons.

Reason 1: Overestimate charged ligand interaction with protein.

Rocklin et al[27] analyzed the binding free energy of charged ligand to a bioengineered protein and found that scaling net charge of ligand and a key residue aspartate in protein to smaller, non-integral values helps decreasing the root-mean-square error for binding affinity prediction. They argued that "current force field have been parameterized to match pure liquid properties and gas-phase electrostatic potentials". This might cause the overestimation the charged ligand interaction with protein due to the lack of polarization effect when ligand binds to receptor. Virtual checking the trajectories of 1Z6S and 1SV9 turned out that the highly negatively charged phosphate group of 1Z6S and carboxyl group of 1SV9 were misplaced to other positively charge residues during the simulation.

However, the unique shape and chemical environment of the 1SV9's binding cavity may also contribute to the its unstable condition. The Leu2 and Ile19 formed a hydrophobic channel which accommodated the aromatic rings of ligand diclofenac in 1SV9. The clash of the hydrophobic channel, which caused by the twisting of the sidechain of Leu2 and Ile19 toward inner side of protein, closed the entrance to the binding cavity and may caused the improper placement of ligand around binding cavity.

Reason 2: Poor reproducible the 'π-π stacking' interaction due to lack of polarization effect.

As for 1Z6S, crystallographic study[28] showed that there is a π-π stacking interaction between the five member ring of adenine in the adenine monophate(AMP) and His 119 of ribonuclease pancreatic A(RNase A) in the crystal structure. Homology study demonstrated that His 119 was preserved among other complexes in the same protein family. This meant that "His 119" played a key role in ligand binding process. However, this specific interaction didn't preserve during the stability test.

Several studies suggested that lack of polarization effect in the force field might contribute to it. Hunter and Sanders' study[29] suggested a polarized pi system theory which provide qualitatively understanding and predicting aromatic-aromatic interactions. They argued that the dislocation of electrons in the aromatic ring forms a quadrupole moment with two partial negatively charge 'electron cloud' above/below the aromatic ring and two partial positively charge around periphery. Then based on this theory, Martinez, C. R.[30] argued that the increased electronic polarizability in aromatic pi system was most often thought to be responsible for 'π-π interaction'. Due to the intrinsic limitation of fixed charged model and the training method of current force field, either the ligand or the protein is fully polarizable. The lack of polarizability of protein or ligand atoms in the simulation might cause the misplacement of AMP in 1Z6S and the deviation of AMP from its crystal pose during the simulation.

Reason 3: Underestimate the hydrogen bond interaction.

Both the charge ligand in 1Z6S[28], 1SV9[31], 2QVD[32] and the neutral ligand in 1HXB[33], 1UY8[34],1UYC[34],1UYH[34],1UYK[34] formed directly hydrogen bonds or water-mediated hydrogen bonds with protein receptor. For example, in 1Z6S complexes, the charged phosphate group in AMP interacted with the Gln 11 and His 12 through hydrogen bonds. In 1SV9 systems[31], the highly charged carboxyl group stabilized at crystal conformation through three hydrogen bonds: one directly with His48 N and the other two through water48 O and water54 O. Moreover in 2QVD[31], One of the two etherial oxygen atoms of berberine form ed hydrogen bond with the hydrogen attached to the main chain nitrogen of Gly 30, the same oxygen atom also made a water mediated hydrogen bond with His 48. Those hydrogen bonds played an important role in ligand binding process.

While Paton et al[34] study demonstrated that the AMBER force field , which doesn't explicitly include the hydrogen bond interaction, underestimated the stabilization of hydrogen bonded complexes. This kind of effect we believed could explain most of the failure in the stability test. Since all those failure systems were include either directly hydrogen bond or water-mediated hydrogen bonds, our research group was currently working on implementation of TIP3P explicit solvent model into MELD. we believed that some of those unstable systems, which caused by excluding hydrogen bond interactions, could also be analyzed by MELD in the future .

Reason 4: Steric effect of crystal water.

As for 2QVD, we noticed that the berberine inserted deeper into the binding cavity and occupied the place where the crystal water located. For 1HXB,1UY8, 1UYC, 1UYH, 1UYK, the binding cavity was also filled with crystal water in the crystal structure. We thought that treating water implicitly through GB-neck2 implicit model provide more space for ligand wiggling around in the binding cavity in our simulation. Therefore we believed that introducing the explicit solvent model might help those complexes stabilize at crystal structure.

- **Stable Systems:**

This Group was consisted of complexes which their ligands could be stabilized at conformations no more than 4.0Å away from the crystal pose most of the time in the trajectory.These systems were 1HEW(**Fig.6**), 1BB5(**Fig.24**), 1BB7(**Fig.25**), 1TOW(**Fig.26**), 1LMO(**Fig.27**), 1LZB(**Fig.28**), 1LZC(**Fig.29**), 1LZE(**Fig.30**), 1JLD(**Fig.31**), 9HVP(**Fig.32**).

All members used ff12SB protein force field and GB-neck2 implicit solvent model during stability test. All members except 9HVP were applied standard cartesian constraint to alpha

carbon atoms of the protein backbone. 9HVP used cartesian constraint to all heavy atoms(N,CA,C,O) of the backbone. The reason for using new cartesian constraint to 9HVP was that this system couldn't stabilize at crystal-like structure under standard cartesian constraint method.

7 out of 10 complexes in stable systems had protein coming from the lysozyme family. 1 complex was fatty acid-binding protein and 2 complexes were HIV-I protease. The ligand charge condition were almost the same. The Lysozyme family protein and HIV-I protease protein had neutral ligand with zero net charge. While the fatty acid binding protein binds with negatively charged carbazole butanoic acid(-1).

However, not all stable systems were suitable for validate our sampling tools MELD. For 9HVP and 1JLD, both computational and experimental studies had shown that the HIV-protease underwent a large conformation change during the binding process. While our currently protocol for replica exchange simulation required cartesian constraint on the carbon alpha of the protein backbone atoms. Moreover the 9HVP system need additional protein backbone cartesian constraints to keep it stable during the stability test. Both of them introduced more difficulty to sample the crystal conformation. It required novel constraints strategy to balance the rigidity request of HIV protease for stabilizing the ligand at the final crystal pose and the flexible demand of HIV-I protease during the binding process. Therefore it was reasonable to exclude 9HVP and 1JLD and focus on validating MELD on the remaining complex systems in the stable group.

In summary, the polarization effect, hydrogen bond and crystal water contributed a lot to the instability of complex with either highly charged or neutral ligand. The overestimated electrostatic interaction which caused by the lack of polarizability of protein and ligand atoms drove the phosphate group of AMP to the place where it shouldn't be. The lack of polarizability of protein and ligand atoms might cause the reproducible failure of the π-π stacking structure in 1SV9 complex. The underestimated hydrogen bond interaction and the lack of steric effect of the crystal water might be the common problems for all failure case of stability test. Therefore the force field might need to introduce new parameters to account the polarizability of atoms or new training environment to mimic the induced polarize of atoms which involved in the binding process. However, no force field or solvent model was perfect. We couldn't wait for the coming of the perfect force field to validate our sampling method. Therefore in order to minimize the influence from the imperfectness of force field and solvent model, we limited the validation of our sampling method to those systems which could be stabilized under the current best protein force field ff12SB and solvent model Gb-neck2.

The 3D structure of the stable systems which we used for MELD simulation were generated using Chimera[20](**Appendix 8.2** "3D structure of stable system" **Fig.42**).

5.Constraint generating protocols

In the previous section, we find target systems based on computational power of our available resources, complexity of receptor-ligand system. Then we designed a so-called “stability test” to find our the systems which suffered the least from the imperfectness of force field and implicit solvent model. Therefore these systems’ crystal structure could be stabilized at their crystal conformation.

In this section, we aim to design method which could cluster DOCK poses and properly translate their structural information into MELD constraints.

In practice,this designing problem could be split into two parts :

- How to translate structural information into MELD constraints?
- How to cluster DOCK poses?

Three protocols were designed and tested to solve these two problems.

5.1.Manually choose DOCK poses based on ligand RMSD

• Motivation:

In order to isolate the problem of clustering DOCK poses and focus on solving the problem of how to translating structural information of DOCK poses into MELD constraints, the “Manually choose DOCK poses based on ligand rmsd” method was first designed.

• Technical Details:

We used the poses’ RMSD with respect to crystal structure as the principle to select four poses in such way that their ligand RMSD varies from 2Å to 8Å.

Then a contact cut-off 3.5Å was chose to search for all contact heavy atoms between protein and ligand. The atom indices of contact atom pairs were recoded. Duplicated pairs were removed and the remaining ones were stored into a constraint atom pair list.

A harmonic distance MELD constraint was applied to all atom pairs in the constraint atom pair list. Then a MELD constraint group was created to store all those constraints. The constraint details and pose summary are listed in **Table VIII** and Table **IX**.

Table VIII: Constraint manage strategy

Groups	1
Collections	1

Table IX: Pose summary

	Enforced	Pose1	pose2	pose3	pose4	Total Steps	zero constrain energy frame
1BB7	20%	4.7Å (#100)	6.6Å (#87)	7.2Å (#46)	8.0Å (#104)	3150	82%
1HEW	30%	3.1Å (#77)	6.1Å (#75)	8.0Å (#55)	4.4Å (#132)	2898	99%
1LMO	20%	8.6Å (#2)	8.3Å (#3)	8.6Å (#4)	7.2Å (#5)	8589	79%
1LZB	20%	2.3Å (#42)	3.1Å (#28)	4.1Å (#57)	4.8Å (#33)	8998	54%
1LZC	20%	2.7Å (#18)	5.7Å (#75)	6.0Å (#103)	8.3Å (#105)	2199	86%
1LZE	20%	5.9Å (#12)	4.1Å (#13)	5.4Å (#59)	7.0Å (#114)	11498	72%
1BB5	20%	6.1Å (#76)	2.7Å (#44)	6.1Å (#20)	7.1Å (#73)	3534	93%
1TOW	30%	3.8Å (#5)	4.5 Å (#22)	5.1Å (#8)	6.1Å (#25)	4398	100%

- ## Pros and Cons:

Pros: Intuitive and simple;

Cons: Need prior knowledge of poses' rmsd with respect to crystal structure which means that we need to have crystal structure.

- ## Results:

The MELD simulation results of "Manually choose DOCK poses based on ligand RMSD" are listed below.

For each complex systems, six graphs were plotted. Five of them were Population of ligand RMSD(with respect to poses which used for generating constraints) at the lowest temperature replica(300K). The bottom right graph shown the density plot of each population plot. Pose details about those poses are listed in **Table IX**.

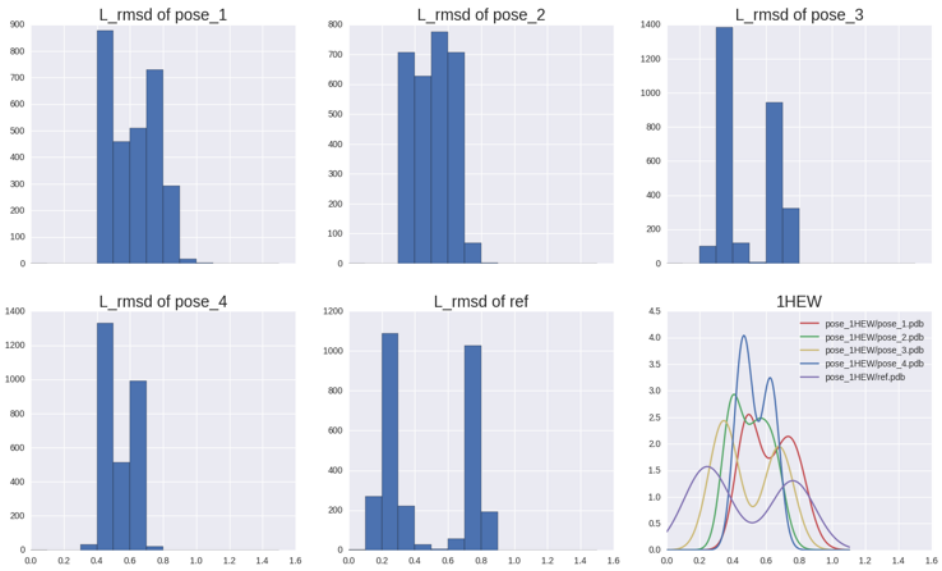

Figure 6. MELD simulation results of 1HEW, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

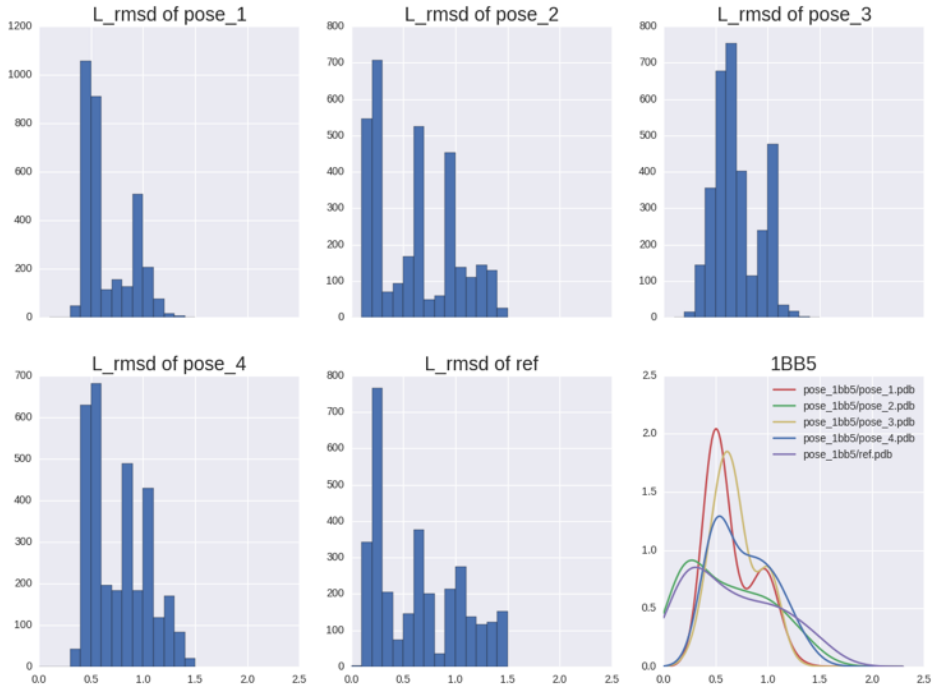

Figure 7. MELD simulation results of 1BB5, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

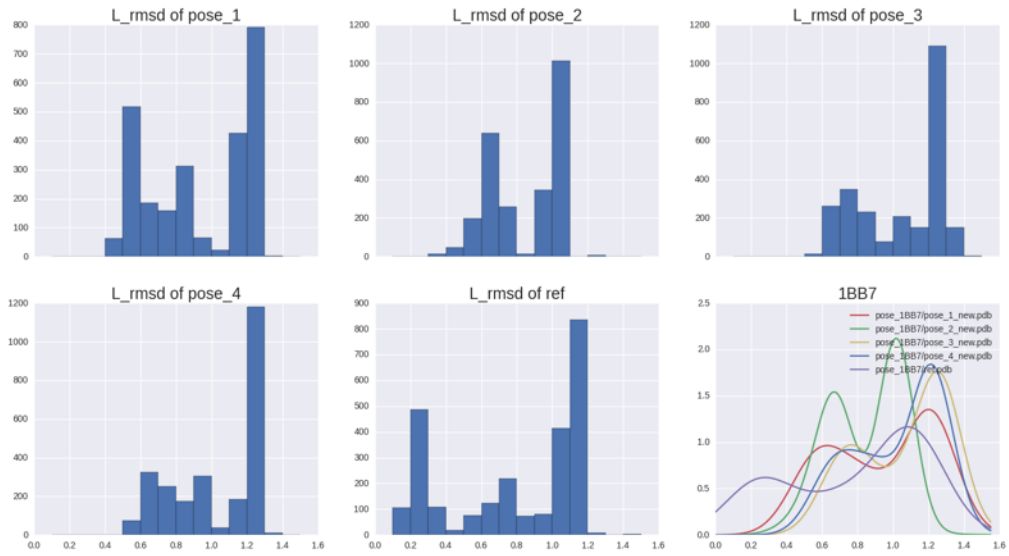

Figure 8. MELD simulation results of 1BB7, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

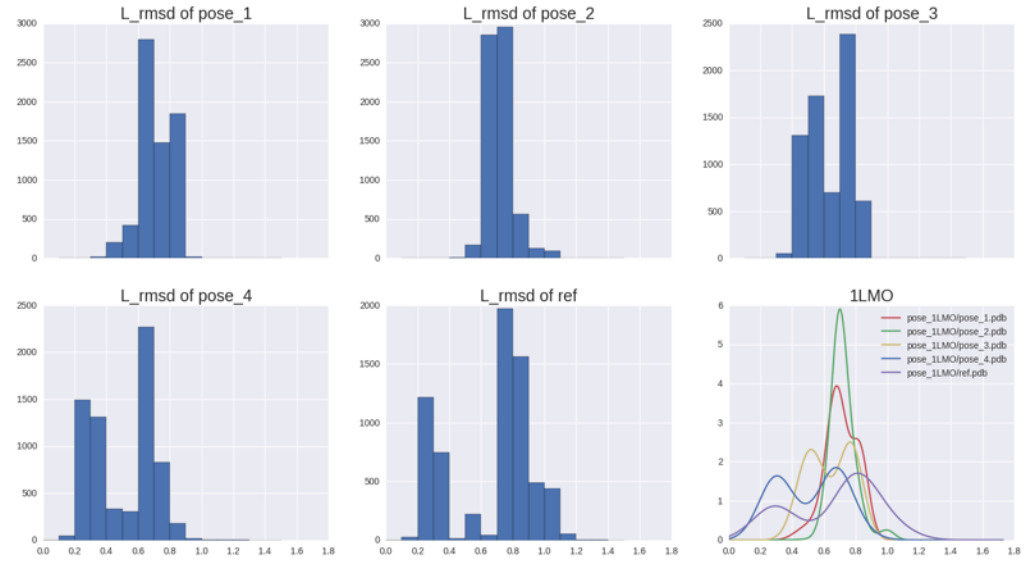

Figure 9. MELD simulation results of 1LMO, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

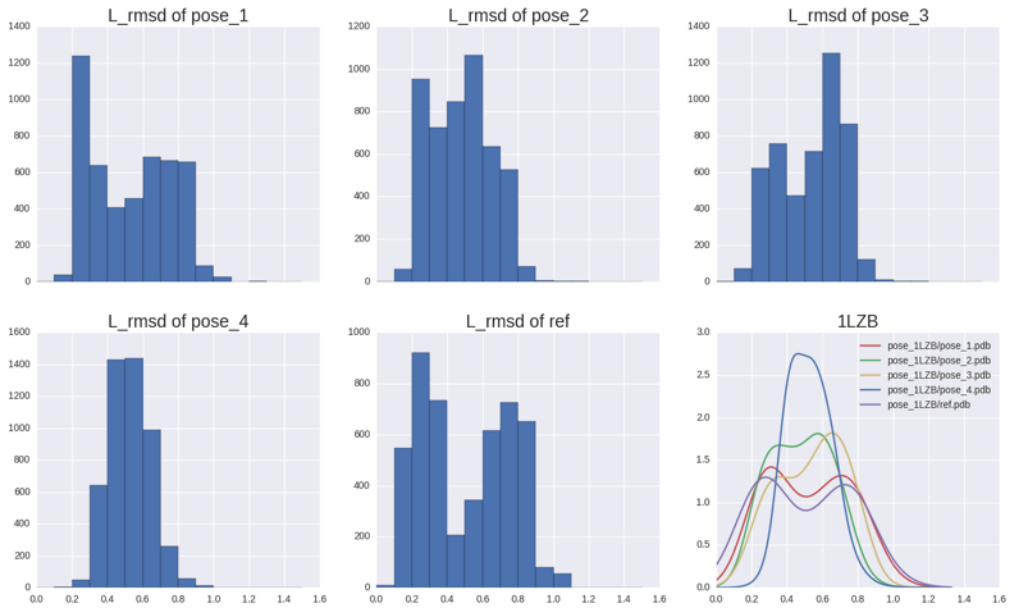

Figure 10. MELD simulation results of 1LZB, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

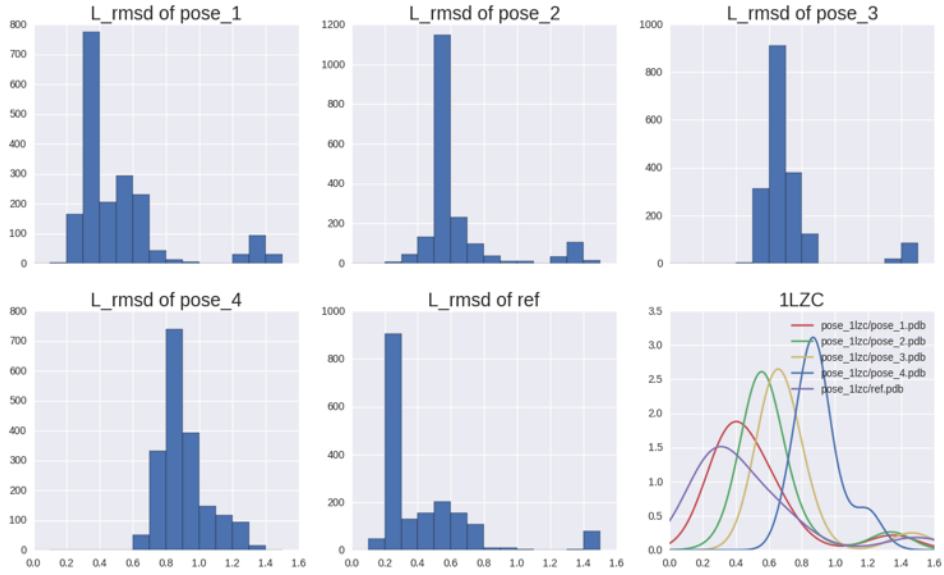

Figure 11. MELD simulation results of 1LZC, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

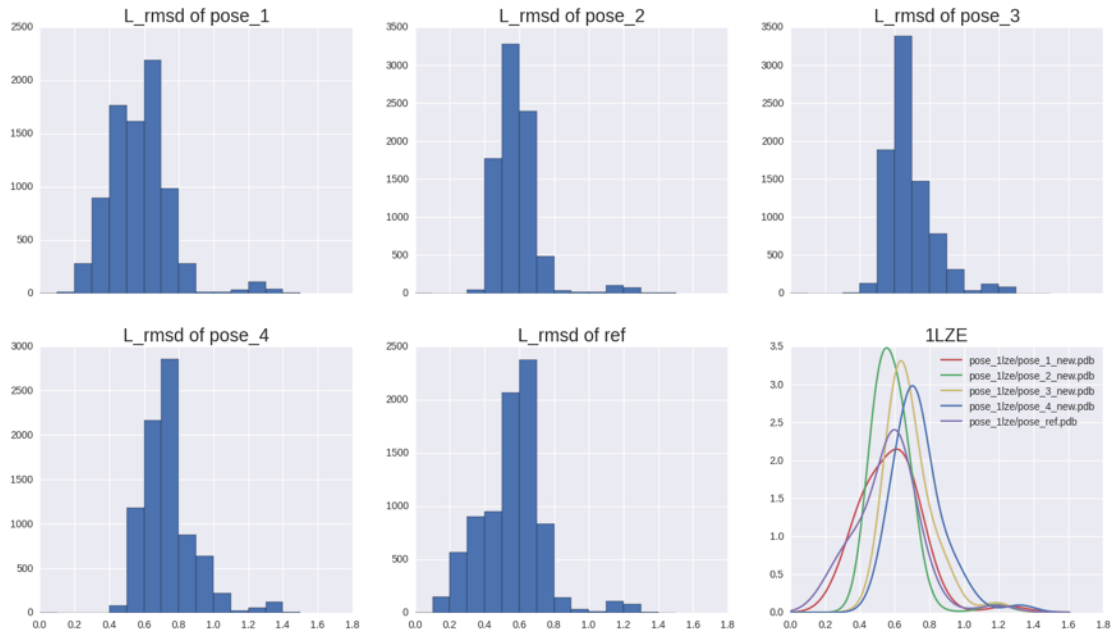

Figure 12. MELD simulation results of 1LZE, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

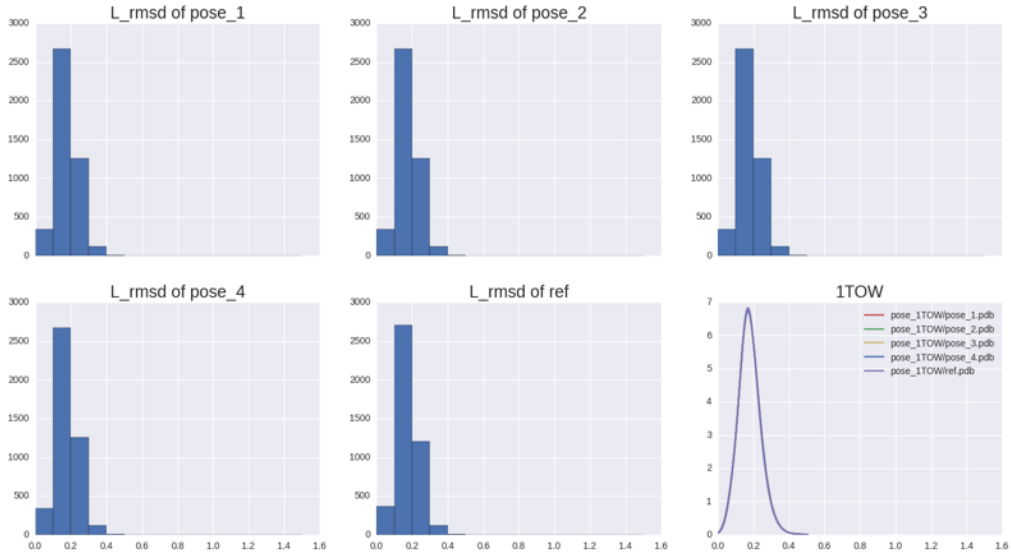

Figure 13. MELD simulation results of 1TOW, five population graphs of ligand RMSD with respect to corresponding poses and one density plot of five population graph are given.

- **Discussion:**

I. MELD could find crystal pose on the ligand-protein interaction landscape

As we could see from Fig.6 to Fig.13, there are lots of frames which the ligand has a RMSD value closer than 3Å to crystal pose.

If we chose 3Å as the cut-off distance to distinguish whether the ligand was sampling at certain pose or not, we could see that the ligand in all system sampled around the crystal poses.

II. MELD could distinguish crystal pose from other DOCK poses.

Using the same criterion of defining whether ligand samples at certain pose or not, we decided to compare the amount of frames which sampled at DOCK poses with those sampled at crystal pose. Qualitatively checking the bars which represented the frames had ligand RMSD closer than 3Å to either DOCK poses or crystal, we could find out that the MELD sampled more often at crystal pose than other DOCK poses. The only exception was 1TOW, 1BB5, 1LZB.

Pose1 in 1LZB, Pose2 in 1BB5, and all four DOCK poses in 1TOW has almost the same sampling frequency. It means that the MELD simulation couldn't distinguish those poses from the crystal pose.

The possible reason for MELD couldn't distinguish these pose might be these pose are thermodynamically equivalent to their corresponding crystal pose. Evidence comes from the stability test results of these system. The thermo-fluctuation region of crystal pose in 1LZB and 1BB5 is 1~4Å in terms of ligand RMSD to their starting conformation. The ligand RMSD of pose1(2.3Å, **TableIX**) in 1LZB and pose2(2.7Å, **TableIX**) in 1BB5 are within this fluctuation region. For 1TOW, all 4 poses are within crystal pose's thermo-fluctuation region(2~6Å). It means that these poses are equivalent to their corresponding crystal poses. Therefore, it won't be surprise to find out that the MELD simulation couldn't distinguish them.

III. No blindly Prediction

This manually choose four DOCK poses method depends on the prior knowledge about the ligand RMSD of each DOCK poses verse crystal pose.

Since we wouldn't like to be able to blindly select DOCK poses, we decided to design another method which could automatically algorithm to cluster the DOCK output and generate poses for MELD.

.

5.2. Automatically cluster and manually choose DOCK poses based on ligand RMSD

- **Motivation:**

The first method "Manually choose DOCK poses based on ligand RMSD" required prior knowledge about the crystal structure. This largely limited its usage in practice.

In order to be able to blindly cluster the DOCK poses, the "automatically cluster and manually choose" method was designed.

- **Technical Details:**

A. Ligand characterization:

Without using each pose's relative RMSD to crystal structure, we designed another quantity, which combined the center of mass and the moment of inertia, to characterize the difference among poses.

Our idea came from the rigid body orientation in physics. In order to describe the state of a rigid body in 3D space, it required six degree of freedom: three cartesian coordinates to describe the exact location and three corresponding axis of inertia tensor to describe its orientation.

Recalled from basic physics[36] that the moment of inertia determines the corresponding torque for a desired angular acceleration about a rotational axis.

$$\tau = I * \alpha \qquad \textbf{(Eq.7)}$$

where τ is the torque, I is the moment of inertia and α is the angular acceleration.

Thus the corresponding rotational axis could be used to describe the orientation of the rigid body.

In order to keep a mathematical strict and consistent way to generate those rotational axis, we decided to use the principal axis of rigid body.

The principal axis coordinate system was the system such that the product terms of the inertia tensor matrix are zero. The existence and the uniqueness of principal axis have been proved.

Mathematically, the problem of determining the principal axis and corresponding moment of inertia was equivalent to find the eigenvector and eigenvalue of the inertia tensor matrix constructed from a arbitrary reference coordinate system[36]

$$\left(I_{reference} - I * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) * \begin{pmatrix} \omega_x \\ \omega_y \\ \omega_z \end{pmatrix} = 0 \quad \textbf{(Eq.8)}$$

where the $I_{reference}$ was the inertia tenser matrix in reference coordinate, ωx , ωy, ωz where represented the x,y,z directions of the principal coordinates were the eigenvector of $I_{reference}$ and the coordinating moment of inertia was the eigenvalue.

In practice, we treated the ligand molecule as rigid body by ignoring all its rotational bonds and used a hybrid quantity, which combined the center of mass and principle axis, to characterize DOCK poses' relative difference.

The exact definition of this hybrid quantity was:

$$\mathrm{M_{i,j}} = \arccos\left(\frac{\overrightarrow{\mathrm{v_i}} \cdot \overrightarrow{\mathrm{v_j}}}{\left|\overrightarrow{\mathrm{v_i}}\right|\left|\overrightarrow{\mathrm{v_j}}\right|}\right) + \left|\mathrm{R_i} - \mathrm{R_j}\right| \quad \textbf{(Eq.9)}$$

The first term represents the relative orientation of the ligand which defined by the cosine angle between the first principal axis. The reason we chose the first principle axis was due to the shape of the ligand and the binding cavity. The second term represents the relative location of the ligand in three dimensional cartesian space which defined by the center of mass distance difference between poses.

Using this hybrid quantity, we could generate a covariance matrix(M) that quantitatively determined the pairwise difference of DOCK poses.

After generating the covariance matrix(M), we used hierarchical clustering algorithm to cluster the DOCK poses.

B.Hierarchical clustering algorithm

Previously, we generated the covariance matrix(M) where each element (M_{ij})represented the relative distance between pose i and pose j.

In this section, the "average" criterion of Hierarchical clustering method was used to cluster DOCK poses.

In Hierarchical clustering algorithm, the algorithm first calculated and compared the "average distance" between all elements of matrix M. The pair which had the shortest distance were merged to form new cluster. The definition of average distance was:

$$\mathrm{D}(r,s) = \frac{\mathrm{T}_{rs}}{(\mathrm{N}_r * \mathrm{N}_s)} \qquad \textbf{(Eq.10)}$$

T_{rs} was the sum of all pairwise distances between cluster r and cluster s. N_r and N_s were the sizes of the cluster r and s, respectively.

Then in the following iteration steps, new members merging to the existing cluster or merging of two existing clusters were based on the same criterion until all members were merged into one big cluster.

After the clustering process completed, a linkage matrix, which described all the merging operation during the clustering process, was returned. Based on empirical data, 8~12 was the optimized cluster amount. A pose from each cluster was used to generate MELD constraints.

The clustering outcomes for each refinement systems are listed below:

Table X: Clustering outcomes

	Enforced	Constraints Number	Cluster Numbers	Total Steps	zero constrain energy fraction
1LMO	10%	155	11	2998	89%
1LZB	10%	127	9	2998	95%
1LZC	10%	200	9	2998	98%
1LZE	10%	247	12	3198	75%
1HEW	10%	212	12	2998	99%
1TOW	10%	67	11	2098	98%
1BB5	10%	332	10	3498	99%
1BB7	10%	132	8	2998	95%

C.Translate DOCK poses structural information into MELD constraints

In previous step, we used hierarchical clustering algorithm to cluster the DOCK poses and a empirical cut-off to determine the optimized number of clusters. One pose per cluster was selected and the heavy atom contact pairs among those poses were generated using the same criterion as the first method.

A slightly difference criterion, which aimed to further reduce the redundancy of constraints, was used to generate the contact atom pairs.

We defined the neighbor of a host constraint as the constraints which had at least one atom that was covalently bonded to atom in host constraint. We further assumed that due to the relative smaller geometric distance between host and neighbor constraints, they were equivalent and simultaneously satisfied during the MELD simulation. Thus it was reasonable to delete those neighbor constraints from the whole constraint list to reduce the redundancy.

In practice, this was done in three steps. First, a N*2 python numpy array, called "topology array", which stored all covalently connected atoms' indices, was generated based on the topology of the DOCK pose. Each row of this array represented the covalently connected atom pairs' indices. Second, all host and neighbor constraint pairs from contact map list, which had at least two covalently connected atoms, were selected through comparing with the topology array. Third, the host constraints, which had the smallest atom indices, were saved and all their neighbor were disregarded. In the end, all those host constraints were assigned MELD constraints.

These constraints were collected into one constraint group and one MELD collection was created to store this constraint groups.

- ## Pros and Cons:

Pros: Blindly cluster the DOCK poses without prior knowledge about crystal structure.

Cons: Since we ignored the internal rotational degree of freedom of the rotational bonds, the robustness of this method was poor. Less strict in term of determining optimized cluster amount.

- ## Results:

The MELD simulation results of "Automatically cluster and manually choose DOCK poses" method are listed below. For each complex systems, a histogram of ligand RMSD with respect to crystal pose was plotted.

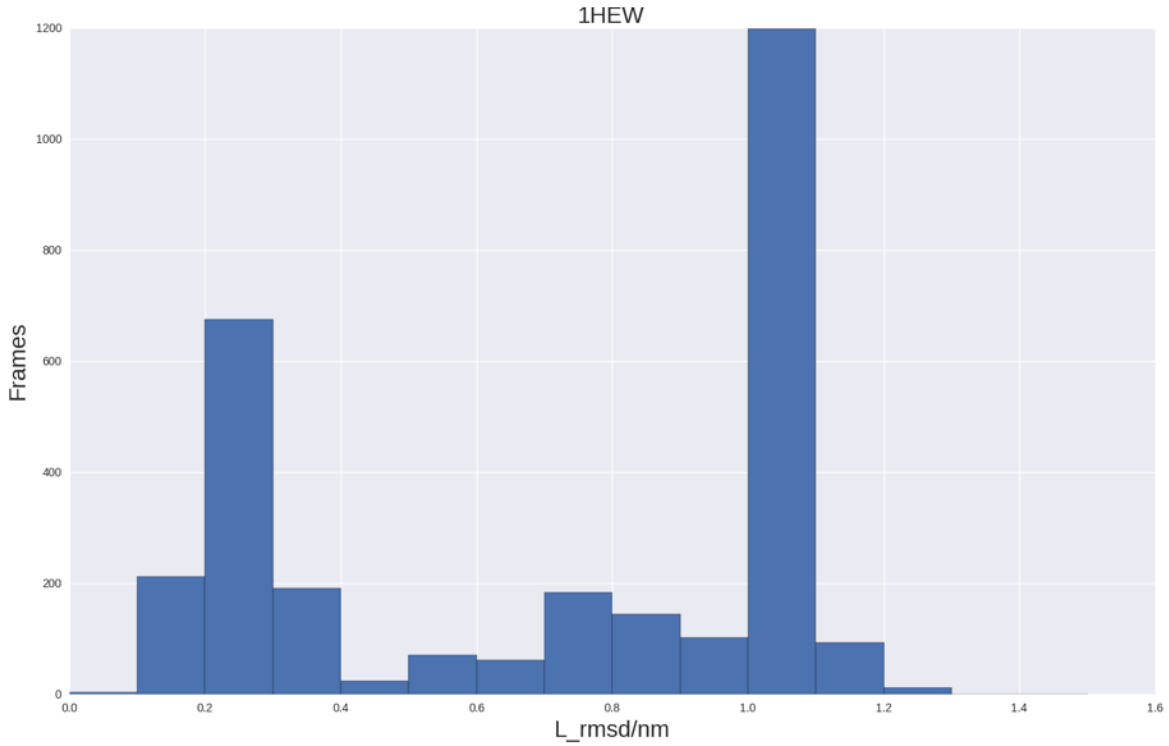

Figure 14. MELD simulation results of 1HEW with 12 DOCK poses;

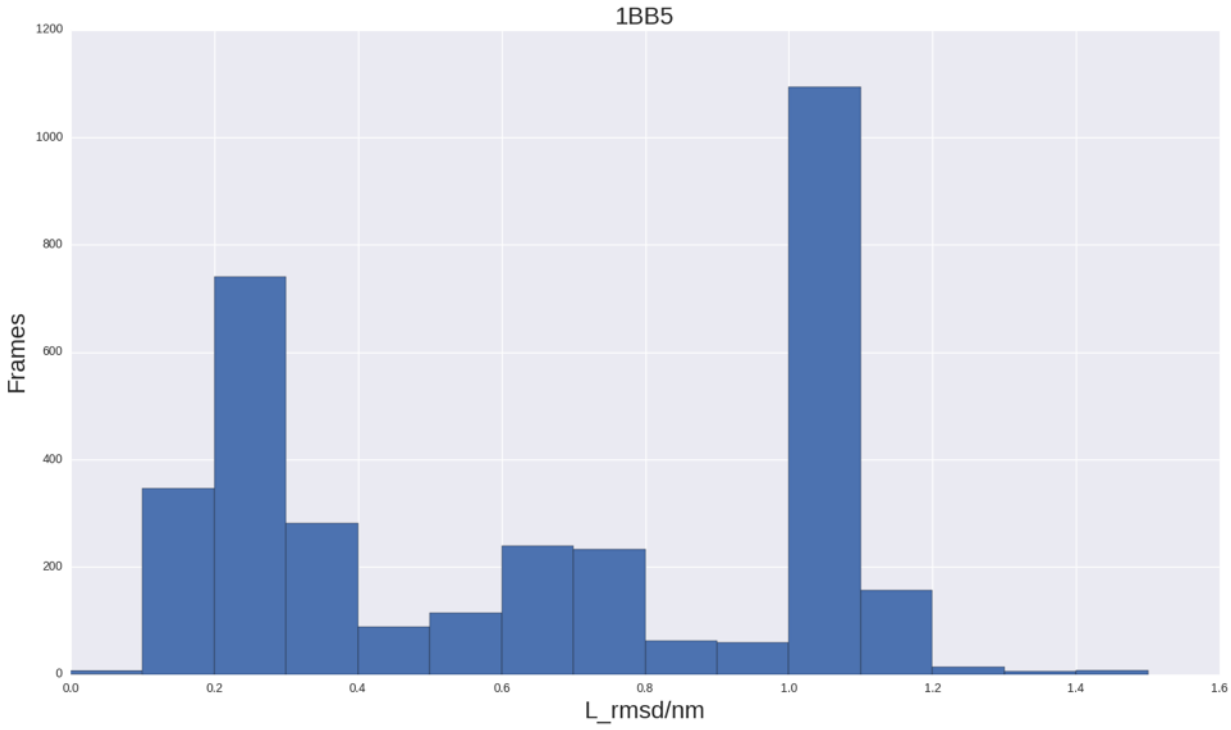

Figure 15. MELD simulation results of 1BB5 with 10 DOCK poses;

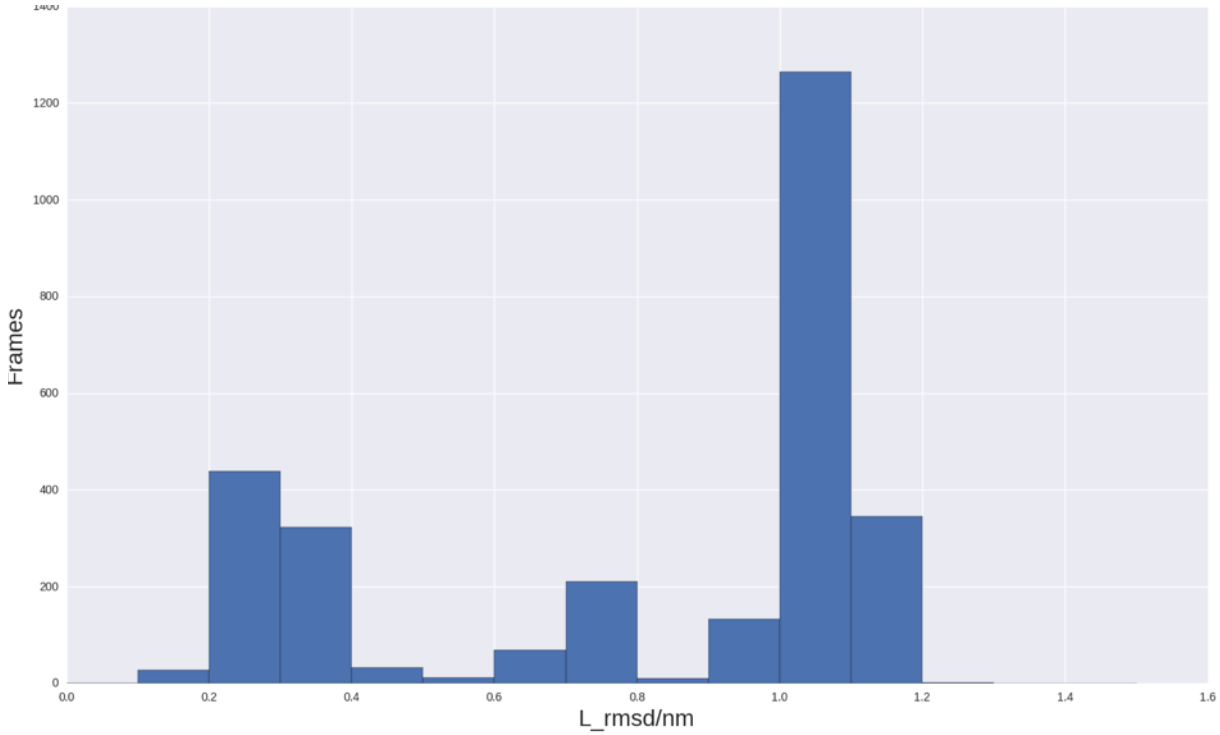

Figure 16. MELD simulation results of 1BB7 with 8 DOCK poses;

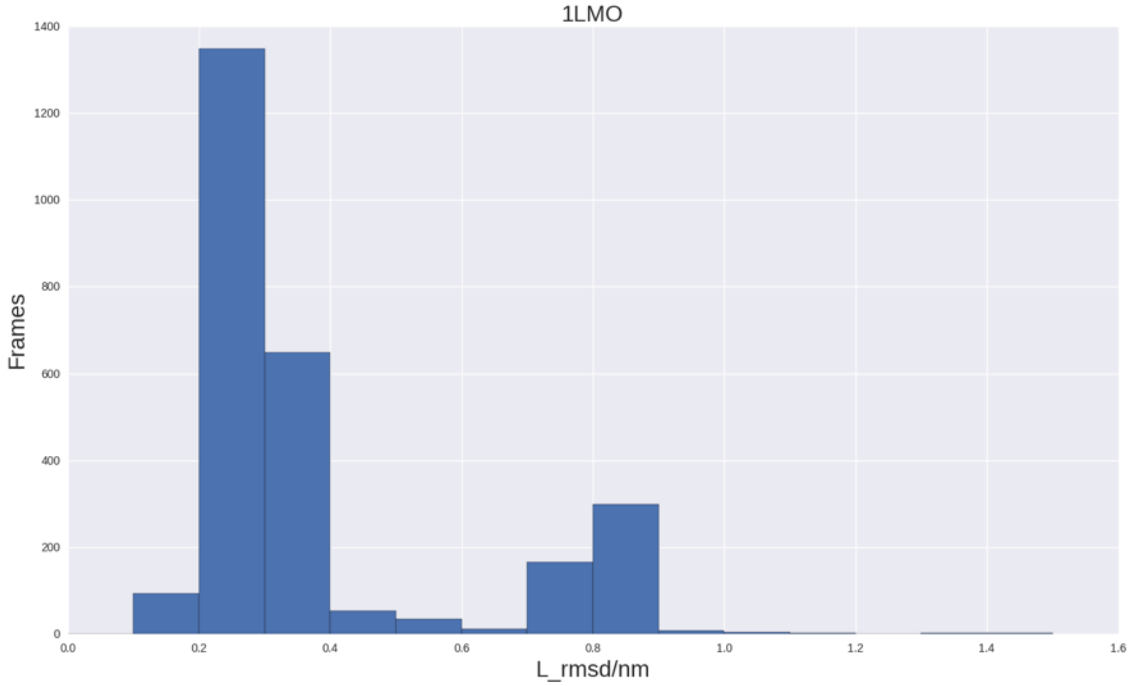

Figure 17. MELD simulation results of 1LMO with 11 DOCK poses

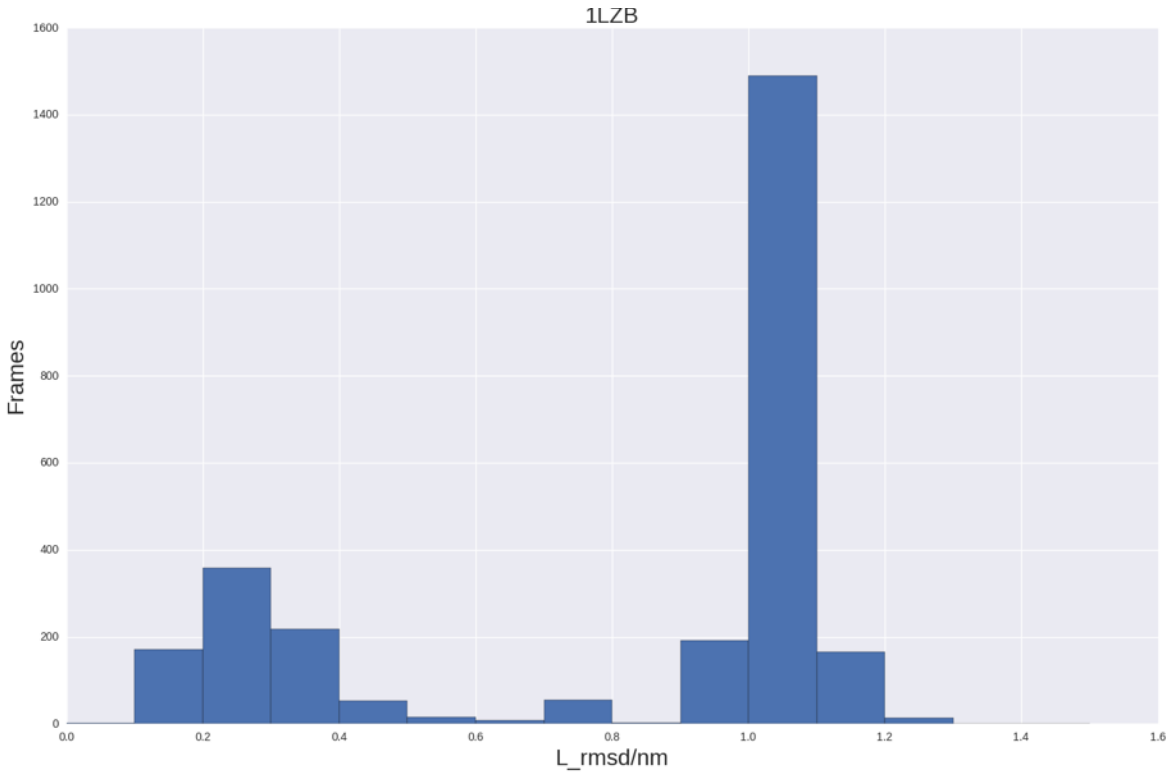

Figure 18. MELD simulation results of 1LZB with 9 DOCK poses;

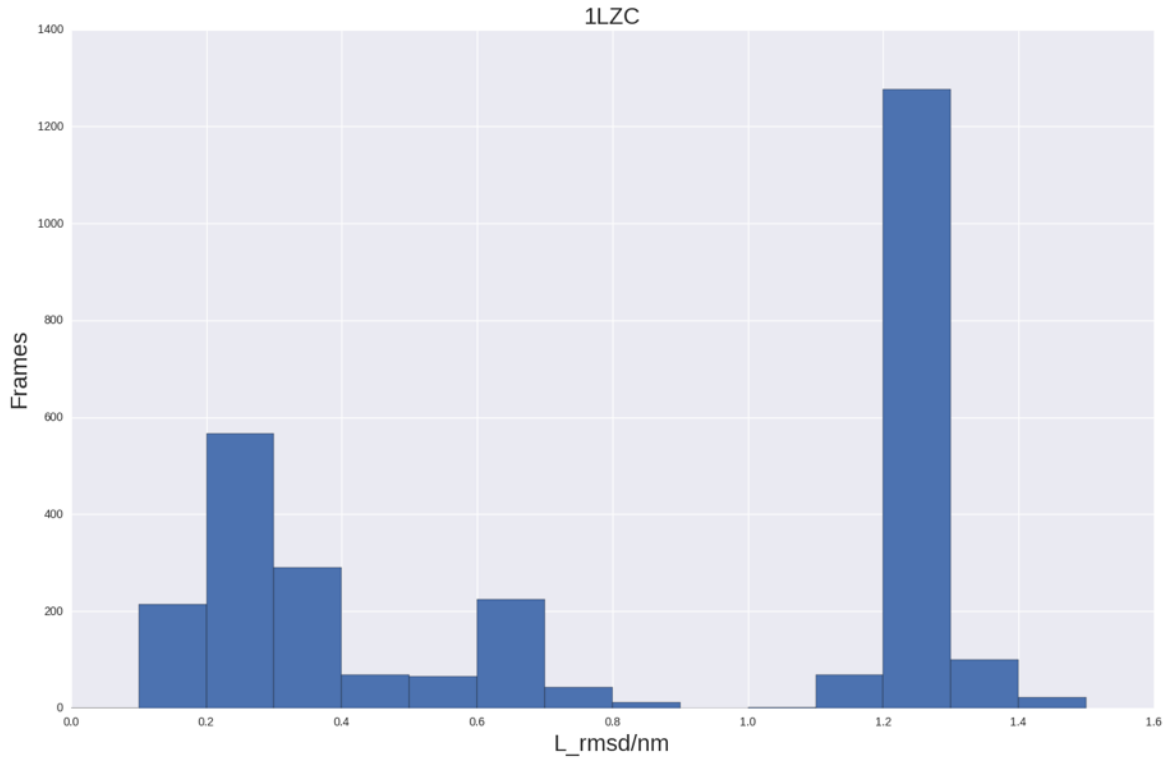

Figure 19. MELD simulation results of 1LZC with 9 DOCK poses

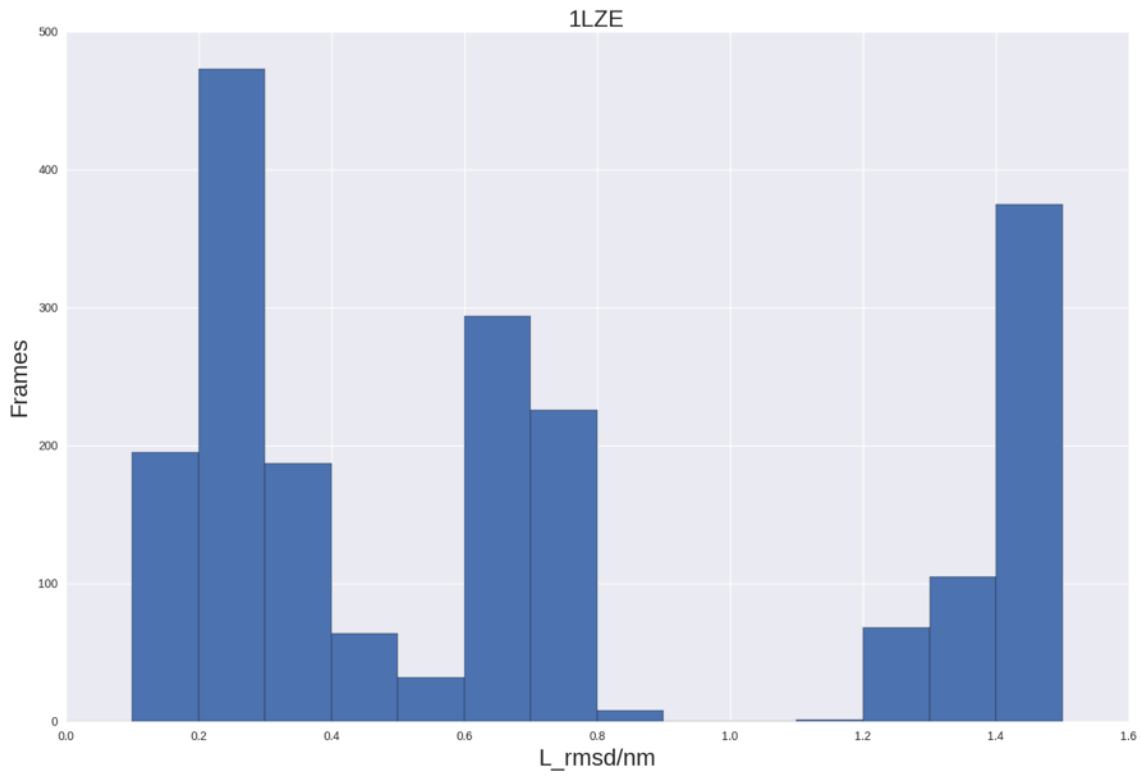

Figure 20. MELD simulation results of 1LZE with 12 DOCK poses

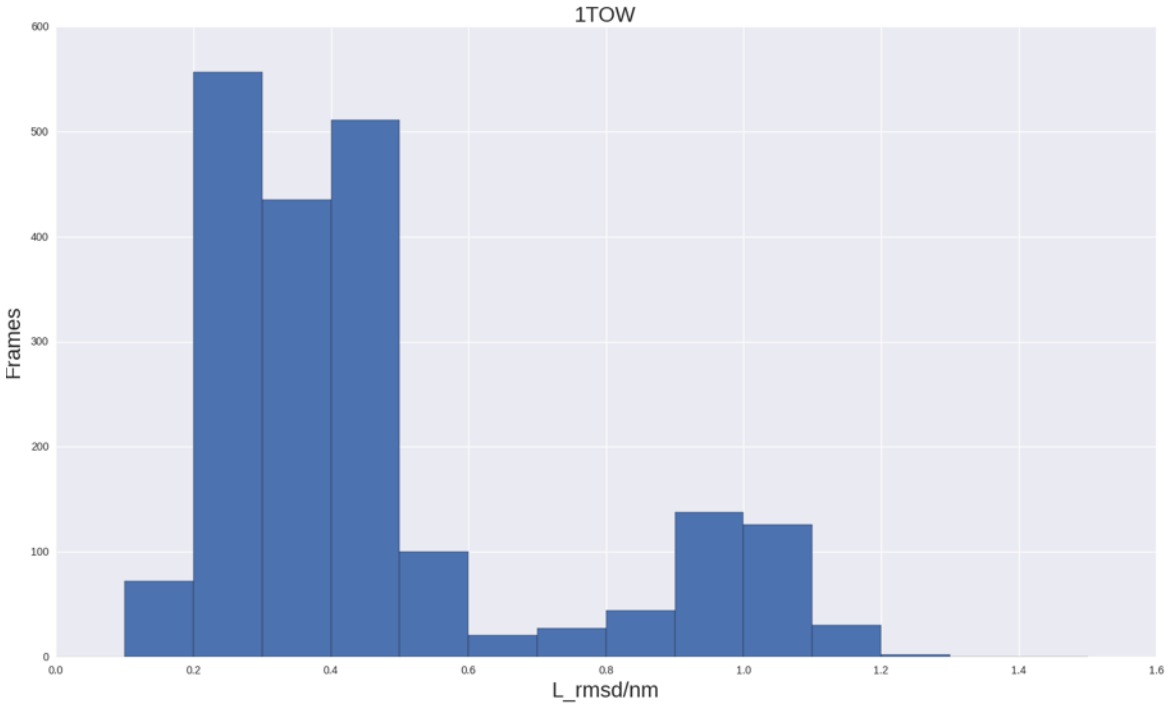

Figure 21. MELD simulation results of 1TOW with 11 DOCK poses

- **Discussion:**

We could see this time, most simulation still find the crystal poses, but the frequency was much less(**Fig.14-Fig.21**). The possible reasons for this were first the total constraint atom pairs were much larger. The 'good' information, which is the constraints appears in crystal-like poses, were much less. The dilution effect which caused by introducing more incorrect information into MELD constraint group largely increased the sampling uncertainty, therefore the crystal pose was no longer appears often in the trajectory. Second we included more DOCK poses. Maybe there were some local minimums which were relative more stable. It will require longer simulation time to get away from it and sample at crystal pose. Third the method of combing center of mass and moment of inertia didn't work as we expected. Sometime, the method couldn't even cluster all crystal-like poses into one cluster. This means that we couldn't guarantee that we could always include the constraint information from the crystal-like poses into our MELD constraints.

In order to better enrich the crystal-like poses, we tested another constraint method.

5.3. Automatically cluster and choose DOCK poses based on Kelley-Gardner-Sutcliffe penalty function

- **Motivation:**

The second method had the problem of poor robustness and lack of strictness in term of optimized cluster number determination. Thus we designed and tested a new method which could automatically determine the cluster amount and had more robustness.

- **Technical Details:**

A. Ligand characterization

Instead of introducing much assumption and estimation to describe the ligand orientation, we simply used the relative root-mean-square-deviation(rmsd) difference among poses to describe their relative difference.

By definition, the relative rmsd of pose i and j were the rmsd between i and j pose without performing any translational and rotational operation.

$$\mathrm{RMSD} = \sqrt{\frac{1}{\mathrm{N}} \sum_{i=1}^{\mathrm{N}} \delta_{\mathrm{i}}^{2}} \qquad \textbf{(Eq.11)}$$

N is the number of atoms involved in rmsd calculation, δ_i is the corresponding atoms' distance between pose i and pose j.

Then using the relative RMSD among poses, a covariance matrix(M) which described the pairwise difference among top30 poses were generated.

B.Hierarchical clustering algorithm with Kelley-Gardner-Sutcliffe penalty function

Like in the second method, a same hierarchical clustering algorithm was used. Then a linkage matrix was returned.

However, instead of empirical determining the optimized cluster number, a strict Kelley-Gardner-Sutcliffe(KGS) penalty function[36,37] was used. This idea came from the observation that in general, the more new members included in a cluster, the lower the homology it was. While less complexity of the whole dataset was due to the reducing of cluster numbers in it.

$$\mathrm{spread_m} = \frac{\sum_{\mathrm{k=1}}^{\mathrm{J}} \sum_{\mathrm{i=1,i<k}}^{\mathrm{J}} \mathrm{d(i,k)}}{\mathrm{J(J\text{-}1)/2}} \qquad 12.1$$

$$\mathrm{AvSp_i} = \frac{\sum_{\mathrm{m=1}}^{\mathrm{cnum_i}} \mathrm{spread_m}}{\mathrm{K_i}} \qquad 12.2 \qquad \textbf{(Eq.12)}$$

$$\mathrm{AvSp(norm)_i} = 2 * \left[\frac{\mathrm{N\text{-}2}}{\mathrm{Max(AvSp) - Min(AvSp)}} (\mathrm{AvSp_i - Min(AvSp)}) + 1 \right] \qquad 12.3$$

$$\mathrm{P_i} = \mathrm{AvSp(norm)_i} + \mathrm{nclus_i} \qquad 12.4$$

Eq12.1 described the "spread" or homology within cluster m. J was the total number of elements in cluster m and d(i,k) was the distance between element i and element k in cluster m;

Eq12.2 described the average "spread" among all clusters in *i-th* iteration step during the clustering process. K_i was the total number of clusters at *i-th* step.

Eq12.3 described the normalized "spread" among the whole process of clustering. N is total number of steps.

Eq12.4 was the penalty function which took into account the "spread" of each cluster and the complexity of the whole dataset at *i-th* step . It equals to the average pairwise distance among all cluster members.

In the end, the clusters in the step which had the lowest accumulated cost was determined as the best cluster assignment of the original dataset.

C.Translate DOCK poses' structural information into MELD constraints

Once the optimized cluster number was determined, all poses in each cluster were used to generate MELD constraints. As for reducing redundancy, please refer to the second method:"Automatically cluster and manually choose DOCK poses"

Once the redundant constraints were removed, all the remaining constraints were assigned harmonic distance constraints. All these constraints were partitioned into 5 different constraint groups and one collections was created to manage these constraint groups. Each time, one out of five constraint groups was selected and the lowest 50% constraints were activated.

- ## Pros and Cons:

Pros: Robustness and including internal rotational degree of freedom of ligand's rotational bonds.

Cons: Complexity in implementing the KGS penalty score function.

- ## Results:

Table XI: Cluster outcomes

Cluster Index	Cluster Members
Cluster1	6,7,11
Cluster2	4,10,21,29
Cluster3	1,5,9,20,26,27
Cluster4	2,3,8,14,17,19,21,22,24
Cluster5	12,13,16,18,30

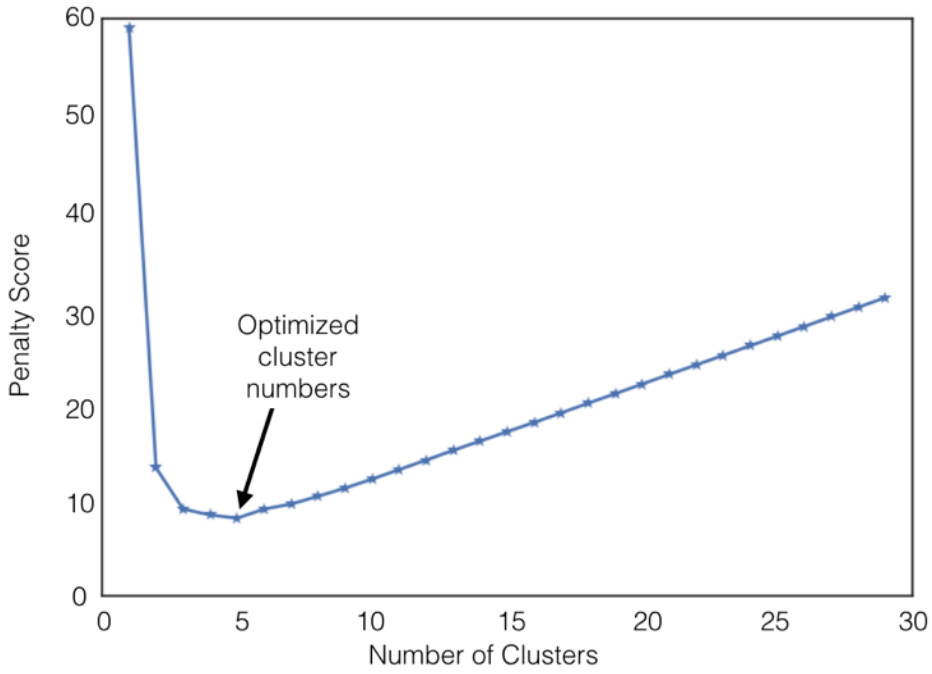

Figure 22. KGS penalty score plot of system 1BB7.

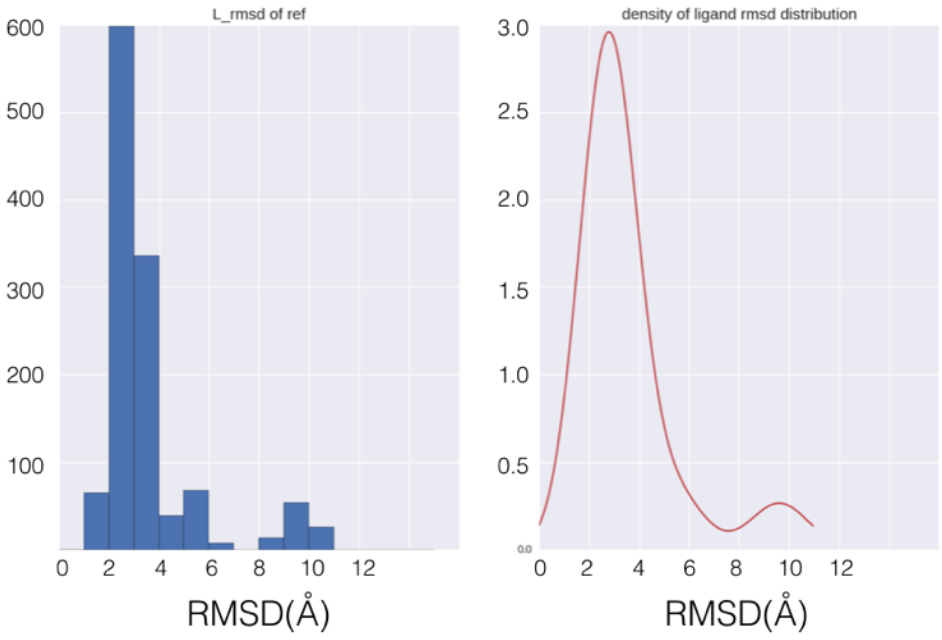

Figure 23. MELD simulation results of 1BB7, the left side is the population of ligand RMSD with respect to crystal structure, the right side is the density plot of the left side graph.

- **Discussion:**

For the KGS penalty score plot, we could see that at the left most, there is only one cluster for the whole dataset. Therefore the homology penalty score is very high which caused the whole penalty score also very high. With the cluster forming in the original dataset, the homology penalty score decreases while the complexity penalty score which represents by nclus increases. However, Since the driving force at this stage is the decreasing of homology penalty score, therefore the whole penalty score continuously decreases until the whole dataset forms the optimized number of clusters.

After total number of clusters in the dataset passed the optimized number of clusters, the complexity penalty score becomes the driving force for penalty increasing. Thus with the increasing of total number of clusters in the dataset, the whole penalty score increases again. Based on the minimum value of penalty score, the optimized number of clusters in the original dataset is 5 for system 1BB7.

As for the MELD simulation results, we could see(**Fig.23**) that this constraint generating method performed very well in 1BB7. The ligand could quickly find the crystal poses and keep stable there. The possible reasons for it works well might came from two aspects: the intrinsic structure of top30 DOCK poses, the properly designed MELD constraint group.

The intrinsic DOCK poses' structure was hard to quantitatively measure, but we could indirectly feel it from the hierarchical cluster results. The top30 DOCK poses were well clustered into five groups(**Table XI**) and crystal-like poses were all concentrated in one cluster. This meant that the DOCK poses were quite different from each other. Moreover the less constraint groups enabled MELD to have higher probability to choose the correct information. This is a incomplete work. There is no more time to test this new constraint generating method on more systems. In the future, we would further test this method on other complexes in the Stable group(1BB7, 1BB5, 1HEW, 1TOW, 1LMO, 1LZB, 1LZC, 1LZE, 1JLD, 9HVP). Moreover, polarized force field and explicit solvent model would also be included in order to target systems, like 1Z6S which has highly charged ligand or 1JLD which has a lot of crystal water involved in the binding site.

6.Conclusion:

In the modern “hit to lead” pharmaceutical industry, the speed request of fast searching a huge small molecule database and generating candidate molecules in the “hit identification” process and the accuracy demand of predicting the binding geometry and binding affinity in the “lead optimization” process were like two far separated extremes and hard to balance in current computational method. Although advanced computational tools have addressed these processes separately, few have bridged the gap.

Here we designed a pipeline which combined external constraint information with T-REMD within the framework of MELD in order to shorten the computational gap between hit and lead process. The constraint information was derived from ensembles of binding poses generated by DOCK. We aim to utilize this information to find the lowest free energy binding pose within the ensemble. Our method is of particular use in cases where DOCK alone fails to correctly score poses in their ensemble.

Parsing the DOCK ensembles into information that can be employed within MELD was a methodological goal of this work. Three strategies which refer to “Manually choose DOCK poses based on ligand RMSD”, “Automatically cluster and choose DOCK poses” and “Automatically cluster and choose DOCK poses based Kelley-Gardner-Sutcliffe penalty function” were designed and tested in this project. After applying the “Manually choose” method, we found that MELD could find the crystal pose and distinguish it from other DOCK poses which were used in the simulation. The low efficiency of sampling in 1LZB and 1BB7, which might caused by the specific ligand conformation in those system, needs further study. Then more complex but automatic method, “Automatically cluster and choose DOCK poses”, was designed and tested for all complex systems. Although the constraint amount increased 2-fold with 1BB5 even increased 5-fold, the MELD could still find the crystal poses. The low enforced factor further demonstrated the robustness of MELD. However, the sampling efficiency still needed improvement. Finally, the “Automatically cluster and choose DOCK poses based on Kelley-Gardner-Sutcliffe penalty function” method, which aimed to improve the sampling efficiency through including more information from crystal-poses, were tested on previous poorly performed 1BB7 complex. Due to the novel constraint method and constraint group strategy, both the sampling efficiency and accuracy of MELD were improved.

Of course, no method is perfect and MELD is not ‘magic.’ It can only infer the globally most likely conformation as defined by the force field and external information. This means that the confidence of MELD results largely depended on the accuracy of the force field and solvent model we used as well as on the external information. Therefore we are currently implementing an explicit solvent model (TIP3P) into MELD and are also planning to add support for polarized force fields. This will allow us to better deal with the highly charged ligand system and co-crystal waters that are often present in the crystal structure. Currently, the third constraint generating protocols, “Automatically cluster and choose based on L_rmsd and Kelley-Gardner-Sutcliffe(KGS) penalty function”, provides the best clustering results. It worked pretty well on 1BB7 and we will further test it on other complexes in the DOCK database.

7.Reference:

1. Kitchen, D. B., et al. (2004). "Docking and scoring in virtual screening for drug discovery: methods and applications." Nat Rev Drug Discov **3**(11): 935-949;
2. Wang, L., et al. (2015). "Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field." J Am Chem Soc 137(7): 2695-2703
3. De Vivo, M., et al. (2016). "Role of Molecular Dynamics and Related Methods in Drug Discovery." J Med Chem.
4. MacCallum, J. L., et al. (2015). "Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference." Proc Natl Acad Sci U S A **112**(22): 6985-6990.
5. Perez, A., et al. (2015). "Accelerating molecular simulations of proteins using Bayesian inference on weak information." Proc Natl Acad Sci U S A **112**(38): 11846-11851.
6. Kuntz, I. D. (1992). "Structure-Based Strategies for Drug Design and Discovery." Scicence **257**(5073): 5
7. Mukherjee, S., et al. (2010). "Docking Validation Resources: Protein Family and Ligand Flexibility Experiments." Journal of chemical information and modeling **50**(11): 15.
8. Jorgensen, W. L. (2004). "The Many Roles of Computation in Drug Discovery." Science **303**(5665): 6
9. Teague, S. J.(2003) “Implications of protein flexibility for drug discovery”. Nat. Rev. Drug Discovery , 2, 527-541.
10.Allen, W. J., et al. (2015). "DOCK 6: Impact of new features and current docking performance." J Comput Chem **36**(15): 1132-1156.
11.Dill, K. A. and H. S. Chan (1997). "From Levinthal to pathwas to funnels." Nature **4**: 10-19.
12.Bernardi, R. C., et al. (2015). "Enhanced sampling techniques in molecular dynamics simulations of biological systems." Biochim Biophys Acta **1850**(5): 872-877.
13.Abrams, C. and G. Bussi (2013). "Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration." Entropy **16**(1): 163-199.
14.Sugita, Y. and Y. Okamoto (1999). "Replica Exchange Molecular Dynamics Method for Protein Folding Simulation." Chemical physics letters **314**: 11.
15.He, S., et al. (2013). "Discovery of highly potent microsomal prostaglandin e2 synthase 1 inhibitors using the active conformation structural model and virtual screen." J Med Chem **56**(8): 3296-3309.
16.Rastelli, G. et al.(2009). “Binding estimation after refinement, a new automated procedure for the refinement and restoring of docked ligands in virtual screening”. Chem. Bill. Drug Des. 73, 283-286
17.Gavalli, A., et al. (2004). "A Computational Study of the Binding of Propidium to the Peripheral Anionic Site of Human Acetylcholinesterase." J.Med. Chem. **47**: 3991-3999.
18.Case, A., Simmerling, C. et al, Amber 14, 2014

19.Nguyen, H. et al.(2015).“Refinement of Generalized Born Implicit Solvation Parameters for Nucleic Acids and Their Complexes with Proteins”. J. Chem. Theory Comput., 2015, 11(8), 3714-3728.
20.Jakalian, A., et al. (2002). "Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation." J Comput Chem **23**(16): 1623-1641.
21.Pettersen, E. F., et al. (2004). "UCSF Chimera--a visualization system for exploratory research and analysis." J Comput Chem **25**(13): 1605-1612.
22.Case, D.A., Betz, R.M., Botello-Smith, W., Cerutti, D.S., Cheatham, III, T.E., Darden, T.A., Duke, R.E., Giese, T.J., Gohlke, H., Goetz, A.W., Homeyer, N., Izadi, S., Janowski, P., Kaus, J., Kovalenko, A., Lee, T.S., LeGrand, S., Li,P., Lin, C., Luchko, T., Luo, R., Madej, B., Mermelstein, D., Merz, K.M., Monard,G., Nguyen, H., Nguyen, H.T., Omelyan, I., Onufriev, A., Roe, D.R., Roitberg, A., Sagui, C., Simmerling, C.L., Swails, J., Walker, R.C., Wang, J., Wolf, R.M., Wu, X., Xiao, L., and Kollman P.A. (2016), AMBER 2016, University of California, San Francisco.
23.Salomon-Ferrer, R., et al. (2013). "An overview of the Amber biomolecular simulation package." Wiley Interdisciplinary Reviews: Computational Molecular Science **3**(2): 198-210.
24.Walker, R., et al. “Using Antechamber to Create Leap Input Files for Simulating Sustiva (efavirenz)-RT complex using the General Amber Force Field.” http://ambermd.org/tutorials/basic/tutorial4b/
25.Wang, J. M., et al. (2004). "Development and Testing of a General Amber Force Field." Journal of Computational Chemistry **25**(9): 1157-1174.
26.Maier, J., et al.(2015). “ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB.” J. Chem. Theory Comput., 2015, 11(8), 3696-3713.
27.Rocklin, G. J., et al. (2013). "Blind prediction of charged ligand binding affinities in a model binding site." J Mol Biol **425**(22): 4569-4583.
28.Hatzopoulos, G. N., et al. (2005). "The binding of IMP to ribonuclease A." FEBS J **272**(15): 3988-4001.
29.Hunter, C. A. and J. K. M. Sanders (1990). "The Nature of π-π Interactions." J. Am. Chem. Soc **112**: 5525-5535.
30.Martinez, C. R. and B. L. Iverson (2012). "Rethinking the term “pi-stacking”." Chemical Science **3**(7): 2191
31.Singh, N., et al. (2006). "Specific binding of non-steroidal anti-inflammatory drugs (NSAIDs) to phospholipase A2: structure of the complex formed between phospholipase A2 and diclofenac at 2.7 A resolution." Acta Crystallogr D Biol Crystallogr **62**(Pt 4): 410-416.
32.Chandra, D. N., et al. (2011). "Identification of a novel and potent inhibitor of phospholipase A(2) in a medicinal plant: crystal structure at 1.93A and Surface Plasmon Resonance analysis of phospholipase A(2) complexed with berberine." Biochim Biophys Acta **1814**(5): 657-663.
33.Krohn, A., et al. (1991). "Novel Binding Mode of Highly Potent HIV-Proteinase Inhibitors Incorporating the (R)-Hydroxyethylamine Isostere." J.Med. Chem. **34**(11): 3340-3342.
34.Wright, L., et al. (2004). "Structure-activity relationships in purine-based inhibitor binding to HSP90 isoforms." Chem Biol **11**(6): 775-785.

35.Paton, R. S. and J. M. Goodman (2009). "Hydrogen Bonding and π-Stacking- How Reliable are Force Fields? A Critical Evaluation of Force Field Descriptions of Non-bonded Interactions." J. Chem. Inf. Model **49**: 944-955. Widnall,S.,et al.(2008). “Lecture L26 - 3D Rigid Body Dynamics: The Inertia Tensor” MIT Aeronautic Dynamics.
36.Bottegoni, G., et al. (2006). "A Comparative Study on the Application of Hierarchical-Agglomerative Clustering Approaches to Organize Outputs of Reiterated Docking Runs." Journal of chemical information and modeling **46**(2): 11.
37.Kelley, L. A., et al. (1996). "An automated approach for clustering an ensemble of NMRderived protein structures into conformationally related subfamilies." Protein Engineering **9**(11): 1062-1065.

8.Appendix

The Appendix contains figures of all simulation results. All figures are grouped into two section:

- Testing the stability of crystal poses in force field;
- 3D crystal structure of Stable systems;

8.1.Testing the stability of crystal poses in force field

The simulation results of each complex system in **TableVI** except 1HEW and 1Z6S are listed below. Two graphs were generated from each complex. One was the ligand rmsd plot. The Ligand rmsd, which superposed the protein heavy atoms to crystal conformation and calculated the in-place root-mean-square deviation of all heavy atoms in ligand, was plotted against frames. Another one was the density of ligand rmsd distribution. The probability density of its rmsd value, which proportional to its frequency appearing in the trajectory, was plotted against the rmsd value of each frames.

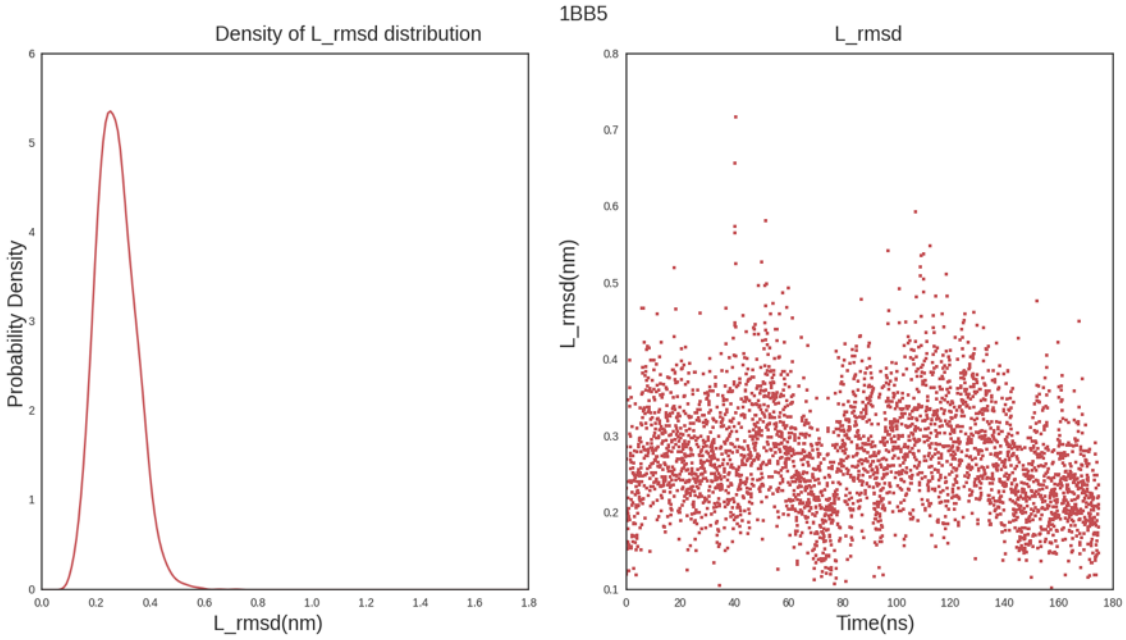

Figure 24. Stability Test results of 1BB5, simulation parameters are listed in Table II, the left is the probability density plot of the L_rmsd distribution, the right one is ligand rmsd plot.

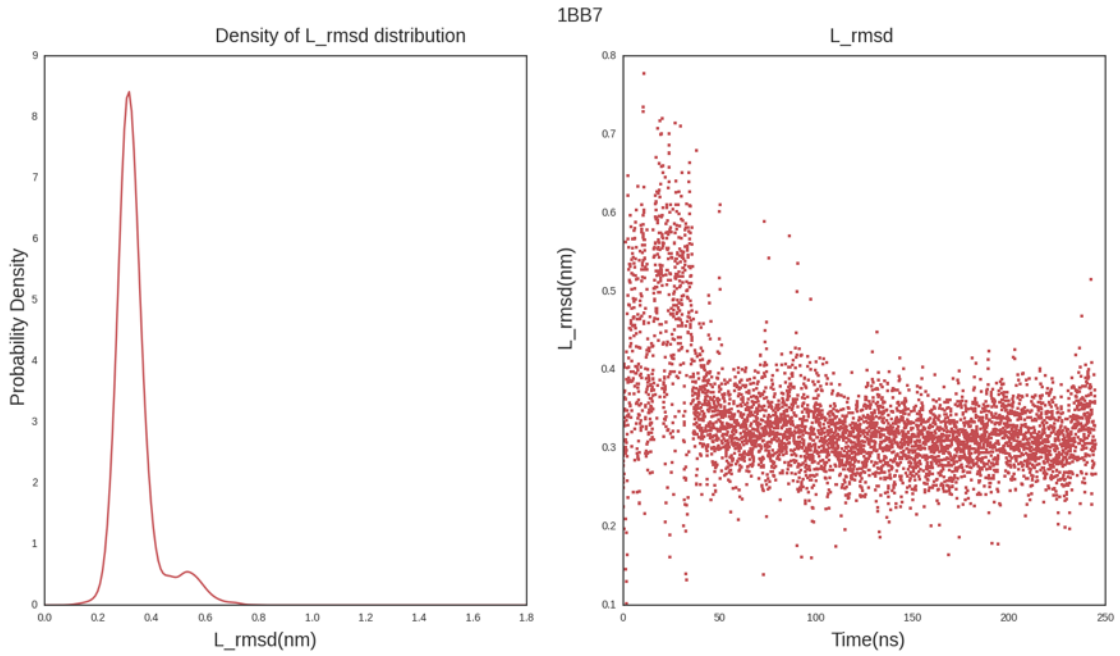

Figure 25. Stability Test results of 1BB7, simulation conditions same as **Figure 24.**

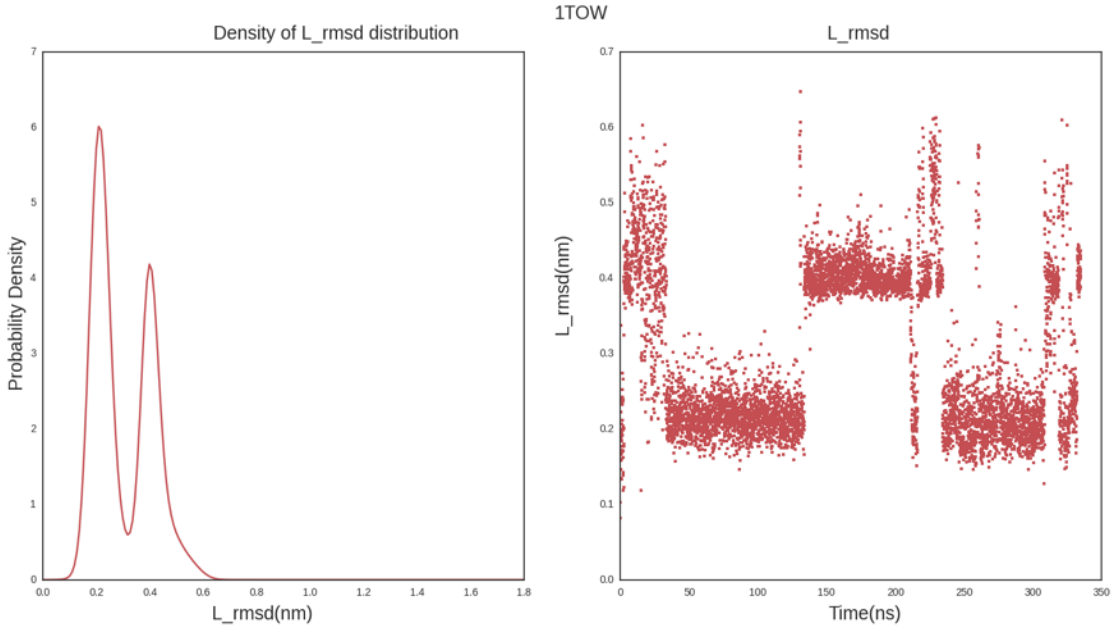

Figure 26. Stability Test results of 1TOW, simulation conditions same as **Figure 24.**

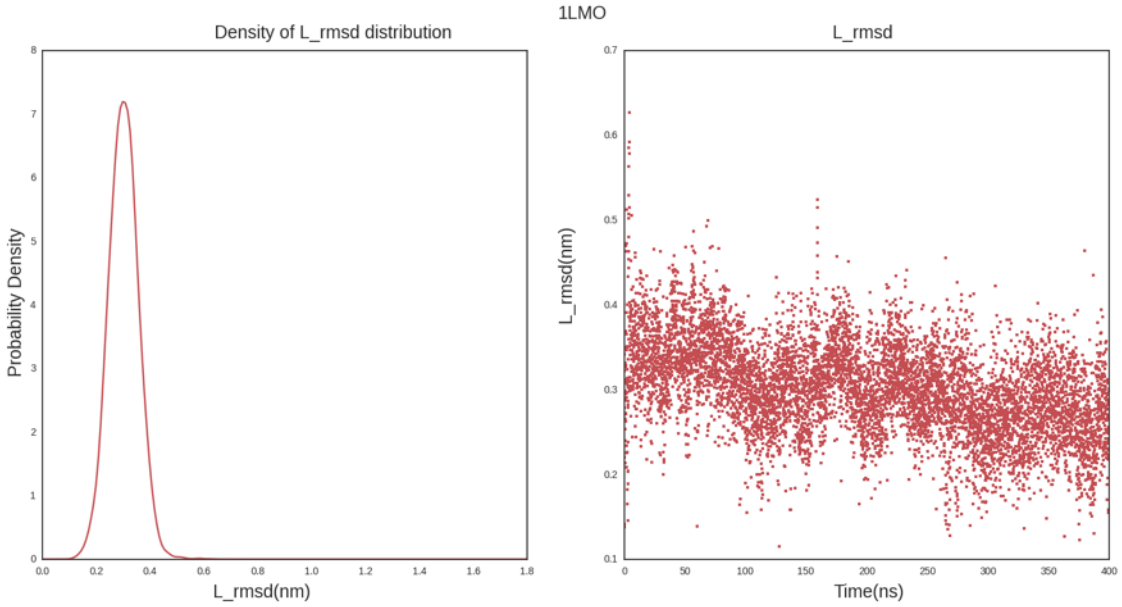

Figure 27. Stability Test results of 1LMO, simulation conditions same as **Figure 24.**

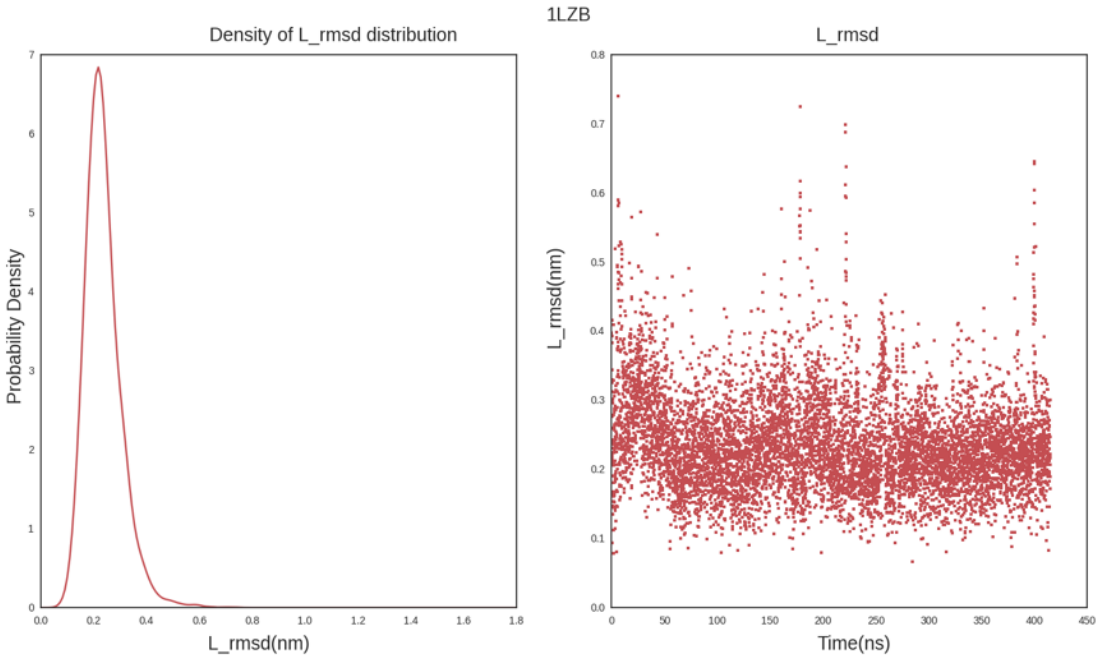

Figure 28. Stability Test results of 1LZB, simulation conditions same as **Figure 24.**

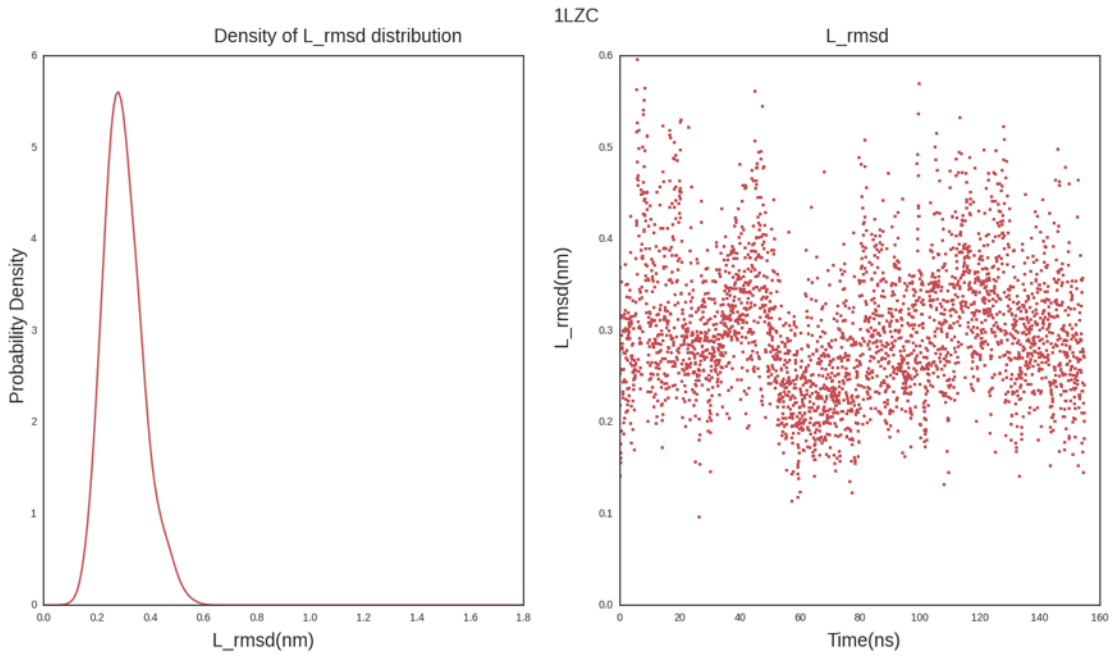

Figure 29. Stability Test results of 1LZC, simulation conditions same as **Figure 24.**

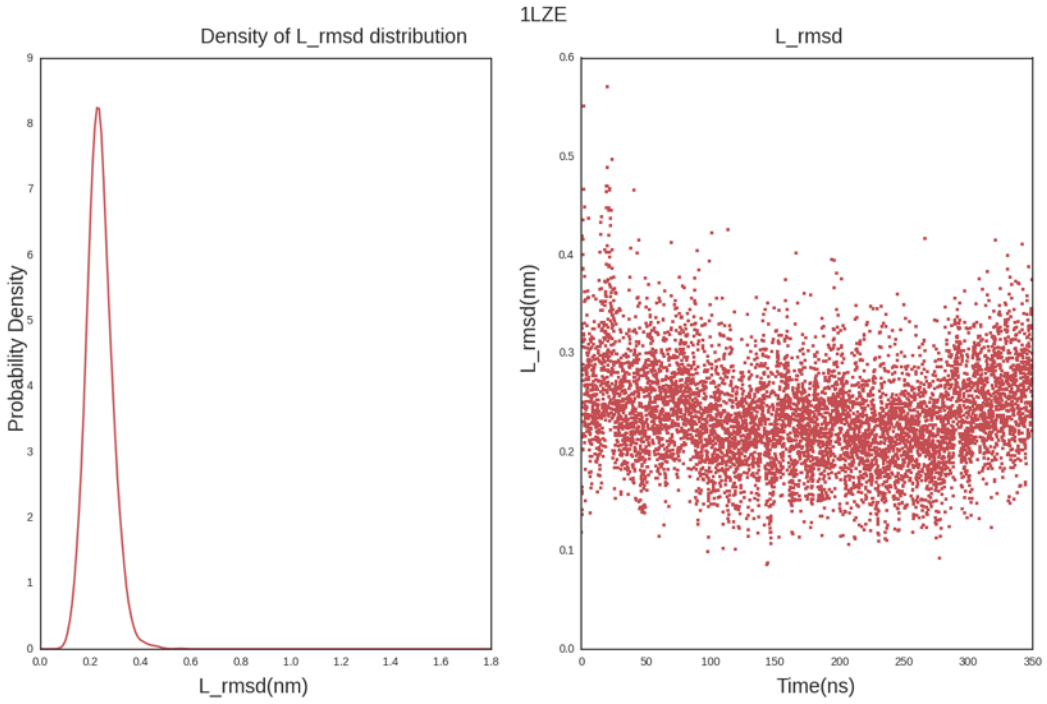

Figure 30. Stability Test results of 1LZE, simulation conditions same as **Figure 24.**

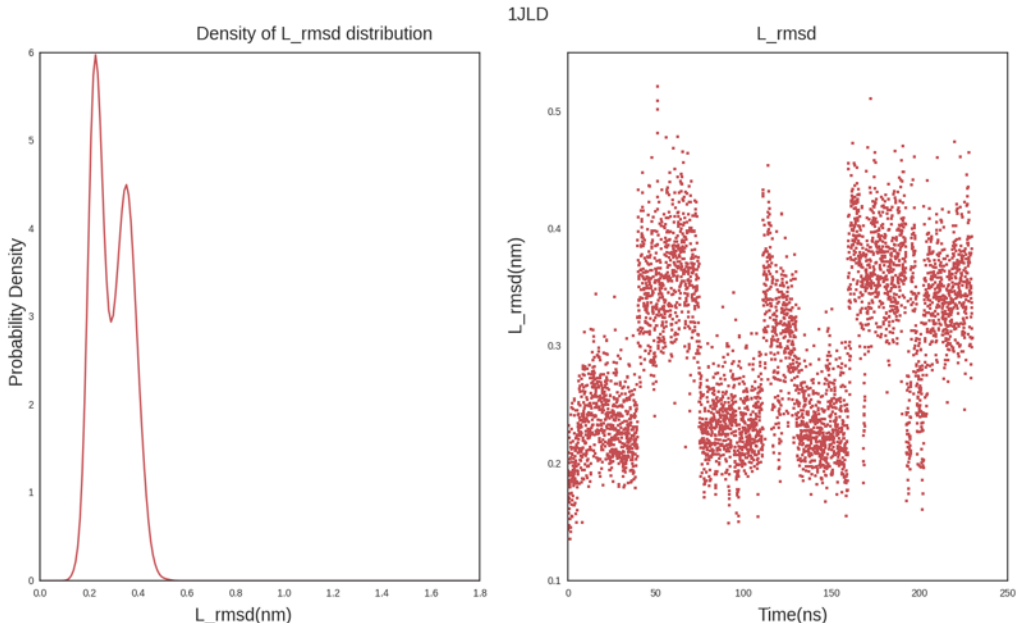

Figure 31. Stability Test results of 1JLD, simulation conditions same as **Figure 24.**

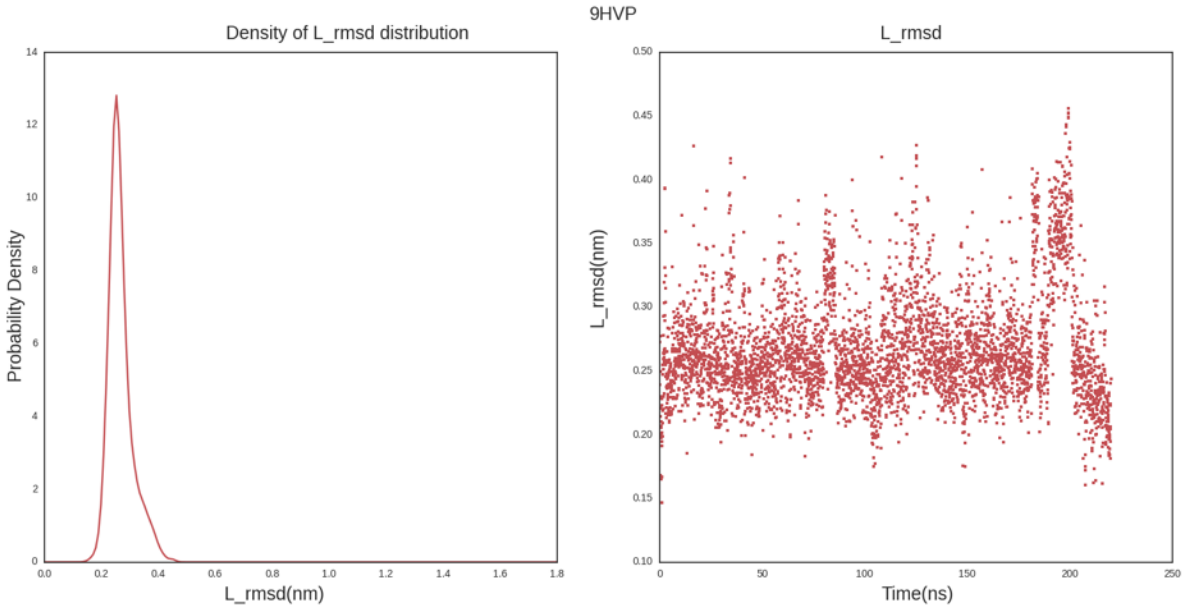

Figure 32. Stability Test results of 9HVP, simulation conditions same as **Figure 24.**

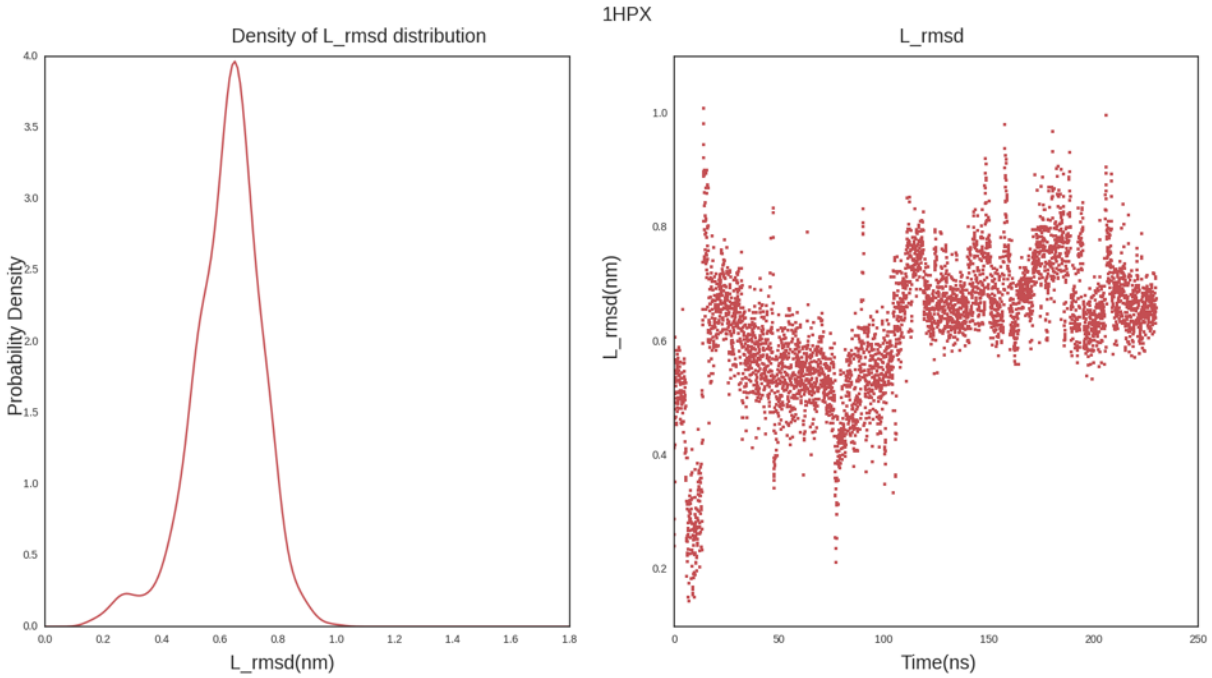

Figure 33. Stability Test results of 1HPX, simulation conditions same as **Figure 24**.

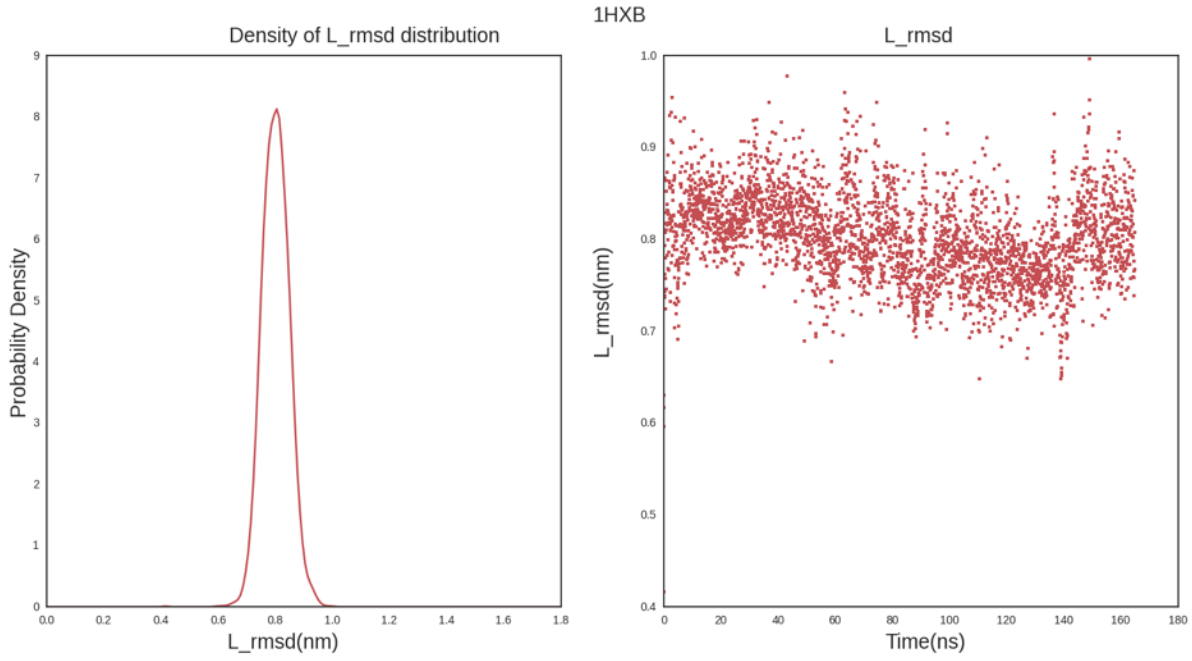

Figure 34. Stability Test results of 1HXB, simulation conditions same as **Figure 24.**

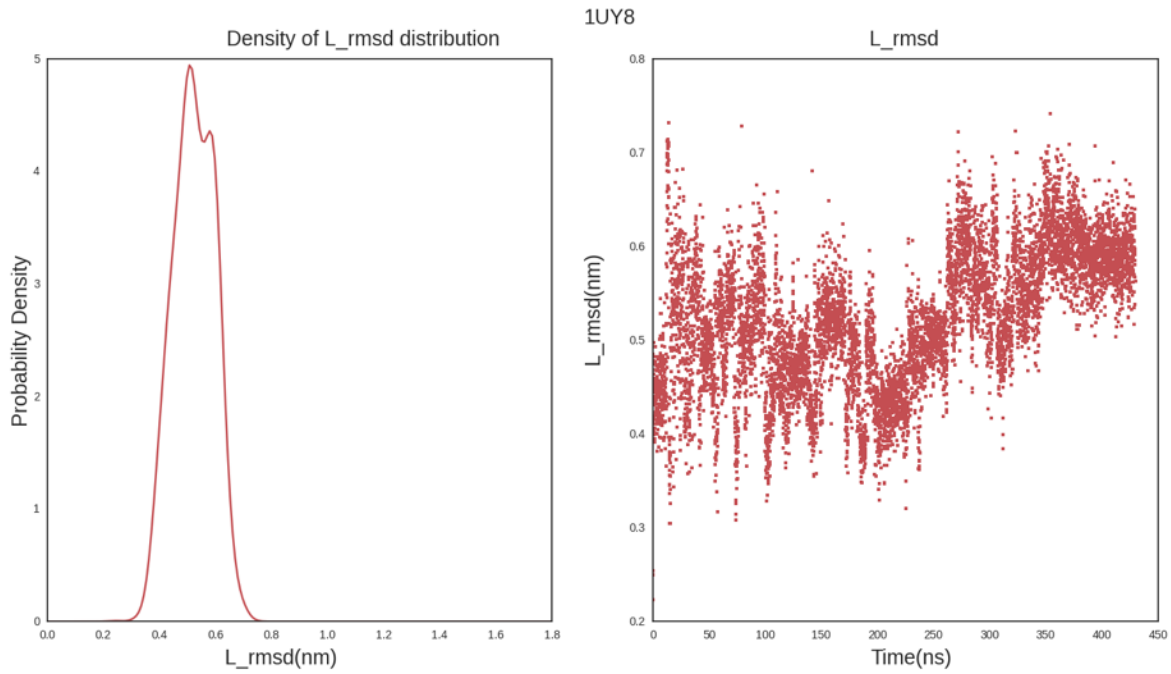

Figure 35. Stability Test results of 1UY8, simulation conditions same as **Figure 24.**

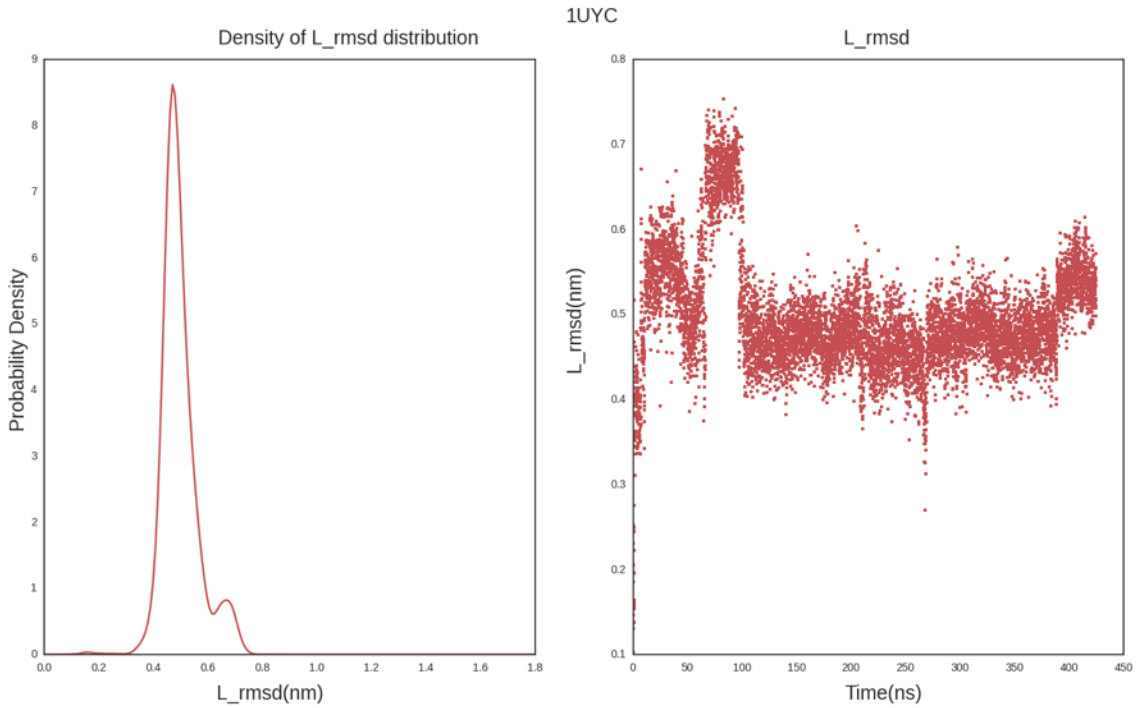

Figure 36. Stability Test results of 1UYC, simulation conditions same as **Figure 24.**

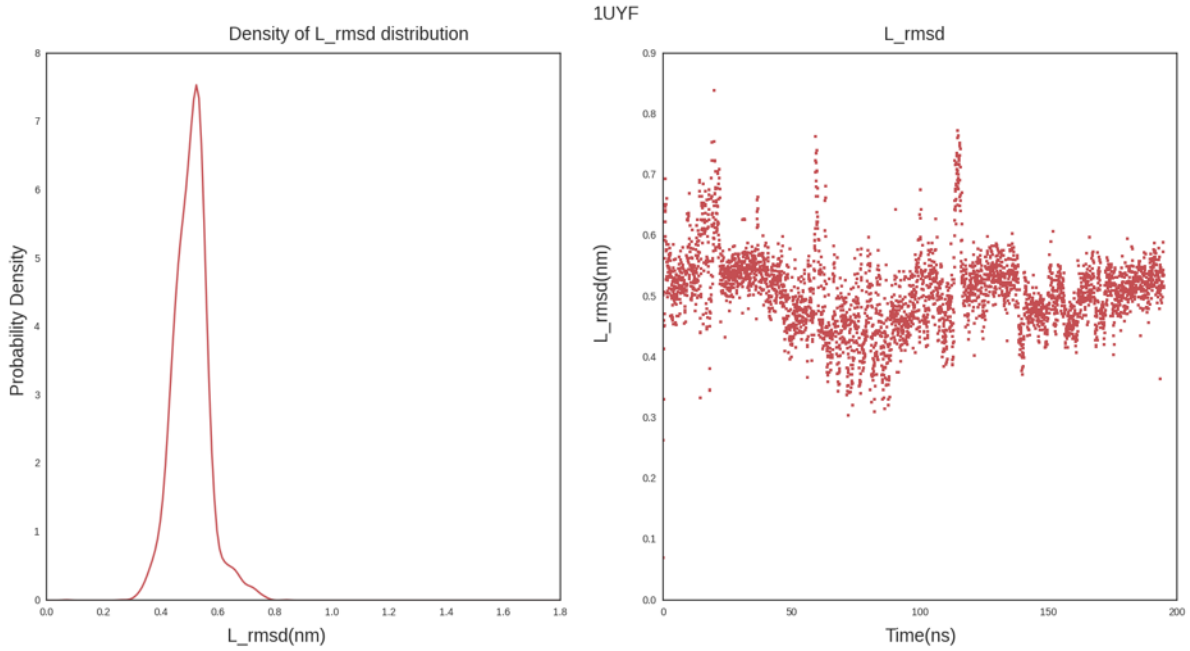

Figure 37. Stability Test results of 1UYF, simulation conditions same as **Figure 24.**

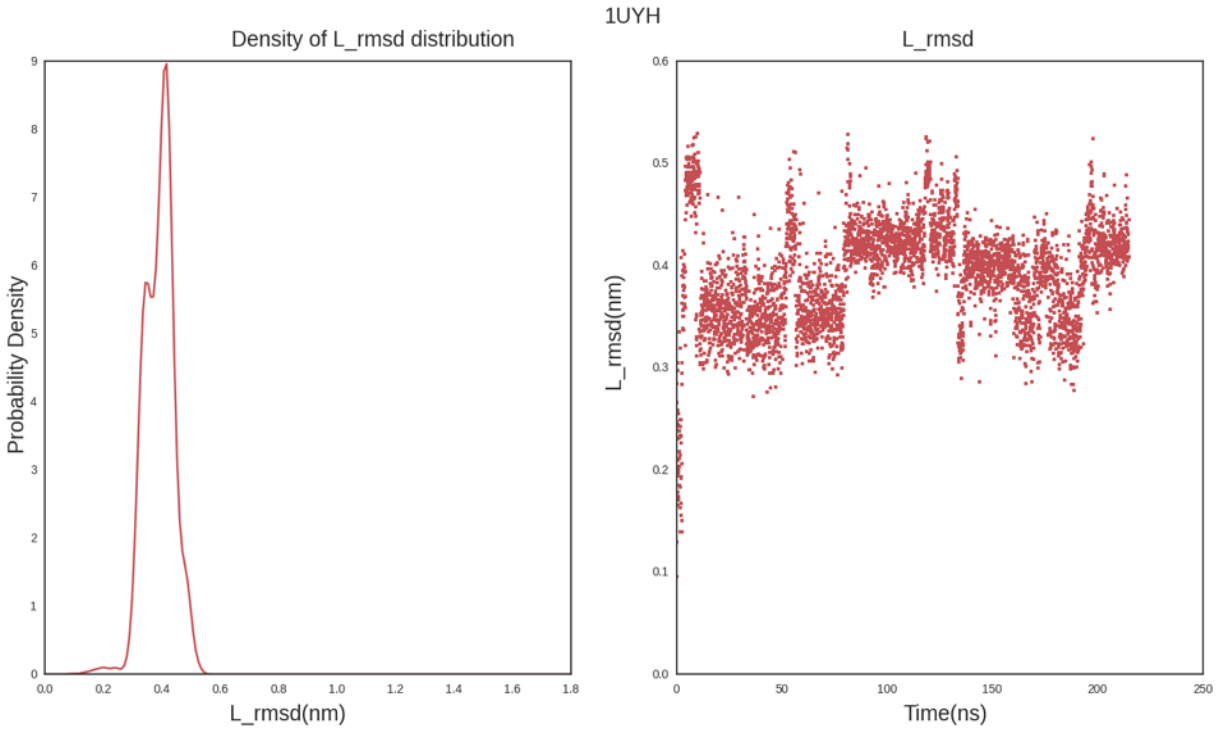

Figure 38. Stability Test results of 1UYH, simulation conditions same as **Figure 24.**

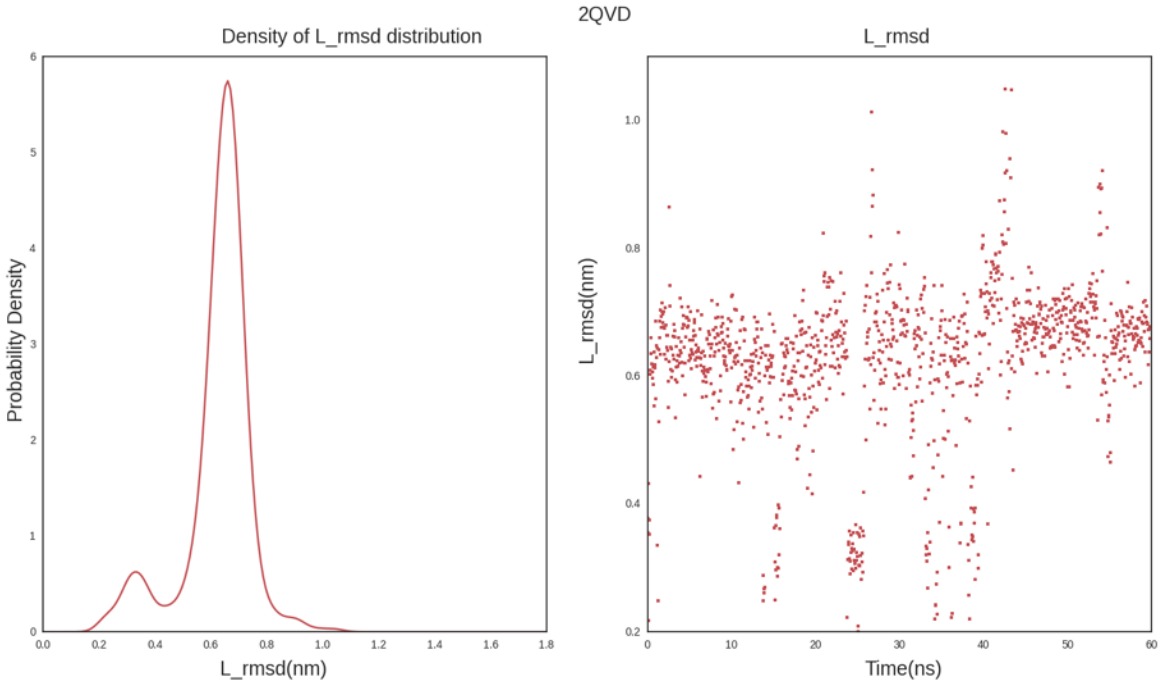

Figure 39. Stability Test results of 2QVD, simulation conditions same as **Figure 24.**

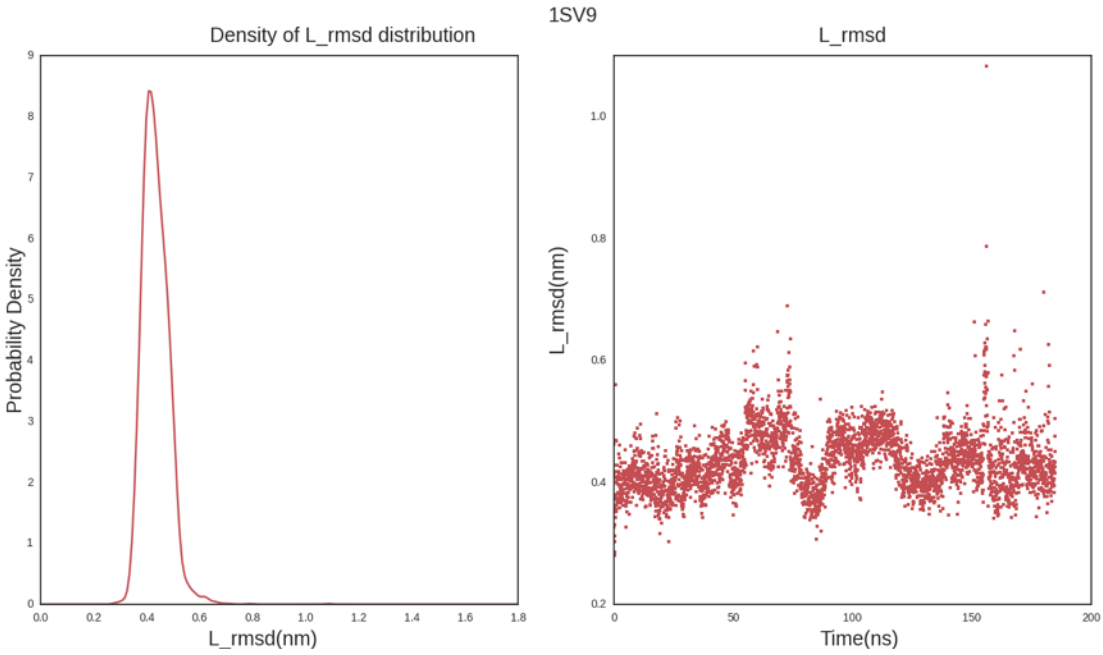

Figure 40. Stability Test results of 1SV9, simulation conditions same as **Figure 24.**

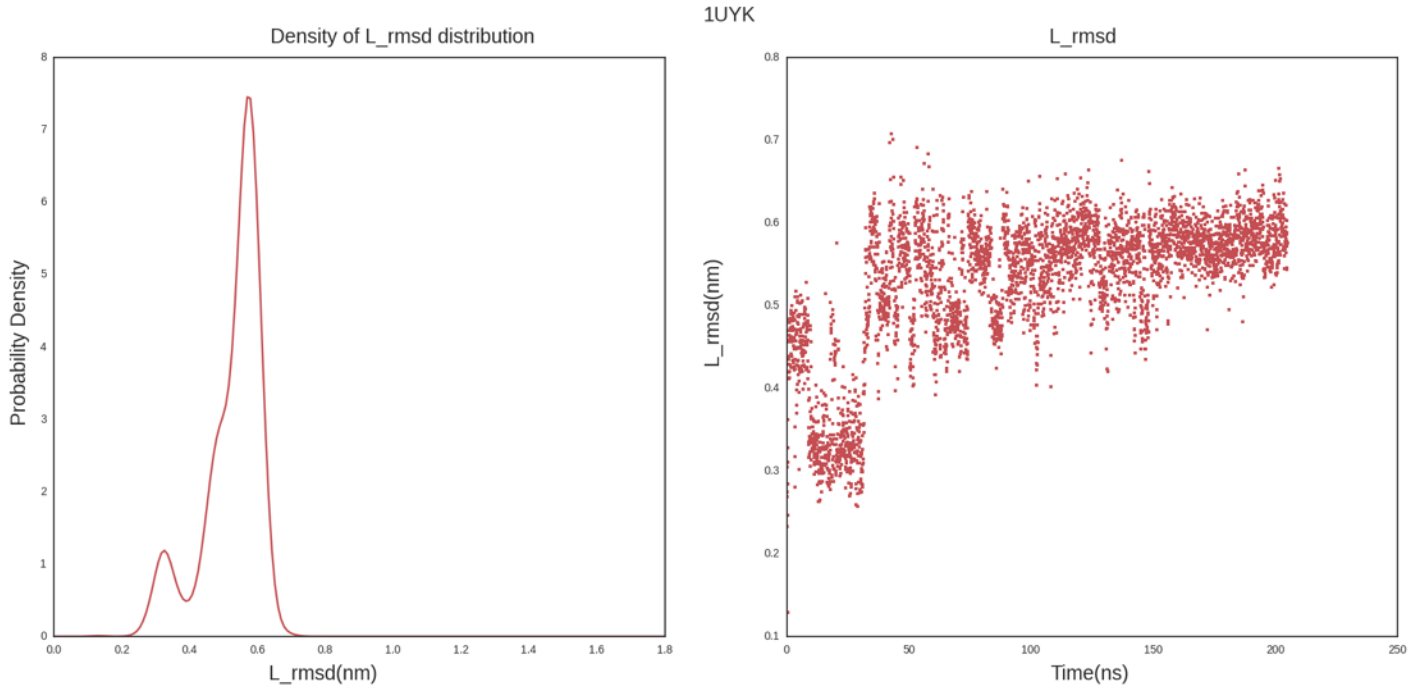

Figure 41. Stability Test results of 1UYK, simulation conditions same as **Figure 24.**

8.2.3D crystal structure of stable systems

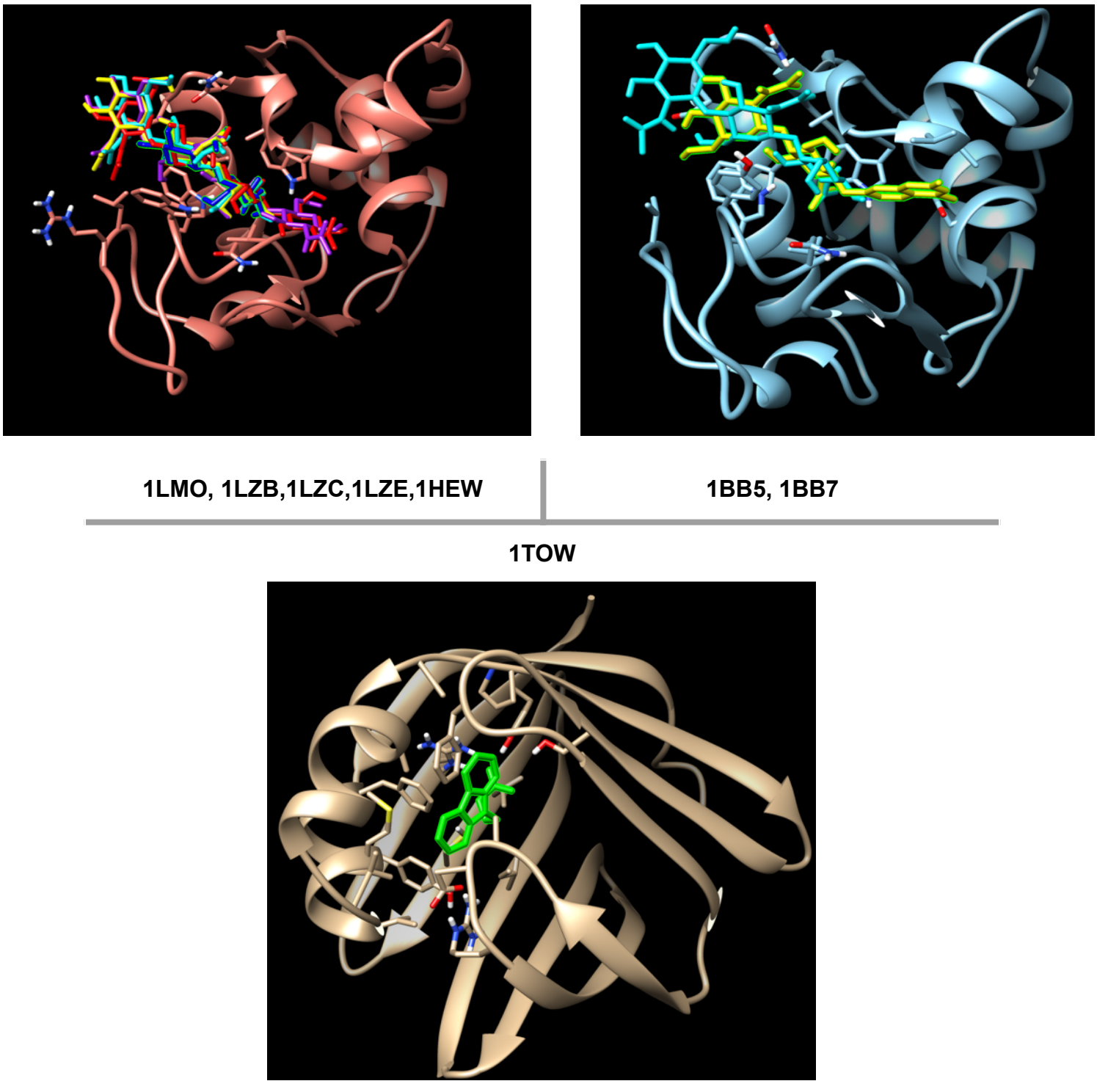

Figure 42. 3D structure of all stable systems