

2021-05-01

The Hybridizing Ions Treatment (hit) Method Development And Computational Study On SARS-CoV-2 E Protein.

Shengjie Sun
University of Texas at El Paso

Follow this and additional works at: https://scholarworks.utep.edu/open_etd



Part of the [Biology Commons](#), [Biophysics Commons](#), and the [Biostatistics Commons](#)

Recommended Citation

Sun, Shengjie, "The Hybridizing Ions Treatment (hit) Method Development And Computational Study On SARS-CoV-2 E Protein." (2021). *Open Access Theses & Dissertations*. 3356.
https://scholarworks.utep.edu/open_etd/3356

This is brought to you for free and open access by ScholarWorks@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of ScholarWorks@UTEP. For more information, please contact lweber@utep.edu.

THE HYBRIDIZING IONS TREATMENT (HIT) METHOD DEVELOPMENT AND
COMPUTATIONAL STUDY ON SARS-COV-2 E PROTEIN.

SHENGJIE SUN

Master's Program in Computational Science

APPROVED:

Lin Li, Ph.D., Chair

Jianjun Sun, Ph.D.

Jorge A. Munoz, Ph.D.

Stephen L. Crites, Jr.
Dean of the Graduate School

Copyright ©

by
Shengjie Sun
2021

THE HYBRIDIZING IONS TREATMENT (HIT) METHOD DEVELOPMENT AND
COMPUTATIONAL STUDY ON SARS-COV-2 E PROTEIN.

by

SHENGJIE SUN, MSc

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Computational Science Program

THE UNIVERSITY OF TEXAS AT EL PASO

May 2021

Acknowledgments

The research has benefited from the valuable inspiration and supports from many people.

I would first like to express my deepest appreciation to my master advisor, Prof. Lin Li, whose outstanding academic guidance was invaluable in my UTEP research life. His expertise in academic research extended my horizons in computational biophysics and his charming personality also educated my sociability.

Then, I also would like to appreciate my committee co-chair Prof. Jianjun Sun. Prof. Sun helped me a lot in the biological area for the envelope protein project. The professional suggestions from Prof. Sun inspired me a lot for the study on biomolecules. Meanwhile, I would like to thank my committee co-chair Prof. Jorge A. Munoz. Many thanks to him for attending my thesis defense and providing important comments and useful suggestions on my current work and future study.

Additionally, I would like to thank Prof. Ming-Ying Leung from Computational Science Program for serving as my director in my qualifier exam. Prof. Leung provided many instructional suggestions on the course selection in every semester. Besides, her course of mathematical and computer modeling enlightened me in the computational science area.

My thanks also go to my friends in my research group: Mr. Chitra Karki, Dr. Yuejiao Xian, Miss. Yixin Xie, Mr. Alan, Miss. Branda, Miss. Wenhan Guo, Mr. Jiadi He, Mr. Tolulope A Ale, and Mr. Juan Lopez. Thank you for being supportive and accompanying me in my master's studies. You make my life in UTEP enjoyable and memorable. I also thank my friends in El Paso: Mr. Honglun Xu, Mr. Yongteng Dong, Mr. Jixuan Pan, and Mr. Yi Xie. During my El Paso life, they give me a lot of help and supports so that I can focus on my study and research.

Lastly and most importantly, I want to express my special thanks to my parents for their unconditional love, support, and encouragement. Without my parents' support and loving care, I would not have got this achievement in my life. Whether I have a good time or a tough time, they provide me support and endless love to realize my dream with no hesitation. I am very happy to have them as my parents and so proud of them.

Abstract

Fast and accurate calculations of the electrostatic features for highly charged biomolecules such as DNA, RNA, highly charged proteins, are crucial but challenging tasks. Traditional implicit solvent methods calculate the electrostatic features fast, but they are not able to balance the high net charges in the biomolecules effectively. Explicit solvent methods add unbalanced ions to neutralize the highly charged biomolecules in molecular dynamic simulations, which require more expensive computing resources. Here we developed a novel method, the Hybridizing Ions Treatment (HIT) method, which hybridizes the implicit solvent method with the explicit method to realistically calculate the electrostatic potential for highly charged biomolecules. This HIT method utilizes the ionic distribution from an explicit method to predict the bound ions. Such bound ions are then added in the implicit solvent method to perform the electrostatic potential calculations. In this study, two training sets were developed to optimize the parameters for the HIT method. The performance on the testing set demonstrates that the HIT method significantly improves the electrostatic calculations. The results on two molecular motors, myosin, and kinesin, reveal some mechanisms for the two molecular motor proteins and explain some previous experimental findings well. This HIT program can be widely used to study highly charged biomolecules, including DNA, RNA, molecular motors, and other highly charged biomolecules. The HIT package is available at http://compbio.utep.edu/static/downloads/download_hit.zip.

The SARS-CoV-2 that caused Covid-19 has spread since the end of 2019. Its major effects resulted in nearly half a million deaths around the whole world. Therefore, understanding virulence mechanisms is important to prevent future outbreaks and for COVID-19 drug development. The envelope protein is an important structural protein affecting virus assembly and budding. The E protein pentamer is a viroporin, serving as ions transferring channel in cells. In this work, we applied MD simulations, topological and electrostatic analyses to study the effects of palmitoylation on the E protein pentamer. The results indicate the cation transferring direction is more like lumen to the cytosol. The structure of palmitoylated E protein pentamer is more stable while the loss of palmitoylation caused the pore radius to reduce and even collapse. The electrostatic forces on the two sides of palmitoylated E protein pentamer are more beneficial to attract cations in the lumen and to release cations into the cytosol. The results indicate the importance of palmitoylation, which can help the drug design for the treatment of COVID-19.

Table of Contents

Acknowledgments.....	iv
Abstract.....	v
List of figures.....	vii
Chapter 1: Hybrid method for representing ions in implicit solvation calculations.	1
1.1 Introduction.....	1
1.2 Methodology.....	4
1.3. Results and discussion.....	13
1.4. Conclusions.....	28
Chapter 2: The computational study on the function of palmitoylation on the envelope protein in SARS-CoV-2.....	30
2.1 Introduction.....	30
2.2 Methodology.....	33
2.2 Result and discussion.....	37
2.4 Conclusion.....	44
Chapter 3: Future work.....	46
3.1 Myosin project.....	46
3.2 HIT2 project.....	48
3.3 Time Schedule of Future Research.....	50
Curriculum Vitae.....	63

List of figures

Figure 1.1 A: Schematic presentation of 1000 frames of ions; B: Ionic cloud distribution of the combination of 1000 frames from A; binding sites are marked by black circles; C: Cubic partition of the ionic	5
Figure 1.2 Myosin-actin complex (A) and Kinesin-tubulin dimer complex (B) and their sodium ionic cloud distribution. C and D are the sodium ionic cloud distribution of myosin complex (A) and kinesin complex (B) within 10 ns simulation (1000 frames).....	6
Figure 1.3 Diagram of ion counting, cube sorting, and clustering process. The number indicates the ions contained in the cube. Note that all numbers are invented for demonstration, and the $4 \times 4 \times 4$ cutting of the solvate box is for better visualization.....	8
Figure 1.4 Diagram of the differences between the center of the initial cluster cube and the centroid of the clusters	10
Figure 1.5 The original binding sites are marked as red (A, B and C) while the calculated binding sites after initial (A), clustering (B), and optimal (C) steps are marked as green, pink, and blue. Average error (D) is calculated by the average distance.	11
Figure 1.6 Average number of ions of top 20 clusters per frame with different cube sizes of 2.0 Å, 2.5 Å, 2.7 Å, 3.0 Å, 3.3 Å, 3.5 Å, 4.0 Å, 4.5 Å, 5.0 Å, 5.5 Å, and 6.0 Å, and the average error (B) based on different cube size of the range from 2.0 Å to 6.0 Å.	12
Figure 1.7 The diagram of binding area and cube size selection (Eq. (1.6))......	16
Figure 1.8 The average error of the calculated 8 binding sites for different simulation times of the NAMD training set.	17

Figure 1.9 The figures of the NAMD training set for different simulation running times where the red balls represent the original binding sites, and the yellow balls represent the calculated binding sites. A, B, C, D, and E represent the original binding sites vs calculated binding sites..... 18

Figure 1.10 The sensitivity study of the calculated binding sites based on the Kinesin dataset. The coverage is the ratio of the number of correctly identified calculated binding sites (Distance smaller than the length of the side of binding sites) divided by the number of total calculated binding sites. Reference is the result of HIT based on 10ns simulation. 19

Figure 1.11 The comparisons of the proteins' position and structure before and after 10ns simulation of both myosin-actin complex and kinesin-tubulin complex (The red represents the position of proteins before simulation while the blue represents that after simulations). 20

Figure 1.12 The errors of five testing positions' potential values without HIT and with HIT in the kinesin testing set. The testing positions for potential testing are 10 Å away from calculated binding sites. We took the top 5 positions without clashes with protein and ions to calculate potentials by DelphiForce. The potential of the five positions was tested based on the structure without explicitly bound ions (without HIT) and with explicitly bound ions (with HIT) for comparison. The error of a certain position's potential value is the difference between the potential in reference and that based on a certain method (without HIT and with HIT). The reference potential values are the average potential values of 200 frames (obtained from 5 ns – 10 ns simulations), where the proteins and all ions (Na⁺ and Cl⁻) were saved as the structure. The concentration of salt was set as 0.00 for the potential calculation of reference while that was set as 150 mM for both structures without explicitly bound ions (without HIT) and with explicitly bound ions (with HIT). 21

Figure 1.13 Electrostatic surface representation of myosin dataset in front (A and B) and back side (C and D), in which A and C represent the electronic surface without explicit sodium ions (yellow

balls) by the traditional method, and B and D represent the electronic surface with explicit sodium ions by the HIT. The images are rendered by Chimera with a color scale from -1.0 to 1.0 kT/e.22

Figure 1.14 Electrostatic surface representation for kinesin dataset in front (A and B) and back side (C and D), in which A and C represent the electronic surface without explicit sodium ions (yellow balls) by the traditional method, and B and D represent the electronic surface with explicit sodium ions by the HIT. Images were rendered by Chimera with a color scale from -1.0 to 1.0 kT/e 23

Figure 1.15 Electrostatic surface representation of the interface between myosin motor domain and actin filament in two directions. A and C represent the electronic surface without explicit sodium ions (yellow balls) by the traditional method, and B and D represent the electronic surface with explicit sodium ions by the HIT. The images are rendered by Chimera with a color scale from -1.0 to 1.0 kT/e. 24

Figure 1.16 Electrostatic surface representation for the interface between kinesin motor domain and tubulin dimer in two directions. A and C represent the electronic surface without explicit sodium ions (yellow balls) by the traditional method while B and D represent the electronic surface with explicit sodium ions by the HIT. The images are rendered by Chimera with a color scale from -1.0 to 1.0 kT/e. 25

Figure 1.17 Electrostatic field line for the interface of myosin motor domain with tropomyosin (Left enlarged view) and actin (Right enlarged view). Yellow balls represent explicit sodium ions added by the HIT. 27

Figure 1.18 Electrostatic field line for the interface of kinesin motor domain with α -tubulin (Left enlarged view) and kinesin motor domain with β -tubulin (Right enlarged view). Yellow balls represent explicit sodium ions added by the HIT. 28

Figure 2.1 The structure alignment and the sequence alignment of SARS-CoV2 and SARS-CoV.	32
Figure 2.2 The diagram of E protein pentamer in the membrane.	33
Figure 2.3 The diagram of the placement of simulated H ⁺ for electrostatic force testing.	36
Figure 2.4 The RMSD of E-protein with and without Palm during 50 ns simulation.....	38
Figure 2.5 The general comparison of the E protein pentamer on the top and bottom view.....	38
Figure 2.6 The minimal pore radius of the E-protein during the simulation.	39
Figure 2.7 The number of salt bridges of E-protein pentamer with and without Palm during 50 ns simulation.....	39
Figure 2.8 The RMSF of the residues in the transmembrane domain of E-protein. Zone A, B, C, D, D, and E represent transmembrane domain residues in 5 monomers of E-protein.....	40
Figure 2.9 The electrostatic surface of E-protein with and without palmitoylation in the top and bottom view	41
Figure 2.10 The electrostatic surface and electric field line of membrane/E-protein with and without palmitoylation.	42
Figure 2.11 The electrostatic force on H ⁺ in both cytosolic and luminal sides of the E-protein. The distance is the distance with the membrane surface.	43
Figure 2.12 The electrostatic force on H ⁺ in both cytosolic and luminal sides of the E-protein. The distance is the distance with the Z-axis, which passes the mass center.....	44
Figure 3.1 Electrostatic surfaces for the actomyosin models in rigor (A and C) and post-rigor state (B and D).....	47
Figure 3.2 Electrostatic field line for the interface of actomyosin model in rigor (A and C) and post-rigor state (B and D).	48

Chapter 1: Hybrid method for representing ions in implicit solvation calculations.

1.1 Introduction

In computational biology, the electrostatic calculation of biomolecules is fundamental and challenging. The electrostatic interactions play significant roles in protein folding(Ganguly et al., 2012), protein stability(Stigter, Alonso, & Dill, 1991; Strickler et al., 2006), protein-protein interactions(L. Li, Jia, et al., 2017; L. Li, Wang, & Alexov, 2015; Zondlo, 2013), protein-DNA/RNA interactions(Akhtar, Zink, & Becker, 2000; Richardson, Richardson, & Bioinformatics, 1988), and many other fields. However, in vivo, the water, ions, and small biomolecules, make the environment of protein extremely complicated for electrostatic calculations. The highly charged biomolecules, including DNAs, RNAs, motor proteins, utilize ions surrounding their surfaces to balance the net charges so that they can well interact with other molecules. The trapped ions directly affect the electrostatic surfaces, which have significant impacts on the interactions between biomolecules. Currently, there are two types of models to handle the ions and water molecules surrounding biomolecules: implicit solvent models and explicit solvent models. The most popular implicit solvent methods include the Poisson-Boltzmann (PB) model(Nicholls & Honig, 1991) and Generalized Born (GB) model(Jayaram, Sprous, & Beveridge, 1998). In implicit solvent models, the electrostatic features are calculated for biomolecules by treating ions implicitly(Jia, Li, Chakravorty, & Alexov, 2017; Klapper et al., 1986; Nicholls & Honig, 1991). On the other side, explicit solvent models such as TIP3P, TIP4P, with explicit ions, are widely used in Molecular Dynamic (MD) simulations by handling ions and molecules explicitly(Florová, Sklenovsky, Banáš, Otyepka, & Computation, 2010). Explicit solvent models can neutralize the highly charged biomolecules by adding unbalanced amounts of positive and negative ions into a system. Implicit models, such as DelPhi(L. Li et al., 2012; L. Li, Li, Zhang, Alexov, & computation, 2013), are widely used to calculate the electrostatic potential, electric field lines, and electrostatic surfaces for biomolecules. However, traditional implicit solvent models treat the solvation as neutral with the same amounts of positive and negative ionic charges, which causes troubles in the electrostatic calculations of highly charged biomolecules and results in unrealistic interaction analyses. Under this circumstance, how to improve the implicit models to handle the highly charged biomolecules is a challenging problem to be solved.

The highly charged biomolecules have been studied for decades for their special functions, such as the motion of motor proteins(L. Li, Alper, & Alexov, 2016a; M. Li & Zheng, 2013; Wei, Yang, & disease, 2019), tRNA binding of ribosomes(Korennykh, Correll, & Piccirilli, 2007), and roles of cells aging-related proteins(De Graff, Hazoglou, & Dill, 2016; Lee, Fitch, Lecomte, & García-Moreno E, 2002). In living cells, the binding of oppositely charged ions by proteins is a common but principal phenomenon, related to enzyme-activations(Dayton, Reville, Goll, & Stromer, 1976) and conformational changes of proteins(Yamada, Namba, & Fujii, 2020). The classic implicit model does not consider those bound ions and thus cannot balance the net charges of the highly charged biomolecules. The loss of bound ions not only causes the net charge of the system unbalanced but also leads to biased electrostatic calculations surrounding the ionic binding sites. To perform realistic electrostatic calculations of highly charged biomolecules, we developed a novel method, which adds the bound ions explicitly and hybridizes with implicit ions to compensate the net charges in highly charged biomolecular systems.

The bound ions in binding sites are crucial for electrostatic calculations of highly charged biomolecules. Such bound ions are not represented in the implicit solvent models. Many methods have been developed to predict such bound ions or the corresponding cavities: Variational Implicit-Solvent Model (VISM) successfully captures the surface of local hydrophobic cavities for ligand-receptor bindings(Zhou et al., 2014; Zhou et al., 2015; Zhou et al., 2019); BION program(Shashikala, Chakravorty, Panday, & Alexov, 2021) implements electrostatic features and geometric information to predict the bound ions. Here we developed a novel algorithm that utilizes information from MD simulations to identify the ion-binding sites. The explicit solvent models treat the ions explicitly in calculations such as MD simulations. Therefore, the trajectories from MD simulations with explicit models contain the dynamic information of the bound ions. However, it is difficult to determine which ions are bound ions based on a single frame from simulations. Combining all the frames from a simulation trajectory into an ionic cloud distribution and properly analyzing the ionic cloud distribution may lead to the identification of the binding sites and explicit ions adding. In this work, we introduce a novel method that uses the information from the explicit solvation modeled MD simulations to identify the binding sites around the highly charged biomolecules. This method hybridizes the explicit ions on binding sites and implicit solvent models to calculate the electrostatic potentials of the highly charged biomolecules, so it is named Hybridizing Ions Treatment (HIT). We tested this method by using NAMD(Phillips et al., 2005)

and Delphi to do the explicit solvent simulations and implicit solvent calculations respectively, which proves that it significantly improves the performance of pure implicit models on highly charged biomolecules.

The method was optimized against two training sets and applied to the testing set with two biological applications. One of the biological applications is the cardiac myosin-actin complex with the net charge of $-77e$, while the other is a kinesin-5 (cut7)- $\alpha\beta$ -tubulin complex with the net charge of $-36e$. Both net charges are calculated by pdb2pqr(Dolinsky, Nielsen, McCammon, & Baker, 2004). The myosin is a superfamily of motor proteins, and it is famous for its role in muscle contraction(Houdusse & Sweeney, 2016), especially on heart diseases(Burghardt & Sikkink, 2013; Geisterfer-Lowrance et al., 1990; Inchingolo, Previs, Previs, Warshaw, & Kad, 2019). Here the type of myosin-actin complex we studied is β cardiac myosin(Robert-Paganin, Auguin, & Houdusse, 2018) with a part of cardiac actin filament(Risi et al., 2017). Usually, an actin filament is assembled by globular-actin (G-actin), tropomyosin (TM), the troponin complex (Tn)(Al-Khayat & Practice, 2013), and myosin-binding protein C (MyBP-C)(Luther et al., 2011). In this study, the TM, G-actin, and myosin motor domain were assembled and applied for MD simulations and related analysis by HIT. The kinesin, a type of well-studied molecular motors(L. Li, Alper, & Alexov, 2016b; L. Li, Jia, et al., 2017), is a superfamily of motor proteins moving along microtubules, which is crucial for mitosis(Kapitein et al., 2005). It is recently identified as an important target for cancer treatments(Huszar, Theoclitou, Skolnik, Herbst, & Reviews, 2009). The yeast kinesin-5 (cut7) motor domain with an $\alpha\beta$ -tubulin heterodimer complex was selected in this study(von Loeffelholz, Peña, Drummond, Cross, & Moores, 2019). The success of the testing set reveals that our method can be widely applied to highly charged biomolecules and can obtain reliable electrostatic calculations. By taking advantage of explicit and implicit solvent models, this novel approach utilizes a hybrid method to realistically simulate the solution environment surrounding biomolecules, which is a promising direction for simulations of highly charged biomolecules. Such a method paves the way for future drug design, DNA/RNA simulations, protein-protein interactions, and other computational biophysics research fields.

1.2 Methodology

1.2.1 Dataset

1.2.1.1 training sets

To optimize our program, two training sets were designed to mimic the sodium distribution surrounding biomolecules in saline solution (150 mM NaCl). One of the training sets was generated by a random generation algorithm (Random ions training set), the other was achieved by MD simulation (NAMD training set).

Random ions training set: To model this 150 mM NaCl concentration into the training dataset, 90 randomly generated sodium were placed in a $100 \text{ \AA} \times 100 \text{ \AA} \times 100 \text{ \AA}$ box (Fig. 1.1 A), which includes 8 bound sodium ions ($8/90 < 10\%$ for a bigger solvent box). In the generation process, the minimal distance between sodium ions was set as 5 \AA due to the exclusion of the same charged ions. 82 randomly generated ions represented free ions while the 8 bound ions represented the ions that were trapped on the surface or in the cavities of the biomolecules. After this, the simulation started for 1000 steps. Each step allowed free ions to move anywhere in the box and restrained bound ions in $5 \text{ \AA} \times 5 \text{ \AA} \times 5 \text{ \AA}$ cubic binding sites. The frames for every step were saved (Fig. 1.1 A). All of 1000 frames were combined into an ionic cloud distribution for further analysis (Fig. 1.1 B and Fig. 1.2 C and D), where the binding sites were marked by black circles. This ionic cloud distribution is the Random ions training set.

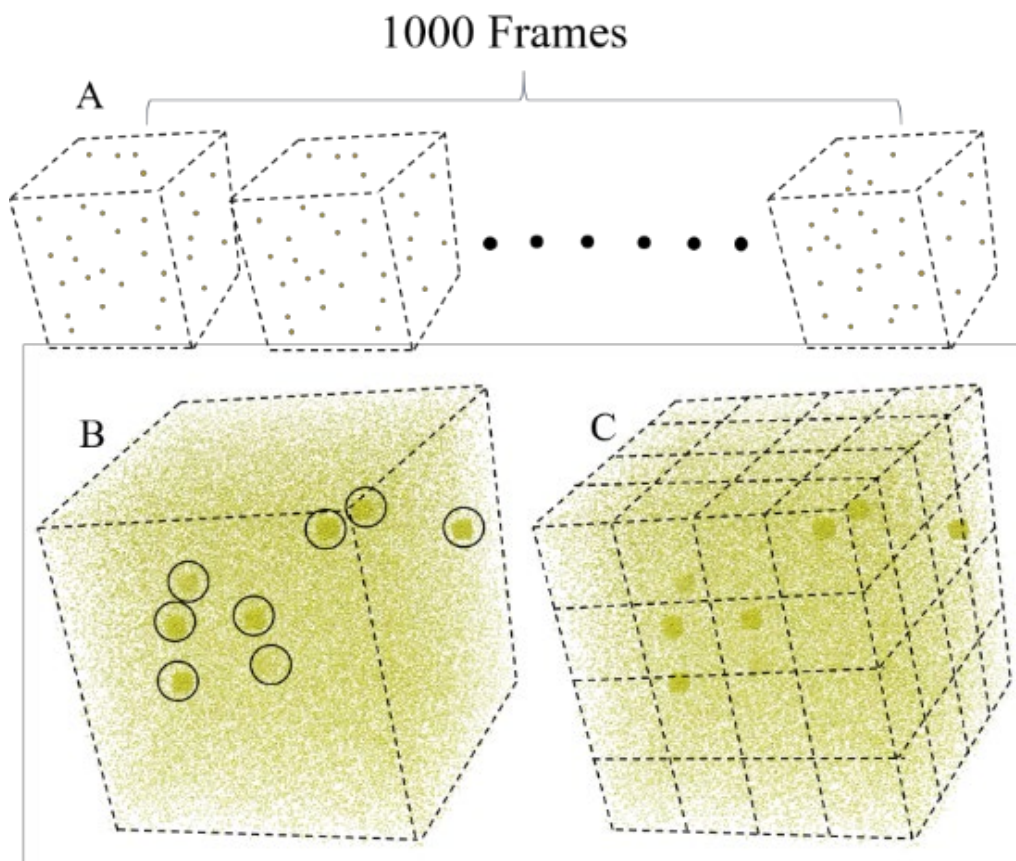


Figure 1.1 A: Schematic presentation of 1000 frames of ions; B: Ionic cloud distribution of the combination of 1000 frames from A; binding sites are marked by black circles; C: Cubic partition of the ionic

NAMD training set: In Visual Molecular Dynamics (VMD)(Humphrey, Dalke, & Schulten, 1996), $100 \text{ \AA} \times 100 \text{ \AA} \times 100 \text{ \AA}$ solvated box with 150 mM NaCl were generated, which included 90 sodium ions and 90 chlorine ions. In this model, 8 sodium ions were randomly selected to be restrained, simulating 8 bound ions in 8 binding sites. After that, MD simulation was achieved by 1000 steps minimization and 0.5 ns (2fs/step) MD simulation. The temperature was set as 300K and the pH was set as 7.0. The CHARMM(Vanommeslaeghe et al., 2010) was used for the force field and the periodic boundary conditions were applied to the system. The frames were saved per 100 fs. After simulations, the ionic cloud distribution for all ions was saved as the NAMD training set.

1.2.1.2 Myosin and kinesin testing sets

In the myosin testing set, five G-actins, a TM (PDB: 5NOJ), and a β -cardiac myosin motor domain (PDB: 6FSA) were assembled based on the rigor-like state model (PDB: 5JLH) (Fig. 1.2 A). The hydrogen atoms were added by VMD. The myosin-actin complex was immersed in a rectangular solvated box (TIP3P). The net charge of the myosin-actin complex model is -77 e. To ensure 150 mM NaCl and to neutralize the system, 570 Na^+ and 493 Cl^- were added into the system. In NAMD simulations, the pH was set as 7.0 and the temperature was set as 300K. The CHARMM was used for the force field and the periodic boundary conditions were applied to the system. The number of steps for energy minimization was 20,000, and the MD simulation was run for 10 ns (1fs/step). After simulation, the sodium ions within 10 Å from proteins in 2000 frames (5000 steps/frame) were assembled as an ionic cloud distribution in the myosin testing set (Fig. 1.2 C).

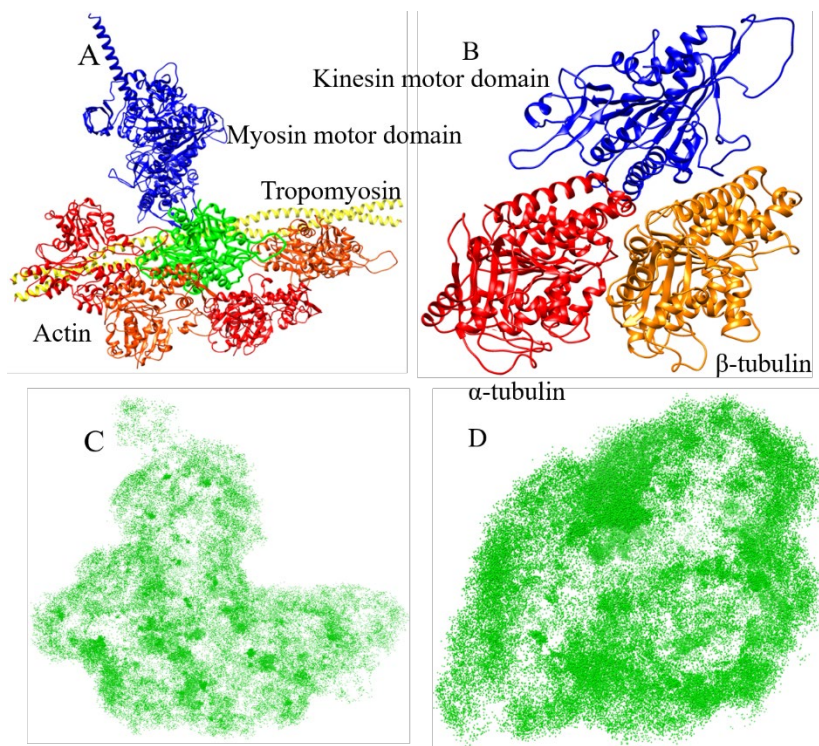


Figure 1.2 Myosin-actin complex (A) and Kinesin-tubulin dimer complex (B) and their sodium ionic cloud distribution. C and D are the sodium ionic cloud distribution of myosin complex (A) and kinesin complex (B) within 10 ns simulation (1000 frames).

Similarly, a complex formed by an $\alpha\beta$ -tubulin heterodimer and a kinesin motor domain (kinesin-5 cut7) was selected as a kinesin testing set (PDB: 5MLV) (Fig. 1.2 B). The hydrogen atoms were added by VMD. The complex was immersed in an explicit solvated box (TIP3P). To ensure 150 mM NaCl and to neutralize the system, 143 Na^+ and 107 Cl^- were added into the system. The MD simulation setting is the same as that of the myosin testing set. After simulation, the sodium ions within 10 Å from proteins in 2000 frames (5000 steps/frame) were assembled as an ionic cloud distribution in the kinesin dataset (Fig. 1.2 D)

The reason why only the sodium ions within 10 Å of protein were selected is that some ions located far away from biomolecules are relatively rigid due to the lack of strong electrostatic forces. If all ions were selected, they could generate a lot of noise, affecting the accuracy of the calculation.

1.2.2 Algorithm

Our Hybridizing Ions Treatment (HIT) method utilizes the frequency of ions occurrence around biomolecules to identify possible binding sites and, to place explicit ions at the centroids of calculated binding sites to compensate the net charge. Ion binding sites are around the molecule surfaces or inside the molecule cavities, where ions will be trapped. After MD simulation, overlapping the frames of ions (Fig. 1.1 A and B) into ionic cloud distribution generates some dense positions, which represent the binding sites with the high frequency of bound ions occurrence (Fig. 1.1 B, Fig. 1.2 C and D). The centroids in the dense positions are the locations for placing explicit ions. The whole process included 4 steps: preparation step, the initial step, clustering step, and optimal step, where the clustering step is the supplement for the initial step while the optimal step is the supplement for the clustering step. The training sets were generated by uniform distribution (random ions training set) and MD simulation (NAMD training set) respectively for accuracy testing and parameters optimization while the testing sets were generated by MD simulations of myosin-actin complex and kinesin-tubulin complex.

1.2.2.1 preparation: the combination of all frames and the solvate box cutting.

First, the target ions in each frame were assembled as ionic cloud distribution. To find the frequency of occurrence in the different areas, the ionic cloud distribution was randomly and equally divided into cubes (Fig. 1.1 B).

1.2.2.2 initial step: ions counting and cube sorting.

The ions are counted in all cubes, and the number of ions is used to sort the cubes from the maximum to the minimum. Then, all cubes are marked following the rank, as the 1st, 2nd, 3rd, cube (Fig. 1.3 (ions counting and cube sorting)).

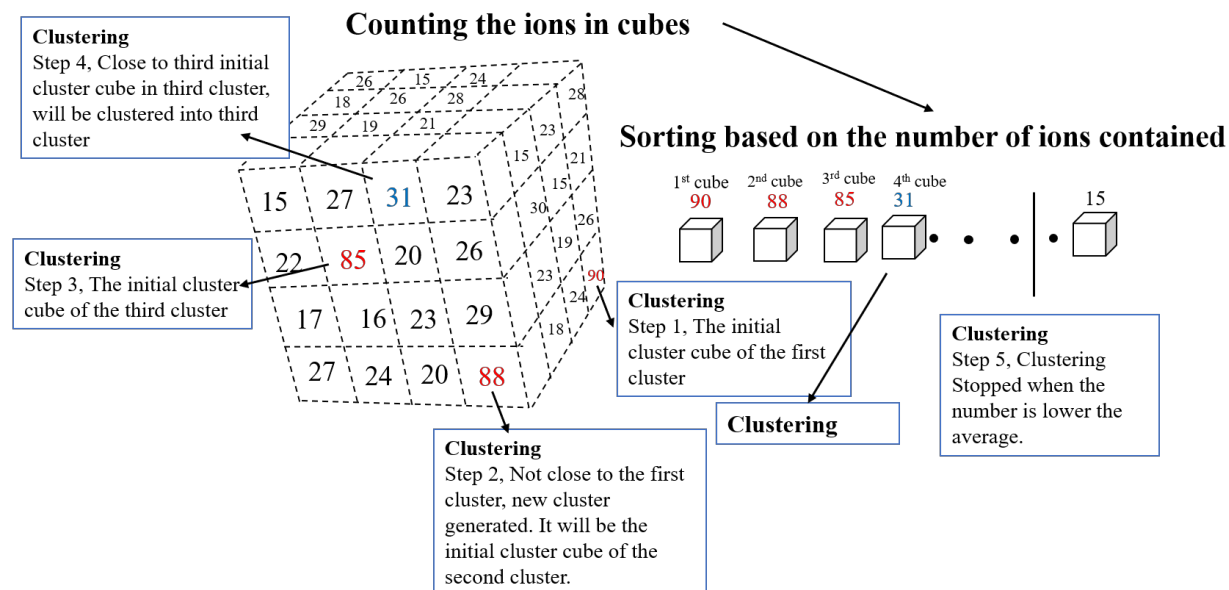


Figure 1.3 Diagram of ion counting, cube sorting, and clustering process. The number indicates the ions contained in the cube. Note that all numbers are invented for demonstration, and the $4 \times 4 \times 4$ cutting of the solvate box is for better visualization.

A certain number of top cubes after sorting are selected. Cubes with a high average number of ions (Eq. 1.1) indicate binding sites. The number of selected cubes is determined by the net charge of the protein and the charge of target ions (Eq. 1.5). The centers of top cubes are the positions for explicit ions placement based on the initial step.

$$N_{\text{ions}_i} = \frac{n_i}{N_{\text{frame}}} \quad (1.1)$$

Where the N_{ion_i} represents the average number of ions in the i^{th} cube and the n_i , N_{frame} represents the total number of ions in i^{th} cubes and the number of frames respectively.

1.2.2.3 clustering step: clustering

However, if the binding site selection is purely based on the initial step, it might cause a potential problem. A single binding site may cover multiple close cubes, which need to be clustered, as shown in Fig. 1.5 A. Otherwise, redundancy and incorrect calculations will happen (Fig. 1.5 A). The redundancy is two or more calculated binding sites nearby one original binding site while the incorrect calculation is the wrong calculated binding site. To solve this problem, the clustering method was implemented. The close cubes are clustered into a cluster, which contains 1-27 ($3 \times 3 \times 3$) cubes, where the first cube in a new cluster is the initial cluster cube (Rule 1). The clustering complies with the following rules and runs based on the rank of the cubes (Fig. 1.3 (clustering)).

Rule 1: If the cube is not close to any previous initial cluster cubes of previous clusters, a new cluster will be generated, and this cube will be the initial cluster cube in the new cluster. For example, the first cube in the rank is the initial cluster cube in the first cluster.

Rule 2: If the cube is close to previous initial cluster cubes of previous clusters, it will be clustered into the corresponding cluster, which contains the highest-average-number-of-ions initial cluster cube.

Rule 3: The clustering will stop when the average number of ions is lower than the average of all cubes.

After clustering, clusters are sorted based on the average number of ions (Eq. 1.2) from maximum to minimum and marked following the rank. The initial cluster cubes of top clusters are selected as binding sites, where the explicit ions will be placed at the center.

$$N_{\text{ion}_j} = \frac{n_j}{N_{\text{frame}}} \quad (1.2)$$

Where the N_{ion_j} represents the average number of ions in the j^{th} cluster, and the n_j and N_{frame} represent the number of ions in the j^{th} clusters and the number of frames respectively.

The centers of the top initial cluster cubes are the positions for placing explicit ions based on the clustering step (Fig. 1.4).

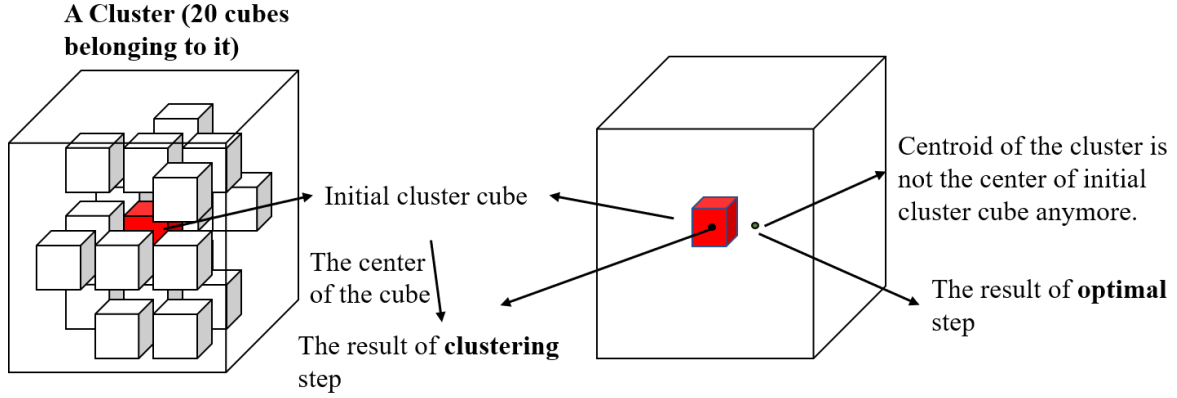


Figure 1.4 Diagram of the differences between the center of the initial cluster cube and the centroid of the clusters

1.2.2.4 optimal step: centroid optimization

Using the centroids of top clusters is a better option than the center of initial cluster cubes. The calculation of the centroid of the cluster is the optimal step, which is based on the weight ratio of cubes in clusters.

The centroid of the cluster is calculated based on the weight ratio R (Eq. 1.3). The centroid of the cluster (X, Y, Z) is calculated by equation (1.4)

$$R_i = \frac{n_i}{N_t} \quad i \in [1,27] \quad (1.3)$$

Where R_i is the weight ratio of the i^{th} cube in the cluster while the n_i, N_t represent the number of ions in the i^{th} cube and the total number of ions in the cluster.

$$X = \sum_{i=0}^n R_i X_i; \quad Y = \sum_{i=0}^n R_i Y_i; \quad Z = \sum_{i=0}^n R_i Z_i \quad n \in [1,27] \quad (1.4)$$

Where X, Y, Z are the coordinates of the centroid of the cluster. The R_i is the weight ratio of the i^{th} cube in the cluster while the X_i, Y_i, Z_i are the coordinates of the center of the i^{th} cube belonging to the cluster. The centroids of top clusters are the positions for placing explicit ions based on optimal step (Fig. 1.4).

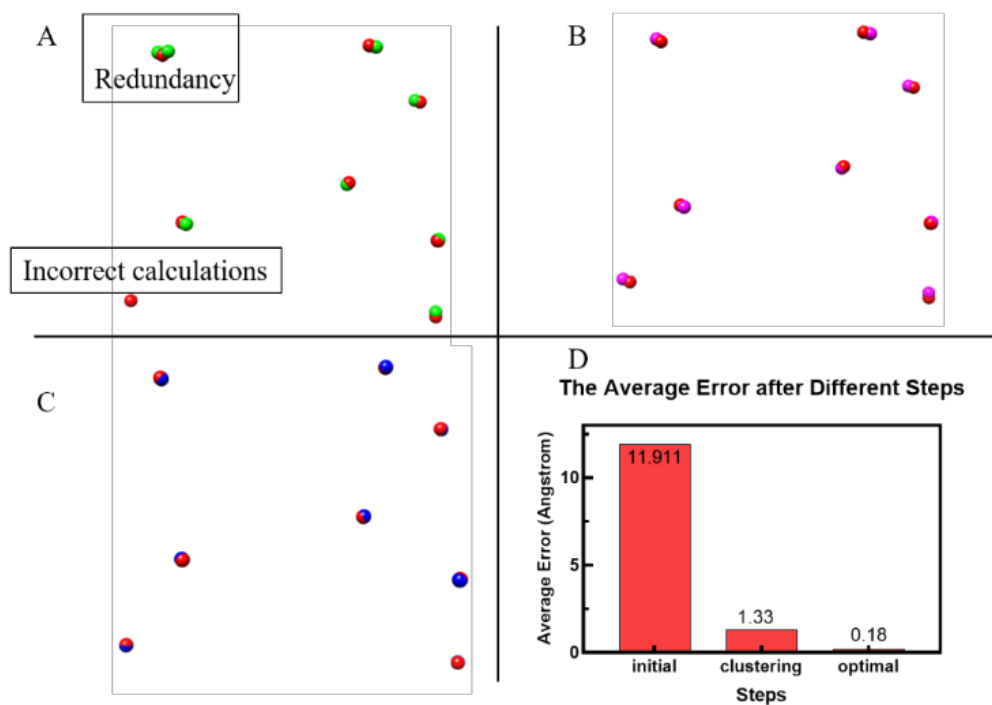


Figure 1.5 The original binding sites are marked as red (A, B and C) while the calculated binding sites after initial (A), clustering (B), and optimal (C) steps are marked as green, pink, and blue.

Average error (D) is calculated by the average distance.

1.2.2.5, The number of clusters selection.

Due to their sizes and charge distributions, the ionic binding sites may attract different numbers of ions. The distribution curves of an average number of ions on the testing set (Fig. S2) were not always as distinct as that of the training set (Fig. 6 A) when classifying clusters into two groups: true binding site predictions (high average number of ions) and false binding site predictions (low average number of ions). In this case, the net charge compensation is satisfied preferentially. It means the number of selected clusters multiplied by the charge of the target ion should equal the opposite net charge of the system (Eq. 1.5). The average number of ions represents the possibility of the occurrence of ions. The cluster selection should follow the rank of the average number of ions of clusters from the maximum to the minimum.

$$n_{ei} = \frac{N_{tc}}{N_{ti}} \quad (1.5)$$

Where the n_{ei} is the number of explicit ions. In another word, it is the number of clusters that should be selected. N_{tc} and N_{ti} are the net charges of the system and the charge of the target ions, respectively. However, the number of bound ions sometimes is smaller than the n_{ei} . Because when the surface of the protein is nearly neutral, some of the “bound” ions may be semi-bound ions, which are not always bound at their binding sites. The HIT method provides the average number of ions (occupancy) in each binding position (clusters), which can be utilized for users to set thresholds and to select the certain number of bound ions.

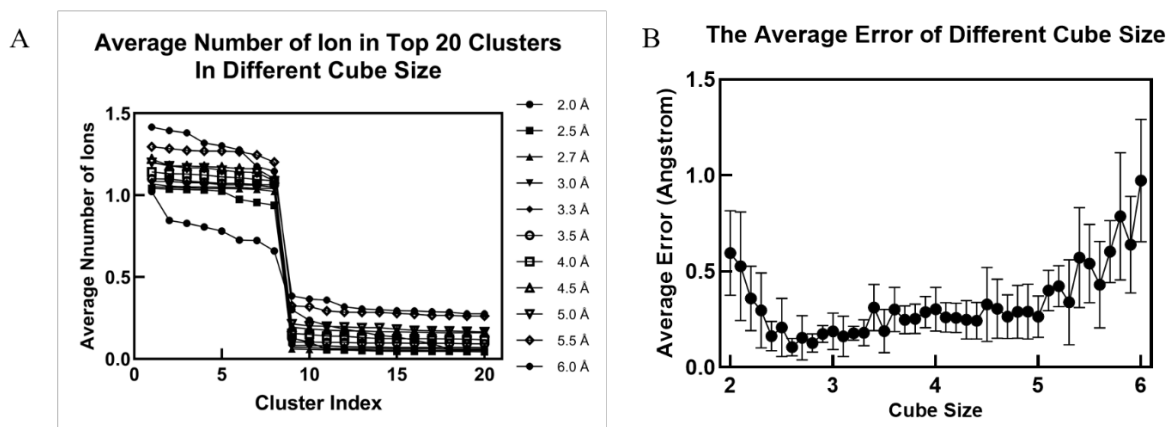


Figure 1.6 Average number of ions of top 20 clusters per frame with different cube sizes of 2.0 Å, 2.5 Å, 2.7 Å, 3.0 Å, 3.3 Å, 3.5 Å, 4.0 Å, 4.5 Å, 5.0 Å, 5.5 Å, and 6.0 Å, and the average error (B) based on different cube size of the range from 2.0 Å to 6.0 Å.

1.2.3 Training

1.2.3.1, The cube size optimization

The selection of cube size was tested from 2 Å to 6 Å with the interval of 0.1 Å in a random ion training set. The distance between the calculated binding site and the corresponding original binding site was regarded as the error. The average error of eight binding sites based on different cube sizes was used for comparison to determine the optimal cube size in HIT. The optimal cube size was analyzed, and the explanatory equation was discussed.

1.2.3.2, The simulation time and accuracy.

The sensitivity of HIT on MD simulation time was trained by the NAMD training set and kinesin dataset. The NAMD training set was split into 0.1 ns, 0.2 ns, 0.3 ns, 0.4 ns, and 0.5 ns simulations to test the minimal simulation time for the HIT method to successfully find all binding sites. The result of HIT on 10 ns MD simulation of kinesin dataset was used as the reference to compare the result of HIT on 1 ns, 2 ns, 8ns, 9ns simulation of kinesin. This experiment was applied to show the stability of HIT on MD simulation time and minimal necessary simulation time for HIT in real cases.

1.2.4 Testing

After the training, the optimized program was applied to the testing set to verify its functions in real biological cases. For the myosin dataset with the net charge of $-77e$, 77 sodium ions were placed at the centroids of the top 77 sodium clusters to neutralize the net charge. The myosin motor domain together with surrounding explicit ions was separated from actin filament by 20 Å for better visualization of electrostatic surface and electrostatic field lines (C. Li et al., 2019; C. Li, Li, Zhang, & Alexov, 2012; C. Li, Petukh, Li, & Alexov, 2013; L. Li et al., 2012). Similarly, the 36 sodium ions were placed at the centroids of the top 36 sodium clusters in the kinesin dataset. The following steps are as same as that of the myosin dataset.

The electrostatic potential maps of myosin-actin complex and kinesin-tubulin complex were generated by Delphi. The electrostatic potential on the surface was visualized by Chimera (Pettersen et al., 2004). To visualize interactions, electric field lines were rendered by VMD (Humphrey et al., 1996).

1.3. Results and discussion

1.3.1 Accuracy

The program is optimized based on the training set.

1.3.1.1 Comparison among initial, clustering, optimal steps based on Random ions training set.

In Fig. 1.5 A, B, and C, the red balls represent the center of the original binding sites while the green, pink and blue balls represent the calculated binding sites based on the results after the initial step, clustering step, and optimal step respectively. There were eight binding sites in the training

set, where the distance between each calculated binding site and the corresponding original binding site is regarded as the error of this binding site. Fig. 1.5 D demonstrates the average errors of the 8 binding sites after the initial, clustering, and optimal steps.

The initial step (Fig. 1.5 A) has two problems: redundancy and incorrect calculation. The redundancy is two or more calculated binding sites nearby one original binding site while the incorrect calculation is the wrong calculated binding site. Both redundancy and incorrect calculation problems are caused by the unexpected partition of binding sites in the ionic cloud distribution. When the ionic cloud was cut into cubes, some original binding sites were approximately equally divided into several cubes. In this case, these cubes yielded a similar average number of ions for each cube (Eq. 1.1) and were comparable in rank. The initial step took these cubes as several binding sites, which should be a single binding site. This is the reason for the redundancy problem (Fig. 1.5 A). Besides, the redundant cubes also took the spot of other binding sites with a lower average number of ions in each cube, causing incorrect calculations (Fig. 1.5 A). In detail, some binding sites were divided into too many cubes, diluting the average number of ions, causing the low-ranking situation, resulting in incorrect calculations. Thus, clustering adjacent cubes into clusters to represent binding sites is a necessary step to avoid such problems.

The clustering step successfully recognized all 8 binding sites, as shown in Fig. 1.5 B. Because the clustering step combined adjacent cubes, covering each binding site into an individual cluster. After that, using the initial cluster cube (the first cube in a new cluster) of its cluster to represent the corresponding binding site avoided redundancy and incorrect calculations. Additionally, the clustering step was stopped when the average number of ions in cubes was lower than the average, avoiding over-clustering. However, there is still a distance between each calculated binding site and the corresponding original binding site. It was further optimized to reduce the error by the optimal step.

The optimal step is based on the ion's distribution in clusters, to use the centroids of clusters to represent the calculated binding sites. The ions distribution in clusters is represented by the weight ratio of cubes. The original binding sites are fully covered by calculated binding sites after the optimal step (Fig. 1.5 C). In Fig. 1.5 D, the average error after the optimal step (0.18 \AA) is far smaller than that after the clustering step (1.33 \AA), which demonstrates that this optimal step significantly improves the accuracy of the method.

1.3.1.2 cube size optimization based on Random ions training set.

The cube size is an important parameter for cutting the ionic cloud. The average error was benchmarked based on the cube sizes from 2.0 Å to 6.0 Å with the interval of 0.1 Å (Fig. 1.6 B). The cube sizes of 2.0 Å, 2.5 Å, 2.7 Å, 3.0 Å, 3.3 Å, 3.5 Å, 4.0 Å, 4.5 Å, 5.0 Å, 5.5 Å and 6.0 Å were selected to show the average number of ions in the top 20 clusters (Fig. 1.6 A). From figure 6 A, the average number of ions of the first 8 clusters is far higher than the others. The pattern classified the clusters into true binding site predictions (high average number of ions) and false binding site predictions (low average number of ions). The average number of ions of the top 8 clusters is close to 1.0 (Fig. 1.6 A), revealing that each of these clusters, representing its corresponding binding site, always contains at least an ion trapped in the area, which is consistent with our setting in the training set. By contrast, clusters with a much lower average number of ions (the tail of curves (Fig. 1.6 A), clustering the non-binding-site-related cubes, should be abandoned.

The ideal cube size should satisfy the full coverage of the binding site by clusters. In another word, the average number of ions in each cluster of calculated binding sites should be 1.00 and the area should be the same as binding sites. In the experiment, the minimal average error appeared during the cube sizes from 2.6 Å to 3.5 Å (Fig. 1.6 B). In the training set, the minimal distance between ions is 5 Å, so the side length of a cubic binding site is 10 Å, thus the best selection of cube size should be 3.3 Å (Fig. 1.7) to satisfy the full coverage of binding sites by clusters, leading to the optimal cube size selection equation (Eq. 1.4).

$$L = \frac{2 \times D}{3} \quad (1.4)$$

Where the L is the optimal length of the cube (cube size) and the D represents the minimal distance between ions. The theoretical optimal cube size of 3.3 Å appeared in the range from 2.6 Å to 3.5 Å, consisting of the experiment. In Fig. 1.6 A, the average number of ions of the top 8 clusters with cube sizes from 2.7 Å to 5.0 Å show stable lines slightly over 1.0, while the curves of cube size smaller than 2.5 Å or over 5 Å are very unstable. The instability is caused by defective clustering, due to the improper cube sizes. In summary, the cube size of 3.3 Å is the optimal selection, which was used for the analysis of the testing set.

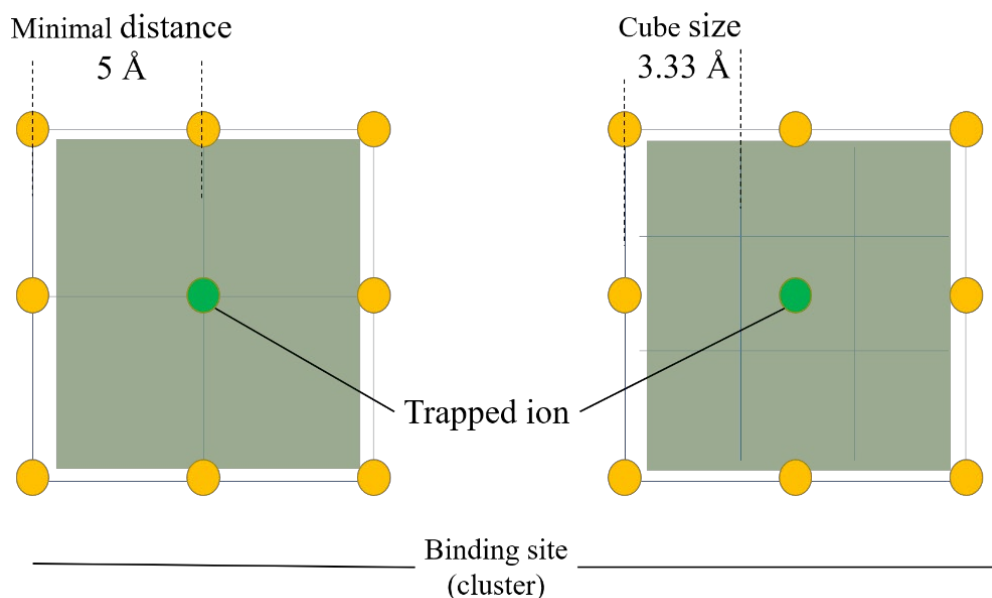


Figure 1.7 The diagram of binding area and cube size selection (Eq. (1.6)).

1.3.1.3 Effects of simulation time on the accuracy

Since the HIT is based on the ionic information from MD simulation, the running time of MD simulation is crucial for HIT to get reliable results. In theory, the MD simulation should be as long as possible to get accurate results for applying HIT. However, running an infinite MD simulation is impossible and impractical. Hence, the necessary simulation running time is demanded applying HIT. Here NAMD training set was applied to test how long the MD simulations can provide enough information for HIT to successfully identify all binding sites. The number of successfully identified binding sites divided by the total number of original binding sites is the accuracy (Fig. 1.8). The wrong calculated binding sites are regarded as incorrect calculations (Fig. 1.9). In the real case of the kinesin dataset, we used the 36 calculated binding sites from HIT based on 10 ns simulations of kinesin as references (Fig. 1.10). The number of calculated binding sites from HIT based on certain simulation running time, divided by total number of original binding sites is the coverage (Fig. 1.10).

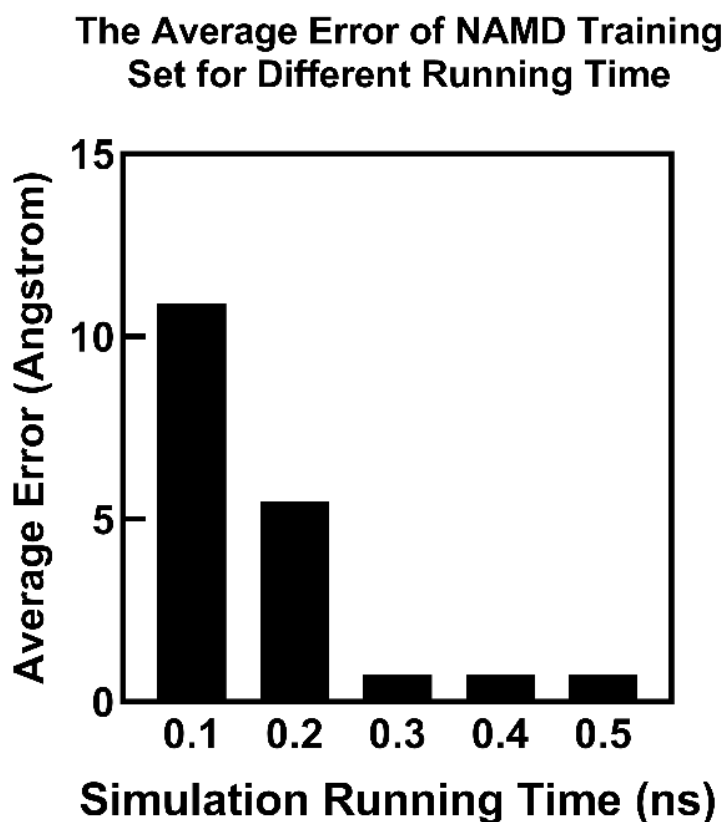


Figure 1.8 The average error of the calculated 8 binding sites for different simulation times of the NAMD training set.

It is obvious that two incorrect calculations happened in the 0.1 ns simulation while one incorrect calculation happened in the 0.2 ns simulation (Fig. 1.9 A and B). After 0.3 ns simulation, all binding sites were identified correctly by the HIT method (Fig. 1.9 C, D, and E). The average error was reduced from 11 Å to 0.73 Å (Fig. 1.8) with the simulation time from 0.1 ns to 0.5 ns. This result shows that the HIT method only needs 0.3 ns simulation to achieve 100% accuracy. In the real case (Fig. 1.10), the HIT method achieved stable results (70%-75% coverage) when the simulation is longer than 4 ns (Fig. 1.10). The 25%-30% coverage loss is because that when the surface of the protein is nearly neutral, some of the “bound” ions may be semi-bound ions, which are not always bound at their binding sites. Before 4 ns, the coverage increased with the simulation time while the coverage kept stable after 4 ns. It means that the HIT method is non-sensitive to simulation time after 4 ns. Note that the necessary simulation running time of the kinesin dataset

(4 ns) is different from that of the NAMD training set (0.3 ns). Because the binding sites in the NAMD training set are stable and strong binding sites. However, in real cases such as the kinesin dataset, not all the binding sites are stable and strong. Although there are weak binding sites in real cases, the HIT still achieved 70%-75% coverage and stabilized after 4 ns. In a nutshell, to get reliable results from the HIT method, a 4 ns simulation is enough, but the longer simulation is always recommended.

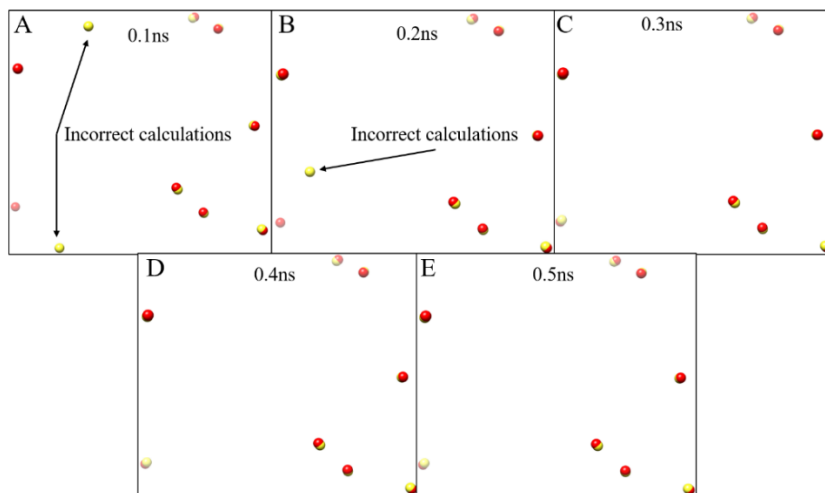


Figure 1.9 The figures of the NAMD training set for different simulation running times where the red balls represent the original binding sites, and the yellow balls represent the calculated binding sites. A, B, C, D, and E represent the original binding sites vs calculated binding sites.

The coverage of finally calculated bound ions

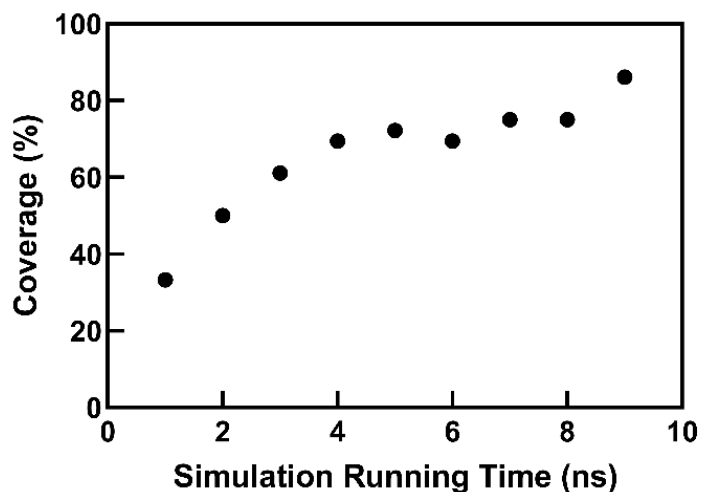


Figure 1.10 The sensitivity study of the calculated binding sites based on the Kinesin dataset.

The coverage is the ratio of the number of correctly identified calculated binding sites (Distance smaller than the length of the side of binding sites) divided by the number of total calculated binding sites. Reference is the result of HIT based on 10ns simulation.

1.3.2. Testing and applications

Two biological applications were tested by HIT, including a myosin-actin complex and a kinesin-tubulin complex. In the MD simulations, even though no atoms were constrained or fixed, both complexes were still relatively rigid in the simulations (Fig. 1.11). The movement of the centers of the myosin-actin complex was 4.78 Å and that of the kinesin-tubulin complex was only 4.14 Å, respectively (Fig. 1.11), which is much smaller than the side length of the expected binding sites (10 Å Fig 1.7). Before the electrostatic calculation, the 20 Å separation on the interface (Fig. 1.13, 1.14, 1.15, and 1.16) was applied on both myosin-actin complex and kinesin-tubulin complex. The figures were visualized by Chimera(Petterson et al., 2004).

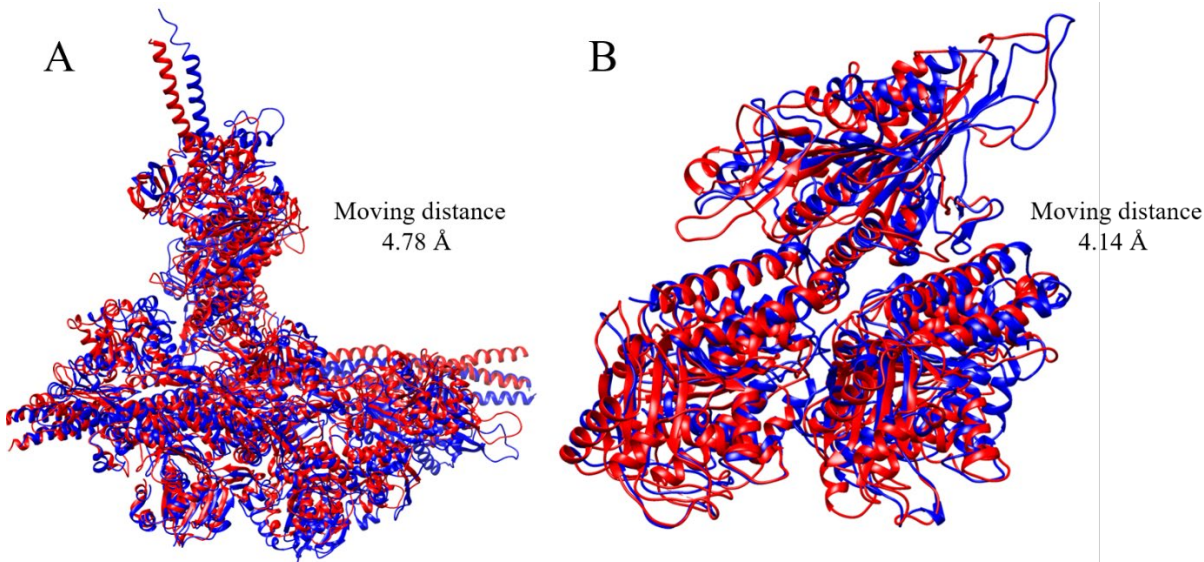


Figure 1.11 The comparisons of the proteins' position and structure before and after 10ns simulation of both myosin-actin complex and kinesin-tubulin complex (The red represents the position of proteins before simulation while the blue represents that after simulations).

1.3.2.1 The compensation of net charge.

As shown in Fig. 1.13 and Fig. 1.14, the comparison between the electrostatic surfaces calculated by the traditional method and the HIT method shows that the HIT method significantly improves the electrostatic calculations. In Fig. 1.13 A and C, the actin filament is highly negatively charged, where some positive ions should be bound surrounding the actin filament. The lack of bound positive ions causes unpredictable errors and bias in electrostatic calculations. By contrast, the HIT method (Fig. 1.13 B and D) added the bound ions based on the ionic cloud distribution, neutralizing the actin filament. Similarly, as shown in Fig. 1.14, the HIT method improved the electrostatic calculation of a highly negatively charged tubulin dimer. Without the HIT calculations, the electrostatic potential was calculated without any bound ions compensating the highly charged system. For such highly charged systems, it is unrealistic to have no bound ions. Additionally, we selected five positions (Fig. 1.12) to quantitatively compare the electrostatic potential calculations without and with HIT. These five points' potential values were calculated by DelphiForce(L. Li, Chakravorty, & Alexov, 2017). The results show that the average potential error was reduced from

0.93 to 0.53 kT/e in kinesin testing set by HIT (Fig. 1.12). Such bound ions added by HIT improve the electrostatic potential calculations and make them more realistic.

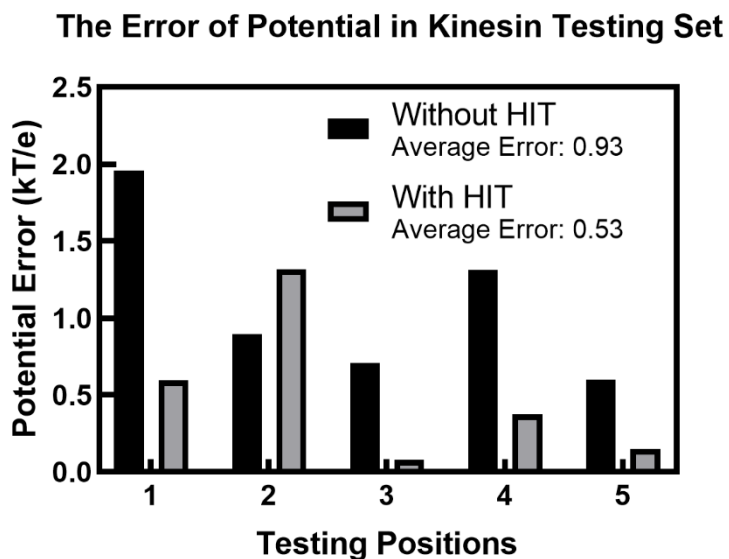


Figure 1.12 The errors of five testing positions' potential values without HIT and with HIT in the kinesin testing set. The testing positions for potential testing are 10 Å away from calculated binding sites. We took the top 5 positions without clashes with protein and ions to calculate potentials by DelphiForce. The potential of the five positions was tested based on the structure without explicitly bound ions (without HIT) and with explicitly bound ions (with HIT) for comparison. The error of a certain position's potential value is the difference between the potential in reference and that based on a certain method (without HIT and with HIT). The reference potential values are the average potential values of 200 frames (obtained from 5 ns – 10 ns simulations), where the proteins and all ions (Na⁺ and Cl⁻) were saved as the structure. The concentration of salt was set as 0.00 for the potential calculation of reference while that was set as 150 mM for both structures without explicitly bound ions (without HIT) and with explicitly bound ions (with HIT).

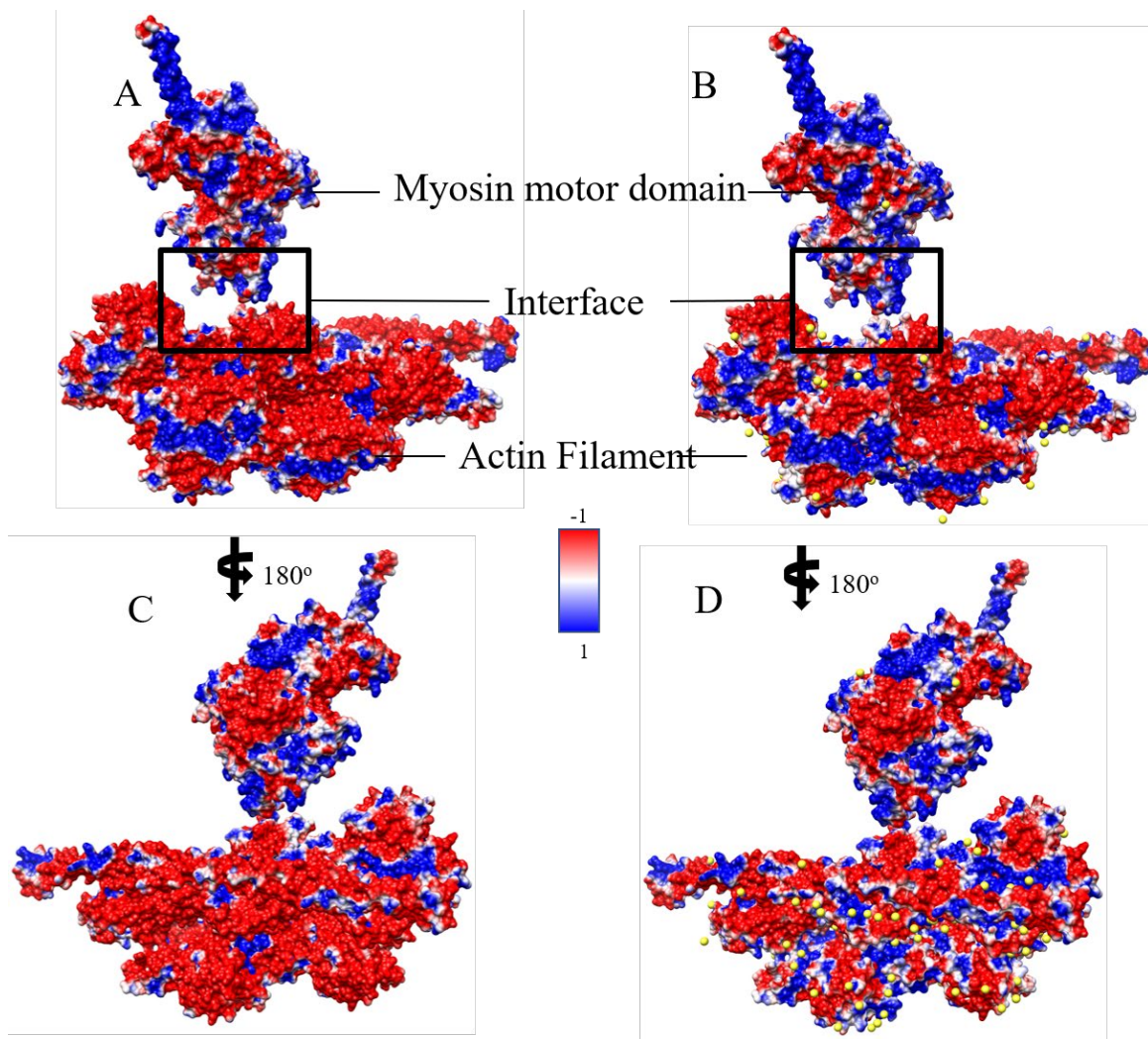


Figure 1.13 Electrostatic surface representation of myosin dataset in front (A and B) and back side (C and D), in which A and C represent the electronic surface without explicit sodium ions (yellow balls) by the traditional method, and B and D represent the electronic surface with explicit sodium ions by the HIT. The images are rendered by Chimera with a color scale from -1.0 to 1.0 kT/e.

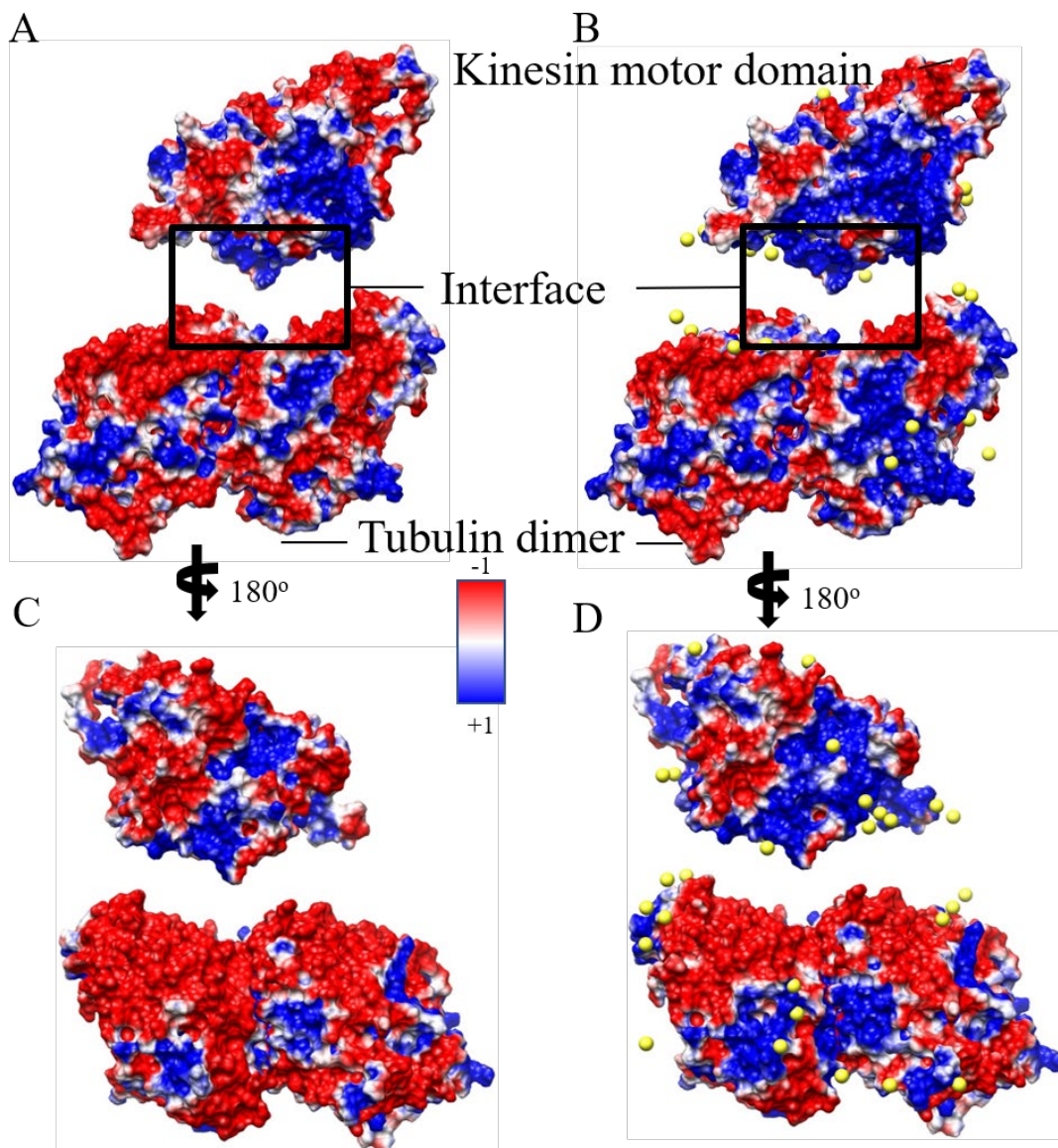


Figure 1.14 Electrostatic surface representation for kinesin dataset in front (A and B) and back side (C and D), in which A and C represent the electronic surface without explicit sodium ions (yellow balls) by the traditional method, and B and D represent the electronic surface with explicit sodium ions by the HIT. Images were rendered by Chimera with a color scale from -1.0 to 1.0 kT/e

Fig. 1.15 illustrates the details on the interfaces of the myosin-actin complex. The myosin-binding interface is positive while the actin filament binding interface is highly negative. Such electrostatic

distributions generate attractive forces between the myosin motor domain and actin filament. With the traditional method, the interface of myosin is positive, as shown in blue regions in Fig. 1.15 A and C. The HIT method added bound ions, which enlarge the positive area on the interface of myosin (Fig. 1.15 B and D). Such bound ions may enhance the binding forces between the myosin and actin filament. The actin filament is highly negatively charged, the HIT method added bound ions, therefore, shrinks the negative binding surfaces on the actin filament. In previous studies, the adjustment of the binding process is widely accepted and controlled by myosin-binding protein C (Burghardt & Sikkink, 2013). It is regarded to happen during the pre-power stroke state (Houdusse & Sweeney, 2016; Llinas et al., 2015). The shrunk negatively charged surface may be also related to the adjustment of the binding process, making the binding more specific.

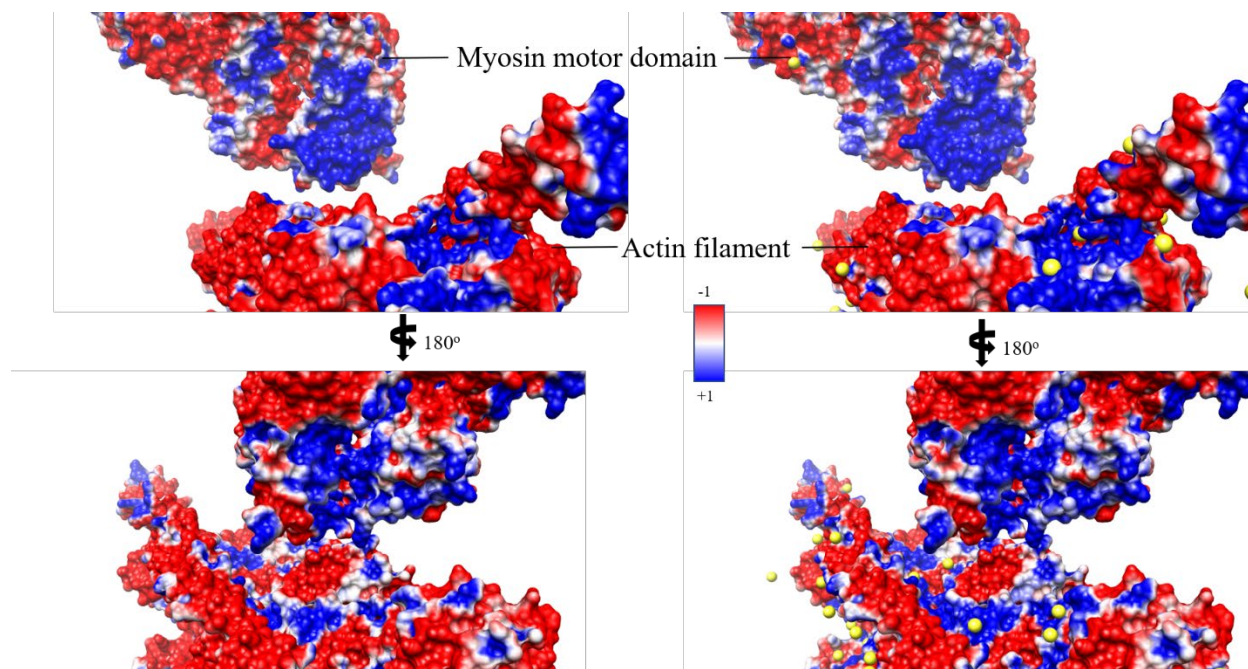


Figure 1.15 Electrostatic surface representation of the interface between myosin motor domain and actin filament in two directions. A and C represent the electronic surface without explicit sodium ions (yellow balls) by the traditional method, and B and D represent the electronic surface with explicit sodium ions by the HIT. The images are rendered by Chimera with a color scale from -1.0 to 1.0 kT/e.

Fig. 1.16 illustrates the details on the interfaces of the kinesin-tubulin dimer complex. As shown in the traditional method (Fig. 1.16 A and C), the interface of the kinesin motor domain is positively charged while that of tubulin dimer is negatively charged. The oppositely charged interfaces generate intensive binding forces. Like the effects on myosin, the HIT method also enlarged the positive area on the kinesin interface by adding bound ions (Fig. 1.16 B and D), strengthening the binding force on the interface. On the other hand, added bound ions shrunk the positively charged area of the interface of the tubulin dimer, enhancing the specificity of the binding site.

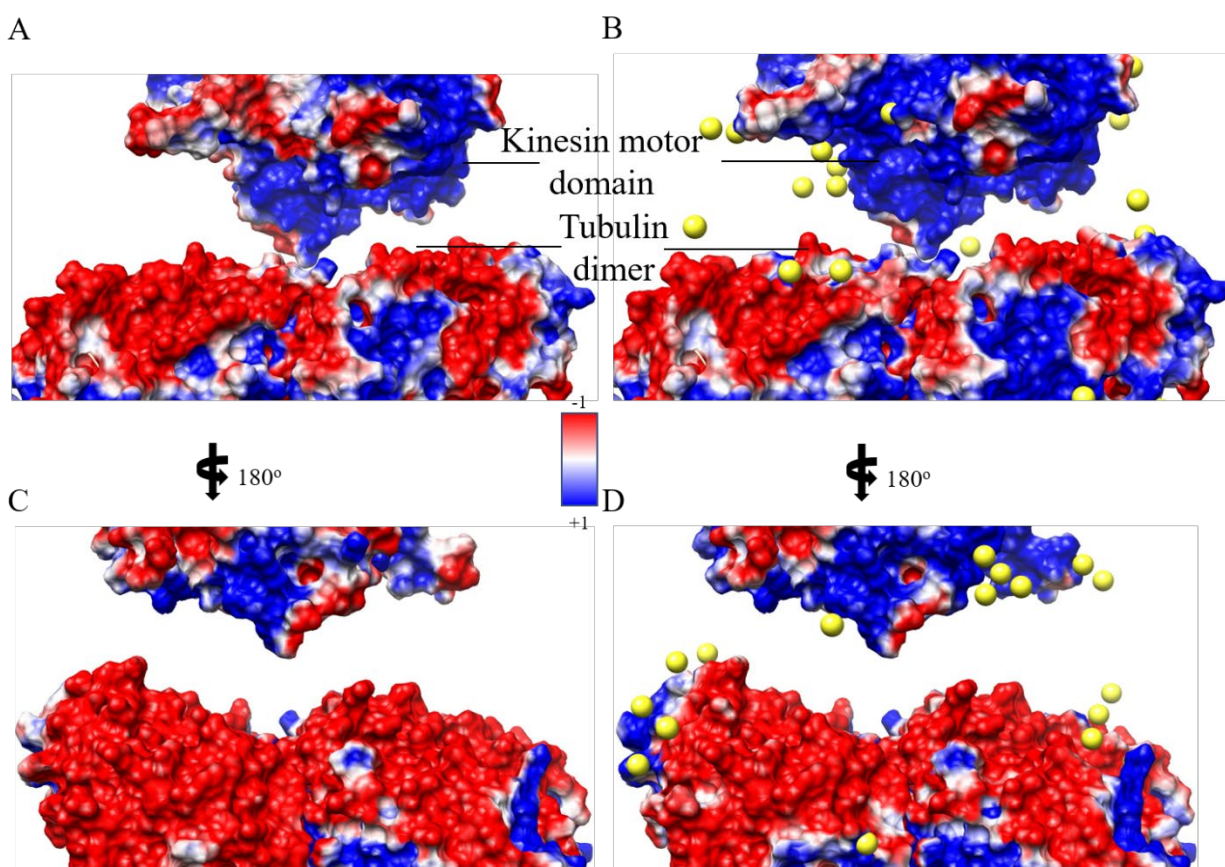


Figure 1.16 Electrostatic surface representation for the interface between kinesin motor domain and tubulin dimer in two directions. A and C represent the electronic surface without explicit sodium ions (yellow balls) by the traditional method while B and D represent the electronic surface with explicit sodium ions by the HIT. The images are rendered by Chimera with a color scale from -1.0 to 1.0 kT/e.

3.2.2 The interactions between myosin motor domain and actin filament

The myosin motor domain and actin filament were separated by 20 Å for better visualization of electrostatic field lines (Fig. 1.17). The electrostatic figures were rendered by VMD. The density of the electrostatic field lines represents the strengths of interactions between proteins. The actin filament includes G-actin and tropomyosin (TM). The interfaces between the myosin motor domain and TM have intensively attractive interactions (Fig. 1.17 Left) while the interfaces between the myosin motor domain and G-actin yield much weaker interactions (Fig. 1.17 Right). In previous TM studies (Yamada et al., 2020), the traditional opinion about the movements of TM includes three states: open, close, and block states. They are regulated by Ca^{2+} activating troponin (Tn) to shift the position of TM. In some studies, myosin induces another movement of TM, which is about 10° (Risi et al., 2017) or 23 Å (Behrmann et al., 2012), after myosin binding to the actin filament. The interactions between the myosin motor domain and TM provide evidence of the myosin-regulated movement of TM (Behrmann et al., 2012). By contrast, there are no distinct electrostatic interactions between the myosin motor domain and G-actins (Fig. 1.17 Right), which are regarded as the main binding sites for the myosin motor domain. Julian von der Ecken showed that the HLH motif of myosin enters the hydrophobic groove between actins to generate a strong binding force (von der Ecken, Heissler, Pathan-Chhatbar, Manstein, & Raunser, 2016). Additionally, the main binding part of cardiomyopathy loop (CM-loop) is also mainly stabilized by hydrophobic interactions (von der Ecken et al., 2016). With the support from the electrostatic studies, it is more reliable that electrostatic force does not dominate the interaction between myosin and G-actins.

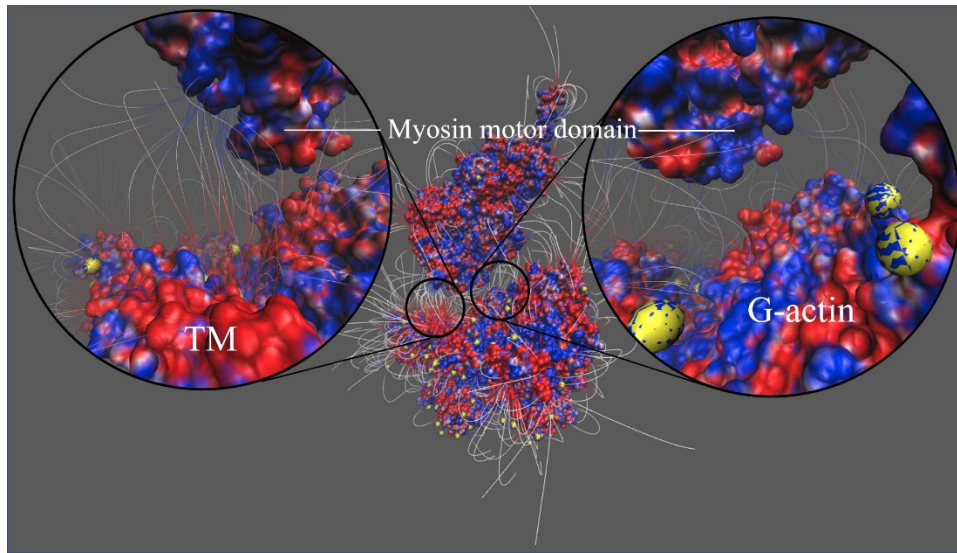


Figure 1.17 Electrostatic field line for the interface of myosin motor domain with tropomyosin (Left enlarged view) and actin (Right enlarged view). Yellow balls represent explicit sodium ions added by the HIT.

1.3.2.3 The interactions between kinesin motor domain and tubulin dimer.

The kinesin motor domain and tubulin dimer were separated by 20 Å to better visualize the electrostatic field lines (Fig. 1.18). There were two intensively attractive interactions between the cut7 kinesin motor domain and tubulin dimer. One is on the interface of cut7/ α -tubulin and the other is on the interface of cut7/ β -tubulin. The binding strengths of cut7/ α -tubulin and cut7/ β -tubulin are similar. It is consistent with the distribution of charges on the interface of the kinesin motor domain and tubulin dimer, as shown in Fig. 1.15 B and D. Most of the kinesins only interact with β -tubulin (Hunter & Allingham, 2020; Woehlke et al., 1997) while kinesin-5 (cut7) interact with both α - and β -tubulin. This may hint at the bidirectional characteristic of kinesin-5 (cut-7) (von Loeffelholz et al., 2019).

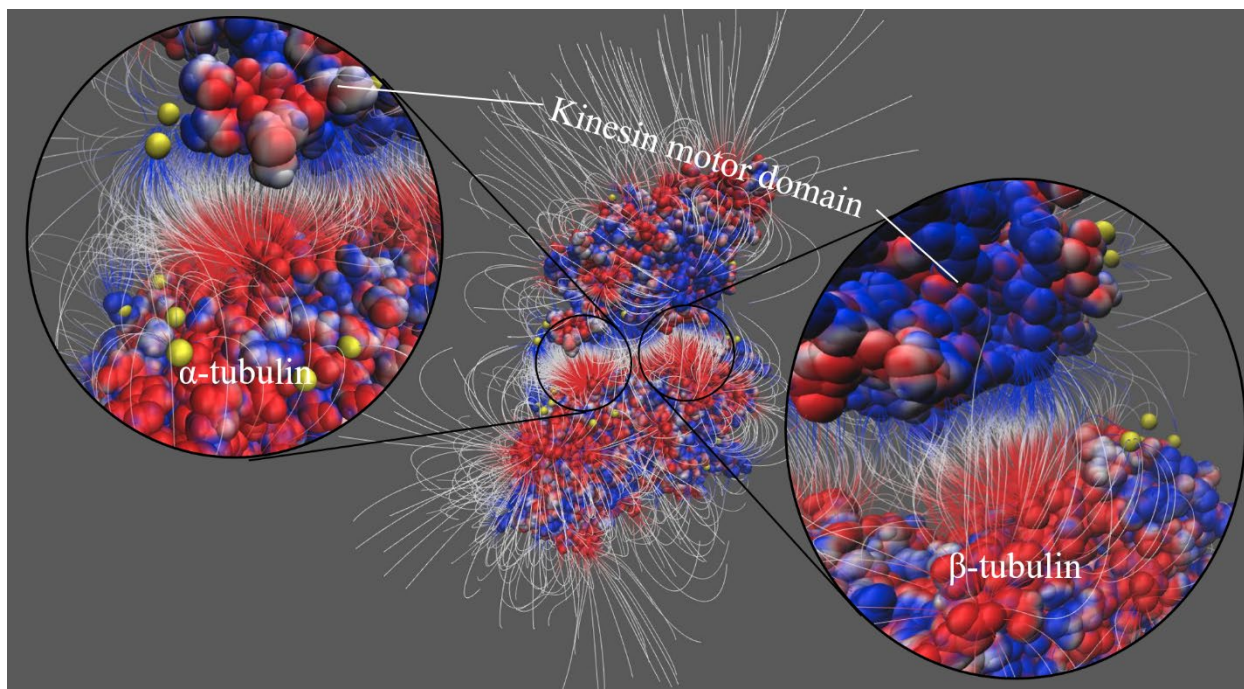


Figure 1.18 Electrostatic field line for the interface of kinesin motor domain with α -tubulin (Left enlarged view) and kinesin motor domain with β -tubulin (Right enlarged view). Yellow balls represent explicit sodium ions added by the HIT.

1.4. Conclusions

Ions are important to balance the net charges of highly charged biomolecules and bounded ions are crucial for the functions of highly charged biomolecules, such as DNAs, RNAs, and other biomolecules. In computational simulations, treating the ions properly is a challenging and essential task. In this work, we developed a novel method, the Hybridizing Ions Treatment (HIT) method, which hybridizes the implicit solvent method and explicit method to realistically calculate the electrostatic potential for highly charged biomolecules.

The implementation of this HIT method on two multiprotein complexes shows that this method improves the electrostatic calculations significantly. It predicts the positions of bound ions and then utilizes the bound ions to neutralize the biomolecules, therefore provides more realistic electrostatic calculations. The electrostatic interaction between actin filament and myosin motor domain proved that the electrostatic interactions between the myosin motor domain and TM is stronger than that between the myosin motor domain and G-actin, revealing the mechanism of the

myosin-regulated motion of TM, which has been observed by experiments (Behrmann et al., 2012; Risi et al., 2017). The interaction between cut7 kinesin motor domain and tubulin dimer was calculated, which demonstrated that the binding strengths of cut7/ α -tubulin and cut7/ β -tubulin are similar. Such similar electrostatic binding interactions may be a factor of the bidirectional motility feature for cut7(von Loeffelholz et al., 2019).

Besides the two applications in this work, the HIT method is also useful and helpful to be applied in many other studies related to highly charged biomolecules, including DNAs, RNAs, molecular motors, and other biomolecules. In this work, we just took Na^+ as the testing case. The performance of HIT is independent of ion types. Because HIT utilizes the information of ions distribution from MD simulations to analyze which ions are bound ions. Current MD algorithms treat different types of ions very reliably. Therefore, the HIT can handle the other types of ions as well as Na^+ . However, the limitation of HIT is that it can only be applied to biomolecules that don't have large conformational changes. For structures that have large conformational changes, more comprehensive algorithms are demanded to predict the bound ions. In our testing sets, the proteins' conformations did not change a lot. Our future work will focus on the problems of biomolecules that undergo large conformational changes. The supplementary material and the data in this work (including training set and results) are available online. The HIT package is available at http://compbio.utep.edu/static/downloads/download_hit.zip

Chapter 2: The computational study on the function of palmitoylation on the envelope protein in SARS-CoV-2

2.1 Introduction

Coronaviruses (CoVs) are notorious as the pathogens of numerous diseases in a wide range of vertebrates, including human beings. In 2002, the SARS-CoV-related severe acute respiratory syndrome (SARS) took 4 months to overwhelm 29 countries and killed 774 reported patients, and it almost paralyzed the Asian economy(De Wit, Van Doremalen, Falzarano, & Munster, 2016; Stadler et al., 2003). After 10 years, the outbreak of a new coronavirus, named the Middle East Respiratory Syndrome coronavirus (MERS-CoV) spread in 27 countries, causing a 35% death rate(De Wit et al., 2016; Organization, 2019; Park, Jung, & Kim, 2018). In December 2019, a novel coronavirus named Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) was discovered and now has already spread across the whole world. The disease caused by it is known as coronavirus disease 2019 (COVID-19)(Wu et al., 2020; Yang, Wang, & immunology, 2020). So far, the disease has infected 106 million people and caused 2.31 million deaths in the whole world. In addition to severe pneumonia, the SARS-CoV-2 can also cause multi-organ damages, including cardiovascular disease(Madjid, Safavi-Naeini, Solomon, & Vardeny, 2020), reproductive risk(Hall et al., 2020), mental illness(Yao, Chen, & Xu, 2020), and smell dysfunction(Moein et al., 2020). Thus it is urgent to understand the virulence mechanisms to prevent future outbreaks and to develop the remedies for COVID-19.

The SARS-CoV-2 is a positive RNA strand virus that originates from the family Coronaviridae(Ramphul & Mejias, 2020; Wu et al., 2020). The genome of coronavirus encodes four major structural proteins, which are the spike (S) protein, nucleocapsid (N) protein, membrane (M) protein, and the envelope (E) protein(Masters, 2006). The S protein mainly regulates the binding with the ACE2 (Angiotensin-Converting Enzyme) receptor of cells(Lan et al., 2020; Xie et al., 2020). The N protein and the M protein serve as RNA genome binding protein(de Haan & Rottier, 2005) and the most abundant structural protein of viral envelope, respectively(Neuman et al., 2011; Schoeman & Fielding, 2019). The E protein is the smallest among four major structural proteins. Other than the S, N, and M proteins, the E protein is abundantly expressed in the infected cells but only a small percentage is assembled into viral envelope(Venkatagopalan, Daskalova,

Lopez, Dolezal, & Hogue, 2015). The majority of the protein is located at the ER (endoplasmic reticulum), Golgi complex, and ERGIC (ER-Golgi intermediate compartment)(Nieto-Torres et al., 2011; Nieva, Madan, & Carrasco, 2012). The E protein is related to intracellular trafficking and participates in the viral assembly and budding(Venkatagopalan et al., 2015; Westerbeck & Machamer, 2019). The recombinant CoVs which lack E protein express crippled viral maturation and incompetent progeny(Curtis, Yount, & Baric, 2002; DeDiego et al., 2007; Ortego, Ceriani, Patiño, Plana, & Enjuanes, 2007). The study about the E protein is important to understand virulence mechanisms.

The SARS-CoV E protein contains a short hydrophilic terminal, a large hydrophobic transmembrane domain (TMD), and a long hydrophilic carboxyl end(Y. Li, Surya, Claudine, & Torres, 2014). Synthetic peptides of TMD of E protein can form dimers, trimers, and pentamers(Torres, Wang, Parthasarathy, & Liu, 2005), among which the pentameric structure is widely accepted and studied as viroporin(Nieva et al., 2012; Schoeman & Fielding, 2019; Venkatagopalan et al., 2015). Similarly, the TMD of SARS-CoV-2 E protein was also found to form a five-helix bundle surrounding a narrow pore(Torres et al., 2005). The narrow pore is suggested as viroporin to regulate multiple stages of viral life cycles. Viroporins can transport different ions but usually prefer positively charged ions (H^+ , K^+ , Na^+ , and Ca^{2+})(Mould et al., 2003; Nieto-Torres et al., 2015; Pham, Perry, Dosey, Delcour, & Hyser, 2017; Wang, Takeuchi, Pinto, & Lamb, 1993; Wilson, Mckinlay, Gage, & Ewart, 2004). As reported, the cell infected by coronavirus exhibited a remarkable increase of pH in the Golgi complex. The increased pH protects spike protein and promotes the release of the virus from the cell(Westerbeck & Machamer, 2019). This finding supports that the E protein of coronavirus could be H^+ channel, changing luminal pH and helping the release of the virus. To modulate protein functions, Post-Translational Modifications (PTMs) are necessary. The PTMs of the E protein of coronavirus include glycosylation and palmitoylation (Palm)(Fung & Liu, 2018; Schoeman & Fielding, 2019). N-glycosylation, as reported, appears in a minor form of SARS-CoV E protein, which has both C- and N- terminals exposed on the luminal side(Chen, Lo, Ma, & Li, 2009; Yuan, Liao, Torres, Tam, & Liu, 2006). Nevertheless, the E protein with both C- and N- terminal on the cytoplasmic side is not modified by glycosylation. By contrast, all three cysteine residues (C40, C43, C44) in the SARS-CoV E protein are modified by palmitoylation(Liao, Yuan, Torres, Tam, & Liu, 2006). The mutant of these cysteines in E protein on mouse hepatitis virus A59 (MHV-A59) weakens the

ability to form virus-like particles(Boscarino, Logan, Lacny, & Gallagher, 2008; Lopez, Riffle, Pike, Gardner, & Hogue, 2008). Additionally, MHV was prone to degradation when carrying triple mutants(Lopez et al., 2008). These findings proved the importance of palmitoylation on the E protein, but the mechanism is still enigmatic.

To better understand how palmitoylation affects the function of coronavirus E protein, SARS-CoV-2 E protein pentameric structure was used as the reference (E protein without Palm). The palmitoylation was added on the cysteine residues 40, 43, and 44 to each monomer as the experimental group (E protein with Palm). Afterward, Molecular Dynamic (MD) simulations were run to observe the effects of palmitoylation. Besides, the topological and electrostatic studies were applied to analyze the function of palmitoylation. In the end, we simulated several H⁺ on two sides to detect the differences of the force on cations. These studies were summarized in the conclusion section to discuss the effects of palmitoylation on the function of passing ions on the E protein pentamer.

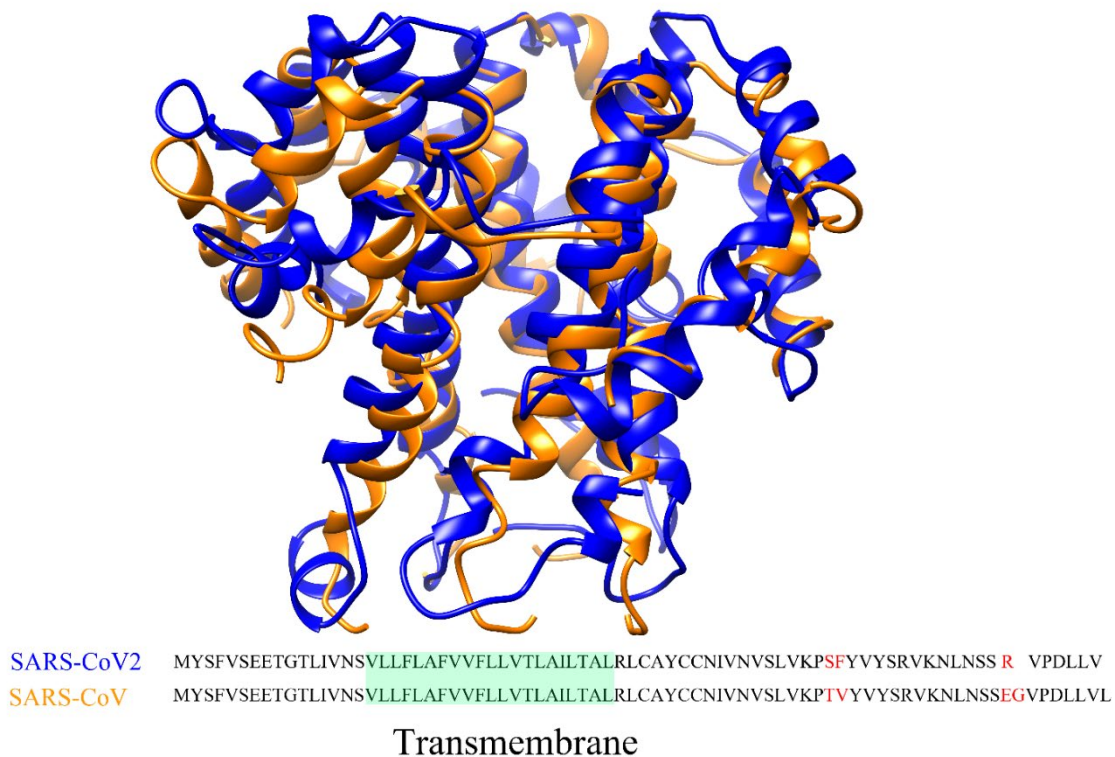


Figure 2.1 The structure alignment and the sequence alignment of SARS-CoV2 and SARS-CoV.

2.2 Methodology

2.2.1 Modeling and simulation

The sequence of E protein monomer of SARS-CoV-2 from 2019 was used as the query sequence (RefSeq: YP_009724392.1) (Fig. 2.1). The 3-dimensional pentameric structure of the E protein of the SARS-CoV-2 structure was based on the pentameric structure of the E protein of SARS-CoV of the PDB 5X29(Surya, Li, & Torres, 2018) and was achieved by Robetta server(Kim, Chivian, & Baker, 2004). Figure 2.1 shows the sequence and structural alignment of SARS-CoV-2 and SARS-CoV. The sequence similarity is 96 %. The modeled pentameric structure was used as the initial configuration (E protein without Palm) for the MD simulation. The palmitoylation was attached to the cysteine residues 40, 43, and 44 to each monomer (Fig. 2.2) as the E protein with Palm.

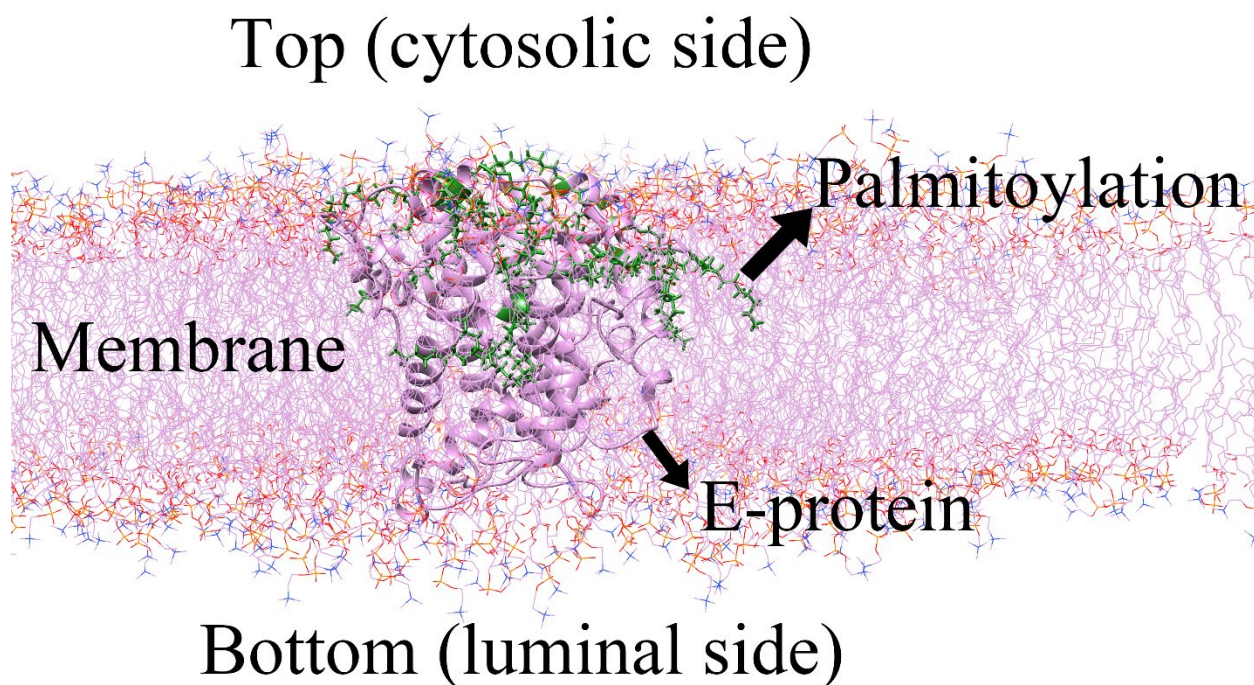


Figure 2.2 The diagram of E protein pentamer in the membrane.

The CHARMM-GUI(Jo, Kim, Iyer, & Im, 2008) webserver was used to create the simulating system (Fig. 2.2). The two structures were then embedded in the membrane dioleoyl phosphatidylcholine (DOPC) around the center along the z-axis. The membrane is a square with a side length of 150 Å long along the x and y-axis. The system was solvated with water of type

TIP3(Mackerell Jr, Feig, & Brooks III, 2004) of thickness 15 Å on either side of the membrane. The NaCl was used to ionize the system with a concentration of 150 mM. The temperature was set to 300 K and pressure to 1 atm. Parameterization of the atoms in the system was attained with the CHARMM36 force field(Huang & MacKerell Jr, 2013). Periodic boundary condition was applied to the simulating box and the Particle Mesh Ewald (PME)(Essmann et al., 1995) was used for the long-range electrostatic interactions. The final system was then subjected to the MD simulation with NAMD 2.12(Phillips et al., 2005). 10,000 steps were performed for minimization and 1ns for equilibration where constraints were imposed to the E protein complex and the head groups of DOPC lipids. Finally, 50 ns of the production running time was performed after removing constraints.

2.2.2 Topological study

The transmembrane domain (α -helix) of E protein pentamer is marked from 17V to 37L (21 residues) in five monomers for RMSF (Root-Mean-Square Fluctuations) analysis. The diagram (Fig. 2.2) showed that the cytosolic side (C-terminal) is on the top while the luminal side (N-terminal) is on the bottom.

The RMSF of the residues (21 residues) on the transmembrane domain is achieved based on the 50 ns simulation (500 frames) by Visual Molecular Dynamics (VMD)(Humphrey et al., 1996) (Eq. 2.1).

$$\text{RMSF}_i = \left[\frac{1}{T} \sum_{t_j=1}^T |r_i(t_j) - r_i^{\text{ref}}| \right]^{1/2} \quad (2.1)(\text{Van Der Spoel et al., 2005})$$

Where i represents the residue ID, the T represents the total simulation time (Here is the number of frames), $r_i(t_j)$ represents the residues i in the time of t_j position. The r_i^{ref} is the reference position of residue i , calculated by the time-average position.

The RMSD (Root-Mean-Square Deviation) is to measure the average distance between two protein structures, calculated by equation 2.2(Humphrey et al., 1996)

$$\text{RMSD}(t) = \left[\frac{1}{WN} \sum_{i=1}^N w_i |r_i(t) - r_i^{\text{ref}}|^2 \right]^{1/2} \quad (2.2)(\text{Van Der Spoel et al., 2005})$$

Where $W = \sum w_i$ is the weighting factor, and N is the total number of atoms. The $r_i(t)$ is the position of atom i at time t after least square fitting the structure to the reference structure. The r_i^{ref} is the reference position of residue i defined by the reference structure (Here we used the initial structure as the reference).

2.2.2.1 General comparison of the projections of the E protein's top and bottom views

The E protein pentamers with and without Palm were compared to show the conformational differences in both top and bottom views. The structures are from the last frame of the simulation. The observation plane is parallel to the membrane surface.

2.2.2.2 Minimal pore radius testing

The pore radius is tested by HOLE2(Smart, Neduvilil, Wang, Wallace, & Sansom, 1996) for the E protein pentamer with and without Palm in 50 ns simulation. The minimal pore radius is the minimum radius of the ion channel formed by the E protein pentamer. To avoid the influence from the C- and N- terminal residues, residues from 13I to 38R were extracted from the whole structure for the pore radius testing. To avoid the spherical probe ball running in the wrong direction, the moving direction is restrained to follow Z-axis(Smart et al., 1996), and the pore radius was tested six times for a single frame to take the maximum number of minimal pore radius.

2.2.3 Electrostatic study

The electrostatic potential calculations of E protein pentamer were calculated by Delphi(L. Li et al., 2012). The charge and the radius of atoms were calculated by the force field CHARMM36 and assigned by pdb2pqr(Dolinsky et al., 2007). The dielectric constants were set as 2 and 80 for protein and water respectively. The salt concentration was set as 150 mM, and the probe radius was set as 1.4 Å, and the protein filling percentage was set as 70%. The resolution was set as 1.5 grids/Å. The electrostatic potentials on the surfaces were visualized by Chimera(Pettersen et al., 2004) for both cytosolic (top view) and luminal sides (bottom view). The color range was set from -1.0 kT/e (red) to 1.0 kT/e (blue). The potential surfaces were visualized by Chimera(Pettersen et al., 2004) and the electric field lines were visualized by VMD(Humphrey et al., 1996) to demonstrate the interactions.

To quantitatively demonstrate the effects of the electrostatic force on the hydrogen ions, two sets of H^+ are simulated. One is a series of H^+ positions that were perpendicular to the membrane and

across the mass center of the E protein pentamer (Fig. 2.3). The other is the two layers of H^+ which were 10 Å far away from the membrane on both sides. To simulate the two layers of H^+ , the positions of H^+ at 10 Å on two sides were chosen as the center to draw several circles of H^+ for electrostatic force testing (Fig. 2.3). The radiuses of circles are from 2 Å to 15 Å. In a certain circle, for every 10° , a hydrogen ion is placed. The electrostatic force on H^+ was split to demonstrate the component force of the direction along the Z-axis. The component force direction was set from the bottom (luminal side) to the top (cytosolic side). The electrostatic force was calculated by DelphiForce(L. Li, Chakravorty, et al., 2017).

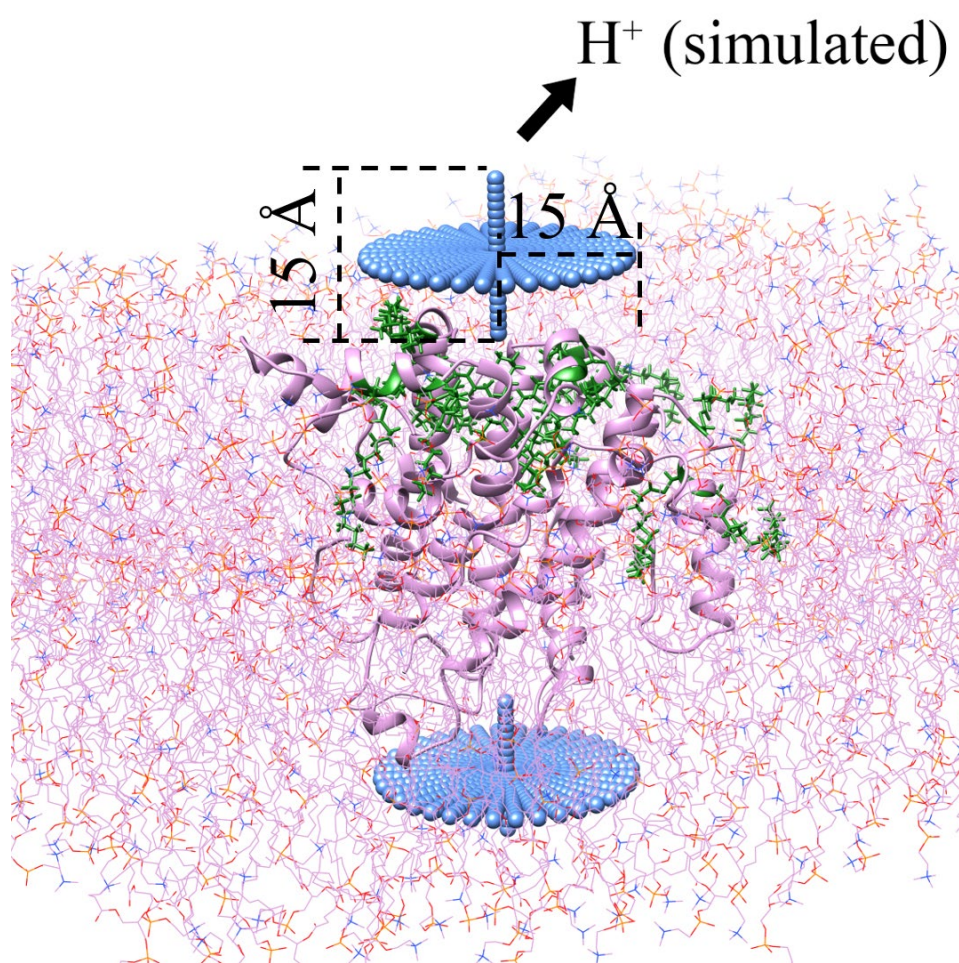


Figure 2.3 The diagram of the placement of simulated H^+ for electrostatic force testing.

2.2 Result and discussion

2.3.1 Topological study

From the RMSD study (Fig. 2.4), both systems have achieved stability after a 50 ns simulation. The structures in Fig. 2.5 were the E protein pentamers extracted from the last frame of the MD simulations of both systems. The structures were matched by Chimera(Pettersen et al., 2004) and the two sides of the ion channel were compared by the top and bottom views. The E protein with Palm stabilizes the pentameric structure in the larger pore size on the two sides while the loss of palmitoylation reduces the pore radius in the two sides and reshapes the structure into an irregular conformation. To further investigate the effects of palmitoylation on the whole ion channel during simulations, the HOLE(Smart et al., 1996) program was applied to test the minimal pore radius of the ion channel. As shown in Fig. 2.6, the E protein pentameric structure with Palm exhibits a minimal pore radius of 0.45 Å. In contrast, the non-palmitoylated E protein pentamer only has a 0.33 Å minimal pore radius. The loss of palmitoylation directly reduces the pore radius. In theory, 0.33 Å is impossible for ions passing, even for H⁺. Additionally, without palmitoylation, the collapse (The minimal pore radius acutely decreases 15% in 1 ns) of the ion channel happened 2 times during the simulations (Fig. 2.7). To investigate the reason for the collapses, the RMSF and salt bridges were calculated for the E protein pentamers. The average salt bridge numbers for the E protein without Palm is around 0~1 while that with Palm is around 1~2 (Fig. 2.6). It means the structure of E protein without Palm is unstable compared with the E protein with Palm. However, the RMSF study shows the α -Carbons in the transmembrane domain residues of the palmitoylated E protein pentameric structure are more flexible (Fig. 2.8). The RMSF studies of the C α proved the flexibility of the structure, which may be important in ion transfers. The “opposite” results from salt bridges and the RMSF study exactly proved that the characterization of dynamic equilibrium of palmitoylated E protein pentameric structure. The loss of the palmitoylation breaks the equilibrium, leading to the narrow pore radius and collapses of the ion channel. In previous studies, the palmitoylation of the E protein has also been shown with high importance in murine coronavirus assembling(Boscarino et al., 2008). The related palmitoylated residues have also been studied in the previous works(Lopez et al., 2008). However, they did not apply the mechanism studies. Unlike previous works, this work focused on the mechanism of how palmitoylation affects the E protein in computational studies. The function of palmitoylation about stabilizing proteins

has also been found in Rous Sarcoma Virus (RSV) transmembrane glycoprotein(Ochsenbauer-Jambor, Miller, Roberts, Rhee, & Hunter, 2001), CCR5 Receptor(Percherancier et al., 2001), and TEAD proteins(Noland et al., 2016). The mechanism found in this work may be also related to the other palmitoylated transmembrane proteins.

The RMSD of E-protein with and without Palm during 50ns simulation

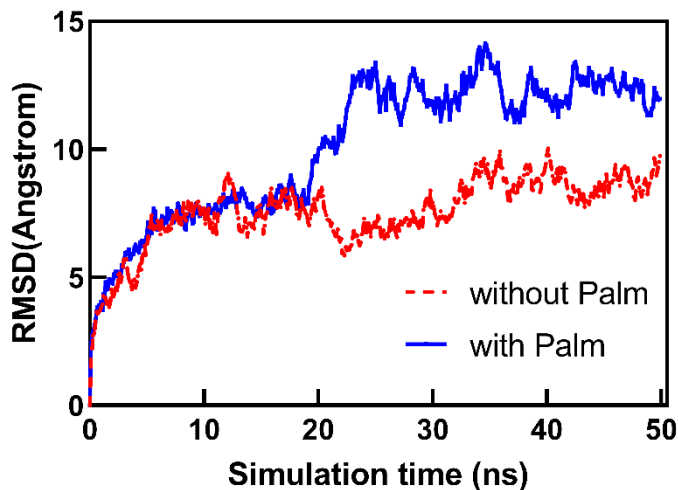


Figure 2.4 The RMSD of E-protein with and without Palm during 50 ns simulation

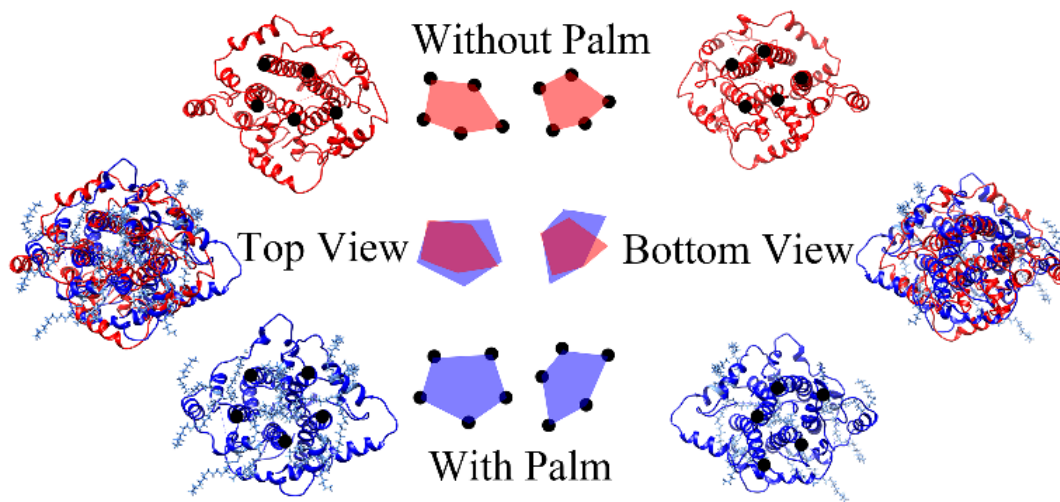


Figure 2.5The general comparison of the E protein pentamer on the top and bottom view.

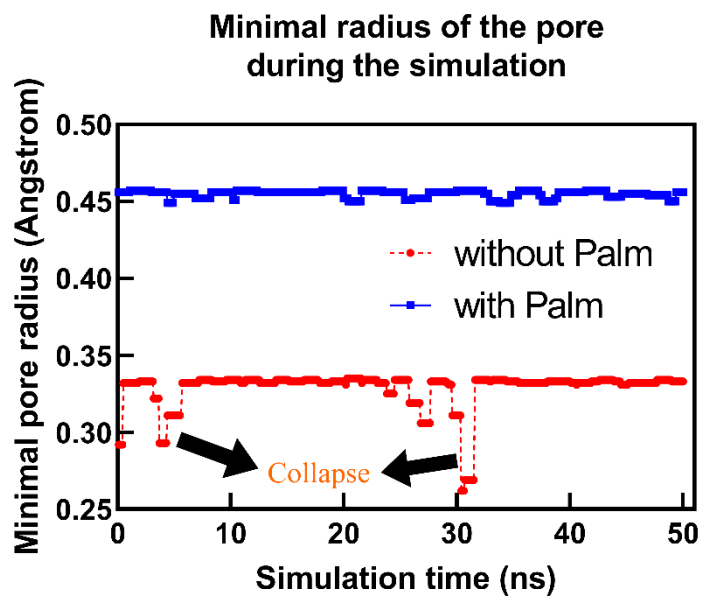


Figure 2.6 The minimal pore radius of the E-protein during the simulation.

The number of salt bridges of E-protein during 50ns simulation

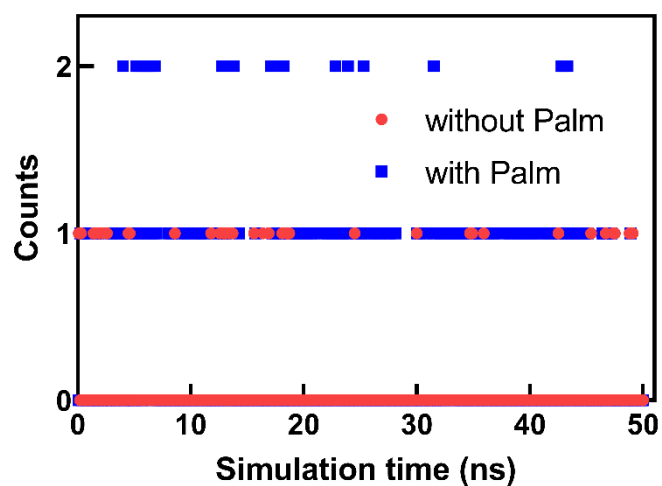


Figure 2.7 The number of salt bridges of E-protein pentamer with and without Palm during 50 ns simulation.

RMSF of the C α in residues in Transmembrane domain

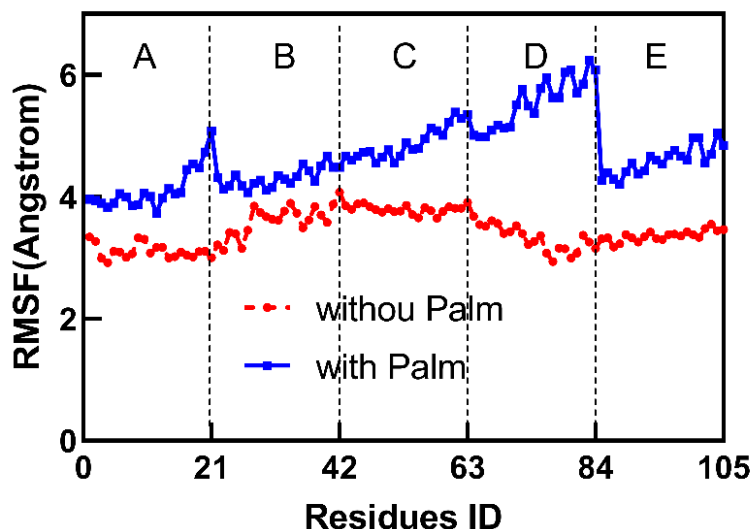


Figure 2.8 The RMSF of the residues in the transmembrane domain of E-protein. Zone A, B, C, D, D, and E represent transmembrane domain residues in 5 monomers of E-protein.

2.3.2 Electrostatic study

As shown in Fig. 2.9, the top side (cytosolic side) of the electrostatic surface of the E protein is positively charged while the bottom side (luminal side) of that is negatively charged. The distribution of surface charge is suitable for attracting cation on the luminal side and releasing cation on the cytosolic side. This finding is consistent with previous studies, which is that the E protein pentamer is a viroporin, preferring cation transfer (Wilson et al., 2004). If the cation is H^+ , the release of H^+ from lumen to cytoplasm by E protein pentamer would cause higher pH in the lumen. This finding is also consistent with that the loss of E protein causes the high pH in the Golgi complex (Westerbeck & Machamer, 2019). From the top view, the center of the E protein is full of palmitoylation whose surface is positively charged (Fig. 2.9). The palmitoylation on the cytosolic side increases the positively charged area. The enlarged area strengthens the repulsive force on the cation on the top side. By contrast, even the luminal side (Fig. 2.9 bottom view) of E protein is not palmitoylated, the surface of that also showed significant differences. Without palmitoylation, the center of E protein is blocked by the negatively charged surface. It is related to the shrunk pore radius in topology studies. When the E protein is palmitoylated, the entry of the

channel is stabilized and a neutral surface appears in the center, surrounded by the negatively charged surface. In theory, this structure not only provided opposite charges for attracting cations but also weakened the repulsion when the cations arrive at the exit. To better understand the advantages and disadvantages of this structure, more sophisticated biological and computational experiments are needed.

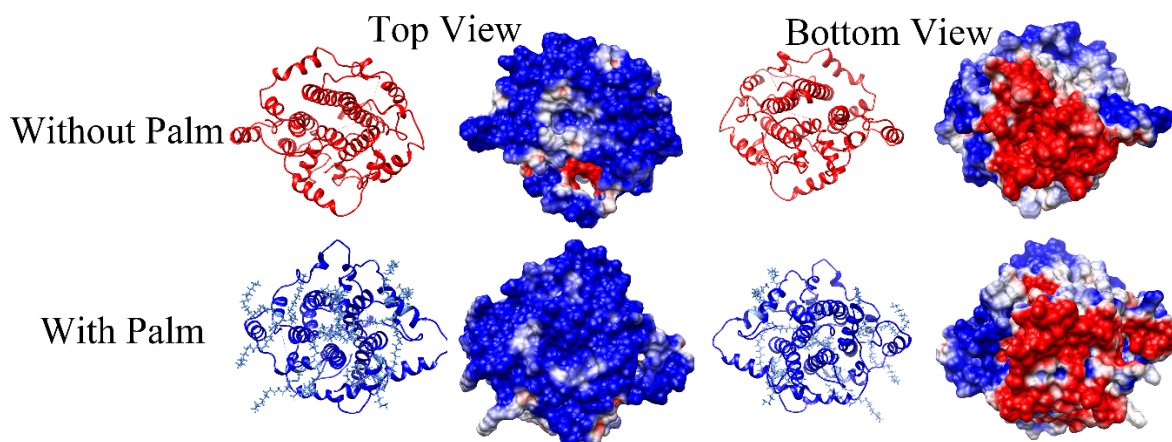


Figure 2.9 The electrostatic surface of E-protein with and without palmitoylation in the top and bottom view

The function of the E protein pentamer is exerted when the protein is on the membrane. The electrostatic surface and electric field lines of the membrane with the E protein are visualized in Fig. 2.10. From the top view, the positively charged area is extended by palmitoylation. It is consistent with the results of the E protein pentamer in Fig 2.9. However, for the electric field line studies, there are no significant differences. For the bottom view of the membranes, the E proteins with and without Palm show no significant differences in the electrostatic potential at the surfaces. However, the electric field lines in the E protein with Palm show intensively interactions with the surrounding membrane. Compared with findings in Fig. 2.10, the palmitoylation on the top can enhance the interactions on the bottom. This effect may be related to charge redistribution caused by the protein palmitoylation.

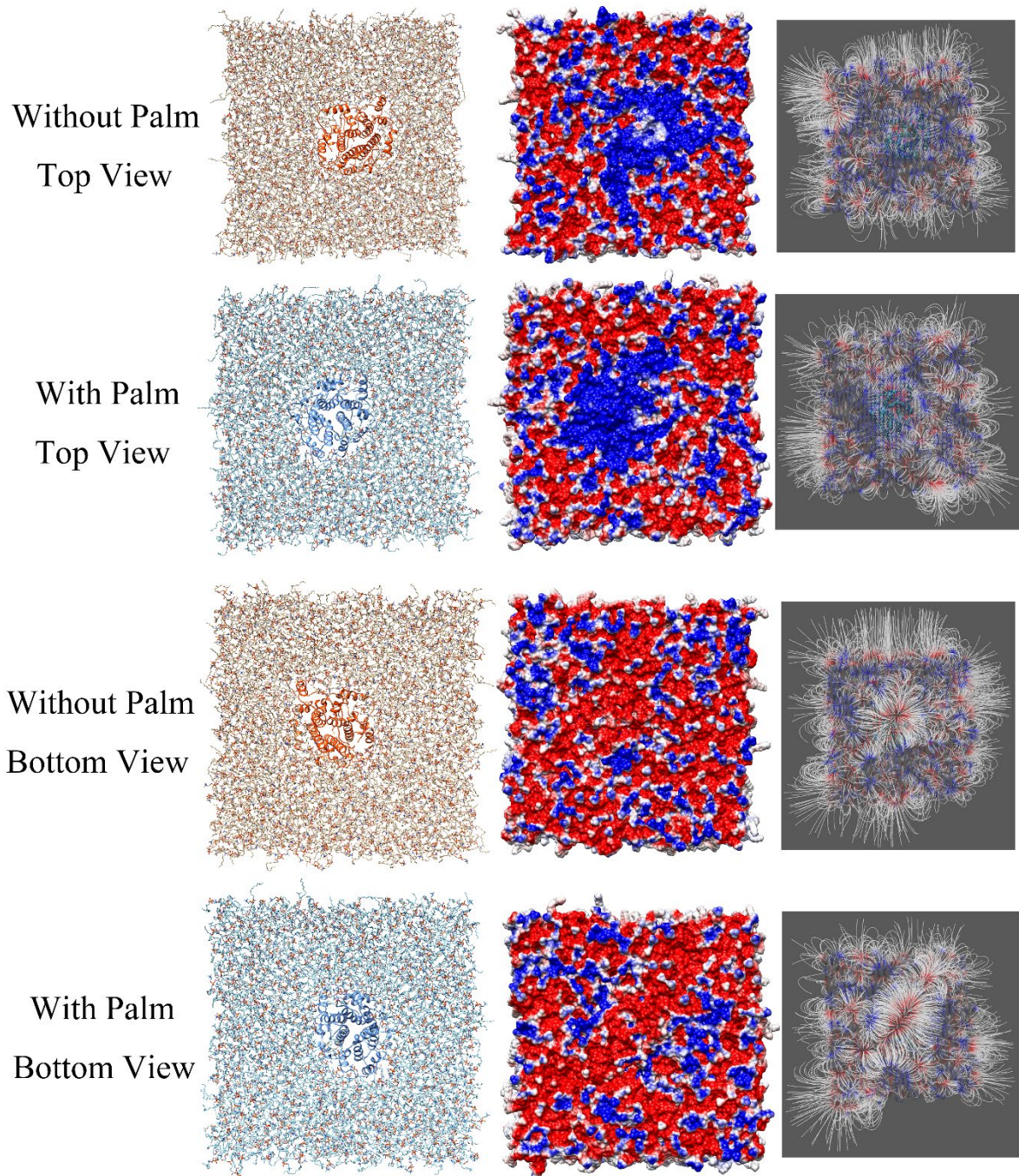


Figure 2.10 The electrostatic surface and electric field line of membrane/E-protein with and without palmitoylation.

2.3.3 The electrostatic force on H^+

To better analyze the electrostatic force on the cations when approaching the E protein, a column of H^+ ions (along the Z-axis) was simulated to place on both sides of the membrane (Fig. 2.3). The

farthest position of H^+ is 15 Å away from the membrane. The electrostatic forces were split into the X, Y, Z components. The Z component is the main force to attract or repel ions (Fig. 2.11). On the cytosolic side, the z component of the electrostatic force shows a stable increase on palmitoylated E protein pentamer when the H^+ approaching the membrane and the direction is from lumen to cytoplasm. By contrast, the loss of the palmitoylation caused the reverses of the direction of the Z components when H^+ ions are approaching the membrane. Similarly, when the H^+ is approaching the luminal side, the palmitoylated E protein also shows a similar and stable increase of the Z component while the E protein without Palm shows the variational direction of the Z components. The direction reverse of the Z component will strongly affect the E protein to attract the H^+ ions in the lumen and weaken the repulsive force of E protein to release the H^+ in the cytoplasm. If the H^+ ions are accumulated on the cytosolic side, the efficiency of the ion channel would be strongly influenced by the concentration gradient. In general, the electrostatic force of palmitoylated E protein is more beneficial for transferring cation from lumen to cytoplasm.

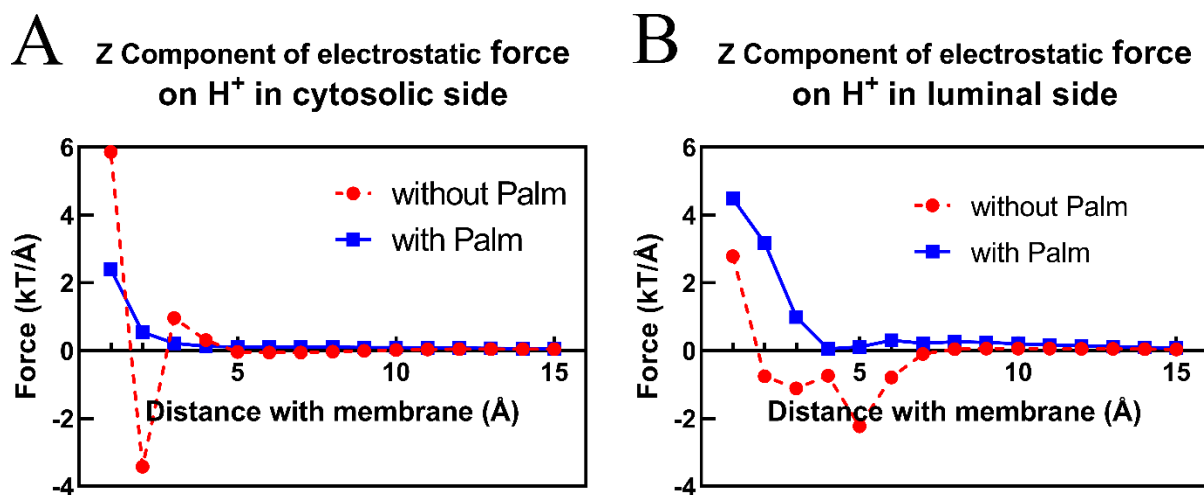


Figure 2.11 The electrostatic force on H^+ in both cytosolic and luminal sides of the E-protein.

The distance is the distance with the membrane surface.

On the 10 Å away from the membrane, several circles of H^+ ions were placed on the two sides (Fig. 2.3) to test the fluctuation of the Z component of the electrostatic force along with the distance with the Z-axis which crosses the mass center of E protein. The results are shown in Fig. 2.12. On the cytosolic side, the Z components on H^+ ions of palmitoylated E protein pentamer are around 0.1 kT/Å, and the short error bars show that the force on H^+ in each circle did not intensively vary.

By comparison, the Z component of the non-palmitoylated E protein pentamer at 2 Å is much lower than that of the palmitoylated E protein pentamer. However, it increased and drop-down rapidly after 9 Å. Besides, for a certain circle of H⁺, especially the circle of 10 Å radius, the error bar is huge, meaning that the force on that circle strongly varies along with the position of H⁺. On the luminal side, the Z component of the membrane with a palmitoylated E protein pentamer is higher than that without palmitoylation. Notably, the N-terminal (luminal sides) in both systems are not palmitoylated, the C-terminal palmitoylated E protein caused more intensive fluctuations in the luminal side (Fig. 2.12 B) but lower than that in the cytosolic side of the non-palmitoylated E protein pentamer. The electrostatic forces study on the simulated H⁺ ions proved that the electrostatic force on certain panels is more consistent on the C-terminal of E protein pentamer. The C-terminal palmitoylation also affects the force on the luminal side, enhancing the electrostatic force but causing force inconsistent on planes.

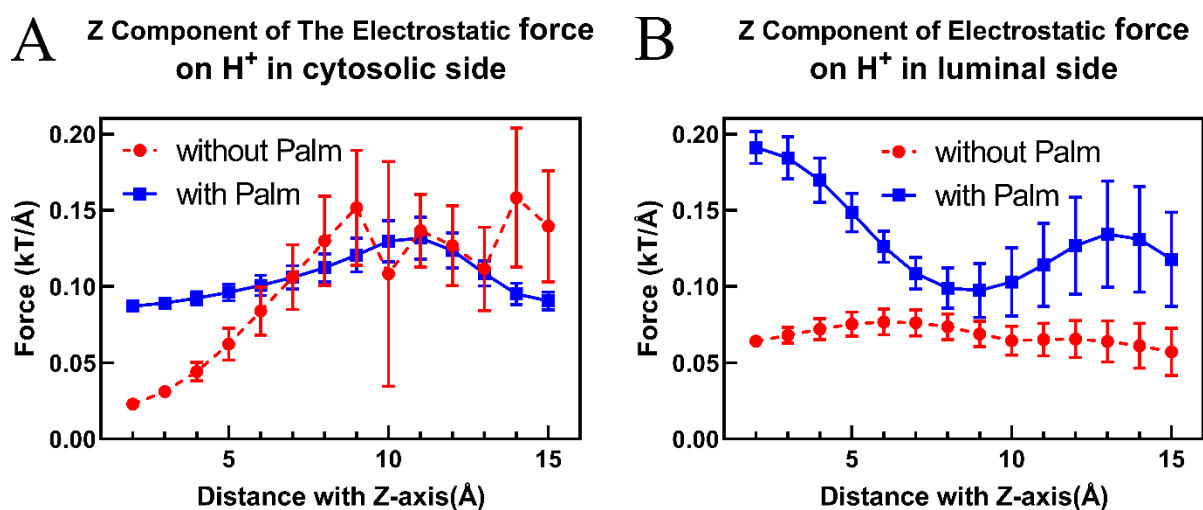


Figure 2.12 The electrostatic force on H⁺ in both cytosolic and luminal sides of the E-protein.

The distance is the distance with the Z-axis, which passes the mass center.

2.4 Conclusion

The E protein is an important SARS-CoV-2 structural protein. The pentameric structure of E protein serves as viroporin. Palmitoylation is a common and important post-translational modification on the E protein. MD simulations are performed on the E protein pentamers with and without Palm in this study. Based on the MD simulations, the topological and electrostatic studies on the E protein reveal how palmitoylation affects the E protein. The palmitoylation on the E

protein stabilizes the pentameric structure, possessing a higher number of salt bridges. Without the palmitoylation, the pentameric structure loses dynamic equilibrium and the pore radius was reduced from 0.45 Å to 0.33 Å. In the electrostatic studies, the luminal side of the E protein is highly negatively charged while the cytosolic side is highly positively charged. The surface potential supports the attraction of cations on the luminal side and the repelling of cations on the cytosolic side. This finding is consistent with the finding of previous works that the viroporin prefer cation transfer. The palmitoylation on the E protein extends the positively charged area on the cytosolic side and increases the electric interaction with the membrane on the luminal side. In the end, we simulated H⁺ ions approaching the E protein on the luminal side and leaving E protein on the cytosolic side. The results indicate that the electrostatic force on palmitoylated E protein pentamer benefits cation transfer from lumen to the cytoplasm. This study reveals the importance of palmitoylation on the E protein, which is helpful for the treatment development of COVID-19 and other corona virus-related diseases. The electrostatic studies indicate the cations transferring direction is from lumen to cytosol and palmitoylation on E protein benefits this process. To go further and more accurately prove the ion transferring direction, more sophisticated biological experiments are required.

Chapter 3: Future work

3.1 Myosin project

I will study cardiac myosin in my future research work. There are two important objects I will start with. First is the binding process of the myosin filament and actin filament. The second is the effects of mutations on myosin, which causes serious diseases in human beings. These two objects mutually reinforce each other. I will start the two paths together for a better understanding of myosin.

3.1.1 background

Myosin is the superfamily of motor protein and famous for its role in muscle contractions and the motility process. The binding and detaching process is highly associated with ATP hydrolysis and Ca^{2+} concentration in cells.

The kinetic cycle is roughly classified into 4 stages: pre-rigor, rigor stage, post-rigor stage, and relaxed stage(Al-Khayat & Practice, 2013). In detail, after binding with ATP, the myosin will be detached from the actin filament. Then it is in a relaxed stage. In the next cycle, the contact with actin triggered the release of P_i and force generation(Llinas et al., 2015). This process started the power-stroke process. When the ADP is released from the myosin cleft, the process is in the rigor stage, which is the most strong binding between myosin filament and actin filament(Llinas et al., 2015).

Additionally, the binding between myosin filament and actin filament mainly happens between the globular actin (G-actin) and myosin motor domain. The tropomyosin in the actin filament is an important “switch” to regulate the binding between the myosin motor domain and G-actin. So far, there are 3 phases of tropomyosin have been observed and well-accepted, which are open, close, and blocked phase(Risi et al., 2017). The Ca^{2+} regulation on myosin-binding with actin is achieved by the troponin to move the tropomyosin into different phases(Yamada et al., 2020).

So far, the experimental biophysical studies on the myosin have shown a lot of high-resolution structures for computational studies, including the myosin motor domain in rigor (PDB: 5JLH)(von der Ecken et al., 2016) and post-rigor states (PDB: 6FSA)(Robert-Paganin et al., 2018). Besides, the corresponding actin filament with 3 stages of tropomyosin have also been published

(PDB: 5NOJ (open), 5NOL(close), and 5NOG(block))(Risi et al., 2017). These crucial materials will be beneficial for my further computational work on the myosin project.

3.1.2 Current progress and future work

The the progress from rigor stage to the post-rigor stage of myosin is the first sub-project I am working on. The actomyosin models in rigor and post-rigor stages have been achieved by the swiss-model(Guex & Peitsch, 1997). The 20 ns simulations have been done for both models. Besides, the surface potential calculation (Fig. 3.1 and 3.2) has been finished. Soon, the free energy and binding force in two stages of actomyosin models should be finished.

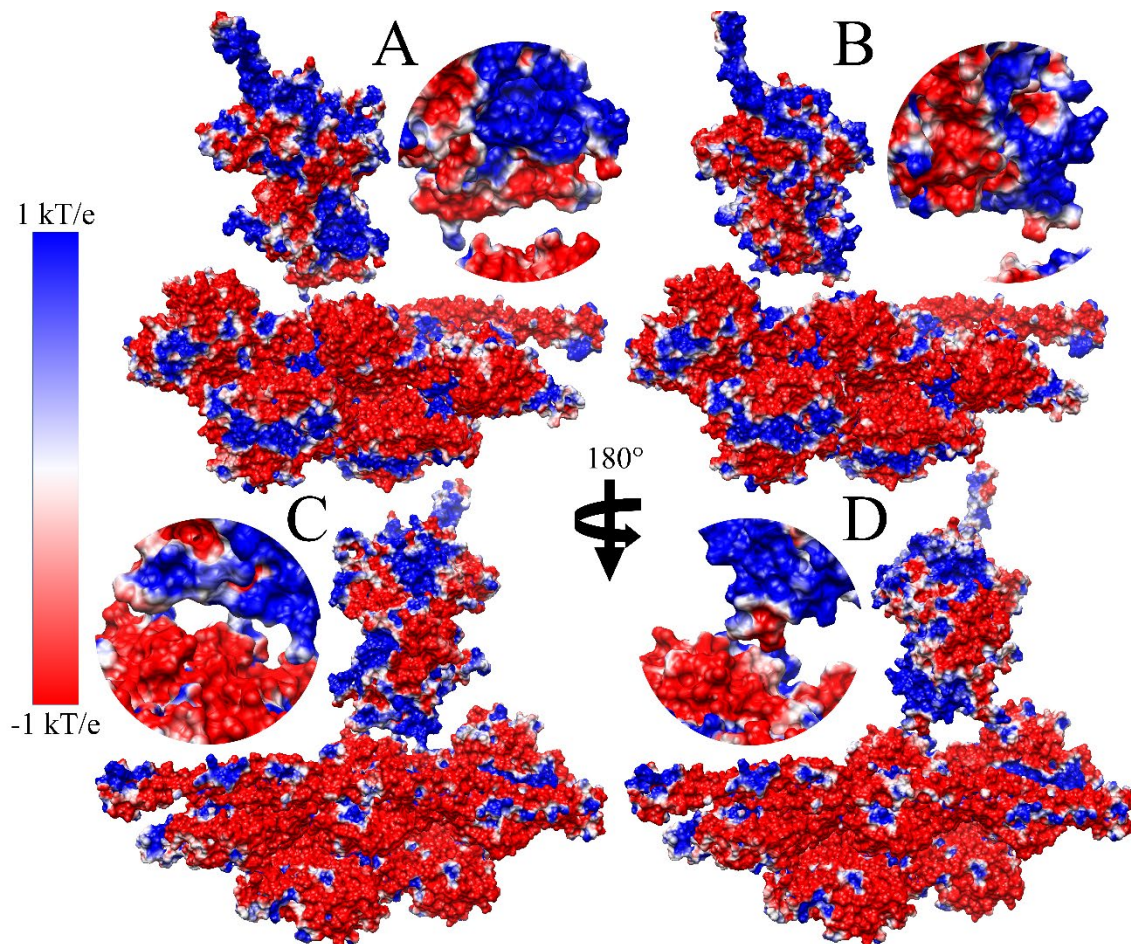


Figure 3.1 Electrostatic surfaces for the actomyosin models in rigor (A and C) and post-rigor state (B and D).

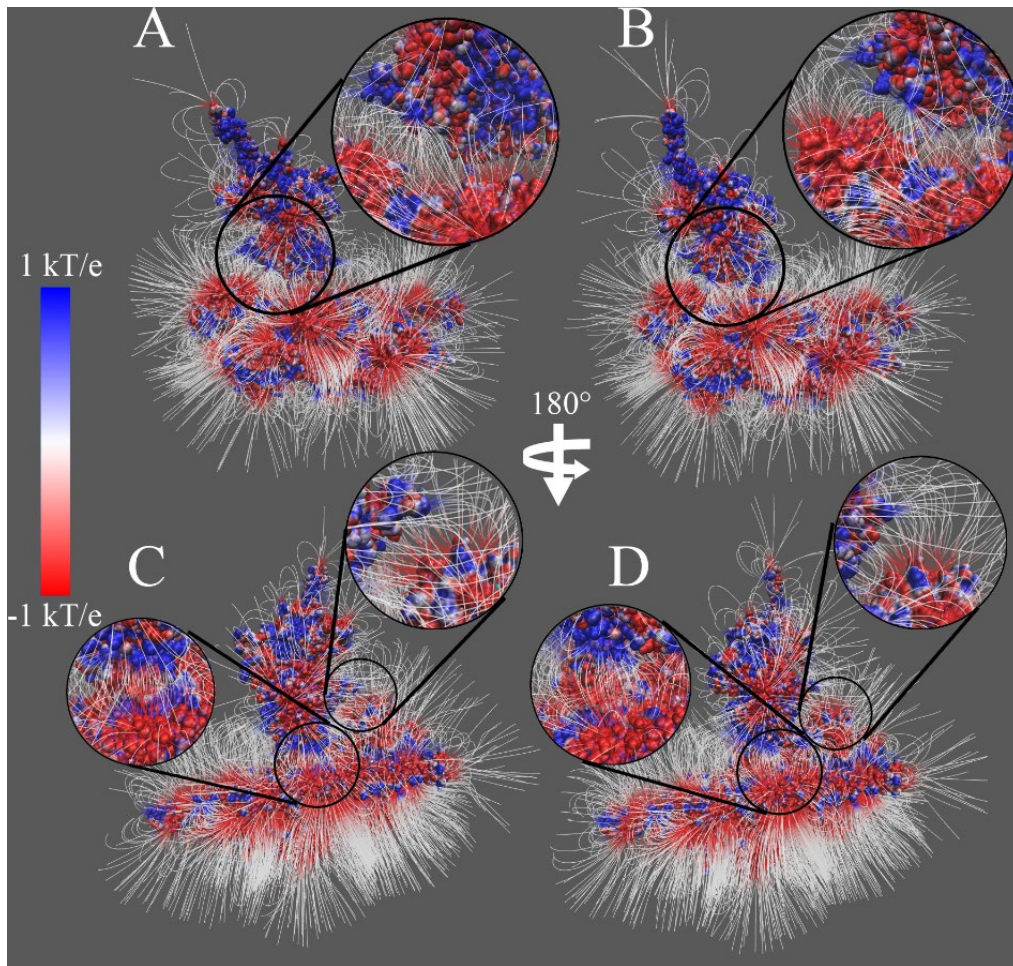


Figure 3.2 Electrostatic field line for the interface of actomyosin model in rigor (A and C) and post-rigor state (B and D).

In the next two years, I will build the actomyosin model for other stages in the myosin kinetic circles and explain the binding and detaching mechanism by sophisticated computation in energy and force associated with the conformational changes.

3.2 HIT2 project

The current HIT program can handle most cases in MD simulation except for the highly charged protein with huge conformational change. This drawback will limit the application of the program in computational biophysics. I will further develop this software into a more user-friendly and accurate tool.

3.2.1 Background

In computational biology, the electrostatic calculation of biomolecules is fundamental and challenging. The electrostatic interactions play significant roles in protein folding(Ganguly et al., 2012), protein stability(Stigter et al., 1991; Strickler et al., 2006), protein-protein interactions(L. Li, Jia, et al., 2017; L. Li et al., 2015; Zondlo, 2013), protein-DNA/RNA interactions(Akhtar et al., 2000; Richardson et al., 1988), and many other fields. However, the trapped ions directly affect the electrostatic surfaces, which have significant impacts on the interactions between biomolecules. Currently, there are two types of models to handle the ions and water molecules surrounding biomolecules: implicit solvent models and explicit solvent models. In implicit solvent models, the electrostatic features are calculated for biomolecules by treating ions implicitly(Jia et al., 2017; Klapper et al., 1986; Nicholls & Honig, 1991). On the other side, explicit solvent models such as TIP3P, TIP4P, with explicit ions, are widely used in Molecular Dynamic (MD) simulations by handling ions and molecules explicitly(Florová et al., 2010). Explicit solvent models can neutralize the highly charged biomolecules by adding unbalanced amounts of positive and negative ions into a system. However, traditional implicit solvent models treat the solvation as neutral with the same amounts of positive and negative ionic charges, which causes troubles in the electrostatic calculations of highly charged biomolecules and results in unrealistic interaction analyses. Under this circumstance, how to improve the implicit models to handle the highly charged biomolecules is a challenging problem to be solved.

The current HIT program(Sun et al., 2021) can find most bound ions with occurrence frequency as reference but cannot provide an accurate number of bound ions. It means the users need their efforts to decide the threshold for the selection of bound ions. Besides, the opposite ions should also be selected for the certain oppositely charged area. In this case, the HIT needs to be run twice to find them. It is not convenient for users. Above all, the crucial limitation of this program is that it cannot handle the biomolecules with significant movement or conformational change. This limitation is the most crucial part that should be eliminated.

3.2.2 future work

The limitation of the HIT program is caused by the usage of the absolute position of ions instead of relative positions. In the new version of the HIT (HIT2) program, the relative position to

residues or mass center will be tested and applied. The change will strongly improve the accuracy for biomolecules with movements or conformational changes.

The number of bound ions will be predicted by deep learning, several deep learning methods, including but not limited, random forest, neural network, lasso (least absolute shrinkage and selection operator) regression, and PCA (principal component analysis), will be tested to find the best number of bound ions. The HIT2 will provide a more accurate suggested number of bound ions. Besides, the opposite bound ions would also be considered in the HIT2 program. The parameters for prediction include but are not limited to cube size, occurrence frequency, and charge of residues.

3.3 Time Schedule of Future Research

Table 3.1: Time Schedule

Tasks to complete project	Approximate time
Myosin project: rigor to the post-rigor process	March-August 2021
HIT2 project: Algorithm modification	August-October 2021
HIT2 project: Dataset creation	October-December 2021
HIT2 project: HIT2 application testing	January-May 2022
Myosin project: pre-power stroke process	May-August 2022
Myosin project: Mutation analysis	August-December 2022
Thesis Writing: Literature review	January-February 2023
Thesis writing: Methodology	February-March 2023
Thesis writing: Results	March-August 2023
Thesis writing: First draft	September 2023
Thesis writing: Revising	October 2023
Defense	November 2023
Submission and publication	December 2023

References

- Akhtar, A., Zink, D., & Becker, P. B. J. N. (2000). Chromodomains are protein–RNA interaction modules. *407*(6802), 405-409.
- Al-Khayat, H. A. J. G. C. S., & Practice. (2013). Three-dimensional structure of the human myosin thick filament: clinical implications. *2013*(3), 36.
- Behrmann, E., Müller, M., Penczek, P. A., Mannherz, H. G., Manstein, D. J., & Raunser, S. J. C. (2012). Structure of the rigor actin-tropomyosin-myosin complex. *150*(2), 327-338.
- Boscarino, J. A., Logan, H. L., Lacny, J. J., & Gallagher, T. M. J. J. o. v. (2008). Envelope protein palmitoylations are crucial for murine coronavirus assembly. *82*(6), 2989-2999.
- Burghardt, T. P., & Sikkink, L. A. J. B. (2013). Regulatory light chain mutants linked to heart disease modify the cardiac myosin lever arm. *52*(7), 1249-1259.
- Chen, S.-C., Lo, S.-Y., Ma, H.-C., & Li, H.-C. J. V. g. (2009). Expression and membrane integration of SARS-CoV E protein and its interaction with M protein. *38*(3), 365-371.
- Curtis, K. M., Yount, B., & Baric, R. S. J. J. o. v. (2002). Heterologous gene expression from transmissible gastroenteritis virus replicon particles. *76*(3), 1422-1434.
- Dayton, W. R., Reville, W., Goll, D. E., & Stromer, M. J. B. (1976). A calcium (2+) ion-activated protease possibly involved in myofibrillar protein turnover. Partial characterization of the purified enzyme. *15*(10), 2159-2167.
- De Graff, A. M., Hazoglou, M. J., & Dill, K. A. J. S. (2016). Highly charged proteins: the Achilles' heel of aging proteomes. *24*(2), 329-336.
- de Haan, C. A., & Rottier, P. J. J. A. i. v. r. (2005). Molecular interactions in the assembly of coronaviruses. *64*, 165-230.

- De Wit, E., Van Doremalen, N., Falzarano, D., & Munster, V. J. J. N. R. M. (2016). SARS and MERS: recent insights into emerging coronaviruses. *14*(8), 523.
- DeDiego, M. L., Álvarez, E., Almazán, F., Rejas, M. T., Lamirande, E., Roberts, A., . . . Enjuanes, L. J. J. o. v. (2007). A severe acute respiratory syndrome coronavirus that lacks the E gene is attenuated in vitro and in vivo. *81*(4), 1701-1713.
- Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., & Baker, N. A. J. N. a. r. (2007). PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *35*(suppl_2), W522-W525.
- Dolinsky, T. J., Nielsen, J. E., McCammon, J. A., & Baker, N. A. J. N. a. r. (2004). PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *32*(suppl_2), W665-W667.
- Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. J. T. J. o. c. p. (1995). A smooth particle mesh Ewald method. *103*(19), 8577-8593.
- Florová, P., Sklenovsky, P., Banáš, P., Otyepka, M. J. J. o. C. T., & Computation. (2010). Explicit water models affect the specific solvation and dynamics of unfolded peptides while the conformational behavior and flexibility of folded peptides remain intact. *6*(11), 3569-3579.
- Fung, T. S., & Liu, D. X. J. F. v. (2018). Post-translational modifications of coronavirus proteins: roles and function. *13*(6), 405-430.
- Ganguly, D., Otieno, S., Waddell, B., Iconaru, L., Kriwacki, R. W., & Chen, J. J. J. o. m. b. (2012). Electrostatically accelerated coupled binding and folding of intrinsically disordered proteins. *422*(5), 674-684.

- Geisterfer-Lowrance, A. A., Kass, S., Tanigawa, G., Vosberg, H.-P., McKenna, W., Seidman, C. E., & Seidman, J. J. C. (1990). A molecular basis for familial hypertrophic cardiomyopathy: a β cardiac myosin heavy chain gene missense mutation. *62*(5), 999-1006.
- Guex, N., & Peitsch, M. C. J. e. (1997). SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *18*(15), 2714-2723.
- Hall, K. S., Samari, G., Garbers, S., Casey, S. E., Diallo, D. D., Orcutt, M., . . . McGovern, T. J. T. I. (2020). Centring sexual and reproductive health and justice in the global COVID-19 response. *395*(10231), 1175-1177.
- Houdusse, A., & Sweeney, H. L. J. T. i. b. s. (2016). How myosin generates force on actin filaments. *41*(12), 989-997.
- Huang, J., & MacKerell Jr, A. D. J. J. o. c. c. (2013). CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *34*(25), 2135-2145.
- Humphrey, W., Dalke, A., & Schulten, K. J. J. o. m. g. (1996). VMD: visual molecular dynamics. *14*(1), 33-38.
- Hunter, B., & Allingham, J. S. J. P. S. (2020). These motors were made for walking. *29*(8), 1707-1723.
- Huszar, D., Theoclitou, M.-E., Skolnik, J., Herbst, R. J. C., & Reviews, M. (2009). Kinesin motor proteins as targets for cancer therapy. *28*(1), 197-208.
- Inchingolo, A. V., Previs, S. B., Previs, M. J., Warshaw, D. M., & Kad, N. M. J. P. o. t. N. A. o. S. (2019). Revealing the mechanism of how cardiac myosin-binding protein C N-terminal fragments sensitize thin filaments for myosin binding. *116*(14), 6828-6835.

- Jayaram, B., Sprous, D., & Beveridge, D. J. T. J. o. P. C. B. (1998). Solvation free energy of biomacromolecules: Parameters for a modified generalized Born model consistent with the AMBER force field. *102*(47), 9571-9576.
- Jia, Z., Li, L., Chakravorty, A., & Alexov, E. (2017). Treating ion distribution with Gaussian-based smooth dielectric function in DelPhi. In: Wiley Online Library.
- Jo, S., Kim, T., Iyer, V. G., & Im, W. J. J. o. c. c. (2008). CHARMM-GUI: a web-based graphical user interface for CHARMM. *29*(11), 1859-1865.
- Kapitein, L. C., Peterman, E. J., Kwok, B. H., Kim, J. H., Kapoor, T. M., & Schmidt, C. F. J. N. (2005). The bipolar mitotic kinesin Eg5 moves on both microtubules that it crosslinks. *435*(7038), 114-118.
- Kim, D. E., Chivian, D., & Baker, D. J. N. a. r. (2004). Protein structure prediction and analysis using the Robetta server. *32*(suppl_2), W526-W531.
- Klapper, I., Hagstrom, R., Fine, R., Sharp, K., Honig, B. J. P. S., Function,, & Bioinformatics. (1986). Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *1*(1), 47-59.
- Korennykh, A. V., Correll, C. C., & Piccirilli, J. A. J. R. (2007). Evidence for the importance of electrostatics in the function of two distinct families of ribosome inactivating toxins. *13*(9), 1391-1396.
- Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., . . . Zhang, L. J. N. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *581*(7807), 215-220.

- Lee, K. K., Fitch, C. A., Lecomte, J. T., & García-Moreno E, B. J. B. (2002). Electrostatic effects in highly charged proteins: salt sensitivity of p K a values of histidines in staphylococcal nuclease. *41*(17), 5656-5667.
- Li, C., Jia, Z., Chakravorty, A., Pahari, S., Peng, Y., Basu, S., . . . Li, L. J. J. o. c. c. (2019). DelPhi suite: New developments and review of functionalities. *40*(28), 2502-2508.
- Li, C., Li, L., Zhang, J., & Alexov, E. J. J. o. c. c. (2012). Highly efficient and exact method for parallelization of grid-based algorithms and its implementation in DelPhi. *33*(24), 1960-1966.
- Li, C., Petukh, M., Li, L., & Alexov, E. J. J. o. c. c. (2013). Continuous development of schemes for parallel computing of the electrostatics in biological systems: Implementation in DelPhi. *34*(22), 1949-1960.
- Li, L., Alper, J., & Alexov, E. J. S. r. (2016a). Cytoplasmic dynein binding, run length, and velocity are guided by long-range electrostatic interactions. *6*(1), 1-12.
- Li, L., Alper, J., & Alexov, E. J. S. r. (2016b). Multiscale method for modeling binding phenomena involving large objects: application to kinesin motor domains motion along microtubules. *6*(1), 1-12.
- Li, L., Chakravorty, A., & Alexov, E. J. J. o. c. c. (2017). DelPhiForce, a tool for electrostatic force calculations: Applications to macromolecular binding. *38*(9), 584-593.
- Li, L., Jia, Z., Peng, Y., Godar, S., Getov, I., Teng, S., . . . Alexov, E. J. S. r. (2017). Forces and Disease: Electrostatic force differences caused by mutations in kinesin motor domains can distinguish between disease-causing and non-disease-causing mutations. *7*(1), 1-12.
- Li, L., Li, C., Sarkar, S., Zhang, J., Witham, S., Zhang, Z., . . . Alexov, E. J. B. b. (2012). DelPhi: a comprehensive suite for DelPhi software and associated resources. *5*(1), 1-11.

- Li, L., Li, C., Zhang, Z., Alexov, E. J. J. o. c. t., & computation. (2013). On the dielectric “constant” of proteins: smooth dielectric function for macromolecular modeling and its implementation in DelPhi. *9*(4), 2126-2136.
- Li, L., Wang, L., & Alexov, E. J. F. i. m. b. (2015). On the energy components governing molecular recognition in the framework of continuum approaches. *2*, 5.
- Li, M., & Zheng, W. J. B. (2013). All-Atom Molecular Dynamics Simulations of Actin–Myosin Interactions: A Comparative Study of Cardiac α Myosin, β Myosin, and Fast Skeletal Muscle Myosin. *52*(47), 8393-8405.
- Li, Y., Surya, W., Claudine, S., & Torres, J. J. J. o. B. C. (2014). Structure of a conserved Golgi complex-targeting signal in coronavirus envelope proteins. *289*(18), 12535-12549.
- Liao, Y., Yuan, Q., Torres, J., Tam, J., & Liu, D. J. V. (2006). Biochemical and functional characterization of the membrane association and membrane permeabilizing activity of the severe acute respiratory syndrome coronavirus envelope protein. *349*(2), 264-275.
- Llinas, P., Isabet, T., Song, L., Ropars, V., Zong, B., Benisty, H., . . . Safer, D. J. D. c. (2015). How actin initiates the motor activity of Myosin. *33*(4), 401-412.
- Lopez, L. A., Riffle, A. J., Pike, S. L., Gardner, D., & Hogue, B. G. J. J. o. v. (2008). Importance of conserved cysteine residues in the coronavirus envelope protein. *82*(6), 3000-3010.
- Luther, P. K., Winkler, H., Taylor, K., Zoghbi, M. E., Craig, R., Padrón, R., . . . Liu, J. J. P. o. t. N. A. o. S. (2011). Direct visualization of myosin-binding protein C bridging myosin and actin filaments in intact muscle. *108*(28), 11423-11428.
- Mackerell Jr, A. D., Feig, M., & Brooks III, C. L. J. J. o. c. c. (2004). Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics

- in reproducing protein conformational distributions in molecular dynamics simulations. *25*(11), 1400-1415.
- Madjid, M., Safavi-Naeini, P., Solomon, S. D., & Vardeny, O. J. J. c. (2020). Potential effects of coronaviruses on the cardiovascular system: a review. *5*(7), 831-840.
- Masters, P. S. J. A. i. v. r. (2006). The molecular biology of coronaviruses. *66*, 193-292.
- Moein, S. T., Hashemian, S. M., Mansourafshar, B., Khorram-Tousi, A., Tabarsi, P., & Doty, R. L. (2020). *Smell dysfunction: a biomarker for COVID-19*. Paper presented at the International forum of allergy & rhinology.
- Mould, J. A., Paterson, R. G., Takeda, M., Ohigashi, Y., Venkataraman, P., Lamb, R. A., & Pinto, L. H. J. D. c. (2003). Influenza B virus BM2 protein has ion channel activity that conducts protons across membranes. *5*(1), 175-184.
- Neuman, B. W., Kiss, G., Kunding, A. H., Bhella, D., Baksh, M. F., Connelly, S., . . . Sawicki, S. G. J. J. o. s. b. (2011). A structural analysis of M protein in coronavirus assembly and morphology. *174*(1), 11-22.
- Nicholls, A., & Honig, B. J. J. o. c. c. (1991). A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *12*(4), 435-445.
- Nieto-Torres, J. L., DeDiego, M. L., Álvarez, E., Jiménez-Guardeño, J. M., Regla-Nava, J. A., Llorente, M., . . . Enjuanes, L. J. V. (2011). Subcellular location and topology of severe acute respiratory syndrome coronavirus envelope protein. *415*(2), 69-82.
- Nieto-Torres, J. L., Verdiá-Báguena, C., Jimenez-Guardeño, J. M., Regla-Nava, J. A., Castaño-Rodríguez, C., Fernandez-Delgado, R., . . . Enjuanes, L. J. V. (2015). Severe acute respiratory syndrome coronavirus E protein transports calcium ions and activates the NLRP3 inflammasome. *485*, 330-339.

- Nieva, J. L., Madan, V., & Carrasco, L. J. N. R. M. (2012). Viroporins: structure and biological functions. *10*(8), 563-574.
- Noland, C. L., Gierke, S., Schnier, P. D., Murray, J., Sandoval, W. N., Sagolla, M., . . . Cunningham, C. N. J. S. (2016). Palmitoylation of TEAD transcription factors is required for their stability and function in Hippo pathway signaling. *24*(1), 179-186.
- Ochsenbauer-Jambor, C., Miller, D. C., Roberts, C. R., Rhee, S. S., & Hunter, E. J. J. o. v. (2001). Palmitoylation of the Rous sarcoma virus transmembrane glycoprotein is required for protein stability and virus infectivity. *75*(23), 11544-11554.
- Organization, W. H. (2019). *WHO MERS global summary and assessment of risk, July 2019*. Retrieved from
- Ortego, J., Ceriani, J. E., Patiño, C., Plana, J., & Enjuanes, L. J. V. (2007). Absence of E protein arrests transmissible gastroenteritis coronavirus maturation in the secretory pathway. *368*(2), 296-308.
- Park, J.-E., Jung, S., & Kim, A. J. B. p. h. (2018). MERS transmission and risk factors: a systematic review. *18*(1), 574.
- Percherancier, Y., Planchenault, T., Valenzuela-Fernandez, A., Virelizier, J.-L., Arenzana-Seisdedos, F., & Bachelier, F. J. J. o. B. C. (2001). Palmitoylation-dependent control of degradation, life span, and membrane expression of the CCR5 receptor. *276*(34), 31936-31944.
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. J. J. o. c. c. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *25*(13), 1605-1612.

- Pham, T., Perry, J. L., Dosey, T. L., Delcour, A. H., & Hyser, J. M. J. S. r. (2017). The rotavirus NSP4 viroporin domain is a calcium-conducting ion channel. *7*(1), 1-11.
- Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., . . . Schulten, K. J. J. o. c. c. (2005). Scalable molecular dynamics with NAMD. *26*(16), 1781-1802.
- Ramphul, K., & Mejias, S. G. J. C. (2020). Coronavirus disease: a review of a new threat to public health. *12*(3).
- Richardson, J. S., Richardson, D. C. J. P. S., Function,, & Bioinformatics. (1988). Helix lap-joints as ion-binding sites: DNA-binding motifs and Ca-binding “EF hands” are related by charge and sequence reversal. *4*(4), 229-239.
- Risi, C., Eisner, J., Belknap, B., Heeley, D. H., White, H. D., Schröder, G. F., & Galkin, V. E. J. P. o. t. N. A. o. S. (2017). Ca²⁺-induced movement of tropomyosin on native cardiac thin filaments revealed by cryoelectron microscopy. *114*(26), 6782-6787.
- Robert-Paganin, J., Auguin, D., & Houdusse, A. J. N. c. (2018). Hypertrophic cardiomyopathy disease results from disparate impairments of cardiac myosin function and auto-inhibition. *9*(1), 1-13.
- Schoeman, D., & Fielding, B. C. J. V. j. (2019). Coronavirus envelope protein: current knowledge. *16*(1), 1-22.
- Shashikala, H., Chakravorty, A., Panday, S. K., & Alexov, E. J. I. J. o. M. S. (2021). BION-2: Predicting Positions of Non-Specifically Bound Ions on Protein Surface by a Gaussian-Based Treatment of Electrostatics. *22*(1), 272.
- Smart, O. S., Neduvélil, J. G., Wang, X., Wallace, B., & Sansom, M. S. J. J. o. m. g. (1996). HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *14*(6), 354-360.

- Stadler, K., Massignani, V., Eickmann, M., Becker, S., Abrignani, S., Klenk, H.-D., & Rappuoli, R. J. N. R. M. (2003). SARS—beginning to understand a new virus. *1*(3), 209-218.
- Stigter, D., Alonso, D., & Dill, K. A. J. P. o. t. N. A. o. S. (1991). Protein stability: electrostatics and compact denatured states. *88*(10), 4176-4180.
- Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., . . . Makhatadze, G. I. J. B. (2006). Protein stability and surface electrostatics: a charged relationship. *45*(9), 2761-2766.
- Sun, S., Karki, C., Xie, Y., Xian, Y., Guo, W., Gao, B. Z., . . . journal, s. b. (2021). Hybrid method for representing ions in implicit solvation calculations. *19*, 801-811.
- Surya, W., Li, Y., & Torres, J. J. B. e. B. A.-B. (2018). Structural model of the SARS coronavirus E channel in LMPG micelles. *1860*(6), 1309-1317.
- Torres, J., Wang, J., Parthasarathy, K., & Liu, D. X. J. B. j. (2005). The transmembrane oligomers of coronavirus protein E. *88*(2), 1283-1290.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. J. J. o. c. c. (2005). GROMACS: fast, flexible, and free. *26*(16), 1701-1718.
- Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., . . . Vorobyov, I. J. J. o. c. c. (2010). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *31*(4), 671-690.
- Venkatagopalan, P., Daskalova, S. M., Lopez, L. A., Dolezal, K. A., & Hogue, B. G. J. V. (2015). Coronavirus envelope (E) protein remains at the site of assembly. *478*, 75-85.
- von der Ecken, J., Heissler, S. M., Pathan-Chhatbar, S., Manstein, D. J., & Raunser, S. J. N. (2016). Cryo-EM structure of a human cytoplasmic actomyosin complex at near-atomic resolution. *534*(7609), 724-728.

- von Loeffelholz, O., Peña, A., Drummond, D. R., Cross, R., & Moores, C. A. J. J. o. m. b. (2019). Cryo-EM structure (4.5-Å) of Yeast Kinesin-5–Microtubule Complex reveals a distinct binding footprint and mechanism of drug resistance. *431*(4), 864-872.
- Wang, C., Takeuchi, K., Pinto, L., & Lamb, R. A. J. J. o. v. (1993). Ion channel activity of influenza A virus M2 protein: characterization of the amantadine block. *67*(9), 5585-5594.
- Wei, Y.-L., Yang, W.-X. J. C. d., & disease. (2019). Kinesin-14 motor protein KIFC1 participates in DNA synthesis and chromatin maintenance. *10*(6), 1-14.
- Westerbeck, J. W., & Machamer, C. E. J. J. o. v. (2019). The infectious bronchitis coronavirus envelope protein alters Golgi pH to protect the spike protein and promote the release of infectious virus. *93*(11).
- Wilson, L., Mckinlay, C., Gage, P., & Ewart, G. J. V. (2004). SARS coronavirus E protein forms cation-selective ion channels. *330*(1), 322-331.
- Woehlke, G., Ruby, A. K., Hart, C. L., Ly, B., Hom-Booher, N., & Vale, R. D. J. C. (1997). Microtubule interaction site of the kinesin motor. *90*(2), 207-216.
- Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., . . . Pei, Y.-Y. J. N. (2020). A new coronavirus associated with human respiratory disease in China. *579*(7798), 265-269.
- Xie, Y., Karki, C. B., Du, D., Li, H., Wang, J., Sobitan, A., . . . Li, L. J. F. i. m. b. (2020). Spike proteins of SARS-CoV and SARS-CoV-2 utilize different mechanisms to bind with human ACE2. *7*.
- Yamada, Y., Namba, K., & Fujii, T. J. N. c. (2020). Cardiac muscle thin filament structures reveal calcium regulatory mechanism. *11*(1), 1-9.

- Yang, P., Wang, X. J. C., & immunology, m. (2020). COVID-19: a new challenge for human beings. *17(5)*, 555-557.
- Yao, H., Chen, J.-H., & Xu, Y.-F. (2020). Patients with mental health disorders in the COVID-19 epidemic.
- Yuan, Q., Liao, Y., Torres, J., Tam, J., & Liu, D. J. F. I. (2006). Biochemical evidence for the presence of mixed membrane topologies of the severe acute respiratory syndrome coronavirus envelope protein expressed in mammalian cells. *580(13)*, 3192-3200.
- Zhou, S., Cheng, L.-T., Dzubiella, J., Li, B., McCammon, J. A. J. J. o. c. t., & computation. (2014). Variational implicit solvation with Poisson–Boltzmann theory. *10(4)*, 1454-1467.
- Zhou, S., Cheng, L. T., Sun, H., Che, J., Dzubiella, J., Li, B., & McCammon, J. A. J. J. o. c. c. (2015). LS-VISM: A software package for analysis of biomolecular solvation. *36(14)*, 1047-1059.
- Zhou, S., Weiß, R. G., Cheng, L.-T., Dzubiella, J., McCammon, J. A., & Li, B. J. P. o. t. N. A. o. S. (2019). Variational implicit-solvent predictions of the dry–wet transition pathways for ligand–receptor binding and unbinding kinetics. *116(30)*, 14989-14994.
- Zondlo, N. J. J. A. o. c. r. (2013). Aromatic–proline interactions: electronically tunable CH/ π interactions. *46(4)*, 1039-1049.

Curriculum Vitae

Shengjie Sun was born on August 26, 1993, on a stormy night with thousands of lightning. He is the only child in the house of Sun, which is one of the most ancient families in China.

In 2012, He is enrolled in The Central South University for bioengineering study. In the third year, he got a chance to do research work in the lab with Dr. Min Gan who was a Ph.D. student at that time. Nobody knows this will be the major turning point in his life. Shengjie Published two papers in the following two years and was found of research. The published two papers are: “Adsorption of Cr (VI) and Cu (II) by AlPO_4 modified biosynthetic Schwertmannite” and “The enhanced effect of *Acidithiobacillus ferrooxidans* on pyrite-based Cr (VI) reduction”.

After graduation in 2016, he received a bachelor's degree in bioengineering and decided to see the world. He directly applied for an environmental Ph.D. program at the University of Queensland in 2017 but stopped the study in 2018 due to economic problems. During the one year's study, He has two publications: “Biosynthesis of β -cyclodextrin modified Schwertmannite and the application in heavy metals adsorption” and “Physiological and transcriptomic analyses reveal CuO nanoparticle inhibition of anabolic and catabolic activities of sulfate-reducing bacterium”.

The leave of Australia did not make him abandon the dream to be a researcher. He successfully applied for a computational science program at The University of Texas at El Paso. Under Dr. Lin Li's supervision, He not only solve the problem of transdisciplinary but also produced some novel ideas by combination with previous biological and chemical knowledge. During the tough two years. He has a publication “Hybrid method for representing ions in implicit solvation calculations” and another paper is under review “The computational study on the function of palmitoylation on the envelope protein in SARS-CoV-2.”

Now he is continuing his Ph.D. program working on the mechanism of the myosin protein kinetic circles and the optimization of the HIT software development.