UNIVERSITY OF CALIFORNIA,
IRVINE

Computational Studies of Pantetheine-Containing Ligands (PCLs) and Advancements of the Polarizable Gaussian Multipole (pGM) Model

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Mathematical, Computational and Systems Biology

by

Shiji Zhao

Dissertation Committee:
Professor Ray Luo, Chair
Professor Shiou-Chuan (Sheryl) Tsai
Assistant Professor Jin Yu

2022

DEDICATION

To my parents, Xianhu Zhao and Kai Mi,
for their unconditional love and support in my life

to my wife, Yirui Li
for always being by my side, sharing every happy and unhappy moments, and inspiring me to be a better person

TABLE OF CONTENTS

	Page
LIST OF FIGURES	iv
LIST OF TABLES	ix
ACKNOWLEDGEMENTS	xiii
VITA	xiv
ABSTRACT OF THE DISSERTATION	xvi
CHAPTER 1: Introduction	1
CHAPTER 2: Molecular Basis for Polyketide Ketoreductase–Substrate Interactions	12
CHAPTER 3: Development of a Pantetheine Force Field Library for Molecular Modeling	48
CHAPTER 4: *PyRESP*: A Program for Electrostatic Parameterizations of Additive and Induced Dipole Polarizable Force Fields	94
CHAPTER 5: Accurate Reproduction of Quantum Mechanical Many-Body Interactions in Peptide Mainchain Hydrogen Bonding Oligomers by the Polarizable Gaussian Multipole Model	158
CHAPTER 6: Transferability of the Electrostatic Parameters of the Polarizable Gaussian Multipole Model	211
APPENDIX A: Proof of Many-Body Interaction Energies Decomposition	266
APPENDIX B: The Singularity Problem of the pGM-perm and pGM-perm-v Models and Solutions	268

LIST OF FIGURES

		Page
Figure 2.1	Synthesis pathways of reducing type II PKSs, using hedamycin and actinorhodin as examples	14
Figure 2.2	Previously solved co-crystals of DM-*Act*KR bound with isoxazole-based linear poly-β-ketone mimics revealed two potential substrate-binding residue patches	16
Figure 2.3	DM-*Act*KR monomer conformation comparison and two major binding motif clusters generated from docking analysis on octaketide-bound DM-*Act*KR monomer (closed conformation)	22
Figure 2.4	Stability Score *SS* analysis of 8 ligands bound to front-patch and back-patch binding positions of DM-*Act*KR	26
Figure 2.5	Stability Score *SS* and MMPBSA comparison of DM-*Act*KR, WT-*Act*KR and WT-*Hed*KR bound with tetraketides and octaketides grouped by ligands	30
Figure 2.6	Stability Score *SS* and MMPBSA comparison of DM-*Act*KR, WT-*Act*KR and WT-*Hed*KR bound with tetraketides and octaketides grouped by KR	33
Figure S2.1	Sequence alignment among various type II PKS KRs	39
Figure S2.2	The phosphopantetheine fragment used in molecular docking	40
Figure S2.3	Front view of DM-*Act*KR displaying the relative positions of front patch, back patch and catalytic residues	41
Figure S2.4	Hanging chain effect comparison between DM-*Act*KR-tet-pp binding and DM-*Act*KR-tet-p binding	42
Figure S2.5	Electrostatic binding free energy comparison of DM-*Act*KR, WT-*Act*KR and WT-*Hed*KR bound with octaketides and tetraketides grouped by ligands	43
Figure S2.6	The position of the front patch and back patch in a native *Act*KR tetramer	44
Figure 3.1	“Plug-and-play” fragmentation strategy of PFF library development	64

Figure 3.2	Comparison of normal mode frequencies of fragments calculated with PFF/RESP and B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels of theories	68
Figure 3.3	Comparison of RESP charges, Gasteiger charges, AM1-BCC charges with unconstrained fragmental partial charges for standalone phosphopantetheine	70
Figure 3.4	Gaussian kernel density estimates (KDEs) of computed RMSD values of heavy atoms of contact residues and PCLs relative to the experimental structures	73
Figure 3.5	Correlation analysis of standardized simulated and experimental B-factors of the contact residues and the PCLs for the PPAT-Ppant system	74
Figure 3.6	Visual comparison of standardized simulated and experimental B-factors for the PPAT-Ppant system	76
Figure 3.7	T-test of binding stability scores of the last 50 ns trajectories of PPAT-Ppant, HGMS/ACP-Ppant-Ser, and EctA-CoA	77
Figure S3.1	Φ/Ψ values of covalently bound phosphopantetheinyl-serine (Ppant-Ser) from the protein data bank for a total of 320 data points	80
Figure S3.2	Comparison of QM and MM minimized structures of “plug and play” fragments	81
Figure S3.3	Comparison of adenosine structures minimized with PFF parameter sets, OL3 parameter sets, and B3LYP/6-311+G(2d,p) level of theory	82
Figure S3.4	Comparison of normal mode frequencies of fragments calculated with PFF/Gasteiger and B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels of theory	84
Figure S3.5	Comparison of normal mode frequencies of fragments calculated with PFF/AM1-BCC and B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels of theory	85
Figure S3.6	Correlation analysis of standardized simulated and experimental B-factors of contact residues and PCLs in the HGMS/ACP-Ppant-Ser system	86

Figure S3.7	Visual comparison of standardized simulated and experimental B-factors of the HGMS/ACP-Ppant-Ser system	87
Figure S3.8	Correlation analysis of standardized simulated and experimental B-factors of contact residues and PCLs in the EctA-CoA system	88
Figure S3.9	Visual comparison of standardized simulated and experimental B-factors of the EctA-CoA system	89
Figure 4.1	Schematic representation of local frame permanent dipole moments of water molecule fitted with RESP-perm and RESP-perm-v electrostatic models	117
Figure 4.2	Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for water molecule	118
Figure 4.3	Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for methanol, ethane, and benzene molecules	125
Figure 4.4	Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for *N*-methyl acetamide (NMA), dimethyl phosphate (DMP) and adenine molecules	131
Figure 4.5	Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for alanine dipeptide using single- and double-conformation fittings	136
Figure S4.1	Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for serine dipeptide using single- and double-conformation fittings	149
Figure S4.2	Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for arginine dipeptide using single- and double-conformation fittings	151
Figure S4.3	Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for aspartic acid dipeptide using single- and double-conformation fittings	153
Figure 5.1	Glycine dipeptide oligomers (in the parallel β-sheet conformation) used to calculate the interaction energy, many-body interaction energy, and the non-additive and additive contributions to the many-body interaction of the Gly_2:Gly_2 oligomer	169
Figure 5.2	Overall RMSEs of the interaction energies IE(Gly_m: Gly_n), many-body interaction energies ME(Gly_m: Gly_n) as well as the non-	

	additive contribution $ME_{NA}(Gly_m: Gly_n)$ and the additive contribution $ME_A(Gly_m: Gly_n)$ to the many-body interaction energies of the tested force fields with the ωB97X-D/aug-cc-pVTZ calculated results	198
Figure S5.1	The formamide dimer hydrogen bonding conformation	199
Figure S5.2	The Gly:Gly dimer hydrogen bonding conformations	200
Figure S5.3	The Gly:Gly_2 trimer hydrogen bonding conformations	200
Figure S5.4	The Gly:Gly_3 tetramer hydrogen bonding conformations	201
Figure S5.5	The Gly_2:Gly_2 tetramer hydrogen bonding conformations	202
Figure S5.6	The Gly_3:Gly_3 hexamer hydrogen bonding conformations	203
Figure 6.1	The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water tetramer clusters	228
Figure 6.2	Visualization of QM ESPs surrounding water tetramer clusters and the differences between QM and MM calculated ESPs of the additive, pGM-ind, pGM-perm, and pGM-perm-v models	230
Figure 6.3	The $RRMS_\mu$ and $ARRMS_V$ of the WAT4, WAT6, WAT8, and WAT10 data sets of the additive, pGM-ind, pGM-perm and pGM-perm-v models parameterized with water monomer	232
Figure 6.4	The $RRMS_\mu$ and $ARRMS_V$ of the ALA-di and ALA-tet data sets of the additive, pGM-ind, and pGM-perm models parameterized with alanine dipeptides from the ALA-di data set in 1-5 conformations	236
Figure 6.5	The $RRMS_\mu$ and $ARRMS_V$ of the ALA-tet data set of the additive, pGM-ind, and pGM-perm models parameterized with alanine tetrapeptides from the ALA-tet data set in 1-5 conformations	239
Figure 6.6	The $RRMS_\mu$ and $ARRMS_V$ of the ALA-poly and GLY-poly data sets of the additive, pGM-ind, and pGM-perm models parameterized with alanine or glycine tetrapeptides	242
Figure 6.7	The transferability tests of the additive, pGM-ind, and pGM-perm models from A, T, G, and C monomers to WC base pair tetramers	247

Figure 6.8 Scatterplots of MM ESPs of the additive, pGM-ind, and pGM-perm models versus QM ESPs for representative WC base pair tetramers 248

Figure S6.1 The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water hexamer clusters 257

Figure S6.2 The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water octamer clusters 258

Figure S6.3 The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water decamer clusters 260

LIST OF TABLES

		Page
Table 2.1	MD Simulation Round 1, including 24 DM-*Act*KR-ligand complexes prepared through structure alignment using DM-*Act*KR-(m-oct-pp) and DM-*Act*KR-(m-tet-p) co-crystal structures as templates	23
Table 2.2	MD Simulation Round 2, including 4 WT-*Act*KR-ligand complexes and 4 WT-*Hed*KR-ligand complexes prepared through structure alignment, using DM-*Act*KR-(m-oct-pp) and DM-*Act*KR-(m-tet-p) co-crystal structures as templates	27
Table 3.1	Pantetheine-Containing Ligands included in the Pantetheine Force Field Library	57
Table 3.2	RMSD Between QM and PFF/RESP Optimized Fragments	66
Table S3.1	RMSD (Angstrom) between QM and PFF/Gasteiger or PFF/AM1-BCC optimized "plug and play" fragments	82
Table 4.1	Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Water Fitted with Four Electrostatic Models	116
Table 4.2	Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Methanol Fitted with Three Electrostatic Models	119
Table 4.3	Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Ethane Fitted with Three Electrostatic Models	121
Table 4.4	Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Benzene Fitted with Three Electrostatic Models	123
Table 4.5	Charges, RRMS and Molecular Dipole/Quadrupole Moments of *N*-methyl Acetamide (NMA) Fitted with Three Electrostatic Models	126
Table 4.6	Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Dimethyl Phosphate (DMP) Fitted with Three Electrostatic Models	127
Table 4.7	Charges, RRMS and Molecular Dipole/Quadrupole Moments of Adenine Fitted with Three Electrostatic Models	129

Table 4.8	RRMS and Molecular Dipole/Quadrupole Moments of Alanine Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models	134
Table 4.9	RRMS and Molecular Dipole/Quadrupole Moments of Serine Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models	137
Table 4.10	RRMS and Molecular Dipole/Quadrupole Moments of Arginine Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models	138
Table 4.11	RRMS and Molecular Dipole/Quadrupole Moments of Aspartic Acid Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models	140
Table S4.1	Permanent Dipole Moments (a.u.) of *N*-methyl Acetamide (NMA) Fitted with the RESP-perm Electrostatic Model	146
Table S4.2	Permanent Dipole Moments (a.u.) of Dimethyl Phosphate (DMP) Fitted with the RESP-perm Electrostatic Model	146
Table S4.3	Permanent Dipole Moments (a.u.) of Adenine Fitted with the RESP-perm Electrostatic Model	147
Table 5.1	The Quantum Mechanical Interaction Energies of the Formamide Dimer and the Glycine Dipeptide Dimers Calculated by the CCSD(T)/CBS and Density Functional Theory Methods (kcal/mol)	175
Table 5.2	The Interaction Energies IE(Gly_m: Gly_n) of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Molecular Mechanical Force Fields (kcal/mol)	178
Table 5.3	The Many-Body Interaction Energies ME(Gly_m: Gly_n) of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Molecular Mechanical Force Fields (kcal/mol)	183
Table 5.4	The Non-additive Contributions ME_{NA}(Gly_m: Gly_n) to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Polarizable Force Fields (kcal/mol)	187
Table 5.5	The Additive Contributions ME_A(Gly_m: Gly_n) to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers Gly_m:Gly_n	

	Calculated by ωB97X-D/aTZ and Polarizable Force Fields (kcal/mol)	188
Table 5.6	The Interaction Energies IE(Gly_m: Gly_n) and Many-Body Interaction Energies ME(Gly_m: Gly_n) of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by the pGM-perm Model with the Alternative Polarizabilities (kcal/mol)	194
Table 5.7	The Non-additive Contributions ME_{NA}(Gly_m: Gly_n) and Additive Contributions ME_A(Gly_m: Gly_n) to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by the pGM-perm Model with the Alternative Polarizabilities (kcal/mol)	194
Table S5.1	The Interaction Energies of the Two Middle Peptides IE_{mid}(Gly_m: Gly_n) in the Presence of the Neighboring Peptides of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Polarizable Force Fields (kcal/mol)	203
Table S5.2	The Interaction Energies IE(Gly_m: Gly_n) and Many-Body Interaction Energies ME(Gly_m: Gly_n) of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by the pGM-ind Model with the Alternative Polarizabilities (kcal/mol)	204
Table S5.3	The Non-additive Contributions ME_{NA}(Gly_m: Gly_n) and Additive Contributions ME_A(Gly_m: Gly_n) to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by the pGM-ind Model with the Alternative Polarizabilities (kcal/mol)	205
Table 6.1	The Mainchain Torsional Angle Constraints for Geometry Optimizations of the Alanine Dipeptides from the ALA-di Data Set and their QM Molecular Dipole Moments	220
Table 6.2	The Mainchain Torsional Angles of the Optimized Alanine Tetrapeptides in Conf1-Conf10 Conformations and ab, aL, aR, b and pII Conformations from the ALA-tet Data Set and their QM Molecular Dipole Moments	221
Table 6.3	Molecular Dipole/Quadrupole Moments and $RRMS_V$ of the A-T and G-C WC Base Pair Dimers Fitted with A, T, G, and C Monomers with the Additive, pGM-ind, and pGM-perm Models	245
Table S6.1	The QM Molecular Dipole Moments (Debye) of Alanine Polypeptides (ACE-ALA_n-NME) from the ALA-poly Data Set	253

Table S6.2 The QM Molecular Dipole Moments (Debye) of Glycine Polypeptides (ACE-ALA_n-NME) from the GLY-poly Data Set 254

Table S6.3 The QM Molecular Dipole Moments of WC Base Pair Tetramers from the BASE Data Set 256

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to my Ph.D. advisors, Prof. Ray Luo and Prof. Sheryl Tsai, whose guidance and mentoring made me an independent and mature researcher. Without them this dissertation would not have been possible. I would also like to thank my dissertation committee member, Prof. Jin Yu, and my advancement committee members, Prof. Ioan Andricioaei and Prof. Yilin Hu, for providing constructive advice to my projects.

I am also thankful to my collaborators, Prof. Yong Duan from UC Davis, Dr. Piotr Cieplak from SBP Medical Discovery Institute, Prof. Wenqi Wang from UC Irvine, and Prof. Chris Soon Heng Tan from Southern University of Science and Technology. It was my great pleasure to work with them, who made me a more open-minded and collaborative researcher.

Previous and current members of Luo lab and Tsai lab have been invaluable resources for support and advice. I am grateful to Dr. Haixin Wei, Dr. Andrew Schaub, Dr. Ruxi Qi, Dr. Vy Duong, Dr. Edward King, Dr. D'Artagnan Greene, Dr. Erick Aitchison, Dr. Gabriel Moreno, Fanglue Ni, Tianyin Qiu, Jacob Wolff, and Hau Truong. They have been not only my colleagues, but also my lifetime friends.

I would also like to thank the 2017 MCSB cohort. We came from different background, but we made a great team to study together and learn from each other during the first year of my Ph.D. study. In addition, a big thank you to previous and current MCSB administrators, including Karen Martin, Cely Dean, Naomi Carreon, Tina Rimal, and Austin Berryman, for being responsive to any questions regarding the policies of my Ph.D. training. I also want to thank my pet dog, Nemo, for giving me emotional support while working on this dissertation.

I thank American Chemical Society for permission to include Chapter Three of my dissertation, which was originally published in *Journal of Chemical Information and Modeling*, and Chapter Four and Chapter Five of my dissertation, which were originally published in *Journal of Chemical Theory and Computation*. Financial support was provided by NIH Grant GM100305, GM076330, GM093040, GM130367, and GM79383.

VITA

Shiji Zhao

EDUCATION

2017-2022 University of California, Irvine

Ph.D. in Mathematical, Computational and Systems Biology

2016-2017 Queen Mary University of London

Full Year Study Abroad Program

2013-2017 Sichuan University

B.S. in Biological Sciences

WORK EXPERIENCE

2021 Genentech

Intern – Pharma Technical Operations – Pharmaceutical Development

PUBLICATIONS

Zhao, Shiji, Piotr Cieplak, Yong Duan, and Ray Luo. "Transferability of the Electrostatic Parameters of the Polarizable Gaussian Multipole Model." (2022, in Preparation).

Zhao, Shiji, Haixin Wei, Piotr Cieplak, Yong Duan, and Ray Luo. "Accurate Reproduction of Quantum Mechanical Multi-Body Effects in Peptide Mainchain Hydrogen Bonding Oligomers of the Polarizable Gaussian Multipole Model." *Journal of Chemical Theory and Computation* (2022).

Zhao, Shiji, Haixin Wei, Piotr Cieplak, Yong Duan, and Ray Luo. "*PyRESP*: A Program for Electrostatic Parameterizations of Additive and Induced Dipole Polarizable Force Fields." *Journal of Chemical Theory and Computation* 18, no. 6 (2022): 3654-3670.

Zhao, Shiji, Andrew J. Schaub, Shiou-Chuan Tsai, and Ray Luo. "Development of Pantetheine Force Field Library for Molecular Modeling." *Journal of Chemical Information and Modeling* 61, no. 2 (2021): 856-868.

Zhao, Shiji, Fanglue Ni, Tianyin Qiu, Jacob T. Wolff, Shiou-Chuan Tsai et al. "Molecular Basis for Polyketide Ketoreductase–Substrate Interactions." *International Journal of Molecular Sciences* 21, no. 20 (2020): 7562.

Ji, Hongchao, Xue Lu, **Shiji Zhao**, Qiqi Wang, Ray Luo et al. "Matrix-Augmented Pooling Strategy for High-Throughput Drug Target Deconvolution." *Angewandte Chemie* (2022, Submitted).

Vargas, Rebecca E., Vy Thuy Duong, Han Han, Albert Paul Ta, Yuxuan Chen, **Shiji Zhao** et al. "Elucidation of WW Domain Ligand Binding Specificities in the Hippo Pathway Reveals STXBP4 as YAP Inhibitor." *The EMBO Journal* 39, no. 1 (2020): e102406.

Schaub, Andrew J., Gabriel O. Moreno, **Shiji Zhao**, Hau V. Truong, Ray Luo et al. "Computational Structural Enzymology Methodologies for the Study and Engineering of Fatty Acid Synthases, Polyketide Synthases and Nonribosomal Peptide Synthetases." *Methods in Enzymology* 622 (2019): 375-409.

Wan, Xiaoping, Jianlin Chen, Chi Cheng, Huabing Zhang, **Shiji Zhao** et al. "Improved Expression of Recombinant Fusion Defensin Gene Plasmids Packed with Chitosan-Derived Nanoparticles and Effect on Antibacterial and Mouse Immunity." *Experimental and Therapeutic Medicine* 16, no. 5 (2018): 3965-3972.

Tang, Jianxue, Yongle Xiao, Junjie Peng, **Shiji Zhao**, Xiaoping Wan et al. "Expression of Fusion Antibacterial Peptide in Recombinant Pichia pastoris and Its Bioactivity In Vitro." *China Biotechnology* 38, no. 6 (2018): 9-16.

Xiong, Qi, Jianlin Chen, Feilin Li, **Shiji Zhao**, Xiaoping Wan et al. "Co-expression of mFat-1 and pig IGF-1 Genes by Recombinant Plasmids in Modified Chitosan Nanoparticles and Its Synergistic Effect on Mouse Immunity." *Scientific Reports* 7, no. 1 (2017): 17136.

ABSTRACT OF THE DISSERTATION

Computational Studies of Pantetheine-Containing Ligands (PCLs) and Advancements of the Polarizable Gaussian Multipole (pGM) Model

by

Shiji Zhao

Doctor of Philosophy in Mathematical, Computational and Systems Biology

University of California, Irvine, 2022

Professor Ray Luo, Chair

Pantetheine-containing ligands (PCLs) play key roles in the biosynthesis of polyketides, a large family of natural products with various bioactivities. A major hurdle that remains is our poor understanding of the protein-substrate interactions of ketoreductase (KR), a key component of polyketide synthases (PKSs). Since the poly-β-ketone intermediates are highly reactive and cannot be isolated, the first project of this dissertation employed molecular dynamics (MD) simulations to interpret the transient KR-substrate interactions. Several key factors guiding KR-substrate interactions were identified, which will help further engineering of PKSs and directing biosynthesis of novel polyketides.

The reliability of MD simulations depends on the quality of the employed force field, which comprises a mathematical formula and a set of parameters to represent the potential energy of molecular systems. The parameter sets for simulating amino acids, nucleic acids, sugars, and lipids are already available. However, lack of scalable parameter sets for PCLs has hampered the computational studies of various biomolecules containing PCLs. Therefore, in the second project of this dissertation, the first Pantetheine Force Field (PFF)

library containing parameter sets for various PCLs compatible with additive force fields was developed and validated.

The extensively used additive force fields use fixed atom-centered partial charges to model atomic electrostatic interactions. However, additive force fields cannot accurately model atomic polarization effects, leading to unrealistic simulations in polarization-sensitive processes. The polarizable Gaussian Multipole (pGM) model with all atomic multipoles represented by Gaussian densities has been recently developed. In the third project of this dissertation, an electrostatic parameterization scheme for the pGM model was developed, and the accuracy and transferability of the pGM model were assessed. Encouragingly, the pGM model was shown to be accurate and transferable, which has the potential to serve as the template for developing the next-generation polarizable force field for modeling various biological systems.

CHAPTER 1

INTRODUCTION

1.1 The Molecular Basis of Type II PKS Ketoreduction Regiospecificity Remains Unexplained

Polyketides have long been recognized as an important class of natural products for medicinal applications. The total sales of polyketide-related pharmaceuticals exceed $15 billion a year.[1] In the past decade, over 5000 papers were devoted to polyketide research, with the continuing increasing rate of publications. Such an impact on healthcare, in conjunction with the growing academic interest, underscores the significance of polyketide research. At the center of polyketide biosynthesis are polyketide synthases (PKSs), the multi-domain enzyme complexes that produce a huge variety of possible products via the controlled variation of building blocks, chain lengths, and modification reactions such as reduction and cyclization. PKSs are categorized into three types based on their architectures: type I, type II, and type III,[2] among which the type II PKS, also called iterative PKS, is important yet relatively less well-understood.[3-4] Examples of type II polyketide therapeutics include but not limited to: (1) antibiotics, such as tetracyclines,[5] tetracenomycin[6] and actinorhodin[7]; (2) anticancer drugs, such as resistomycin,[8] doxorubicin[9] and mithramycin[10]; (3) anti-fungal drugs, such as pradimicin[11]; and (4) anti-HIV drugs, such as rubromycin[12] and griseorhodin[13]. Despite their extensive medicinal applications, type II polyketides are typically difficult to obtain via organic synthesis.[14-16] In contrast, by transforming the PKS gene into a host system followed by industrial fermentation, kilogram quantities of

polyketides are routinely biosynthesized.[17] As a result, there is a great interest in developing biosynthetic systems for the production of type II polyketides.

The first focus of this dissertation is the ketoreduction chemical process in polyketide synthesis. Reducing iterative type II polyketide biosynthesis proceeds through 4 common steps: (1) chain elongation, catalyzed by ketosynthase/chain length factor; (2) regiospecific reduction, catalyzed by ketoreductase (KR); (3) aromatization/cyclization, catalyzed by aromatase/cyclase; and (4) system-specific chemical modification, carried out by a variety of other enzymes.[18] During the entire process, the growing poly-β-ketone intermediates is covalently attached to the acyl-carrier protein (ACP) through the phosphopantetheine-serine linker.[19] Despite the well-known big picture of polyketides synthesis, there is still a knowledge gap in correlating the PKS structures with enzyme regiospecificity. Although the ketoreduction of type II polyketides by KR is identical to the that of fatty acid ketoreduction, their regiospecificity are drastically different: fatty acid KR reduces every carbonyl group, while type II polyketide KR specifically reduces the C9-carbonyl group.[20] In addition, different chain length specificities for polyketide KRs were observed, including rigid substrate specificity of actinorhodin KR (*Act*KR) and doxorubicin KR (*Dox*KR), and promiscuous specificity for hedamycin KR (*Hed*KR).[21-24] Since the growing poly-β-ketone intermediates are highly reactive and cannot be isolated,[25] isoxazole-based poly-β-ketone mimics[26] and computational approaches such as molecular dynamics (MD) simulations with time resolution of picosecond range have been employed to capture the transient protein-substrate interactions of KR. Previous sequencing, co-crystallization and MD studies suggested a "chain length filter" region formed by two histidine residues exists in specific KRs such as *Act*KR and *Dox*KR, while the residues at corresponding sites become tyrosine

and glycine in promiscuous *Hed*KR.[27] This dissertation aims to dive deeper into the molecular basis interpretation of KR-substrate recognition and interactions. Outcomes will identify factors important for polyketide ketoreduction, which can significantly help rationally engineer type II PKS with chemical modification and provide the basis for production of "unnatural" type II polyketides with engineered ketoreduction patterns.

1.2 PCLs are Important but Lacks Sufficient Study Due to the Absence of PFF Library

Pantetheine is the cysteamine amide analog of pantothenic acid (vitamin B_5), which is ubiquitous in nature in various forms of pantetheine-containing ligands (PCLs), including coenzyme A (CoA)[28-31] and phosphopantetheine (Ppant).[32-35] CoA is a PCL that exist in all organisms with genome sequenced to date, and around 4% of known enzymes use either CoA or CoA thioester as a substrate.[36] In all living organisms, CoA synthesis from pantothenate requires the following five steps[37-38]: (1) Pantothenate phosphorylation by pantothenate kinase to phosphopantothenate; (2) Cysteine addition catalyzed by phosphopantothenoylcysteine synthetase to phospho-N-pantothenoylcysteine (PPC); (3) PPC decarboxylation to Ppant catalyzed by phosphopantothenoylcysteine decarboxylase; (4) Ppant adenylation to dephospho-CoA by phosphopantetheine adenylyl transferase; (5) Dephospho-CoA phosphorylation to form CoA by dephosphocoenzyme A kinase. CoA is important in all living organisms as it plays two majors roles in metabolism[28-31]: (1) energy production, by participating two key steps of citric acid cycle in the form of acetyl-CoA and succinyl-CoA; and (2) fatty acid synthesis, by acting as acyl group carrier that assists in transferring fatty acid from cytosol to mitochondria during fatty acid oxidation, and from

mitochondria to cytosol during fatty acid synthesis. Another PCL focused by this dissertation is Ppant, which functions as a prosthetic group in the form of phosphopantetheinyl-serine (Ppant-Ser) by covalently linking to carrier proteins (CPs), such as acyl carrier protein (ACP) for polyketide synthases (PKSs) or fatty acid synthases (FASs), and peptidyl carrier proteins (PCP) or aryl carrier proteins (ArCP) for nonribosomal peptide synthetases (NRPSs).[32-35] The Ppant moiety is post-translationally transferred from CoA to a conserved serine residue on CPs by the action of phosphopantetheinyl transferase.[34] By forming an energy-rich thioester linkage with polyketides, fatty acids or nonribosomal peptides intermediates in their biosynthetic pathways, Ppant-Ser fulfills the demand of providing flexibility and sufficient length (approximately 2 nm) that allows the covalently tethered intermediates to have access to spatially distinct enzyme active sites.

The reliability of MD simulations depends on the quality of the employed force field, which comprises a mathematical formula and a set of parameters to represent the potential energy of molecular systems.[39] Most current force fields are additive force fields that use fixed partial charges centered on atoms to model electrostatic interactions. Additive force field parameter sets supporting simulations with standard amino acids, nucleic acids, sugars, and lipids are already available.[40-43] At the time of this writing, a search on Protein Data Bank (PDB) database revealed about 1800 PDB entries containing CoA, CoA thioesters, Ppant or Ppant thioesters, the majority of which contain CoA (689 entries) and acetyl-CoA (250 entries).[44-45] However, a search on PubMed for keywords “molecular dynamics” together with “coenzyme A” or “pantetheine” revealed only 182 and 14 publications respectively. The lack of works for MD studies on PCLs is directly linked to the lack of force field parameter sets for PCLs. In order to increase the computational accessibility to molecular interactions

related with PCLs, this dissertation developed a scalable additive PCL force field parameter set and was hosted in Pantetheine Force Field (PFF) library (http://rayluolab.org/pff-library/). Outcomes from this work can have a significant impact in facilitating researchers to conduct MD simulations of systems containing PCLs.

1.3 High Quality Polarizable Force Fields are Required to Accurately Model Polarization Effects

MD simulations of macromolecules on atomic level have been applied in a wide range of biological systems.[46] Additive force fields, also known as nonpolarizable force fields, typically use fixed atom-centered partial charges to model electrostatics. Numerous attempts have been made to modify classical additive force fields to more reliable polarizable force fields reproducing the way a molecule responds to changing environments, such as by incorporating lone pairs or other extra points, point multipoles and polarizabilities,[39] attempting to include explicit nonadditive polarization effects, i.e. the redistribution of the electron density due to an electric field exerted by other molecules.[47]

The widely adopted additive force fields cannot accurately model polarization effects. One reason is that additive force fields include the polarization response to the environment only in an averaged, mean-field manner. For example, the gas-phase water dimer interaction energy is overestimated by more than 30% in the polarizable TIP5P model.[48] Similarly, for large biomolecular systems, there are concerns that such models cannot correctly account for situations where the same nonpolarizable moiety is exposed to different electrostatic solvents.[49] Another limitation of nonpolarizable models is their use of partial atomic charges in the electrostatic models, which often lack sufficient mathematical flexibility to describe

the electrostatic potential (ESP) around molecules. Previous researches showed that the least-squares fitting of atom-centered partial charges resulted in relative root-mean-square errors of 3%–10% over a set of grid points outside the surface of representative polar small molecules.[50] While these errors were reduced by 2–3 orders of magnitude via the use of higher atomic multipoles.[48]

A great deal of effort has been directed to developing polarizable models, including the fluctuating charge models,[51-52] the Drude oscillator model,[53-56] and models incorporating induced dipoles.[57-58] The use of induced point dipoles is a classical approach with a long history.[59] The original induced dipole model of Applequist places the induced point dipoles on atom centers.[60] However, this model suffers from the so-called "polarization catastrophe": the interaction between two induced dipoles diverges at a finite distance. Thole proposed a solution by applying damping functions to the induced dipole–induced dipole interactions.[61] However, a drawback to Thole's model is that it does not describe how the induced dipoles and permanent charges interact. The polarizable Gaussian Multipole (pGM) model has been recently proposed, where all charges and multipoles are represented by Gaussian densities,[49] which naturally leads to more consistent physics, better accuracy, higher transferability, and easier coarse graining. In addition, because distributed dipole densities instead of point dipoles are induced at atomic centers are used, the pGM model can naturally avoid polarization catastrophe. This dissertation aims to develop an electrostatic parameterization scheme for the pGM model and assess the accuracy and transferability of the pGM model. These works will greatly contribute to the development of general-purpose polarizable force fields.

1.4 Overview of this Dissertation

In the subsequent five chapters of this dissertation, three projects will be introduced in detail.

Chapter 2: Computational Studies of Molecular Basis of Polyketide Ketoreductase (KR)-Substrate Interactions

- Revealed the roles of phosphorylation, pantetheine and polyketide length in KR-polyketide interactions.
- Identified key factors causing different chain length specificity of *Act*KR and *Hed*KR.

Chapter 3: Developments of Pantetheine Force Field (PFF) Library for Modeling Pantetheine-Containing Ligands (PCLs)

- Parameterized PCLs with a "plug-and-play" strategy using appropriately sized PCL fragments.
- Validated the PFF library by performing MD simulations on representative systems containing PCLs.

Chapter 4-6: Advancements of Polarizable Gaussian Multipole (pGM) Models for Accurate Modeling of Polarization Effects

- Implemented the *PyRESP* program for flexible electrostatic parameterizations for the pGM models.
- Validated the pGM models in terms of the accuracy of modeling many-body interactions and transferability.

References

1. Borchardt, J., Combinatorial biosynthesis panning for pharmaceutical gold. *Mod Drug Discov* **1999,** *2*, 22-29.
2. Shen, B., Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current opinion in chemical biology* **2003,** *7* (2), 285-295.
3. Tsai, S. C. S.; Ames, B. D., Structural enzymology of polyketide synthases. *Methods in enzymology* **2009,** *459*, 17-47.
4. Crawford, J. M.; Townsend, C. A., New insights into the formation of fungal aromatic polyketides. *Nature Reviews Microbiology* **2010,** *8* (12), 879-889.
5. Kim, E.-S.; Bibb, M. J.; Butler, M. J.; Hopwood, D. A.; Sherman, D. H., Sequences of the oxytetracycline polyketide synthase-encoding otc genes from Streptomyces rimosus. *Gene* **1994,** *141* (1), 141-142.
6. Motamedi, H.; Hutchinson, C. R., Cloning and heterologous expression of a gene cluster for the biosynthesis of tetracenomycin C, the anthracycline antitumor antibiotic of Streptomyces glaucescens. *Proceedings of the National Academy of Sciences* **1987,** *84* (13), 4445-4449.
7. Malpartida, F.; Hopwood, D., Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host. *Nature* **1984,** *309* (5967), 462-464.
8. Jakobi, K.; Hertweck, C., A gene cluster encoding resistomycin biosynthesis in Streptomyces resistomycificus; exploring polyketide cyclization beyond linear and angucyclic patterns. *Journal of the American Chemical Society* **2004,** *126* (8), 2298-2299.
9. Otten, S. L.; Stutzman-Engwall, K.; Hutchinson, C. R., Cloning and expression of daunorubicin biosynthesis genes from Streptomyces peucetius and S. peucetius subsp. caesius. *Journal of bacteriology* **1990,** *172* (6), 3427-3434.
10. Blanco, G.; Fu, H.; Mendez, C.; Khosla, C.; Salas, J. A., Deciphering the biosynthetic origin of the aglycone of the aureolic acid group of anti-tumor agents. *Chemistry & biology* **1996,** *3* (3), 193-196.
11. Dairi, T.; Hamano, Y.; Igarashi, Y.; Furumai, T.; Oki, T., Cloning and nucleotide sequence of the putative polyketide synthase genes for pradimicin biosynthesis from Actinomadura hibisca. *Bioscience, biotechnology, and biochemistry* **1997,** *61* (9), 1445-1453.
12. Martin, R.; SIERNER, O.; Alvarez, M. A.; De Clercq, E.; Bailey, J. E.; Minas, W., Collinone, a new recombinant angular polyketide antibiotic made by an engineered Streptomyces strain. *The Journal of antibiotics* **2001,** *54* (3), 239-249.
13. Li, A.; Piel, J., A gene cluster from a marine Streptomyces encoding the biosynthesis of the aromatic spiroketal polyketide griseorhodin A. *Chemistry & biology* **2002,** *9* (9), 1017-1026.
14. Woodward, R.; Logusch, E.; Nambiar, K.; Sakan, K.; Ward, D.; Au-Yeung, B.; Balaram, P.; Browne, L.; Card, P.; Chen, C., Asymmetric total synthesis of erythromcin. 1. Synthesis of an erythronolide A secoacid derivative via asymmetric induction. *Journal of the American Chemical Society* **1981,** *103* (11), 3210-3213.
15. Woodward, R.; Au-Yeung, B.; Balaram, P.; Browne, L.; Ward, D.; Au-Yeung, B.; Balaram, P.; Browne, L.; Card, P.; Chen, C., Asymmetric total synthesis of erythromycin. 2. Synthesis of an erythronolide A lactone system. *Journal of the American Chemical Society* **1981,** *103* (11), 3213-3215.
16. Woodward, R.; Logusch, E.; Nambiar, K.; Sakan, K.; Ward, D.; Au-Yeung, B.; Balaram, P.; Browne, L.; Card, P.; Chen, C., Asymmetric total synthesis of erythromycin. 3. Total synthesis of erythromycin. *Journal of the American Chemical Society* **1981,** *103* (11), 3215-3217.
17. Pfeifer, B. A.; Admiraal, S. J.; Gramajo, H.; Cane, D. E.; Khosla, C., Biosynthesis of complex polyketides in a metabolically engineered strain of E. coli. *Science* **2001,** *291* (5509), 1790-1792.
18. McDaniel, R.; Ebert-Khosla, S.; Fu, H.; Hopwood, D. A.; Khosla, C., Engineered biosynthesis of novel polyketides: influence of a downstream enzyme on the catalytic specificity of a minimal aromatic polyketide synthase. *Proceedings of the National Academy of Sciences* **1994,** *91* (24), 11542-11546.

19. Crosby, J.; Crump, M. P., The structural role of the carrier protein–active controller or passive carrier. *Natural product reports* **2012,** *29* (10), 1111-1137.
20. O'Hagan, D., Biosynthesis of fatty acid and polyketide metabolites. *Natural product reports* **1995,** *12* (1), 1-32.
21. Das, A.; Khosla, C., In vivo and in vitro analysis of the hedamycin polyketide synthase. *Chemistry & biology* **2009,** *16* (11), 1197-207.
22. Korman, T. P.; Hill, J. A.; Vu, T. N.; Tsai, S.-C., Structural analysis of actinorhodin polyketide ketoreductase: cofactor binding and substrate specificity. *Biochemistry* **2004,** *43* (46), 14529-14538.
23. Javidpour, P.; Das, A.; Khosla, C.; Tsai, S. C., Structural and biochemical studies of the hedamycin type II polyketide ketoreductase (HedKR): molecular basis of stereo- and regiospecificities. *Biochemistry* **2011,** *50* (34), 7426-39.
24. Javidpour, P.; Korman, T. P.; Shakya, G.; Tsai, S. C., Structural and biochemical analyses of regio- and stereospecificities observed in a type II polyketide ketoreductase. *Biochemistry* **2011,** *50* (21), 4638-49.
25. Harris, T. M.; Harris, C.; Hindley, K., Biogenetic-type syntheses of polyketide metabolites. In *Fortschritte der Chemie Organischer Naturstoffe/Progress in the Chemistry of Organic Natural Products*, Springer: 1974; pp 217-282.
26. Shakya, G.; Rivera Jr, H.; Lee, D. J.; Jaremko, M. J.; La Clair, J. J.; Fox, D. T.; Haushalter, R. W.; Schaub, A. J.; Bruegger, J.; Barajas, J. F., Modeling linear and cyclic PKS intermediates through atom replacement. *Journal of the American Chemical Society* **2014,** *136* (48), 16792-16799.
27. Tsai, S. C., The Structural Enzymology of Iterative Aromatic Polyketide Synthases: A Critical Comparison with Fatty Acid Synthases. *Annual review of biochemistry* **2018,** *87*, 503-531.
28. Hoagland, M. B.; Novelli, G. D., Biosynthesis of coenzyme A from phosphopantetheine and of pantetheine from pantothenate. *J. biol. Chem* **1954,** *207*, 767-773.
29. Harwood, J. L., Fatty acid metabolism. *Annual Review of Plant Physiology and Plant Molecular Biology* **1988,** *39* (1), 101-138.
30. Daugherty, M.; Polanuyer, B.; Farrell, M.; Scholle, M.; Lykidis, A.; de Crécy-Lagard, V.; Osterman, A., Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *Journal of Biological Chemistry* **2002,** *277* (24), 21431-21439.
31. Shi, L.; Tu, B. P., Acetyl-CoA and the regulation of metabolism: mechanisms and consequences. *Current opinion in cell biology* **2015,** *33*, 125-131.
32. Zhou, Z.; Lai, J. R.; Walsh, C. T., Directed evolution of aryl carrier proteins in the enterobactin synthetase. *Proceedings of the National Academy of Sciences* **2007,** *104* (28), 11621-11626.
33. Qiao, C.; Wilson, D. J.; Bennett, E. M.; Aldrich, C. C., A mechanism-based aryl carrier protein/thiolation domain affinity probe. *Journal of the American Chemical Society* **2007,** *129* (20), 6350-6351.
34. Elovson, J.; Vagelos, P. R., Acyl carrier protein X. Acyl carrier protein synthetase. *Journal of Biological Chemistry* **1968,** *243* (13), 3603-3611.
35. Byers, D. M.; Gong, H., Acyl carrier protein: structure–function relationships in a conserved multifunctional protein family. *Biochemistry and Cell Biology* **2007,** *85* (6), 649-662.
36. Mishra, P. K.; Drueckhammer, D. G., Coenzyme A analogues and derivatives: Synthesis and applications as mechanistic probes of coenzyme A ester-utilizing enzymes. *Chemical reviews* **2000,** *100* (9), 3283-3310.
37. Leonardi, R.; Jackowski, S., Biosynthesis of Pantothenic Acid and Coenzyme A. *EcoSal Plus* **2007,** *2* (2), 10.1128/ecosalplus.3.6.3.4.
38. Leonardi, R.; Zhang, Y.-M.; Rock, C. O.; Jackowski, S., Coenzyme A: back in action. *Progress in lipid research* **2005,** *44* (2-3), 125-153.
39. Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J., Polarization effects in molecular mechanical force fields. *Journal of Physics: Condensed Matter* **2009,** *21* (33), 333102.

40. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015,** *11* (8), 3696-3713.
41. Galindo-Murillo, R.; Robertson, J. C.; Zgarbova, M.; Sponer, J.; Otyepka, M.; Jurečka, P.; Cheatham III, T. E., Assessing the current state of amber force field modifications for DNA. *Journal of chemical theory and computation* **2016,** *12* (8), 4114-4127.
42. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *Journal of computational chemistry* **2008,** *29* (4), 622-655.
43. Dickson, C. J.; Madej, B. D.; Skjevik, Å. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C., Lipid14: the amber lipid force field. *Journal of chemical theory and computation* **2014,** *10* (2), 865-879.
44. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The protein data bank. *Nucleic acids research* **2000,** *28* (1), 235-242.
45. Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S., RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* **2019,** *47* (D1), D464-D474.
46. Burkert, U.; Allinger, N., Molecular Mechanics American Chemical Society. *Washington, DC* **1982**.
47. Rick, S. W.; Stuart, S. J., Potentials and algorithms for incorporating polarizability in computer simulations. *Reviews in computational chemistry* **2002,** *18*, 89-146.
48. Ren, P.; Ponder, J. W., Polarizable atomic multipole water model for molecular mechanics simulation. *The Journal of Physical Chemistry B* **2003,** *107* (24), 5933-5947.
49. Wei, H.; Qi, R.; Wang, J.; Cieplak, P.; Duan, Y.; Luo, R., Efficient formulation of polarizable Gaussian multipole electrostatics for biomolecular simulations. *The Journal of chemical physics* **2020,** *153* (11), 114116.
50. Williams, D. E., Representation of the molecular electrostatic potential by atomic multipole and bond dipole models. *Journal of computational chemistry* **1988,** *9* (7), 745-763.
51. Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X. X.; Murphy, R. B.; Zhou, R. H.; Halgren, T. A., Development of a polarizable force field for proteins via ab initio quantum chemistry: First generation model and gas phase tests. *Journal of Computational Chemistry* **2002,** *23* (16), 1515-1531.
52. Friesner, R. A., Modeling Polarization in Proteins and Protein–ligand Complexes: Methods and Preliminary Results. In *Advances in Protein Chemistry*, Academic Press: 2005; Vol. 72, pp 79-104.
53. Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., A polarizable model of water for molecular dynamics simulations of biomolecules. *Chemical Physics Letters* **2006,** *418* (1-3), 245-249.
54. Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. *Journal of Physical Chemistry B* **2007,** *111* (11), 2873-2885.
55. Patel, S.; Mackerell, A. D.; Brooks, C. L., CHARMM fluctuating charge force field for proteins: II - Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *Journal of Computational Chemistry* **2004,** *25* (12), 1504-1514.
56. Jiang, W.; Hardy, D. J.; Phillips, J. C.; MacKerell, A. D., Jr.; Schulten, K.; Roux, B., High-Performance Scalable Molecular Dynamics Simulations of a Polarizable Force Field Based on Classical Drude Oscillators in NAMD. *Journal of Physical Chemistry Letters* **2011,** *2* (2), 87-92.
57. Cieplak, P.; Caldwell, J.; Kollman, P. A., Molecular Mechanical Models for Organic and Biological Systems Going Beyond the Atom Centered Two Body Additive Approximation: Aqueous Solution Free Energies of Methanol and N-Methyl Acetamide, Nucleic Acid Base, and Amide Hydrogen

Bonding and Chloroform/Water Partition Coefficients of the Nucleic Acid Bases. *J Comput Chem* **2001,** *22* (10), 1048-1057.
58. Wang, Z. X.; Zhang, W.; Wu, C.; Lei, H. X.; Cieplak, P.; Duan, Y., Strike a balance: Optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *Journal of Computational Chemistry* **2006,** *27* (6), 781-790.
59. Vesely, F. J., N-particle dynamics of polarizable Stockmayer-type molecules. *Journal of Computational Physics* **1977,** *24* (4), 361-371.
60. Applequist, J.; Carl, J. R.; Fung, K.-K., Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *Journal of the American Chemical Society* **1972,** *94* (9), 2952-2960.
61. Thole, B. T., Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics* **1981,** *59* (3), 341-350.

CHAPTER 2

Molecular Basis for Polyketide Ketoreductase–Substrate Interactions

2.1. Introduction

Polyketides form a large family of natural products with a diverse array of chemical structure and bioactivity.[1] Many polyketides have important pharmaceutical properties and can be used as anticancer, antibiotic, and antihypercholesterol drugs.[2-4] In nature, polyketides are biosynthesized by multi-enzyme complexes called polyketide synthases (PKSs) in plants, fungi and bacteria. Because of its medical importance, there has been a vigorous effort to engineer PKSs to produce new polyketides with therapeutic potential.[1] PKSs are genetically, structurally, and enzymatically homologous to fatty acid synthases (FASs),[5] and are categorized into three types based on their architectures: type I, type II, and type III.[6] This study focuses on reducing type II PKSs found mostly in bacteria, whose products are aromatic polyketides such as actinorhodin.[2] Reducing type II polyketide biosynthesis proceeds through 4 common steps: (1) chain elongation, catalyzed by ketosynthase/chain length factor (KS/CLF); (2) regiospecific reduction, catalyzed by ketoreductase (KR); (3) aromatization/cyclization, catalyzed by aromatase/cyclase (ARO/CYC); and (4) system-specific chemical modification, carried out by a variety of other enzymes.[7] During the entire process, the growing polyketide chain is covalently attached to the acyl carrier protein (ACP) at the conserved active site serine using the phosphopantetheine linker.[8] In total, polyketide production can involve more than 20 enzyme-catalyzed reactions to produce one major product. It is the controlled selection by

PKS of starter units, chain length, reduction and cyclization patterns that result in the huge diversity of polyketides observed in nature.

Many polyketide engineering attempts have tried to take advantage of the discrete nature of each step to mix and match proteins from different systems to produce novel products. Understanding the molecular factors controlling selection is needed to successfully engineer PKS that synthesize "unnatural" natural products that can be developed into new therapeutics. Despite past research into type II PKS synthesis, how KS/CLF, ACP, and KR choreograph their respective reactions while maintaining precise chain length, regio-, and stereo-specificity remains a mystery. Such a lack of knowledge has greatly hampered type II polyketide engineering efforts.[1] Therefore, there is a need to understand the molecular basis for the chain length and regiospecificity observed in type II PKSs. This chapter focuses on elucidating the binding mechanism of the poly-β-ketone intermediate with the KR, which catalyzes the first carbonyl to hydroxyl reduction of a single carbon group to a hydroxyl group.[9] In addition, KR is also hypothesized to be able to catalyze first ring cyclization. However, it is highly selective in reducing polyketide with certain chain lengths.[10-11] Therefore, the study of the polyketide intermediates selection mechanism by KR is essential to understand how PKS controls its product outcome.

In this study, chain length specificity distinct actinorhodin KR (*Act*KR) and hedamycin KR (*Hed*KR) were used as model KRs (**Figure 2.1**). Actinorhodin is a pigmented antibiotic produced by a type II PKS from *Streptomyces coelicolor*,[7, 9] and hedamycin is a pluramycin-type antitumor antibiotic produced by *Streptomyces griseoruber*.[12-13] The actinorhodin PKS is the model system of type II PKS, and the first type II KR structure reported was *Act*KR co-

crystallized with the cofactor NADPH.[9] *Act*KR specifically reduce the C9 carbonyl group of a 16-carbon poly-β-ketone intermediate.[14] In contrast, *Hed*KR is able to reduce tetra-, octa-, undeca-, and dodeca-ketides.[15-16] *Act*KR and *Hed*KR has high sequence homology (61% sequence identity), and both KRs specifically reduce the C9 carbonyl group.[17] It remains a mystery how *Hed*KR may have higher promiscuity in terms of chain length control than that of *Act*KR. Based on sequence alignment (**Figure S2.1**), in and around the KR active site, H153 and H201 are conserved among many type II KRs but not *Hed*KR, which has Tyrosine and Glycine at these two positions. This difference led us to hypothesize that these residues could be responsible for controlling what length of poly-β-ketone intermediates could successfully enter the active site. To test our hypothesis, we created H153Y/H201G double mutant *Act*KR with the expectation that DM-*Act*KR will have similar promiscuity as *Hed*KR. (Manuscript in preparation)

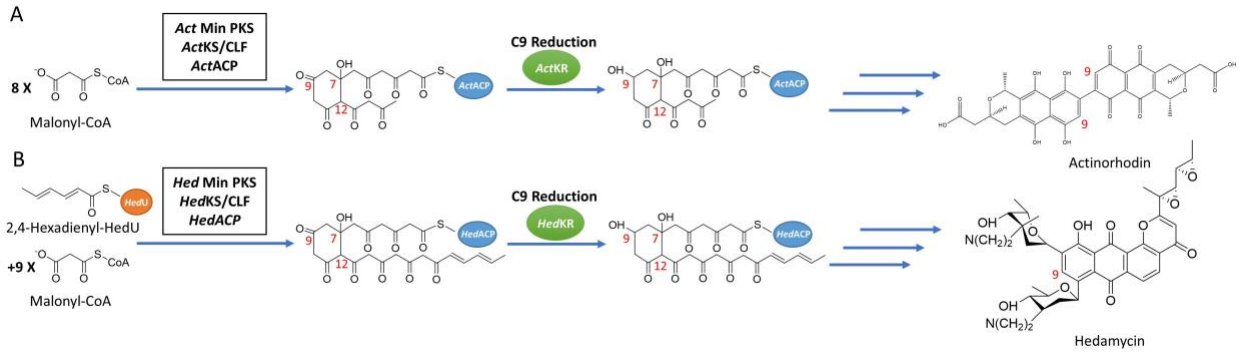


Figure 2.1. Synthesis pathways of reducing type II PKSs, using hedamycin and actinorhodin as examples. **A.** The synthesis pathway of actinorhodin, in which C9 reduction of the 16-carbon poly-β-ketone intermediate is catalyzed by *Act*KR. **B.** The synthesis pathway of hedamycin, in which C9 reduction of the 24-carbon poly-β-ketone intermediate is catalyzed

by *Hed*KR. Abbreviations: ACP, acyl carrier protein; CLF, chain length factor; CoA, coenzyme A; KR, ketoreductase; KS, ketosynthase; PKS, polyketide synthase.

A persisting problem of type II PKS research is that the polyketide intermediates of type II PKS are highly reactive, which are apt to have spontaneous aldol cyclization, resulting in great difficulty to isolate the intermediates and use them experimentally for X-ray crystallography studies or enzymological analysis.[18] To fully understand the binding mechanism of the KR with its polyketide intermediates, a series of stable isoxazole-based polyketide mimics were synthesized.[19] These mimics substituted some of the polyketide carbonyl groups with sulfur and isoxazole to achieve stability (**Figure 2.2A**). After extensive crystallization effort of both wild type and double mutant *Act*KRs with the mimic probes, we were able to crystallize and solve the co-crystal structures of the double mutant (DM-*Act*KR) bound with tetraketide-pantetheine and octaketide-phosphopantetheine mimics, which were used as templates for computational studies in this chapter.

The positions of the polyketide substrates binding raise an interesting question. In the structure of DM-*Act*KR bound with octaketide-phosphopantetheine mimic, the phosphate group binds closely to previously proposed positively-charged arginine patch (defined as the front-patch), which is formed by a cluster of arginine residues (R38, R65, R93) that interact with the phosphate moiety of the phosphopantetheinyl group of the incoming polyketide intermediate (**Figure 2.2B**).[9] However, in the structure bound with the tetraketide-pantetheine mimic, the pantetheine is close to another cluster of positive and amidic residues (Q149, R220, N260), which was defined as the back-patch (**Figure 2.2C**). It

would be of great interest to analyze if polyketide intermediates with different chain lengths and with or without phosphorylation would bind in different position, and if different KR conformation causes any change in the binding motif.

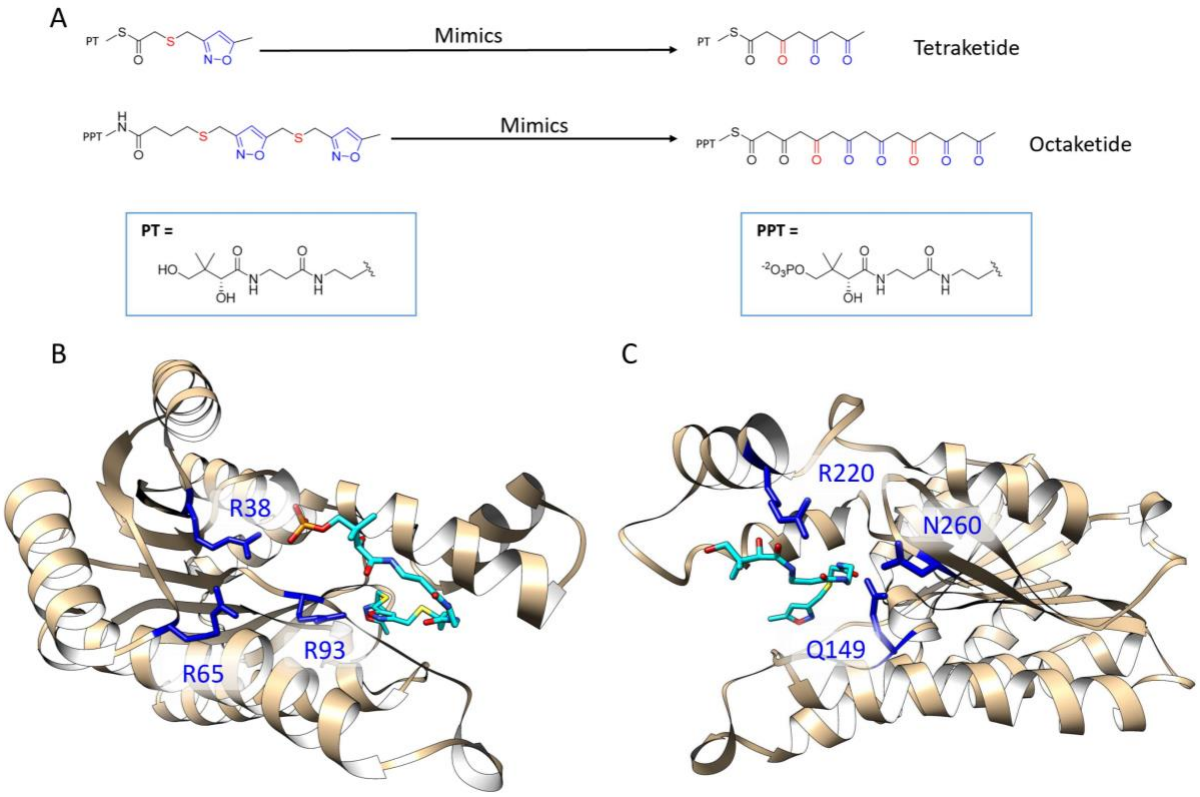


Figure 2.2. Previously solved co-crystals of DM-*Act*KR bound with isoxazole-based linear poly-β-ketone mimics revealed two potential substrate-binding residue patches (Monomers shown). **A.** Pantetheinylated (PT) tetraketide (8 carbons) and phosphopantetheinylated (PPT) octaketide (16 carbons) mimics synthesized to probe PKS active sites. Sulfur and isoxazole substitutions to replace the native carbonyls are displayed in red and blue respectively. **B.** DM-*Act*KR-octaketide-PPT co-crystal structure indicated the mimic's phosphate bound to a "front patch": R38, R65, R93. **C.** DM-*Act*KR-tetraketide-PT co-crystal

structure showed interactions between PT and a "back patch": Q149, R220, N260. Mimics are displayed in cyan; patch residues are displayed in blue.

In this chapter, we applied molecular dynamics (MD) simulations to investigate the polyketide binding mechanism from two perspectives. First, the effect of polyketide length or substrate phosphorylation on the binding orientation of polyketide intermediates (front-patch or back-patch). Second, the effect of double mutation on *Act*KR on polyketide binding. To evaluate if isoxazole-based mimics are comparable with the actual polyketide intermediates, MD simulations were conducted using both the actual polyketide intermediates and the polyketide mimics. The results from MD simulations help us understand how KR recognizes polyketide intermediates with different chain length, which will help further engineering of type II PKS and directed biosynthesis of new polyketides.

2.2 Methods

2.2.1. Molecular Docking

Molecular docking analysis was conducted using three DM-*Act*KR crystal structures as the docking templates: the previously solved DM-*Act*KR co-crystal structure bound with the octaketide-phosphopantetheine mimic, the DM-*Act*KR co-crystal structure bound with the tetraketide-pantetheine mimic, and the *apo* structure of DM-*Act*KR mutated *in silico*. Modeller[20] was used to generate the *apo* structure of DM-*Act*KR mutant *in silico* from a previously solved *apo* structure of WT-*Act*KR (PDB ID: 1X7H).[9] The models were prepared

by selecting the monomer subunits that contain the substrate mimic in each tetramer structure and manually deleting the mimic coordinates. In order to enhance the exhaustiveness and specificity of the docking analysis by limiting the amount of degrees of freedom on conformation space, a fragment containing the phosphate group and part of the pantetheine (**Figure S2.2**) was designed as the docking ligand. AutoDock Vina[21] was used with the default scoring function. The dimensions of the search box were 25.04 × 25.74 × 30.82 Å, centered to include the entire model in each run to avoid biasing binding position. Search exhaustiveness was set to 10,000 to sufficiently sample ligand binding modes. The first 200 binding modes with the highest scores were visually assessed using UCSF Chimera.[22]

2.2.2. MD Preparations

To prepare each KR-ligand complex for MD simulation, two previously solved DM-*Act*KR co-crystal structures (Manuscript in preparation) were used as templates to place ligands into the binding pocket by alignment. All ketoreductases models were prepared as tetramers to match their native multimeric state. To parameterize small molecules including the 4 co-enzyme NADPHs associated with each KR monomer and each ligand, the AM1-BCC charging method, derived from the *antechamber* program, was used,[23-24] and the *parmchk2* program was used to prepare the missing parameters. Topology and coordinate files for the KR-ligand complexes were prepared using the *tleap* module. Following parametrization, the KR-ligand complexes were solvated in an octahedral box of TIP3P water molecules with

thickness extending 10 Å from the protein surface[25] and complexes were neutralized by adding sodium ions.

2.2.3. MD Simulations

All MD simulations were performed using the *pmemd.cuda* program from the *Amber 18* software suite.[26-27] A 10 Å cutoff was used for nonbonded interactions and short-range electrostatic corrections. Long-range electrostatic interactions were handled by the particle mesh Ewald (PME) method.[28-29] The hydrogen atom bond lengths were fixed with the SHAKE algorithm.[30-31] Minimization was performed in two steps to relieve any possible atomic overlaps. The first step involved relaxing only water molecules, while the second step minimized the whole system. Langevin dynamics with a 1 ps^{-1} collision frequency were used to gradually increase system temperature from 0 to 300 K over 200 ps.[32] Prior to production stage simulations, the system was equilibrated for 100 ns under constant pressure and temperature (NPT) to adjust the system density. Finally, 100 ns production simulations without any restraint were performed under constant volume and temperature (NVT) conditions. Each simulation was repeated three times with a different random seed, starting from identical minimized structures. A 2fs integration time step was utilized with structural snapshots extracted every 1 ns.

2.2.4. MD Analysis

All simulation trajectories were visualized using the software VMD.[33] The Stability Score (*SS*) was developed to determine how stable a receptor-ligand interaction is during a simulation. The native atom pairs are defined as the heavy atom pairs that are within the distance of 7 Å in the initial frame. In any subsequent frame, the stability score is the fraction of the amount of these pairs that remain within 7Å of each other, with 1 indicating that the ligand position closely matches the initial frame, and 0 indicating ligand exit from its original binding site. Thus, *SS* of the first frame is always 1 for each trajectory, and SS is less than or equal to 1 for subsequent frames. RMSD and Stability Score *SS* of each simulation trajectories were calculated using the *cpptraj* module in *AmberTools18*.[34]

MMPBSA calculations[35-39] were conducted on the last 100 ns of each MD trajectory (frame interval is 1 ns) using the *MMPBSA.py* module in *AmberTools18*. The ionic strength was set at 0.100 M to reflect the sodium ions originally present in the simulations. Because KR active sites are highly charged, the internal dielectric constant was set to 4, which is suitable for charged receptor-ligand systems.[40] Due to time and computational resource limitations, the normal-mode-based entropy corrections to these values were not calculated as they do not improve agreement with measured affinities.

All statistical analyses were conducted by using *R* statistical packages.

2.3. Results and Discussion

2.3.1 Fragment Docking Identified Front- and Back- Patches as Two Major Binding Motifs

Three DM-*Act*KR crystal structures were used as docking templates for molecular docking analysis: a previously solved DM-*Act*KR co-crystal structure bound with the octaketide-phosphopantetheine mimic, a DM-*Act*KR co-crystal structure bound with the tetraketide-pantetheine mimic, and an apo structure of DM-*Act*KR mutated *in silico* from WT-*Act*KR (PDB ID: 1X7H). In particular, the monomer subunit that contain the substrate mimic in each tetrameric cocrystal structure was used. Three major conformations are present in the three DM-*Act*KR structures used as templates: closed, half-closed, and open conformations, corresponding to octaketide mimic-bound, tetraketide mimic-bound, and apo structures, respectively (**Figure 2.3A**). This trend in conformational variation between different ligands could be explained by the fact that a larger ligand can form more protein-ligand interactions and create a stronger ligand-enzyme interaction, leading to more closed conformation.

In order to enhance the exhaustiveness and specificity of the docking analysis by limiting the amount of degrees of freedom on conformation space, a fragment that contains the entire phosphate group and a part of pantetheine were used as the docking ligand (**Figure S2.2**). All three conformations were docked with the fragment for 10,000 independent rounds, and the first 200 binding modes with the highest scores were analyzed. The docking results reveal two major binding motifs in all three conformations, which are consistent with the previously identified front-patch and back-patch (**Figure 2.3B** and **2.3C**). There are a few other sites detected from the docking result, but they are either buried under the tetrameric interface, or the frequency is too low to be considered significant. In the open conformation, all binding poses are located at the back-patch, indicating that the open conformation provides a highly exposed binding pocket that the probe can bind to instead of

the front-patch. In the analysis of the half-closed conformation docking simulation, 3.3% of the poses were at the front-patch and 96.0% were at the back-patch. This trend is repeated in the closed conformation docking analysis, with 1.0% front-patch poses and 98.0% back-patch poses. In total, 98.0% of high-scoring binding poses appear at the back-patch regardless of conformation. This reveals the trend that binding pockets in the closed and half-closed forms tend to accept more ligands in front-patch binding poses (**Figure 2.3B**), which can be explained by a narrower back-patch binding site in combination with a wider front-patch binding site in the closed conformation binding pocket.

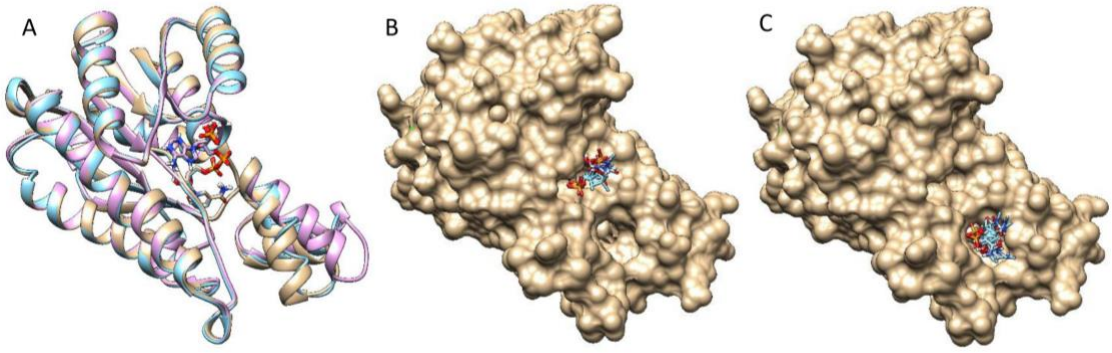


Figure 2.3. DM-*Act*KR monomer conformation comparison and two major binding motif clusters generated from docking analysis on octaketide-bound DM-*Act*KR monomer (closed conformation). **A.** Visualization of aligned monomers among the structures originally containing: octaketide-bound (gold), tetraketide-bound (cyan), and no ligand (pink). The NADPH present in all three structures is displayed; ligands deleted for clarity. **B.** The front-patch binding motif. **C.** The back-patch binding motif. The protein and NADPH surface are displayed in gold.

2.3.2 Pantetheine or Phosphopantetheine Moiety is Necessary for Ligand Binding

Twenty-four DM-*Act*KR-ligand complexes were prepared for MD simulations through structure alignment using the two DM-*Act*KR co-crystal structures solved previously as templates (**Table 2.1**), among which DM-*Act*KR-(m-oct-pp) (ligand binds to front-patch) and DM-*Act*KR-(m-tet-p) (ligand binds to back-patch) are experimental structures. Framewise Stability Score (*SS*) was developed as a measure to evaluate the binding stability of each KR-ligand pair, with *SS* close to 0 indicating weak binding, and *SS* close to 1 indicating strong binding. Each system was simulated in triplicate using identical minimized structures. 200 ns MD simulation were performed on each minimized structure, with RMSD and Stability Score *SS* plots showing that all trajectories had reached equilibrium by 100 ns (Data not shown). Surprisingly, all ligands without a pantetheine or phosphopantetheine moiety (m-tet, tet, m-oct, oct) exited the DM-*Act*KR binding pocket within 200 ns, regardless of initial binding position. These results strongly indicate that pantetheine or phosphopantetheine are essential for KR-ligand binding for any polyketide or polyketide mimic and might explain why none of our previous attempts to co-crystalize KRs with mimics lacking these moieties have ligand electron density. Thus, for further simulations, only ligands with pantetheine or phosphopantetheine moiety were prepared.

Table 2.1. MD Simulation Round 1, including 24 DM-*Act*KR-ligand complexes prepared through structure alignment using DM-*Act*KR-(m-oct-pp) and DM-*Act*KR-(m-tet-p) co-crystal structures as templates.

Aligned to DM-*Act*KR front-patch[a, b]			Aligned to DM-*Act*KR back-patch[a, b]		
m-tet	m-tet-p	m-tet-pp	m-tet	m-tet-p[c]	m-tet-pp
tet	tet-p	tet-pp	tet	tet-p	tet-pp
m-oct	m-oct-p	m-oct-pp[c]	m-oct	m-oct-p	m-oct-pp
oct	oct-p	oct-pp	oct	oct-p	oct-pp

[a] Each DM-*Act*KR-ligand pair were simulated in triplicate.

[b] Ligand nomenclatures explained. Prefix: "m" means isoxazole mimic, without "m" means natural structure; Body: "tet" means tetraketide, "oct" means octaketide; Suffix: "p" means (unphosphorylated) pantetheine, "pp" means phosphopantetheine, and without suffix means the ligand only has polyketide moiety.

[c] DM-*Act*KR-(m-oct-pp) (ligand binds to front-patch) and DM-*Act*KR-(m-tet-p) (ligand binds to back-patch) are experimental structures.

2.3.3 Polyketide Length Determines Ligand Binding Position

As discussed above, all ligands without pantetheine or phosphopantetheine do not remain bound in the DM-*Act*KR binding pocket. Therefore, there are only 16 DM-*Act*KR-ligand complexes left to be considered from **Table 2.1**. Because two potential binding sites (front-patch and back-patch) have been revealed by previous experimental structures, we investigated what key factor(s) determines the ligand binding site, i.e., given a specific polyketide ligand, which binding site the ligand would go to.

The two previously solved co-crystal structures have shown that the two mimics bind to the binding pocket of DM-*Act*KR at different sites. The phosphopantetheine moiety of m-oct-pp binds to the front-patch, while the pantetheine moiety of m-tet-p binds to the back-patch. There are two major differences between the m-oct-pp and m-tet-p mimics: polyketide length (16 and 8 carbons) and pantetheine phosphorylation (phosphorylated and not phosphorylated). It is reasonable to assume that one of the two factors determine the ligand binding position. The Stability Score *SS* of the last 100 ns of each trajectory were extracted and compared pairwise, grouped by ligand identity. Three of four octaketide ligands (m-oct-p, oct-pp, oct-p) showed higher average *SS* towards front-patch, while three of four tetraketide ligands (m-tet-p, tet-pp, tet-p) showed higher average *SS* towards back-patch (**Figure 2.4**). This indicates that polyketide length is a consistent and significant factor determining ligand binding site. Conversely, pantetheine phosphorylation is not significantly correlated with a specific ligand binding site. Surface visualization of the DM-*Act*KR binding pocket shows that the front-patch (R38, R65, R93) and the back-patch (Q149, R220, N260) form two opposite entrances of a long channel, in which the active site catalytic residues (N114, S144, Y157, K161) are located at the center (**Figure S2.3**). Shorter polyketide substrates, such as a tetraketide, may enter the active site more easily through the back-patch compared to longer substrates.

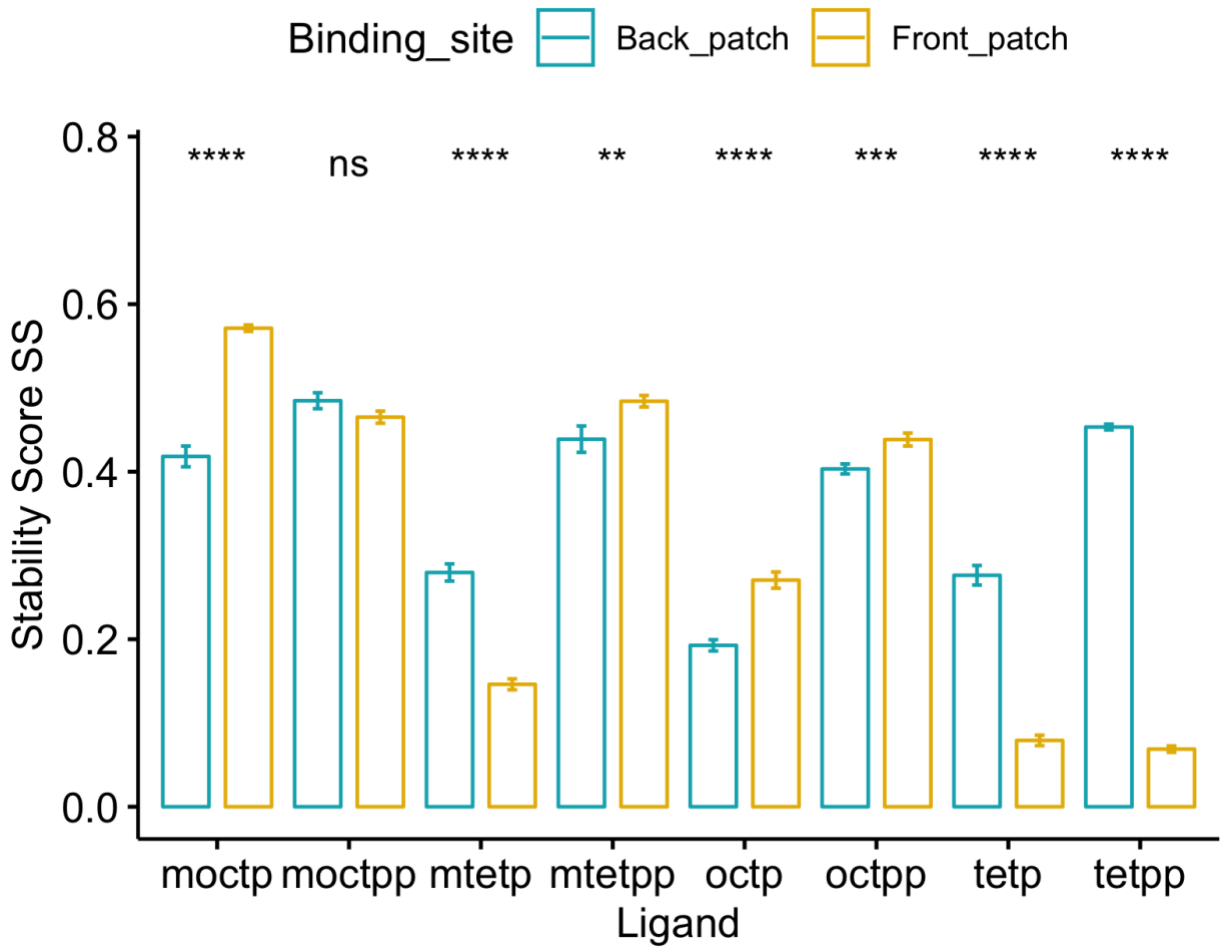


Figure 2.4. Stability Score *SS* analysis of 8 ligands bound to front-patch and back-patch binding positions of DM-*Act*KR. Among all 4 octaketide ligands, 3 of them (m-oct-p, oct-pp, oct-p) showed a significantly higher *SS* at front-patch. While among all 4 tetraketide ligands, 3 of them (m-tet-p, tet-pp, tet-p) show significantly higher *SS* at back-patch. Significance levels: ***, p <= 0.001; ****, p <= 0.0001.

2.3.4 *Act*KR H153Y/H201G Double Mutation Increases Ligand Binding Affinity

Histidines 153 and 201 near the *Act*KR active site were identified as potentially enforcing a minimum chain length based on conservation with other KRs and substitutions at those positions in the apparently more promiscuous *Hed*KR (**Figure S2.1**).[15, 41] Thus, a H153Y/H201G DM-*Act*KR was generated to test the hypothesis that DM-*Act*KR will show higher binding affinity towards polyketides with lengths that differ from *Act*KR's canonical 16 carbon substrate. Eight new KR-ligand complexes were prepared through structural alignment, including 4 WT-*Act*KR-ligand complexes as negative control and 4 WT-*Hed*KR-ligand complexes as positive control (**Table 2.2**). The structure of WT-*Act*KR was prepared by mutating Y153 and G201 of the DM-*Act*KR cocrystal structure to Histidine, and the WT-*Hed*KR structure was obtained from the Protein Data Bank (PDB ID: 3SJU). All octaketide ligands were aligned to the front-patch, and all tetraketide ligands to the back-patch, in line with the front/back-patch docking results.

Table 2.2. MD Simulation Round 2, including 4 WT-*Act*KR-ligand complexes and 4 WT-*Hed*KR-ligand complexes prepared through structure alignment, using DM-*Act*KR-(m-oct-pp) and DM-*Act*KR-(m-tet-p) co-crystal structures as templates.

Aligned to WT-*Act*KR front-patch[a, b]		Aligned to WT-*Act*KR back-patch[a, b]	
oct-pp	oct-p	tet-pp	tet-p
Aligned to WT-*Hed*KR front-patch[a, b]		Aligned to WT-*Hed*KR back-patch[a, b]	
oct-pp	oct-p	tet-pp	tet-p

[a] Each DM-*Act*KR-ligand pair were simulated in triplicate.

[b] Ligand nomenclatures are the same as **Table 2.1**.

RMSD plots as well as *SS* plots show that all trajectories had reached equilibrium after 100 ns (Data not shown). Therefore, the *SS* of the last 100 ns of each trajectory were extracted for *t*-test analysis, grouped by KR type (**Figure 2.5A**). For each ligand, the DM-*Act*KR-ligand complexes showed significantly higher average *SS* than the corresponding WT-*Act*KR-ligand complexes, and the average DM-*Act*KR-ligand complex *SS*'s are closer to the corresponding average WT-*Hed*KR-ligand complex *SS*'s than the average WT-*Act*KR-ligand complex *SS*'s.

Furthermore, MMPBSA analysis was performed to complement the Stability Score analysis.[35-39] Among the three trajectories simulated for each KR-ligand complex, the total binding free energy, ΔG_{total}, non-electrostatic binding free energy, ΔG_{vdw}, and electrostatic binding free energy, ΔG_{ele}, from the last 100 ns of the trajectory with the highest average *SS*, were used for MMPBSA *t*-test analysis, grouped by KR type. Non-electrostatic binding free energy ΔG_{vdw} reflects packing/hydrophobic effects of the system and is the sum of the VDWAALS (van der Waals energy change upon binding) and ENPOLAR terms (nonpolar solvation free energy change upon binding). Electrostatic binding free energy ΔG_{ele} reflects the electrostatic effects within the system and is the sum of the EEL (electrostatic energy change upon binding) and EPB terms (electrostatic solvation free energy change upon binding). The total binding free energy ΔG_{total} results show that the *Act*KR double mutation significantly reduces the binding free energy for octaketide ligands (oct-pp, oct-p), performing more similarly to WT-*Hed*KR than WT-*Act*KR (**Figure 2.5**). However, two interesting results were observed for the tetraketide ligands (tet-pp, tet-p).

First, while the binding energy of both ligands is close for the DM-*Act*KR, the tet-pp ligand has decreased binding energy with WT-*Act*KR compared to the tet-p ligand and is similar to the DM-*Act*KR binding energy. This shows that the presence of phosphate group on tetraketide counteract the positive effect on the binding affinity caused by the double mutation on WT-*Act*KR. A possible explanation is the "hanging-chain effect" on linear ligand binding that occurs when both ends of the ligand are tightly constrained by the binding pocket (**Figure S2.4**). This leaves the linear moiety without many interactions with nearby residues, leading to weaker binding affinity compared with those ligands with only one end constrained. The second interesting point is that the total binding free energy of WT-*Hed*KR is not significantly lower than that of WT-*Act*KR as expected, indicating that *Hed*KR is not necessarily more promiscuous than *Act*KR. In fact, whether WT-*Act*KR can reduce short polyketide intermediates is still a debatable question, because it was observed that the same products were generated from the minimal hedamycin PKS (*Hed*KS/CLF and *Hed*ACP) when combined with WT-*Act*KR or *Hed*KR.[17]

The non-electrostatic binding free energy ΔG_{vdw} results show nearly identical patterns as the total binding free energy, indicating that packing/hydrophobic effects are the main contributing factors to KR-substrates binding (**Figure 2.4C)**. On the other hand, the electrostatic binding free energy ΔG_{ele} results show random patterns compared with the total binding free energy results (**Figure S2.5**). Framewise Pearson correlation tests show that for all 12 KR-ligand pairs, the total binding free energy ΔG_{total} has higher correlation with non-electrostatic binding free energy ΔG_{vdw}, rather than electrostatic binding free energy ΔG_{ele} (Data not shown).

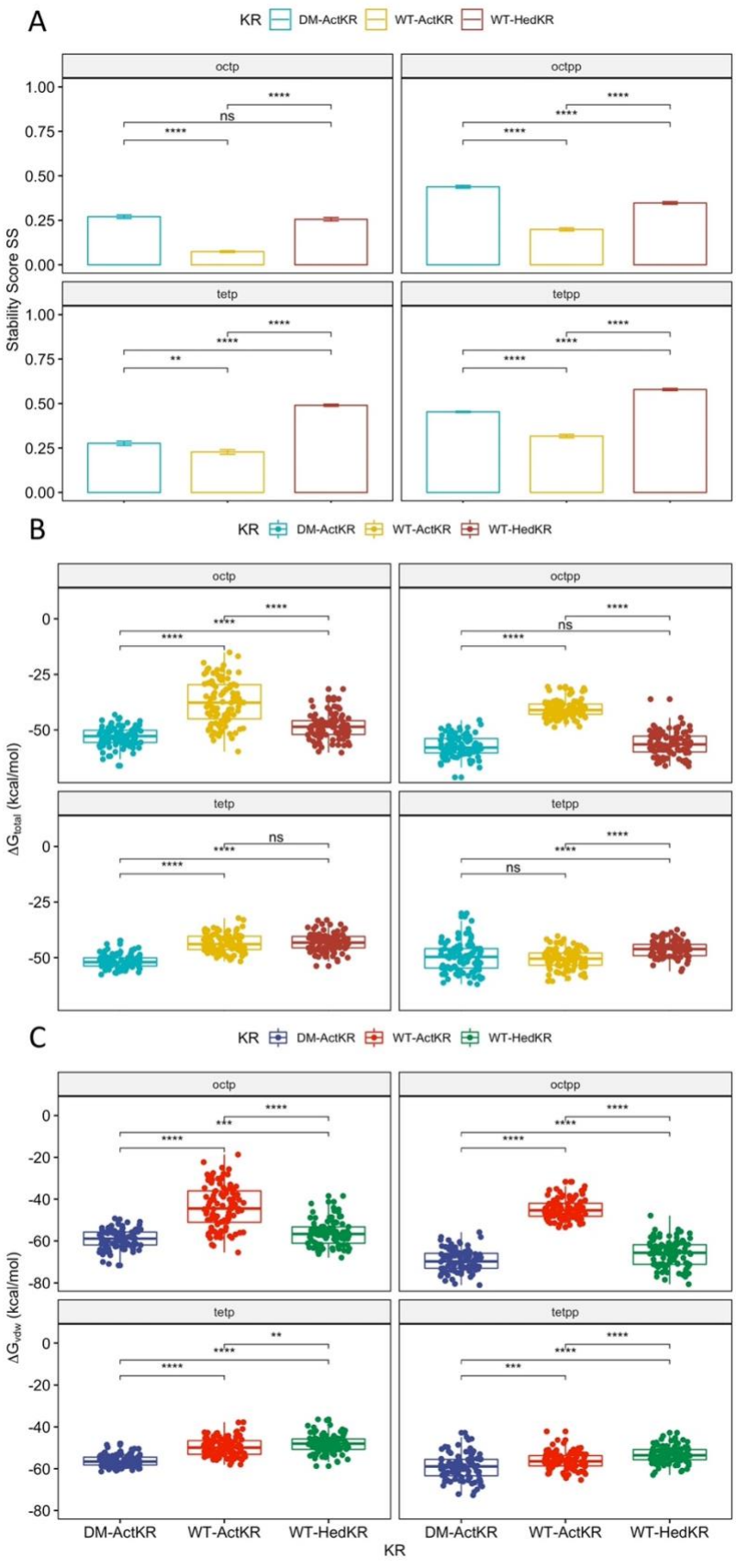

A
KR DM-ActKR WT-ActKR WT-HedKR
octp
octpp
tetp
tetpp
Stability Score SS
1.00
0.75
0.50
0.25
0.00
****
ns
**
B
KR DM-ActKR WT-ActKR WT-HedKR
ΔG_total (kcal/mol)
0
-25
-50
C
KR DM-ActKR WT-ActKR WT-HedKR
ΔG_vdw (kcal/mol)
0
-20
-40
-60
-80
***
DM-ActKR
WT-ActKR
WT-HedKR
KR

Figure 2.5. Stability Score *SS* and MMPBSA comparison of DM-*Act*KR, WT-*Act*KR and WT-*Hed*KR bound with tetraketides and octaketides grouped by ligands. **A.** For each ligand, the average Stability Score *SS* of DM-*Act*KR is significantly increased compared with WT-*Act*KR and is closer to that of WT-*Hed*KR. **B.** The total binding free energy ΔG_{total} results. The total binding free energy of DM-*Act*KR is shifted towards that of WT-*Hed*KR for octaketides, but not tetraketides. **C.** Non-electrostatic binding free energy ΔG_{vdw} shows similar pattern as **B**. Significance levels: ns, p > 0.05; *, p <= 0.05; **, p <= 0.01; ***, p <= 0.001; ****, p <= 0.0001.

2.3.5 Phosphate Group Contributes to Ligand Binding through van der Waals Interactions

Ligand positioning in the DM-*Act*KR-(m-oct-pp) and DM-*Act*KR-(m-tet-p) co-crystal structures provide grounds for the phosphate-front/back patch interaction contributing significantly to the initial ACP-phosphopantetheine-polyketide and KR docking phase. A comparison of the Stability Score *SS* of ligands with pantetheine moiety or phosphopantetheine moiety grouped by polyketide type shows that the presence of the phosphate group significantly increases KR-ligand binding stability in each KR-ligand system (**Figure 2.6A** and **Figure 2.6B**). In addition, the total binding free energy, ΔG_{total}, from the trajectory with the highest average *SS* was analyzed for each KR-ligand system (**Figure 2.6C** and **Figure 2.6D**). This analysis shows that for each ketoreductase system used, ligands with a phosphopantetheine moiety tend to have lower binding free energies than those with a pantetheine moiety, regardless of ligand length. The only exception is the DM-*Act*KR-tetraketide, where tet-pp binding free energy is higher than tet-p which might be due to the

"hanging-chain effect" as mentioned earlier (**Figure S2.4**). The non-electrostatic binding free energy, ΔG_{vdw}, results exhibit nearly identical patterns as the total binding free energy, ΔG_{total} (**Figure 2.6E** and **Figure 2.6F**). It is worth noting that phosphorylated ligand electrostatic binding free energy ΔG_{ele} to WT-*Act*KR is consistently higher than unphosphorylated ligand, while ligand electrostatic binding free energy ΔG_{ele} to DM-*Act*KR and WT-*Hed*KR follows the opposite trend (**Figure 2.6G** and **Figure 2.6H**). This indicates that the "chain length filter" mutation from Histidine to Tyrosine/Glycine swapped *Act*KR's electrostatic affinity for negatively charged phosphorylated ligands. Nonetheless, the high correlation coefficient between the total binding free energy ΔG_{total} and non-electrostatic binding free energy ΔG_{vdw} still suggests that the effect of van der Waals interactions of the phosphate group with the front/back-patch is greater than that of electrostatic interactions (Data not shown).

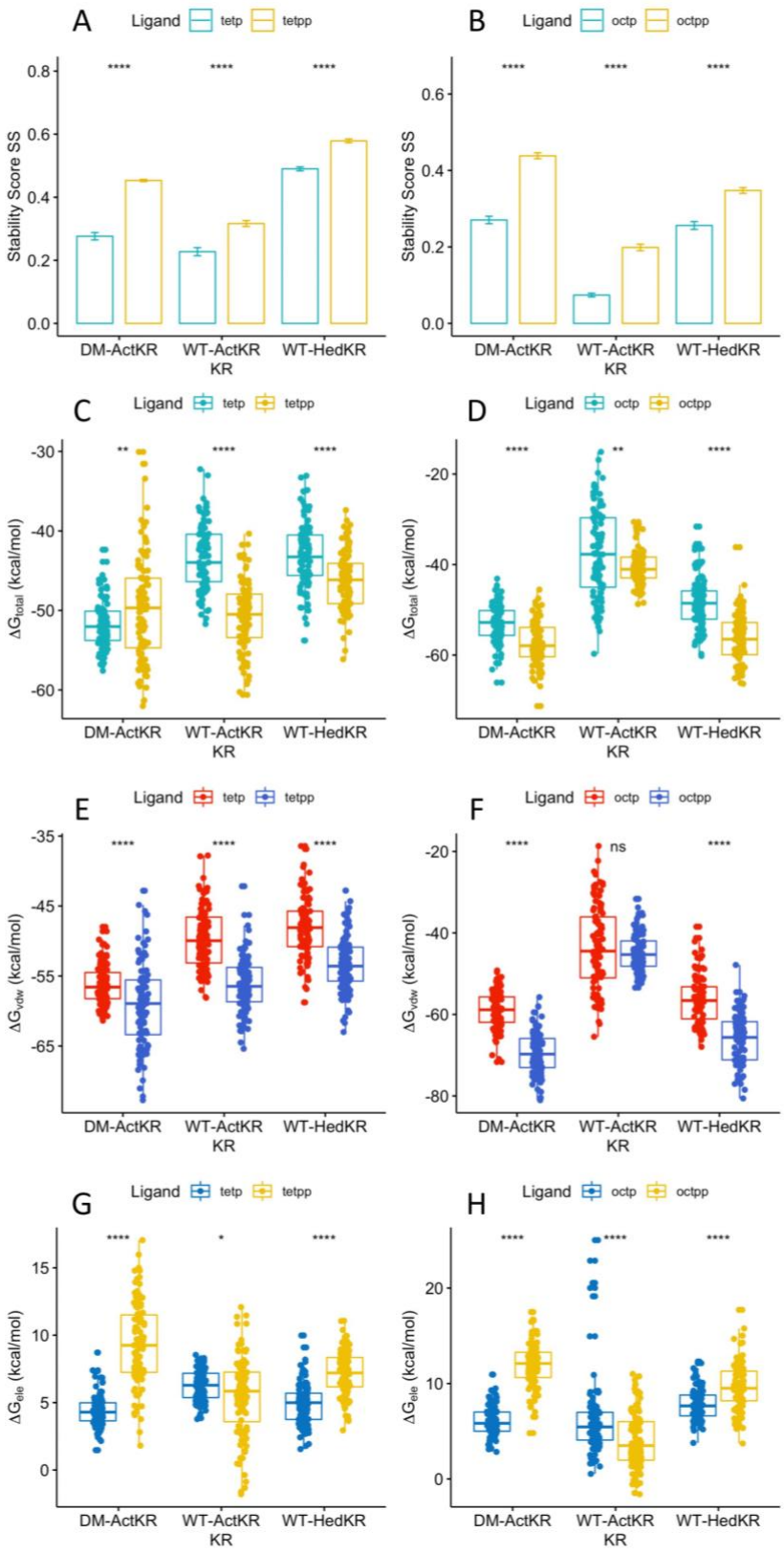

A
Ligand tetp tetpp
Stability Score SS
0.0 0.2 0.4 0.6 0.8
**** **** ****
DM-ActKR WT-ActKR KR WT-HedKR
B
Ligand octp octpp
Stability Score SS
0.0 0.2 0.4 0.6
**** **** ****
DM-ActKR WT-ActKR KR WT-HedKR
C
Ligand tetp tetpp
ΔGtotal (kcal/mol)
-30 -40 -50 -60
** **** ****
DM-ActKR WT-ActKR KR WT-HedKR
D
Ligand octp octpp
ΔGtotal (kcal/mol)
-20 -40 -60
**** ** ****
DM-ActKR WT-ActKR KR WT-HedKR
E
Ligand tetp tetpp
ΔGvdw (kcal/mol)
-35 -45 -55 -65
**** **** ****
DM-ActKR WT-ActKR KR WT-HedKR
F
Ligand octp octpp
ΔGvdw (kcal/mol)
-20 -40 -60 -80
**** ns ****
DM-ActKR WT-ActKR KR WT-HedKR
G
Ligand tetp tetpp
ΔGele (kcal/mol)
0 5 10 15
**** * ****
DM-ActKR WT-ActKR KR WT-HedKR
H
Ligand octp octpp
ΔGele (kcal/mol)
0 10 20
**** **** ****
DM-ActKR WT-ActKR KR WT-HedKR

Figure 2.6. Stability Score *SS* and MMPBSA comparison of DM-*Act*KR, WT-*Act*KR and WT-*Hed*KR bound with tetraketides and octaketides grouped by KR. All systems in **A.** and **B.** showed identical pattern that ligand with phospho-pantetheine moiety has significantly higher *SS* than those with pantetheine moiety, regardless of being octaketide or tetraketide. **C.** Total binding free energy ΔG_{total} of tetraketides binding. **D.** Total binding free energy ΔG_{total} of octaketides binding. **E.** Non-electrostatic binding free energy ΔG_{vdw} of tetraketides binding. **F.** Non-electrostatic binding free energy ΔG_{vdw} of octaketides binding. **G.** Electrostatic binding free energy ΔG_{ele} of tetraketides binding. **H.** Electrostatic binding free energy ΔG_{ele} of octaketides binding. Significance levels: ns, p > 0.05; *, p <= 0.05; **, p <= 0.01; ***, p <= 0.001; ****, p <= 0.0001.

Therefore, packing/hydrophobic effect are the main contributing factors to KR-substrate binding, as shown in MMPBSA results where ΔG_{vdw} and ΔG_{total} consistently show virtually identical patterns for each KR-ligand pair (**Figure 2.5, 2.6, S2.6**). This implies that although the electrostatic interactions between negatively charged phosphopantetheine and positively charged patches play certain role in stabilizing KR-substrate interactions, van der Waals interaction and hydrophobic effects between the uncharged polyketide moiety and binding pockets are still the dominant contributors to KR-ligands binding specificity.

2.4. Discussions

2.4.1. Sequence Analysis of *ActKR* and *Hed*KR

The antibiotic actinorhodin is synthesized by a type II PKS, which generates 16 carbon intermediate (octaketide) that is reduced by *Act*KR at the C9 carbonyl group. *Act*KR has been shown to be highly specific in reducing octaketides over other ketide lengths, with much-reduced activity for a hexaketide (12 carbons).[14] In contrast, *Hed*KR, involved in hedamycin synthesis, is much more promiscuous, reducing tetra-, octa-, nona-, undeca- and dodeca-ketides (8, 16, 18, 22, 24 carbons). It remains unknown what leads to the observed difference in substrate specificity between *Act*KR and *Hed*KR. Our previous studies identified 4 important aspects guiding *Act*KR substrate specificity[15]: (1) An Arg-rich surface patch responsible for ACP and phospho-pantetheine binding, (2) "gate" residues controlling substrate access to the active site, (3) "steering" residues that guide the pantetheine-bound polyketide towards the active site, and (4) cyclizing residues responsible for first ring cyclization. However, sequence alignment shows that some of the identified residues are conserved between *Act*KR and *Hed*KR. For instance, V151 and V154, which belong to the "steering" residues group, and Y202, which may stabilize the flexible α6-α7 helices via π-π interactions with W206, are all conserved, (**Figure S2.1**) indicating those residues are not the reasons why these two proteins have different substrate specificity. A close inspection of the sequence alignment results revealed that H153 is proximal to V151 and V154, and H201 is proximal to Y202, which are not conserved between these two proteins. Thus, the H153Y/H201G DM-*Act*KR was generated with the hope that the double mutation will increase the promiscuity of *Act*KR so that it can accept polyketides of lengths other than 16 carbon.

DM-*Act*KR was co-crystalized with pantetheinylated tetraketide and phosphopantetheinylated octaketide isoxazole mimics. Well-defined electron density of

both mimics can be observed inside the DM-*Act*KR active site pocket. As expected, DM-*Act*KR can accept both long (16 carbon) and short (8 carbon) polyketides. We have previously proposed that flexible and less conserved α6-α7 helices are important for substrate recognition,[15] the double mutant may have removed the hydrogen bonding interactions at the substrate pocket entrance that could trap shorter polyketides outside the active site.

2.4.2. Structures of ACP-polyketide-KR Complexes are Still Needed

The acyl carrier protein (ACP) is a critical component in both fatty acid and polyketide biosynthesis. Throughout synthesis, the growing product chains are bound as thiol esters at the distal thiol of the ACP's phosphopantetheine moiety and are thus transported to required protein for each synthetic step.[42] We note that we performed all the MD simulations in the absence of ACP which would be present *in vivo*. Previous studies have postulated that the positively charged front-patch that promotes complementary interactions with both helix II of the ACP and the phosphopantetheine.[43-45] The co-crystal structure of DM-*Act*KR-(m-tet-p) first identified the back-patch, which is also positively charged. However, it is noticeable that ketoreductases for type II PKSs tend to exist in the form of tetramer; therefore, only the front-patches are exposed to the outer surface, while the back-patches are buried inside the interface between attaching monomers, which may not have enough space for ACP binding (**Figure S2.6**). Thus, the back-patch may only function as ligand binding patch in experimental conditions where the ligands are not attached to ACPs. Therefore, the structures of ACP-polyketide-KR complexes are still urgently needed to reveal the natural mechanism of KR-polyketide binding in detail.

2.5. Conclusions

The regiospecific reduction of a single carbonyl group to a hydroxyl group catalyzed by ketoreductase (KR) is an essential step of reducing-type polyketide synthesis.[41] Several important ketoreductase structures from reducing type II PKS have been solved, including actinorhodin (*Act*KR)[9] and hedamycin (*Hed*KR).[15] However, the mechanism of ketoreductase-substrate interaction is still not well-known due to the fact that the poly-β-ketone intermediates which ketoreductases act on are highly reactive and prone to spontaneous cyclization, making them challenging to isolate. The primary solutions to overcome inherent reactivity to study ketoreductase-substrate interaction include substrate mimics and computer simulation.

Using these two approaches, we made five important observations on KR-substrate binding on co-crystal structures. First, docking results show that the previously proposed back-patch and front-patch residues are two major sites for substrate binding in *Act*KR, regardless of conformation. Second, polyketide length is the key determinant for which of the two sites a substrate will bind to in a KR. Third, H153 and H201 of *Act*KR are key gating residues for substrate chain length specificity, and the mutation of these two residues towards corresponding residues in *Hed*KR increased the binding affinity of *Act*KR towards polyketide substrates with different chain lengths. Fourth, pantetheine or phosphopantetheine are essential for substrate binding, and the binding affinity of most ligands with KR increased significantly in the presence of phosphate group on the ligand.

Finally, packing/hydrophobic effects are the main contributing factors to KR-substrates binding stability.

Understanding the detailed molecular basis for KR-substrate binding is crucial for rationally engineering type II PKS systems. The molecular features identified in this chapter will serve as protein engineering targets for rational control of KR specificity to produce new polyketides with pharmaceutical potential.

2.6. Supporting Information

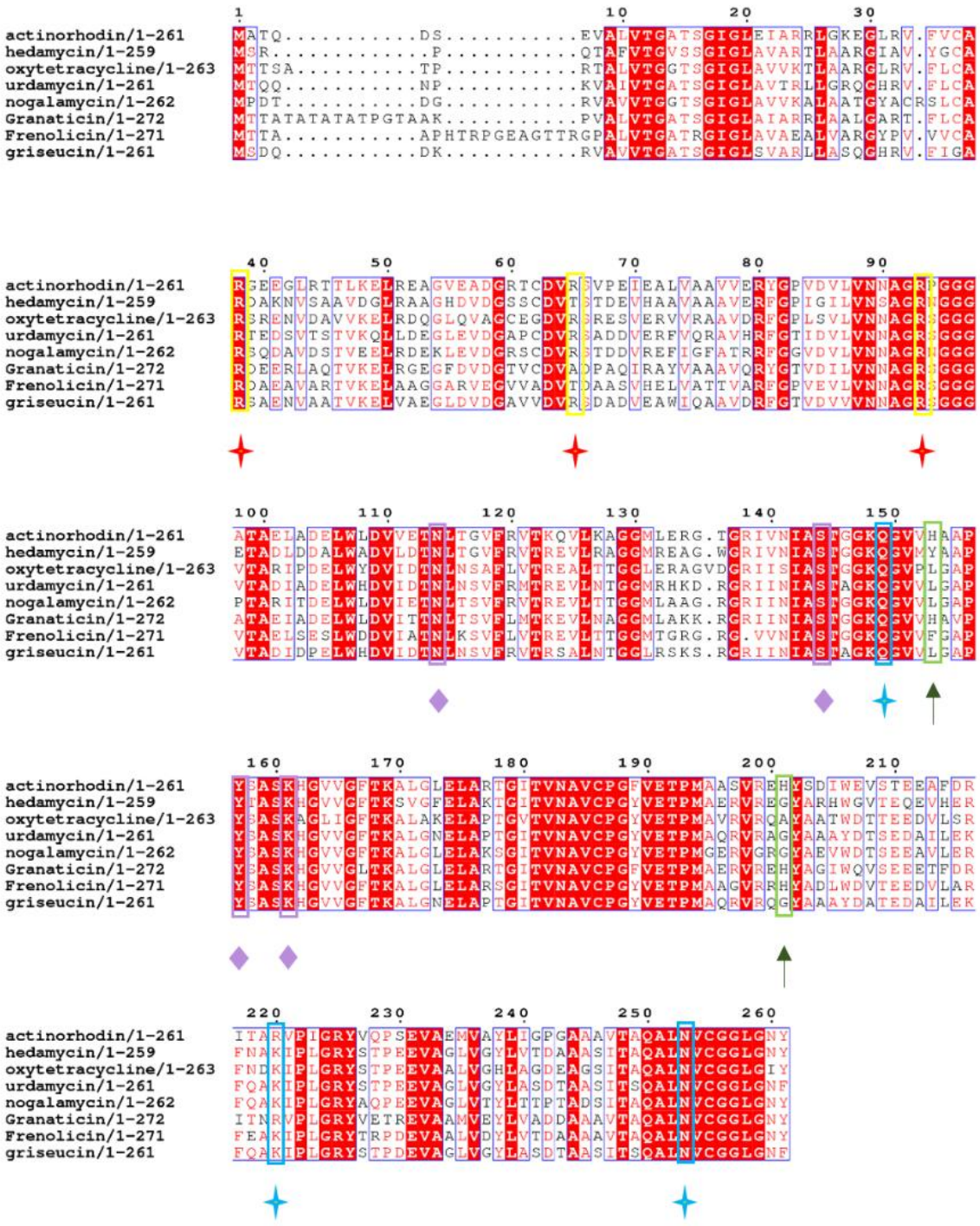
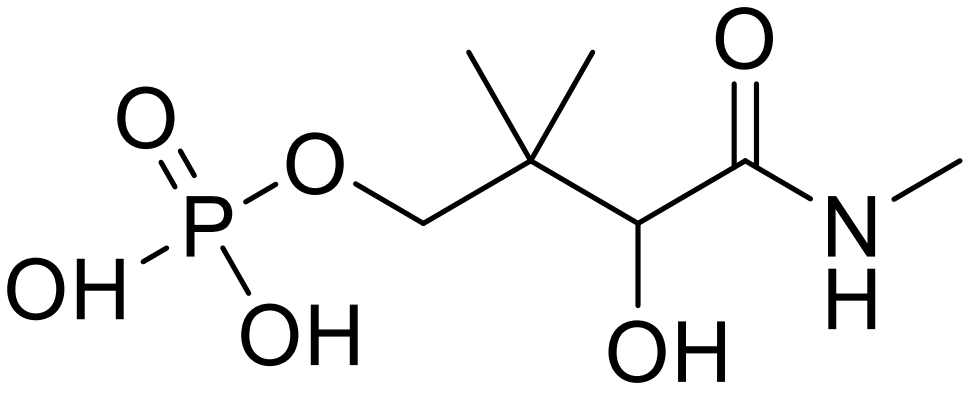


Figure S2.1. Sequence alignment among various type II PKS KRs. Sequences included actinorhodin, hedamycin, oxytetracycline, urdamycin, nogalamycin, granaticin, frenolicin, and griseucin KRs. Key: Red stars, front-patch residues; Cyan stars, back-patch residues;

Purple diamonds, catalytic residues; Green arrows, proposed chain length filter residues (double mutation targets on WT-*Act*KR).



3-hydroxy-2,2-dimethyl-4-(methylamino)-4-oxobutyl dihydrogen phosphate

Figure S2.2. The phosphopantetheine fragment used in molecular docking.

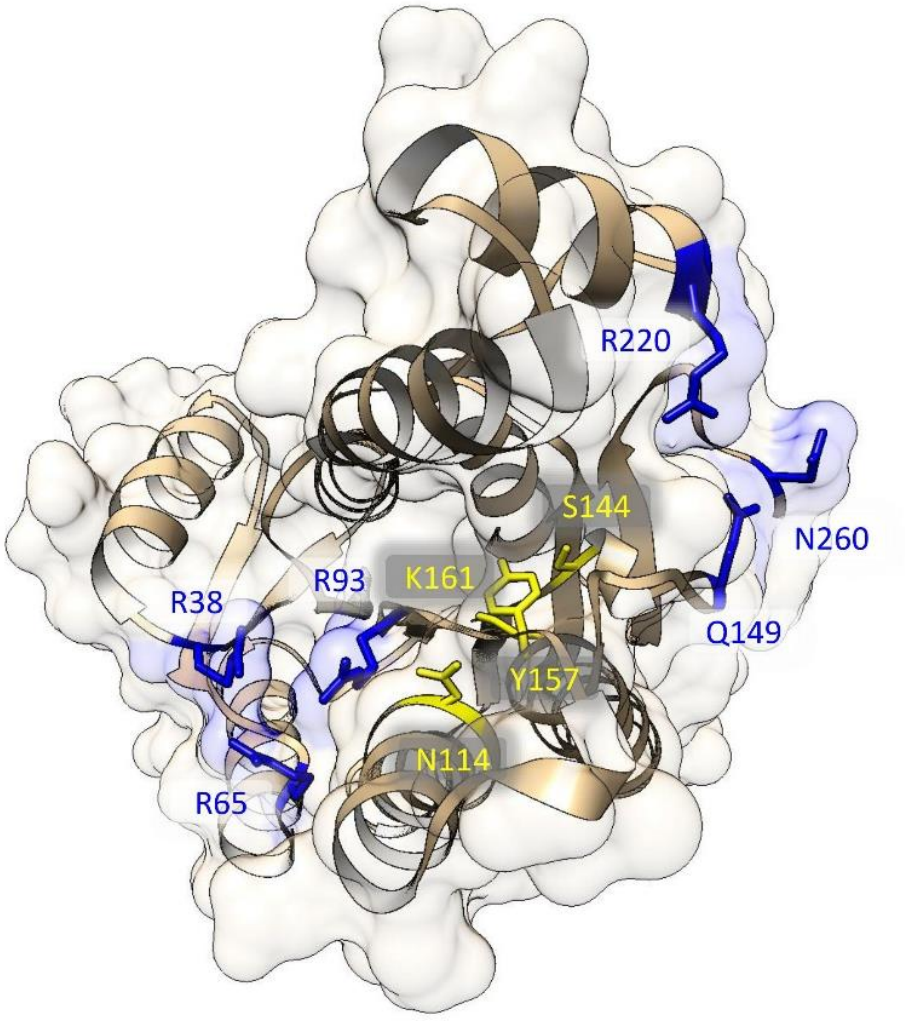


Figure S2.3. Front view of DM-*Act*KR displaying the relative positions of front patch, back patch and catalytic residues. The front-patch (R38, R65, R93) and the back-patch (Q149, R220, N260) form two opposite entrances of a long channel, in which the catalytic residues (N114, S144, Y157, K161) of active site are located at the center. Patch residues are displayed in blue and active site residues are displayed in yellow.

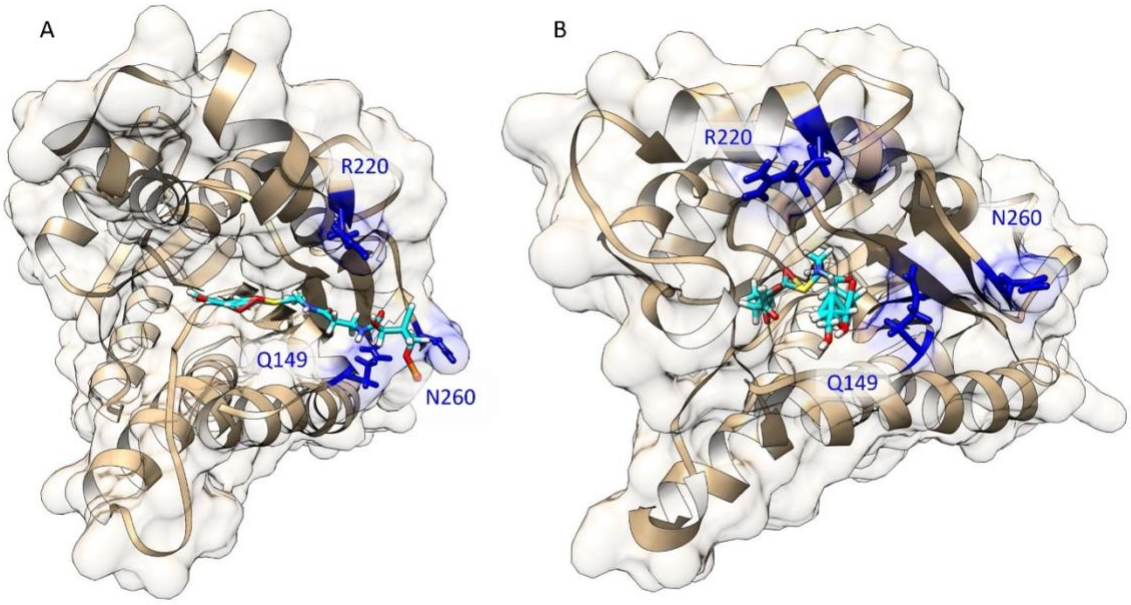


Figure S2.4. Hanging chain effect comparison between DM-*Act*KR-tet-pp binding and DM-*Act*KR-tet-p binding. Hanging chain effect is shown in DM-*Act*KR-tet-pp binding (left) but not in DM-*Act*KR-tet-p binding (right) Both figures show the average structure of the last 100ns of the simulation trajectories. In **A.**, both ends of the ligand are constrained, and DM-*Act*KR is in open form. In **B.**, the pantetheine end of the ligand is not constrained, and DM-*Act*KR is in closed form.

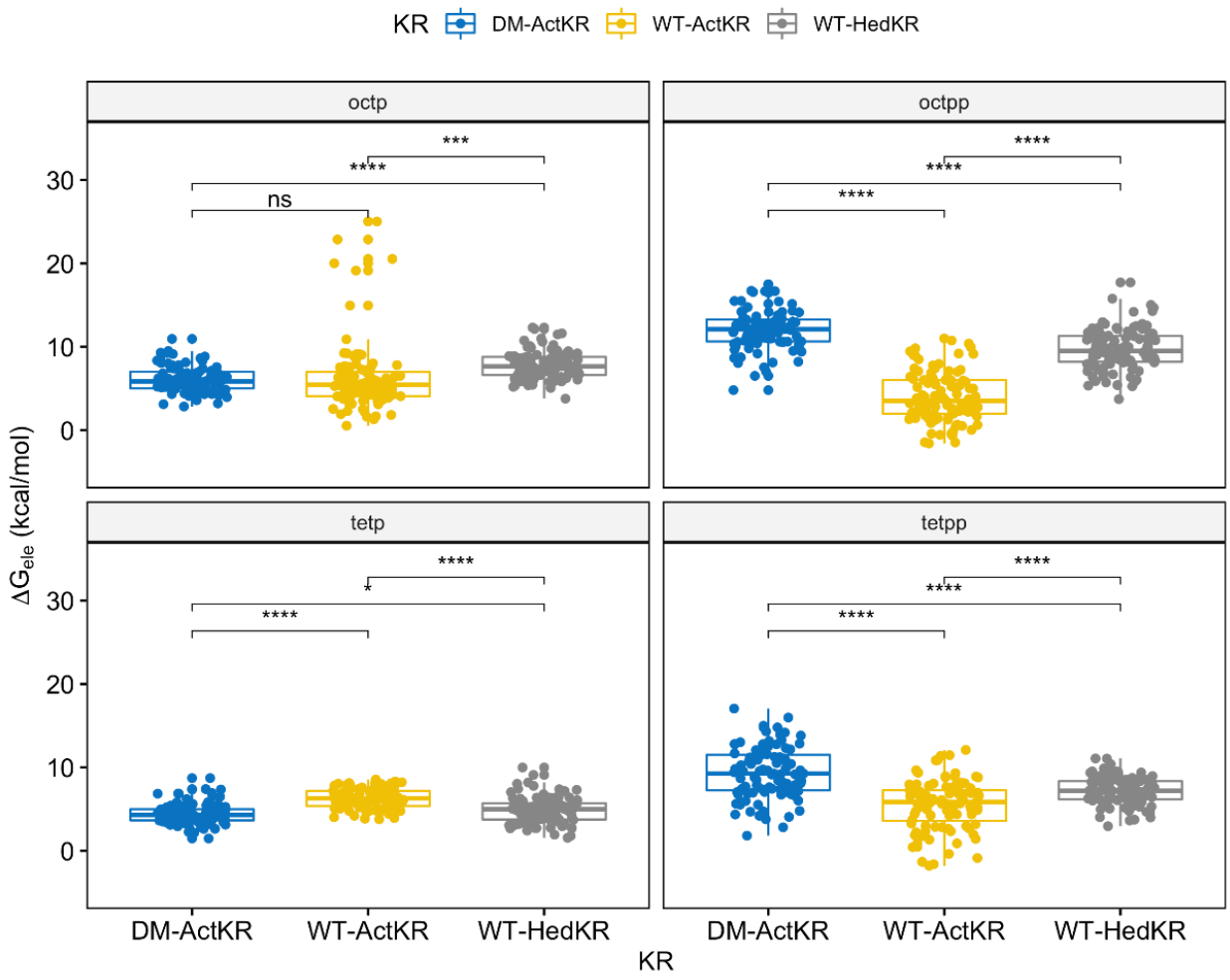


Figure S2.5. Electrostatic binding free energy comparison of DM-*Act*KR, WT-*Act*KR and WT-*Hed*KR bound with octaketides and tetraketides grouped by ligands. Each box plot shows the electrostatic energy ΔG_{ele} results. Significance levels: ns, p > 0.05; ***, p <= 0.001; ****, p <= 0.0001.

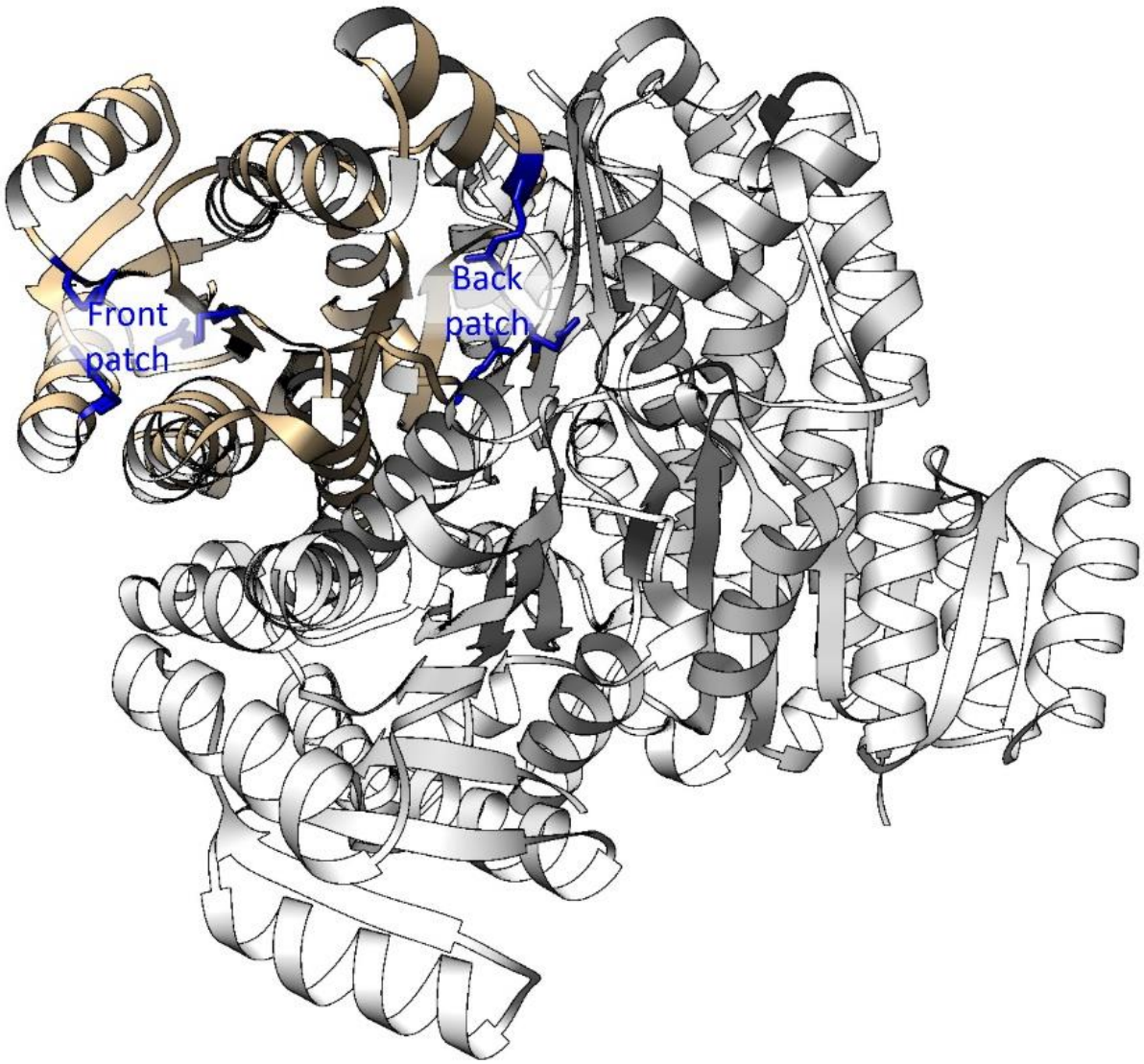


Figure S2.6. The position of the front patch and back patch in a native *Act*KR tetramer. Only the front patches (left) are exposed to the outer surface, while the back patches (right) are buried inside the interface between monomers, potentially occluding ACP binding. Front and back patches are displayed in blue.

References

1. Staunton, J.; Weissman, K. J., Polyketide biosynthesis: a millennium review. *Natural product reports* **2001,** *18* (4), 380-416.
2. Malpartida, F.; Hopwood, D., Molecular cloning of the whole biosynthetic pathway of a Streptomyces antibiotic and its expression in a heterologous host. *Nature* **1984,** *309* (5967), 462-464.
3. Otten, S. L.; Stutzman-Engwall, K.; Hutchinson, C. R., Cloning and expression of daunorubicin biosynthesis genes from Streptomyces peucetius and S. peucetius subsp. caesius. *Journal of bacteriology* **1990,** *172* (6), 3427-3434.
4. Manzoni, M.; Rollini, M., Biosynthesis and biotechnological production of statins by filamentous fungi and application of these cholesterol-lowering drugs. *Applied microbiology and biotechnology* **2002,** *58* (5), 555-564.
5. Hopwood, D. A., Genetic contributions to understanding polyketide synthases. *Chemical reviews* **1997,** *97* (7), 2465-2498.
6. Shen, B., Polyketide biosynthesis beyond the type I, II and III polyketide synthase paradigms. *Current opinion in chemical biology* **2003,** *7* (2), 285-295.
7. McDaniel, R.; Ebert-Khosla, S.; Fu, H.; Hopwood, D. A.; Khosla, C., Engineered biosynthesis of novel polyketides: influence of a downstream enzyme on the catalytic specificity of a minimal aromatic polyketide synthase. *Proceedings of the National Academy of Sciences* **1994,** *91* (24), 11542-11546.
8. Crosby, J.; Crump, M. P., The structural role of the carrier protein–active controller or passive carrier. *Natural product reports* **2012,** *29* (10), 1111-1137.
9. Korman, T. P.; Hill, J. A.; Vu, T. N.; Tsai, S.-C., Structural analysis of actinorhodin polyketide ketoreductase: cofactor binding and substrate specificity. *Biochemistry* **2004,** *43* (46), 14529-14538.
10. McDaniel, R.; Ebert-Khosla, S.; Hopwood, D. A.; Khosla, C., Engineered biosynthesis of novel polyketides. *Science* **1993,** *262* (5139), 1546-1550.
11. O'Hare, H. M.; Baerga-Ortiz, A.; Popovic, B.; Spencer, J. B.; Leadlay, P. F., High-throughput mutagenesis to evaluate models of stereochemical control in ketoreductase domains from the erythromycin polyketide synthase. *Chemistry & biology* **2006,** *13* (3), 287-296.
12. Bradner, W.; Heinemann, B.; Gourevitch, A., Hedamycin, a new antitumor antibiotic. II. Biological properties. *Antimicrobial agents and chemotherapy* **1966,** *6*, 613-618.
13. Schmitz, H.; Crook Jr, K.; Bush, J., Hedamycin, a new antitumor antibiotic. I. Production, isolation, and characterization. *Antimicrobial agents and chemotherapy* **1966,** *6*, 606.
14. Javidpour, P.; Korman, T. P.; Shakya, G.; Tsai, S. C., Structural and biochemical analyses of regio- and stereospecificities observed in a type II polyketide ketoreductase. *Biochemistry* **2011,** *50* (21), 4638-49.
15. Javidpour, P.; Das, A.; Khosla, C.; Tsai, S. C., Structural and biochemical studies of the hedamycin type II polyketide ketoreductase (HedKR): molecular basis of stereo- and regiospecificities. *Biochemistry* **2011,** *50* (34), 7426-39.
16. Bruegger, J. J., *Pantetheine Analogues Reveal Novel Characteristics in Polyketide Synthase Protein-Protein and Protein-Substrate Interactions*. University of California, Irvine: 2013.
17. Das, A.; Khosla, C., In vivo and in vitro analysis of the hedamycin polyketide synthase. *Chemistry & biology* **2009,** *16* (11), 1197-207.
18. Harris, T. M.; Harris, C.; Hindley, K., Biogenetic-type syntheses of polyketide metabolites. In *Fortschritte der Chemie Organischer Naturstoffe/Progress in the Chemistry of Organic Natural Products*, Springer: 1974; pp 217-282.
19. Shakya, G.; Rivera Jr, H.; Lee, D. J.; Jaremko, M. J.; La Clair, J. J.; Fox, D. T.; Haushalter, R. W.; Schaub, A. J.; Bruegger, J.; Barajas, J. F., Modeling linear and cyclic PKS intermediates through atom replacement. *Journal of the American Chemical Society* **2014,** *136* (48), 16792-16799.

20. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M.; Eramian, D.; Shen, M. y.; Pieper, U.; Sali, A., Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics* **2006,** *15* (1), 5.6. 1-5.6. 30.
21. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010,** *31* (2), 455-461.
22. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004,** *25* (13), 1605-1612.
23. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of computational chemistry* **2000,** *21* (2), 132-146.
24. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of computational chemistry* **2002,** *23* (16), 1623-41.
25. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.
26. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005,** *26* (16), 1668-88.
27. Le Grand, S.; Götz, A. W.; Walker, R. C., SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications* **2013,** *184* (2), 374-380.
28. Crowley, M.; Darden, T.; Cheatham, T.; Deerfield, D., Adventures in improving the scaling and accuracy of a parallel molecular dynamics program. *The Journal of Supercomputing* **1997,** *11* (3), 255-278.
29. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **1993,** *98* (12), 10089-10092.
30. Miyamoto, S.; Kollman, P. A., Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry* **1992,** *13* (8), 952-962.
31. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics* **1977,** *23* (3), 327-341.
32. Loncharich, R. J.; Brooks, B. R.; Pastor, R. W., Langevin dynamics of peptides: The frictional dependence of isomerization rates of N - acetylalanyl - N ′ - methylamide. *Biopolymers: Original Research on Biomolecules* **1992,** *32* (5), 523-535.
33. Humphrey, W.; Dalke, A.; Schulten, K., VMD: visual molecular dynamics. *Journal of molecular graphics* **1996,** *14* (1), 33-38.
34. Roe, D. R.; Cheatham III, T. E., PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation* **2013,** *9* (7), 3084-3095.
35. Miller III, B. R.; McGee Jr, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E., MMPBSA. py: an efficient program for end-state free energy calculations. *Journal of chemical theory and computation* **2012,** *8* (9), 3314-3321.
36. Wang, J.; Cai, Q.; Xiang, Y.; Luo, R., Reducing Grid Dependence in Finite-Difference Poisson-Boltzmann Calculations. *Journal of Chemical Theory and Computation* **2012,** *8* (8), 2741-2751.
37. Wang, C.; Nguyen, P. H.; Pham, K.; Huynh, D.; Le, T. B. N.; Wang, H.; Ren, P.; Luo, R., Calculating protein–ligand binding affinities with MMPBSA: Method and error analysis. *Journal of computational chemistry* **2016,** *37* (27), 2436-2446.

38. Wei, H.; Luo, R.; Qi, R., An efficient second-order poisson–boltzmann method. *Journal of Computational Chemistry* **2019,** *40* (12), 1257-1269.
39. Wei, H.; Luo, A.; Qiu, T.; Luo, R.; Qi, R., Improved Poisson–Boltzmann Methods for High-Performance Computing. *Journal of Chemical Theory and Computation* **2019,** *15* (11), 6190-6202.
40. Hou, T.; Wang, J.; Li, Y.; Wang, W., Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of chemical information and modeling* **2010,** *51* (1), 69-82.
41. Tsai, S. C., The Structural Enzymology of Iterative Aromatic Polyketide Synthases: A Critical Comparison with Fatty Acid Synthases. *Annual review of biochemistry* **2018,** *87*, 503-531.
42. Byers, D. M.; Gong, H., Acyl carrier protein: structure-function relationships in a conserved multifunctional protein family. *Biochemistry and cell biology = Biochimie et biologie cellulaire* **2007,** *85* (6), 649-62.
43. Javidpour, P.; Bruegger, J.; Srithahan, S.; Korman, T. P.; Crump, M. P.; Crosby, J.; Burkart, M. D.; Tsai, S. C., The determinants of activity and specificity in actinorhodin type II polyketide ketoreductase. *Chemistry & biology* **2013,** *20* (10), 1225-34.
44. Tang, Y.; Lee, H. Y.; Tang, Y.; Kim, C. Y.; Mathews, I.; Khosla, C., Structural and functional studies on SCO1815: a beta-ketoacyl-acyl carrier protein reductase from Streptomyces coelicolor A3(2). *Biochemistry* **2006,** *45* (47), 14085-93.
45. Korman, T. P.; Tan, Y. H.; Wong, J.; Luo, R.; Tsai, S. C., Inhibition kinetics and emodin cocrystal structure of a type II polyketide ketoreductase. *Biochemistry* **2008,** *47* (7), 1837-47.

CHAPTER 3

Development of a Pantetheine Force Field Library for Molecular Modeling

3.1. Introduction

Pantetheine is the cysteamine amide analog of pantothenic acid (vitamin B_5), which is ubiquitous in nature in various forms of pantetheine-containing ligands (PCLs). Playing a central role in energy metabolism, coenzyme A (CoA) is arguably one of the most important universal PCLs. It is present in all known organisms with genomes sequenced to date, and roughly 4% of known enzymes use either CoA or CoA thioesters as substrates.[1] Coenzyme A is important as it plays two major roles in metabolism:[2-5] (1) energy production, by participating in two key steps of the citric acid cycle in the form of acetyl-CoA and succinyl-CoA; and (2) fatty acid synthesis, by acting as an acyl group carrier that assists in transferring fatty acid from cytosol to mitochondria during fatty acid oxidation, and from mitochondria to cytosol during fatty acid synthesis. Coenzyme A synthesis from pantothenate requires the following five steps[6-7]: (1) Pantothenate phosphorylation to phsphopantothenate by pantothenate kinase; (2) Cysteinylation to phospho-N-pantothenoylcysteine (PPC) by phosphopantothenoylcysteine synthetase; (3) PPC decarboxylation to phosphopantetheine (Ppant) by phosphopantothenoylcysteine decarboxylase; (4) Ppant adenylation to dephospho-CoA by phosphopantetheine adenylyltransferase (PPAT); (5) Dephospho-CoA phosphorylation to form CoA by dephosphocoenzyme A kinase.

Another important PCL is phosphopantetheine (Ppant), which usually functions as a prosthetic group by covalently linked to carrier proteins (CPs), such as acyl carrier

protein (ACP) for fatty acid synthases (FASs) or polyketide synthases (PKSs), and peptidyl carrier proteins or aryl carrier proteins for nonribosomal peptide synthetases (NRPSs).[8-11] The Ppant moiety is post-translationally transferred from CoA to a conserved serine residue on CPs by the action of phosphopantetheinyl transferase.[10] By forming an energy-rich thioester linkage with fatty acids, polyketides, or nonribosomal peptides intermediates in their biosynthetic pathways, Ppant fulfills the demand of providing flexibility and relatively sufficient length (approximately 2 nm) that allows the covalently tethered intermediates to navigate and access spatially distinct and structurally diverse enzyme active sites.

Both CoA and Ppant play central roles in carrier protein-based biosynthesis of fatty acids, polyketides and nonribosomal peptides, ultimately providing a wide array of complex, bioactive natural products including valuable pharmaceuticals and precious commodity chemicals. For fatty acid synthesis, the simplest model system available is the type II FAS in *E. coli*. In this system, ACP interacts with more than 10 different catalytic partners, catalyzing the formation of long fatty acid chains from malonyl-CoA with high efficiency and fidelity.[12] For polyketide synthesis, besides similar mechanism for polyketide chain elongation with the participation of ACP and malonyl-CoA, nature has co-opted the assembly line strategy to produce macrocyclic polyketide natural products by utilizing additional tailoring domains for increased chemical diversity and biological function.[13] Similarly, NRPSs utilize the carrier protein machinery with elongation by amino acids instead of acyl groups.[14] Recent efforts have been made to engineer these systems to expand their product diversity as well as to optimize systems for expression in heterologous hosts.[15-16] A major hurdle that remains is our poor understanding of the transient substrate-protein interactions between the CPs with their Ppant bound intermediates, as well as protein-protein interactions between CPs and

their catalytic partner domains. Molecular dynamics (MD) and other computational techniques can be used to provide models of these transient interactions that are difficult to capture experimentally, thus providing an additional tool to increase yields and expand product diversity for the biosynthesis of "unnatural" natural products.[17-21]

The reliability of MD simulations depends on the availability and quality of molecular mechanics force fields, including both the functional form and parameter sets. Current Amber force fields provides parameter sets support modeling standard amino acids, nucleic acids, sugars, lipids, and other relatively common moieties.[22-26] At present, no scalable force field parameter set exist for PCLs. Performing MD simulations on systems containing PCLs require extra parameterization works each time, thus reducing the computational accessibility to potentially critical information on protein-protein and protein-substrate interactions. In addition, non-standard residues, such as a phosphopantetheinyl-serine (Ppant-Ser) covalently embedded in a protein, require more efforts in parameterization. Furthermore, the size of CoA, Ppant and Ppant-Ser "apo" ligands and their corresponding thioesters are at least 80, 43 and 52 atoms, respectively, making their parameterization processes computationally expensive and time consuming. At the time of this writing, a search on Protein Data Bank (PDB) database returns about 1700 entries containing CoA, CoA thioesters, Ppant or Ppant thioesters, the majority of which contain CoA (603 entries) and acetyl-CoA (222 entries).[27-28] However, a search on PubMed for keywords "molecular dynamics" with "coenzyme A" or "pantetheine" reveals only 141 or 9 publications respectively. The limited literature for MD studies of PCLs is directly linked to the lack of pantetheine force field (PFF) parameters. The availability of a PFF library would allow

researchers to model these enzymes for engineering efforts and provide medicinal chemists better models for drug design efforts.

Here we report a PFF library built specifically to model and simulate systems containing PCLs compatible with standard Amber force fields,[22] including 12 standalone CoA or CoA-thioesters, 9 standalone Ppant or Ppant-thioesters, and 9 covalently linked Ppant-Ser or Ppant-Ser-thioesters with compatible nomenclatures with Protein Data Bank. The atomic partial charge parameters were calculated by one of three charging algorithms, including Gasteiger,[29] AM1-BCC[30-31] and restrained electrostatic potential (RESP) matching similar techniques employed in current Amber force fields.[32-33] Inspired by the development of LIPID11 force field,[34] a "plug-and-play" parameterization scheme utilizing modular splitting was employed to simplify the computational complexity of using the RESP algorithm, resulting in a fragmentation strategy that allows for systematic charging of large molecules sharing common substructural motifs. The remaining parameters such as those for bond terms, angle terms, dihedral angle terms were adopted from either ff14SB[23] or gaff2.[35] This library is expected to have a significant impact on researchers who wish to conduct MD simulations of any system that requires PCLs as either substrates or cofactors.

3.2. Methods

3.2.1. Structural Preparation

Structural files of PCLs in the CIF (Crystallographic Information File) format containing observed and idealized structures (calculated by software such as OpenEye's Omega based on the known covalent geometry) were obtained from the RCSB Protein Data

Bank.[28, 36] The original hydrogens on each PCL structural file were removed and the Amber/*Reduce* program was used to add hydrogens matching its physiological protonation state.[37] A "plug-and-play" fragmentation scheme was employed for the computationally expensive RESP charging method, which splits CoA, Ppant and Ppant-Ser into a pool of 8 fragments: (1) methylphosphate, (2) adenosine, (3) dimethyldiphosphate, (4) pantoic acid, (5) beta-alanine, (6) cysteamine, (7) serine dipeptide and (8) dimethylphosphate. Fragments 1-6 were obtained from the structural file of CoA (PDB ID: COA); Fragments 7 and 8 were obtained from the structural file of phosphoserine (PDB ID: SEP); Extending fragment for each selected PCL was obtained from corresponding structural file directly. Fragments were capped with acetyl, methylamide, methyl, and/or hydroxyl groups using the *Build Structure* feature of UCSF Chimera.[38]

3.2.2. PFF Parameterization

The RESP ESP charge Derive (R.E.D.)III.5 tools were used for RESP charge fitting for each "plug and play" fragment.[39] Gaussian 09 was used to optimize the geometry of each fragment at B3LYP/6-31G* level of theory, and to derive molecular electrostatic potential (MEP) at HF/6-31G* level of theory.[40] Extra care was taken during the optimization of the serine dipeptide fragment (fragment 7), where ϕ and ψ angles were constrained at $-60.70°$ and $-31.32°$ respectively. A four-step RESP fitting strategy was employed to derive final RESP partial charges, including: (1) charge fitting for each fragment independently; (2) pairwise charge fitting for each pair of connecting fragments, with intermolecular charge constraints applied on corresponding caps whose net charge was constrained to 0; (3)

fragment merging by averaging the two different charges of each atom derived at step (2); and (4) charge scaling to ensure integer total charges of intact molecules with the following equation:

$$C_{i,scaled} = C_i \times \frac{C_{tot}}{\sum_1^N C_i}$$

where C_i is the partial charge of the *i*th atom of the molecule before normalization, and C_{tot} is the total integer charge of the molecule. To reduce the charging error, Rigid-Body Reorientation Algorithm (RBRA) embedded in R.E.D.-III.5 were applied in step (1).[39] Amber/*Antechamber* program was used to conduct the Gasteiger and AM1-BCC charge fitting procedures.[41]

Non-charge parameters include those for bond, angle, dihedral angle and van der Waals terms. For covalent Ppant-Ser PCLs, these parameters were first derived from ff14SB force fields where possible.[23] Missing parameters were adopted from gaff2.[35] For standalone CoA and Ppant PCLs, non-charge parameters were derived from gaff2 force field directly. The parameterization process was handled by *parmchk2* program to obtain parameter modification (frcmod) files. Finally, the Amber/*tleap* program was used to generate OFF library (lib) files.[42]

3.2.3. Structural and Normal Mode Analysis of Fragments

Fragment geometries were sequentially minimized using increasing levels of theories in the order of B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ using Gaussian 09, after which the QM normal mode frequencies were obtained.[40] Scaling factors

of 0.967, 0.959 and 0.953 were applied to B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ calculated normal mode frequencies, respectively, as suggested by precomputed scaling factors of Computational Chemistry Comparison and Benchmark DataBase (CCCBDB).[43] For molecular mechanical minimization with PFF and OL3 force fields, the Amber/*pmemd* program was used.[42, 44] PFF normal mode analysis was then performed using the *nmode* function of the Nucleic Acid Builder (NAB) language.[45]

Structural alignment and RMSD calculation between QM minimized structures and PFF or OL3 minimized structures were conducted using the *match* command of UCSF Chimera.[38]

3.2.4. MD Preparation

Three PCL containing systems selected for validation purposes including phosphopantetheine adenylyltransferase-phosphopantetheine (PPAT-Ppant, PDB ID: 1OD6),[46] 3-hydroxy-3-methylglutaryl synthase/acyl carrier protein complex (HGMS/ACP-Ppant-Ser, PDB ID: 5KP7),[47] and diaminobutyrate acetyltransferase-Coenzyme A (EctA-CoA, PDB ID: 6SK1).[48] Missing residues in PPAT and HGMS/ACP were added using *modeller*.[49] Topology and coordinate files were prepared using the Amber/*tleap* program, with standard residues parameterized by the ff14SB force field, and PCLs parameterized by the PFF library.[23, 42] Following parametrization, each system was solvated in an octahedral box of TIP3P water molecules with thickness extending 10 Å from the protein surface.[50] Complexes were neutralized by adding counterions with opposite charges (sodium or chloride), and

extra sodium-chloride ion pairs were added to match reported experimental salt concentrations.

3.2.5. MD Simulations

The Amber/*pmemd.cuda* program was used for all MD simulations.[42, 44] A 10 Å cutoff was set for nonbonded interactions and short-range electrostatic corrections. The SHAKE algorithm was used to constrain the hydrogen atom bond lengths,[51-52] and the particle mesh Ewald (PME) method was use used to handle long-range electrostatic interactions.[53-54] Energy minimization was performed to relieve any possible atomic spatial conflicts in two stages. The first stage was used to relax only water molecules and ions, while the second stage was used to relax the whole system. Langevin dynamics with a 1 ps^{-1} collision frequency were used to gradually increase system temperature from 0 K to reported experimental temperatures over 200 ps.[55] The systems were first equilibrated for 100 ns under constant pressure and temperature (NPT) to adjust the system density, then 100 ns production simulations were performed under constant volume and temperature (NVT) conditions. Both equilibration and production phases employed 2 fs integration time step, and 200 ps interval for simulation snapshot extraction. Each simulation was repeated in triplicates with different random seeds, starting from identical minimized structures.

3.2.6. MD Analysis

MD simulation results were analyzed using 3 metrics: comparisons of RMSD between simulated and experimental conformations, comparisons of simulated and experimental B-factors, and our previously developed binding stability scoring.[56] All metrics of each simulation were calculated using the Amber/*cpptraj* program, employing commands *rmsd*, *atomicfluct*, and *nativecontacts* respectively.[57] Simulated B-factor calculations only included snapshots of the last 10 ns. Both experimental and simulated B-factors were standardized using the following equation:

$$B_{i,standardized} = \frac{B_i - \mu}{\sigma}$$

where μ and σ are the mean value and standard deviations of all B-factors. The stability score (SS) was developed to determine the binding stability of receptor–ligand pair during simulation.[56] The native atom pairs are defined as the heavy atom pairs that are within the distance of 7 Å in the crystal structure, and the stability score is calculated using the following equation:

$$SS = \frac{1}{f_{end} - f_{start} + 1} \sum_{f_{start}}^{f_{end}} SS_i$$

where the stability score of the *i*th frame SS_i is the fraction of the amount of these pairs that remain within 7 Å of each other. f_{start} and f_{end} are the start and end frame numbers, respectively. In this chapter, f_{start} were set as 101 and f_{end} were set as 200 to include the trajectory snapshots of the last 100 ns. Gaussian Kernel Density Estimation (KDE) plots for RMSD and scatterplots for standardized B-factors were generated by *Matplotlib* package of Python. B-factor visualizations were generated using the *Render by Attribute* feature of UCSF

Chimera.[38] Statistical analyses of stability scores were conducted by using the R statistical package.

3.3. Results and Discussion

3.3.1. PFF Library Design

The current pantetheine force field (PFF) library includes parameters for 30 PCLs available in Protein Data Bank. (**Table 3.1**) Besides "apo" CoA, Ppant and Ppant-Ser, the PFF library contains thioesters of CoA, Ppant and Ppant-Ser with extending units from saturated fatty acids, whose lengths range from 3 carbons to 16 carbons, or the intermediates of fatty acid synthesis, including acetyl-, malonyl-, acetoacetyl-CoA, acetyl-Ppant and acetyl-Ppant-Ser. All PCLs included in the CoA Library and the Ppant Library are standalone ligands, and all PCLs included in the Phosphopantetheinyl-Serine (Ppant-Ser) Library are non-standard residues covalently linked to proteins. The URL link to the individual page of each PCL is also shown in **Table 3.1**.

Table 3.1. Pantetheine-Containing Ligands included in the Pantetheine Force Field Library

PCL Name	PDB ID	Description	Entries in PDB	URL Links
Coenzyme A (CoA) Library				

CoA	COA	“apo” coenzyme A	529	http://rayluolab.org/pff-files-for-coenzyme-a/
Acetyl-CoA	ACO	2 Carbon Acyl-CoA	218	http://rayluolab.org/pff-files-for-acetyl-coa/
Propionyl-CoA	1VU	3 Carbon Acyl-CoA	11	http://rayluolab.org/pff-files-for-propionyl-coa/
Butyryl-CoA	BCO	4 Carbon Acyl-CoA	8	http://rayluolab.org/pff-files-for-butyryl-coa/
Hexanoyl-CoA	HXC	6 Carbon Acyl-CoA	10	http://rayluolab.org/pff-files-for-hexanoyl-coa/
Octanoyl-CoA	CO8	8 Carbon Acyl-CoA	14	http://rayluolab.org/pff-files-for-octanoyl-coa/
Decanoyl-CoA	MFK	10 Carbon Acyl-CoA	5	http://rayluolab.org/pff-files-for-decanoyl-coa/
Dodecanoyl-CoA	DCC	12 Carbon Acyl-CoA	7	http://rayluolab.org/pff-files-for-dodecanoyl-coa/
Tetradecanoyl-CoA	MYA	14 Carbon Acyl-CoA	89	http://rayluolab.org/pff-files-for-tetradecanoyl-coa/

Hexadecanoyl-CoA	PKZ	16 Carbon Acyl-CoA	2	http://rayluolab.org/pff-files-for-hexadecanoyl-coa/
Malonyl-CoA	MLC	CoA derivative of malonic acid	12	http://rayluolab.org/pff-files-for-malonyl-coa/
Acetoacetyl-CoA	CAA	Precursor of HMG-CoA in mevalonate pathway	30	http://rayluolab.org/pff-files-for-acetoacetyl-coa/
Phosphopantetheine (Ppant) Library				
Ppant	PNS	"apo" phosphopantetheine	11	http://rayluolab.org/pff-files-for-phosphopantetheine/
Acetyl-Ppant	6VG	2 Carbon Acyl-Ppant	0	http://rayluolab.org/pff-files-for-acetyl-ppant/
Butyryl-Ppant	PSR	4 Carbon Acyl-Ppant	0	http://rayluolab.org/pff-files-for-butyryl-ppant/
Hexanoyl-Ppant	SXH	6 Carbon Acyl-Ppant	0	http://rayluolab.org/pff-files-for-hexanoyl-ppant/

Octanoyl-Ppant	SXO	8 Carbon Acyl-Ppant	0	http://rayluolab.org/pff-files-for-octanoyl-ppant/
Decanoyl-Ppant	PM8	10 Carbon Acyl-Ppant	0	http://rayluolab.org/pff-files-for-decanoyl-ppant/
Dodecanoyl-Ppant	8Q1	12 Carbon Acyl-Ppant	7	http://rayluolab.org/pff-files-for-dodecanoyl-ppant/
Tetradecanoyl-Ppant	ZMP	14 Carbon Acyl-Ppant	25	http://rayluolab.org/pff-files-for-tetradecanoyl-ppant/
Hexadecanoyl-Ppant	G9S	16 Carbon Acyl-Ppant	0	http://rayluolab.org/pff-files-for-hexadecanoyl-ppant/
Phosphopantetheinyl-Serine (Ppant-Ser) Library				
Ppant	PNS	"apo" phosphopantetheinyl-serine	48	http://rayluolab.org/pff-files-for-phosphopantetheinyl-serine/

Acetyl-Ppant	6VG	2 Carbon Acyl-Ppant	1	http://rayluolab.org/pff-files-for-acetyl-ppant-ser/
Butyryl-Ppant	PSR	4 Carbon Acyl-Ppant	2	http://rayluolab.org/pff-files-for-butyryl-ppant-ser/
Hexanoyl-Ppant	SXH	6 Carbon Acyl-Ppant	2	http://rayluolab.org/pff-files-for-hexanoyl-ppant-ser/
Octanoyl-Ppant	SXO	8 Carbon Acyl-Ppant	2	http://rayluolab.org/pff-files-for-octanoyl-ppant-ser/
Decanoyl-Ppant	PM8	10 Carbon Acyl-Ppant	5	http://rayluolab.org/pff-files-for-decanoyl-ppant-ser/
Dodecanoyl-Ppant	8Q1	12 Carbon Acyl-Ppant	9	http://rayluolab.org/pff-files-for-dodecanoyl-ppant-ser/

Tetradecanoyl-Ppant	ZMP	14 Carbon Acyl-Ppant	39	http://rayluolab.org/pff-files-for-tetradecanoyl-ppant-ser/
Hexadecanoyl-Ppant	G9S	16 Carbon Acyl-Ppant	1	http://rayluolab.org/pff-files-for-hexadecanoyl-ppant-ser/

The functional form of a typical force field includes terms responsible for bond stretching, angle bending, dihedral angle torsion, van der Waals, and electrostatic interactions. For example, the additive Amber force field functional form for the total potential energy (E_{total}) is:

$$E_{total} = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} V_n(1 + \cos(n\phi - \gamma))$$

$$+ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right)$$

In this equation, ε is the dielectric constant, which has a default value 1 in Amber and thus can be omitted. A parameter set including the following parameters has to be provided to perform tasks such as minimization and molecular dynamics simulations:

- Bond Parameters: k_b, r_0
- Angle Parameters: k_θ, θ_0
- Torsional Angle Parameters: V_n, γ

- Van der Waals Parameters: A_{ij}, B_{ij}
- Charge Parameters: q_i, q_j

For Ppant-Ser PCLs, both covalent parameters (bond, angle, and torsional angle) and noncovalent van der Waals parameters were first derived from ff14SB where possible to ensure compatibility with parameters for standard amino acid residues,[23] missing parameters were then obtained from the gaff2 force field, which were designed for general organic molecules.[35] For standalone CoA and Ppant PCLs, these parameters were directly derived from the gaff2 force field. Charge parameters have to be treated separately, since individual partial charge has to be assigned to each atom for widely used point-charge electrostatic models. In the PFF library, three common charging algorithms were applied, including Gasteiger,[29] AM1-BCC[30-31] and RESP.[32-33]

The RESP charges depend on molecular geometries provided as input. However, large, flexible molecules tend to form intramolecular interactions such as hydrogen bonds during the geometry optimization step, introducing a bias in fitted charges. Moreover, the CPU time of geometry optimization is positively correlated with molecular sizes. Therefore, a "plug-and-play" fragmentation approach was employed serving as a consistent charging scheme for the PFF library development, which splits common substructures of PCLs: CoA, Ppant and Ppant-Ser, into a fragment pool including 8 components: (1) methylphosphate, (2) adenosine, (3) dimethyldiphosphate, (4) pantoic acid, (5) beta-alanine, (6) cysteamine, (7) serine dipeptide and (8) dimethylphosphate, as shown in **Figure 3.1**. Fragments were capped with acetyl, methylamide, methyl, and/or hydroxyl groups mimicking the natural chemical environments of the fragments, and these caps were constrained to 0 net charge

and removed during the fragment merging process. This approach was deemed necessary due primarily to the flexibility and relatively large size of the pantetheine moiety itself. Indeed, it is common for primed CoA and Ppant-Ser thioesters to achieve sizes greater than 200 atoms.[58] During the geometry optimization step, extra care was taken for the serine dipeptide fragment (fragment 7), where ϕ and ψ angles were constrained at $-60.70°$ and $-31.32°$ respectively, according to the analysis of ϕ and ψ angle distributions of 320 Ppant-Ser conformations from Protein Data Bank. (**Figure S3.1**) In contrast, Gasteiger charges and AM1-BCC charges were obtained with the whole molecule strategy, i.e., the structural files of the intact molecule of each PCL were used as inputs, because of the much higher efficiency of the two charging algorithms than that of the RESP charging method.

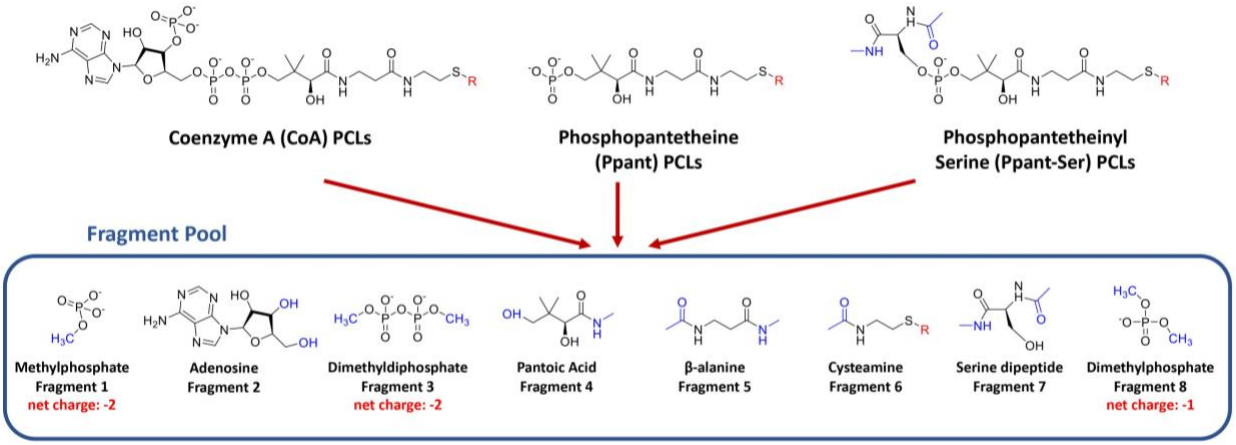


Figure 3.1. "Plug-and-play" fragmentation strategy of PFF library development. Coenzyme A (CoA) PCLs, phosphopantetheine (Ppant) PCLs and phosphopantetheinyl-serine (Ppant-Ser) PCLs can be fragmented into a fragment pool consisting of 8 components: (1) methylphosphate, (2) adenosine, (3) dimethyldiphosphate, (4) pantoic acid, (5) beta-alanine, (6) cysteamine, (7) serine dipeptide and (8) dimethylphosphate. CoA PCLs can be

reconstructed with fragments 1, 2, 3, 4, 5 and 6; Ppant PCLs can be reconstructed with fragments 1, 4, 5 and 6; Ppant-Ser PCLs can be reconstructed with fragments 4, 5, 6, 7 and 8. Various extending units that form thioester bonds with CoA, Ppant or Ppant-Ser are labeled with "R" in red. Acetyl, methylamide, methyl, and hydroxyl caps that were constrained to 0 net charge and removed during the fragment merging process are depicted in blue.

A caveat during the PFF library development is inconsistent atomic nomenclature of common substructures between different PCLs on PDB. For example, the amine nitrogen atom of adenine of coenzyme A (PDB ID: COA), malonyl CoA (PDB ID: MLC) and propionyl CoA (PDB ID: 1VU) are named as N6A, N6, and N4, respectively. The nomenclature inconsistency prevents parameter transferability that is necessary for our fragmentation strategy. To address this problem, an atom renaming program called *PyRenamer* written in Python that enables converting atom names to corresponding atom names of reference molecule was developed. The source code of *PyRenamer* can be obtained by contacting the authors.

3.3.2. Structural Comparisons of QM and PFF Optimized Fragments

To validate PFF parameters, fragments were sequentially minimized with increasing level of QM theories in the order of B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ as benchmark. An acetyl cysteamine fragment was also tested as a representative thioester extending unit. Due to the increasing computational time complexity for larger fragments, the highest level of theory used for fragment 2 (adenosine) was B3LYP/6-

311+G(2d,p). For fragments 3 (dimethyldiphosphate), 4 (pantoic acid), 5 (beta-alanine) and acetyl cysteamine fragment, the highest level of theory used was MP2/aug-cc-pVDZ. MP2/aug-cc-pVTZ were only applied to smaller fragments including fragments 1 (methylphosphate), 6 (cysteamine), and 8 (dimethylphosphate). The RMSD between QM and PFF optimized fragments ranged from 0.095 Å to 0.465 Å when RESP charges were used (denoted as PFF/RESP below). (**Table 3.2**, **Figure S3.2**) In particular, since the structure of fragment 2 (adenosine) matches the adenosine (entry name: AN) available in the OL3 force field, the QM- and *OL3*-optimized fragments were also compared.[59] The RMSD comparison shows that PFF/RESP (0.327 Å) has higher accuracy than OL3 (0.550 Å) for adenosine. (**Figure S3.3**) Additionally, PFF with Gasteiger (PFF/Gasteiger) and AM1-BCC (PFF/AM1-BCC) charges were also validated similarly. The RMSD between QM and PFF/Gasteiger optimized fragments ranged from 0.102 Å to 0.519 Å, and for PFF/AM1-BCC optimized fragments, the RMSD ranged from 0.097 Å to 0.509 Å. (**Table S3.1**) Overall, PFF parameters with all three charging methods perform similarly in reproducing QM-optimized structures for the tested fragments.

Table 3.2. RMSD Between QM and PFF/RESP Optimized Fragments

Fragment No.	Fragment Name	Highest Level of Theory	RMSD
1	methylphosphate	MP2/aug-cc-pVTZ	0.095
2	adenosine	B3LYP/6-311+G(2d,p)	0.327

3	dimethyldiphosphate	MP2/aug-cc-pVDZ	0.465
4	pantoic acid	MP2/aug-cc-pVDZ	0.281
5	beta-alanine	MP2/aug-cc-pVDZ	0.386
6	cysteamine	MP2/aug-cc-pVTZ	0.098
8	dimethylphosphate	MP2/aug-cc-pVTZ	0.113
-	acetyl-cysteamine	MP2/aug-cc-pVDZ	0.401
Average			0.271

3.3.3. Normal Mode Analysis of QM and PFF Optimized Fragments

In order to gain further insights of the quality of PFF parameters, the QM normal mode frequencies of each fragment were obtained with the same level of theories described above. Due to the fact that *ab initio* calculated harmonic vibrational frequencies are typically larger than the experimental vibrational frequencies,[60] scaling factors were applied to QM calculated normal mode frequencies. Normal modes plots agreed well between QM calculations and PFF calculations, except for modes in the high frequency (above 2000 cm^{-1}) region. (**Figure 3.2** and **Figure S3.4-S3.5)**. For example, the frequencies observed in the 450-1100 cm^{-1} range include C-O and O-P bond-stretching, O-P-O twisting, O-P-O wagging, and O-P-O scissoring. S-C bond stretching was observed at 645 and 749 cm^{-1}, O-C-S scissoring observed at 439 cm^{-1}, and the characteristic intense carbonyl stretch for thioesters at 1720

cm^{-1} at the MP2/aug-cc-pVDZ level of theory. The PFF frequencies were in good agreement with QM frequencies for both cases.

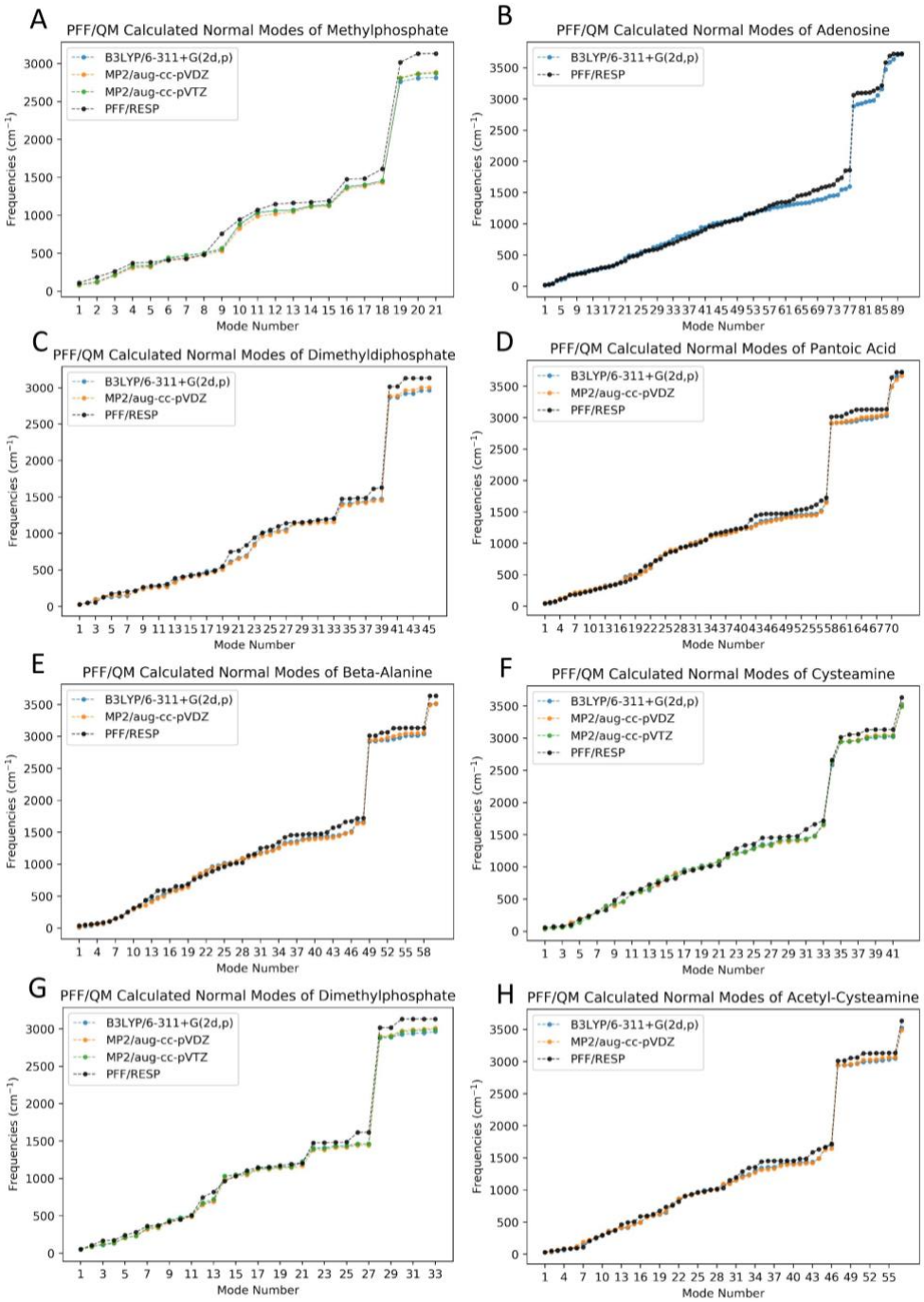


Figure 3.2. Comparison of normal mode frequencies of fragments calculated with PFF/RESP and B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels of theories. Scaling factors of 0.967, 0.959 and 0.953 were applied to B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ calculated normal mode frequencies, respectively.

3.3.4. Partial Charge Comparisons Between Three Charge Fitting Methods

Since the accuracy of PFF parameters for individual fragments has been validated, a four-step RESP fitting strategy was employed to derive final RESP partial charges as stated in **section 3.2**. **Figure 3.3** shows the atom names and partial charges derived by RESP (fragmentation strategy), Gasteiger (whole molecule strategy) and AM1-BCC methods (whole molecule strategy), including their deviations from the unconstrained fragmental partial charges (the "differences" column) for standalone phosphopantetheine (PDB ID: PNS).

It can be observed that the greatest deviations are from charges derived by Gasteiger method, where 17 atoms having differences above 0.15, including O27, P24, O23, O24, O25, C29, O33, H33, C34, O35, N36, H36, C39, O40, N41, H4, S44. This is due to the fact that Gasteiger charges are not derived to reproduce electrostatic potentials (ESP) as the other two methods do.[29] In contrast, ESP based AM1-BCC[30-31] and RESP[32-33] charging methods produced only 7 or 1 atomic partial charges with differences above 0.15, respectively. It is reasonable to set 0.15 partial charge deviation as the "red line", as indicated by LIPID11 force field development involving similar fragmentation approach.[34] Therefore, it is expected that

AM1-BCC and RESP charges perform better than Gasteiger charges in subsequent validation tests.

	Atom No.	Atom Name	No Constraints	RESP	Differences	Gasteiger	Differences	AM1-BCC	Differences
Fragment 1	1	O27	-0.5529	-0.6569	0.104	-0.3539	-0.199	-0.5602	0.0073
	2	P24	1.2291	1.1893	0.0398	0.0598	1.1693	1.3409	-0.1118
	3	O23	-0.9098	-0.8862	-0.0236	-0.6341	-0.2757	-0.9212	0.0114
	4	O25	-0.9098	-0.8862	-0.0236	-0.6341	-0.2757	-0.9212	0.0114
	5	O26	-0.9098	-0.8862	-0.0236	-0.6341	-0.2757	-0.9212	0.0114
Fragment 4	6	C29	0.3762	0.3327	0.0435	0.0233	0.3529	-0.111	0.4872
	7	C28	0.08	0.0052	0.0748	0.0596	0.0204	0.2274	-0.1474
	8	H281	0.0315	0.0527	-0.0212	0.0578	-0.0263	-0.0038	0.0353
	9	H282	0.0315	0.0527	-0.0212	0.0578	-0.0263	-0.0038	0.0353
	10	C30	-0.084	-0.1458	0.0618	-0.0547	-0.0293	-0.0916	0.0076
	11	H301	0.0045	0.0272	-0.0227	0.0238	-0.0193	0.0342	-0.0297
	12	H302	0.0045	0.0272	-0.0227	0.0238	-0.0193	0.0342	-0.0297
	13	H303	0.0045	0.0272	-0.0227	0.0238	-0.0193	0.0342	-0.0297
	14	C31	-0.084	-0.1458	0.0618	-0.0547	-0.0293	-0.0916	0.0076
	15	H311	0.0045	0.0272	-0.0227	0.0238	-0.0193	0.0342	-0.0297
	16	H312	0.0045	0.0272	-0.0227	0.0238	-0.0193	0.0342	-0.0297
	17	H313	0.0045	0.0272	-0.0227	0.0238	-0.0193	0.0342	-0.0297
	18	C32	0.0975	0.0714	0.0261	0.1375	-0.04	0.0941	0.0034
	19	H32	0.0465	0.0531	-0.0066	0.0709	-0.0244	0.0637	-0.0172
	20	O33	-0.6238	-0.6058	-0.018	-0.3828	-0.241	-0.6168	-0.007
	21	H33	0.4	0.3887	0.0113	0.2112	0.1888	0.419	-0.019
	22	C34	0.4392	0.4929	-0.0537	0.2411	0.1981	0.6211	-0.1819
	23	O35	-0.5513	-0.5498	-0.0015	-0.2747	-0.2766	-0.6971	0.1458
Fragment 5	24	N36	-0.5881	-0.5145	-0.0736	-0.3127	-0.2754	-0.5399	-0.0482
	25	H36	0.3476	0.3265	0.0211	0.1494	0.1982	0.3995	-0.0519
	26	C37	-0.0224	-0.0136	-0.0088	0.0196	-0.042	0.113	-0.1354
	27	H371	0.0933	0.0836	0.0097	0.0472	0.0461	0.0222	0.0711
	28	H372	0.0933	0.0836	0.0097	0.0472	0.0461	0.0222	0.0711
	29	C38	0.0023	-0.052	0.0543	0.0399	-0.0376	-0.1754	0.1777
	30	H381	0.0296	0.0407	-0.0111	0.0381	-0.0085	0.0952	-0.0656
	31	H382	0.0296	0.0407	-0.0111	0.0381	-0.0085	0.0952	-0.0656
	32	C39	0.4988	0.542	-0.0432	0.2132	0.2856	0.6571	-0.1583
	33	O40	-0.5487	-0.5485	-0.0002	-0.2777	-0.271	-0.7321	0.1834
Fragment 6	34	N41	-0.7086	-0.5471	-0.1615	-0.3146	-0.394	-0.5359	-0.1727
	35	H41	0.3729	0.3296	0.0433	0.1494	0.2235	0.4115	-0.0386
	36	C42	-0.0438	0.0165	-0.0603	0.0198	-0.0636	0.111	-0.1548
	37	H421	0.1501	0.1156	0.0345	0.0475	0.1026	0.0312	0.1189
	38	H422	0.1501	0.1156	0.0345	0.0475	0.1026	0.0312	0.1189
	39	C43	-0.1023	-0.1079	0.0056	0.0058	-0.1081	-0.0103	-0.092
	40	H431	0.1118	0.1063	0.0055	0.0394	0.0724	0.1287	-0.0169
	41	H432	0.1118	0.1063	0.0055	0.0394	0.0724	0.1287	-0.0169
	42	S44	-0.3742	-0.362	-0.0122	-0.1777	-0.1965	-0.4579	0.0837
	43	H44	0.2077	0.1994	0.0083	0.1023	0.1054	0.1728	0.0349
Total			-2.0561	-2	-0.0561	-2.0002	-0.0559	-2.0001	-0.056

Figure 3.3. Comparison of RESP charges, Gasteiger charges, AM1-BCC charges with unconstrained fragmental partial charges for standalone phosphopantetheine. The "differences" column associated with each charging method shows the differences with

unconstrained partial charges of corresponding atoms (the "No Constraints" column). Color schemes were applied to "differences" columns, where blue indicates negative differences and red indicates positive differences. The bottom row shows the sum of corresponding columns.

3.3.5. Parameter Validations in MD Simulations

Three representative systems containing PCLs with available experimental structures were used for validation purposes: phosphopantetheine adenylyltransferase-phosphopantetheine (PPAT-Ppant, PDB ID: 1OD6),[46] 3-hydroxy-3-methylglutaryl synthase/acyl carrier protein complex (HGMS/ACP-Ppant-Ser, PDB ID: 5KP7),[47] and diaminobutyrate acetyltransferase-Coenzyme A (EctA-CoA, PDB ID: 6SK1).[48] It is notable that (1) Ppant is the substrate of PPAT in PPAT-Ppant; (2) Ppant-Ser is covalently linked to ACP as a prosthetic group in HGMS/ACP-Ppant-Ser; (3) CoA is the cofactor of EctA in EctA-CoA. Since covalent bonds are typically stronger than non-covalent interactions, and cofactors typically remain bound with proteins, it is reasonable to expect that their binding strengths increase in the order of PPAT-Ppant, EctA-CoA and HGMS/ACP-Ppant-Ser.[61] Each system was simulated under reported experimental temperatures and salt concentrations.

RMSD of simulation trajectories to the crystal structure is considered as an important validation metric of the quality of a force field, since it is reasonable to assume that protein crystal structures are typically close to the structures at the physiological condition.[62] Therefore, the RMSD's relative to crystal structures were computed for heavy atoms of both PCLs and protein residues in contact with PCLs (Data not shown), and the probability density

functions estimated were also analyzed via the Gaussian kernel density estimation (KDE), as shown in **Figure 3.4**. For contact residues, the parameter sets of all three charging methods show similar RMSD distribution patterns, as illustrated in the left panel of **Figure 3.4**. For Ppant-Ser and CoA PCLs, PFF/AM1-BCC- parameter set gave significantly lower RMSD than the other two charging methods. (**Figure 3.4D, F**) However, for Ppant PCL, PFF/AM1-BCC and PFF/RESP parameter sets lead to higher RMSD than PFF/Gasteiger parameter set, and the highest RMSD value observed reaches 7 Å. (**Figure 3.4B**) Nevertheless, the PFF/AM1-BCC parameter set is the best in capturing the expected trend that PPAT-Ppant, EctA-CoA, and HGMS/ACP-Ppant-Ser are in the increasing order of binding strengths.

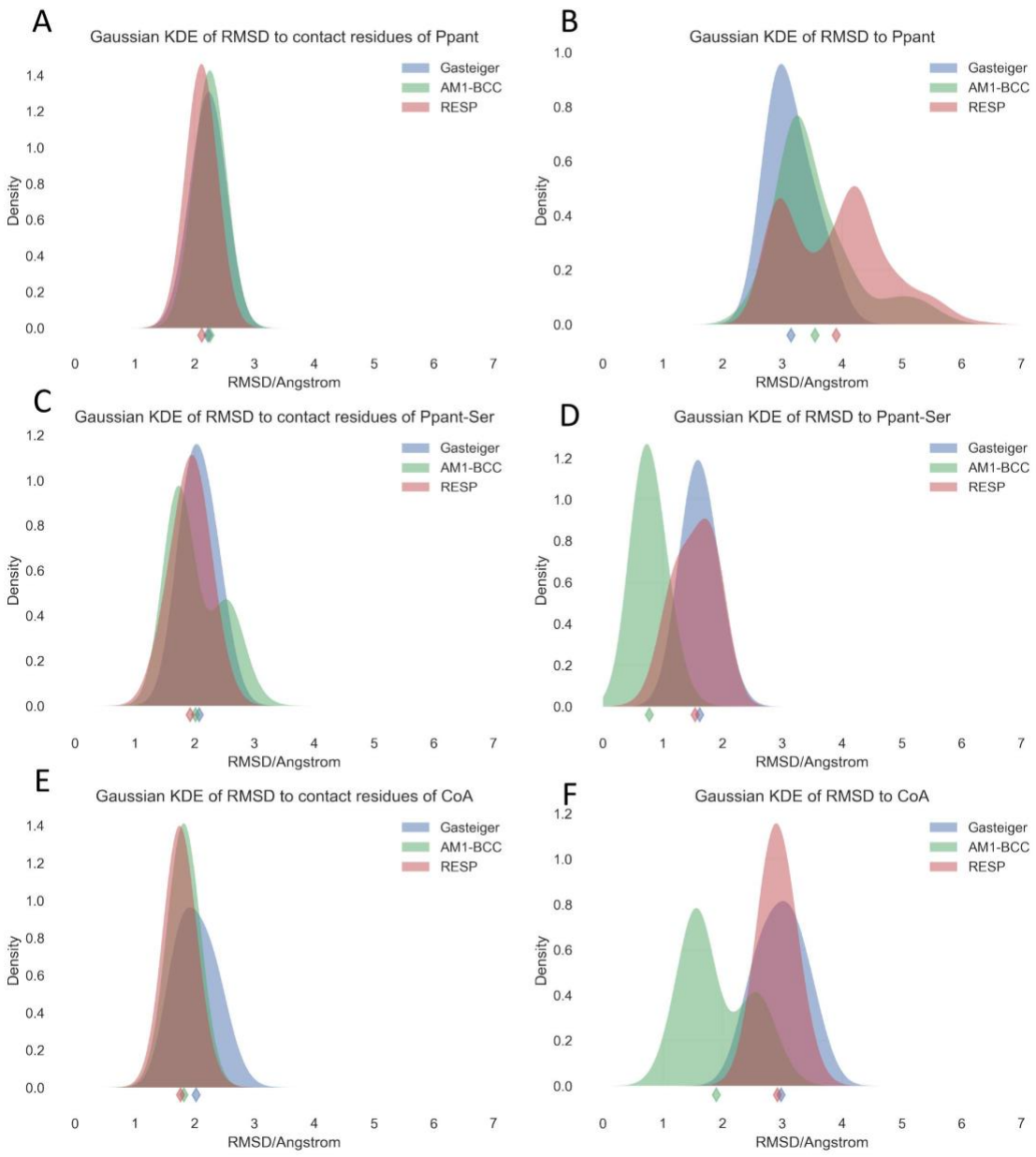


Figure 3.4. Gaussian kernel density estimates (KDEs) of computed RMSD values of heavy atoms of contact residues (left panel) and PCLs (right panel) relative to the experimental structures. The diamond markers indicate the mean RMSD values.

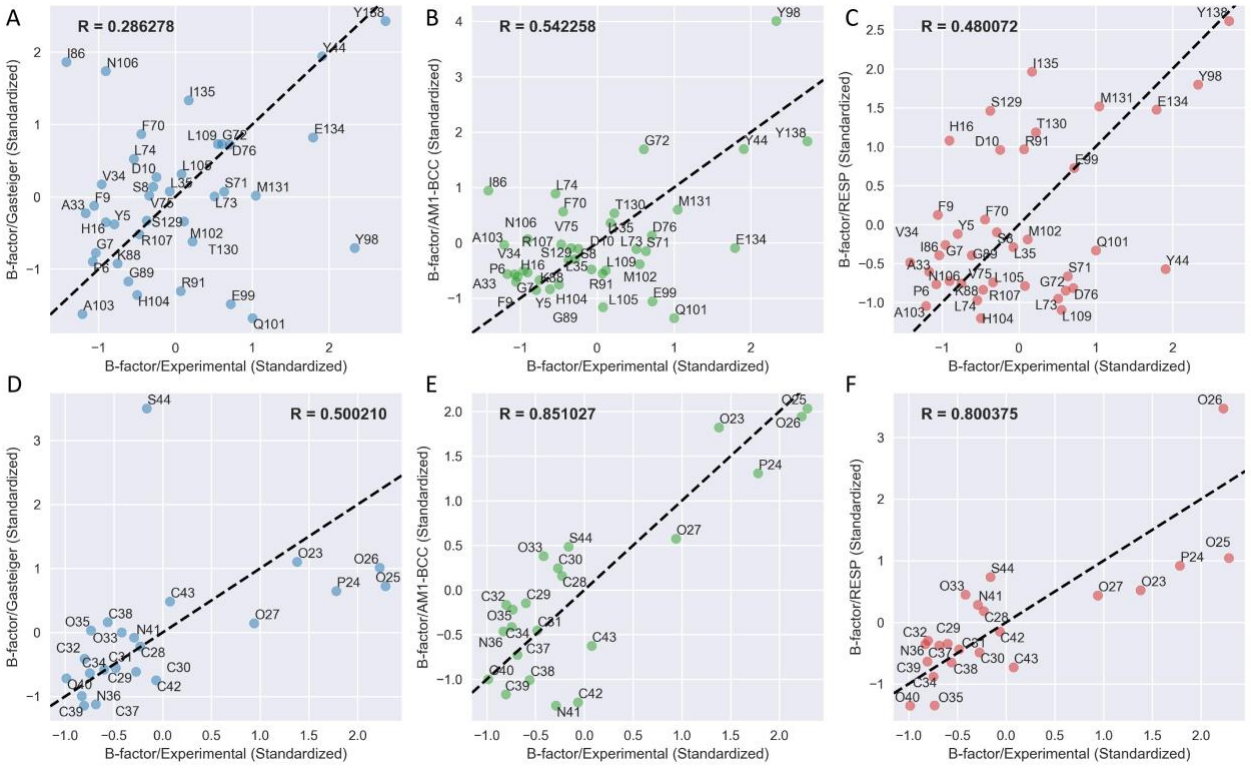


Figure 3.5. Correlation analysis of standardized simulated and experimental B-factors of the contact residues (upper panel) and the PCLs (lower panel) for the PPAT-Ppant system. The residue names of contact residues or atom names of Ppant are annotated. R is the Pearson correlation coefficient.

The second quantitative validation of PFF parameter set is the comparison of experimental and simulated B-factor, or temperature factor, reflecting the mobility or flexibility of various parts of the molecule caused by thermal motion. High B-factors indicate greater uncertainty about the actual atom position. **Figure 3.5** displays the scatterplots of standardized B-factors simulated from three charging methods compared with experimental B-factors of the ligands and contact residues of PPAT-Ppant system. PFF/AM1-BCC set

resulted in highest correlation coefficients for both Ppant and contact residues, and next comes PFF/RESP and PFF/Gasteiger sets. The visualization of standardized experimental B-factor and simulated B-factors of PPAT-Ppant system illustrating the average structures of the last 10 ns are shown in **Figure 3.6**. PFF/AM1-BCC and PFF/RESP simulations yield similar agreement in Ppant conformations with respect to experimental structures, (**Figure 3.6C, D**) although all three charging methods resulted in similar B-factor patterns for Ppant with the two ends of the linear structure having higher flexibility than the middle region. The corresponding scatterplots and structural visualizations of HGMS/ACP-Ppant-Ser and EctA-CoA systems are shown in **Figures S3.6-S3.9**. Highest correlation coefficients with experimental B-factors are always observed in PFF/AM1-BCC simulations. However, the simulated B-factors of CoA with all three charging methods failed to capture the trend that the phosphate group has higher flexibility than the rest of the molecules. (**Figures S3.8, S3.9**) A closer look into the X-ray structures of EctA-CoA revealed the existence of the cation-pi interaction between the adenine ring and Arg 99, which is difficult to be modeled in current non-polarizable Amber force field[63] but is actively investigated in on-going Amber polarizable force field developments.[64-69]

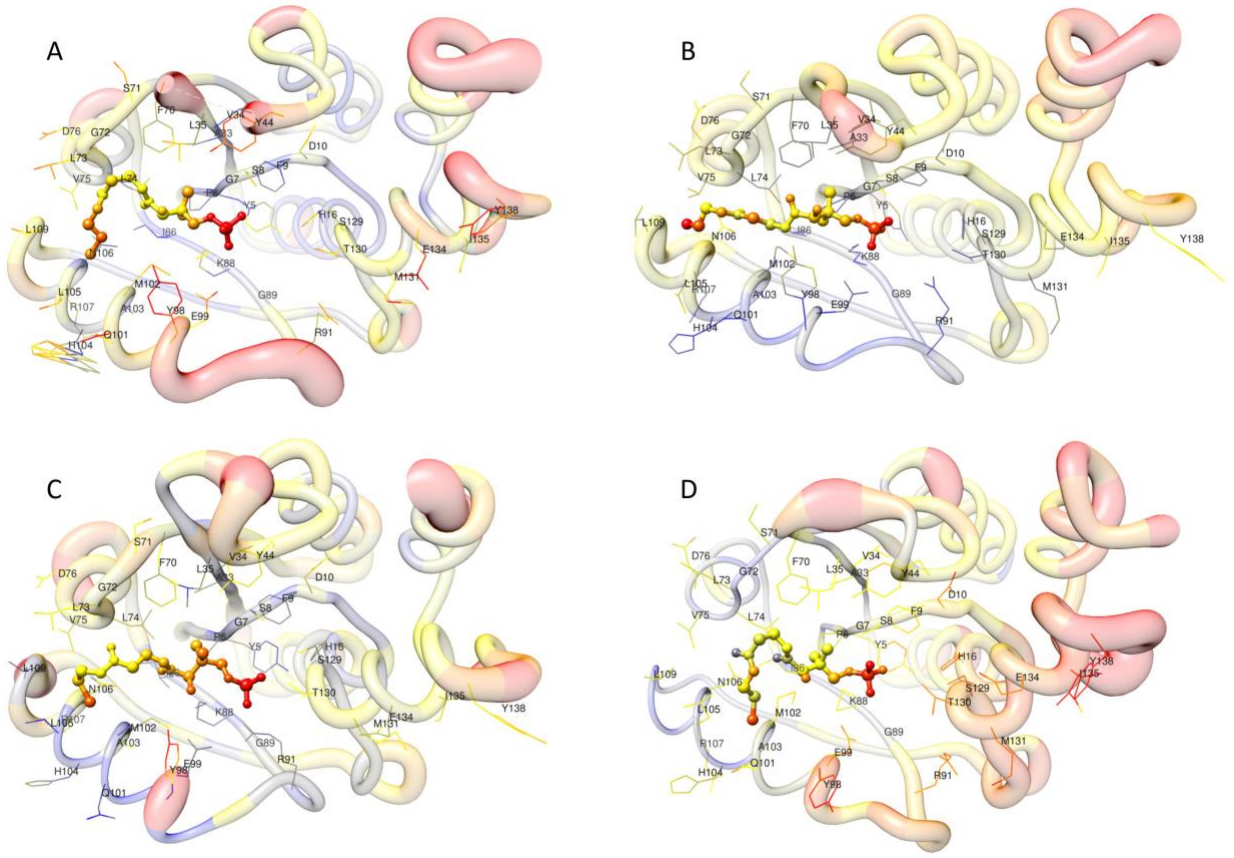


Figure 3.6. Visual comparison of standardized simulated and experimental B-factors for the PPAT-Ppant system. Ppant is depicted as ball-and-stick; contact residues are depicted as wire. Colors (Red color indicates high B-factors, and blue color indicates low B-factors) and thickness of protein backbone also indicate B-factor values. **A.** Experimental structure. **B.** The average structure of the last 10 ns trajectory with Gasteiger charges. **C.** The average structure of the last 10 ns trajectory with AM1-BCC charges. **D.** The average structure of the last 10 ns trajectory with RESP charges.

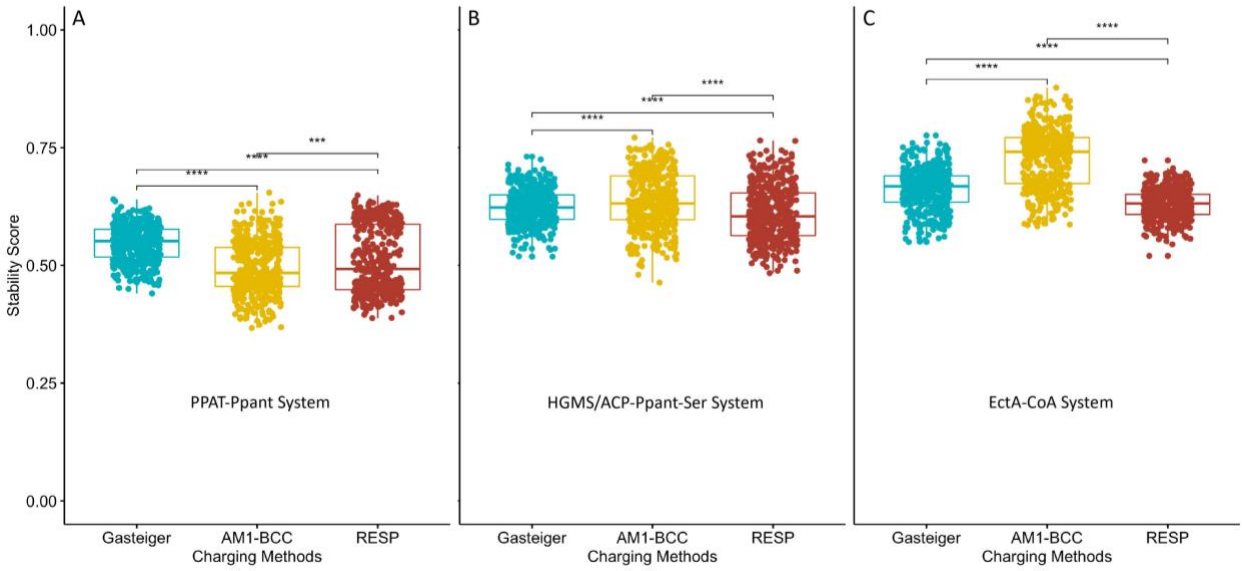


Figure 3.7. T-test of binding stability scores of the last 50 ns trajectories of **A.** PPAT-Ppant, **B.** HGMS/ACP-Ppant-Ser, and **C.** EctA-CoA. Significance levels: ***, p <= 0.001; ****, p <= 0.0001.

The last quantitative validation for MD simulations is our previously defined binding stability scoring, which reflects the binding stability between two molecules (ligands and proteins for example) by counting the native atomic contacts between the two molecules in each trajectory snapshots.[56] Higher stability score indicates stronger binding. The last 50 ns of each trajectory was used for t-test analysis for stability scores. (Data not shown) Consistent with previous expectations, the stability scores in PFF/AM1-BCC simulations are the highest among the three for HGMS/ACP-Ppant-Ser and EctA-CoA, while the lowest for PPAT-Ppant, reflecting the nature of their expected binding strengths. (**Figure 3.7**)

3.3.6. Pantetheine Force Field (PFF) Library Website Interface

A website hosting the pantetheine force field library (http://rayluolab.org/pff-library/) has been developed. Published on the website are three libraries of force fields for CoA PCLs, Ppant PCLs, and Ppant-Ser PCLs. For each PCL, an OFF library (lib) file with all structures and charges, and one or two parameter modification (frcmod) file with all missing non-charge parameters for each charging method are present. OFF library files contain the same atom names and coordinates as present in the Protein Data Bank for compatibility. Only one frcmod file is provided for CoA or Ppant PCL, since they are derived from only gaff2 force field; while two frcmod files are present for Ppant-Ser PCL, due to the fact that non-charge parameters of Ppant-Ser PCLs are first derived from the ff14SB force field, then from the gaff2 force field. Users of PFF files for Ppant-Ser PCLs are expected to load gaff2 frcmod files first, then ff14SB frcmod files to overwrite overlapping parameters. In addition, tutorials are present on the website to provide detailed protocols and input files on how to model and setup simulations containing PCLs. These structures can be used for minimizations, MD simulations, or as part of docking studies.

3.4. Conclusions

In this chapter, we present the first Amber-compatible force field library for various pantetheine containing ligands. The PFF library was parameterized using Gasteiger, AM1-BCC, or RESP charging method in combination with gaff2 parameters. Among three commonly used charging schemes, PFF/AM1-BCC parameter set shows better MD simulation performance than PFF/Gasteiger and PFF/RESP parameter sets, as indicated by

MD validations. Furthermore, a "plug-and-play" fragmentation strategy was designed to enable systematic charge fitting for large molecules sharing common substructures. However, the parameter sets with the RESP charges derived from the fragmentation strategy does not perform better than that with the AM1-BCC charging method that can be applied to whole molecules in terms of MD simulations. In fact, the "plug-and-play" strategy applied in this study generating a fragment pool with extremely small fragments ranging from 9 atoms (methylphosphate) to 32 atoms (adenosine) has the following disadvantages: First, the increased amount of manual work overshadows the benefits of cheaper computational expenses during parametrization. Second, many charging errors were introduced due to the existence of too many merging interfaces between adjacent fragments. Therefore, a natural improvement of the "plug-and-play" strategy is to employ larger fragments. In subsequent version of the PFF library, larger fragments will be explored to reduce errors in the RESP charging method.

This chapter paves the foundation for easy setup of MD simulations of biological systems containing PCLs *in silico*, and it is hoped to be applied in applications such as protein engineering for the production of novel compounds, or drug discovery for targeting certain PCL-containing proteins that play critical roles in diseases.

3.5. Supporting Information

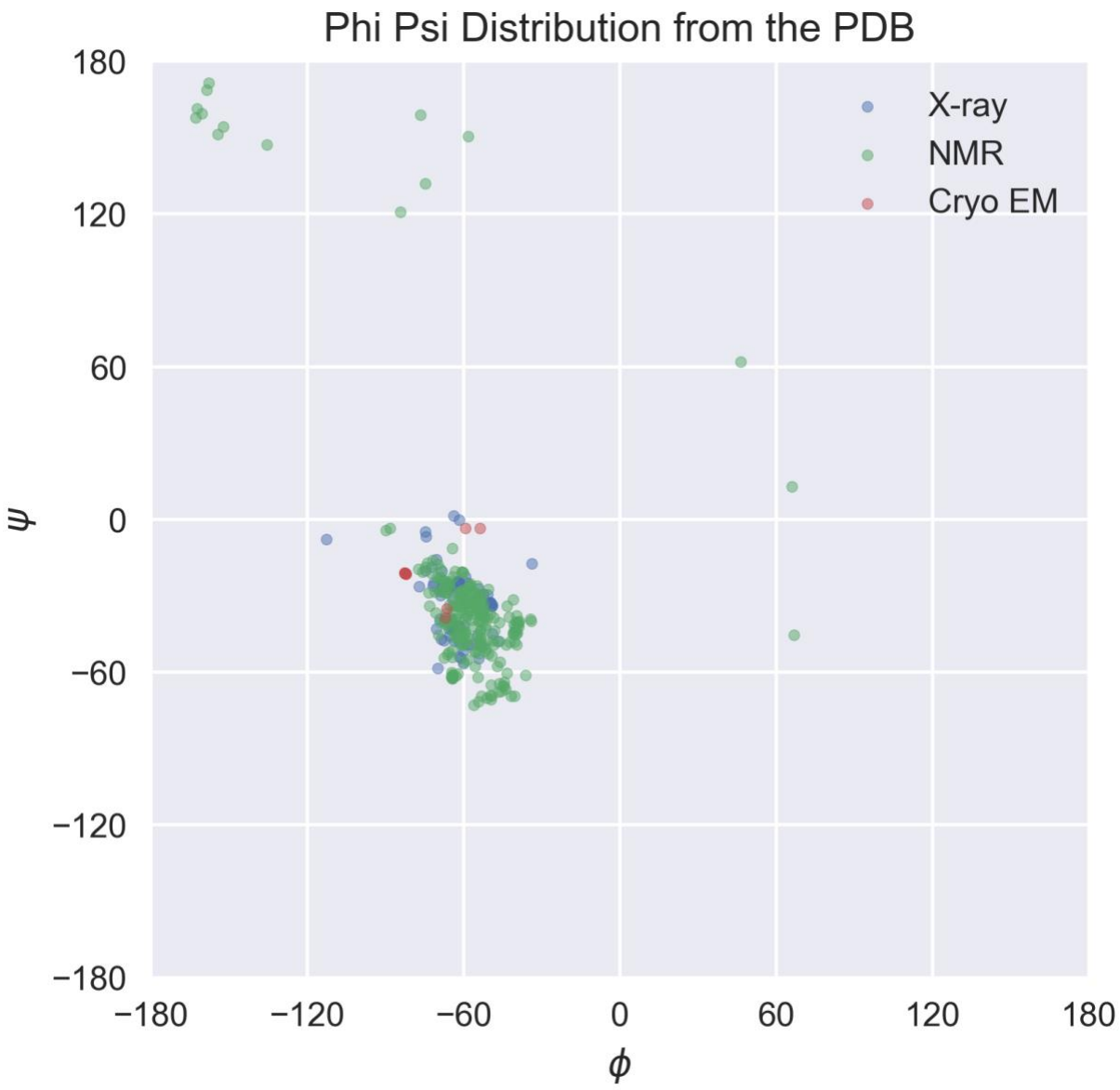


Figure S3.1. Φ/Ψ values of covalently bound phosphopantetheinyl-serine (Ppant-Ser) from the protein data bank (PDB ID: PNS) for a total of 320 data points (240 from NMR structures, 70 from X-ray crystal structures, and 10 from cryo-EM structures). The representative Φ/Ψ angles selected that is closest to the average values are $-60.70°$ and $-31.32°$ for Φ and Ψ angles, respectively.

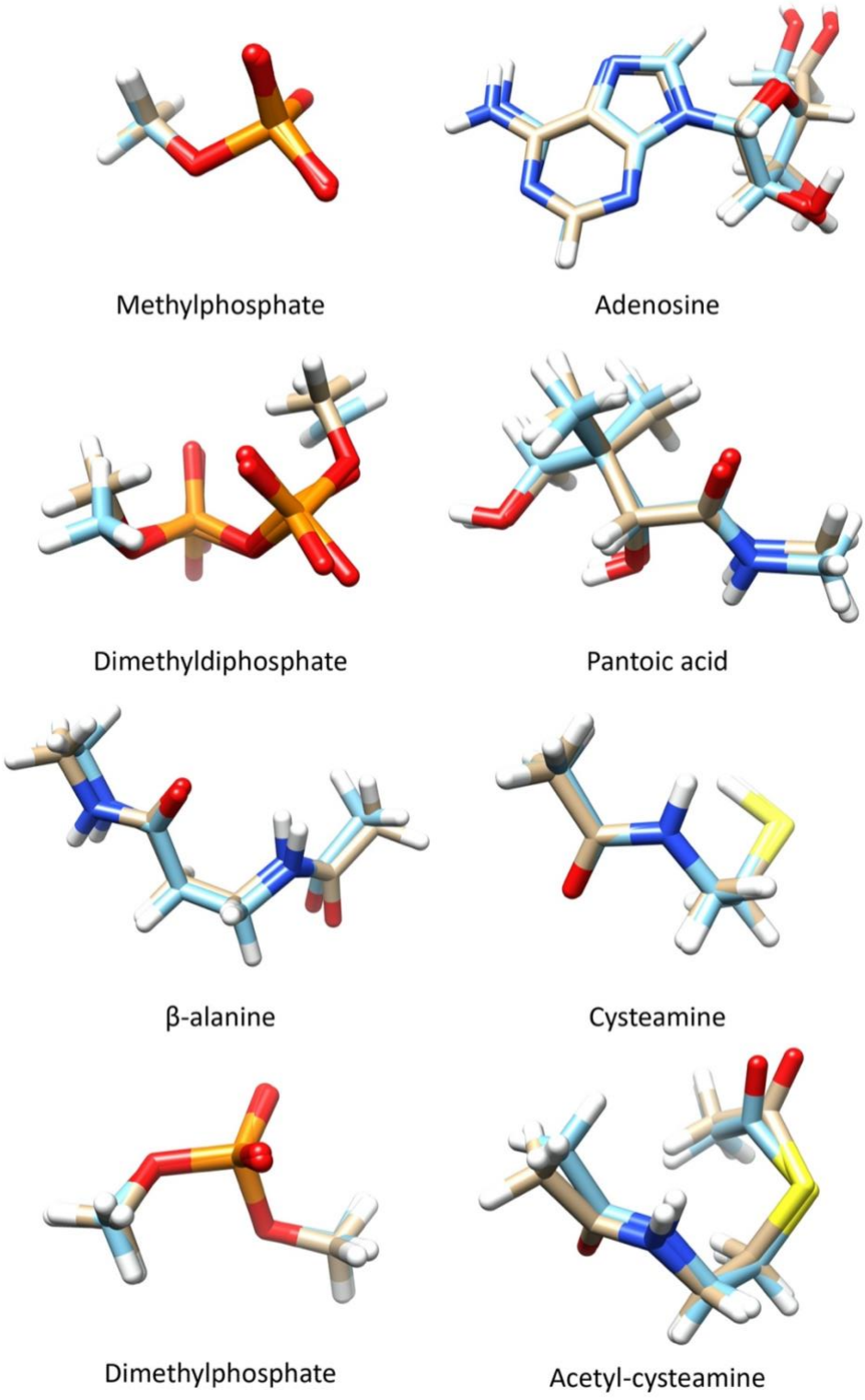


Figure S3.2. Comparison of QM and MM minimized structures of "plug and play" fragments. The carbons in QM minimized structures are depicted in brown, and those in PFF minimized

structures are depicted in cyan. MM minimization was conducted with the RESP-charged PFF. The highest level of QM theory used for each fragment and the respective RMSD values can be found in **Table S3.1**.

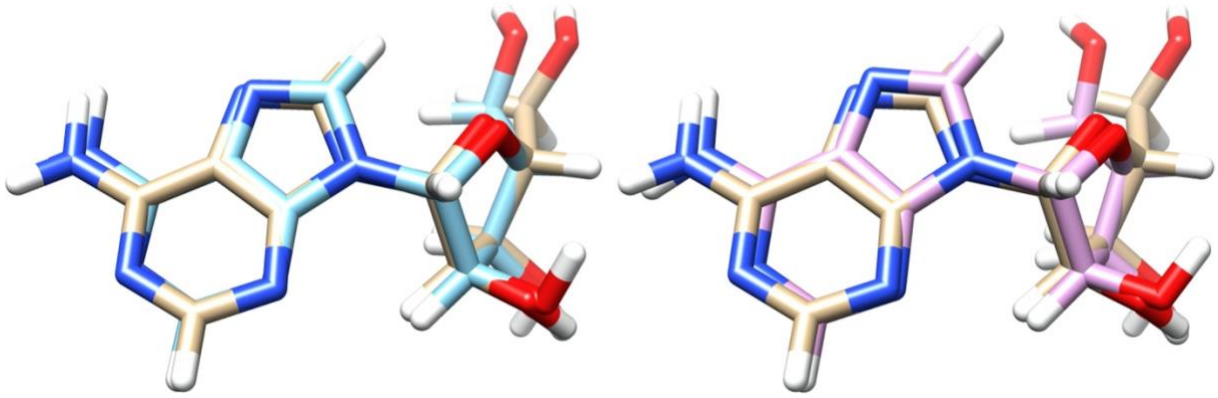


Figure S3.3. Comparison of adenosine structures minimized with PFF parameter sets, OL3 parameter sets, and B3LYP/6-311+G(2d,p) level of theory. The left figure shows the structural alignment of PFF minimized structure (cyan) and B3LYP minimized structure (brown), and the RMSD value is 0.327 Angstrom; The right figure shows the structural alignment of OL3 minimized structure (pink) and B3LYP minimized structure (brown), and the RMSD value is 0.550 Angstrom.

Table S3.1. RMSD (Angstrom) between QM and PFF/Gasteiger or PFF/AM1-BCC optimized "plug and play" fragments

Fragment No.	Fragment Name	Highest Level of Theory	PFF/Gasteiger	PFF/AM1-BCC
1	methylphosphate	MP2/aug-cc-pVTZ	0.102	0.097
2	adenosine	B3LYP/6-311+G(2d,p)	0.176	0.302
3	dimethyldiphosphate	MP2/aug-cc-pVDZ	0.385	0.509
4	pantoic acid	MP2/aug-cc-pVDZ	0.519	0.265
5	beta-alanine	MP2/aug-cc-pVDZ	0.42	0.389
6	cysteamine	MP2/aug-cc-pVTZ	0.138	0.107
7	dimethylphosphate	MP2/aug-cc-pVTZ	0.119	0.108
-	acetyl-cysteamine	MP2/aug-cc-pVDZ	0.247	0.366
Average			0.263	0.268

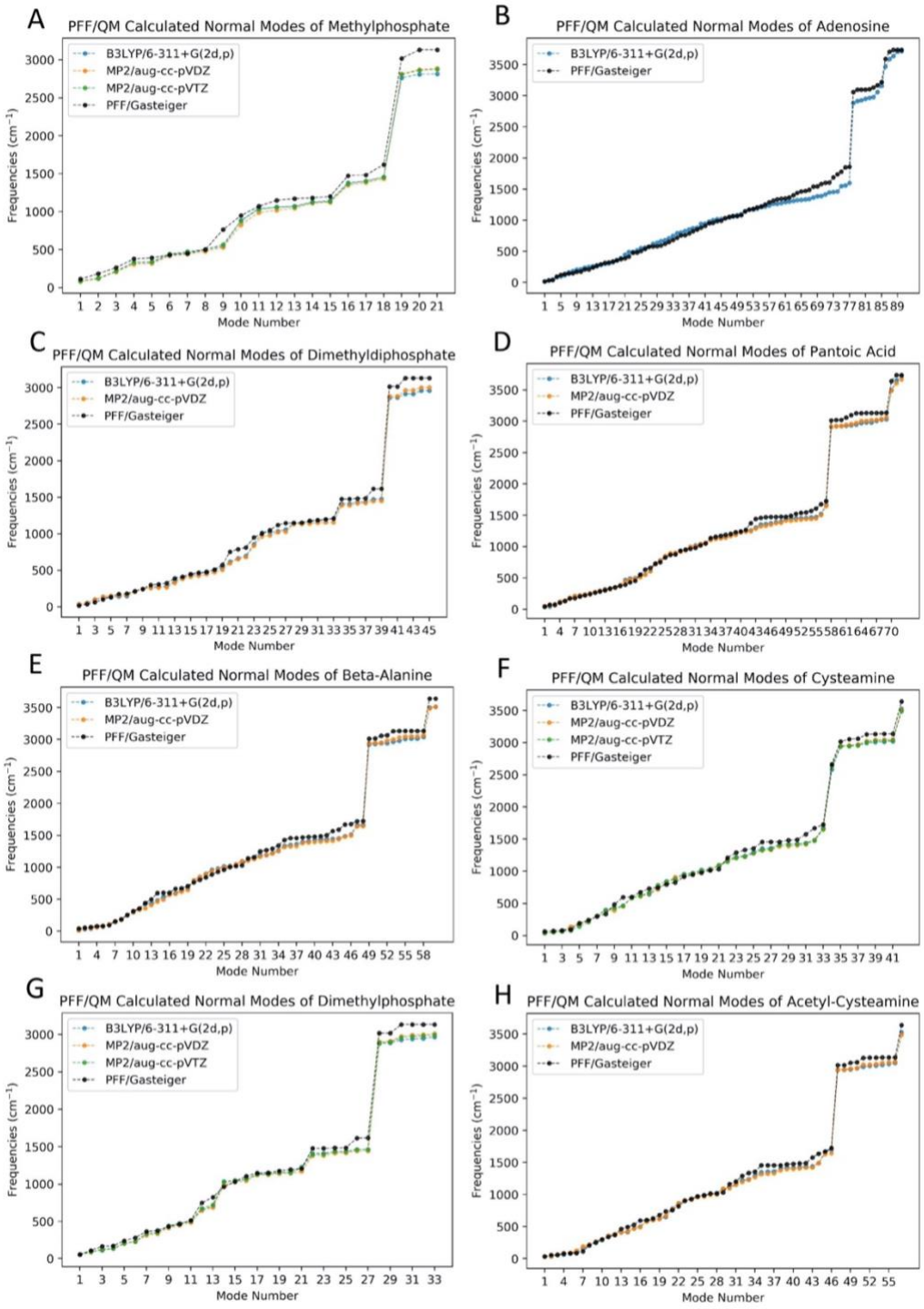


Figure S3.4. Comparison of normal mode frequencies of fragments calculated with PFF/Gasteiger and B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels

of theory. Scaling factors of 0.967, 0.959 and 0.953 were applied to B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ calculated normal mode frequencies, respectively.

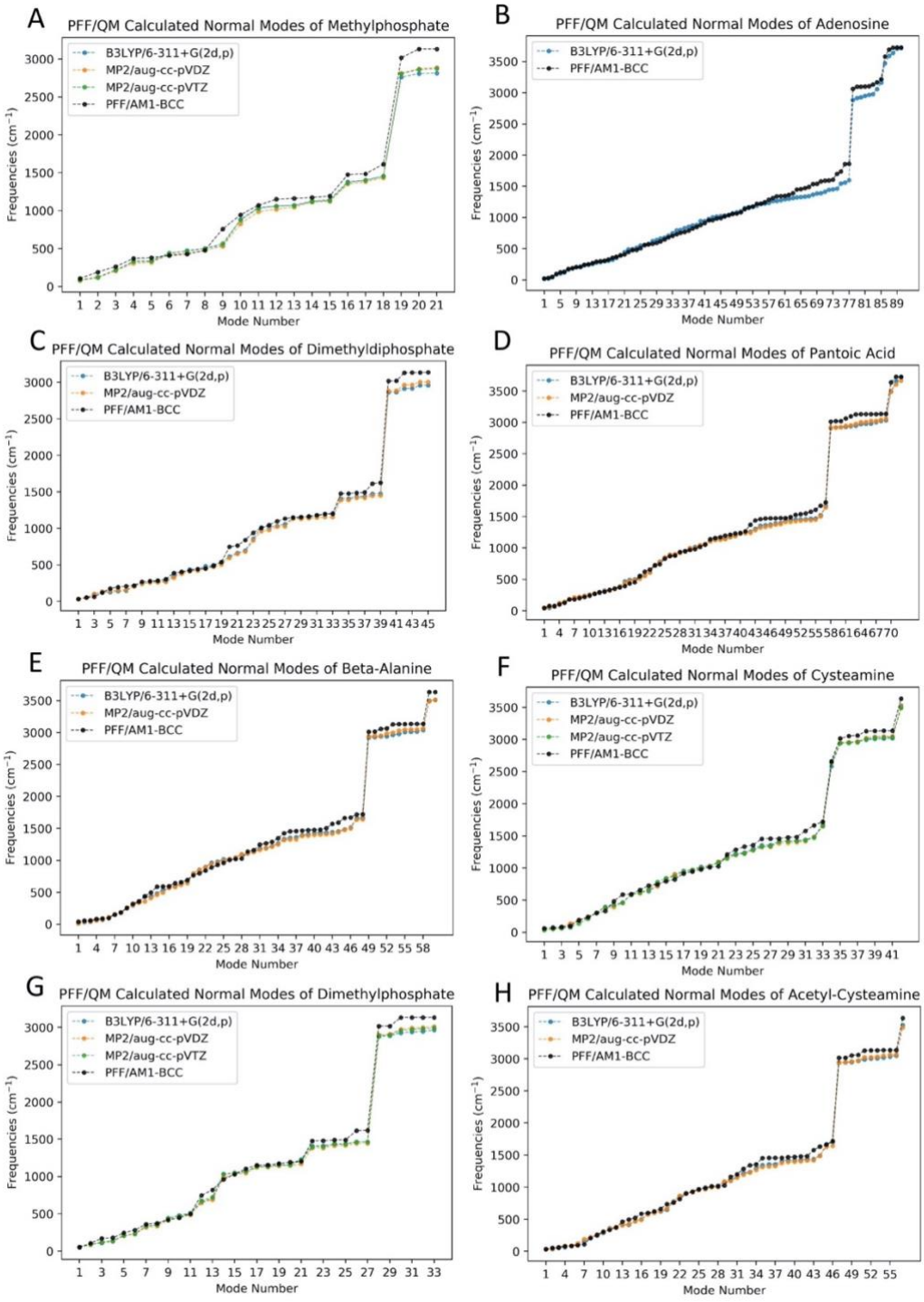


Figure S3.5. Comparison of normal mode frequencies of fragments calculated with PFF/AM1-BCC and B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ levels of theory. Scaling factors of 0.967, 0.959 and 0.953 were applied to B3LYP/6-311+G(2d,p), MP2/aug-cc-pVDZ and MP2/aug-cc-pVTZ calculated normal mode frequencies, respectively.

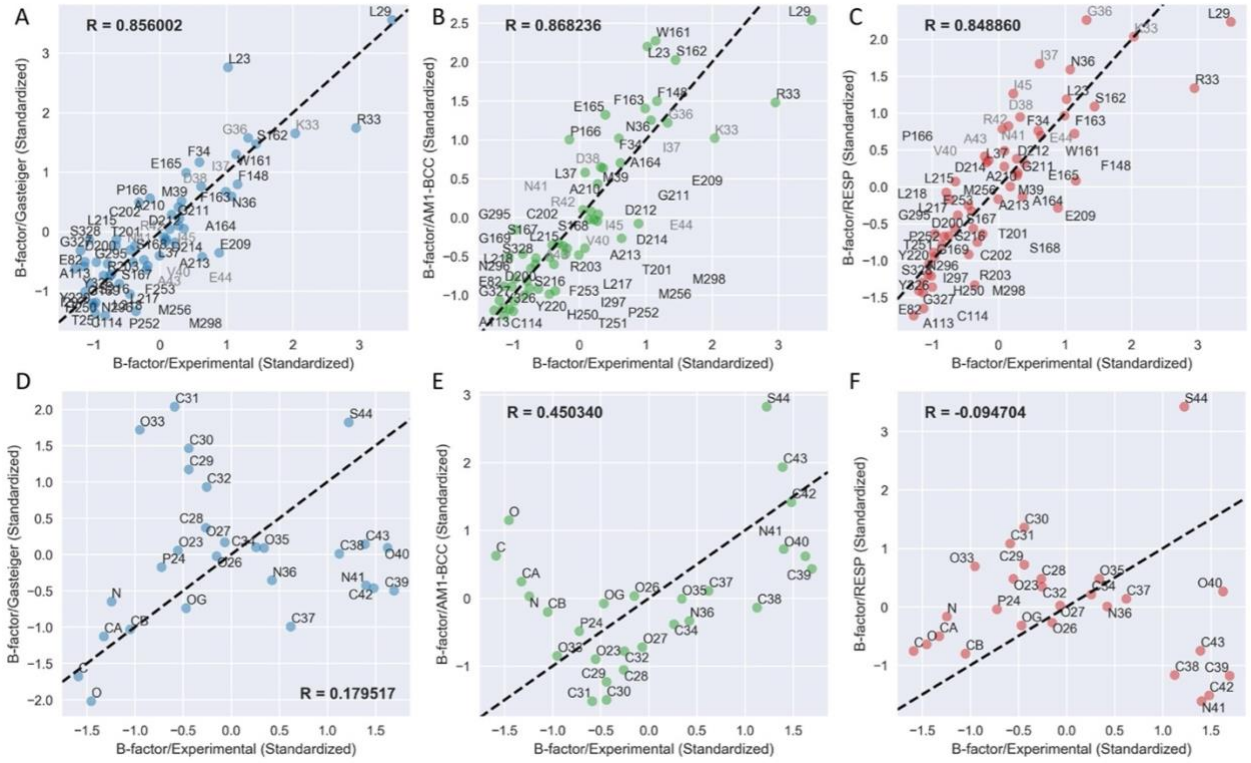


Figure S3.6. Correlation analysis of standardized simulated and experimental B-factors of contact residues (upper panel) and PCLs (lower panel) in the HGMS/ACP-Ppant-Ser system. The residue names of contact residues from HGMS (in black) and ACP (in gray) or atom names of Ppant-Ser are annotated. R is the Pearson correlation coefficient.

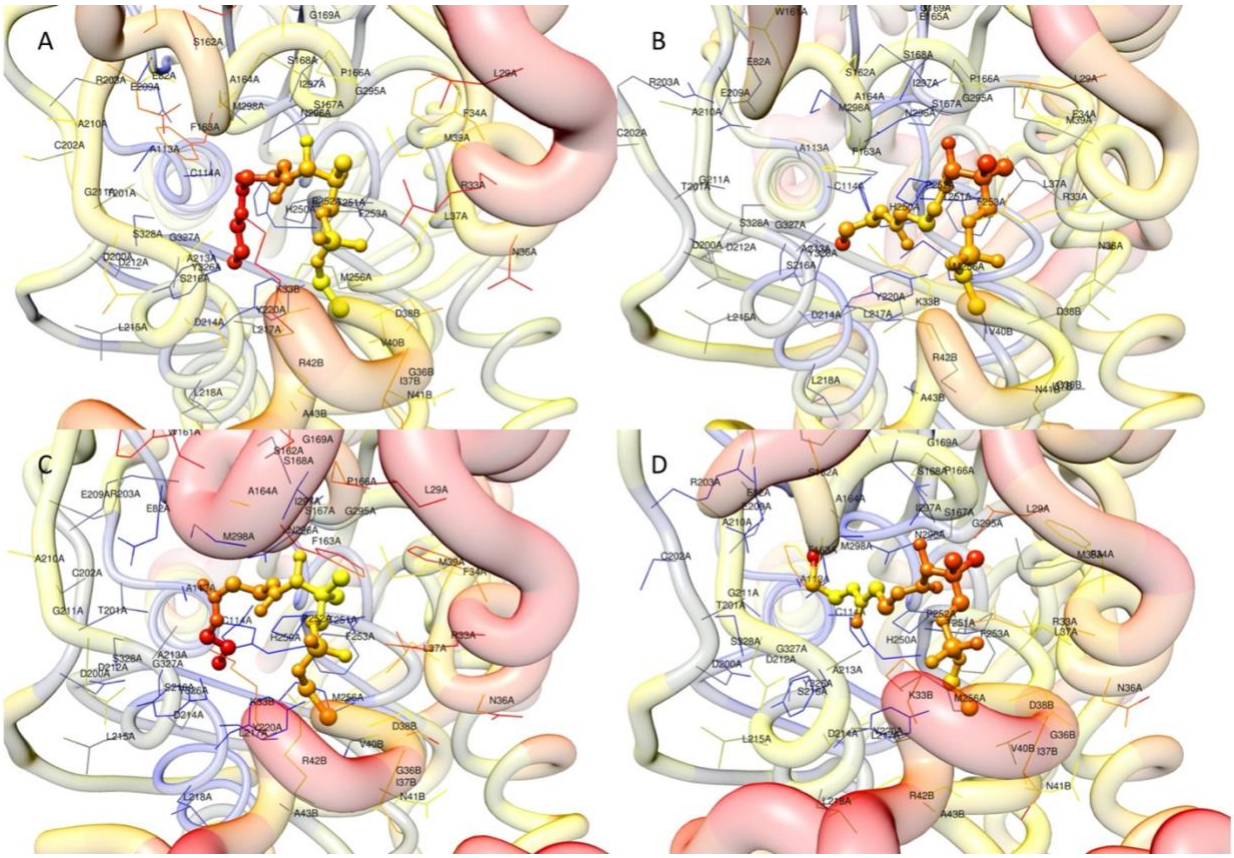


Figure S3.7. Visual comparison of standardized simulated and experimental B-factors of the HGMS/ACP-Ppant-Ser system. Ppant-Ser is depicted as ball-and-stick; contact residues are depicted as wire. Colors (Red color indicates high B-factors, and blue color indicates low B-factors) and thickness of protein backbone also indicate B-factor values. **A.** Experimental structure. **B.** The average structure of the last 10 ns trajectory with Gasteiger charges. **C.** The average structure of the last 10 ns trajectory with AM1-BCC charges. **D.** The average structure of the last 10 ns trajectory with RESP charges.

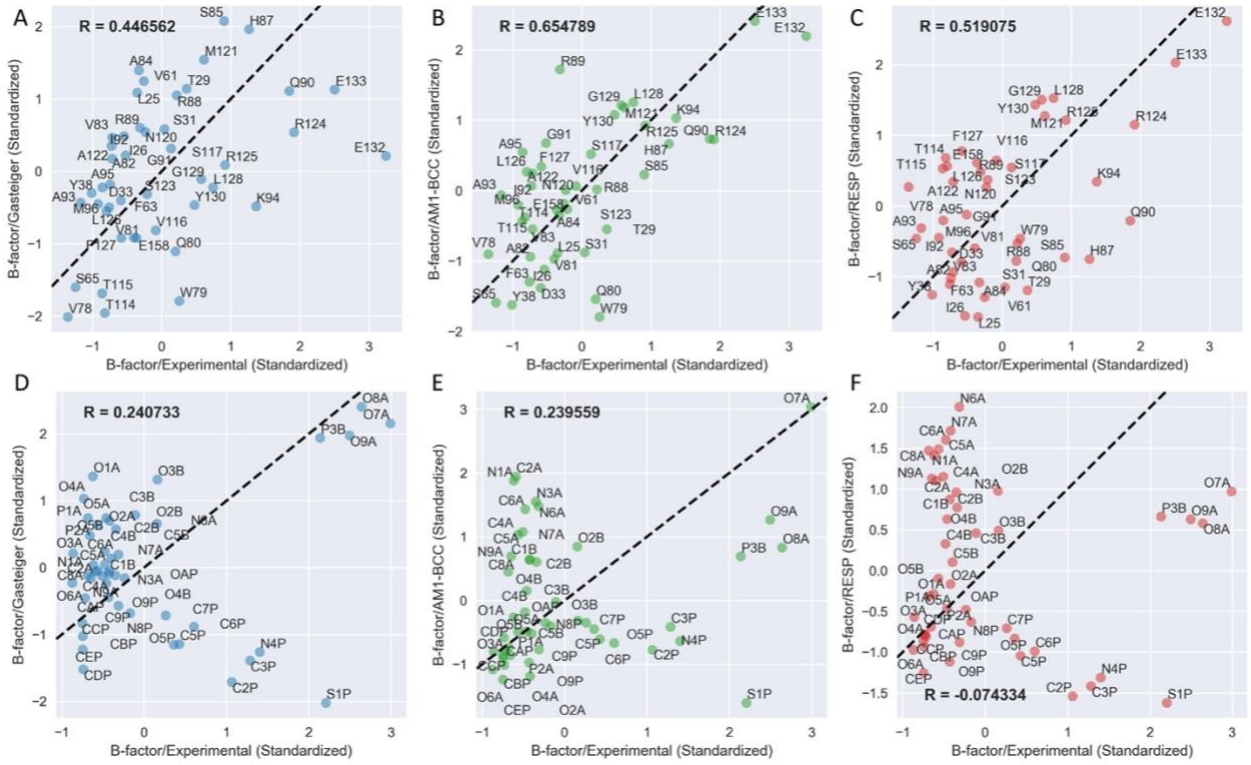


Figure S3.8. Correlation analysis of standardized simulated and experimental B-factors of contact residues (upper panel) and PCLs (lower panel) in the EctA-CoA system. The residue names of contact residues from EctA or atom names of CoA are annotated. R is the Pearson correlation coefficient.

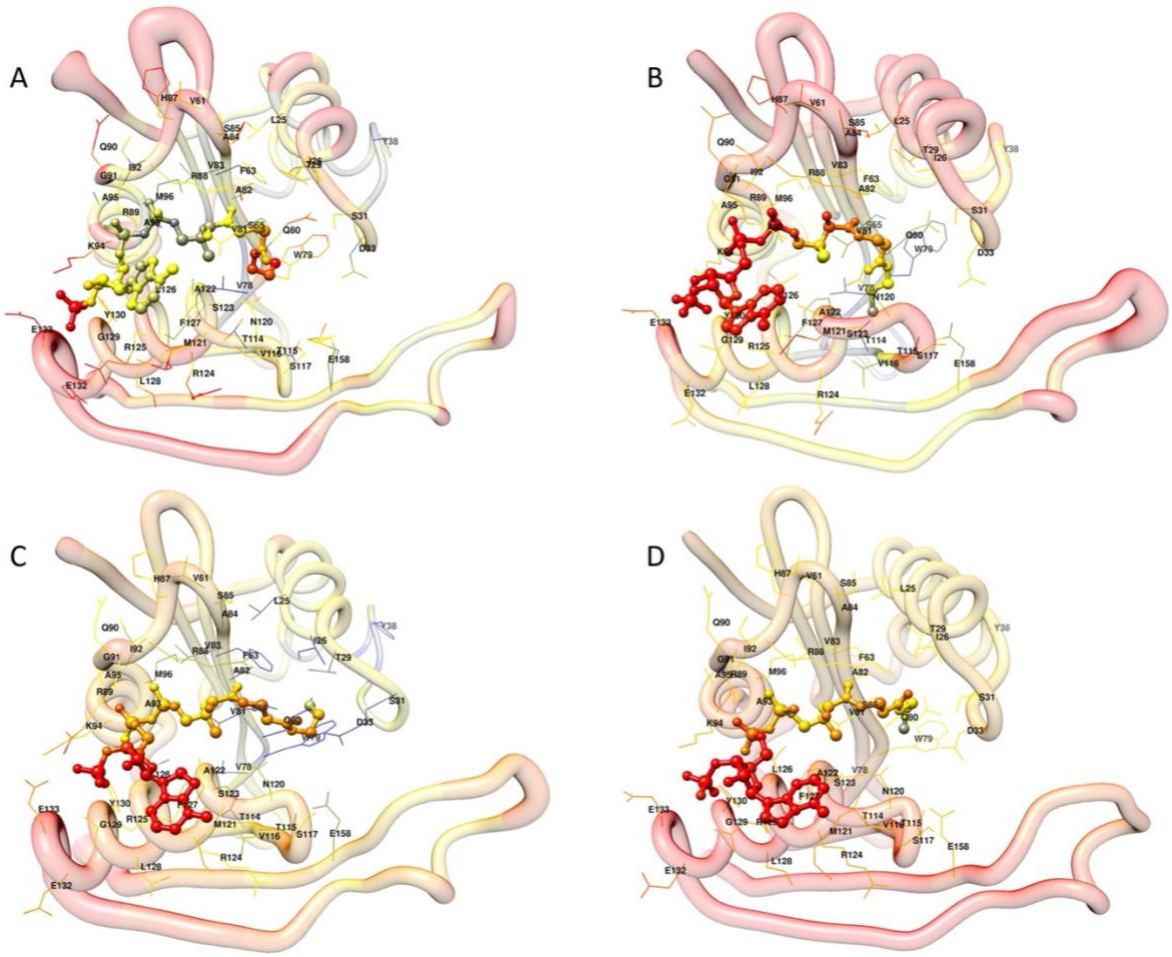


Figure S3.9. Visual comparison of standardized simulated and experimental B-factors of the EctA-CoA system. CoA is depicted as ball-and-stick; contact residues are depicted as wire. Colors (Red color indicates high B-factors, and blue color indicates low B-factors) and thickness of protein backbone also indicate B-factor values. **A.** Experimental structure. **B.** The average structure of the last 10 ns trajectory with Gasteiger charges. **C.** The average structure of the last 10 ns trajectory with AM1-BCC charges. **D.** The average structure of the last 10 ns trajectory with RESP charges.

References

1. Mishra, P. K.; Drueckhammer, D. G., Coenzyme A analogues and derivatives: Synthesis and applications as mechanistic probes of coenzyme A ester-utilizing enzymes. *Chemical reviews* **2000,** *100* (9), 3283-3310.
2. Shi, L.; Tu, B. P., Acetyl-CoA and the regulation of metabolism: mechanisms and consequences. *Current opinion in cell biology* **2015,** *33*, 125-131.
3. Hoagland, M. B.; Novelli, G. D., Biosynthesis of coenzyme A from phosphopantetheine and of pantetheine from pantothenate. *J. biol. Chem* **1954,** *207*, 767-773.
4. Harwood, J. L., Fatty acid metabolism. *Annual Review of Plant Physiology and Plant Molecular Biology* **1988,** *39* (1), 101-138.
5. Daugherty, M.; Polanuyer, B.; Farrell, M.; Scholle, M.; Lykidis, A.; de Crécy-Lagard, V.; Osterman, A., Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *Journal of Biological Chemistry* **2002,** *277* (24), 21431-21439.
6. Leonardi, R.; Jackowski, S., Biosynthesis of Pantothenic Acid and Coenzyme A. *EcoSal Plus* **2007,** *2* (2), 10.1128/ecosalplus.3.6.3.4.
7. Leonardi, R.; Zhang, Y.-M.; Rock, C. O.; Jackowski, S., Coenzyme A: back in action. *Progress in lipid research* **2005,** *44* (2-3), 125-153.
8. Qiao, C.; Wilson, D. J.; Bennett, E. M.; Aldrich, C. C., A mechanism-based aryl carrier protein/thiolation domain affinity probe. *Journal of the American Chemical Society* **2007,** *129* (20), 6350-6351.
9. Zhou, Z.; Lai, J. R.; Walsh, C. T., Directed evolution of aryl carrier proteins in the enterobactin synthetase. *Proceedings of the National Academy of Sciences* **2007,** *104* (28), 11621-11626.
10. Elovson, J.; Vagelos, P. R., Acyl carrier protein X. Acyl carrier protein synthetase. *Journal of Biological Chemistry* **1968,** *243* (13), 3603-3611.
11. Byers, D. M.; Gong, H., Acyl carrier protein: structure–function relationships in a conserved multifunctional protein family. *Biochemistry and Cell Biology* **2007,** *85* (6), 649-662.
12. White, S. W.; Zheng, J.; Zhang, Y.-M.; Rock, C. O., The structural biology of type II fatty acid biosynthesis. *Annu. Rev. Biochem.* **2005,** *74*, 791-831.
13. Sattely, E. S.; Fischbach, M. A.; Walsh, C. T., Total biosynthesis: in vitro reconstitution of polyketide and nonribosomal peptide pathways. *Natural product reports* **2008,** *25* (4), 757-793.
14. Schwarzer, D.; Finking, R.; Marahiel, M. A., Nonribosomal peptides: from genes to products. *Natural product reports* **2003,** *20* (3), 275-287.
15. Nielsen, J.; Keasling, J. D., Engineering cellular metabolism. *Cell* **2016,** *164* (6), 1185-1197.
16. Yuzawa, S.; Kim, W.; Katz, L.; Keasling, J. D., Heterologous production of polyketides by modular type I polyketide synthases in Escherichia coli. *Current opinion in biotechnology* **2012,** *23* (5), 727-735.
17. Nguyen, C.; Haushalter, R. W.; Lee, D. J.; Markwick, P. R.; Bruegger, J.; Caldara-Festin, G.; Finzel, K.; Jackson, D. R.; Ishikawa, F.; O'Dowd, B., Trapping the dynamic acyl carrier protein in fatty acid biosynthesis. *Nature* **2014,** *505* (7483), 427-431.
18. Dowling, D. P.; Kung, Y.; Croft, A. K.; Taghizadeh, K.; Kelly, W. L.; Walsh, C. T.; Drennan, C. L., Structural elements of an NRPS cyclization domain and its intermodule docking domain. *Proceedings of the National Academy of Sciences* **2016,** *113* (44), 12432-12437.
19. Bravo-Rodriguez, K.; Klopries, S.; Koopmans, K. R.; Sundermann, U.; Yahiaoui, S.; Arens, J.; Kushnir, S.; Schulz, F.; Sanchez-Garcia, E., Substrate flexibility of a mutated acyltransferase domain and implications for polyketide biosynthesis. *Chemistry & biology* **2015,** *22* (11), 1425-1430.
20. Barajas, J. F.; Phelan, R. M.; Schaub, A. J.; Kliewer, J. T.; Kelly, P. J.; Jackson, D. R.; Luo, R.; Keasling, J. D.; Tsai, S.-C., Comprehensive structural and biochemical analysis of the terminal myxalamid reductase domain for the engineered production of primary alcohols. *Chemistry & biology* **2015,** *22* (8), 1018-1029.

21. Milligan, J. C.; Lee, D. J.; Jackson, D. R.; Schaub, A. J.; Beld, J.; Barajas, J. F.; Hale, J. J.; Luo, R.; Burkart, M. D.; Tsai, S.-C., Molecular basis for interactions between an acyl carrier protein and a ketosynthase. *Nature chemical biology* **2019,** *15* (7), 669-671.
22. Salomon‐Ferrer, R.; Case, D. A.; Walker, R. C., An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013,** *3* (2), 198-210.
23. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015,** *11* (8), 3696-3713.
24. Galindo-Murillo, R.; Robertson, J. C.; Zgarbova, M.; Sponer, J.; Otyepka, M.; Jurečka, P.; Cheatham III, T. E., Assessing the current state of amber force field modifications for DNA. *Journal of chemical theory and computation* **2016,** *12* (8), 4114-4127.
25. Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González‐Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *Journal of computational chemistry* **2008,** *29* (4), 622-655.
26. Dickson, C. J.; Madej, B. D.; Skjevik, Å. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C., Lipid14: the amber lipid force field. *Journal of chemical theory and computation* **2014,** *10* (2), 865-879.
27. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The protein data bank. *Nucleic acids research* **2000,** *28* (1), 235-242.
28. Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S., RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* **2019,** *47* (D1), D464-D474.
29. Gasteiger, J.; Marsili, M., Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980,** *36* (22), 3219-3228.
30. Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: I. Method. *Journal of computational chemistry* **2000,** *21* (2), 132-146.
31. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of computational chemistry* **2002,** *23* (16), 1623-41.
32. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993,** *97* (40), 10269-10280.
33. Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A., Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *Journal of computational chemistry* **1995,** *16* (11), 1357-1377.
34. Skjevik, Å. A.; Madej, B. D.; Walker, R. C.; Teigen, K., LIPID11: a modular framework for lipid simulations using amber. *The Journal of Physical Chemistry B* **2012,** *116* (36), 11124-11136.
35. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174.
36. Hawkins, P. C.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T., Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. *Journal of chemical information and modeling* **2010,** *50* (4), 572-584.
37. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology* **1999,** *285* (4), 1735-1747.

38. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004,** *25* (13), 1605-1612.
39. Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P., The REd. Tools: Advances in RESP and ESP charge derivation and force field library building. *Physical Chemistry Chemical Physics* **2010,** *12* (28), 7821-7839.
40. Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G., Gaussian 09; Gaussian, Inc. *Wallingford, CT* **2009,** *32*, 5648-5652.
41. Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A., Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling* **2006,** *25* (2), 247-260.
42. Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005,** *26* (16), 1668-88.
43. NIST Computational Chemistry Comparison and Benchmark Database, NIST Standard Reference Database Number 101, Release 20, August 2020. Editor: Russell D. Johnson III. http://cccbdb.nist.gov/.
44. Le Grand, S.; Götz, A. W.; Walker, R. C., SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications* **2013,** *184* (2), 374-380.
45. Macke, T. J.; Case, D. A., Modeling Unusual Nucleic Acid Structures. In *Molecular Modeling of Nucleic Acids*, American Chemical Society: 1997; Vol. 682, pp 379-393.
46. Takahashi, H.; Inagaki, E.; Fujimoto, Y.; Kuroishi, C.; Nodake, Y.; Nakamura, Y.; Arisaka, F.; Yutani, K.; Kuramitsu, S.; Yokoyama, S., Structure and implications for the thermal stability of phosphopantetheine adenylyltransferase from Thermus thermophilus. *Acta Crystallographica Section D: Biological Crystallography* **2004,** *60* (1), 97-104.
47. Maloney, F. P.; Gerwick, L.; Gerwick, W. H.; Sherman, D. H.; Smith, J. L., Anatomy of the β-branching enzyme of polyketide biosynthesis and its interaction with an acyl-ACP substrate. *Proceedings of the National Academy of Sciences* **2016,** *113* (37), 10316-10321.
48. Richter, A. A.; Kobus, S.; Czech, L.; Hoeppner, A.; Zarzycki, J.; Erb, T. J.; Lauterbach, L.; Dickschat, J. S.; Bremer, E.; Smits, S. H., The architecture of the diaminobutyrate acetyltransferase active site provides mechanistic insight into the biosynthesis of the chemical chaperone ectoine. *Journal of Biological Chemistry* **2020,** *295* (9), 2822-2838.
49. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M.; Eramian, D.; Shen, M. y.; Pieper, U.; Sali, A., Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics* **2006,** *15* (1), 5.6. 1-5.6. 30.
50. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.
51. Miyamoto, S.; Kollman, P. A., Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry* **1992,** *13* (8), 952-962.
52. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of computational physics* **1977,** *23* (3), 327-341.
53. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **1993,** *98* (12), 10089-10092.
54. Crowley, M.; Darden, T.; Cheatham, T.; Deerfield, D., Adventures in improving the scaling and accuracy of a parallel molecular dynamics program. *The Journal of Supercomputing* **1997,** *11* (3), 255-278.

55. Loncharich, R. J.; Brooks, B. R.; Pastor, R. W., Langevin dynamics of peptides: The frictional dependence of isomerization rates of N - acetylalanyl - N ' - methylamide. *Biopolymers: Original Research on Biomolecules* **1992,** *32* (5), 523-535.
56. Zhao, S.; Ni, F.; Qiu, T.; Wolff, J. T.; Tsai, S.-C.; Luo, R., Molecular Basis for Polyketide Ketoreductase–Substrate Interactions. *International Journal of Molecular Sciences* **2020,** *21* (20), 7562.
57. Roe, D. R.; Cheatham III, T. E., PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *Journal of chemical theory and computation* **2013,** *9* (7), 3084-3095.
58. Ray, L.; Valentic, T. R.; Miyazawa, T.; Withall, D. M.; Song, L.; Milligan, J. C.; Osada, H.; Takahashi, S.; Tsai, S.-C.; Challis, G. L., A crotonyl-CoA reductase-carboxylase independent pathway for assembly of unusual alkylmalonyl-CoA polyketide synthase extender units. *Nature communications* **2016,** *7* (1), 1-12.
59. Zgarbová, M.; Otyepka, M.; Šponer, J. i.; Mládek, A. t.; Banáš, P.; Cheatham III, T. E.; Jurecka, P., Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of chemical theory and computation* **2011,** *7* (9), 2886-2902.
60. Scott, A. P.; Radom, L., Harmonic vibrational frequencies: an evaluation of Hartree− Fock, Møller− Plesset, quadratic configuration interaction, density functional theory, and semiempirical scale factors. *The Journal of Physical Chemistry* **1996,** *100* (41), 16502-16513.
61. Hashim, O. H.; Adnan, N. A., Coenzyme, cofactor and prosthetic group: ambiguous biochemical jargon. *Biochemical education* **1994,** *22* (2), 93-94.
62. Wang, L.-P.; McKiernan, K. A.; Gomes, J.; Beauchamp, K. A.; Head-Gordon, T.; Rice, J. E.; Swope, W. C.; Martínez, T. J.; Pande, V. S., Building a more predictive protein force field: a systematic and reproducible route to AMBER-FB15. *The Journal of Physical Chemistry B* **2017,** *121* (16), 4023-4039.
63. Lamoureux, G.; Orabi, E. A., Molecular modelling of cation–π interactions. *Molecular Simulation* **2012,** *38* (8-9), 704-722.
64. Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y., Development of Polarizable Models for Molecular Mechanical Calculations I: Parameterization of Atomic Polarizability. *Journal of Physical Chemistry B* **2011,** *115* (12), 3091-3099.
65. Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y., Development of Polarizable Models for Molecular Mechanical Calculations II: Induced Dipole Models Significantly Improve Accuracy of Intermolecular Interaction Energies. *Journal of Physical Chemistry B* **2011,** *115* (12), 3100-3111.
66. Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M. J.; Wang, J. M.; Duan, Y.; Luo, R., Development of Polarizable Models for Molecular Mechanical Calculations. 3. Polarizable Water Models Conforming to Thole Polarization Screening Schemes. *Journal of Physical Chemistry B* **2012,** *116* (28), 7999-8008.
67. Wang, J. M.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M. J.; Luo, R.; Duan, Y., Development of Polarizable Models for Molecular Mechanical Calculations. 4. van der Waals Parametrization. *Journal of Physical Chemistry B* **2012,** *116* (24), 7088-7101.
68. Wang, J.; Cieplak, P.; Luo, R.; Duan, Y., Development of Polarizable Gaussian Model for Molecular Mechanical Calculations I: Atomic Polarizability Parameterization To Reproduce ab Initio Anisotropy. *Journal of chemical theory and computation* **2019,** *15* (2), 1146-1158.
69. Wei, H.; Qi, R.; Wang, J.; Cieplak, P.; Duan, Y.; Luo, R., Efficient formulation of polarizable Gaussian multipole electrostatics for biomolecular simulations. *The Journal of Chemical Physics* **2020,** *153* (11), 114116.

CHAPTER 4

PyRESP: A Program for Electrostatic Parameterizations of Additive and Induced Dipole Polarizable Force Fields

4.1. Introduction

Developing accurate force fields remains to be a great challenge for molecular modeling. One of the key components of force field development is the accurate modeling of atomic electrostatic interactions. The extensively used additive force fields apply fixed atom-centered partial charges to model electrostatic interactions, such as AMBER ff14SB,[1] ff19SB,[2] CHARMM,[3] and OPLS,[4] to name a few. One disadvantage of the additive force fields is that they are unable to model the atomic polarization effects, i.e., the redistribution of the atomic electron density due to the electric field produced by nearby atoms.[5] The importance of modeling polarization effects is well known. For example, during the protein folding process, amino acids forming hydrophobic core must move from the hydrated environment to the more hydrophobic interior, experiencing considerably different dielectric environments.[6-7] Additive force fields are also considered to be unable to capture the important cation-π interactions between aromatic rings and charged amino acids, leading to unrealistic receptor-ligand interaction simulations.[8-9] Therefore, a great deal of effort has been directed to developing polarizable models, including the fluctuating charge models,[10-11] the Drude oscillator models,[12-16] and models incorporating induced dipoles[17-18] or continuum dielectric.[19-20]

The induced point dipole model is the most studied approach with a long history since 1970s.[21-22] To date, it has been incorporated into several polarizable force fields, including

AMOEBA,[23-24] AMBER ff02,[17] ff02pol.rl,[18] and ff12pol.[25-28] The original induced dipole model developed by Applequist et al. places the induced point dipole on each atom center, where the magnitude and direction of the induced dipole moment is determined by the isotropic polarizability of each atom and the electric field on this atom exerted by other atoms.[29] The induced dipole of atom i, subject to external electric field $\boldsymbol{E}_i$, is

$$\boldsymbol{\mu}_i = \alpha_i \left[\boldsymbol{E}_i - \sum_{j \neq i}^{\boldsymbol{n}} \boldsymbol{T}_{ij} \boldsymbol{\mu}_j \right] \tag{4.1}$$

where α_i is the isotropic polarizability of atom i, and $\boldsymbol{T}_{ij}$ is the dipole field tensor with the matrix form

$$\boldsymbol{T}_{ij} = \frac{1}{{r_{ij}}^3} \boldsymbol{I} - \frac{3}{{r_{ij}}^5} \begin{bmatrix} x^2 & xy & xz \\ xy & y^2 & yz \\ xz & yz & z^2 \end{bmatrix} \tag{4.2}$$

where $\boldsymbol{I}$ is the identity matrix, and x, y and z are the Cartesian components along the vector between atoms i and j at distance r_{ij}. However, this model suffers from the so-called "polarization catastrophe" problem: the molecular polarizability diverges due to the cooperative interaction between induced dipoles at short distances.[5, 29] One solution to this problem is to apply distance-dependent damping functions for interactions on short distances. Thole proposed several schemes by modeling the interaction using smeared charge distributions $\rho(u)$ instead of point charges, where $u = r_{ij}/(\alpha_i \alpha_j)^{1/6}$ is the effective distance, in which α_i and α_j are atomic polarizabilities of atoms i and j, and r_{ij} is the distance between them.[30-31] This will modify the dipole field tensor $\boldsymbol{T}_{ij}$ in such a way that it

does not behave as r^{-3} at short distances. Among the proposed schemes, linear scheme (**eq. 4.3**) and exponential scheme (**eq. 4.4**) are shown to be the most effective:

$$\rho(u) = \begin{cases} \frac{3}{\pi}\frac{(a-u)}{a^4} & u < a \\ 0 & u \geq a \end{cases} \tag{4.3}$$

and

$$\rho(u) = \frac{a^3}{8\pi}\exp(-au) \tag{4.4}$$

where a is the damping factor that controls the decay of the smeared charge distribution. Another Thole's scheme (**eq. 4.5**) was adopted in the AMOEBA force field and implemented in Tinker program,[23-24, 32] which has the following form

$$\rho(u) = \frac{3a}{4\pi}\exp(-au^3) \tag{4.5}$$

The recently developed Thole scheme-based polarizable force field ff12pol have been shown to significantly reduce the root-mean-square errors of interaction energies with those calculated at the MP2/aug-cc-pVTZ level of theory, compared with additive force fields.[26]

About a decade ago, Elking et al. proposed a polarizable multipole model with Gaussian charge densities, which was later named as polarizable Gaussian Multipole (pGM) model.[33] The nth order Gaussian multipole at position $\boldsymbol{r}$ generated by an atom located at coordinate $\boldsymbol{R}$ represented by the pGM model is defined as

$$\rho^{(n)}(\boldsymbol{r};\boldsymbol{R}) = \boldsymbol{\Theta}^{(n)} \cdot \nabla_R^{(n)} \left(\frac{\beta}{\sqrt{\pi}}\right)^3 e^{-\beta^2|\boldsymbol{r}-\boldsymbol{R}|^2} \tag{4.6}$$

where $\boldsymbol{\Theta}^{(n)}$ is the nth rank momentum tensor, $\nabla_R^{(n)}$ is the nth rank gradient operator, and β is a Gaussian exponent controlling the "radius" of the distribution with the following form

$$\beta = s\left(\frac{2\alpha}{3\sqrt{2\pi}}\right)^{-\frac{1}{3}} \tag{4.7}$$

where α is the atomic polarizability, and s is the screening factor. Although in the pGM model any order of multipoles can be modeled, only charges (0th order multipole, **eq. 4.8**) and dipoles (1st order multipole, **eq. 4.9**) are retained in the current pGM model design.

$$\rho^{(0)}(\boldsymbol{r};\boldsymbol{R}) = q\left(\frac{\beta}{\sqrt{\pi}}\right)^3 e^{-\beta^2|\boldsymbol{r}-\boldsymbol{R}|^2} \tag{4.8}$$

$$\rho^{(1)}(\boldsymbol{r};\boldsymbol{R}) = \boldsymbol{p}\cdot\nabla_R\left(\frac{\beta}{\sqrt{\pi}}\right)^3 e^{-\beta^2|\boldsymbol{r}-\boldsymbol{R}|^2} \tag{4.9}$$

where q is the permanent charge and $\boldsymbol{p}$ is the permanent dipole. Wei et al. recently proposed a local frame for the permanent dipoles formed by covalent basis vectors (CBVs), which are unit vectors along the direction of covalent bonds or virtual bonds.[34-35] This design is based on the fact that atomic permanent moments mainly result from covalent bonding interactions. Replacing $\boldsymbol{p}$ with $\boldsymbol{\mu}$ in **eq. 4.9** will give the pGM distribution of induced dipole, which has the same form as that of permanent dipole. A key advantage of the pGM model is that all short-range electrostatic interactions can be calculated analytically in a consistent manner, including the interactions of charge-charge, charge-dipole, charge-quadrupole, dipole-dipole, and so on. Consequently, it has been shown that the pGM model notably improves the prediction of molecular polarizability anisotropy compared with that of Thole models.[36]

Each of the four damping schemes discussed above requires parameterization of the atomic isotropic polarizabilities α and damping factors a (and s for the pGM model), which has been done by fitting experimental or *ab initio* molecular polarizability tensors using a genetic algorithm as presented in our recent works.[25, 36] In this chapter, we aim to take one step further toward the development of general and accurate polarizable force fields by developing a computer program for electrostatic parameterizations for the atomic charges and dipoles of various polarizable models.

For additive models, the atomic point partial charges are traditionally derived by performing least-square fitting of the charges to reproduce the quantum mechanically (QM) determined electrostatic potential (ESP) at a large number of grid points lying outside the van der Waals distance of the molecule. Assuming a molecule with n atoms is being parameterized, and there are m ESP points lying outside the van der Waals distance of the molecule, then the least-square fitting aims to minimize the objective function

$$\gamma = \sum_{j=1}^{m} \left(V_j^{QM} - V_j\right)^2 \tag{4.10}$$

where V_j^{QM} is the ESP value evaluated through QM calculations at point j, and V_j is the ESP value calculated from the fitting results. This method was initially used by Momany,[37] further refined by Cox et al.[38] A ESP point sampling scheme that uses points on molecular surfaces constructed using gradually increasing van der Waals radii for the atoms was proposed by Singh et al.[39-40] The CHELP algorithm initially employed a Lagrange multiplier method to perform constrained least-square fitting, in which the Lagrange multiplier (λ) is multiplied by the constraining function (g) and added to the objective function γ to be minimized. In

the context of charge fitting, the Lagrange multiplier method is most used to enforce the total charge constraints, i.e., the charge of all atoms of a molecule should sum to the total molecular charge. Alternatively, it can also be used to specify the total charge of molecule fragments. For example, during amino acid parameterizations, the *N*-acetyl (ACE) and *N*-methylamide (NME) groups are commonly used to cap amino acid dipeptides to mimic the chemical environment within a protein. Both capping fragments need to be constrained to have neutral charge to ensure correct total charge of the amino acid fragments.[41-42]

In general, the ESP-based charge derivation methods perform very well in reproducing QM determined molecular multipole moments, and optimally reproduce intermolecular interaction energy. However, all methods discussed above suffer from the problem that the atomic charges are sensitive to molecular conformations, leading to a lack of transferability of the charges between identical molecules with different conformations, as well as between common functional groups in related molecules. Another problem of this approach is the poor determination of charges on buried atoms that are far from ESP points, which can fluctuate wildly to reach the optimal fitting to the ESP. Both problems have been addressed by the restrained electrostatic potential (RESP) method developed by Bayly et al., which employs restraints by adding a penalty function χ to the objective function during the fitting process.[43-44] Two types of penalty functions were proposed. The first is a simple harmonic function

$$\chi = a \sum_{i=1}^{n} {q_i}^2 \tag{4.11}$$

where a is the scale factor determining the restraining strength. The second penalty function is a hyperbolic function with the form

$$\chi = a\sum_{i=1}^{n}\left(\sqrt{{q_i}^2 + b^2} - b\right) \tag{4.12}$$

where a is again the scale factor that defines the restraining strength, and b determines the "tightness" of the hyperbola around its minimum. b has been recommended to be set to 0.1 by the original RESP work to make the restraint appropriately tight.[43] To this end, assuming there are w different Lagrange constraints imposed on the charges in a molecule, the objective function to be minimized becomes:

$$z = \gamma + \lambda_1 g_1 + \lambda_2 g_2 + \cdots + \lambda_w g_w + \chi \tag{4.13}$$

To date, the computer program *RESP* has been applied in charge derivations of a variety of additive force fields,[41-42] and is still being used actively for charge calculations for small organic molecules.[8, 45-46] Following the idea of charge parameterization by reproducing ESPs, Cieplak et al. extended the RESP method for induced dipole electrostatic models, assuming that ESPs around molecules are determined by both permanent charges and atomic induced dipoles. According to this method, atomic charges are iteratively fitted to the effective ESP, which is the differences between the QM derived ESPs and the ESPs generated by induced dipoles. Iterations stop when the induced molecular dipole moment converges within certain accuracy level.[5, 17] A program named *i_RESP* has been developed to facilitate this iterative charge fitting procedure.

In this chapter, we further extended the RESP method for parameterizations of electrostatic models with induced point dipoles and permanent point dipoles. A Python

program named *PyRESP* was designed and implemented based on its ancestor *RESP* program, providing the parameterization ability for three electrostatic models: (1) the additive RESP model; (2) the polarizable model with induced point dipoles only, named as RESP-ind model; and (3) the polarizable model with both induced point dipoles and permanent point dipoles, named as RESP-perm model. In the next section we present the theory behind the parametrization strategies of the three models, as well as several other features provided by *PyRESP*. We have tested all three models using several representative molecules, and the parameterization results will be evaluated and discussed.

4.2. Theory

In earlier works the objective function z shown in **eq. 4.13** has been minimized using iterative gradient descent approaches, as were done by Momany et al. and Singh et al.[37, 39] Similarly, the *i_RESP* program developed by Cieplak et al. parameterizes the induced dipole polarizable model iteratively by fitting charges to the differences between the QM derived ESPs and the ESPs generated by induced dipoles.[5, 17] In both cases, an initial guess on the atomic charges before the iteration process is required. On the other hand, iterative algorithms suffer from the problem that the convergence of iteration is sensitive to the specified accuracy level. In rare cases, the objective function might jump back and forth near the minimum, leading to a non-convergence problem. Therefore, *PyRESP* takes a direct approach by solving the system of equation in matrix form with the partial derivative of the objective function z against each parameter (permanent charges or dipoles) and each Lagrange multiplier λ set to be equal zero, as were done in CHELP, CHELPG, and the original

RESP works.[43, 47-49] The advantage of the direct approach is that it gives the exact least-squares solution, so that the initial guess on the atomic charges and accuracy level are no longer needed. Another advantage of the direct approach is that the matrix form representations allow us to present each of the following electrostatic models in a consistent and elegant way.

4.2.1. RESP

The original RESP method performs charge fitting for additive electrostatic terms with the assumption that ESPs only come from permanent point charges.[43] For each ESP point j, the following equation need be solved

$$\sum_{i=1}^{n} \frac{q_i}{r_{ij}} = V_j^{QM} \tag{4.14}$$

In matrix form

$$\boldsymbol{Xq} = \boldsymbol{V} \tag{4.15}$$

where $\boldsymbol{X}$ is m by n matrix for charge-ESP interactions between each ESP point j and atom i, $\boldsymbol{q}$ is n-dimensional vector for the partial charge of each atom, $\boldsymbol{V}$ is m-dimensional vector for QM ESP. Typically, there will be many ESP points sampled so that $\boldsymbol{X}$ becomes a rectangular matrix (tall and thin). Consequently, **eq. 4.15** is unlikely to have an exact solution. Therefore, we aim to find the least-squares solution by solving the following equation, the proof of which can be found in most linear algebra textbooks.

$$\boldsymbol{X}^T\boldsymbol{Xq} = \boldsymbol{X}^T\boldsymbol{V} \tag{4.16}$$

where $\boldsymbol{X}^T\boldsymbol{X}$ is a square matrix and is usually positive definite and invertible. The constraints on the charges could also be expressed in the following matrix form

$$\boldsymbol{K}\boldsymbol{q} = \boldsymbol{L} \tag{4.17}$$

where $\boldsymbol{K}$ is a w by n matrix with only 1 and 0 as elements indicating the presence or absence of each charge in each constraint, $\boldsymbol{L}$ is w-dimensional vector for the total charge in each constraint. The constrained least-squares fitting has the following matrix form, whose solution gives constrained RESP fitting results.

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{K}^T \\ \boldsymbol{K} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{V} \\ \boldsymbol{L} \end{bmatrix} \tag{4.18}$$

where $\boldsymbol{\lambda}$ is w-dimensional vectors of all Lagrange multipliers. Finally, the penalty function χ could be applied to restrain fitted charges by adding its partial derivative only to the diagonal terms of the matrix in **eq. 4.18**, and the reasoning can be found in the original RESP work.[43]

4.2.2. RESP-ind (RESP with Induced Point Dipole)

Following Applequist et al,[29] **eq. 4.1** maybe rearranged into

$$\alpha_i^{-1}\boldsymbol{\mu}_i + \sum_{j\neq i}^{\boldsymbol{n}} \boldsymbol{T}_{ij}\boldsymbol{\mu}_j = \boldsymbol{E}_i \tag{4.19}$$

which could be written in the following matrix form

$$\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{E} \tag{4.20}$$

where $\boldsymbol{A}$ is a $3n$ by $3n$ matrix containing the information of polarizability and dipole field tensors, $\boldsymbol{\mu}$ is a $3n$-dimensional vector of the induced dipole of each atom, and $\boldsymbol{E}$ is a $3n$-dimensional vector of the electric field at atom i.

The implicit assumption is that $\boldsymbol{E}_i$ is produced by permanent charges of all atoms other than i, and there are no additional applied external electric fields. Thus, we have

$$\boldsymbol{E}_i = \sum_{j \neq i}^{n} \frac{q_j}{{r_{ij}}^3} \boldsymbol{r}_{ji} \tag{4.21}$$

In matrix form

$$\boldsymbol{E} = \boldsymbol{C}\boldsymbol{q} \tag{4.22}$$

where $\boldsymbol{C}$ is a $3n$ by n matrix of the charge-electric field coefficient between each atom pair. Combining **eq. 4.20** and **eq. 4.22** gives

$$\boldsymbol{\mu} = \boldsymbol{A}^{-1}\boldsymbol{C}\boldsymbol{q} \tag{4.23}$$

In contrast to the RESP model where the permanent charges are the only sources for ESPs, the RESP-ind model assumes that ESP comes from both permanent point charges and induced point dipoles. Therefore, for each ESP point j, we have the following equation

$$\sum_{i=1}^{n} \frac{q_i}{r_{ij}} + \sum_{i=1}^{n} \frac{\boldsymbol{\mu}_i \cdot \boldsymbol{r}_{ij}}{{r_{ij}}^3} = V_j^{QM} \tag{4.24}$$

In matrix form

$$\boldsymbol{X}\boldsymbol{q} + \boldsymbol{Y}\boldsymbol{\mu} = \boldsymbol{V} \tag{4.25}$$

where $\boldsymbol{Y}$ is a m by $3n$ matrix for the dipole-ESP interactions between each ESP point and atom pair. Substitute **eq. 4.23** into **eq. 4.25** gives

$$(\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C})\boldsymbol{q} = \boldsymbol{V} \tag{4.26}$$

Same as we did for the RESP model, solving the following equation gives the least-squares solution

$$(\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C})^T(\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C})\boldsymbol{q} = (\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C})^T\boldsymbol{V} \tag{4.27}$$

and solving the following equation gives the constrained least-squares solution

$$\begin{bmatrix} (\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C})^T(\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C}) & \boldsymbol{K}^T \\ \boldsymbol{K} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{q} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C})^T\boldsymbol{V} \\ \boldsymbol{L} \end{bmatrix} \tag{4.28}$$

Finally, the partial derivative of the penalty function χ can be applied to **eq. 4.28** to restrain atomic charges.

4.2.3. RESP-perm (RESP with Induced and Permanent Point Dipole)

RESP-perm is the electrostatic model with highest degree of freedom implemented in *PyRESP*. It has one additional component compared to the RESP-ind model, the permanent point dipoles $\boldsymbol{p}_i$ of each atom i, which is a 3-dimensional vector. Now the electric field at atom i is produced by both permanent charges and permanent dipoles of all atoms other than i. Thus, we have

$$\boldsymbol{E}_i = \sum_{j \neq i}^{n} \left(\frac{q_j}{{r_{ij}}^3} \boldsymbol{r}_{ji} + \boldsymbol{T}_{ij}\boldsymbol{p}_j \right) \tag{4.29}$$

In matrix form

$$\boldsymbol{E} = \boldsymbol{Cq} + \boldsymbol{Dp} \tag{4.30}$$

where $\boldsymbol{D}$ is a $3n$ by $3n$ matrix of the dipole-electric field coefficients between each atom pair, and $\boldsymbol{p}$ is a $3n$-dimensional vector for the permanent dipole of each atom in global frame. Therefore, the induced dipole vector $\boldsymbol{\mu}$ becomes

$$\boldsymbol{\mu} = \boldsymbol{A^{-1}}(\boldsymbol{Cq} + \boldsymbol{Dp}) \tag{4.31}$$

Now, ESPs come from three sources: permanent point charges, permanent point dipoles, and induced point dipoles. That is

$$\sum_{i=1}^{n} \frac{q_i}{r_{ij}} + \sum_{i=1}^{n} \frac{(\boldsymbol{\mu}_i + \boldsymbol{p}_i) \cdot \boldsymbol{r}_{ij}}{{r_{ij}}^3} = V_j^{QM} \tag{4.32}$$

In matrix form

$$\boldsymbol{Xq} + \boldsymbol{Y}(\boldsymbol{\mu} + \boldsymbol{p}) = \boldsymbol{V} \tag{4.33}$$

Eq. 4.31 can be plugged into **eq. 4.33** and rearranged to

$$(\boldsymbol{X} + \boldsymbol{YA^{-1}C})\boldsymbol{q} + \boldsymbol{Y}(\boldsymbol{A^{-1}D} + \boldsymbol{I})\boldsymbol{p} = \boldsymbol{V} \tag{4.34}$$

The RESP-perm model is designed to be compatible with the pGM model of Wei et al.,[34] where the permanent dipoles are defined in the local frame formed by covalent basis vectors (CBVs). Assume that the molecule to be fitted has z CBVs, i.e., $z/2$ covalent bonds since covalent bonds are bi-directional, then the permanent dipoles in global frame $\boldsymbol{p}$ can be conveniently expressed in the local frame using a $3n$ by z dimensional conversion matrix $\boldsymbol{F}$, with CBVs as its elements. The conversion has the simple matrix form

$$\boldsymbol{p} = \boldsymbol{F}\boldsymbol{p}^{loc} \tag{4.35}$$

where $\boldsymbol{p}^{loc}$ is a z-dimensional vector for permanent dipoles in local frame. Therefore, the RESP-perm model in fact performs least-square fitting on $\boldsymbol{p}^{loc}$ rather than $\boldsymbol{p}$, and **eq. 4.34** should be expressed as

$$(\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C})\boldsymbol{q} + \boldsymbol{Y}(\boldsymbol{A}^{-1}\boldsymbol{D} + \boldsymbol{I})\boldsymbol{F}\boldsymbol{p}^{loc} = \boldsymbol{V} \tag{4.36}$$

One advantage of using matrix $\boldsymbol{F}$ is that the local frame can be easily extended to include non-covalent basis vectors. In the current *PyRESP* implementation, the "virtual" bonds of 1-3 interacting atom pairs are also enabled; all we need to do is to increase the number of columns of $\boldsymbol{F}$ to contain both covalent basis vectors and 1-3 interaction basis vectors, and the number of rows of $\boldsymbol{F}$ will not change, since the number of atoms stays the same. The RESP-perm model considering both 1-2 and 1-3 interacting atom pairs in the local frame is named as RESP-perm-v, where v stands for "virtual".

In order to perform least-squares fitting on both $\boldsymbol{q}$ and $\boldsymbol{p}^{loc}$ directly, we construct a new vector $\boldsymbol{Q}$, which is $(n+z)$-dimensional vector $\begin{bmatrix}\boldsymbol{q}\\ \boldsymbol{p}^{loc}\end{bmatrix}$, and a new matrix $\boldsymbol{M}$, which is m by $(n+z)$ matrix $[(\boldsymbol{X} + \boldsymbol{Y}\boldsymbol{A}^{-1}\boldsymbol{C}) \quad \boldsymbol{Y}(\boldsymbol{A}^{-1}\boldsymbol{D} + \boldsymbol{I})\boldsymbol{F}]$. Then we have

$$\boldsymbol{M}\boldsymbol{Q} = \boldsymbol{V} \tag{4.37}$$

The least-squares solution of $\boldsymbol{Q}$ can be found by solving

$$\boldsymbol{M}^T\boldsymbol{M}\boldsymbol{Q} = \boldsymbol{M}^T\boldsymbol{V} \tag{4.38}$$

and the constrained least-squares fitting has the matrix form

$$\begin{bmatrix}\boldsymbol{M}^T\boldsymbol{M} & \boldsymbol{K}^T\\ \boldsymbol{K} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q}\\ \boldsymbol{\lambda}\end{bmatrix} = \begin{bmatrix}\boldsymbol{M}^T\boldsymbol{V}\\ \boldsymbol{L}\end{bmatrix} \tag{4.39}$$

The current *PyRESP* implementation uses two separate restraining strength for permanent charges and permanent dipoles, which can be set to different values according to users' preferences.

4.2.4. Intra- and Inter-Molecular Equivalence

A reliable force field would require atoms sharing equivalent chemical environments to have identical permanent charges and dipoles. Taking a methyl group as an example, all three hydrogens must have the same charge, and all permanent dipoles pointing from methyl carbon towards hydrogens (and those in reverse directions) must have the same magnitudes, otherwise rotating the methyl to the three degenerate rotamers would give rise to different energies. Intra-molecular equivalencing are applied for this symmetry purpose. One strategy examined by previous studies is by averaging the charges of equivalent atom after the fitting, which were set free to change during the fitting process. However, this so called *a posteriori* strategy was found to have an unsatisfying negative impact on the fitting quality and on the final molecular dipole moments.[43] Thus, the *PyRESP* program employs the improved approach proposed by the original RESP work that performs equivalencing during the fitting process. Depending on the specific electrostatic model selected, the preliminary matrices in eq. 4.18, 4.28 or 4.39 are generated as if there were no equivalent fitting centers. Then, the rows and columns of corresponding equivalent fitting centers were added up to form a single row and column, giving rise to smaller linear equation systems to be solved as usual.

In comparison, inter-molecular equivalencing are often used for fitting one set of parameters for multiple conformations of the same molecule to further reduce the conformation-dependent problem, in addition to applying restraints. Alternatively, it can also be used for fitting the same chemical groups in different molecules. Both intra- and inter-molecular charge equivalencing have already been implemented in the original *RESP* program.[43] In *PyRESP*, the equivalencing algorithm is extended so that both intra- and inter-molecular equivalencing are enabled for permanent charges and dipoles in consistent manner.

4.2.5. Polarization Catastrophe Avoidance

A well-known problem of the point dipole model discussed so far is that it may lead to infinite molecular polarizability by the cooperative interaction between two induced dipoles, known as "polarization catastrophe".[5, 29] One way to avoid this problem is to turn off the polarization interactions between 1-2 and 1-3 interacting atoms pairs, as were done in the AMBER ff02 and ff02pol.rl force fields.[17-18] This can be easily achieved by setting corresponding elements in the charge-electric field coefficient matrix $\boldsymbol{C}$ and the dipole-electric field coefficient matrix $\boldsymbol{D}$ to zero. Alternatively, one can apply distance-dependent damping functions on interacting atom pairs, such as those developed by Thole,[30-31] and the pGM scheme developed by Elking et al.,[33] which will lead to the damped dipole field tensor

$$\boldsymbol{T}_{ij} = \frac{f_e}{{r_{ij}}^3}\boldsymbol{I} - \frac{3f_t}{{r_{ij}}^5}\begin{bmatrix} x^2 & xy & xz \\ xy & y^2 & yz \\ xz & yz & z^2 \end{bmatrix} \tag{4.40}$$

with screening functions f_e and f_t. Consequently, the charge-electric field coefficient matrix $\boldsymbol{C}$ and the dipole-electric field coefficient matrix $\boldsymbol{D}$ will also contain elements damped by f_e and f_t correspondingly. It is easy to see that for the original undamped Applequist model, f_e and f_t are constants

$$f_e = 1.0; f_t = 1.0 \tag{4.41}$$

For the linear model, we have

$$v = u/a$$

$$f_e = \begin{cases} 4v^3 - 3v^4 & v < 1 \\ 1.0 & v \geq 1 \end{cases} \tag{4.42}$$

$$f_t = \begin{cases} v^4 & v < 1 \\ 1.0 & v \geq 1 \end{cases}$$

For the exponential model, we have

$$v = au$$

$$f_e = 1 - \left(\frac{v^2}{2} + v + 1\right)\exp(-v) \tag{4.43}$$

$$f_t = 1 - \left(\frac{v^3}{6} + \frac{v^2}{2} + v + 1\right)\exp(-v)$$

For the Tinker-exponential model, we have

$$v = au^3$$

$$f_e = 1 - \exp(-v) \tag{4.44}$$

$$f_t = 1 - (v + 1)\exp(-v)$$

For the pGM model, we have

$$S_{ij} = \frac{\beta_i \beta_j r_{ij}}{\sqrt{2\left(\beta_i{}^2 + \beta_j{}^2\right)}}$$

$$f_e = \mathrm{erf}\left(S_{ij}\right) - \frac{2}{\sqrt{\pi}} S_{ij} \exp\left(-S_{ij}{}^2\right) \tag{4.45}$$

$$f_t = \mathrm{erf}\left(S_{ij}\right) - \frac{2}{\sqrt{\pi}} S_{ij} \exp\left(-S_{ij}{}^2\right)\left(1 + \frac{2}{3} S_{ij}{}^2\right)$$

$$f_0 = \mathrm{erf}\left(S_{ij}\right)$$

Note that for the pGM model, the charge-ESP interaction matrix $\boldsymbol{X}$ and the dipole-ESP interaction matrix $\boldsymbol{Y}$ should be scaled by f_0 and f_e, respectively, in addition to modifying the dipole field tensor $\boldsymbol{T}_{ij}$.

In the current *PyRESP* release, both polarization catastrophe avoidance strategies have been implemented, including turning off 1-2 and 1-3 interactions, and the four damping schemes (linear, exponential, Tinker-exponential, and pGM schemes).

4.3. Computational Details

4.3.1. Quantum Mechanical Calculations

Several molecules were selected as candidates for testing the *PyRESP* program, including water, methanol (alcohol), ethane (aliphatic), benzene (aromatic), *N*-methyl acetamide (peptide backbone), dimethyl phosphate (nucleic acid backbone), adenine (nucleobase), alanine dipeptide (hydrophobic amino acid), serine dipeptide (polar amino

acid), arginine dipeptide (positively charged amino acid), aspartic acid dipeptide (negatively charged amino acid). For the seven non-amino acid molecules, single-conformation fittings were performed. For the four amino acid molecules, both single-conformation and double-conformation fittings were performed, with the mainchain torsion angles in ($\phi = 300°, \psi = 300°$) and ($\phi = 240°, \psi = 120°$), approximating α-helix and antiparallel β-sheet secondary structure conformations. The geometries of all molecules were optimized at B3LYP/6-311++G(d, p) level of theory, with dihedral angle constraints applied to corresponding amino acid molecules only.

QM ESP values were calculated at MP2/aug-cc-pVTZ level of theory for a set of points fixed in space in the solvent-accessible region around each molecule. The points were generated using the method developed by Singh et al. on molecular surfaces (with a density of 6 points/Å^2) at each of 1.4, 1.6, 1.8 and 2.0 times the van der Waals radii.[39-40] For small molecules such as water, approximately 1800 points were generated, while for large molecules such as arginine dipeptide , more than 9000 points were generated. All quantum mechanical calculations were performed using the Gaussian 09 software.[50]

4.3.2. Parameterizations

A two-stage parameterization procedure has been adopted as the standard approach for RESP parameterization.[43] We extended this procedure for all electrostatic models: RESP, RESP-ind, and RESP-perm (and RESP-perm-v for water molecule), where the hyperbolic function in **eq. 4.12** was applied in all parametrizations. In the first stage, all fitting centers (permanent charges for all models, and permanent dipoles for RESP-perm and RESP-perm-

v) were set free to change, and a weak restraining strength 0.0005 (a in **eq. 4.12**) was applied to all fitting centers. In the second stage, intra-molecular equivalencing was enforced on all fitting centers that share identical chemical environment with others, such as methyl and methylene hydrogens; A stronger restraining strength 0.001 was applied to those fitting centers, and all other fitting centers were set frozen to keep the values obtained from the first stage. The restraints were only applied to non-hydrogen heavy atoms. To get better fitting results, the only Lagrange constraint enforced during parameterization is the total charge constraint, without applying additional intra-molecular charge constraints. Inter-molecular equivalencing was enforced in both the first and the second stages for double-conformation fittings of amino acid molecules.

Previous studies have shown that in the polarizable models with Thole-like damping schemes, it is important to include all atomic pair interactions to have anisotropic molecular response.[36, 51] Therefore, for parameterizations of the RESP-ind, RESP-perm and RESP-perm-v models, both 1-2 and 1-3 polarization interactions were included, and the pGM damping scheme were applied to all models to avoid polarization catastrophe.[33-34] The isotropic atomic polarizabilities derived in the previous work were employed for models considering polarization effects.[36]

The performance of each electrostatic model was evaluated based on the relative root mean square (RRMS) error,[38, 43, 49] given by

$$RRMS = \sqrt{\frac{\sum_{j=1}^{m}\left(V_j^{QM} - V_j\right)^2}{\sum_{j=1}^{m} {V_j^{QM}}^2}} \tag{4.46}$$

The molecular dipole moments and quadrupole moments along the principal axes calculated with each electrostatic model were compared with those calculated using *ab initio* methods as an additional metric in evaluating parameterization results. The Pearson correlation analysis was performed using the Python package *Scipy*. The scatterplots for QM ESPs and ESPs calculated by electrostatic models are plotted using the Python package *Matplotlib*.

4.4. Results

4.4.1. Water

The first molecule we tested is the water molecule. **Table 4.1** shows the parameterization results, RRMS and moments of the water molecule fitted with the RESP, RESP-ind, RESP-perm and RESP-perm-v electrostatic models, respectively. All models fit permanent point charges on oxygen and hydrogen atoms. In addition, the RESP-perm and RESP perm-v models also fit local frame permanent point dipole moments defined on CBVs, i.e., unit vectors along the direction of 1-2 interacting atom pairs (covalent bonds) or 1-3 interacting atom pairs (virtual bonds). For the RESP-perm model, a water molecule has two types of permanent dipoles: $\boldsymbol{p}_{OH}^{loc}$ and $\boldsymbol{p}_{HO}^{loc}$; while the RESP-perm-v model has one additional type of permanent dipole: $\boldsymbol{p}_{HH}^{loc}$, corresponding to the virtual CBV between the two hydrogen atoms. The permanent dipoles $\boldsymbol{p}_{OH}^{loc}$ and $\boldsymbol{p}_{HH}^{loc}$ have negative values, which means they point to the opposite direction of corresponding CBVs. That is, $\boldsymbol{p}_{OH}^{loc}$ points from the oxygen atom against the direction of the hydrogen atom, rather than the default CBV direction which points from oxygen towards hydrogen. Similarly, $\boldsymbol{p}_{HH}^{loc}$ points from the hydrogen atom

against the direction of the neighbor hydrogen atom, rather than the default CBV direction towards the neighbor hydrogen. **Figure 4.1** gives a better illustration of the parameterization results of local frame permanent dipole moments of water molecule. It can be observed that the RESP-perm and RESP-perm-v models produce higher magnitudes of permanent charges than the RESP and RESP-ind models. That is, they assign values to the charge centers in a more aggressive way to reproduce QM ESPs. All models assign negative charges to the oxygen atom, and positive charges to the hydrogen atom, and both the RESP-perm and RESP-perm-v models assign large but negative value to permanent dipole moments $\boldsymbol{p}_{OH}^{loc}$. This agrees with the fact that oxygen has higher electronegativity than hydrogen.

The RESP-perm model produces the lowest RRMS, with its RRMS only 19% of that of the RESP model, a factor of more than 5 folds reduction. The RESP-perm and RESP-perm-v models also produce molecular dipole moments and quadrupole moments with better agreement to the QM moments. The scatterplots of QM ESPs versus calculated ESPs for water are shown in **Figure 4.2**. The Pearson correlation coefficients of the RESP-perm and RESP-perm-v models are the highest among all models, and the RESP-ind model comes next. We can therefore conclude that electrostatic models with induced dipoles and permanent dipoles perform better than the RESP model in terms of all metrics analyzed.

The current RESP-perm-v model enables the virtual bonds between 1-3 interacting atom pairs. In theory, we can also enable virtual bonds between 1-4, 1-5 and atom pairs with even longer distances using a consistent method, giving rise to higher-level RESP-perm-v models. However, as can be seen from **Table 4.1** and **Figure 4.2**, the virtual bonds in the

RESP-perm-v model does not improve the fitting quality for the water molecule. In fact, adding too many virtual bonds may lead to the overfitting problem, and is expected to significantly increase the computational time for both parameterization and MD simulation processes. For these reasons, parametrization with the RESP-perm-v model will only be performed for the water molecule for illustration purpose, and other molecules will only be parameterized with the RESP, RESP-ind and RESP-perm models.

Table 4.1. Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Water Fitted with Four Electrostatic Models

	RESP	RESP-ind	RESP-perm	RESP-perm-v	QM
Charges/a.u.					
H	0.3401	0.5182	0.7576	0.7441	
O	-0.6802	-1.0365	-1.5151	-1.4882	
Permanent Dipole Moments/a.u.					
H-O[a]			0.0753	0.0773	
O-H[a]			-0.2761[b]	-0.2577	
H-H[a]				-0.0121	
RRMS					

	0.2051	0.1244	0.0391	0.0404	
Dipole Moments/Debye					
μ^{c}	1.9141	1.9417	1.8668	1.8660	1.8470
Quadrupole Moments/Debye Angstroms					
$Q_{xx}{}^{d}$	1.0444	1.5151	1.8549	1.8803	1.8389
$Q_{yy}{}^{d}$	-0.1858	-0.3198	-0.2467	-0.2841	-0.2418
$Q_{zz}{}^{d}$	-0.8586	-1.1953	-1.6082	-1.5962	-1.5971

[a] Each permanent dipole moment $\boldsymbol{p}_{AB}^{loc}$ is named in the format A-B, corresponding to the CBV points from atom A to atom B. [b] Negative value indicates pointing the reverse direction of CBV. [c] Dipole moment relative to center of mass. [d] Quadrupole moments along the principal axes.

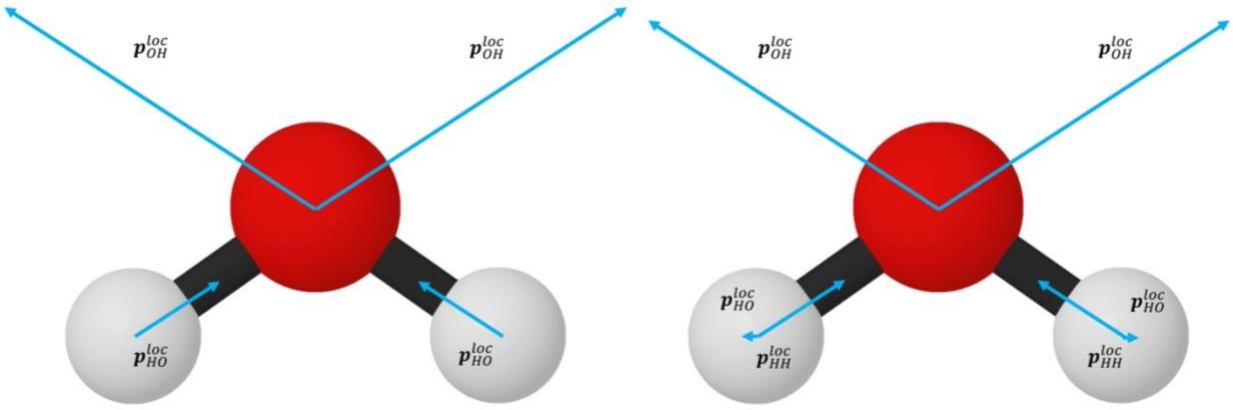


Figure 4.1. Schematic representation of local frame permanent dipole moments of water molecule fitted with RESP-perm (left) and RESP-perm-v (right) electrostatic models. The lengths of permanent dipole moments are shown in scale of their magnitudes. Refer to the text for detailed descriptions.

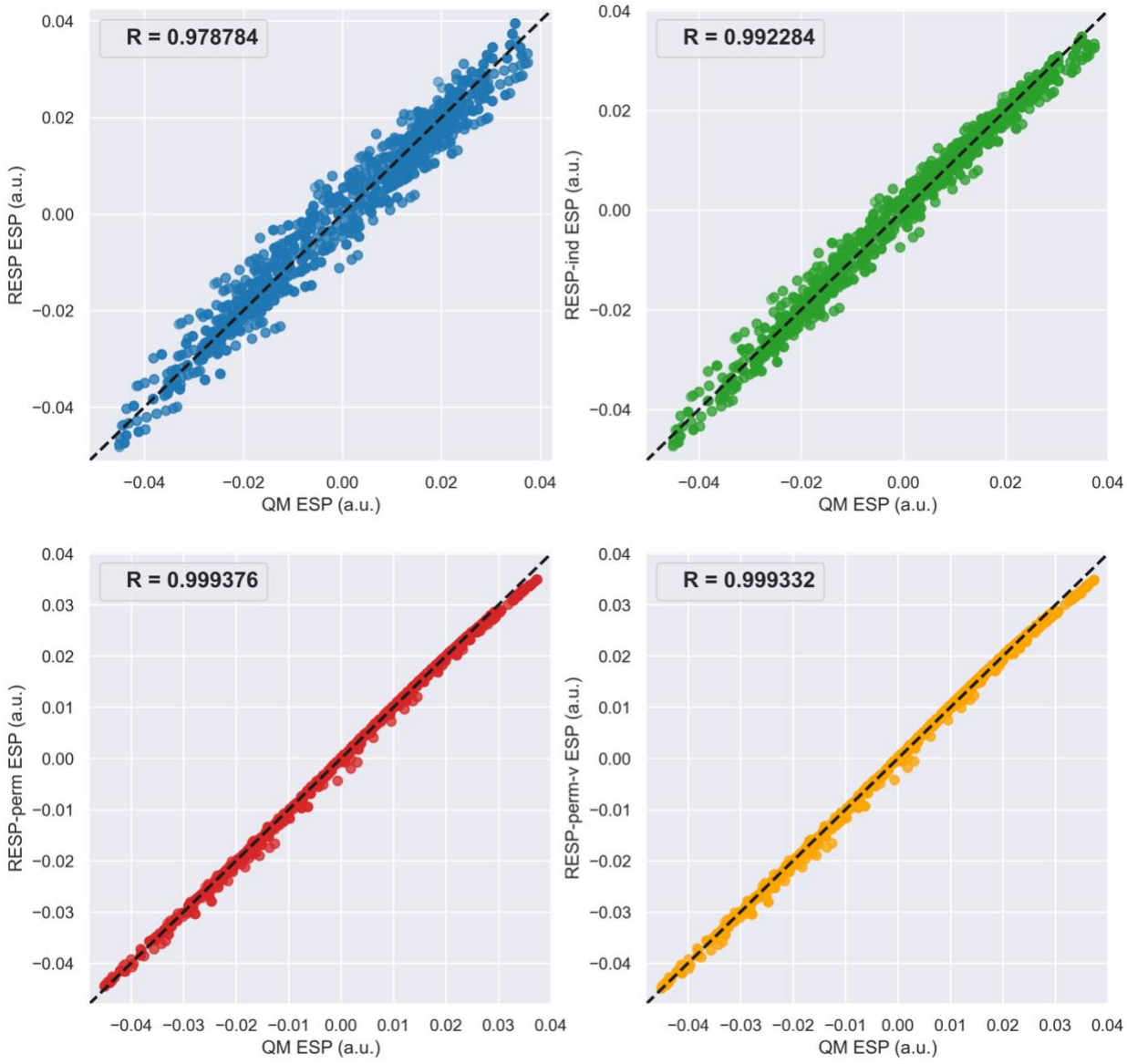


Figure 4.2. Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for water molecule, which was fitted with 1874 ESP data points. The dashed line corresponds to perfect correlation. *R* is the Pearson correlation coefficient.

4.4.2. Methanol, Ethane, and Benzene

We next extend our studies to the molecules methanol (CH_3OH), ethane (CH_3CH_3) and benzene (C_6H_6) to see how the parameterization results for these molecules differ from those for water. Methanol has lower symmetry than water, so it is of interest to see how electrostatic models parameterize this molecule. As shown in **Table 4.2**, all models assigned large negative charges to the highly electronegative oxygen atom and produced low RRMS and high correlation coefficients (**Figure 4.3**). In terms of molecular dipole and quadrupole moments, the RESP-perm model yields the best agreement with QM calculations among all three models. The results of methanol show the importance of induced and permanent dipoles for modeling polar molecules.

Table 4.2. Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Methanol Fitted with Three Electrostatic Models[a]

	RESP	RESP-ind	RESP-perm	QM
Charges/a.u.				
C	0.1609	0.1008	-0.0763	

H (methyl)	0.0194	0.0770	0.1105	
O	-0.6002	-0.8841	-1.0075	
H (hydroxyl)	0.3812	0.5524	0.7524	
Permanent Dipole Moments/a.u.				
C-H (methyl)			-0.0141	
H (methyl)-C			-0.0068	
C-O			0.0158	
O-C			0.1071	
O-H (hydroxyl)			-0.2268	
H (hydroxyl)-O			0.0973	
RRMS				
	0.2519	0.1298	0.0801	
Dipole Moments/Debye				
μ	1.9558	1.7563	1.6786	1.6873
Quadrupole Moments/Debye Angstroms				
Q_{xx}	2.2197	2.5574	2.6684	2.6984
Q_{yy}	-0.7640	-0.7275	-0.6935	-0.8281

Q_{zz}	-1.4557	-1.8299	-1.9749	-1.8703

[a] See **Table 4.1** for notation.

In the case of ethane, all models assign positive charges to hydrogen, and negative charges to carbon, as shown in **Table 4.3**. Among the three models, the RESP-ind model assigns charges with the highest magnitudes, and the RESP model assigns charges with the lowest magnitudes. Ethane is a non-polar molecule, as reflected by the molecular dipole moments calculated by all three models as well as QM calculations. However, the RESP-perm model significantly outperforms the RESP and the RESP-ind models in terms of all other metrics, including RRMS, quadrupole moments, and correlation coefficients, making it the only model that gives reasonable performance. As shown in **Figure 4.3**, the ESPs around the ethane molecule is very close to 0 a.u., with the range between -0.005 a.u. to 0.006 a.u., compared with that of polar molecules such as water (-0.045 a.u. to 0.04 a.u.) and methanol (-0.05 a.u. to 0.04 a.u.). The non-polar nature of ethane makes it particularly difficult to parameterize, so that models with high degree of freedom like RESP-perm perform significantly better than those with low degree of freedom.

Table 4.3. Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Ethane Fitted with Three Electrostatic Models[a]

	RESP	RESP-ind	RESP-perm	QM

Charges/a.u.				
C	-0.0254	-0.2148	-0.0723	
H	0.0085	0.0716	0.0241	
Permanent Dipole Moments/a.u.				
C-H			-0.0201	
C-C			0.0645	
H-C			-0.0787	
RRMS				
	0.9939	0.8808	0.3490	
Dipole Moments/Debye				
μ	0.0000	0.0000	0.0000	0.0000
Quadrupole Moments/Debye Angstroms				
Q_{xx}	0.0403	0.0457	-0.5761	-0.5050
Q_{yy}	-0.0201	-0.0229	0.2881	0.2525
Q_{zz}	-0.0201	-0.0229	0.2880	0.2524

[a] See **Table 4.1** for notation.

Table 4.4 shows the parameterization results, RRMS and moments of benzene. Similar to the ethane molecule, benzene is also a non-polar molecule, and the molecular dipole moment was successfully predicted by all three models. The RESP-ind model again fit charges most aggressively by assigning charges with highest magnitudes, and the RESP model fit charges most conservatively by assigning charges with lowest magnitudes. However, unlike the case of ethane, none of the models perform significantly better in terms of other metrics. The RESP model yields the lowest RRMS, but it is only 14% lower than the highest RRMS (given by the RESP-ind model). All models underestimate the molecular quadrupole moments, although those given by the RESP-ind model have better agreement with QM results than those of the other two models. As shown in **Figure 4.3**, the RESP-perm model has the highest correlation coefficient but is still lower than those for polar molecules such as water and methanol. Modeling aromatics such as benzene is therefore also a difficult task, possibly due to the existence of π orbital that are located outside of the 2-dimensional plane of the aromatics ring.

Table 4.4. Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Benzene Fitted with Three Electrostatic Models[a]

	RESP	RESP-ind	RESP-perm	QM
Charges/a.u.				
C	-0.1123	-0.2464	-0.2227	

H	0.1123	0.2464	0.2227	
Permanent Dipole Moments/a.u.				
C-H			0.0670	
H-C			0.0074	
C-C			-0.0290	
RRMS				
	0.2203	0.2570	0.2432	
Dipole Moments/Debye				
μ	0.0000	0.0000	0.0000	0.0000
Quadrupole Moments/Debye Angstroms				
Q_{xx}	2.2657	2.3738	2.3203	2.6637
Q_{yy}	2.2655	2.3732	2.3199	2.6627
Q_{zz}	-4.5312	-4.7470	-4.6403	-5.3264

[a] See **Table 4.1** for notation.

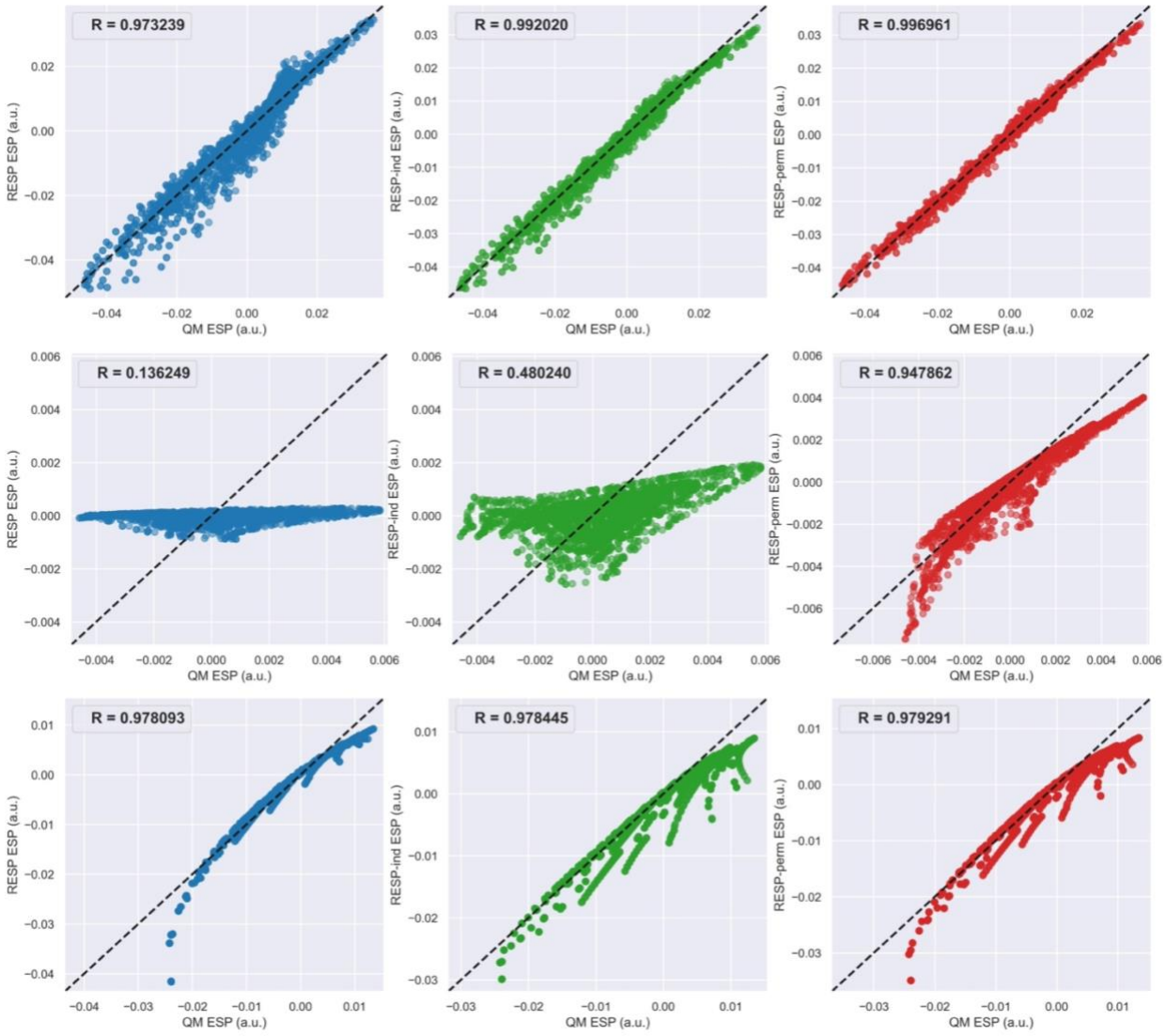


Figure 4.3. Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for methanol (upper panel), ethane (middle panel), and benzene (lower panel) molecules. Methanol, ethane, and benzene molecules were fitted with 2654, 2951 and 4130 ESP data points, respectively. The dashed lines correspond to perfect correlation. *R* is the Pearson correlation coefficient.

4.4.3. NMA, DMP and Adenine

We next turn to *N*-methyl acetamide (NMA), dimethyl phosphate (DMP), and adenine base. These molecules are chosen as they are common model compounds for peptides and nucleic acids. **Table 4.5** and **Table 4.6** show the charges, RRMS and moments of NMA and DMP, respectively, and the permanent dipole moments fitted with the RESP-perm model are shown in **Table S4.1** and **Table S4.2**. All models produce charge sets with consistent signs for NMA. Interestingly, there is significant variation in the atomic charges of DMP fitted by the three models. For example, the charges for the central phosphorus (P) range from -0.4188 a.u. to 1.1047 a.u. Low RRMS and high correlation coefficients (**Figure 4.4** and **Figure 4.5**) are yielded by all models. However, for both NMA and DMP molecules, the molecular dipole and quadrupole moments produced by the RESP-ind and RESP-perm models agree worse to the QM results than those of the RESP model, indicating the potential overfitting problem for the RESP-ind and RESP-perm models.

Table 4.5. Charges, RRMS and Molecular Dipole/Quadrupole Moments of *N*-methyl Acetamide (NMA) Fitted with Three Electrostatic Models[a]

	RESP	RESP-ind	RESP-perm	QM
Charges/a.u.				
C1	-0.4202	-0.3524	-0.4778	
H1	0.1113	0.1422	0.1347	
C	0.6515	1.1283	1.0510	

O	-0.5297	-0.9953	-0.8081	
N	-0.4249	-1.1062	-0.5250	
H	0.2848	0.6127	0.2715	
C2	-0.3419	-0.1267	-0.1219	
H2	0.1488	0.1377	0.0687	
RRMS				
	0.1029	0.0812	0.0786	
Dipole Moments/Debye				
μ	3.8335	3.6657	3.6502	3.8004
Quadrupole Moments/Debye Angstroms				
Q_{xx}	3.6515	3.1427	3.4849	3.6815
Q_{yy}	-0.7200	-0.3841	-0.6802	-0.7850
Q_{zz}	-2.9315	-2.7586	-2.8047	-2.8965

[a] See **Table 4.1** for notation.

Table 4.6. Parameterization Results, RRMS and Molecular Dipole/Quadrupole Moments of Dimethyl Phosphate (DMP) Fitted with Three Electrostatic Models[a]

	RESP	RESP-ind	RESP-perm	QM
Charges/a.u.				
P	1.1047	0.5525	-0.4188	
O1 (O=)	-0.7411	-0.6776	-0.3424	
O2 (-O-)	-0.4399	-0.4920	-0.1201	
C	0.0553	0.0987	-0.2107	
H	0.0244	0.0982	0.1276	
RRMS				
	0.0196	0.0161	0.0117	
Dipole Moments/Debye				
μ	2.4333	2.4494	2.4635	2.5559
Quadrupole Moments/Debye Angstroms				
Q_{xx}	9.2617	7.5526	8.3254	9.0420
Q_{yy}	-3.5225	-2.8853	-3.4178	-3.6665
Q_{zz}	-5.7392	-4.6673	-4.9076	-5.3755

[a] See **Table 4.1** for notation.

The charges, RRMS and moments of the nucleic acid base adenine are shown in **Table 4.7**, and the permanent dipole moments fitted with the RESP-perm model are shown in **Table S4.3**. Among the three electrostatic models, RESP-ind assigns charges with the highest magnitude to most atoms, but results in the worst RRMS, molecular dipole moment agreement, and correlation coefficient. On the other hand, the RESP-perm model yields the lowest RRMS, dipole and quadrupole moments with best agreements, and highest correlation coefficient (**Figure 4.4**). Therefore, permanent dipole moments are necessary components for modeling the adenine molecule.

Table 4.7. Charges, RRMS and Molecular Dipole/Quadrupole Moments of Adenine Fitted with Three Electrostatic Models[a]

	RESP	RESP-ind	RESP-perm	QM
Charges/a.u.				
N1[b]	-0.7086	-2.0082	-0.0586	
C2[b]	0.4549	1.6084	-0.1038	
H2[b]	0.0770	0.3283	0.0701	
N3[b]	-0.7256	-2.5767	-0.1907	
C4[b]	0.6413	2.4364	0.1796	
C5[b]	0.0209	0.0431	0.1477	

C6[b]	0.6856	2.2390	0.4396	
N6[b]	-0.9046	-2.2041	-1.4981	
HN6[b]	0.4054	0.7019	0.5695	
N7[b]	-0.5608	-1.7370	-0.0397	
C8[b]	0.2693	1.2954	-0.0643	
H8[b]	0.1199	0.4007	-0.1734	
N9[b]	-0.5699	-1.9989	-0.3756	
HN9[b]	0.3898	0.7698	0.5283	
RRMS				
	0.1263	0.1661	0.1043	
Dipole Moments/Debye				
μ	2.5562	2.5856	2.4726	2.4994
Quadrupole Moments/Debye Angstroms				
Q_{xx}	12.3287	12.0435	12.5849	12.7410
Q_{yy}	-5.7358	-6.0081	-5.6209	-6.0143
Q_{zz}	-6.5930	-6.0354	-6.9640	-6.7266

[a] See **Table 4.1** for notation. [b] The atom names are from the adenine obtained from Protein Data Bank (ligand ID: ADE).

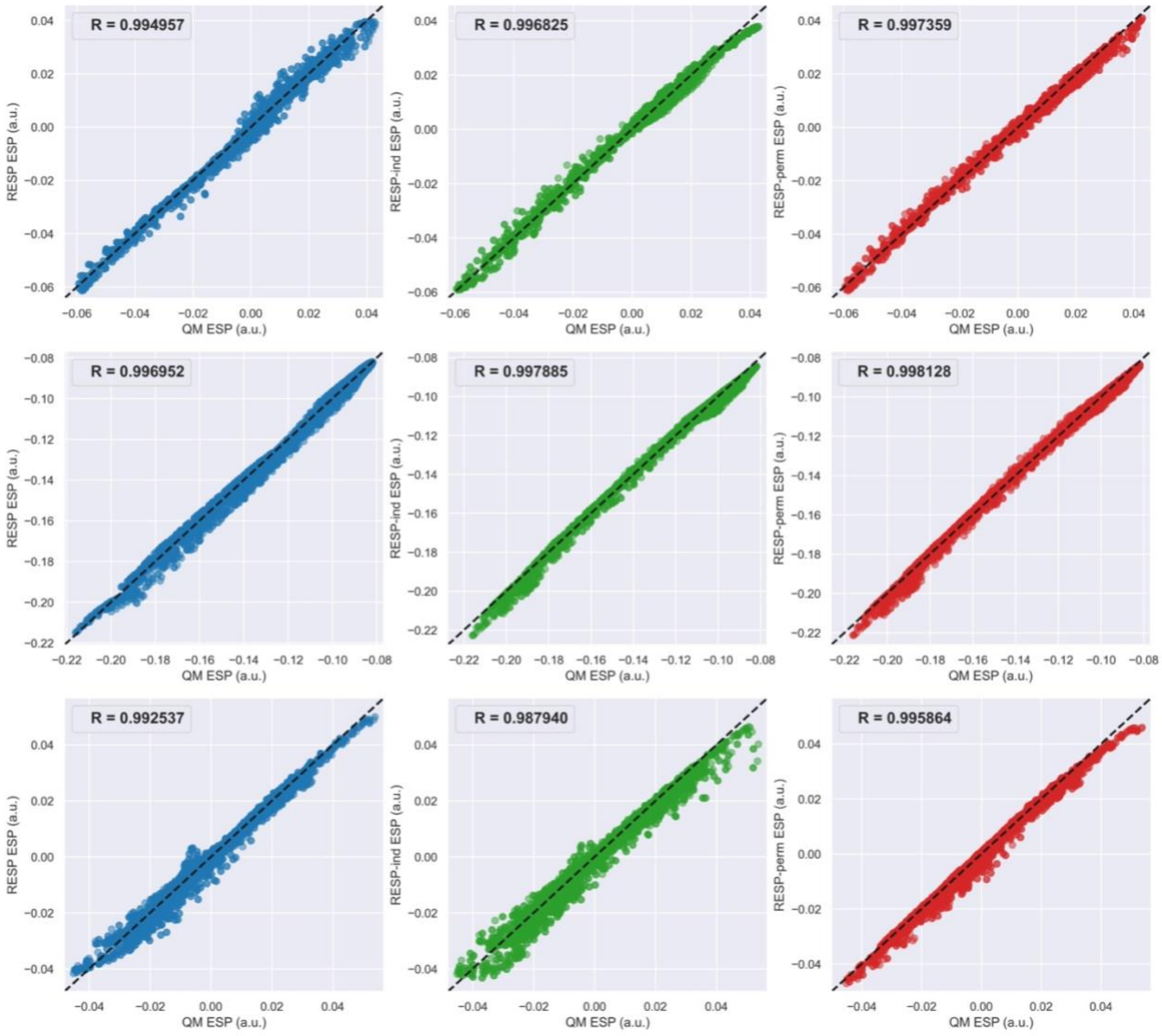


Figure 4.4. Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for *N*-methyl acetamide (NMA, upper panel), dimethyl phosphate (DMP, middle panel) and adenine (lower panel) molecules. NMA, DMP, and adenine molecules were fitted with 4159, 4847 and 5155 ESP data points, respectively. The dashed lines correspond to perfect correlation. *R* is the Pearson correlation coefficient.

4.4.4. Amino Acid Dipeptides

PyRESP was designed as the next generation parameterization tool for polarizable force field development, with the aim to replace its ancestor *RESP* program.[43-44] Amino acids are key molecules for force field development for biomacromolecules, so we next tested the program on several amino acid dipeptides, all capped with *N*-acetyl (ACE) group at the N-terminal, and *N*-methylamide (NME) group at the C-terminal. Selected amino acids include alanine (hydrophobic amino acid), serine (polar amino acid), arginine (positively charged amino acid) and aspartic acid (negatively charged amino acid). Two conformations, approximating α-helix ($\phi = 300°, \psi = 300°$) and antiparallel β-sheets ($\phi = 240°, \psi = 120°$), were used for both single-conformation and double-conformation fittings. Double-conformation fittings were performed with inter-molecular equivalencing applied. For single-conformation fittings, we would like to examine both the differences and consistencies of the parameterizations between the two conformations, and we are interested in which electrostatic model can give the best performance in parameterizing each amino acid. For double-conformation fittings, it can be expected that they will show higher RRMS and lower correlation coefficients compared to single-conformation fittings, since the double-conformation fitting needs to accommodate contributions from both conformations to reduce conformational dependence.

Table 4.8-4.11 show the RRMS and moments of alanine dipeptide, serine dipeptide, arginine dipeptide, and aspartic acid dipeptide, respectively, fitted with both single-conformation and double-conformation fittings. We first focus on the results for single-conformation fittings. For uncharged amino acids alanine and serine, the lowest RRMS is

produced by the RESP-perm model for the α-helix conformation, and by the RESP-ind model for the β-sheet conformation. While for charged amino acids arginine and aspartic acid, the RRMS consistently decreases in the order of RESP, RESP-ind and RESP-perm models for both α-helix and β-sheet conformations. In addition, most α-helix conformation fittings give lower RRMS than that of β-sheet conformation, which might be explained by the fact that amino acids in the α-helix conformation have higher polarity (larger dipole moment) than in the β-sheet conformation. Similar trend was observed in **Figure 4.5** and **Figure S4.1-4.3**, where the correlation coefficients for the α-helix conformation are mostly higher than that of the β-sheet conformation. The correlation coefficients of single-conformation fittings consistently increase in the order of RESP, RESP-ind and RESP-perm models for all amino acids in both conformations. The molecular dipole and quadrupole moments show interesting patterns. The RESP-ind model consistently yields the best agreement with QM moments for amino acids in the α-helix conformation. On the other hand, the RESP model yields the worst agreement for the α-helix conformation but yields the best agreement for the β-sheet conformation.

Next, we compare the results of double-conformation fittings with those of single-conformation fittings. Surprisingly, in contrast to the expectation that double-conformation fittings will always produce higher RRMS and lower correlation coefficients compared to single-conformation fittings, the double-conformation fittings of the RESP-perm model consistently give lower RRMS and higher correlation coefficients than those of single-conformation fittings for all amino acids in both conformations, so is the RESP model for amino acids in the α-helix conformation. Next, the molecular dipole and quadrupole

moments of double- and single- conformation fittings are compared. Interestingly, most double-conformation fittings result in better agreement with the QM calculated moments than those of single-conformation fittings for the α-helix conformation but result in worse agreements for the β-sheet conformation. In particular, the RESP-perm model is the only model that improve the molecular moment qualities for all amino acids in both α-helix and β-sheet conformations.

Table 4.8. RRMS and Molecular Dipole/Quadrupole Moments of Alanine Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models[a]

		Single-Conformation Fitting			Double-Conformation Fitting			
Conformation		RESP	RESP-ind	RESP-perm	RESP	RESP-ind	RESP-perm	QM
RRMS								
α-helix		0.0929	0.0551	0.0552	0.0854	0.0602	0.0432	
β-sheet		0.1210	0.0852	0.0870	0.1431	0.0939	0.0732	
Dipole Moments/Debye								
α-helix	μ	7.2117	6.9530	6.8602	7.1200	6.9617	6.9060	7.0313
β-sheet	μ	0.7805	0.6641	0.6809	0.7759	0.6016	0.6166	0.6963

Quadrupole Moments/Debye Angstroms								
α-helix	Q_{xx}	8.5529	7.5041	7.4211	8.3689	7.8608	7.9172	8.1763
	Q_{yy}	-0.8103	0.1479	0.7630	-0.3067	-0.2380	0.1786	0.1868
	Q_{zz}	-7.7425	-7.6519	-8.1841	-8.0622	-7.6229	-8.0958	-8.3630
β-sheet	Q_{xx}	14.6902	14.1491	14.1046	13.8200	14.0687	14.0705	14.9055
	Q_{yy}	3.9444	3.4097	3.1843	3.7454	3.4612	3.5775	3.4394
	Q_{zz}	-18.6346	-17.5588	-17.2889	-17.5654	-17.5299	-17.6481	-18.3449

[a] See **Table 4.1** for notation.

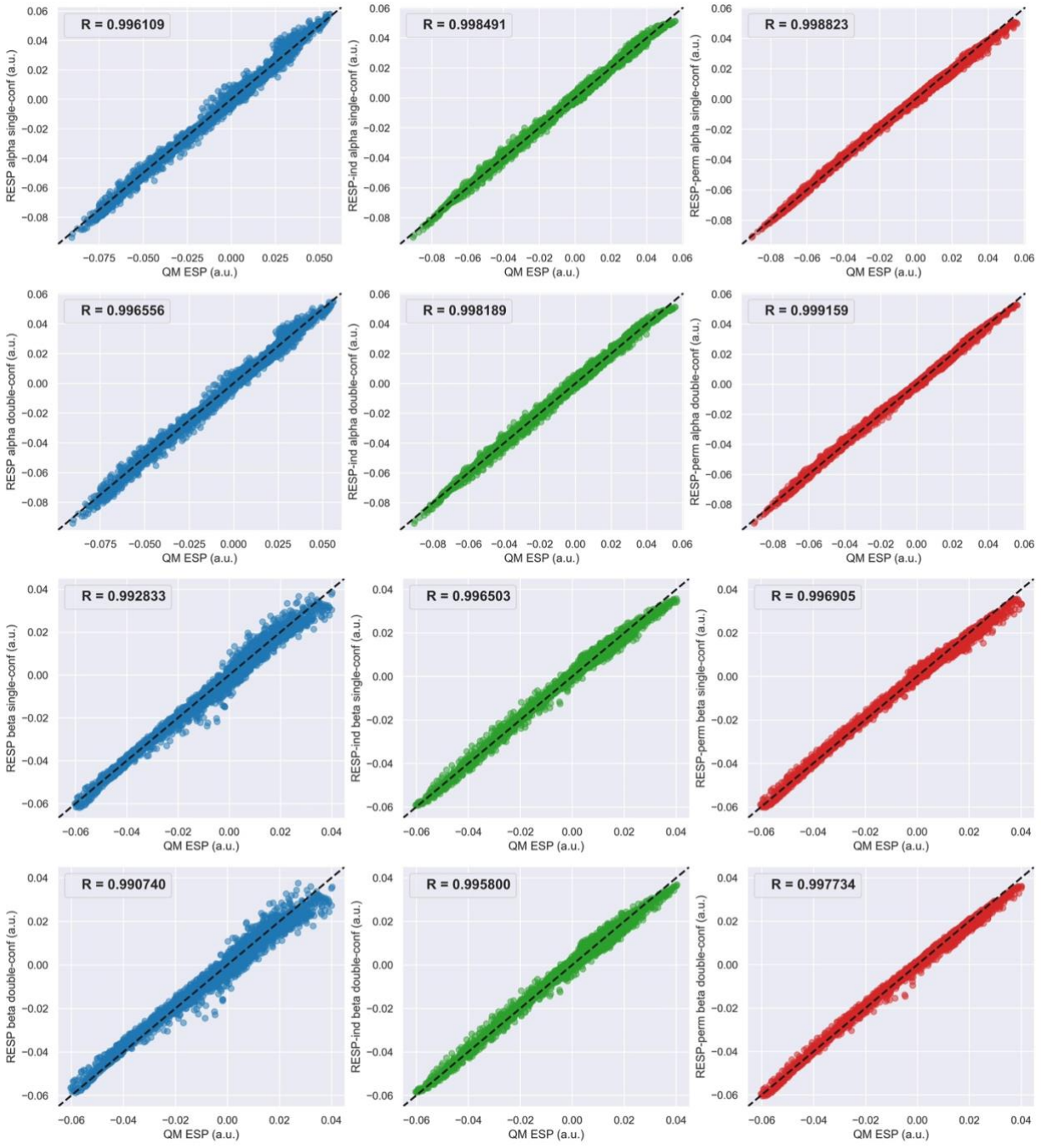


Figure 4.5. Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for alanine dipeptide using single- and double-conformation fittings. 1st row: α-helix conformation fitted with single-conformation; 2nd row: α-helix conformation fitted with double-conformation; 3rd row: β-sheet conformation fitted with single-conformation; 4th

row: β-sheet conformation fitted with double-conformation. The α-helix conformation was fitted with 6292 ESP data points, and the β-sheet conformation was fitted with 6460 ESP data points. The dashed lines correspond to perfect correlation. *R* is the Pearson correlation coefficient.

Table 4.9. RRMS and Molecular Dipole/Quadrupole Moments of Serine Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models[a]

		Single-Conformation Fitting			Double-Conformation Fitting			
Conformation		RESP	RESP-ind	RESP-perm	RESP	RESP-ind	RESP-perm	QM
RRMS								
α-helix		0.1092	0.0583	0.0544	0.1015	0.0638	0.0456	
β-sheet		0.1169	0.0719	0.0768	0.1283	0.0800	0.0627	
Dipole Moments/Debye								
α-helix	μ	7.2984	7.0225	6.8966	7.1918	7.0907	6.9997	7.0311
β-sheet	μ	1.6728	1.7070	1.6607	1.6197	1.6176	1.6159	1.6838
Quadrupole Moments/Debye Angstroms								

α-helix	Q_{xx}	4.9007	4.6936	5.0199	4.8222	4.4699	4.6349	4.5426
	Q_{yy}	3.1930	3.3143	3.2288	3.3371	3.6943	3.6715	3.9228
	Q_{zz}	-8.0937	-8.0079	-8.2487	-8.1593	-8.1642	-8.3065	-8.4653
β-sheet	Q_{xx}	14.1477	13.2178	13.0126	13.0852	13.1921	13.3675	14.0962
	Q_{yy}	6.7942	6.8604	6.8887	6.8816	6.7819	6.7311	6.5504
	Q_{zz}	-20.9419	-20.0782	-19.9014	-19.9668	-19.9740	-20.0986	-20.6466

[a] See **Table 4.1** for notation.

Table 4.10. RRMS and Molecular Dipole/Quadrupole Moments of Arginine Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models[a]

		Single-Conformation Fitting			Double-Conformation Fitting			
Conformation		RESP	RESP-ind	RESP-perm	RESP	RESP-ind	RESP-perm	QM
RRMS								

α-helix		0.0236	0.0163	0.0133	0.0256	0.0176	0.0129	
β-sheet		0.0185	0.0164	0.0148	0.0226	0.0177	0.0128	
Dipole Moments/Debye								
α-helix	μ	24.6060	24.6416	24.5407	24.5512	24.5354	24.4555	24.4666
β-sheet	μ	17.0782	17.2996	17.2833	17.2463	17.4137	17.4077	17.0905
Quadrupole Moments/Debye Angstroms								
α-helix	Q_{xx}	70.9219	71.3031	71.9070	70.4120	71.3074	71.3896	71.3708
	Q_{yy}	-26.1898	-25.8375	-26.0332	-25.4176	-25.5352	-25.5844	-25.8057
	Q_{zz}	-44.7321	-45.4656	-45.8738	-44.9944	-45.7722	-45.8052	-45.5651
β-sheet	Q_{xx}	79.1043	79.2100	79.3335	79.4498	79.1450	79.5268	79.4586

	Q_{yy}	-27.5596	-28.3617	-28.3538	-28.8008	-28.7891	-28.5127	-27.9418
	Q_{zz}	-51.5447	-50.8484	-50.9797	-50.6490	-50.3559	-51.0142	-51.5168

[a] See **Table 4.1** for notation.

Table 4.11. RRMS and Molecular Dipole/Quadrupole Moments of Aspartic Acid Dipeptide (Single- and Double- Conformation) Fitted with Three Electrostatic Models[a]

		Single-Conformation Fitting			Double-Conformation Fitting			
Conformation		RESP	RESP-ind	RESP-perm	RESP	RESP-ind	RESP-perm	QM
RRMS								
α-helix		0.0238	0.0159	0.0126	0.0232	0.0164	0.0118	
β-sheet		0.0253	0.0156	0.0134	0.0259	0.0162	0.0125	
Dipole Moments/Debye								

α-helix	μ	10.1754	9.8314	9.8469	10.0132	9.7958	9.8477	9.9939
β-sheet	μ	9.3896	9.2885	9.2847	9.5458	9.3331	9.2629	9.4228
Quadrupole Moments/Debye Angstroms								
α-helix	Q_{xx}	21.5346	22.1268	22.4423	21.8971	22.0499	22.4577	22.6697
	Q_{yy}	16.9975	16.1860	15.7439	15.9289	16.2698	16.0475	16.3449
	Q_{zz}	-38.5321	-38.3128	-38.1862	-37.8260	-38.3197	-38.5053	-39.0145
β-sheet	Q_{xx}	27.8352	26.9809	26.9421	27.8597	26.9899	27.0867	27.8433
	Q_{yy}	-4.5477	-3.2223	-3.2344	-4.3845	-3.0961	-3.2715	-3.5950
	Q_{zz}	-23.2874	-23.7587	-23.7077	-23.4752	-23.8938	-23.8152	-24.2483

[a] See **Table 4.1** for notation.

4.5. Discussion and Conclusions

We have developed and implemented the *PyRESP* program for flexible force field parameterizations with four electrostatic models: RESP, RESP-ind, RESP-perm and RESP-perm-v. The RESP model is a Python implementation of the original *RESP* program in the Fortran language.[43-44] Compared with previous ESP-based charge derivation methods,[37-39, 47-48] the RESP model reduces the overall magnitude of the charges using a simple hyperbolic restraining function, which improves the transferability of fitted charges and reduces the conformational dependency problem. The RESP-ind, RESP-perm and RESP-perm-v models were designed and implemented in a consistent manner as the RESP model, with the additional modeling of atomic induced dipole moments, atomic permanent dipole moments and atomic permanent virtual dipole moments, respectively. The Lagrange constraints as well as the intra- and inter- molecular equivalencing schemes developed in the original RESP work were also implemented for the latter three models in *PyRESP*.

A variety of molecules were tested with various electrostatic models implemented in *PyRESP*. All molecules were parameterized using the standard two-stage approach proposed by the original RESP work.[43] The 1-2 and 1-3 interactions were included for all polarizable models, and the pGM damping function was applied to all electrostatic interactions both to avoid the polarization catastrophe and to achieve adequate anisotropic molecular response.[33-34, 36] It can be observed that for each molecule, most charges fitted with the RESP-ind model have higher magnitude than those of the RESP model. This is due to the polarization effect among atoms. Taking the water molecule as an example, the electric field at the position of the oxygen atom caused by the positively charged hydrogen atom points

outside the molecule along the symmetric axis, which generates an induced dipole in the same direction. The dipole generates positive ESP at the outward direction of oxygen atom, which cancels out certain amounts of ESP caused by the negatively charged oxygen atom. To compensate this effect, a negative charge with higher magnitude was fitted to the oxygen atom. On the other hand, the magnitudes of charges fitted by the RESP-perm model does not show consistent trend when compared to those of the RESP-ind model. The charges with the RESP-perm model have higher magnitudes than those of the RESP-ind model for the water molecule, but the opposite is true for ethane and benzene molecules. The magnitude of charges with the RESP-perm model is directly affected by the directions of induced dipole moments and permanent dipole moments. If they point to the same direction, the charge magnitude will increase to compensate the combined effects of induced and permanent dipole moments. If they point to opposite directions, the cancel-off effect of polarization becomes weaker, leading to lower magnitude of charges.

Among the molecules tested in this chapter, the parameterizations of ethane molecule resulted in the highest RRMS and lowest correlation coefficients. This is not only because of its non-polar nature, but also because of the fact that it contains only weak electronegative elements carbon and hydrogen. **Figure 4.3** shows that the ESPs around ethane molecule is very close to 0 a.u., with the range between -0.005 a.u. to 0.006 a.u. The low magnitude of ESP makes the parametrization process sensitive to noise, so that models with high degree of freedom like RESP-perm are needed to give reasonable fitting. Another molecule that none of the model gave satisfactory performances is benzene, also a non-polar molecule. The difficulty for parameterizing benzene likely comes from the existence of π orbital lying outside the ring plane, which cannot be modeled adequately even with the induced and

permanent dipole moments, since they are both located on the 2-dimensional plane. This is an inherent limitation of the current model, which may be improved by adding additional fitting centers outside the aromatic ring, or by fitting permanent quadrupole moments in addition to permanent charges and dipoles. Therefore, modeling aromatic molecules remains a challenge even for polarizable force field developments.

The RESP-perm model has higher degree of freedom than the RESP and RESP-ind models, due to the addition of permanent dipole moments; addition of virtual bonds increases the degree of freedom for the RESP-perm-v model even further. For most molecules tested here, the parameterizations with the RESP-perm/RESP-perm-v models resulted in lower RRMS, higher correlation coefficients, and molecular moments agree better with QM calculations. However, the quadrupole moments of methanol, NMA and DMP molecules fitted by the RESP-perm model clearly agree worse with QM results than those fitted by the RESP model. This raises the concern of overfitting problem when the model degree of freedom is so high that noise start to diminish fitting accuracy, leading to deteriorated overall fitting quality. Among the metrics used here to evaluate models, the RRMS and correlation coefficients are highly correlated with the objective function to be minimized in **eq. 4.13**, so that low RRMS and high correlation coefficients are not reliable enough to eliminate the concerns of overfitting. Therefore, while performing molecule parameterizations using electrostatic models with high degree of freedom, it is critical to inspect the final molecular dipole and quadrupole moments to determine if the overfitting occurred.

We tested several amino acid dipeptide molecules using both single- and double-conformation fittings. The α-helix ($\phi = 300°, \psi = 300°$) and antiparallel β-sheet ($\phi = 240°, \psi = 120°$) conformations were selected since they are two of the most frequently found conformations for amino acids in proteins, and they represent considerably different electrostatic properties (e.g., notably different dipole moments). For single-conformation fittings, the RESP-ind model consistently yields the best agreement with QM moments for amino acids in the α-helix conformation, while the RESP model yields the best agreement for the β-sheet conformation. The RESP-perm model that has the highest degree of freedom shows the lowest RRMS and highest correlation coefficients but does not outperform other models in terms of reproducing QM molecular moments. Double-conformation fittings were expected to have poorer performances than those of single-conformation fittings. Surprisingly, double-conformation fittings with the RESP-perm model consistently shows better overall performances than the single-conformation fittings for amino acids in both conformations, as illustrated by the lower RRMS, higher correlation coefficients, and moments agree better with QM results. This shows that the double-conformation fittings are necessary for amino acids fitted with the RESP-perm model. For future polarizable force field parameterizations, more conformations are expected to be included to further reduce conformational dependence of the parameters.

In conclusion, the *PyRESP* program developed here is a flexible, efficient, and user-friendly tool that is recommended for parameterizations of various additive and polarizable force fields. *PyRESP* has been released as an open-source software within AmberTools 2022 under the GNU General Public License, available for download from http://ambermd.org/.[52] Documentation and tutorials will also be made available on the Amber website.

Alternatively, the standalone version of *PyRESP* with the latest updates is available through https://github.com/ShijiZ/PyRESP.

4.6. Supporting Information

Table S4.1. Permanent Dipole Moments (a.u.) of *N*-methyl Acetamide (NMA) Fitted with the RESP-perm Electrostatic Model[a]

C1-H1	**H1-C1**	**C1-C**	**C-C1**	**C-O**
0.0175	-0.0534[b]	-0.1356	0.0323	-0.0335
O-C	**C-N**	**N-C**	**N-H**	**H-N**
0.1508	0.0031	-0.1423	0.0084	-0.2141
N-C2	**C2-N**	**C2-H2**	**H2-C2**	
0.1094	-0.0525	0.0331	-0.0490	

[a] Each permanent dipole moment is named in the format A-B, corresponding to the CBV points from atom A to atom B. [b] Negative value indicates pointing the reverse direction of CBV.

Table S4.2. Permanent Dipole Moments (a.u.) of Dimethyl Phosphate (DMP) Fitted with the RESP-perm Electrostatic Model[a]

P-O1	O1-P	P-O2	O2-P	O2-C
-0.0392	0.2779	0.0592	0.1353	0.2213
C-O2	**C-H**	**H-C**		
-0.0896	0.0399	0.0196		

[a] See **Table S4.1** for notation.

Table S4.3. Permanent Dipole Moments (a.u.) of Adenine Fitted with the RESP-perm Electrostatic Model[a]

N1-C2	C2-N1	C2-H2	H2-C2	C2-N3
0.5299	-0.0137	0.1008	-0.0403	-0.0547
N3-C2	**N3-C4**	**C4-N3**	**C4-C5**	**C5-C4**
0.3527	0.5373	-0.0655	0.1054	0.0567
C4-N9	**N9-C4**	**C5-C6**	**C6-C5**	**C5-N7**
-0.0044	0.0373	0.0339	0.1488	-0.1087
N7-C5	**C6-N1**	**N1-C6**	**C6-N6**	**N6-C6**
0.4430	0.0089	0.5776	-0.2238	-0.0288
N6-HN6	**HN6-N6**	**N7-C8**	**C8-N7**	**C8-H8**

0.0112	-0.1024	0.3586	-0.1499	0.4969
H8-C8	**C8-N9**	**N9-C8**	**N9-HN9**	**HN9-N9**
-0.2386	-0.0241	0.0349	-0.0692	-0.0229

[a] See **Table S4.1** for notation.

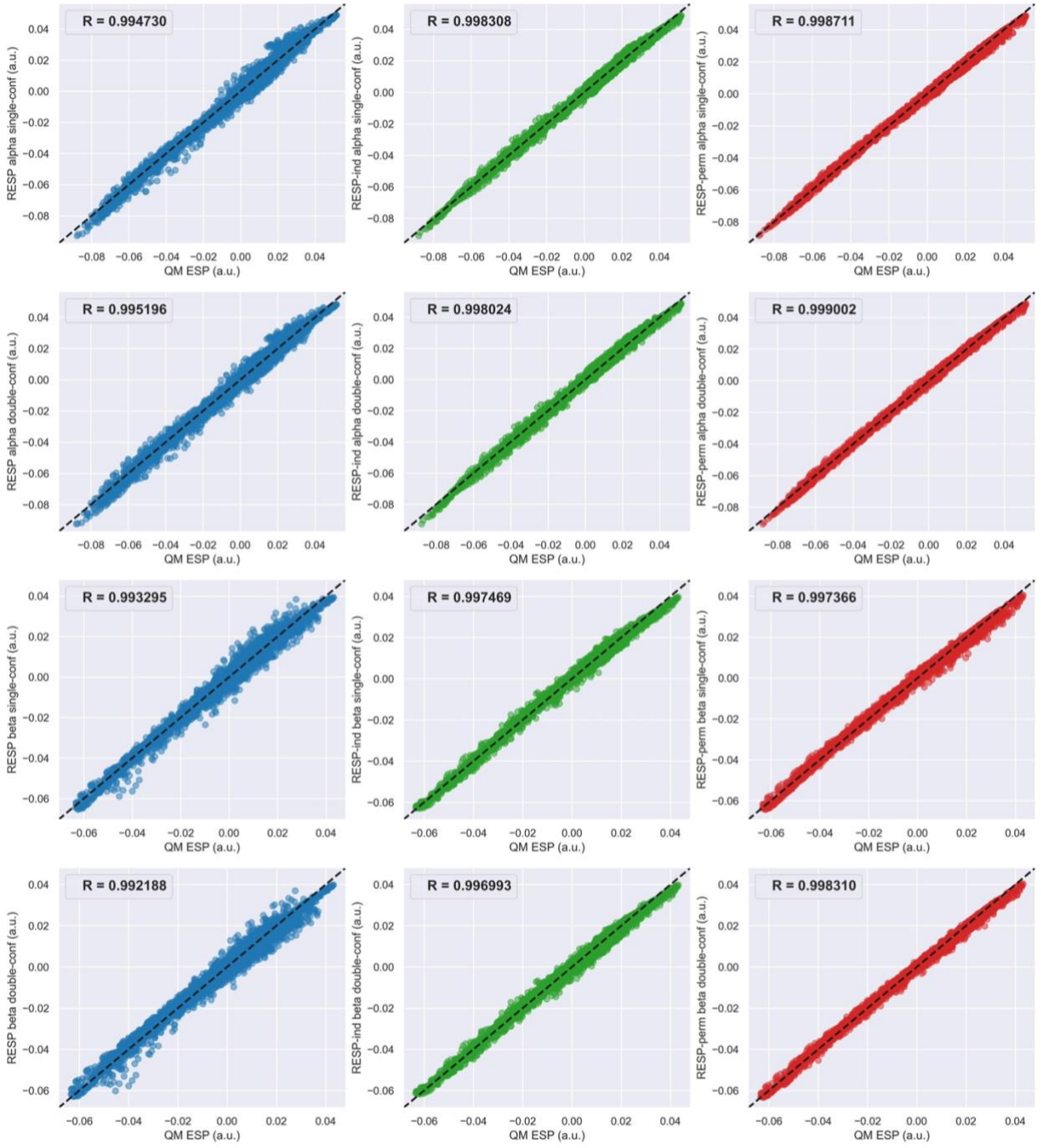


Figure S4.1. Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for serine dipeptide using single- and double-conformation fittings. 1st row: α-helix conformation fitted with single-conformation; 2nd row: α-helix conformation fitted with double-conformation; 3rd row: β-sheet conformation fitted with single-conformation; 4th

row: β-sheet conformation fitted with double-conformation. The α-helix conformation was fitted with 6542 ESP data points, and the β-sheet conformation was fitted with 6709 ESP data points. The dashed lines correspond to perfect correlation. *R* is the Pearson correlation coefficient.

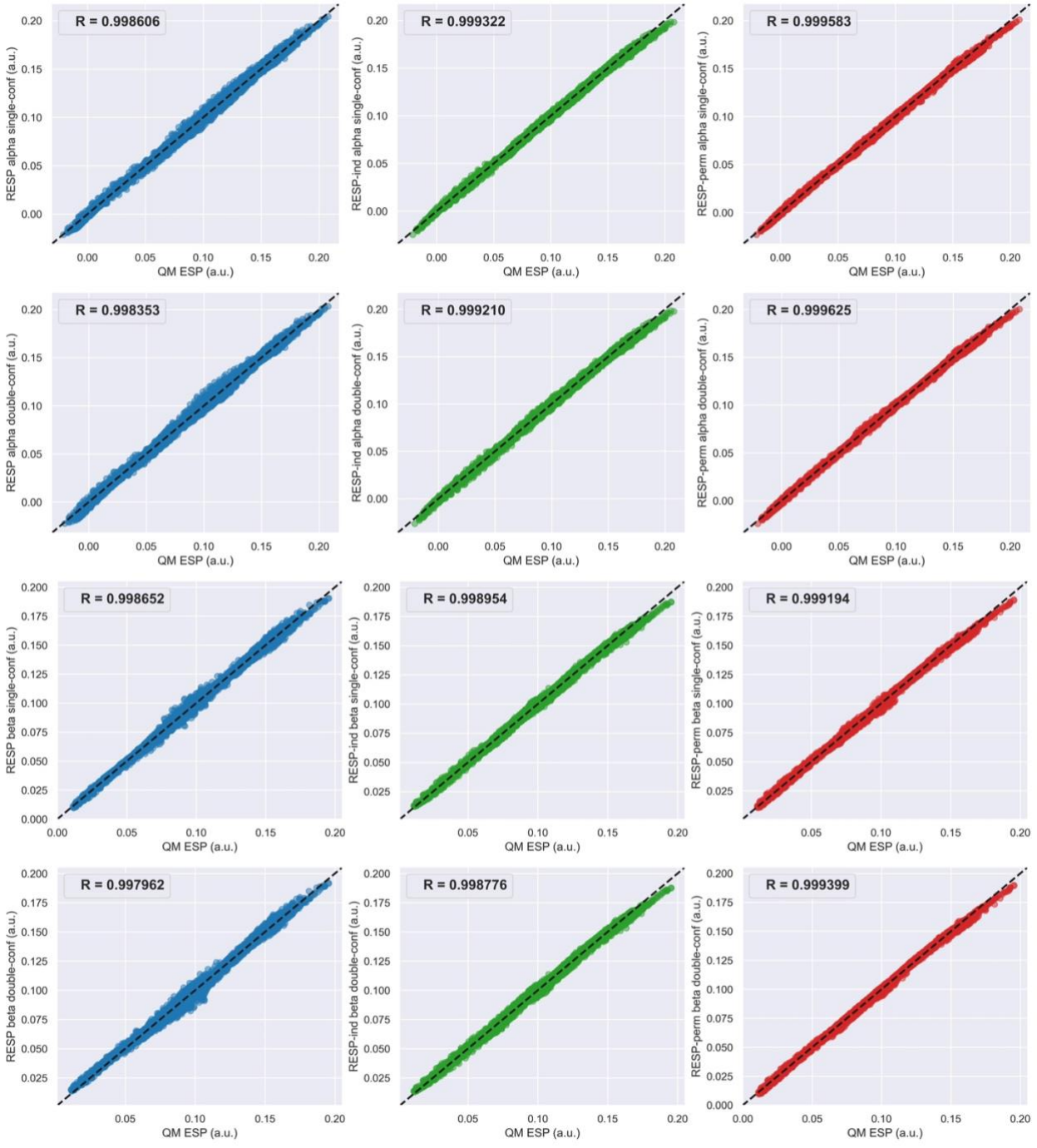


Figure S4.2. Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for arginine dipeptide using single- and double-conformation fittings. 1st row: α-helix conformation fitted with single-conformation; 2nd row: α-helix conformation fitted with double-conformation; 3rd row: β-sheet conformation fitted with single-conformation;

4th row: β-sheet conformation fitted with double-conformation. The α-helix conformation was fitted with 9165 ESP data points, and the β-sheet conformation was fitted with 9323 ESP data points. The dashed lines correspond to perfect correlation. *R* is the Pearson correlation coefficient.

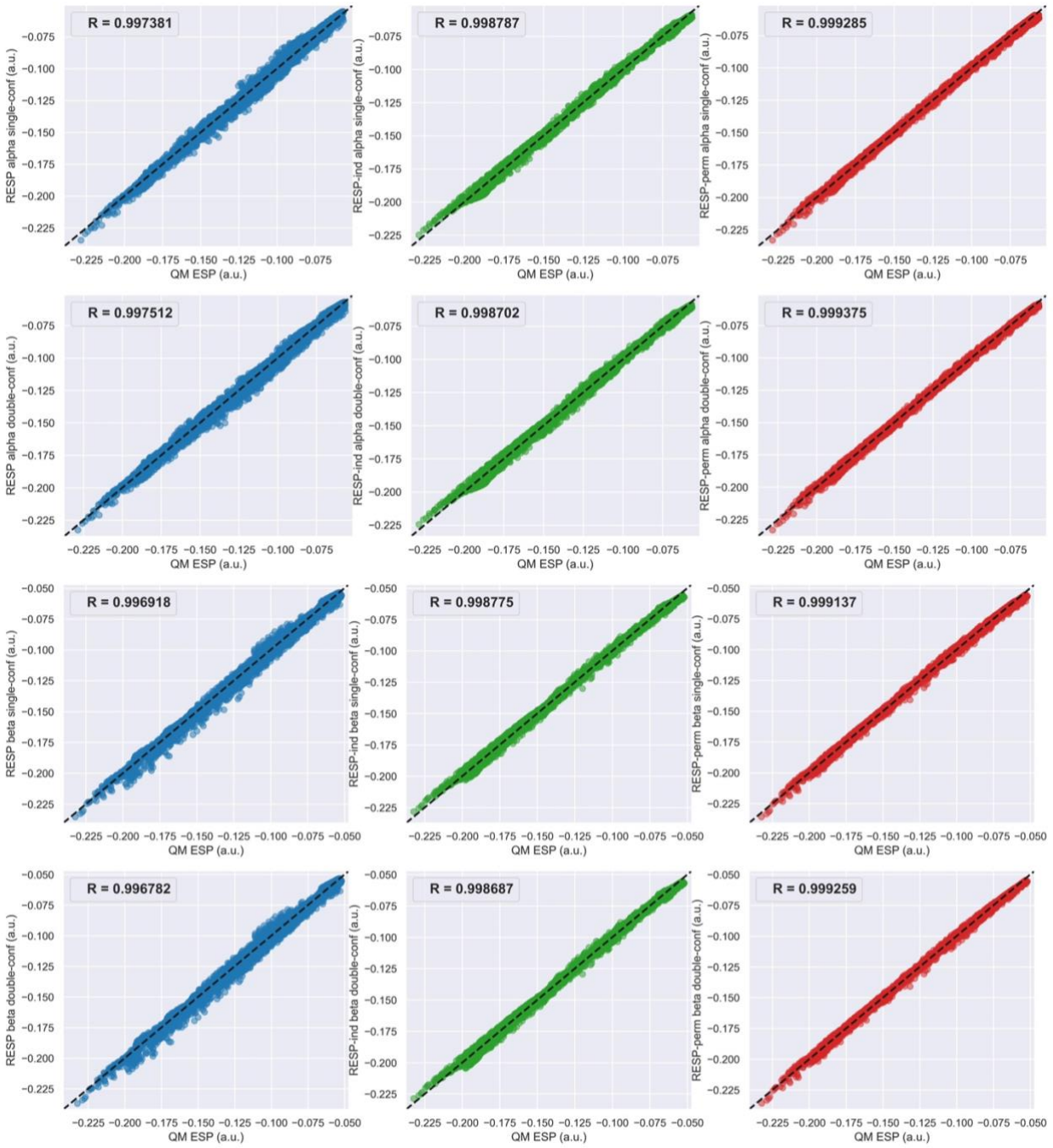


Figure S4.3. Correlation analysis of QM ESPs and ESPs calculated with various electrostatic models for aspartic acid dipeptide using single- and double-conformation fittings. 1st row: α-helix conformation fitted with single-conformation; 2nd row: α-helix conformation fitted with double-conformation; 3rd row: β-sheet conformation fitted with single-conformation;

4th row: β-sheet conformation fitted with double-conformation. The α-helix conformation was fitted with 7077 ESP data points, and the β-sheet conformation was fitted with 7047 ESP data points. The dashed lines correspond to perfect correlation. *R* is the Pearson correlation coefficient.

References

1. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015,** *11* (8), 3696-3713.
2. Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q., ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of chemical theory and computation* **2019,** *16* (1), 528-552.
3. Brooks, B. R.; Brooks III, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S., CHARMM: the biomolecular simulation program. *Journal of computational chemistry* **2009,** *30* (10), 1545-1614.
4. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996,** *118* (45), 11225-11236.
5. Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J., Polarization effects in molecular mechanical force fields. *Journal of Physics: Condensed Matter* **2009,** *21* (33), 333102.
6. Dill, K. A.; Bromberg, S.; Yue, K.; Chan, H. S.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D., Principles of protein folding—a perspective from simple exact models. *Protein science* **1995,** *4* (4), 561-602.
7. Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B., Experimental pKa values of buried residues: analysis with continuum methods and role of water penetration. *Biophysical journal* **2002,** *82* (6), 3289-3304.
8. Zhao, S.; Schaub, A. J.; Tsai, S.-C.; Luo, R., Development of a Pantetheine Force Field Library for Molecular Modeling. *Journal of chemical information and modeling* **2021,** *61* (2), 856-868.
9. Lin, F. Y.; MacKerell Jr, A. D., Improved Modeling of Cation-π and Anion-Ring Interactions Using the Drude Polarizable Empirical Force Field for Proteins. *Journal of computational chemistry* **2020,** *41* (5), 439-448.
10. Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X. X.; Murphy, R. B.; Zhou, R. H.; Halgren, T. A., Development of a polarizable force field for proteins via ab initio quantum chemistry: First generation model and gas phase tests. *Journal of Computational Chemistry* **2002,** *23* (16), 1515-1531.
11. Friesner, R. A., Modeling Polarization in Proteins and Protein–ligand Complexes: Methods and Preliminary Results. In *Adv. Protein Chem.*, Academic Press: 2005; Vol. 72, pp 79-104.
12. Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., A polarizable model of water for molecular dynamics simulations of biomolecules. *Chemical Physics Letters* **2006,** *418* (1-3), 245-249.

13. Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. *Journal of Physical Chemistry B* **2007,** *111* (11), 2873-2885.
14. Patel, S.; Mackerell, A. D.; Brooks, C. L., CHARMM fluctuating charge force field for proteins: II - Protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *Journal of Computational Chemistry* **2004,** *25* (12), 1504-1514.
15. Jiang, W.; Hardy, D. J.; Phillips, J. C.; MacKerell, A. D., Jr.; Schulten, K.; Roux, B., High-Performance Scalable Molecular Dynamics Simulations of a Polarizable Force Field Based on Classical Drude Oscillators in NAMD. *Journal of Physical Chemistry Letters* **2011,** *2* (2), 87-92.
16. Kumar, A.; Pandey, P.; Chatterjee, P.; MacKerell Jr, A. D., Deep Neural Network Model to Predict the Electrostatic Parameters in the Polarizable Classical Drude Oscillator Force Field. *Journal of Chemical Theory and Computation* **2022,** *18* (3), 1711-1725.
17. Cieplak, P.; Caldwell, J.; Kollman, P., Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of computational chemistry* **2001,** *22* (10), 1048-1057.
18. Wang, Z. X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y., Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *Journal of computational chemistry* **2006,** *27* (6), 781-790.
19. Tan, Y.-H.; Luo, R., Continuum treatment of electronic polarization effect. *J. Chem. Phys.* **2007,** *126* (9), 094103.
20. Tan, Y.-H.; Tan, C.; Wang, J.; Luo, R., Continuum polarizable force field within the Poisson-Boltzmann framework. *J. Phys. Chem. B* **2008,** *112* (25), 7675-7688.
21. Warshel, A.; Levitt, M., Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of molecular biology* **1976,** *103* (2), 227-249.
22. Vesely, F. J., N-particle dynamics of polarizable Stockmayer-type molecules. *Journal of Computational Physics* **1977,** *24* (4), 361-371.
23. Ren, P.; Ponder, J. W., Consistent treatment of inter-and intramolecular polarization in molecular mechanics calculations. *Journal of computational chemistry* **2002,** *23* (16), 1497-1506.
24. Ren, P.; Ponder, J. W., Polarizable atomic multipole water model for molecular mechanics simulation. *The Journal of Physical Chemistry B* **2003,** *107* (24), 5933-5947.
25. Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations I: parameterization of atomic polarizability. *The journal of physical chemistry B* **2011,** *115* (12), 3091-3099.
26. Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations II: induced dipole models significantly improve accuracy of intermolecular interaction energies. *The journal of physical chemistry B* **2011,** *115* (12), 3100-3111.
27. Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Luo, R., Development of polarizable models for molecular mechanical calculations. 3. Polarizable water models conforming to Thole polarization screening schemes. *The Journal of Physical Chemistry B* **2012,** *116* (28), 7999-8008.
28. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.-J.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *The Journal of Physical Chemistry B* **2012,** *116* (24), 7088-7101.
29. Applequist, J.; Carl, J. R.; Fung, K.-K., Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *Journal of the American Chemical Society* **1972,** *94* (9), 2952-2960.

30. Thole, B. T., Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics* **1981,** *59* (3), 341-350.
31. Van Duijnen, P. T.; Swart, M., Molecular and atomic polarizabilities: Thole's model revisited. *The journal of physical chemistry A* **1998,** *102* (14), 2399-2407.
32. Ponder, J. W., TINKER: Software tools for molecular design. *Washington University School of Medicine, Saint Louis, MO* **2004,** *3*.
33. Elking, D.; Darden, T.; Woods, R. J., Gaussian induced dipole polarization model. *Journal of computational chemistry* **2007,** *28* (7), 1261-1274.
34. Wei, H.; Qi, R.; Wang, J.; Cieplak, P.; Duan, Y.; Luo, R., Efficient formulation of polarizable Gaussian multipole electrostatics for biomolecular simulations. *The Journal of chemical physics* **2020,** *153* (11), 114116.
35. Wei, H.; Cieplak, P.; Duan, Y.; Luo, R., Stress Tensor and Constant Pressure Simulation for Polarizable Gaussian Multipole Model. *The Journal of chemical physics* **2022**.
36. Wang, J.; Cieplak, P.; Luo, R.; Duan, Y., Development of Polarizable Gaussian Model for Molecular Mechanical Calculations I: Atomic Polarizability Parameterization To Reproduce ab Initio Anisotropy. *J. Chem. Theory Comput.* **2019,** *15* (2), 1146-1158.
37. Momany, F. A., Determination of partial atomic charges from ab initio molecular electrostatic potentials. Application to formamide, methanol, and formic acid. *The Journal of Physical Chemistry* **1978,** *82* (5), 592-601.
38. Cox, S.; Williams, D., Representation of the molecular electrostatic potential by a net atomic charge model. *Journal of computational chemistry* **1981,** *2* (3), 304-323.
39. Singh, U. C.; Kollman, P. A., An approach to computing electrostatic charges for molecules. *Journal of computational chemistry* **1984,** *5* (2), 129-145.
40. Connolly, M. L., Analytical molecular surface calculation. *Journal of applied crystallography* **1983,** *16* (5), 548-558.
41. Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A., Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *Journal of computational chemistry* **1995,** *16* (11), 1357-1377.
42. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.
43. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993,** *97* (40), 10269-10280.
44. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A., Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* **1993,** *115* (21), 9620-9631.
45. Konishi, S.; Kashiwagi, Y.; Watanabe, G.; Osaki, M.; Katashima, T.; Urakawa, O.; Inoue, T.; Yamaguchi, H.; Harada, A.; Takashima, Y., Design and mechanical properties of supramolecular polymeric materials based on host–guest interactions: the relation between relaxation time and fracture energy. *Polymer Chemistry* **2020,** *11* (42), 6811-6820.
46. Wang, X.; Gao, J., Atomic partial charge predictions for furanoses by random forest regression with atom type symmetry function. *RSC Advances* **2020,** *10* (2), 666-673.
47. Chirlian, L. E.; Francl, M. M., Atomic charges derived from electrostatic potentials: A detailed study. *Journal of computational chemistry* **1987,** *8* (6), 894-905.
48. Breneman, C. M.; Wiberg, K. B., Determining atom - centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of computational chemistry* **1990,** *11* (3), 361-373.

49. Besler, B. H.; Merz Jr, K. M.; Kollman, P. A., Atomic charges derived from semiempirical methods. *Journal of computational chemistry* **1990,** *11* (4), 431-439.
50. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J., Gaussian 09, Revision D.01, Gaussian. *Inc., Wallingford CT* **2009,** *201*.
51. Xie, W.; Pu, J.; Gao, J., A coupled polarization-matrix inversion and iteration approach for accelerating the dipole convergence in a polarizable potential function. *The Journal of Physical Chemistry A* **2009,** *113* (10), 2109-2116.
52. Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Jin, C.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xue, Y.; York, D. M.; Zhao, S.; Kollman, P. A., *Amber 2021*. University of California, San Francisco: 2021.

CHAPTER 5

Accurate Reproduction of Quantum Mechanical Many-Body Interactions in Peptide Mainchain Hydrogen Bonding Oligomers by the Polarizable Gaussian Multipole Model

5.1. Introduction

Development of molecular mechanical force fields has been at the forefront of molecular modeling research due to the critical roles that force fields play in applications such as molecular dynamics (MD) simulations, Monte Carlo (MC) simulations, and protein structure prediction.[1-4] Force fields that have the ability to provide accurate energy calculations, and are highly transferable to a wide range of molecular systems have become highly desirable. With GPU-accelerated and specialized high-performance computational platforms,[5-6] it becomes increasingly feasible to conduct simulations at time scales of biological relevance. The extensively used point-charge additive force fields, such as Amber ff19SB,[7] CHARMM,[8] and OPLS,[9] share similar functional forms. In the additive Amber force fields, the following general functional form is used to calculate the potential energies of molecular systems

$$E_{total} = E_{bond} + E_{angle} + E_{dihedral} + E_{ele} + E_{vdW} \tag{5.1}$$

The first three terms are short-range bonded terms, including the bond stretching terms E_{bond}, the angle bending terms E_{angle}, and the dihedral angle torsion terms $E_{dihedral}$, with the following formulas

$$E_{bond} = \sum_{bonds} k_b (r - r_0)^2 \tag{5.2}$$

$$E_{angle} = \sum_{angles} k_\theta(\theta - \theta_0)^2 \tag{5.3}$$

$$E_{dihedral} = \sum_{dihedrals} V_n(1 + \cos(n\phi - \gamma)) \tag{5.4}$$

The last two terms are non-bonded terms between any two atoms i and j. The electrostatic term E_{ele}, usually modeled by the interactions between fixed atom-centered partial charges (Coulomb's law), is a long-range term; whereas the van der Waals term E_{vdW}, modeled by the 6-12 Lennard-Jones potential, is nominally also a long-range term, although it decays rather quickly with increasing distance. E_{ele} and E_{vdW} are formulated as

$$E_{vdW} = \sum_{i<j} \left(\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} \right) \tag{5.5}$$

$$E_{ele} = \sum_{i<j} \frac{q_i q_j}{\varepsilon R_{ij}} \tag{5.6}$$

both of which are pairwise and additive. Therefore, in this framework, the interaction between any two atoms is not affected by the presence or absence of other non-bonded atoms.

While additive force fields will continue to play important roles, polarizable force fields are expected to extend our ability to study biomolecular systems more adequately, due to their ability to model the atomic polarization effects, which are the redistribution of atomic electron density due to the electric field produced by nearby atoms.[10] Polarization effects are important in biological processes such as ligand-receptor interactions,[11-14] the interactions of ions with nucleic acids,[15-16] the dielectric environmental changes during

protein folding,[17-18] and enzymatic mechanisms.[19] If more than two atoms are involved, polarization effects lead to non-additivity, since when polarized by a third atom, any two atoms interact differently from the situation where the third atom is absent. Lacking proper representation of the polarization effects is considered a major shortcoming of the additive force fields. For over five decades, many attempts have been directed to properly incorporating polarization effects into polarizable force fields. A variety of methods have been explored, including the induced dipole models,[20-28] the fluctuating charge models,[29-30] the Drude oscillator models[31-32], and the continuum dielectric models.[33-34]

The induced point dipole model is one of the most studied approaches with a long history since the 1970s.[35-36] In this approach, the induced dipole of atom i subject to the external electric field $\boldsymbol{E}_i$, is

$$\boldsymbol{\mu}_i = \alpha_i \left[\boldsymbol{E}_i - \sum_{j \neq i}^{\boldsymbol{n}} \boldsymbol{T}_{ij} \boldsymbol{\mu}_j \right] \tag{5.7}$$

where α_i is the isotropic polarizability of atom i, and $\boldsymbol{T}_{ij}$ is the dipole field tensor with the matrix form

$$\boldsymbol{T}_{ij} = \frac{f_e}{{r_{ij}}^3} \boldsymbol{I} - \frac{3f_t}{{r_{ij}}^5} \begin{bmatrix} x^2 & xy & xz \\ xy & y^2 & yz \\ xz & yz & z^2 \end{bmatrix} \tag{5.8}$$

where $\boldsymbol{I}$ is the identity matrix; x, y and z are the Cartesian components along the vector between atoms i and j at distance r_{ij}; f_e and f_t are distance-dependent damping functions that modify $\boldsymbol{T}_{ij}$ to avoid the so-called "polarization catastrophe" problem, i.e. the phenomenon that induced dipole diverges due to the cooperative induction between

induced dipoles at short distances.[10, 37] Several damping schemes have been proposed by Thole using a smeared charge distributions $\rho(u)$, where $u = r_{ij}/(\alpha_i\alpha_j)^{1/6}$ is the effective distance.[38-39] Thole's damping schemes have been incorporated into several important polarizable force fields. For example, in the ff12pol force field,[22-25] the linear damping scheme is adopted

$$\rho(u) = \begin{cases} \frac{3}{\pi}\frac{(a-u)}{a^4} & u < a \\ 0 & u \geq a \end{cases} \tag{5.9}$$

and the damping functions f_e and f_t have the form

$$v = u/a$$

$$f_e = \begin{cases} 4v^3 - 3v^4 & v < 1 \\ 1.0 & v \geq 1 \end{cases} \tag{5.10}$$

$$f_t = \begin{cases} v^4 & v < 1 \\ 1.0 & v \geq 1 \end{cases}$$

In the Amoeba polarizable force field,[26-28] an exponential damping scheme is used

$$\rho(u) = \frac{3a}{4\pi}\exp(-au^3) \tag{5.11}$$

and the damping functions f_e and f_t become

$$v = au^3$$

$$f_e = 1 - \exp(-v) \tag{5.12}$$

$$f_t = 1 - (v+1)\exp(-v)$$

However, since Thole's schemes only screen the interactions between induced dipoles, leaving the polarization due to fixed charges and permanent multipoles unaffected, one caveat is the possibility of producing large atomic induced dipoles when other highly charged species are nearby. About a decade ago, Elking et al. developed a scheme that models atomic electric multipoles using Gaussian electron densities,[40-42] which was originally proposed by Wheatley,[43-44] and this model was later named as the polarizable Gaussian Multipole (pGM) model.[45-47] The pGM model can overcome the potential problem of Thole's scheme by screening all short-range electrostatic interactions in a consistent manner, including the interactions of charge-charge, charge-dipole, charge-quadrupole, dipole-dipole, and so on, eliminating a potential source of singularity in the electrostatic term E_{ele}. Consequently, it has been shown that the pGM model notably improves the prediction of molecular polarizability anisotropy compared with that of Thole models.[45] In the pGM model, the nth order Gaussian multipole at distance r with atom i is defined as

$$\rho^{(n)}(r) = \boldsymbol{\Theta}^{(n)} \cdot \nabla^{(n)} \left(\frac{\beta_i}{\sqrt{\pi}}\right)^3 \exp\left(-{\beta_i}^2 r^2\right) \tag{5.13}$$

where $\boldsymbol{\Theta}^{(n)}$ is the nth rank momentum tensor, $\nabla^{(n)}$ is the nth rank gradient operator, and β_i is the Gaussian exponent controlling the "radius" of the distribution with the following formula

$$\beta_i = s\left(\frac{2\alpha_i}{3\sqrt{2\pi}}\right)^{-\frac{1}{3}} \tag{5.14}$$

where α_i is the atomic polarizability, and s is a constant screening factor. Although any order of multipoles can be modeled by the pGM model, only charges (0th order multipole, **eq 5.15**) and dipoles (1st order multipole, **eq 5.16**) are considered in the current pGM model design

$$\rho^{(0)}(r) = q_i \left(\frac{\beta_i}{\sqrt{\pi}}\right)^3 \exp\left(-{\beta_i}^2 r^2\right) \tag{5.15}$$

$$\rho^{(1)}(r) = \boldsymbol{p}_i \cdot \nabla \left(\frac{\beta_i}{\sqrt{\pi}}\right)^3 \exp\left(-{\beta_i}^2 r^2\right) \tag{5.16}$$

where q_i is the permanent charge and $\boldsymbol{p}_i$ is the permanent dipole of atom i. Replacing $\boldsymbol{p}_i$ in **eq 5.16** with $\boldsymbol{\mu}_i$ in **eq 5.7** will give the pGM distribution of the induced dipole, which has the same form as that of the permanent dipole. For the pGM model, we have the following formula of damping functions f_e and f_t

$$S_{ij} = \frac{\beta_i \beta_j r_{ij}}{\sqrt{2\left({\beta_i}^2 + {\beta_j}^2\right)}}$$

$$f_e = \mathrm{erf}\left(S_{ij}\right) - \frac{2}{\sqrt{\pi}} S_{ij} \exp\left(-{S_{ij}}^2\right) \tag{5.17}$$

$$f_t = \mathrm{erf}\left(S_{ij}\right) - \frac{2}{\sqrt{\pi}} S_{ij} \exp\left(-{S_{ij}}^2\right)\left(1 + \frac{2}{3}{S_{ij}}^2\right)$$

where $\mathrm{erf}\left(S_{ij}\right)$ is the error function of S_{ij}.

In a series of recent works, the functional form and parameterization schemes for the pGM model have been designed and implemented. First, a set of isotropic atomic polarizabilities and radii for the pGM model were obtained by fitting to molecular polarizability tensors of 1405 molecules or dimers calculated at the B3LYP/aug-cc-

pVTZ//B3LYP/aug-cc-pVTZ level of theory using an optimization method based on the genetic algorithm (GA).[45] Second, a local frame for the pGM permanent dipoles formed by covalent basis vectors (CBVs), which are unit vectors along the direction of covalent bonds or virtual bonds, has been proposed based on the fact that atomic permanent moments mainly result from covalent bonding interactions.[46] Third, the analytical formula of the electrostatic term of the pGM models have been derived,[46] which is the sum of a permanent electrostatic term $E_{ele-perm}$ and an induced electrostatic term $E_{ele-ind}$

$$E_{ele} = E_{ele-perm} + E_{ele-ind} \tag{5.18}$$

with the following formula

$$E_{ele-perm} = \sum_{i<j} (q_i + \boldsymbol{p}_i \cdot \nabla_i)(q_j + \boldsymbol{p}_j \cdot \nabla_j) \frac{\mathrm{erf}(S_{ij})}{r_{ij}} \tag{5.19}$$

$$E_{ele-ind} = \sum_{i<j} \boldsymbol{\mu}_i (q_j + \boldsymbol{p}_j \cdot \nabla_j) \nabla_i \frac{\mathrm{erf}(S_{ij})}{r_{ij}} \tag{5.20}$$

Therefore, in the functional form of the pGM models, the electrostatic term in **eq 5.6** is replaced by **eq 5.18-5.20**, and the rest of terms remain unchanged (**eq 5.2-5.5**). In addition, the pGM electrostatic term has been interfaced with the particle mesh Ewald (PME) method for molecular simulations under the periodic boundary conditions.[46, 48-51] Fourth, the pGM internal stress tensor expression for constant pressure MD simulations of both the flexible and rigid body molecular system has been derived.[47] Finally, the *PyRESP* program enabling parameterizations for the pGM models with and without atomic permanent dipoles by reproducing quantum mechanical (QM) electrostatic potential (ESP) around molecules has been implemented.[52] All of the components mentioned above, including the pGM

polarizabilities and radii, the *sander* program enabling MD simulations for the pGM models, and the *PyRESP* parametrization program, are available in the AmberTools22 program suite that can be downloaded from http://ambermd.org/.[53]

In this chapter, we assessed the ability of the pGM models to reproduce QM many-body interaction energies in peptide oligomers, specifically the influences of neighboring peptides upon a pair of interacting peptide monomers. For polarizable force fields, the many-body interaction energies can be decomposed into non-additive and additive contributions. The detailed definitions of the many-body interaction energy as well as its non-additive and additive contributions will be presented in **section 5.2**. Glycine dipeptide oligomers arranged in three mainchain hydrogen bonding conformations were used as the model peptide systems, because glycine has the minimalist side chain so that we can focus on mainchain hydrogen bonding interactions. Two types of pGM models were considered, including pGM-perm, in which the atomic dipoles are represented by a combination of both induced and permanent dipoles, and pGM-ind, in which the atomic dipoles are represented by the induced dipoles only. We compared the performances of the pGM-perm and pGM-ind models with several other widely used force fields in terms of reproducing QM interaction energies and many-body interaction energies, including four Amber force fields: ff12pol,[22-25] ff19SB,[7] ff15ipq[54] and ff03,[55] as well as the 2013 version of the Amoeba protein force field (Amoeba13).[28] Among the seven force fields tested, pGM-perm, pGM-ind, Amoeba13 and ff12pol are polarizable force fields, while ff19SB, ff15ipq and ff03 are classical point-charge additive force fields. The results show that the pGM models perform significantly better than all other force fields in terms of reproducing QM interaction energies, many-body interaction energies, and the non-additive and additive contributions to the many-body interactions. In

addition, we tested the robustness of the pGM models against parameterization errors by employing alternative atomic polarizabilities, including the pGM polarizabilities scaled by a factor of 0.9,[45] the Amoeba polarizabilities,[26] and the ff12pol polarizabilities.[22] The results show that the pGM models are highly robust and perform well even with those "wrong" polarizabilities.

5.2. Theory

In this chapter, each oligomer is arranged in a general form of m glycine dipeptides interacting with n glycine dipeptides, named Gly_m:Gly_n, where m and n are ranged from 1 to 3. Each "Gly" in this chapter represents a glycine dipeptide (ACE-GLY-NME) capped with an *N*-acetyl (ACE) group at the N-terminal, and an *N*-methylamide (NME) group at the C-terminal. For example, **Figure 5.1A** shows the Gly_2:Gly_2 oligomer.

The interaction energy $IE(Gly_m: Gly_n)$ between Gly_m and Gly_n of the Gly_m:Gly_n oligomer can be calculated by the following equation

$$IE(Gly_m: Gly_n) = E(Gly_m: Gly_n) - E(Gly_m) - E(Gly_n) \quad (5.21)$$

where $E(Gly_m: Gly_n)$ is the potential energy of the entire Gly_m:Gly_n oligomer, and $E(Gly_m)$, $E(Gly_n)$ are the potential energies of isolated Gly_m and Gly_n, respectively.

More importantly, we intend to study the many-body effects in the Gly_m:Gly_n oligomer, specifically, the influence of the neighboring glycine dipeptides Gly_{m-1} and Gly_{n-1} upon the interaction between the two middle glycine dipeptides Gly:Gly in the Gly_m:Gly_n

oligomer. Here we define the many-body interaction energy $ME(Gly_m: Gly_n)$ as the difference between $IE(Gly_m: Gly_n)$ and IE(Gly: Gly). That is

$$ME(Gly_m: Gly_n) = IE(Gly_m: Gly_n) - IE(Gly: Gly) \tag{5.22}$$

Taking the Gly_2:Gly_2 oligomer in **Figure 5.1A** as an example, the difference between $IE(Gly_2: Gly_2)$ and IE(Gly: Gly) of the two middle peptides (displayed in brown) is the many-body interaction energy $ME(Gly_2: Gly_2)$ caused by the presence of the two neighboring peptides (displayed in cyan).

The many-body interaction energy $ME(Gly_m: Gly_n)$ can be decomposed into the non-additive contribution $ME_{NA}(Gly_m: Gly_n)$ and additive contribution $ME_A(Gly_m: Gly_n)$. Before showing their formulas, we first need to define the interaction energies of the two middle peptides $IE_{mid}(Gly_m: Gly_n)$ in the presence of the neighboring peptides Gly_{m-1} and Gly_{n-1}

$$\begin{aligned} IE_{mid}(Gly_m: Gly_n) = IE(Gly_m: Gly_n) - IE(Gly_m: XGly_{n-1}) \\ -IE(Gly_{m-1}X: Gly_n) + IE(Gly_{m-1}X: XGly_{n-1}) \end{aligned} \tag{5.23}$$

where X indicates the absence of either one of the two middle peptides. $IE(Gly_m: XGly_{n-1})$ is the interaction energy between Gly_m and the neighboring peptides Gly_{n-1}; $IE(Gly_{m-1}X: Gly_n)$ is the interaction energy between the neighboring peptides Gly_{m-1} and Gly_n; and $IE(Gly_{m-1}X: XGly_{n-1})$ is the interaction energy between the neighboring peptides Gly_{m-1} and Gly_{n-1} on both sides. Then we have the formulas of $ME_{NA}(Gly_m: Gly_n)$ and $ME_A(Gly_m: Gly_n)$, the proof of which can be found in the **Appendix A**

$$ME_{NA}(Gly_m: Gly_n) = IE_{mid}(Gly_m: Gly_n) - IE(Gly: Gly) \tag{5.24}$$

$$ME_A(Gly_m: Gly_n) = ME(Gly_m: Gly_n) - ME_{NA}(Gly_m: Gly_n) \tag{5.25}$$

For the Gly_2:Gly_2 oligomer example, **Figure 5.1B** shows the oligomer Gly_2:XGly with the interaction energy $IE(Gly_2: XGly)$; **Figure 5.1C** shows the oligomer GlyX:Gly_2 with the interaction energy $IE(GlyX: Gly_2)$; and **Figure 5.1D** shows the oligomer GlyX:XGly with the interaction energy $IE(GlyX: XGly)$. The interaction energy of the two middle peptides in the presence of the neighboring peptides is $IE_{mid}(Gly_2: Gly_2) = IE(Gly_2: Gly_2) - IE(Gly_2: XGly) - IE(GlyX: Gly_2) + IE(GlyX: XGly)$. The non-additive and additive contributions to the many-body interaction energy $ME(Gly_2: Gly_2)$ are $ME_{NA}(Gly_2: Gly_2) = IE_{mid}(Gly_2: Gly_2) - IE(Gly: Gly)$ and $ME_A(Gly_2: Gly_2) = ME(Gly_2: Gly_2) - ME_{NA}(Gly_2: Gly_2)$, respectively.

For additive force fields, $ME_{NA}(Gly_m: Gly_n)$ is guaranteed to be zero, so that $ME(Gly_m: Gly_n)$ is equivalent to $ME_A(Gly_m: Gly_n)$; while for polarizable force fields, $ME_{NA}(Gly_m: Gly_n)$ is non-zero, so that $ME(Gly_m: Gly_n)$ has both additive and non-additive contributions.

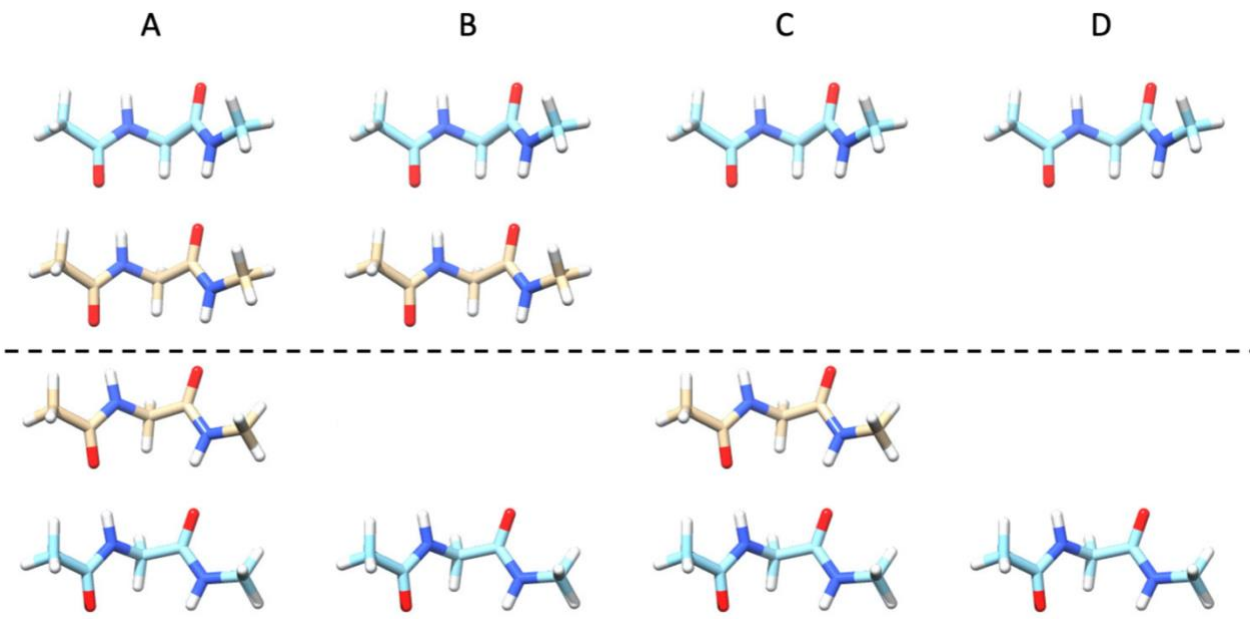


Figure 5.1. Glycine dipeptide oligomers (in the parallel β-sheet conformation) used to calculate the interaction energy, many-body interaction energy, and the non-additive and additive contributions to the many-body interaction of the Gly_2:Gly_2 oligomer. For each oligomer, the interaction energies between glycine dipeptides above and below the dashed line are calculated using **eq 5.21**. **A.** Gly_2:Gly_2, **B.** Gly_2:XGly, **C.** GlyX:Gly_2, **D.** GlyX:XGly. Refer to **section 5.2** for detailed descriptions.

5.3. Computational Details

5.3.1. Geometry Preparations

The formamide dimer and three glycine dipeptide dimers were used to select density functional theory (DFT) methods for subsequent QM energy calculations. The formamide dimer was first arranged into hydrogen bonding conformation, and the geometry was

optimized at the B3LYP/6-311++G(d, p) level of theory. A total of fifteen glycine dipeptide oligomers were constructed. First, three glycine dipeptide dimers were configured and arranged into α-helix, anti-parallel β-sheet and parallel β-sheet hydrogen bonding conformations observed in proteins. Then, the geometries were optimized at the B3LYP/6-311++G(d, p) level of theory with the mainchain torsion angles fixed at (ϕ, ψ) = (-57°, -47°), (-140°, 135°) and (-119°, 113°), corresponding to the α-helix, anti-parallel β-sheet, and parallel β-sheet conformations, respectively. Higher-order oligomers were constructed from these three optimized dimers by rigid-body translations and rotations. For example, to produce a Gly_3:Gly_3 dipeptide hexamer (in the conformation of an interacting pair of trimers) while maintaining the central dimer in the optimized conformation, both dipeptides of the Gly:Gly dimer are rotated and moved towards both sides along the hydrogen-bond direction. The structures of formamide dimer and glycine dipeptide oligomers are presented in **Figures S5.1-5.6**.

5.3.2. Quantum Mechanical Calculations

Three DFT methods were tested to calculate the QM interaction energies of the formamide dimer and the glycine dipeptide dimers, including ωB97X-D,[56] M062X,[57] and B3LYP,[58-59] all with the aug-cc-pVTZ basis set. The basis set superposition errors (BSSEs) were corrected through the counterpoise corrections.[60] To select the most suitable DFT method for our systems, the CCSD(T)/CBS interaction energies $IE_{CCSD(T)/CBS}$ were calculated as the reference energies using Helgaker's extrapolation method.[61-62] First, the HF and MP2 interaction energies were calculated with aug-cc-pVTZ (aTZ) and aug-cc-pVQZ (aQZ) basis

sets, and the correlation (CORR) energies IE_{CORR} were defined as the difference between the MP2 and HF energies $IE_{CORR} = IE_{MP2} - IE_{HF}$. Next, $IE_{HF/CBS}$ and $IE_{CORR/CBS}$ were calculated using the following equations

$$IE_{HF/CBS} = \frac{IE_{HF/aTZ} \times \exp(-1.63 \times 4) - IE_{HF/aQZ} \times \exp(-1.63 \times 3)}{\exp(-1.63 \times 4) - \exp(-1.63 \times 3)} \tag{5.26}$$

$$IE_{CORR/CBS} = \frac{IE_{CORR/aTZ} \times 3^3 - IE_{CORR/aQZ} \times 4^3}{3^3 - 4^3} \tag{5.27}$$

and $IE_{MP2/CBS}$ can be calculated as

$$IE_{MP2/CBS} = IE_{HF/CBS} + IE_{CORR/CBS} \tag{5.28}$$

Note that the average of $IE_{MP2/CBS}$ with and without counterpoise corrections was used as the final $IE_{MP2/CBS}$. Finally, $IE_{CCSD(T)/CBS}$ were calculated by adding a CCSD(T) correction calculated at a small basis set to the averaged $IE_{MP2/CBS}$

$$IE_{CCSD(T)/CBS} = IE_{MP2/CBS} + (IE_{CCSD(T)} - IE_{MP2})_{\text{small basis set}} \tag{5.29}$$

For formamide dimers, aug-cc-pVTZ was used as the small basis set; for glycine dipeptide dimers, cc-pVDZ was used as the small basis set.

Following the strategy that has been successfully used in Amber force field development in which the partial charges were fit to QM electrostatic potentials (ESPs), the QM ESPs were calculated at the MP2/aug-cc-pVTZ level of theory for a set of points in the solvent-accessible region around each glycine dipeptide molecule in the α-helix, anti-parallel β-sheet, and parallel β-sheet conformations. The points were generated using the method developed by Singh et al. on molecular surfaces (with a density of 6 points/Å^2) at each of 1.4,

1.6, 1.8 and 2.0 times the van der Waals radii.[63-64] All QM calculations were performed using the Gaussian 16 software.[65]

5.3.3. pGM Parameterizations

To assess the robustness of the pGM models against errors in polarizability parameterization, four sets of atomic polarizabilities were employed to parameterize the pGM models, including the pGM polarizabilities,[45] the pGM polarizabilities scaled by a factor of 0.9, the Amoeba polarizabilities,[26] and the ff12pol polarizabilities,[22] for a combined total of two pGM models and six variants. The recently developed *PyRESP* program was used to parameterize the point charges and permanent point dipoles of the glycine dipeptide molecule for the pGM-perm and pGM-ind models, and a two-stage parameterization procedure was adopted.[52] In the first stage, all charges and permanent dipoles were set free to change, and a weak restraining strength 0.0005 was applied. In the second stage, intra-molecular equivalencing was enforced on all charges and permanent dipoles that share identical chemical environment with others, such as those of methyl and methylene hydrogens. A stronger restraining strength 0.001 was applied, and all other fitting centers were set frozen to keep the values obtained from the first stage. In both stages, the restraints were only applied to non-hydrogen heavy atoms. Only the total charge constraint was enforced in the parameterization process, and no additional intra-molecular charge constraint was applied. Inter-molecular equivalencing was enforced in both the first and the second stages for the three conformations of glycine dipeptides. For the parameterizations

of both the pGM-perm and pGM-ind models, both 1-2 and 1-3 polarization interactions were included for reasons elucidated before.[45, 66]

The parameters of bonded terms (bond stretching terms, angle bending terms, dihedral angle torsion terms) and the van der Waals terms for both the pGM-perm and pGM-ind models were obtained from the ff12pol force field without any change.[25]

5.3.4. Molecular Mechanics Calculations

Seven force fields were explored for calculating the molecular mechanics energies of glycine dipeptide oligomers, including four polarizable force fields: pGM-perm, pGM-ind, ff12pol,[22-25] and Amoeba13,[28] and three additive force fields: ff19SB,[7] ff15ipq[54] and ff03.[55] The parameter and topology files for the Amber force fields (pGM-perm, pGM-ind, ff12pol, ff19SB, ff15ipq and ff03) were generated using the *tleap* program from the AmberTools22 program suite.[53] The coordinate files for the Amber force fields and the .xyz files for the Amoeba13 force field were generated from the geometries optimized by the Gaussian 16 software.[65] The single point energies of the Amber force fields were calculated by the *sander* program with extensions to accommodate the pGM models.[46, 53] The *dynamic* program from the Tinker 8.6.1 software package was used to calculate the single point energies of the Amoeba13 force field.[67] All nonbonded interactions were calculated in gas phase without distance cutoff.

The performance of each force field in each energy calculation task was evaluated by the root-mean-squared-error (RMSE) and mean-absolute-error (MAE), given by

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(E_i^{QM} - E_i\right)^2}{N}} \tag{5.30}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N}\left|E_i^{QM} - E_i\right|}{N} \tag{5.31}$$

where E_i^{QM} is the energy given by QM calculations, and E_i is the energy calculated by molecular mechanics force fields.

5.4. Results and Discussion

5.4.1. ωB97X-D without Counterpoise Correction Most Accurately Reproduces CCSD(T)/CBS Interaction Energies

There have been numerous works documenting the performances of various DFT methods in their ability to model the dispersion effect. Among these DFT methods, ωB97X-D[56] and M062X[57] exhibit great trade-offs between computation speed and accuracy.[68] One observation is that the accuracy of DFT methods depends on the particular molecular systems being studied. To determine which one of these DFT methods is the most suitable to our systems, we compared the interaction energies of the formamide dimer and the glycine dipeptide dimers obtained from three DFT methods, including ωB97X-D,[56] M062X,[57] and B3LYP,[58-59] with those calculated at the CCSD(T)/CBS level of theory, which have been considered as the "gold standard" of computational chemistry. **Table 5.1** shows the interaction energies calculated with these DFT methods with and without counterpoise BSSE corrections. We can see that the interaction energies by ωB97X-D without counterpoise

correction are the closet to the CCSD(T)/CBS results (**eq 5.29**), with an RMSE of 0.17 kcal/mol and an MAE of 0.12 kcal/mol. Not surprisingly, B3LYP interaction energies consistently exhibit the highest deviations with CCSD(T)/CBS results, since B3LYP lacks proper consideration of dispersion contributions. Without counterpoise corrections, the RMSEs of M062X and B3LYP are 0.32 and 3.37 kcal/mol, respectively and the MAEs are 0.30 and 3.12 kcal/mol, respectively. With counterpoise corrections, the RMSEs of ωB97X-D, M062X and B3LYP are 0.31, 0.71 and 3.70 kcal/mol, respectively and the MAEs are 0.29, 0.67 and 3.42 kcal/mol, respectively. Therefore, for the formamide dimer and the glycine dipeptide dimers, ωB97X-D without counterpoise BSSE correction best reproduces the CCSD(T)/CBS interaction energies. For this reason, ωB97X-D without counterpoise correction was chosen as the QM reference method to evaluate various molecular mechanical force fields in the following discussions.

Table 5.1. The Quantum Mechanical Interaction Energies of the Formamide Dimer and the Glycine Dipeptide Dimers Calculated by the CCSD(T)/CBS and Density Functional Theory Methods (kcal/mol)

CCSD(T)/CBS[a]		ωB97X-D/aTZ		M062X/aTZ		B3LYP/aTZ	
		noCP[b]	CP[b]	noCP[b]	CP[b]	noCP[b]	CP[b]
Formamide Dimer							
-7.04		-6.96 (0.08)[d]	-6.83 (0.21)	-6.91 (0.13)	-6.77 (0.27)	-6.15 (0.89)	-6.05 (0.99)
Glycine Dipeptide Dimer (Gly:Gly)							
αR[c]	-14.40	-14.73 (-0.33)	-14.29 (0.11)	-14.00 (0.40)	-13.57 (0.83)	-10.37 (4.03)	-10.01 (4.39)
aβ[c]	-14.49	-14.53 (-0.04)	-14.10 (0.39)	-14.22 (0.27)	-13.77 (0.72)	-10.54 (3.95)	-10.17 (4.32)
β[c]	-17.34	-17.36 (-0.02)	-16.91 (0.43)	-16.93 (0.41)	-16.47 (0.87)	-13.75 (3.59)	-13.37 (3.97)
RMSE		0.17	0.31	0.32	0.71	3.37	3.70
MAE		0.12	0.29	0.30	0.67	3.12	3.42

[a] The formula for the CCSD(T)/CBS interaction energies is shown in **eq 5.29**. [b] noCP means no counterpoise BSSE corrections; CP means with counterpoise BSSE corrections. [c] αR, aβ,

and β represent α-helix, anti-parallel β-sheet, and parallel β-sheet conformations, respectively. [d] The values in parentheses are differences between the values calculated by DFT methods and corresponding values calculated by the CCSD(T)/CBS method.

5.4.2. pGM Models Show the Best Performances in Interaction Energy Calculations

Listed in **Table 5.2** are the interaction energies calculated at the ωB97X-D/aug-cc-pVTZ level of theory without counterpoise corrections, and seven molecular mechanical force fields. The interaction energies IE(Gly_m: Gly_n) between Gly_m and Gly_n of the Gly_m:Gly_n oligomers were calculated following **eq 5.21** without consideration of the deformation energies to avoid complications in energy calculations that may arise from structural changes. The interaction energies calculated by the ωB97X-D method exhibit an increasing trend in the order of Gly:Gly, Gly:Gly_2, Gly:Gly_3, Gly_2:Gly_2, Gly_3:Gly_3 for all three conformations. It is notable that IE(Gly_3: Gly_3) is 11.00 kcal/mol stronger than IE(Gly: Gly) in the α-helix conformation. In comparison, for the anti-parallel β-sheet and parallel β-sheet conformations, IE(Gly_3: Gly_3) are only 3.50 and 4.11 kcal/mol stronger than IE(Gly: Gly), respectively. The larger difference of the α-helix conformation is attributable to the strong polarization effect caused by the alignments of the main chain peptide hydrogen bonds. Another observation is that although Gly:Gly_3 and Gly_2:Gly_2 are both tetramers, Gly_2:Gly_2 consistently show stronger interaction energies than Gly:Gly_3 in all three conformations. This shows that the inner parts of peptide secondary structures are expected to have stronger mainchain hydrogen bonding than the outer parts.

The interaction energies calculated by the pGM-perm and pGM-ind models stand out as the closest to the DFT results, with RMSEs of 1.35 and 1.37 kcal/mol, respectively. The similarity between the performances of the pGM-perm and pGM-ind models indicates that, with the pGM damping schemes, the induced dipoles are sufficient for calculating the interaction energies of glycine dipeptides. The next best performance is given by the ff15ipq force field with an RMSE of 1.87 kcal/mol, which is an additive force field whose charges were fit to the ESP of peptides in the presence of explicit solvent water.[54] The polarizable force field ff12pol, the additive force field ff19SB, and the polarizable force field Amoeba13 are ranked forth to sixth, with RMSE of 2.28, 2.67, and 2.91 kcal/mol, respectively. The observation that Amoeba13 performs worse than ff12pol in this test set is somewhat surprising, given that Amoeba13 is such an elaborate force field that includes atomic permanent dipoles and quadrupoles, in addition to the polarizable induced dipoles.[26, 28] In contrast, the ff12pol force field is a minimalist polarizable induced dipole force field with neither permanent dipoles nor quadrupoles.[22-25]

Another interesting observation is that, compared with the ωB97X-D results, the pGM-perm and pGM-ind models systematically overestimate the interaction energies $IE(Gly_m:Gly_n)$. On average, pGM-perm and pGM-ind overestimate the interaction energies by 1.32 and 1.34 kcal/mol, respectively. Also interesting is the consistency of the deviations between the interaction energies of ωB97X-D and the pGM models across different conformations. Since the mainchain hydrogen bonds contribute to peptide secondary structure formations, the balance across different conformations can influence the peptide secondary structure preference and the capability of modeling relative strength of different hydrogen bonding systems and in peptide mainchain secondary structures. In this regard,

both pGM models show good balance and their differences in the interaction energies are about the same magnitude across the three conformations, with pGM-perm exhibiting slightly better consistency than pGM-ind. For example, the Gly:Gly interaction energies are overestimated by 1.48 to 1.88 kcal/mol by pGM-perm and overestimated by 1.36 to 1.92 kcal/mol by pGM-ind. In contrast, all other force fields show non-uniform deviations across different conformations. Taking ff12pol as an example, the largest deviation of interaction energies among the three conformations consistently come from the oligomers in the α-helix conformation. For Amoeba13, the deviations of α-helix and parallel β-sheet conformers are comparable, whereas the deviations of anti-parallel β-sheet conformer is notably smaller. Since uniform deviations across different conformers naturally avoid introducing conformational bias, it is much more preferred than non-uniform deviations. Furthermore, when taking all Gly_m:Gly_n oligomers into account, the deviations of interaction energies range between -0.96 and -1.88 kcal/mol for pGM-perm, and between -0.90 to -1.92 kcal/mol for pGM-ind, which are more consistent than other force fields. On the other hand, all other force fields (except ff15ipq) exhibit a tendency of growing deviations with increasing size of oligomers, suggesting that they underestimate the many-body interactions when there are multiple peptides in the oligomers. Therefore, it is encouraging that the pGM models outperform all other five force fields in terms of interaction energy calculations across oligomers with different conformations and with different sizes.

Table 5.2. The Interaction Energies IE(Gly_m: Gly_n) of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Molecular Mechanical Force Fields (kcal/mol)[a]

ωB97X-D/aTZ		pGM-perm	pGM-ind	Amoeba13	ff12pol	ff19SB	ff15ipq	ff03
Gly:Gly								
αR[b]	-14.73	-16.21 (-1.48)[c]	-16.09 (-1.36)	-12.98 (1.75)	-12.91 (1.81)	-16.04 (-1.31)	-19.08 (-4.35)	-15.35 (-0.62)
aβ[b]	-14.53	-16.01 (-1.48)	-16.13 (-1.60)	-13.82 (0.71)	-14.67 (-0.15)	-13.60 (0.92)	-16.31 (-1.78)	-11.83 (2.70)
β[b]	-17.36	-19.24 (-1.88)	-19.28 (-1.92)	-14.16 (3.20)	-17.82 (-0.46)	-15.69 (1.67)	-18.85 (-1.48)	-14.78 (2.58)
Gly:Gly_2								
αR[b]	-17.90	-19.21 (-1.31)	-19.11 (-1.22)	-15.38 (2.52)	-14.94 (2.96)	-17.90 (-0.01)	-21.30 (-3.40)	-17.09 (0.80)
aβ[b]	-16.00	-17.26 (-1.26)	-17.38 (-1.38)	-14.89 (1.11)	-15.68 (0.32)	-14.06 (1.94)	-16.83 (-0.83)	-12.28 (3.72)
β[b]	-18.99	-20.64 (-1.65)	-20.67 (-1.69)	-15.42 (3.57)	-18.97 (0.02)	-16.19 (2.80)	-19.45 (-0.46)	-15.26 (3.73)
Gly:Gly_3								
αR[b]	-19.06	-20.32 (-1.26)	-20.23 (-1.18)	-16.26 (2.80)	-15.68 (3.37)	-18.42 (0.63)	-21.93 (-2.87)	-17.57 (1.48)
aβ[b]	-16.12	-17.36 (-1.24)	-17.47 (-1.35)	-14.95 (1.17)	-15.74 (0.38)	-14.09 (2.03)	-16.86 (-0.74)	-12.31 (3.81)
β[b]	-19.16	-20.79 (-1.63)	-20.83 (-1.67)	-15.56 (3.61)	-19.08 (0.08)	-16.25 (2.91)	-19.52 (-0.36)	-15.31 (3.85)
Gly_2:Gly_2								
αR[b]	-22.11	-23.20 (-1.09)	-23.12 (-1.01)	-18.54 (3.57)	-17.71 (4.40)	-20.30 (1.81)	-24.16 (-2.04)	-19.36 (2.75)
aβ[b]	-17.72	-18.75 (-1.03)	-18.85 (-1.13)	-16.21 (1.51)	-16.88 (0.84)	-14.61 (3.11)	-17.49 (0.23)	-12.80 (4.92)
β[b]	-21.02	-22.34 (-1.33)	-22.40 (-1.38)	-16.98 (4.03)	-20.27 (0.75)	-16.77 (4.25)	-20.14 (0.87)	-15.80 (5.22)
Gly_3:Gly_3								
αR[b]	-25.73	-26.68 (-0.96)	-26.63 (-0.90)	-21.27 (4.45)	-20.09 (5.64)	-21.91 (3.82)	-26.11 (-0.39)	-20.86 (4.87)
aβ[b]	-18.03	-18.99 (-0.97)	-19.09 (-1.07)	-16.37 (1.66)	-17.04 (0.98)	-14.69 (3.33)	-17.57 (0.46)	-12.90 (5.12)
β[b]	-21.47	-22.74 (-1.27)	-22.80 (-1.32)	-17.33 (4.14)	-20.56 (0.91)	-16.92 (4.55)	-20.33 (1.14)	-15.94 (5.53)
RMSE		1.35	1.37	2.91	2.28	2.67	1.87	3.78
MAE		1.32	1.34	2.65	1.54	2.34	1.43	3.45
Max Dev[c]		-0.96	-0.90	4.45	5.64	4.55	1.14	5.53
Min Dev[c]		-1.88	-1.92	0.71	-0.46	-1.31	-4.35	-0.62

[a] The formula for the interaction energies IE(Gly_m: Gly_n) is shown in **eq 5.21**. [b] αR, aβ, and β represent α-helix, anti-parallel β-sheet, and parallel β-sheet conformations, respectively. [c] The maximum and minimum deviations with ωB97X-D/aTZ results. [c] The values in parentheses are differences between the values calculated by molecular mechanical force fields and corresponding values calculated by the ωB97X-D/aTZ method.

5.4.3. pGM Models Most Accurately Reproduce QM Many-Body Interaction Energies

Subtracting corresponding rows of the Gly:Gly dimers from other rows in **Table 5.2** gives **Table 5.3**, which lists the many-body interaction energies ME(Gly_m: Gly_n) calculated by the seven force fields compared with those calculated by the ωB97X-D method. The many-body interaction energies defined in **eq 5.22** describe the overall (additive and non-additive) contributions from the neighboring glycine dipeptides (Gly_{m-1} and Gly_{n-1}, excluding the

Gly:Gly at the interface) to the dimerization energy at the interface of Gly_m: Gly_n. As shown in the ωB97X-D results, the many-body interaction energies again increase in the order of Gly:Gly_2, Gly:Gly_3, Gly_2:Gly_2, Gly_3:Gly_3 for all the three conformations. Comparing oligomers in the anti-parallel and parallel β-sheet conformations, we can see that the addition of outer peptides does not significantly increase the many-body interactions. Thus, the cross-strand effects in the β-sheet conformations are mainly limited to those in close contact and diminish rather quickly with distance. For the α-helix conformation, notably stronger many-body interaction is observed, because the hydrogen bonds are aligned in the same directions. An interesting observation is that the many-body interaction energy of tetramer Gly:Gly_3 is only marginally stronger than that of trimer Gly:Gly_2 by 1.16 kcal/mol, which is much smaller increase compared to the 4.21 kcal/mol increase of tetramer Gly_2:Gly_2. By adding one more peptide at each side, the many-body interaction energy of hexamer Gly_3:Gly_3 becomes stronger by 3.62 kcal/mol than that of tetramer Gly_2:Gly_2. Therefore, for the α-helix conformation, in contrast to the marginal effect of adding peptides to one side of the interface, symmetric addition makes the interaction at the interface significantly stronger, so that much stronger many-body interaction is expected in the inner part of α-helices. Therefore, the outer and inner parts of α-helices could have considerably different stabilities. This effect could be significant in non-polar environments such as transmembrane proteins.

Among the seven force fields tested, pGM-perm and pGM-ind again show the lowest RMSEs (0.40 and 0.38 kcal/mol, respectively) and the lowest MAEs (0.37 and 0.35 kcal/mol, respectively), making them the best force fields in terms of many-body interaction energy

calculations. It is encouraging that the RMSEs and MAEs of both pGM models are lower than the thermal fluctuation energy of 0.60 kcal/mol at 300K. Similar to the case of interaction energy calculations, both pGM models give similar performances in terms of many-body interaction energy calculations. This indicates a potential advantage of the ESP fitting strategy employed for pGM parameterizations. Both the interaction energies and many-body interaction energies between molecules are largely dependent on the electrostatic interactions, and the ESP surrounding molecules are one of the most important electrostatic properties. Since both the pGM-ind and pGM-perm models are able to reproduce QM ESPs with low errors,[52] it is expected that both models can accurately reproduce QM interaction energies and many-body interaction energies, therefore giving similar performances.

The next best performing force field is the Amoeba13 force field, which exhibits the largest improvement compared with interaction energy calculations in **Table 5.2**, with the RMSE reduced from 2.91 kcal/mol to 1.16 kcal/mol. The significant improvement of Amoeba13 shows that the short-range interactions are the main cause of the large errors observed in the interaction energies. The ff12pol force field is ranked the third best performing force field, with an RMSE of 1.62 kcal/mol. It is remarkable that all polarizable force fields perform better than all additive force fields. In fact, all additive force fields notably underestimate the many-body interactions by more than 2.00 kcal/mol. Among the additive force fields tested, ff15ipq once again shows the best performance, with an RMSE of 2.04 kcal/mol. However, this RMSE is slightly higher than that of interaction energies (1.87 kcal/mol) given by ff15ipq.

Another observation in **Table 5.3** is that all force fields consistently underestimate the many-body interaction energies ME(Gly_m: Gly_n) when compared with the ωB97X-D results. For the pGM models, this is in sharp contrast to the systematic overestimations in the interaction energies IE(Gly_m: Gly_n) as shown in **Table 5.2**, indicating that the long-range terms are still under-represented in the pGM models. Moreover, this suggests that the overestimations of IE(Gly_m: Gly_n) of the pGM models are primarily due to the short-range van der Waals terms. Because of this, we anticipate that the present van der Waals parameters, which were taken directly from ff12pol without optimization, need to be tuned to make the short-range terms less attractive. For Amoeba13 and ff12pol, however, since their IE(Gly_m: Gly_n) and ME(Gly_m: Gly_n) are systematically weaker than the ωB97X-D results and the errors in IE(Gly_m: Gly_n) are larger than ME(Gly_m: Gly_n), it appears that these two force fields could be improved by strengthening both their short-range and polarization terms.

Similar to the case of interaction energies, the underestimations of the pGM models compared to ωB97X-D across different conformations and across different oligomers are consistent, which range between 0.17 and 0.61 kcal/mol for pGM-perm, and between 0.15 to 0.59 kcal/mol for pGM-ind. In contrast, all other force fields once again show non-uniform deviations across different conformations, and the deviations increase with the size of oligomers. It is notable that the additive ff15ipq force field, which exhibit relatively consistent deviations in terms of interaction energies, also show gradually increasing deviations with the oligomer size in terms of many-body interaction energies. Based on above observations, we conclude that the polarization effects play critical roles in the many-

body interactions, and the additive force fields are, in general, incapable of modeling them accurately.

Table 5.3. The Many-Body Interaction Energies ME(Gly_m:Gly_n) of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Molecular Mechanical Force Fields (kcal/mol)[a, b]

ωB97X-D/aTZ		pGM-perm	pGM-ind	Amoeba13	ff12pol	ff19SB	ff15ipq	ff03
Gly:Gly_2								
αR	-3.17	-3.00 (0.17)	-3.02 (0.15)	-2.40 (0.77)	-2.03 (1.14)	-1.87 (1.30)	-2.22 (0.95)	-1.75 (1.42)
aβ	-1.47	-1.25 (0.22)	-1.24 (0.23)	-1.07 (0.40)	-1.00 (0.47)	-0.46 (1.02)	-0.53 (0.95)	-0.44 (1.03)
β	-1.63	-1.39 (0.23)	-1.39 (0.23)	-1.25 (0.37)	-1.14 (0.48)	-0.50 (1.12)	-0.60 (1.02)	-0.48 (1.15)
Gly:Gly_3								
αR	-4.33	-4.11 (0.22)	-4.14 (0.19)	-3.28 (1.05)	-2.77 (1.56)	-2.39 (1.94)	-2.85 (1.48)	-2.23 (2.10)
aβ	-1.59	-1.35 (0.24)	-1.34 (0.25)	-1.14 (0.46)	-1.07 (0.52)	-0.49 (1.11)	-0.55 (1.04)	-0.48 (1.11)
β	-1.80	-1.55 (0.25)	-1.55 (0.25)	-1.39 (0.41)	-1.26 (0.54)	-0.56 (1.24)	-0.67 (1.13)	-0.53 (1.27)
Gly_2:Gly_2								
αR	-7.38	-6.99 (0.39)	-7.03 (0.35)	-5.57 (1.82)	-4.80 (2.58)	-4.27 (3.12)	-5.08 (2.31)	-4.01 (3.37)
aβ	-3.19	-2.74 (0.45)	-2.72 (0.47)	-2.39 (0.80)	-2.20 (0.99)	-1.01 (2.19)	-1.19 (2.01)	-0.97 (2.22)
β	-3.66	-3.10 (0.55)	-3.12 (0.54)	-2.82 (0.84)	-2.45 (1.21)	-1.08 (2.57)	-1.30 (2.36)	-1.02 (2.64)
Gly_3:Gly_3								
αR	-11.00	-10.47 (0.53)	-10.53 (0.47)	-8.29 (2.71)	-7.18 (3.82)	-5.88 (5.12)	-7.03 (3.97)	-5.51 (5.49)
aβ	-3.50	-2.98 (0.51)	-2.96 (0.53)	-2.55 (0.95)	-2.37 (1.13)	-1.09 (2.41)	-1.26 (2.24)	-1.07 (2.43)
β	-4.11	-3.50 (0.61)	-3.52 (0.59)	-3.17 (0.94)	-2.74 (1.37)	-1.23 (2.87)	-1.48 (2.63)	-1.16 (2.95)
RMSE		0.40	0.38	1.16	1.62	2.45	2.04	2.58
MAE		0.37	0.35	0.96	1.32	2.17	1.84	2.26
Max Dev		0.61	0.59	2.71	3.82	5.12	3.97	5.49
Min Dev		0.17	0.15	0.37	0.47	1.02	0.95	1.03

[a] The formula for the many-body interaction energies ME(Gly_m:Gly_n) is shown in **eq 5.22**. [b] See **Table 5.2** and text for notation.

5.4.4. pGM Models Perform the Best in Reproducing QM Non-additive and Additive Contributions to the Many-Body Interactions

The many-body interaction energies in **Table 5.3** depends on the non-bonded terms in the functional form of each molecular mechanical force fields. For additive force fields such

as Amber ff19SB, ff15ipq and ff03, the many-body interactions only have contributions from the additive electrostatic and van der Waals terms. In polarizable force fields ff12pol and pGM-ind, the non-additive induced dipole polarization energy is also involved. The pGM-perm model has additional energy contributions from atomic permanent dipoles, and the Amoeba13 force field also has contributions from atomic permanent quadrupoles, which are both additive terms. It is difficult to decipher which of these terms play more important role if we just look at the total many-body interaction energies shown in **Table 5.3**. Therefore, we decompose the many-body interaction energies into non-additive and additive contributions to gain insight into these force fields. For each Gly_m:Gly_n oligomer, the formulas of the non-additive contributions $ME_{NA}(Gly_m: Gly_n)$ and the additive contributions $ME_A(Gly_m: Gly_n)$ are given in **eq 5.24** and **eq 5.25**, respectively. The proof of the decomposition is shown in the **Appendix A**. Note that the functional forms of additive force fields only have additive terms, so that the non-additive contribution $ME_{NA}(Gly_m: Gly_n)$ of any additive force field is guaranteed to be zero. For this reason, we will only compare the performances of polarizable force fields pGM-perm, pGM-ind, Amoeba13, and ff12pol in this section.

The interaction energies of the two peptides at the interface, $IE_{mid}(Gly_m: Gly_n)$, in the presence of the neighboring peptides Gly_{m-1} and Gly_{n-1} defined in **eq 5.23**, calculated by the seven force fields and by ωB97X-D are shown in **Table S5.1**. For additive force fields, the values will be identical to those of $IE(Gly: Gly)$ in absence of neighboring peptides Gly_{m-1} and Gly_{n-1} as shown in **Table 5.2**, if listed. For polarizable force fields, a trend similar to that in **Table 5.2** is observed. First, the pGM-perm and pGM-ind models outperform the other two polarizable force fields, with RMSEs of 1.39 and 1.43 kcal/mol, respectively. The RMSEs of

the Amoeba13 force field and the ff12pol force field are 2.54 and 1.69 kcal/mol, respectively. Second, compared with the ωB97X-D results, the pGM-perm and pGM-ind models systematically overestimate the interaction energies (by 1.37 and 1.41 kcal/mol, respectively), whereas both Amoeba13 and ff12pol underestimate (by 2.31 and 1.18 kcal/mol, respectively). Third, the deviations between the interaction energies given by ωB97X-D and the pGM models across different conformations and different oligomers are highly consistent. For pGM-perm, the largest spread (0.41 kcal/mol) comes from Gly:Gly_3 between the anti-parallel β-sheet (-1.25 kcal/mol) and parallel β-sheet (-1.66 kcal/mol) conformers. For pGM-ind, the largest spread (0.56 kcal/mol) is between the α-helix (-1.36 kcal/mol) and parallel β-sheet (-1.92 kcal/mol) conformers of Gly:Gly. Overall, the deviations range between -1.02 and -1.88 kcal/mol for pGM-perm, and between -1.15 to -1.92 kcal/mol for pGM-ind, for all five oligomers and all three conformations. Whereas Amoeba13 and ff12pol show non-uniform deviations across different conformations, and these deviations tend to increase with the size of oligomers.

As shown in the **Appendix A**, the many-body interaction energies $ME(Gly_m:Gly_n)$ can be decomposed to the non-additive contributions $ME_{NA}(Gly_m:Gly_n)$ defined in **eq 5.24** and the additive contributions $ME_A(Gly_m:Gly_n)$ defined in **eq 5.25**. **Table 5.4** shows the non-additive contributions $ME_{NA}(Gly_m:Gly_n)$ of ωB97X-D and the four polarizable force fields obtained by subtracting corresponding rows of the Gly:Gly dimers from other rows in **Table S5.1**, and **Table 5.5** shows the additive contributions $ME_A(Gly_m:Gly_n)$ obtained by subtracting the corresponding non-additive contributions $ME_{NA}(Gly_m:Gly_n)$ from the many-body interaction energies $ME(Gly_m:Gly_n)$ in **Table 5.3**. For non-additive

contributions $ME_{NA}(Gly_m:Gly_n)$, pGM-ind produces the lowest RMSE (0.30 kcal/mol) among all four polarizable force fields, and pGM-perm is the second best, with an RMSE of 0.33 kcal/mol. However, for additive contributions $ME_A(Gly_m:Gly_n)$, pGM-perm performs the best with an RMSE of 0.09 kcal/mol, and pGM-ind is the second best with an RMSE of 0.10 kcal/mol. The significantly lower RMSEs of the additive contributions of the pGM models than other polarizable force fields show the robustness of the ESP fitting scheme of *PyRESP* since the additive contribution is mainly due to the interactions involving fixed point charges and permanent dipoles.[52] For both non-additive and additive contributions, the ff12pol force field gives the worst performance, with RMSEs of 0.95 kcal/mol and 0.72 kcal/mol for the non-additive and additive contributions, respectively. Amoeba13 yields RMSEs of 0.59 kcal/mol for both the non-additive and additive contributions, which are better than ff12pol in terms of both contributions, but still notably worse than the pGM models.

Interestingly, compared with the ωB97X-D results, all four polarizable force fields underestimate both the non-additive and additive contributions to varying degrees. As measured by MAE, the non-additive contributions are underestimated by 0.30, 0.27, 0.54, and 0.85 kcal/mol by pGM-perm, pGM-ind, Amoeba13 and ff12pol, respectively, and the additive contributions are underestimated by 0.06, 0.08, 0.42, and 0.46 kcal/mol by pGM-perm, pGM-ind, Amoeba13 and ff12pol, respectively. Therefore, MAEs show the same performance trend as RMSEs, where pGM-ind performs the best for calculating the non-additive contributions, and pGM-perm performs the best for calculating the additive contributions. The consistent underestimations could potentially come from two sources: inadequate short-range damping and smaller-than-needed polarizabilities, both of which may be improved by further parameterizations. Encouragingly, the pGM models again

exhibit consistent deviations from the ωB97X-D results for both contributions across different conformations and different sized oligomers. For the pGM-perm model, the deviations range between 0.12 and 0.52 kcal/mol for the non-additive contributions, and between 0.01 to 0.22 kcal/mol for the additive contributions; For the pGM-ind model, the deviations range between 0.09 and 0.49 kcal/mol for the non-additive contributions, and between 0.02 to 0.25 kcal/mol for the additive contributions. In contrast, Amoeba13 and ff12pol show significant variances in the deviations from the ωB97X-D results for both non-additive and additive contributions across different conformations. Overall, **Table 5.4-5.5** show that, compared with the other two polarizable force fields tested here, the pGM models (with or without permanent atomic dipoles) perform the best in terms of reproducing QM non-additive and additive contributions to the many-body interaction energies.

Table 5.4. The Non-additive Contributions $ME_{NA}(Gly_m: Gly_n)$ to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Polarizable Force Fields (kcal/mol)[a, b]

ωB97X-D/aTZ		pGM-perm	pGM-ind	Amoeba13	ff12pol
			Gly:Gly_2		
αR	-1.70	-1.57 (0.12)	-1.61 (0.09)	-1.30 (0.39)	-0.98 (0.71)
aβ	-1.04	-0.83 (0.21)	-0.83 (0.21)	-0.76 (0.28)	-0.64 (0.40)
β	-1.12	-0.91 (0.21)	-0.92 (0.20)	-0.86 (0.26)	-0.73 (0.40)
			Gly:Gly_3		
αR	-2.17	-2.02 (0.15)	-2.07 (0.10)	-1.65 (0.52)	-1.27 (0.90)
aβ	-1.12	-0.89 (0.23)	-0.89 (0.22)	-0.80 (0.31)	-0.68 (0.44)
β	-1.21	-0.99 (0.22)	-1.00 (0.21)	-0.95 (0.26)	-0.79 (0.42)
			Gly_2:Gly_2		
αR	-3.52	-3.27 (0.26)	-3.34 (0.18)	-2.68 (0.84)	-2.10 (1.43)
aβ	-2.13	-1.71 (0.42)	-1.72 (0.42)	-1.56 (0.57)	-1.32 (0.81)
β	-2.48	-1.99 (0.49)	-2.02 (0.47)	-1.90 (0.58)	-1.51 (0.97)
			Gly_3:Gly_3		
αR	-4.62	-4.32 (0.31)	-4.41 (0.21)	-3.48 (1.14)	-2.79 (1.83)
aβ	-2.29	-1.83 (0.46)	-1.83 (0.45)	-1.65 (0.64)	-1.40 (0.89)
β	-2.68	-2.16 (0.52)	-2.19 (0.49)	-2.05 (0.63)	-1.62 (1.06)
RMSE		0.33	0.30	0.59	0.95
MAE		0.30	0.27	0.54	0.85
Max Dev		0.52	0.49	1.14	1.83
Min Dev		0.12	0.09	0.26	0.40

[a] The formula for the non-additive contributions $ME_{NA}(Gly_m:Gly_n)$ to the many-body interaction energies is shown in **eq 5.24**. [b] See **Table 5.2** and text for notation.

Table 5.5. The Additive Contributions $ME_A(Gly_m:Gly_n)$ to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Polarizable Force Fields (kcal/mol)[a, b]

ωB97X-D/aTZ		pGM-perm	pGM-ind	Amoeba13	ff12pol
			Gly:Gly_2		
αR	-1.47	-1.43 (0.05)	-1.41 (0.06)	-1.10 (0.38)	-1.04 (0.43)
aβ	-0.43	-0.42 (0.01)	-0.41 (0.02)	-0.31 (0.12)	-0.36 (0.07)
β	-0.50	-0.48 (0.02)	-0.47 (0.03)	-0.39 (0.11)	-0.42 (0.08)
			Gly:Gly_3		
αR	-2.16	-2.09 (0.07)	-2.08 (0.08)	-1.63 (0.53)	-1.50 (0.66)
aβ	-0.48	-0.46 (0.02)	-0.45 (0.03)	-0.33 (0.15)	-0.39 (0.09)
β	-0.59	-0.56 (0.03)	-0.55 (0.04)	-0.44 (0.14)	-0.47 (0.11)
			Gly_2:Gly_2		
αR	-3.86	-3.73 (0.14)	-3.69 (0.17)	-2.88 (0.98)	-2.70 (1.16)
aβ	-1.06	-1.03 (0.03)	-1.01 (0.05)	-0.83 (0.23)	-0.88 (0.17)
β	-1.17	-1.11 (0.06)	-1.10 (0.07)	-0.92 (0.25)	-0.94 (0.23)
			Gly_3:Gly_3		
αR	-6.37	-6.16 (0.22)	-6.12 (0.25)	-4.81 (1.57)	-4.38 (1.99)
aβ	-1.21	-1.15 (0.06)	-1.13 (0.08)	-0.90 (0.31)	-0.97 (0.24)
β	-1.43	-1.34 (0.09)	-1.33 (0.10)	-1.11 (0.32)	-1.12 (0.31)
RMSE		0.09	0.10	0.59	0.72
MAE		0.06	0.08	0.42	0.46
Max Dev		0.22	0.25	1.57	1.99
Min Dev		0.01	0.02	0.11	0.07

[a] The formula for the additive contributions $ME_A(Gly_m : Gly_n)$ to the many-body interaction energies is shown in **eq 5.25**. [b] See **Table 5.2** and text for notation.

5.4.5. pGM Models are Robust with Altered Atomic Polarizabilities

Because of the inherent approximations to either experimental observations or QM calculations, all mechanical force fields are subject to errors that can come from both the functional forms and parameterization processes. In the development of Amber force fields, a consistent electrostatic parameterization approach is to fit the QM calculated ESPs of small molecules or fragments of large molecules to obtain atomic charges and multipoles. Technically, this approach is rather straightforward and allow development of consistent parameters across a wide variety of chemistry. In the cases of polarizable force fields, another advantage is that the errors in the initial fitting of polarizabilities can be partially compensated at the stage when charges and permanent multipoles are calculated, yielding a more robust force field. This feature is potentially advantageous because of the non-linear nature of the polarization energy.

In our previous work, the pGM atomic polarizabilities and radii were obtained by fitting QM molecular polarizability tensors of 1405 molecules or dimers.[45] In this chapter, we further evaluated the robustness of the pGM models by re-parameterizing the glycine dipeptide charges and permanent dipoles using the recently developed *PyRESP* program,[52] with the alternative polarizabilities including the pGM polarizabilities scaled by a factor of 0.9, the Amoeba13 polarizabilities,[26] and the Amber ff12pol polarizabilities.[22] These alternative polarizability sets are either scaled or taken from different sources and have been

developed for different polarization schemes. Therefore, we expect that the energies related to polarization calculated with these three polarizability sets be less accurate than those produced by the pGM models with original pGM polarizabilities shown in **Table 5.2-5.5**. Our objective is to see whether these "wrong" polarizabilities would lead to intolerable errors in energy calculations.

The interaction energies IE(Gly_m:Gly_n) as well as the many-body interaction energies ME(Gly_m:Gly_n) of each Gly_m:Gly_n oligomer of the pGM-perm and pGM-ind models calculated with the alternative polarizabilities are shown in **Table 5.6** and **Table S5.2**, respectively. Interestingly, for both pGM models, the "wrong" polarizabilities produce interaction energies IE(Gly_m:Gly_n) with lower RMSEs compared with those obtained by the "correct" pGM polarizabilities shown in **Table 5.2**. For the pGM-perm model, the overall RMSE of interaction energies decreased from 1.35 kcal/mol to 0.80, 0.82, and 0.62 kcal/mol for the scaled pGM, Amoeba13, and ff12pol polarizabilities, respectively. For the pGM-ind model, the overall RMSE of interaction energies decreased from 1.37 kcal/mol to 0.81, 0.97, and 0.57 kcal/mol for the scaled pGM, Amoeba13, and ff12pol polarizabilities, respectively. The higher RMSEs associated with the "correct" pGM polarizabilities compared with that of the "wrong" polarizabilities might be explained by the fact that the van der Waals parameters for the pGM models were taken directly from the ff12pol force field without any optimization. As shown in **Table 5.2**, with the original polarizabilities, both pGM-perm and pGM-ind overestimate the interaction energies. Since the interaction energies can be decomposed to the electrostatic and van der Waals contributions, the overestimation in the interaction energies can be explained by the overestimation of the van der Waals term in the current pGM models. Specifically, the dispersion effect of the van der Waals term might be

too attractive. As shown in **Table 5.6** and **Table S5.2**, the amount of overestimation in the interaction energies is reduced with the alternative polarizabilities, indicating weaker electrostatic attractions. Consequently, the overestimation of the van der Waals term is compensated by the underestimation of the electrostatic term with the alternative polarizabilities, leading to lower overall RMSEs. Therefore, there is a need to re-parameterize the van der Waals terms, and we anticipate the pGM models with re-parameterized van der Waals terms will give interaction energies with better agreement with QM results than those with the "wrong" polarizabilities.

In contrast to the interaction energies IE(Gly_m: Gly_n), as expected, the many-body interaction energies ME(Gly_m: Gly_n) calculated by the pGM models with the "wrong" polarizabilities are consistently worse than those with the "correct" pGM polarizabilities shown in **Table 5.3**. For the pGM-perm model, the overall RMSE of the many-body interaction energies increased from 0.40 kcal/mol to 0.70, 0.69, and 1.11 kcal/mol for the scaled pGM, Amoeba13, and ff12pol polarizabilities, respectively. Similarly, for the pGM-ind model, the overall RMSE increased from 0.38 kcal/mol to 0.69, 0.68, and 1.10 kcal/mol for the scaled pGM, Amoeba13, and ff12pol polarizabilities, respectively. Remarkably, the many-body interaction energies produced by the pGM models with alternative polarizabilities consistently outperform the Amoeba13 and ff12pol force fields with their respective native polarizabilities shown in **Table 5.3**. With the Amoeba13 polarizabilities, the RMSEs of the Amoeba13 force field, the pGM-perm and pGM-ind models are 1.16, 0.69, and 0.68 kcal/mol, respectively; with the ff12pol polarizabilities, the RMSEs of the ff12pol force field, the pGM-perm and pGM-ind models are 1.62, 1.11, and 1.10 kcal/mol, respectively. Because short-range terms contribute much less to ME(Gly_m: Gly_n) than to IE(Gly_m: Gly_n), these

improvements are likely attributable to the differences in the treatment of long-range terms, including the electrostatic and polarization terms. This shows that the pGM models are highly robust in terms of modeling the many-body interactions of peptide mainchain hydrogen bonding structures. The improvement is remarkable, given the substantial differences among the different force fields in their functional forms of the electrostatic and polarization terms. Both the Amoeba13 and ff12pol force fields are based on the Thole screening schemes, in which only the cross induction between the induced dipoles is screened to avoid polarization catastrophe. In the pGM models, all electrostatic terms are represented as Gaussian densities. Consequently, all electrostatic interactions are screened, including charge-charge, charge-dipole, and dipole-dipole interactions. The improvement observed in this comparison is likely attributable to the inherent consistency of the treatment of electrostatic terms in the pGM models.

The non-additive and additive contributions to the many-body interactions calculated by the pGM-perm and pGM-ind models with the alternative polarizabilities are shown in **Table 5.7** and **Table S5.3**, respectively. With the scaled pGM polarizabilities, the RMSEs of the non-additive contributions of the pGM-perm and pGM-ind models increase from 0.33 and 0.30 kcal/mol to 0.60 and 0.58 kcal/mol, respectively. Despite the fact that the functional forms of the polarization terms of the pGM models are different from those of the Amoeba13 and ff12pol force fields, the non-additive contributions of the pGM models with the Amoeba13 and ff12pol polarizabilities are remarkably similar to Amoeba13 and ff12pol with their respective native polarizabilities. With the Amoeba13 polarizabilities, the RMSEs of the Amoeba13 force field, the pGM-perm and pGM-ind models are 0.59, 0.59, and 0.57 kcal/mol, respectively; with the ff12pol polarizabilities, the RMSEs of the ff12pol force field,

the pGM-perm and pGM-ind models are 0.95, 0.96, and 0.95 kcal/mol, respectively. We therefore conclude that the changes in the functional forms in the calculations of induction energies from their respective native forms did not lead to intolerable level of error.

For the additive contribution calculations, the RMSEs of the pGM-perm and pGM-ind models with the scaled pGM polarizabilities increase from 0.09 and 0.10 kcal/mol to 0.12 and 0.13 kcal/mol, respectively, which are essentially unchanged. Remarkably, both pGM models with the Amoeba13 and ff12pol polarizabilities notably outperform their respective native counterparts (Amoeba13 and ff12pol force fields) in the calculation of the additive contributions. With the Amoeba13 polarizabilities, the RMSEs of the Amoeba13 force field, the pGM-perm and pGM-ind models are 0.59, 0.12, and 0.13 kcal/mol, respectively; with the ff12pol polarizabilities, the RMSEs of the ff12pol force field, the pGM-perm and pGM-ind models are 0.72, 0.17, and 0.18 kcal/mol, respectively. In fact, for the additive contribution calculations, the RMSEs of both pGM models with the alternative polarizabilities are comparable to those with the original polarizabilities. The notably better performance of the pGM models with the Amoeba13 and ff12pol polarizabilities than the Amoeba13 and ff12pol force fields in the additive contribution calculations suggest that the *PyRESP* scheme of fitting charges and permanent multipoles from QM calculated ESPs is a reliable approach in the development of molecular mechanical force fields and has the ability to compensate the errors in the initial parameterization of polarizabilities.

Table 5.6. The Interaction Energies IE(Gly_m: Gly_n) and Many-Body Interaction Energies ME(Gly_m: Gly_n) of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by the pGM-perm Model with the Alternative Polarizabilities (kcal/mol)[a, b]

	pGM-perm/scaled[c]		pGM-perm/Amoeba[c]		pGM-perm/ff12pol[c]	
	IE(Gly_m: Gly_n)	ME(Gly_m: Gly_n)	IE(Gly_m: Gly_n)	ME(Gly_m: Gly_n)	IE(Gly_m: Gly_n)	ME(Gly_m: Gly_n)
Gly:Gly						
αR	-16.06 (-1.33)		-16.13 (-1.40)		-15.86 (-1.13)	
aβ	-15.58 (-1.06)		-15.60 (-1.07)		-15.01 (-0.48)	
β	-18.64 (-1.28)		-18.62 (-1.26)		-17.75 (-0.39)	
Gly:Gly_2						
αR	-18.89 (-0.99)	-2.83 (0.34)	-18.93 (-1.04)	-2.80 (0.36)	-18.44 (-0.54)	-2.58 (0.59)
aβ	-16.72 (-0.72)	-1.13 (0.34)	-16.75 (-0.75)	-1.14 (0.33)	-15.99 (0.01)	-0.98 (0.49)
β	-19.90 (-0.91)	-1.26 (0.37)	-19.89 (-0.90)	-1.27 (0.36)	-18.84 (0.15)	-1.09 (0.54)
Gly:Gly_3						
αR	-19.91 (-0.85)	-3.85 (0.48)	-19.96 (-0.91)	-3.83 (0.50)	-19.35 (-0.30)	-3.49 (0.84)
aβ	-16.80 (-0.68)	-1.21 (0.38)	-16.83 (-0.71)	-1.23 (0.36)	-16.06 (0.06)	-1.05 (0.54)
β	-20.04 (-0.88)	-1.40 (0.40)	-20.03 (-0.87)	-1.41 (0.39)	-18.96 (0.20)	-1.21 (0.59)
Gly_2:Gly_2						
αR	-22.64 (-0.53)	-6.58 (0.80)	-22.67 (-0.56)	-6.54 (0.84)	-21.86 (0.25)	-6.00 (1.38)
aβ	-18.06 (-0.34)	-2.47 (0.72)	-18.10 (-0.38)	-2.50 (0.69)	-17.15 (0.57)	-2.14 (1.05)
β	-21.43 (-0.42)	-2.79 (0.86)	-21.43 (-0.41)	-2.81 (0.85)	-20.15 (0.87)	-2.40 (1.25)
Gly_3:Gly_3						
αR	-25.85 (-0.12)	-9.79 (1.21)	-25.90 (-0.17)	-9.77 (1.23)	-24.72 (1.01)	-8.86 (2.14)
aβ	-18.27 (-0.25)	-2.69 (0.81)	-18.33 (-0.30)	-2.73 (0.77)	-17.33 (0.69)	-2.32 (1.17)
β	-21.78 (-0.31)	-3.14 (0.97)	-21.80 (-0.33)	-3.18 (0.93)	-20.45 (1.02)	-2.71 (1.40)
RMSE	0.80	0.70	0.82	0.69	0.62	1.11
MAE	0.71	0.64	0.74	0.63	0.51	1.00
Max Dev	-0.12	1.21	-0.17	1.23	1.02	2.14
Min Dev	-1.33	0.34	-1.40	0.33	-1.13	0.49

[a] See **Table 5.2** and text for notation. [b] The values in parentheses are differences between the values calculated by the pGM-perm model with the alternative polarizabilities and corresponding values calculated by the ωB97X-D/aTZ method, which can be found in **Table 5.2** and **Table 5.3**, respectively. [c] pGM-perm/scaled, pGM-perm/Amoeba, and pGM-perm/ff12pol represent the pGM-perm models with the pGM polarizabilities scaled by a factor of 0.9, the Amoeba13 polarizabilities and the ff12pol polarizabilities, respectively.

Table 5.7. The Non-additive Contributions ME_{NA}(Gly_m: Gly_n) and Additive Contributions ME_A(Gly_m: Gly_n) to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers

Gly_m:Gly_n Calculated by the pGM-perm Model with the Alternative Polarizabilities (kcal/mol)[a, b]

	pGM-perm/scaled[c]		pGM-perm/Amoeba[c]		pGM-perm/ff12pol[c]	
	$ME_{NA}(Gly_m:Gly_n)$	$ME_A(Gly_m:Gly_n)$	$ME_{NA}(Gly_m:Gly_n)$	$ME_A(Gly_m:Gly_n)$	$ME_{NA}(Gly_m:Gly_n)$	$ME_A(Gly_m:Gly_n)$
$Gly:Gly_2$						
αR	-1.38 (0.31)	-1.44 (0.03)	-1.37 (0.33)	-1.44 (0.04)	-1.12 (0.58)	-1.46 (0.01)
aβ	-0.71 (0.33)	-0.42 (0.01)	-0.73 (0.32)	-0.42 (0.01)	-0.56 (0.48)	-0.42 (0.01)
β	-0.78 (0.35)	-0.48 (0.02)	-0.79 (0.33)	-0.48 (0.02)	-0.61 (0.51)	-0.48 (0.02)
$Gly:Gly_3$						
αR	-1.77 (0.40)	-2.08 (0.08)	-1.75 (0.42)	-2.08 (0.08)	-1.42 (0.75)	-2.07 (0.09)
aβ	-0.76 (0.36)	-0.46 (0.02)	-0.78 (0.34)	-0.45 (0.02)	-0.60 (0.52)	-0.45 (0.03)
β	-0.84 (0.37)	-0.55 (0.03)	-0.86 (0.35)	-0.55 (0.03)	-0.66 (0.55)	-0.55 (0.04)
$Gly_2:Gly_2$						
αR	-2.87 (0.65)	-3.71 (0.15)	-2.84 (0.68)	-3.70 (0.16)	-2.31 (1.21)	-3.69 (0.17)
aβ	-1.46 (0.67)	-1.02 (0.04)	-1.49 (0.64)	-1.01 (0.05)	-1.15 (0.99)	-1.00 (0.06)
β	-1.69 (0.79)	-1.10 (0.07)	-1.71 (0.77)	-1.10 (0.07)	-1.32 (1.17)	-1.08 (0.09)
$Gly_3:Gly_3$						
αR	-3.76 (0.86)	-6.03 (0.35)	-3.74 (0.89)	-6.03 (0.34)	-3.00 (1.62)	-5.86 (0.52)
aβ	-1.55 (0.73)	-1.13 (0.08)	-1.59 (0.69)	-1.13 (0.08)	-1.22 (1.07)	-1.10 (0.11)
β	-1.82 (0.85)	-1.32 (0.11)	-1.86 (0.82)	-1.32 (0.11)	-1.42 (1.26)	-1.29 (0.14)
RMSE	0.60	0.12	0.59	0.12	0.96	0.17
MAE	0.56	0.08	0.55	0.09	0.89	0.11
Max Dev	0.86	0.35	0.89	0.34	1.62	0.52
Min Dev	0.31	0.01	0.32	0.01	0.48	0.01

[a] See **Table 5.6** and text for notation. [b] The values in parentheses are differences between the values calculated by the pGM-perm model with the alternative polarizabilities and corresponding values calculated by the ωB97X-D/aTZ method, which can be found in **Table 5.4** and **Table 5.5**, respectively.

5.5. Conclusions

In this chapter, we assessed the capabilities of the recently developed pGM models[46-47] in modeling the many-body interactions of glycine dipeptides mainchain hydrogen bonding conformers. Two types of pGM models were considered, including that with (pGM-perm) and without (pGM-ind) permanent atomic dipoles. The performances of the pGM models were compared with several other widely used force fields, including Amoeba13,[28]

ff12pol,[22-25] ff19SB,[7] ff15ipq[54] and ff03.[55] The glycine dipeptide oligomers were selected as the model systems since glycine has the minimalist side chain so that we can focus on mainchain hydrogen bonding interactions.

We first identified ωB97X-D/aug-cc-pVTZ without counterpoise BSSE correction as the most suitable DFT method for our molecular systems. Compared with other DFT methods tested (M062X and B3LYP) with and without counterpoise corrections, ωB97X-D without counterpoise correction produced the interaction energies of the formamide dimer and the glycine dipeptide dimers with the best agreement to those calculated at the CCSD(T)/CBS level of theory.

Next, we compared the interaction energies IE(Gly_m: Gly_n) and many-body interaction energies ME(Gly_m: Gly_n) calculated at the ωB97X-D/aug-cc-pVTZ level of theory and those calculated by the seven molecular mechanical force fields. The overall RMSEs of the interaction energies and many-body interaction energies of the seven force fields are shown in **Figure 5.2**. Encouragingly, the overall RMSEs of the interaction energies IE(Gly_m: Gly_n) calculated by the pGM-perm and pGM-ind models are 1.35 and 1.37 kcal/mol, respectively, which significantly outperform other polarizable (Amoeba13, 2.91 kcal/mol; ff12pol, 2.28 kcal/mol) and additive (ff19SB, 2.67 kcal/mol; ff15ipq, 1.87 kcal/mol; ff03, 3.78 kcal/mol) force fields. For the many-body interaction energies ME(Gly_m: Gly_n), the overall RMSEs of the pGM-perm and pGM-ind models are 0.40 and 0.38 kcal/mol, respectively. In comparison, the RMSEs of other polarizable (Amoeba13, 1.16 kcal/mol; ff12pol, 1.62 kcal/mol) and additive (ff19SB, 2.45 kcal/mol; ff15ipq, 2.04 kcal/mol; ff03, 2.58 kcal/mol) force fields are notably higher than that of the pGM models. In addition, for both interaction

energies and many-body interaction energies, the deviations between the ωB97X-D results and the pGM models results across different conformations and oligomers with different sizes are highly consistent, while all other force fields exhibit non-uniform deviations across different conformations, and these deviations increase with the size of oligomers. Therefore, our data show that the pGM models perform the best among all seven tested force fields in terms of calculating interaction energy and many-body interaction energy.

For polarizable force fields, the many-body interaction energy can be decomposed into the non-additive contribution $ME_{NA}(Gly_m:Gly_n)$ and the additive contribution $ME_{A}(Gly_m:Gly_n)$, so that we compared both contributions calculated by the four polarizable force fields with those of ωB97X-D calculations. **Figure 5.2** shows the overall RMSEs of the non-additive and additive contributions to the many-body interaction energies of the four polarizable force fields. Encouragingly, the pGM models result in the lowest RMSEs for both non-additive (pGM-perm, 0.33 kcal/mol; pGM-ind, 0.30 kcal/mol) and additive (pGM-perm, 0.09 kcal/mol; pGM-ind, 0.10 kcal/mol) contributions. In comparison, the Amoeba13 force field gives RMSEs of 0.59 kcal/mol for both the non-additive and additive contributions. The ff12pol force field gives RMSEs of 0.95 kcal/mol for the non-additive contribution, and 0.72 kcal/mol for the additive contribution. Therefore, the pGM models perform the best among all tested polarizable force fields in terms of modeling both the non-additive and additive contributions to the many-body interactions.

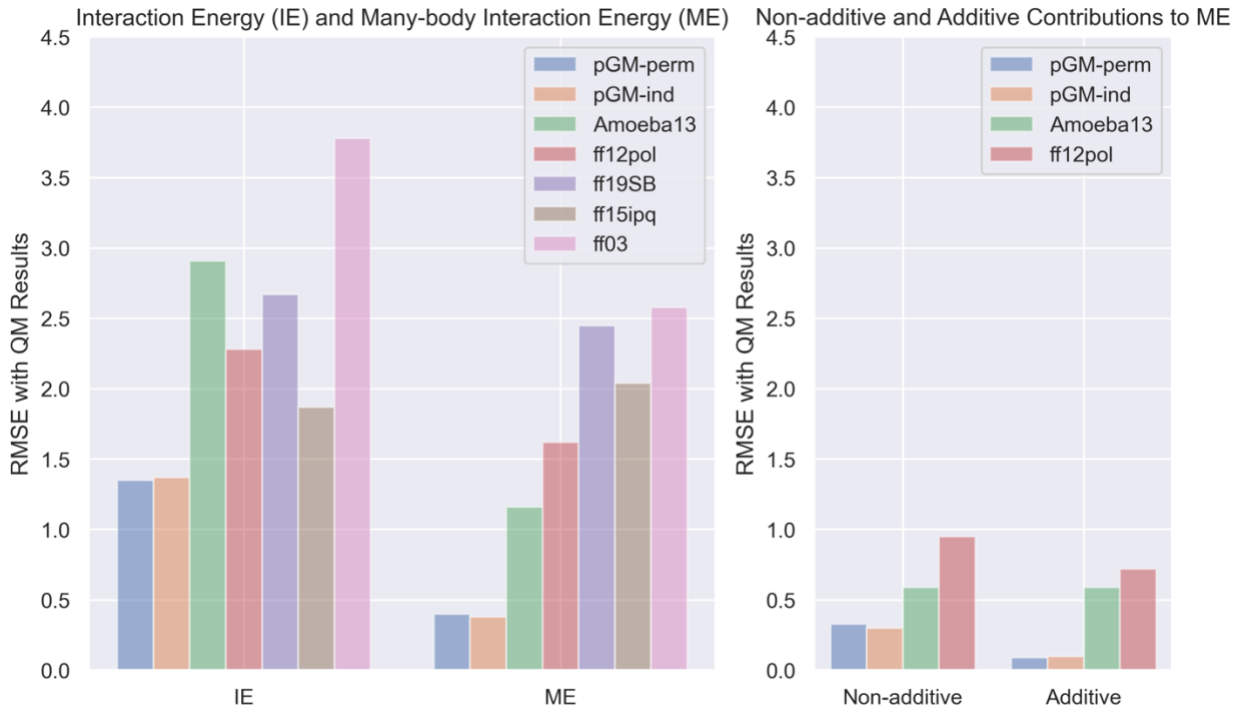


Figure 5.2. Overall RMSEs of the interaction energies IE(Gly_m:Gly_n), many-body interaction energies ME(Gly_m:Gly_n) (left) as well as the non-additive contribution ME_{NA}(Gly_m:Gly_n) and the additive contribution ME_A(Gly_m:Gly_n) to the many-body interaction energies (right) of the tested force fields with the ωB97X-D/aug-cc-pVTZ calculated results.

Finally, we tested the robustness of the pGM models against parameterization errors by employing alternative polarizabilities. Interestingly, the pGM models with the alternative polarizabilities produce interaction energies with lower RMSEs compared with those produced by the original pGM polarizabilities. This might be explained by the fact that the current pGM models share identical van der Waals parameters as the ff12pol force field, and the overestimation of the van der Waals term is compensated by the underestimation of the electrostatic term with the alternative polarizabilities. In future works, the van der Waals

parameters of the pGM models will be re-parameterized using similar ways as we did for parameterizing the ff12pol force field.[25] On the other hand, the pGM models with the alternative polarizabilities produce many-body interaction energies as well as the non-additive and additive contributions to the many-body interactions with higher RMSEs compared with those with the original pGM polarizabilities. Even so, both pGM models with the alternative polarizabilities still give better or similar performances compared with the Amoeba13 and ff12pol force fields. Our data show that the pGM models are robust against polarizability errors and perform well even with those "wrong" polarizabilities.

In summary, this chapter validates that the pGM models have the capabilities to accurately model the interaction energies, many-body interaction energies, as well as the non-additive and additive contributions to the many-body interactions of peptide mainchain hydrogen bonding structures. We expect that the pGM models have the potential to serve as templates for developing the next-generation polarizable force fields for modeling various polarization-sensitive biological processes.

5.6. Supporting Information

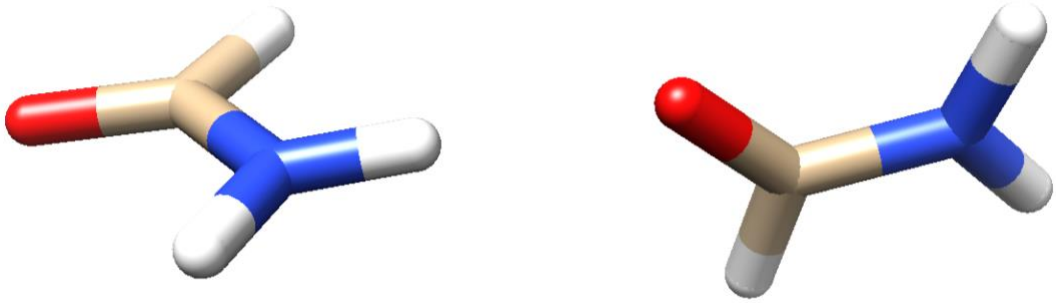


Figure S5.1. The formamide dimer hydrogen bonding conformation.

Figure S5.2. The Gly:Gly dimer hydrogen bonding conformations. Left: α-helix conformation; Middle: parallel β-sheet conformation; Right: anti-parallel β-sheet conformation.

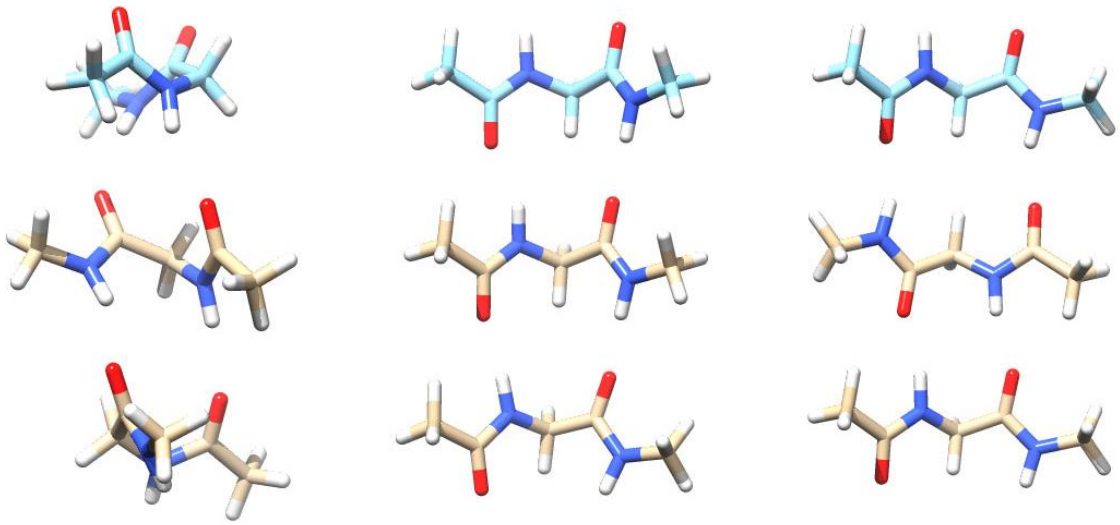


Figure S5.3. The Gly:Gly_2 trimer hydrogen bonding conformations. Left: α-helix conformation; Middle: parallel β-sheet conformation; Right: anti-parallel β-sheet conformation.

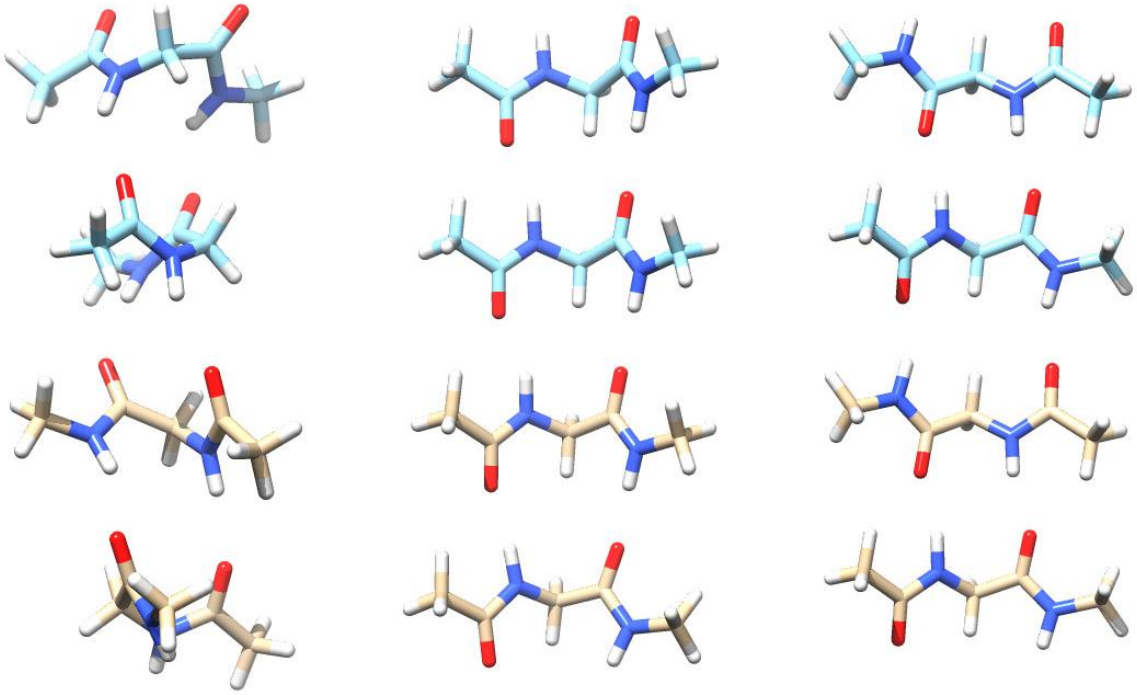


Figure S5.4. The Gly:Gly_3 tetramer hydrogen bonding conformations. Left: α-helix conformation; Middle: parallel β-sheet conformation; Right: anti-parallel β-sheet conformation.

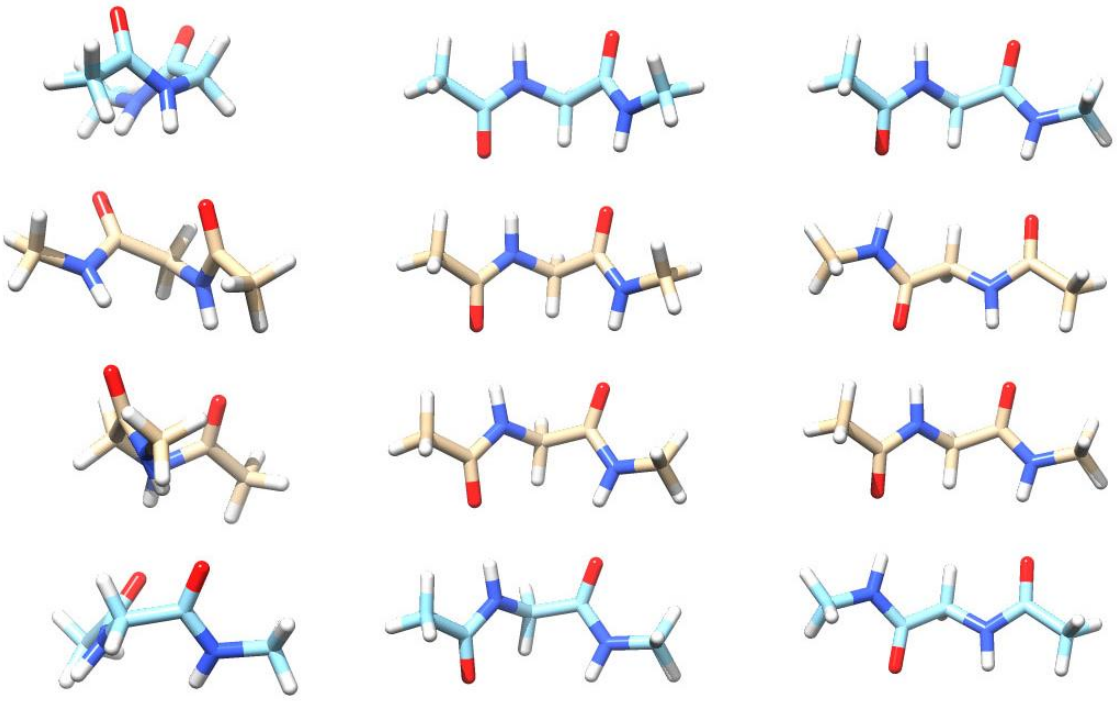


Figure S5.5. The Gly_2:Gly_2 tetramer hydrogen bonding conformations. Left: α-helix conformation; Middle: parallel β-sheet conformation; Right: anti-parallel β-sheet conformation.

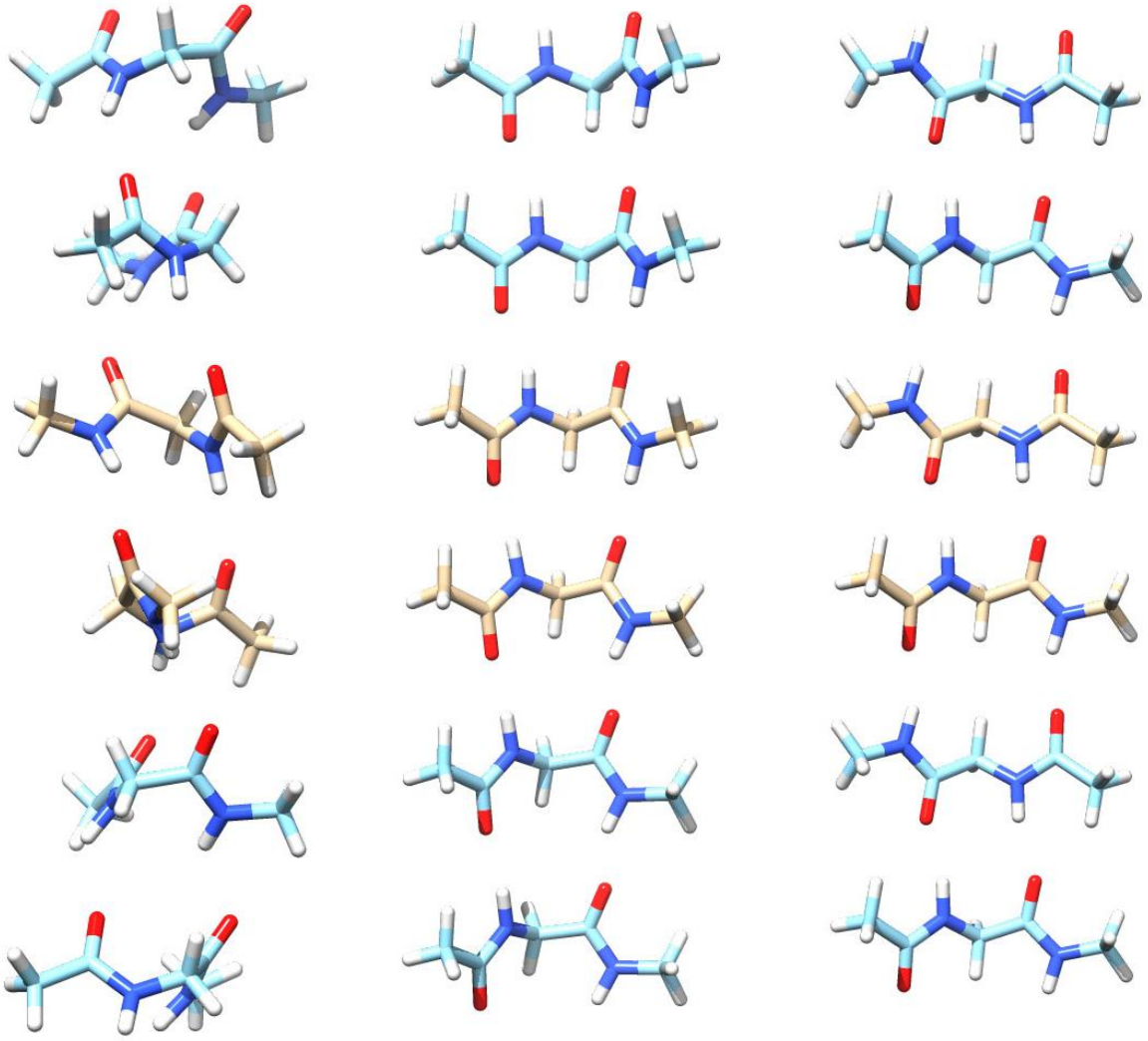


Figure S5.6. The Gly_3:Gly_3 hexamer hydrogen bonding conformations. Left: α-helix conformation; Middle: parallel β-sheet conformation; Right: anti-parallel β-sheet conformation.

Table S5.1. The Interaction Energies of the Two Middle Peptides $IE_{mid}(Gly_m:Gly_n)$ in the Presence of the Neighboring Peptides of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by ωB97X-D/aTZ and Polarizable Force Fields (kcal/mol)[a, b]

ωB97X-D/aTZ		pGM-perm	pGM-ind	Amoeba13	ff12pol
Gly:Gly					
αR	-14.73	-16.21 (-1.48)	-16.09 (-1.36)	-12.98 (1.75)	-12.91 (1.81)
aβ	-14.53	-16.01 (-1.48)	-16.13 (-1.60)	-13.82 (0.71)	-14.67 (-0.15)
β	-17.36	-19.24 (-1.88)	-19.28 (-1.92)	-14.16 (3.20)	-17.82 (-0.46)
Gly:Gly_2					
αR	-16.42	-17.78 (-1.36)	-17.70 (-1.28)	-14.28 (2.14)	-13.90 (2.53)
aβ	-15.57	-16.84 (-1.27)	-16.97 (-1.40)	-14.58 (0.99)	-15.32 (0.25)
β	-18.49	-20.15 (-1.67)	-20.20 (-1.71)	-15.03 (3.46)	-18.55 (-0.06)
Gly:Gly_3					
αR	-16.90	-18.23 (-1.33)	-18.16 (-1.26)	-14.63 (2.27)	-14.19 (2.71)
aβ	-15.64	-16.90 (-1.25)	-17.02 (-1.38)	-14.62 (1.02)	-15.35 (0.29)
β	-18.58	-20.23 (-1.66)	-20.28 (-1.71)	-15.11 (3.46)	-18.61 (-0.04)
Gly_2:Gly_2					
αR	-18.25	-19.48 (-1.22)	-19.43 (-1.18)	-15.66 (2.59)	-15.01 (3.24)
aβ	-16.66	-17.72 (-1.06)	-17.85 (-1.18)	-15.38 (1.28)	-15.99 (0.67)
β	-19.85	-21.23 (-1.39)	-21.30 (-1.45)	-16.07 (3.78)	-19.33 (0.51)
Gly_3:Gly_3					
αR	-19.35	-20.53 (-1.17)	-20.50 (-1.15)	-16.46 (2.89)	-15.71 (3.65)
aβ	-16.82	-17.84 (-1.02)	-17.96 (-1.15)	-15.47 (1.35)	-16.07 (0.75)
β	-20.04	-21.40 (-1.36)	-21.47 (-1.43)	-16.22 (3.82)	-19.44 (0.60)
RMSE		1.39	1.43	2.54	1.69
MAE		1.37	1.41	2.31	1.18
Max Dev		-1.02	-1.15	3.82	3.65
Min Dev		-1.88	-1.92	0.71	-0.46

[a] The formula for the interaction energies of the two middle peptides $IE_{mid}(Gly_m:Gly_n)$ in the presence of the neighboring peptides Gly_{m-1} and Gly_{n-1} is shown in **eq 5.23**. [b] See **Table 5.2** and text for notation.

Table S5.2. The Interaction Energies $IE(Gly_m:Gly_n)$ and Many-Body Interaction Energies $ME(Gly_m:Gly_n)$ of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by the pGM-ind Model with the Alternative Polarizabilities (kcal/mol)[a, b]

	pGM-ind/scaled[c]		pGM-ind/Amoeba[c]		pGM-ind/ff12pol[c]	
	$IE(Gly_m:Gly_n)$	$ME(Gly_m:Gly_n)$	$IE(Gly_m:Gly_n)$	$ME(Gly_m:Gly_n)$	$IE(Gly_m:Gly_n)$	$ME(Gly_m:Gly_n)$
Gly:Gly						
αR	-15.91 (-1.18)		-16.10 (-1.37)		-15.77 (-1.05)	
aβ	-15.72 (-1.19)		-15.90 (-1.37)		-15.25 (-0.72)	
β	-18.70 (-1.34)		-18.83 (-1.47)		-18.01 (-0.65)	
$Gly:Gly_2$						
αR	-18.75 (-0.85)	-2.84 (0.33)	-18.92 (-1.02)	-2.81 (0.35)	-18.35 (-0.46)	-2.58 (0.59)
aβ	-16.84 (-0.84)	-1.13 (0.35)	-17.04 (-1.04)	-1.14 (0.33)	-16.23 (-0.23)	-0.99 (0.49)
β	-19.96 (-0.97)	-1.26 (0.36)	-20.11 (-1.12)	-1.28 (0.35)	-19.12 (-0.13)	-1.10 (0.52)
$Gly:Gly_3$						
αR	-19.77 (-0.72)	-3.87 (0.46)	-19.95 (-0.89)	-3.85 (0.48)	-19.27 (-0.21)	-3.49 (0.83)
aβ	-16.93 (-0.80)	-1.21 (0.39)	-17.13 (-1.01)	-1.23 (0.36)	-16.30 (-0.18)	-1.06 (0.54)
β	-20.10 (-0.94)	-1.40 (0.40)	-20.26 (-1.10)	-1.42 (0.37)	-19.24 (-0.08)	-1.22 (0.57)
$Gly_2:Gly_2$						
αR	-22.51 (-0.40)	-6.60 (0.79)	-22.66 (-0.55)	-6.56 (0.82)	-21.78 (0.33)	-6.01 (1.38)
aβ	-18.18 (-0.46)	-2.46 (0.73)	-18.40 (-0.67)	-2.50 (0.70)	-17.40 (0.32)	-2.16 (1.03)
β	-21.51 (-0.49)	-2.81 (0.85)	-21.67 (-0.65)	-2.84 (0.82)	-20.44 (0.57)	-2.43 (1.22)
$Gly_3:Gly_3$						
αR	-25.72 (0.01)	-9.81 (1.19)	-25.90 (-0.17)	-9.80 (1.20)	-24.63 (1.09)	-8.86 (2.14)
aβ	-18.39 (-0.37)	-2.68 (0.82)	-18.62 (-0.60)	-2.72 (0.77)	-17.58 (0.44)	-2.34 (1.16)
β	-21.86 (-0.39)	-3.17 (0.94)	-22.05 (-0.57)	-3.21 (0.90)	-20.75 (0.72)	-2.74 (1.37)
RMSE	0.81	0.69	0.97	0.68	0.57	1.10
MAE	0.73	0.63	0.91	0.62	0.48	0.99
Max Dev	0.01	1.19	-0.17	1.20	1.09	2.14
Min Dev	-1.34	0.33	-1.47	0.33	-1.05	0.49

[a] See **Table 5.6** and text for notation. [b] The values in parentheses are differences between the values calculated by the pGM-ind model with the alternative polarizabilities and corresponding values calculated by the ωB97X-D/aTZ method, which can be found in **Table 5.2** and **Table 5.3**, respectively. [c] pGM-ind/scaled, pGM-ind/Amoeba, and pGM-ind/ff12pol represent the pGM-ind models with the pGM polarizabilities scaled by a factor of 0.9, the Amoeba13 polarizabilities and the ff12pol polarizabilities, respectively.

Table S5.3. The Non-additive Contributions $ME_{NA}(Gly_m:Gly_n)$ and Additive Contributions $ME_A(Gly_m:Gly_n)$ to the Many-Body Interaction Energies of Glycine Dipeptide Oligomers Gly_m:Gly_n Calculated by the pGM-ind Model with the Alternative Polarizabilities (kcal/mol)[a, b]

	pGM-ind/scaled[c]		pGM-ind/Amoeba[c]		pGM-ind/ff12pol[c]	
	$ME_{NA}(Gly_m:Gly_n)$	$ME_A(Gly_m:Gly_n)$	$ME_{NA}(Gly_m:Gly_n)$	$ME_A(Gly_m:Gly_n)$	$ME_{NA}(Gly_m:Gly_n)$	$ME_A(Gly_m:Gly_n)$
$Gly:Gly_2$						
αR	-1.41 (0.29)	-1.43 (0.04)	-1.39 (0.31)	-1.43 (0.05)	-1.13 (0.57)	-1.46 (0.02)
aβ	-0.71 (0.33)	-0.41 (0.02)	-0.73 (0.31)	-0.41 (0.02)	-0.57 (0.47)	-0.42 (0.01)
β	-0.79 (0.34)	-0.48 (0.03)	-0.80 (0.32)	-0.48 (0.03)	-0.63 (0.50)	-0.48 (0.02)
$Gly:Gly_3$						
αR	-1.79 (0.37)	-2.07 (0.09)	-1.78 (0.39)	-2.07 (0.09)	-1.43 (0.74)	-2.07 (0.09)
aβ	-0.76 (0.36)	-0.45 (0.03)	-0.78 (0.34)	-0.45 (0.03)	-0.60 (0.51)	-0.45 (0.03)
β	-0.86 (0.36)	-0.55 (0.04)	-0.87 (0.34)	-0.55 (0.04)	-0.68 (0.54)	-0.55 (0.04)
$Gly_2:Gly_2$						
αR	-2.91 (0.61)	-3.69 (0.17)	-2.88 (0.64)	-3.68 (0.18)	-2.33 (1.19)	-3.68 (0.18)
aβ	-1.46 (0.67)	-1.00 (0.06)	-1.50 (0.64)	-1.00 (0.06)	-1.16 (0.97)	-1.00 (0.06)
β	-1.72 (0.77)	-1.09 (0.08)	-1.74 (0.74)	-1.10 (0.08)	-1.35 (1.14)	-1.09 (0.09)
$Gly_3:Gly_3$						
αR	-3.81 (0.81)	-6.00 (0.37)	-3.79 (0.83)	-6.01 (0.37)	-3.02 (1.60)	-5.84 (0.54)
aβ	-1.56 (0.73)	-1.12 (0.09)	-1.60 (0.69)	-1.12 (0.09)	-1.24 (1.05)	-1.10 (0.11)
β	-1.85 (0.83)	-1.31 (0.12)	-1.89 (0.79)	-1.32 (0.11)	-1.45 (1.23)	-1.29 (0.14)
RMSE	0.58	0.13	0.57	0.13	0.95	0.18
MAE	0.54	0.09	0.53	0.09	0.88	0.11
Max Dev	0.83	0.37	0.83	0.37	1.60	0.54
Min Dev	0.29	0.02	0.31	0.02	0.47	0.01

[a] See **Table S5.2** and text for notation. [b] The values in parentheses are differences between the values calculated by the pGM-ind model with the alternative polarizabilities and corresponding values calculated by the ωB97X-D/aTZ method, which can be found in **Table 5.4** and **Table 5.5**, respectively.

References

1. Leach, A. R., *Molecular modelling: principles and applications*. 2nd ed.; Pearson education: 2001.
2. Monticelli, L.; Tieleman, D. P., Force fields for classical molecular dynamics. *Biomolecular simulations* **2013**, 197-213.
3. Vitalis, A.; Pappu, R. V., Methods for Monte Carlo simulations of biomacromolecules. *Annual reports in computational chemistry* **2009,** *5*, 49-76.
4. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583-589.
5. Le Grand, S.; Götz, A. W.; Walker, R. C., SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications* **2013,** *184* (2), 374-380.
6. Lee, T.-S.; Cerutti, D. S.; Mermelstein, D.; Lin, C.; LeGrand, S.; Giese, T. J.; Roitberg, A.; Case, D. A.; Walker, R. C.; York, D. M., GPU-accelerated molecular dynamics and free energy methods in

Amber18: performance enhancements and new features. *Journal of chemical information and modeling* **2018,** *58* (10), 2043-2050.
7. Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q., ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of chemical theory and computation* **2019,** *16* (1), 528-552.
8. Brooks, B. R.; Brooks III, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S., CHARMM: the biomolecular simulation program. *Journal of computational chemistry* **2009,** *30* (10), 1545-1614.
9. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* **1996,** *118* (45), 11225-11236.
10. Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J., Polarization effects in molecular mechanical force fields. *Journal of Physics: Condensed Matter* **2009,** *21* (33), 333102.
11. Friesner, R. A., Modeling polarization in proteins and protein–ligand complexes: Methods and preliminary results. *Advances in protein chemistry* **2005,** *72*, 79-104.
12. Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P., Anisotropic, polarizable molecular mechanics studies of inter-and intramolecular interactions and ligand– macromolecule complexes. A bottom-up strategy. *Journal of chemical theory and computation* **2007,** *3* (6), 1960-1986.
13. Zhao, S.; Schaub, A. J.; Tsai, S.-C.; Luo, R., Development of a Pantetheine Force Field Library for Molecular Modeling. *Journal of chemical information and modeling* **2021,** *61* (2), 856-868.
14. King, E.; Qi, R.; Li, H.; Luo, R.; Aitchison, E., Estimating the roles of protonation and electronic polarization in absolute binding affinity simulations. *Journal of chemical theory and computation* **2021,** *17* (4), 2541-2555.
15. Draper, D. E.; Grilley, D.; Soto, A. M., Ions and RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **2005,** *34*, 221-243.
16. Lipfert, J.; Doniach, S.; Das, R.; Herschlag, D., Understanding nucleic acid–ion interactions. *Annual review of biochemistry* **2014,** *83*, 813-841.
17. Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B., Experimental pKa values of buried residues: analysis with continuum methods and role of water penetration. *Biophysical journal* **2002,** *82* (6), 3289-3304.
18. Dill, K. A.; Bromberg, S.; Yue, K.; Chan, H. S.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D., Principles of protein folding—a perspective from simple exact models. *Protein science* **1995,** *4* (4), 561-602.
19. Greatbanks, S. P.; Gready, J. E.; Limaye, A. C.; Rendell, A. P., Enzyme polarization of substrates of dihydrofolate reductase by different theoretical methods. *Proteins: Structure, Function, and Bioinformatics* **1999,** *37* (2), 157-165.
20. Cieplak, P.; Caldwell, J.; Kollman, P., Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of computational chemistry* **2001,** *22* (10), 1048-1057.
21. Wang, Z. X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y., Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *Journal of computational chemistry* **2006,** *27* (6), 781-790.
22. Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations I: parameterization of atomic polarizability. *The journal of physical chemistry B* **2011,** *115* (12), 3091-3099.
23. Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations II: induced dipole models significantly

improve accuracy of intermolecular interaction energies. *The journal of physical chemistry B* **2011,** *115* (12), 3100-3111.
24. Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Luo, R., Development of polarizable models for molecular mechanical calculations. 3. Polarizable water models conforming to Thole polarization screening schemes. *The Journal of Physical Chemistry B* **2012,** *116* (28), 7999-8008.
25. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.-J.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *The Journal of Physical Chemistry B* **2012,** *116* (24), 7088-7101.
26. Ren, P.; Ponder, J. W., Consistent treatment of inter - and intramolecular polarization in molecular mechanics calculations. *Journal of computational chemistry* **2002,** *23* (16), 1497-1506.
27. Ren, P.; Ponder, J. W., Polarizable atomic multipole water model for molecular mechanics simulation. *The Journal of Physical Chemistry B* **2003,** *107* (24), 5933-5947.
28. Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P., Polarizable atomic multipole-based AMOEBA force field for proteins. *Journal of chemical theory and computation* **2013,** *9* (9), 4046-4063.
29. Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B.; Friesner, R. A., Parametrizing a polarizable force field from ab initio data. I. The fluctuating point charge model. *The Journal of chemical physics* **1999,** *110* (2), 741-754.
30. Patel, S.; Brooks III, C. L., CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *Journal of computational chemistry* **2004,** *25* (1), 1-16.
31. Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell Jr, A. D., A polarizable model of water for molecular dynamics simulations of biomolecules. *Chemical Physics Letters* **2006,** *418* (1-3), 245-249.
32. Lopes, P. E.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. *The Journal of Physical Chemistry B* **2007,** *111* (11), 2873-2885.
33. Tan, Y.-H.; Luo, R., Continuum treatment of electronic polarization effect. *J. Chem. Phys.* **2007,** *126* (9), 094103.
34. Tan, Y.-H.; Tan, C.; Wang, J.; Luo, R., Continuum polarizable force field within the Poisson-Boltzmann framework. *J. Phys. Chem. B* **2008,** *112* (25), 7675-7688.
35. Warshel, A.; Levitt, M., Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *Journal of molecular biology* **1976,** *103* (2), 227-249.
36. Vesely, F. J., N-particle dynamics of polarizable Stockmayer-type molecules. *Journal of Computational Physics* **1977,** *24* (4), 361-371.
37. Applequist, J.; Carl, J. R.; Fung, K.-K., Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *Journal of the American Chemical Society* **1972,** *94* (9), 2952-2960.
38. Thole, B. T., Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics* **1981,** *59* (3), 341-350.
39. Van Duijnen, P. T.; Swart, M., Molecular and atomic polarizabilities: Thole's model revisited. *The journal of physical chemistry A* **1998,** *102* (14), 2399-2407.
40. Elking, D.; Darden, T.; Woods, R. J., Gaussian induced dipole polarization model. *Journal of computational chemistry* **2007,** *28* (7), 1261-1274.
41. Elking, D. M.; Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A.; Pedersen, L. G., Gaussian multipole model (GMM). *Journal of chemical theory and computation* **2010,** *6* (1), 190-202.

42. Elking, D. M.; Perera, L.; Duke, R.; Darden, T.; Pedersen, L. G., Atomic forces for geometry-dependent point multipole and Gaussian multipole models. *Journal of computational chemistry* **2010,** *31* (15), 2702-2713.
43. Wheatley, R. J., Gaussian multipole functions for describing molecular charge distributions. *Molecular Physics* **1993,** *79* (3), 597-610.
44. Wheatley, R. J.; Mitchell, J. B., Gaussian multipoles in practice: Electrostatic energies for intermolecular potentials. *Journal of computational chemistry* **1994,** *15* (11), 1187-1198.
45. Wang, J.; Cieplak, P.; Luo, R.; Duan, Y., Development of polarizable Gaussian model for molecular mechanical calculations I: Atomic polarizability parameterization to reproduce ab initio anisotropy. *Journal of chemical theory and computation* **2019,** *15* (2), 1146-1158.
46. Wei, H.; Qi, R.; Wang, J.; Cieplak, P.; Duan, Y.; Luo, R., Efficient formulation of polarizable Gaussian multipole electrostatics for biomolecular simulations. *The Journal of chemical physics* **2020,** *153* (11), 114116.
47. Wei, H.; Cieplak, P.; Duan, Y.; Luo, R., Stress tensor and constant pressure simulation for polarizable Gaussian multipole model. *The Journal of chemical physics* **2022,** *156* (11), 114114.
48. Darden, T.; York, D.; Pedersen, L., Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **1993,** *98* (12), 10089-10092.
49. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A smooth particle mesh Ewald method. *The Journal of chemical physics* **1995,** *103* (19), 8577-8593.
50. Crowley, M.; Darden, T.; Cheatham, T.; Deerfield, D., Adventures in improving the scaling and accuracy of a parallel molecular dynamics program. *The Journal of Supercomputing* **1997,** *11* (3), 255-278.
51. Duke, R. E.; Cisneros, G. A., Ewald-based methods for Gaussian integral evaluation: application to a new parameterization of GEM*. *Journal of molecular modeling* **2019,** *25* (10), 1-9.
52. Zhao, S.; Wei, H.; Cieplak, P.; Duan, Y.; Luo, R., PyRESP: A Program for Electrostatic Parameterizations of Additive and Induced Dipole Polarizable Force Fields. *Journal of Chemical Theory and Computation* **2022,** *18* (6), 3654-3670.
53. Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cisneros, G. A.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shajan, A.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiong, Y.; Xue, Y.; York, D. M.; Zhao, S.; Kollman, P. A., *Amber 2022*. University of California, San Francisco: 2022.
54. Debiec, K. T.; Cerutti, D. S.; Baker, L. R.; Gronenborn, A. M.; Case, D. A.; Chong, L. T., Further along the road less traveled: AMBER ff15ipq, an original protein force field built on a self-consistent physical model. *Journal of chemical theory and computation* **2016,** *12* (8), 3926-3947.
55. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T., A point - charge force field for molecular mechanics simulations of proteins based on condensed - phase quantum mechanical calculations. *Journal of computational chemistry* **2003,** *24* (16), 1999-2012.
56. Chai, J.-D.; Head-Gordon, M., Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Physical Chemistry Chemical Physics* **2008,** *10* (44), 6615-6620.
57. Zhao, Y.; Truhlar, D. G., The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical chemistry accounts* **2008,** *120* (1), 215-241.

58. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical review B* **1988,** *37* (2), 785.
59. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *The Journal of chemical physics* **1993,** *98* (7), 5648-5652.
60. Boys, S. F.; Bernardi, F., The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Molecular Physics* **1970,** *19* (4), 553-566.
61. Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K., Basis-set convergence in correlated calculations on Ne, N2, and H2O. *Chemical Physics Letters* **1998,** *286* (3-4), 243-252.
62. Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Olsen, J., Basis-set convergence of the energy in molecular Hartree–Fock calculations. *Chemical Physics Letters* **1999,** *302* (5-6), 437-446.
63. Connolly, M. L., Analytical molecular surface calculation. *Journal of applied crystallography* **1983,** *16* (5), 548-558.
64. Singh, U. C.; Kollman, P. A., An approach to computing electrostatic charges for molecules. *Journal of computational chemistry* **1984,** *5* (2), 129-145.
65. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J., Gaussian 16 Revision A. 03. 2016; Gaussian Inc. *Wallingford CT* **2016,** *2* (4).
66. Xie, W.; Pu, J.; Gao, J., A coupled polarization-matrix inversion and iteration approach for accelerating the dipole convergence in a polarizable potential function. *The Journal of Physical Chemistry A* **2009,** *113* (10), 2109-2116.
67. Ponder, J. W., TINKER: Software tools for molecular design. *Washington University School of Medicine, Saint Louis, MO* **2004,** *3*.
68. Li, A.; Muddana, H. S.; Gilson, M. K., Quantum mechanical calculation of noncovalent interactions: a large-scale evaluation of PMx, DFT, and SAPT approaches. *Journal of chemical theory and computation* **2014,** *10* (4), 1563-1575.

CHAPTER 6

Transferability of the Electrostatic Parameters of the Polarizable Gaussian Multipole Model

6.1. Introduction

Molecular modeling techniques at atomic level such as molecular dynamics (MD) simulations and Monte Carlo (MC) simulations rely on the development of accurate and transferable molecular mechanical force fields.[1-3] The ability to transfer parameters from one molecule to another molecule or across different conformations of the same molecule is crucial for general purpose force fields that aim at applications to a wide range molecular systems. For this type of force fields, it is of critical importance to accurately reproduce the properties and behaviors of not only the training molecules and conformations used for parameterizations, but also larger testing systems (such as oligomer clusters, molecule complexes, or polymers) and different conformations that are absent from the parameterization process. For example, the AMBER force fields are general purpose force fields that were designed for modeling biomolecules such as proteins and nucleic acids,[4] whose parameterizations were performed on smaller training molecules such as amino acid dipeptides and nucleotides in selected representative conformations.[5-7]

One of the most important components of force field developments is the treatment of electrostatic interactions. In the extensively used point-charge additive force fields, the electrostatic terms are modeled by the interactions between fixed atom-centered point partial charges that obey the Coulomb's law. One commonly used parameterization method for obtaining the atomic partial charges is to use least-squares fitting to reproduce the

quantum mechanically (QM) determined electrostatic potential (ESP) at a large number of grid points around the molecule.[8-12] However, these fixed-point charges suffer from two disadvantages of being lack of both accuracy and transferability. First, charges on atoms that are buried by the other atoms are often poorly determined and their values often have high degree of uncertainty while fitting to QM ESPs. Consequently, unphysically large charges may be assigned to these buried atoms. Second, the ESP derived atomic charges are often sensitive to molecular conformations, leading to a lack of transferability of the charges across different conformations of identical molecules, as well as among common functional groups in related molecules. The problems of the ESP fitting strategy have been addressed by the restrained electrostatic potential (RESP) method developed by Bayly et al., which restrains the atomic charges towards zero using a hyperbolic penalty function to avoid impractically large charges.[13-14] Additionally, the multiple-conformation fitting strategy further improved the transferability of ESP fitted charges.[15-16] Using the combination of the multiple-conformation fitting strategy and the RESP method, Cieplak et al. derived the charges for all ribonucleotides, deoxyribonucleotides, and amino acids using ESPs calculated at the HF/6-31G* level of theory, which were incorporated into the AMBER ff94 force field.[5-6] Since then, the charge set of the ff94 force field has become the foundation of various subsequent AMBER force fields, including the AMBER ff99 force field,[7] the AMBER SB (Stony Brook) family force fields for modeling proteins[17-19] and the AMBER OL (Olomouc) family force fields for modeling nucleic acids.[20-22] The changes made by these subsequent force fields are mainly in torsional parameters, while the charges remain mostly unchanged.

Despite the improved accuracy and transferability of the additive AMBER force fields with the charge parameters derived using the RESP method, the additive force fields suffer

from a major disadvantage of being unable to model the atomic polarization effects, i.e., the redistribution of the atomic electron density due to the electric field produced by nearby charged atoms.[23] Polarization effects are important in various biological processes such as protein-ligand bindings,[24-26] nucleic acid-ion interactions,[27-28] the dielectric environmental changes during protein folding,[29-30] and ion transport through transmembrane ion channels.[31-32] Therefore, a variety of methods have been proposed to properly incorporate polarization effects into polarizable force fields, including the induced dipole models,[33-38] the fluctuating charge models,[39-40] the Drude oscillator models[41-42], and the continuum dielectric models.[43-44]

The induced dipole model is one of the most studied polarizable models, which has been incorporated into various AMBER polarizable force fields, including ff02,[33] ff02rl,[34] and ff12pol.[35-38] In this model, the induced dipole $\boldsymbol{\mu}_i$ of atom i subject to the external electric field $\boldsymbol{E}_i$ that comes from all atoms other than i is

$$\boldsymbol{\mu}_i = \alpha_i \left[\boldsymbol{E}_i - \sum_{j \neq i}^{\boldsymbol{n}} \boldsymbol{T}_{ij} \boldsymbol{\mu}_j \right] \tag{6.1}$$

where α_i is the isotropic polarizability of atom i, and $\boldsymbol{T}_{ij}$ is the dipole field tensor with the matrix form

$$\boldsymbol{T}_{ij} = \frac{f_e}{{r_{ij}}^3} \boldsymbol{I} - \frac{3 f_t}{{r_{ij}}^5} \begin{bmatrix} x^2 & xy & xz \\ xy & y^2 & yz \\ xz & yz & z^2 \end{bmatrix} \tag{6.2}$$

where $\boldsymbol{I}$ is the identity matrix; x, y and z are the Cartesian components along the vector between atoms i and j at distance r_{ij}; f_e and f_t are distance-dependent damping functions

that modify $\boldsymbol{T}_{ij}$ to avoid the so-called "polarization catastrophe" problem, which is the phenomenon that induced dipole diverges due to the cooperative induction between induced dipoles at short distances.[23, 45] Various damping schemes have been proposed by Thole,[46] which have been incorporated into the AMBER ff12pol force field.[35-38] However, one disadvantage of Thole's schemes is that they only screen the interactions between induced dipoles, leading to an inconsistent treatment of the polarizations due to fixed charges and permanent multipoles. About a decade ago, a damping scheme that models atomic electric multipoles using Gaussian electron densities was proposed by Elking et al.,[47-49] which was later named as the polarizable Gaussian Multipole (pGM) model.[50-53] The pGM model overcomes the disadvantage of Thole's schemes by screening all short-range electrostatic interactions in a physically consistent manner, including the interactions of charge-charge, charge-dipole, charge-quadrupole, dipole-dipole, and so on. The formula of damping functions f_e and f_t for the pGM model are as follows

$$S_{ij} = \frac{\beta_i \beta_j r_{ij}}{\sqrt{2\left(\beta_i{}^2 + \beta_j{}^2\right)}}$$

$$f_e = \mathrm{erf}\left(S_{ij}\right) - \frac{2}{\sqrt{\pi}} S_{ij} \exp\left(-S_{ij}{}^2\right) \tag{6.3}$$

$$f_t = \mathrm{erf}\left(S_{ij}\right) - \frac{2}{\sqrt{\pi}} S_{ij} \exp\left(-S_{ij}{}^2\right)\left(1 + \frac{2}{3} S_{ij}{}^2\right)$$

where $\beta_i = s\left(\frac{2\alpha_i}{3\sqrt{2\pi}}\right)^{-\frac{1}{3}}$ is the pGM "radius" of the Gaussian density distribution of atom i; s is a constant screening factor; $\mathrm{erf}\left(S_{ij}\right)$ is the error function of S_{ij}.

In the current pGM model design, the atomic charges and atomic induced dipoles are always present, while the inclusion of the atomic permanent dipoles is optional, leading to two distinct pGM models. The pGM model without atomic permanent dipoles is named pGM-ind, indicating the atomic dipoles of this pGM model only have contributions from atomic induced dipoles; The pGM model with atomic permanent dipoles is named pGM-perm, indicating the atomic dipoles of this pGM model have contributions from both induced dipoles and permanent dipoles. Based on the observation that atomic permanent dipole moments mainly exist along the direction of covalent bonding interactions, a local frame for the permanent dipoles formed by covalent basis vectors (CBVs) that are unit vectors along the directions of covalent bonds has been proposed for the pGM-perm model, so that the atomic permanent dipoles of the pGM-perm model always exist along the directions of covalent bonds.[51] An alternative pGM-perm model is called pGM-perm-v, where "v" stands for "virtual". In the pGM-perm-v model, the CBVs exist not only along the directions of covalent bonds (1-2 connecting atoms), but also along the directions of virtual bonds (1-3 connecting atoms) such as between the two hydrogen atoms of a water molecule. Consequently, in the pGM-ind model, the electric field $\boldsymbol{E}_i$ at the position of atom i in **eq 6.1** is only produced by fixed point charges of all atoms other than i. While in the pGM-perm and pGM-perm-v models, the electric field $\boldsymbol{E}_i$ is produced by both point charges and permanent dipoles of all atoms other than i. The formula of the electric field $\boldsymbol{E}_i$ for the pGM-ind model is shown in **eq 6.4**, and that for the pGM-perm and pGM-perm-v models is shown in **eq 6.5**.

$$\boldsymbol{E}_i = \sum_{j \neq i}^{n} f_e \frac{q_j}{{r_{ij}}^3} \boldsymbol{r}_{ji} \tag{6.4}$$

$$\boldsymbol{E}_i = \sum_{j \neq i}^{n} \left(f_e \frac{q_j}{{r_{ij}}^3} \boldsymbol{r}_{ji} + \boldsymbol{T}_{ij} \boldsymbol{p}_j \right) \tag{6.5}$$

where q_j is the point charge of atom j, $\boldsymbol{p}_j$ is the permanent dipole of atom j in the global frame, $\boldsymbol{r}_{ji}$ is the unit vector pointing in the direction from atom j to atom i.

In a series of recent works, the pGM models have been further developed and made available to the molecular modeling community. First, using an optimization method based on the genetic algorithm, we obtained a set of isotropic atomic polarizabilities and radii for the pGM models by fitting to molecular polarizability tensors of 1405 molecules or dimers calculated at the B3LYP/aug-cc-pVTZ level of theory.[50] Second, the closed-form analytical formula of the electrostatic energy and forces of the pGM models have been derived, and has been interfaced with the particle mesh Ewald (PME) method for molecular simulations under the periodic boundary conditions.[51] Third, the pGM internal stress tensor expression for constant pressure MD simulations of both flexible and rigid body molecular systems has been derived.[52] Finally, following the idea of charge parameterization by reproducing QM ESPs of the RESP method, we implemented the *PyRESP* program enabling the electrostatic parameterizations of the point-charge additive model and various induced dipole polarizable models, including the pGM-ind, pGM-perm, and pGM-perm-v models.[54]

The accuracy of the pGM models has been demonstrated by various previous works. It has been shown that even without atomic permanent dipoles, the pGM-ind model can notably improve the prediction of molecular polarizability anisotropy compared with the AMBER ff12pol force field that is based on the Thole's damping schemes.[50] Moreover, the electrostatic parameterizations on various molecules with various electrostatic models

using the *PyRESP* program show that the pGM models consistently produce ESPs and molecular electric moments with better agreement with QM calculated results than the additive point charge model.[54] A recent work assessed the accuracy of the pGM models in reproducing QM interaction energies, many-body interaction energies, as well as the non-additive and additive contributions to the many-body interactions for peptide mainchain hydrogen-bonding conformers, which shows the pGM models outperform all other tested widely used polarizable and additive force fields.[53]

However, there has been no work assessing the transferability of the pGM models. That is, whether the pGM models can accurately reproduce the electrostatic properties of larger molecular systems or different molecular conformations other than the molecules or conformations used for parametrizations? This is the first focus of this work. Another focus of this work is to find the optimal parameterization strategy for developing the next generation polarizable force fields based on the pGM models. Specifically, we aim to identify how many and what conformations should be applied for parameterizing amino acids for the pGM-ind and pGM-perm models that can give optimal accuracy and transferability for modeling long amino acid peptides or proteins. The performances of the pGM models were compared with that of the additive point charge model, which we call the "additive model" for short. The electrostatic parameterizations of the additive, pGM-ind, and pGM-perm models were performed by fitting to the same QM ESPs of each data set. One caveat of the pGM-perm and the pGM-perm-v models is that their parameterizations suffer from the so-called "singularity problem", which originates from the use of the permanent dipole local frame formed by CBVs. Fortunately, the restrained fitting strategy and the multiple-conformation fitting strategy implemented in the *PyRESP* program can theoretically address

the singularity problem, both of which have been demonstrated to successfully improve the accuracy and transferability of the electrostatic parameters of the additive model. The details of the singularity problem of the pGM-perm and pGM-perm-v models as well as the discussion of how restrained fitting and multiple-conformation fitting can address this problem can be found in the **Appendix B**. Therefore, extra attention will be paid to the performance of the pGM-perm an pGM-perm-v models with different parametrization strategies in this chapter.

6.2. Computational Details

6.2.1. Data Sets and Geometry Preparations

A total of nine data sets were generated and used for testing the transferability of the pGM models in this chapter, including WAT4, WAT6, WAT8, WAT10, ALA-di, ALA-tet, ALA-poly, GLY-poly, and BASE. The WAT4, WAT6, WAT8, WAT10 data sets are comprised of 100 water tetramer clusters, 72 water hexamer clusters, 13 water octamer clusters and 10 water decamer clusters, respectively. The initial geometries of the water clusters were extracted from 1 ns of MD simulations of a periodic box of 322 TIP3P waters.[55] 100 snapshots were saved at 10 ps intervals, and all clusters were extracted from these 100 TIP3P water boxes by randomly selecting a water molecule together with those closest water molecules. The MD simulation was conducted using the *sander* program from the AmberTools22 program suite.[56] Next, the WAT4 data set were optimized at the MP2/6-311++G(d, p) level of theory, and the WAT6, WAT8 and WAT10 data sets were optimized at the B3LYP/6-311++G(d, p) level of theory.

The ALA-di data set is comprised of 14 alanine dipeptides (ACE-ALA-NME) capped with an *N*-acetyl (ACE) group at the N-terminal, and an *N*-methylamide (NME) group at the C-terminal. The ACE and NME caps are used to mimic the chemical environment within peptides. Each alanine dipeptide was optimized at the MP2/6-311++G(d, p) level of theory with the mainchain torsional angles ϕ and ψ fixed according to **Table 6.1**. The ALA-tet data set is comprised of a total of 15 alanine tetrapeptides (ACE-ALA_3-NME), including (1) those in the conf1-conf10 conformations optimized at the HF/6-31G** level of theory by Beachy et al.,[57] which were further optimized at the MP2/6-311++G(d, p) level of theory without any constraints, and (2) those in aβ, αL, αR, β and pII conformations optimized at the MP2/6-311++G(d, p) level of theory with all mainchain torsional angles ϕ and ψ constrained. The mainchain torsional angles ϕ and ψ of each of the optimized conf1-conf10 conformations and the torsional angle constraints of the aβ, αL, αR, β and pII conformations are given in **Table 6.2**.

The ALA-poly and GLY-poly data sets are comprised of 60 alanine polypeptides (ACE-ALA_n-NME) and 60 glycine polypeptides (ACE-GLY_n-NME), respectively, where n is the number of repetitive ALA or GLY residues, ranging from 1 to 20. Each ACE-ALA_n-NME and ACE-GLY_n-NME have 3 conformations: aβ, αR and β. To prepare the ALA-poly and GLY-poly data sets, three alanine dipeptides (ACE-ALA-NME) and three glycine dipeptides (ACE-GLY-NME) were optimized at the ωB97X-D/6-311++G(d, p) level of theory with the mainchain torsional angles fixed at (ϕ, ψ) = (-140°, 135°), (-57°, -47°) and (-119°, 113°), corresponding to the aβ, αR and β conformations, respectively. Next, all ACE-ALA_n-NME and ACE-GLY_n-NME with n greater than or equal to 2 were generated from optimized alanine and glycine dipeptides by rigid body translation and rotation with the same ϕ and ψ torsional angles.

The BASE data set is comprised of 4 individual DNA nucleobases, including adenine (A), thymine (T), guanine (G), and cytosine (C), each capped with a methyl group to mimic the chemical environment within nucleosides, 2 Watson-Crick (WC) base pairs (A-T and C-G), and 8 stacked WC base pair tetramers (A-T/A-T, A-T/T-A, A-T/C-G, A-T/G-C, G-C/A-T, G-C/T-A, G-C/C-G, and G-C/G-C). The WC base pair tetramers are named as follows: the A-T/C-G tetramer means an A-T base pair stacked onto a C-G base pair, where A and T is stacked with C and G, respectively. To prepare the BASE data set, the two WC base pair dimers were first optimized at the MP2/6-311++G(d, p) level of theory. The individual nucleobases were extracted from the WC base pair dimers without further optimization. The tetramers were constructed from the WC base pairs by rigid body alignment of the base pair dimers to the B-DNA geometry created using the *nucgen* program,[58] without further optimization.

All QM geometry optimizations were performed using the Gaussian 16 software.[59]

Table 6.1. The Mainchain Torsional Angle Constraints for Geometry Optimizations of the Alanine Dipeptides from the ALA-di Data Set and their QM Molecular Dipole Moments

Conformation	ϕ/°	ψ/°	μ/Debye[a]
C5	-140	120	1.8190
$C7_{eq}$	-80	80	2.5090
$C7_{ax}$	60	-70	3.1220
a_1	-60	-40	5.9446

a_2	-52	-53	5.9848
al	70	30	5.5989
ap	-160	-40	5.1311
β_1	-161.9	166.4	3.0836
β_2	-130	20	4.5831
aβ	-140	135	2.2315
αL	57	47	5.7158
αR	-57	-47	5.9860
β	-119	113	0.8758
pII	-79	150	2.0894

[a] The QM molecular dipole moments are calculated at the MP2/aug-cc-pVTZ level of theory.

Table 6.2. The Mainchain Torsional Angles of the Optimized Alanine Tetrapeptides in Conf1-Conf10 Conformations and aβ, αL, αR, β and pII Conformations from the ALA-tet Data Set and their QM Molecular Dipole Moments

Conformation	ϕ_1/°	ψ_1/°	ϕ_2/°	ψ_2/°	ϕ_3/°	ψ_3/°	μ/Debye[a]
Conf1	-158.4	157.1	-158.1	156.5	-157.5	154.0	5.8104

Conf2	-158.3	155.7	-158.9	152.6	-80.1	84.7	1.7551
Conf3	-76.9	95.1	73.8	-59.0	-75.4	85.1	2.8191
Conf4	-159.1	156.0	-79.9	87.9	-160.7	143.3	3.4964
Conf5	-157.4	164.0	-59.9	-35.8	-76.7	90.1	3.0076
Conf6	-85.5	64.8	51.8	28.1	-179.0	139.2	6.0550
Conf7	52.2	-160.6	-88.0	71.2	-166.6	-53.3	10.6094
Conf8	69.3	-74.8	-52.8	134.4	54.3	33.9	3.7092
Conf9	74.3	-54.9	74.5	-53.4	74.4	-50.5	10.1255
Conf10	66.8	20.1	45.6	42.3	68.6	-74.5	9.6006
aβ	-140.0	135.0	-140.0	135.0	-140.0	135.0	4.2870
αL	57.0	47.0	57.0	47.0	57.0	47.0	14.7002
αR	-57.0	-47.0	-57.0	-47.0	-57.0	-47.0	15.2678
β	-119.0	113.0	-119.0	113.0	-119.0	113.0	0.8339
pII	-79.0	150.0	-79.0	150.0	-79.0	150.0	5.0536

[a] The QM molecular dipole moments are calculated at the MP2/aug-cc-pVTZ level of theory.

6.2.2. Electrostatic Parameterizations

The electrostatic parameterizations of the additive, pGM-ind, pGM-perm and pGM-perm-v models require the QM ESPs of a set of points in the solvent-accessible region around molecules as input. The QM ESPs of the molecules from the data sets WAT4, WAT6, WAT8, WAT10, ALA-di, and ALA-tet were calculated at the MP2/aug-cc-pVTZ level of theory, and that of the data sets ALA-poly, GLY-poly, and BASE were calculated at the ωB97X-D/aug-cc-pVTZ level of theory. The points were generated using the strategy developed by Singh et al. on molecular surfaces (with a density of 6 points/Å^2) at each of 1.4, 1.6, 1.8 and 2.0 times the van der Waals radii.[60-61] The QM molecular dipole moments of alanine dipeptides and alanine tetrapeptides from the ALA-di and ALA-tet data sets are shown in **Table 6.1-6.2**, and the QM molecular dipole moments of alanine polypeptides, glycine polypeptides, and WC base pair tetramers from the ALA-poly, GLY-poly, and BASE data sets are shown in **Table S6.1-S6.3**. All QM ESPs and molecular dipole moments were calculated using the Gaussian 16 software.[59]

The recently developed *PyRESP* program was used to parameterize the atomic charges (and permanent dipoles) of the molecules from each data set for each electrostatic model.[54] For polarizable models pGM-ind, pGM-perm and pGM-perm-v, the isotropic atomic polarizabilities derived in our previous work were used to calculate the induced dipoles.[50] A two-stage parameterization procedure was adopted.[54] In the first stage, all charges (and permanent dipoles) were set free to change, and a weak restraining strength 0.0005 was applied. In the second stage, intra-molecular equivalencing was enforced on all charges (and permanent dipoles) that share identical chemical environment with others, such as those of methyl and methylene hydrogens. A stronger restraining strength 0.001 was applied, and all other fitting centers were set frozen to keep the values obtained from the first stage. In both

stages, the restraints were only applied to non-hydrogen heavy atoms. The parameters of individual water molecule for the WAT4, WAT6, WAT8 and WAT10 data sets have been derived in our previous work.[54] The parameters for the ALA-di, ALA-tet, ALA-poly and GLY-poly data sets were obtained by constraining the total molecular charge to be zero, and the intra-molecular charge of the ACE and NME groups sum to zero in order to ensure zero net charge of the central amino acid fragments (-NH-CHR-CO-). For the parameterizations of amino acid tetrapeptides, intra-molecular equivalencing was enforced in both the first and the second stages to ensure identical parameters across the three repetitive central amino acid fragments. For multiple-conformational fittings, inter-molecular equivalencing was enforced in both stages to ensure identical atomic charges and permanent dipoles of the same molecule in different conformations. The parameters for the BASE data set were derived using single-conformation fittings with zero total molecular charge constraint only, and no additional intra-molecular charge constraint. For the parameterizations of the pGM-ind, pGM-perm and pGM-perm-v models, both 1-2 and 1-3 polarization interactions were included for reasons elucidated before.[50, 62]

6.2.3. Transferability Tests

The transferability of the electrostatic parameters of all electrostatic models were measured by the relative-root-mean-square errors of the overall molecular dipoles ($RRMS_{\mu}$) of each data set and the relative-root-mean-square errors of ESPs ($RRMS_{V}$) of each molecule (or molecule oligomer), given by

$$RRMS_{\mu} = \sqrt{\frac{\sum_{i=1}^{m}\left(\mu_i^{QM} - \mu_i\right)^2}{\sum_{i=1}^{m}\left(\mu_i^{QM}\right)^2}} \tag{6.6}$$

$$RRMS_{V} = \sqrt{\frac{\sum_{j=1}^{n_i}\left(V_{ij}^{QM} - V_{ij}\right)^2}{\sum_{j=1}^{n_i}\left(V_{ij}^{QM}\right)^2}} \tag{6.7}$$

and the average-relative-root-mean-square errors of ESPs ($ARRMS_V$) of each data set is

$$ARRMS_V = \frac{\sum_{i=1}^{m} RRMS_V}{m} \tag{6.8}$$

where m is the number of molecules for each data set; n_i is the number of ESP points surrounding molecule (or molecule oligomer) i; μ_i^{QM} and μ_i are the overall molecular dipoles of molecule/oligomer i given by QM calculations and molecular mechanics (MM) calculations, respectively; V_{ij}^{QM} and V_{ij} are the ESP values at point j of molecule/oligomer i given by QM calculations and MM calculations, respectively.

To calculate the total molecular dipole and ESP values of molecule A with the electrostatic parameters transferred from the parameterization results of molecule B, the input file (-i) and qin file (-q) of molecule A are created manually using the parameters from molecule B, which are provided as the inputs for the *PyRESP* program. The control parameter *irstrnt* of the *PyRESP* program is set to 2 so that no parameterization on molecule A is carried out, and the total molecular dipole and ESP values of molecule A with the transferred parameters from molecule B are printed in the output file (-o) of the *PyRESP* program.[54]

All scatterplots, boxplots and line plots are plotted using the Python package *Matplotlib*. The QM ESPs surrounding water tetramer clusters and the differences between QM and MM calculated ESPs are visualized using the UCSF Chimera software.[63]

6.3. Results and Discussion

6.3.1. The pGM-perm and pGM-perm-v Models Show the Best Transferability from Water Monomer to Water Oligomer Clusters

The transferability of the additive, pGM-ind, pGM-perm and pGM-perm-v models from water monomer to water oligomer clusters is tested by investigating the quality of the overall water cluster dipoles and ESPs calculated by MM calculations in comparison to those calculated at the MP2/aug-cc-pVTZ QM level of theory, measured by $RRMS_{\mu}$ and $ARRMS_{V}$, respectively. The parameters of water monomer for each electrostatic model have been derived in the original *PyRESP* work.[54] As discussed in the **Appendix B**, the water molecule is nonsingular, so that single-conformation fitting is sufficient for the parameterization of the pGM-perm and pGM-perm-v models for the water molecule. The single set of water monomer parameters are used in testing all WAT4, WAT6, WAT8 and WAT10 data sets. **Figure 6.1A** shows the scatterplots of MM dipoles calculated by each electrostatic model for the 100 water tetramer clusters from the WAT4 data sets versus those calculated by QM methods. It can be observed that all three pGM models outperform the additive model, as the $RRMS_{\mu}$ of the pGM-ind (0.0711), pGM-perm (0.0817), and pGM-perm-v (0.0823) models are only 34%, 39%, and 39% of that of the additive model (0.2110). **Figure 6.1B** shows the boxplots of the $RRMS_{V}$ of each electrostatic model for the WAT4 data sets, and we can see

that the $ARRMS_V$ of both the pGM-perm (0.0788) and pGM-perm-v (0.0790) models are 34% of that of the additive (0.2319) model and are 53% of that of the pGM-ind (0.1481) model. Interestingly, adding the virtual dipoles along the H-H direction in the pGM-perm-v model does not improve the quality of calculated overall dipoles and ESPs, as both the $RRMS_\mu$ and $ARRMS_V$ of the pGM-perm-v model are slightly higher than those of the pGM-perm model. To further explore the transferability difference among different models, the scatterplots of MM versus QM ESPs for the water tetramer clusters with the highest QM overall dipole (**Figure 6.1C**, dipole = 4.2850 Debye) and with the lowest QM overall dipole (**Figure 6.1D**, dipole = 0.0008 Debye) are shown. The pGM-perm and pGM-perm-v models produce the lowest $RRMS_V$ for both water clusters. For the water cluster with the highest QM dipole, the pGM-perm and pGM-perm-v models produce $RRMS_V$ of 0.0745 and 0.0743, respectively, both of which are 31% of that of the additive model (0.2393) and 56% of that of the pGM-ind model (0.1324). For the water cluster with the lowest QM dipole, the pGM-perm and pGM-perm-v models produce $RRMS_V$ of 0.0785 and 0.0788, respectively, both of which are 37% of that of the additive model (0.2138) and 52% of that of the pGM-ind model (0.1526). Once again, the $RRMS_V$ of the pGM-perm and pGM-perm-v models are very similar.

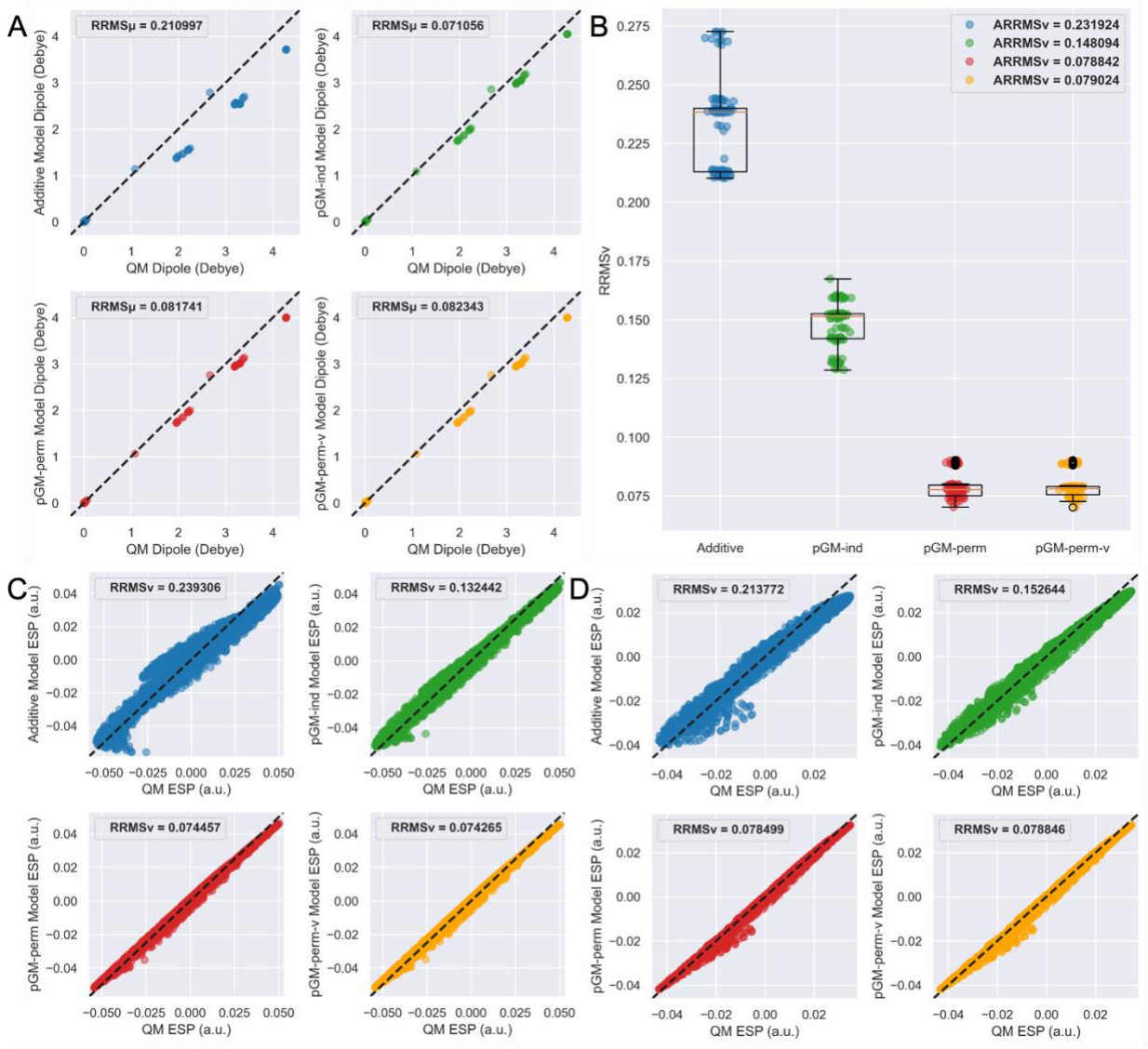


Figure 6.1. The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water tetramer clusters. **A.** Scatterplots of MM dipoles of each electrostatic model versus QM dipoles. Each plot shows a total of 100 data points, with each point representing a water tetramer. **B.** Boxplots of the $RRMS_V$ of each electrostatic model with QM results. Each plot shows a total of 100 data points, with each point representing a water tetramer. **C.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water tetramer with the highest QM dipole (dipole = 4.2850 Debye). Each

plot shows a total of 4660 data points, with each point representing an ESP point. **D.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water tetramer with the lowest QM dipole (dipole = 0.0008 Debye). Each plot shows a total of 4339 data points, with each point representing an ESP point. For **A**, **C**, and **D**, the dashed lines correspond to perfect matching.

Figure 6.2 illustrates the QM ESPs surrounding the water tetramer clusters with the highest and lowest QM overall dipoles, as well as the differences between QM ESPs and MM ESPs calculated by each electrostatic model. It can be observed that the additive model is unable to accurately reproduce the ESP of polar regions, i.e., regions with high ESP absolute values. Specifically, the additive model tends to generate ESPs with lower values than QM results where the QM ESPs have large positive values but generate ESPs with higher values than QM results where the QM ESPs have large negative values. The pGM-ind model improves the ESP fitting significantly. It is noteworthy that both the pGM-ind and additive models have identical number of electrostatic parameters. Therefore, the significant improvement observed in the pGM-ind model over the additive model is strong evidence to the critical roles that intra-molecular polarization plays in transferability. The pGM-perm and pGM-perm-v models give ESPs nearly identical to QM results in both polar and nonpolar regions. Note that the ESP differences with QM results given by the pGM-perm and pGM-perm-v models are almost indistinguishable. Therefore, we conclude that the additional dipoles along the H-H virtual bond in the pGM-perm-v model do not improve the ESP fitting

quality and transferability compared with the pGM-perm model for the water tetramer clusters.

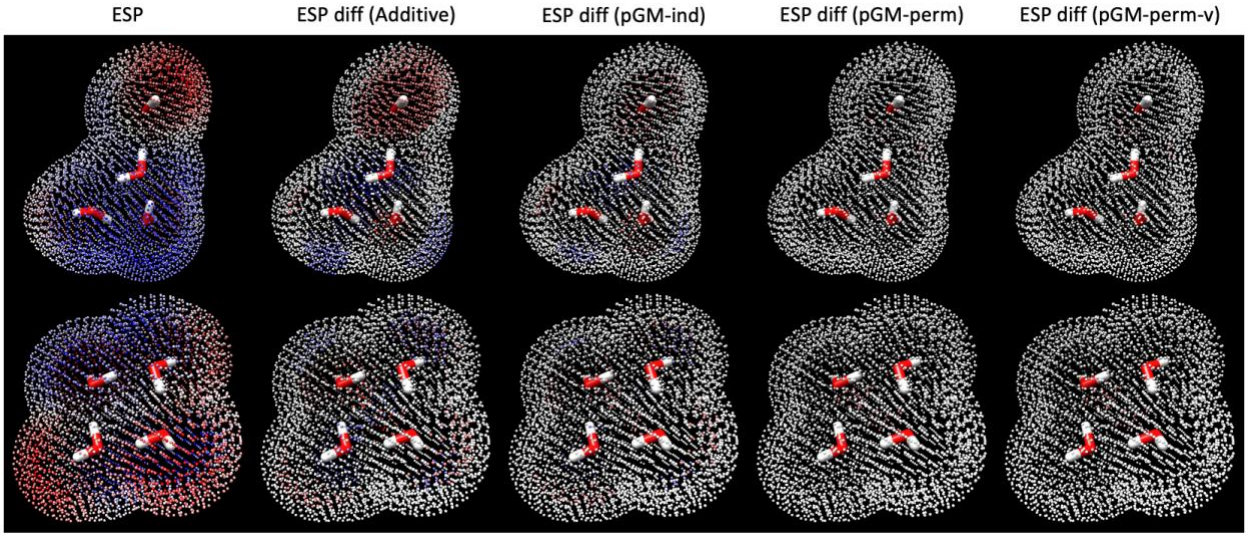


Figure 6.2. Visualization of QM ESPs surrounding water tetramer clusters and the differences between QM and MM calculated ESPs of the additive, pGM-ind, pGM-perm, and pGM-perm-v models. The upper panel shows the water tetramer with the highest QM dipole (4.2850 Debye) and the lower panel shows the water tetramer with the lowest QM dipole (0.0008 Debye). The leftmost column shows the QM ESPs, with red color indicating positive ESP value and blue color indicating negative ESP value. All other columns show the differences between QM ESPs and MM ESPs, with red color indicating QM ESP is greater than MM ESP and blue color indicating QM ESP is less than MM ESP.

After analyzing the transferability from water monomer to water tetramer clusters, we examined the transferability of each electrostatic model from water monomer to water

oligomers with larger sizes, including hexamer, octamer, and decamer clusters from the WAT6, WAT8, and WAT10 data sets, respectively. The scatterplots of MM dipoles of each electrostatic model versus QM dipoles, the boxplots of the $RRMS_V$ of each electrostatic model with QM results, and the scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water hexamer, octamer, and decamer clusters with the highest and lowest QM dipoles are shown in **Figure S6.1-S6.3**. The $RRMS_\mu$ and $ARRMS_V$ of each water oligomer cluster size produced by each electrostatic model are summarized in **Figure 6.3**. The pGM-ind, pGM-perm and pGM-perm-v models consistently outperform the additive models in terms of both $RRMS_\mu$ and $ARRMS_V$, regardless of the water oligomer cluster sizes. Although the pGM-ind model performs slightly better than the pGM-perm and pGM-perm-v models in terms of $RRMS_\mu$, the latter two models significantly outperform the pGM-ind model in terms of $ARRMS_V$. Another observation is that the $RRMS_\mu$ and $ARRMS_V$ of each water oligomer cluster data set produced by the pGM-perm and pGM-perm-v models are essentially indistinguishable, as their plots overlap each other, consistent with the earlier observations in the case of water tetramers. In fact, as discussed in the original *PyRESP* work,[54] the virtual dipoles in the pGM-perm-v model may lead to overfitting problem and is expected to increase the computational time in simulations. Furthermore, the virtual dipole may cause additional singularity problems during parameterization, as discussed in **Appendix B**. For these reasons, the transferability test of the pGM-perm-v model will only be performed for the water oligomer clusters for illustration purpose. For other data sets, we will only test the transferability of the additive, pGM-ind and pGM-perm models.

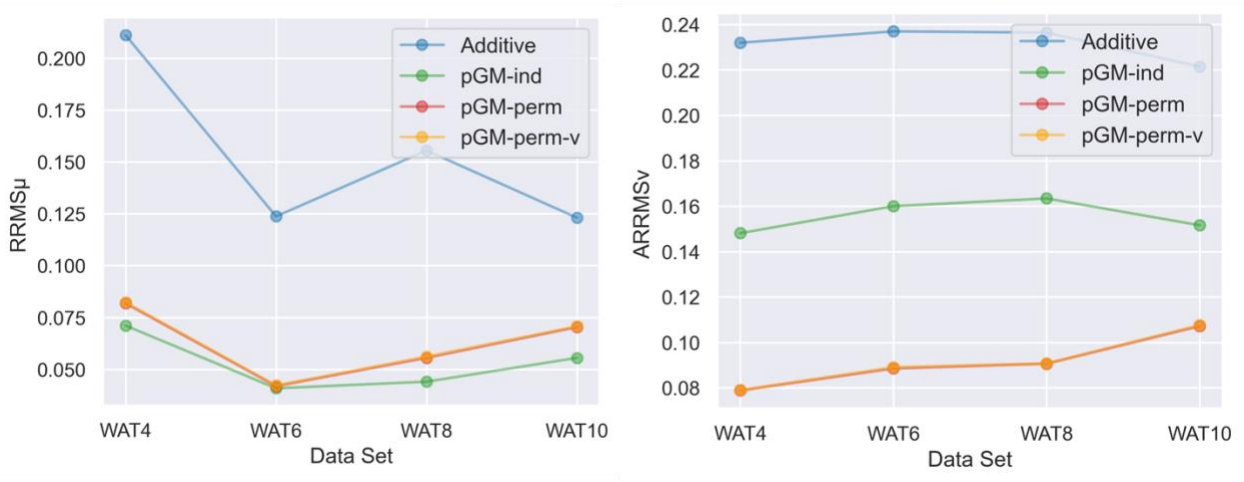


Figure 6.3. The $RRMS_{\mu}$ and $ARRMS_{V}$ of the WAT4, WAT6, WAT8, and WAT10 data sets of the additive, pGM-ind, pGM-perm and pGM-perm-v models parameterized with water monomer. Note that the plots of the pGM-perm and pGM-perm-v models overlap each other.

6.3.2. Electrostatic Parameterization of the pGM Models with Amino Acid Dipeptides Leads to Lack of Transferability to Tetrapeptides

In the previous subsection, we have shown that the pGM-perm and pGM-ind models outperform the additive model in terms of the transferability from water monomer (training molecule) to water oligomer clusters (testing molecules). Next, we move on to compare the transferability of the additive, pGM-ind, and pGM-perm models across different conformations of amino acids, as well as from short amino acid dipeptides (training molecules) to longer amino acid tetrapeptides (testing molecules). The reason why we are interested in amino acids is that they are the building blocks of proteins, so that the electrostatic parameterizations of amino acids are of critical importance for the development of force fields for modeling biomolecules. Therefore, we aim to explore the best

parametrization strategy of amino acids for developing the next generation polarizable Amber force field based on the pGM models. As discussed in the **Appendix B**, every amino acid molecule is a singular molecule in the context of the parameterization of the pGM-perm model, due to the existence of the sp^3 alpha carbon in every amino acid backbone. Therefore, the combination of restrained fitting and multiple-conformation fitting implemented in the *PyRESP* program will be explored for the electrostatic parameterizations of each model, which are expected to improve the transferability of each model and to mitigate the singularity problem of the pGM-perm model. Alanine was selected as the model amino acid for testing the transferability of each electrostatic model. In this test, five alanine dipeptides (ACE-ALA-NME) in αR (QM dipole = 5.9860 Debye), β (0.8758 Debye), $C7_{eq}$ (2.5090 Debye), aβ (2.2315 Debye), and C5 (1.8190 Debye) conformations from the ALA-di data set were used for electrostatic parameterization because of their wide range of molecular dipole moments as shown in **Table 6.1**. A total of nine parameterization combinations of the five conformations were tested, including three single-conformation fittings (αR, β, $C7_{eq}$), three double-conformation fittings (αR/β, αR/$C7_{eq}$, β/$C7_{eq}$), one triple-conformation fitting (αR/β/$C7_{eq}$), one 4-conformation fitting (αR/β/$C7_{eq}$/aβ), and one 5-conformation fitting (αR/β/$C7_{eq}$/aβ/C5).

We first tested the transferability of the additive, pGM-ind, and pGM-perm models across different conformations of alanine dipeptides within the ALA-di data set, which contains a total of 14 conformations. Among all the three single-conformation fittings, the $C7_{eq}$ conformation gives the best overall performance for the transferability of the pGM-ind and pGM-perm models, with $RRMS_{\mu}$ of 0.0386 and 0.0641, and $ARRMS_{V}$ of 0.1074 and

0.1237, respectively. (Data not shown) Among all the three double-conformation fittings, the combination of the αR and β conformations gives the best overall performance for the pGM-ind and pGM-perm models, with $RRMS_\mu$ of 0.0244 and 0.0239, and $ARRMS_V$ of 0.1004 and 0.0858, respectively. (Data not shown) **Figure 6.4A-B** summarizes the $RRMS_\mu$ and $ARRMS_V$ of the ALA-di data set of each electrostatic model parameterized with alanine dipeptides in 1-5 conformations, where the single-conformation fitting and double-conformation fitting are $C7_{eq}$ and αR/β, respectively. One observation is that for all the additive, pGM-ind, and pGM-perm models, both $RRMS_\mu$ and $ARRMS_V$ reached convergence with double-conformation fittings, and multiple-conformation fittings with more than two conformations do not significantly improve the transferability across different conformations of alanine dipeptides in the ALA-di data set. Another observation is that the pGM-perm model performs the best among the three models in terms of both $RRMS_\mu$ and $ARRMS_V$. Taking double-conformation fitting as an example, the $RRMS_\mu$ and $ARRMS_V$ of the pGM-perm model are 0.0239 and 0.0858, respectively, which are 98% and 85% of those of the pGM-ind model (0.0244 and 0.1004), and 39% and 54% of those of the additive model (0.0607 and 0.1601). One exception is the case of single-conformation fitting, where the pGM-perm model shows worse transferability than the pGM-ind model, as the $RRMS_\mu$ and $ARRMS_V$ of the pGM-ind model (0.0386 and 0.1074) are 60% and 87% of those of the pGM-perm model (0.0641 and 0.1237). The worse performance of the pGM-perm model with single-conformation fitting might be explained by its singularity problem. Even so, the pGM-perm model still performs much better than the additive model, as the $RRMS_\mu$ and $ARRMS_V$ of the pGM-perm model are only 38% and 62% of those of the additive model (0.1687 and 0.1994).

Next, we tested the transferability of each electrostatic model from alanine dipeptides (ACE-ALA-NME) to longer alanine tetrapeptides (ACE-ALA_3-NME). Specifically, the electrostatic parameters derived with the nine combinations of alanine dipeptides in the previously used five conformations (αR, β, $C7_{eq}$, αR/β, αR/$C7_{eq}$, β/$C7_{eq}$, αR/β/$C7_{eq}$, αR/β/$C7_{eq}$/aβ, and αR/β/$C7_{eq}$/aβ/C5) from the ALA-di data set were used to calculate the $RRMS_{\mu}$ and $ARRMS_V$ of alanine tetrapeptides from the ALA-tet data set, which contains a total 15 conformations. Among all the three single-conformation fittings, the $C7_{eq}$ conformation again gives the best overall performance for the transferability of the pGM-ind and pGM-perm models, with $RRMS_{\mu}$ of 0.1306 and 0.1892, and $ARRMS_V$ of 0.1652 and 0.2066, respectively. (Data not shown) This is consistent with the transferability test across alanine dipeptides in different conformations. Among all the three double-conformation fittings, the combination of the αR and $C7_{eq}$ conformations gives the best overall performance for the pGM-ind and pGM-perm models, with $RRMS_{\mu}$ of 0.1268 and 0.1634, and $ARRMS_V$ of 0.1663 and 0.1836, respectively. (Data not shown) This is different from the transferability test across different alanine dipeptide conformations where the best performance is given by the combination of the αR and β conformations. **Figure 6.4C-D** summarizes the $RRMS_{\mu}$ and $ARRMS_V$ of the ALA-tet data set of each electrostatic model parameterized with alanine dipeptides in 1-5 conformations, where the single-conformation fitting and double-conformation fitting are $C7_{eq}$ and αR/$C7_{eq}$, respectively. It can be observed that the transferability of the additive, pGM-ind, and pGM-perm models from alanine dipeptides to alanine tetrapeptides show very different patterns compared with the transferability across alanine dipeptides in different conformations. First, the transferability of each model tends to get worse with more conformations used for parametrization, as

shown by the results of both $RRMS_\mu$ and $ARRMS_V$. Second, the pGM-ind model consistently gives the lowest $RRMS_\mu$ and $ARRMS_V$, which outperforms both the additive and pGM-perm models. However, with the $RRMS_\mu$ of each model consistently greater than 0.13, and the $ARRMS_V$ of each model consistently greater than 0.16, none of the three models give satisfactory transferability from alanine dipeptides to alanine tetrapeptides. Therefore, we conclude that parameterization using amino acid dipeptides is inadequate for developing polarizable force fields based on the pGM models, as those parameters have the risk of lacking transferability to polypeptides or proteins.

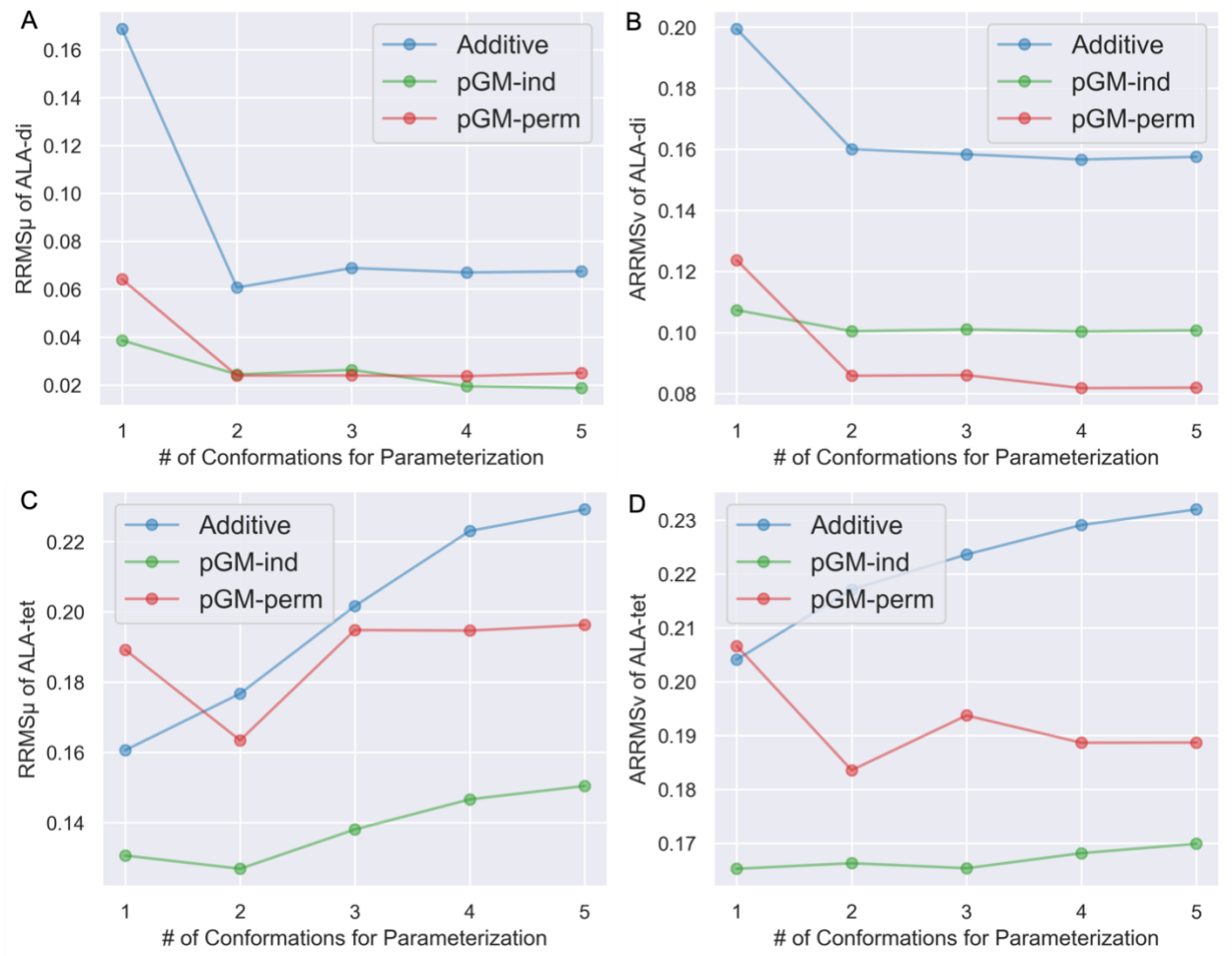


Figure 6.4. The $RRMS_{\mu}$ and $ARRMS_{V}$ of the ALA-di and ALA-tet data sets of the additive, pGM-ind, and pGM-perm models parameterized with alanine dipeptides from the ALA-di data set in 1-5 conformations. **A-B.** The $RRMS_{\mu}$ and $ARRMS_{V}$ of the ALA-di data set. The 1-5 conformations are $C7_{eq}$, αR/β, αR/β/$C7_{eq}$, αR/β/$C7_{eq}$/aβ, and αR/β/$C7_{eq}$/aβ/C5. **C-D.** The $RRMS_{\mu}$ and $ARRMS_{V}$ of the ALA-tet data set. The 1-5 conformations are $C7_{eq}$, αR/$C7_{eq}$, αR/β/$C7_{eq}$, αR/β/$C7_{eq}$/aβ, and αR/β/$C7_{eq}$/aβ/C5.

6.3.3. Electrostatic Parameters of the pGM Models Derived with Amino Acid Tetrapeptides Are Highly Transferable Across Different Conformations

The transferability of the electrostatic parameters derived from dipeptides is unsatisfactory for all three models, particularly from dipeptides to tetrapeptides. We hypothesize that part of the reason is that there are two terminal groups (ACE and NME) in each dipeptide, making the terminal/amino acid ratio to be 2, much higher than that in polypeptides in which this ratio can be orders of magnitude lower. Therefore, we attempted to perform parameterizations using tetrapeptides in which three repetitive amino acid residues are present, making it possible to mimic multiple chemical environments and multiple conformations. The alanine tetrapeptides (ACE-ALA_3-NME) in αR (QM dipole = 15.2678 Debye), β (0.8339 Debye), pII (5.0536 Debye), aβ (4.2870 Debye), and αL (14.7002 Debye) conformations from the ALA-tet data set were selected for parameterizations because of their wide range of molecular dipole moments as shown in **Table 2**. A total of nine parameterization combinations of the above five conformations were tested, including three single-conformation fittings (αR, β, pII), three double-conformation fittings (αR/β,

αR/pII, β/pII), one triple-conformation fitting (αR/β/pII), one 4-conformation fitting (αR/β/pII/aβ), and one 5-conformation fitting (αR/β/pII/aβ/αL).

We first tested the transferability of each electrostatic model across different conformations of alanine tetrapeptides within the ALA-tet data set, which contains a total of 15 conformations. Among all the three single-conformation fittings, the pII conformation gives the best overall performance for the transferability of the pGM-ind and pGM-perm models, with $RRMS_\mu$ of 0.0197 and 0.0978, and $ARRMS_V$ of 0.0893 and 0.1073, respectively. (Data not shown) Among all the three double-conformation fittings, the combination of the αR and β conformations gives the best overall performance for the pGM-ind and pGM-perm models, with $RRMS_\mu$ of 0.0182 and 0.0249, and $ARRMS_V$ of 0.0881 and 0.0744, respectively. (Data not shown) **Figure 6.5** summarizes the $RRMS_\mu$ and $ARRMS_V$ of each electrostatic model of the ALA-tet data set parameterized with alanine tetrapeptides using 1-5 conformations, where the single-conformation fitting and double-conformation fitting are pII and αR/β, respectively. Similar to the transferability test across different conformations of the ALA-di data set, both the $RRMS_\mu$ and $ARRMS_V$ of the additive and pGM-perm models reached convergence with double-conformation fittings, and multiple-conformation fitting with more than two conformations do not significantly improve the transferability across different conformations in the ALA-tet data set. Interestingly, the pGM-ind model consistently shows the lowest $RRMS_\mu$ (less than 0.02) among all three models, regardless of the number of alanine tetrapeptide conformations used for parameterizations. The best performance shown by the pGM-ind model is somewhat surprising, given that the pGM-ind model does not take atomic permanent dipoles into account, in contrast to the pGM-perm

model. In terms of $ARRMS_V$, the pGM-perm model shows the best performance with multiple-conformation fittings. In contrast, with single-conformation fitting, the pGM-ind model again outperforms the pGM-perm model, as the $ARRMS_V$ of the pGM-ind model (0.0893) is 83% of that of the pGM-perm model (0.1073). This is consistent with the transferability test across different conformations of the ALA-di data set, which can be explained again by the singularity problem of the pGM-perm model. Furthermore, the additive model consistently gives the worst transferability as measured by both $RRMS_\mu$ and $ARRMS_V$. For example, with double-conformation fittings, the $RRMS_\mu$ of the pGM-ind (0.0182) and pGM-perm (0.0249) models are only 21% and 28% of that of the additive model (0.0875), and the $ARRMS_V$ of the pGM-ind (0.0881) and pGM-perm (0.0744) models are 46% and 39% of that of the additive model (0.1909).

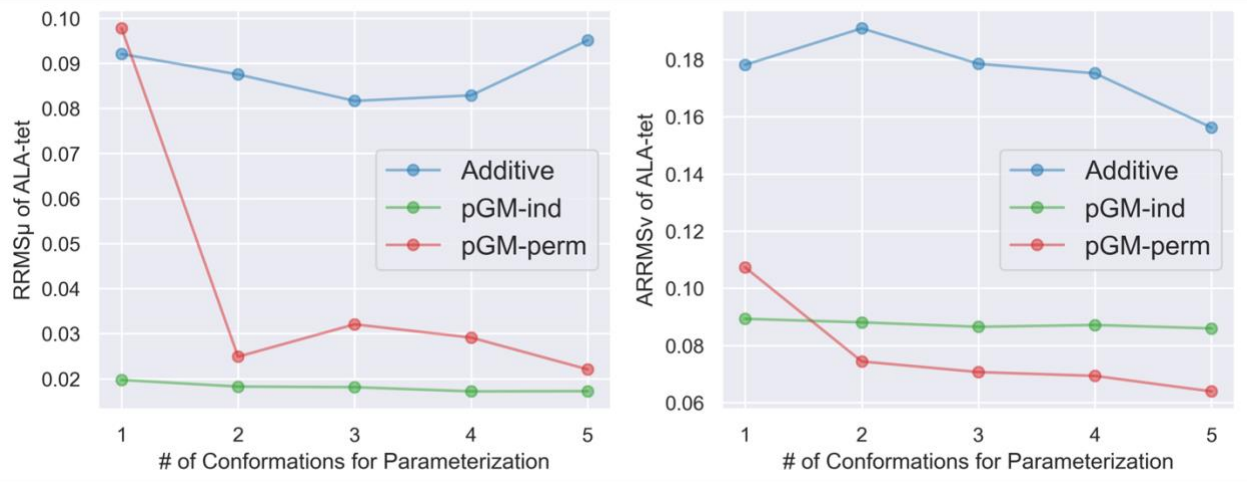


Figure 6.5. The $RRMS_\mu$ and $ARRMS_V$ of the ALA-tet data set of the additive, pGM-ind, and pGM-perm models parameterized with alanine tetrapeptides from the ALA-tet data set in 1-

5 conformations. The 1-5 conformations are pII, αR/β, αR/β/pII, αR/β/pII/aβ, and αR/β/pII/aβ/αL.

6.3.4. Electrostatic Parameterization of the pGM Models with Amino Acid Tetrapeptides Leads to Satisfactory Transferability to Longer Polypeptides

In addition to the transferability across different conformations, another question that needs to be addressed is the transferability across polypeptide chains with different lengths. This is a rather critical question because for practical purposes, all protein force fields are parameterized using short peptides or model compounds and are applied to proteins that can be hundreds-amino acid long. Therefore, we need to know how well the electrostatic parameters obtained from parameterizing tetrapeptides transfer to longer polypeptides. To answer this question, transferability tests were performed using the ALA-poly data set containing a total of 60 alanine polypeptides (ACE-ALA_n-NME), and the GLY-poly data set containing a total of 60 glycine polypeptides (ACE-GLY_n-NME), where n ranges between 1 and 20. Each ACE-ALA_n-NME and ACE-GLY_n-NME is represented by 3 conformations: aβ, αR and β. Due to the large molecule size of long polypeptides such as ACE-ALA_{20}-NME (212 atoms) and ACE-GLY_{20}-NME (152 atoms), the ωB97X-D DFT method was used for both geometry optimizations and ESP calculations for the two data sets to save computational resources. The electrostatic parameters (atomic charges and permanent dipoles) of alanine polypeptides and glycine polypeptides were both obtained by αR/β double-conformation fittings to the ESPs calculated at the ωB97X-D/aug-cc-pVTZ level of theory, using alanine tetrapeptides (ACE-ALA_3-NME) from the ALA-poly data set and glycine

tetrapeptides (ACE-GLY_3-NME) from the GLY-poly data set, respectively. The re-parameterization of alanine tetrapeptides is necessary to ensure that the parameters are consistent with other alanine polypeptides, since the ESPs of alanine tetrapeptides in the ALA-tet data set were calculated using a different QM method (MP2/aug-cc-pVTZ), which leads to slightly different ESPs. The $RRMS_\mu$ and $ARRMS_V$ of the ALA-poly data set and the GLY-poly data set of each electrostatic model are shown in **Figure 6.6A-B** and **Figure 6.6C-D**, respectively. Encouragingly, with αR/β double conformation fittings, both the pGM-ind and pGM-perm models show great transferability to alanine and glycine polypeptides with lengths range from 1 to 20, although the pGM-perm model performs slightly better than the pGM-ind model. Interestingly, both the pGM-ind and pGM-perm models exhibit higher $ARRMS_V$ at the shorter end compared to longer polypeptides. This is indicative that the underlying chemical environment in peptides of 1-2 amino acids are somewhat different from that of longer polypeptides. This explains why electrostatic parameterization with dipeptides leads to unsatisfactory transferability to longer polypeptides. The additive model consistently shows the worst transferability to alanine and glycine polypeptides among all the three electrostatic models. In general, the longer the polypeptides are, the higher $RRMS_\mu$ and $ARRMS_V$ the additive model produces. Therefore, we conclude that double-conformation fitting using amino acid tetrapeptides in the αR and β conformations is a sound strategy for amino acid electrostatic parametrizations for the pGM models. In the future development of the pGM force fields for proteins, this strategy is expected to be applied to the systematic electrostatic parameterizations for all amino acids.

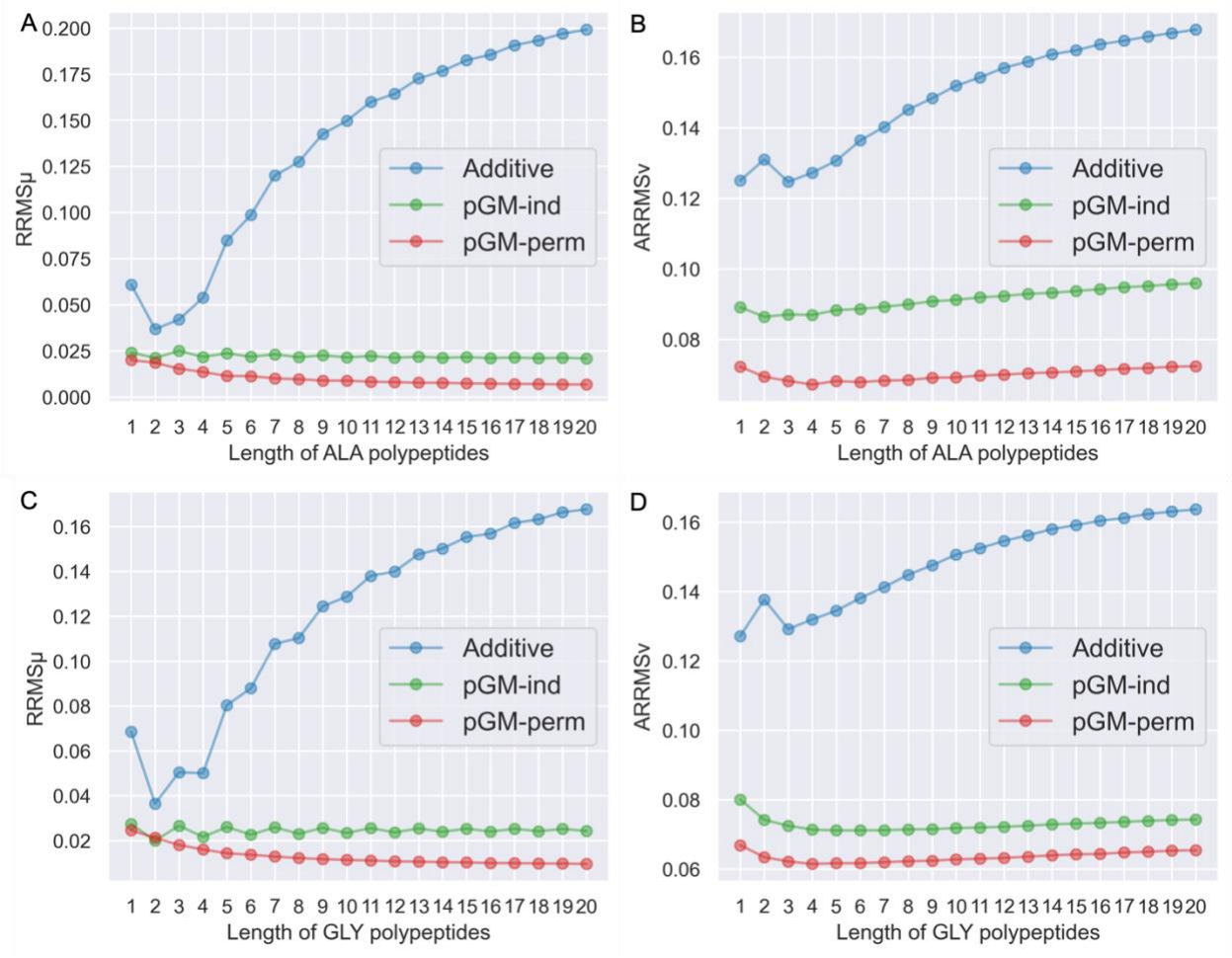


Figure 6.6. The $RRMS_{\mu}$ and $ARRMS_{V}$ of the ALA-poly and GLY-poly data sets of the additive, pGM-ind, and pGM-perm models parameterized with alanine or glycine tetrapeptides. **A-B.** The $RRMS_{\mu}$ and $ARRMS_{V}$ against the length of alanine polypeptides from the ALA-poly data set. Each model is parameterized with alanine tetrapeptides from the ALA-poly data set using αR/β double-conformation fitting. **C-D.** The $RRMS_{\mu}$ and $ARRMS_{V}$ against the length of glycine polypeptides from the GLY-poly data set. Each model is parameterized with glycine tetrapeptides from the GLY-poly data set using αR/β double-conformation fitting.

6.3.5. The pGM Models Outperforms the Additive Model in Transferability from Nucleobase Monomers to WC Base Pair Dimers and Tetramers

Besides amino acids, another key component of force field development for modeling biomolecules is the electrostatic parameterizations of nucleotides, the building blocks of nucleic acids including DNA and RNA. Nucleotides are composed of three subunits, including a nucleobase, a five-carbon sugar, and a phosphate group. The ability of nucleobases to form hydrogen-bonding WC base pairs and to stack upon each other through π-π interactions leads directly to the double-stranded helical structures of DNA molecules. Therefore, in this subsection, we aim to compare the transferability of the additive, pGM-ind, and pGM-perm models from the DNA nucleobase monomers, including adenine (A), thymine (T), guanine (G), and cytosine (C), to the WC base pair dimers and stacked WC base pair tetramers formed by the four DNA nucleobases. All monomers, dimers, and tetramers used in this chapter are from the BASE data set. Each nucleobase is capped with a methyl group to mimic the chemical environment within nucleosides. The two WC base pair dimers include the A-T base pair with two hydrogen bonds and the G-C base pair with three hydrogen bonds. The eight stacked WC base pair tetramers include A-T/A-T, A-T/T-A, A-T/C-G, A-T/G-C, G-C/A-T, G-C/T-A, G-C/C-G, and G-C/G-C. For instance, the A-T/C-G tetramer is formed by stacking the A-T base pair onto the C-G base pair, where A and T is stacked with C and G, respectively.

Since the nucleobases are rigid molecules in nature, each DNA nucleobase monomer was parameterized using single-conformation fitting to ESPs calculated at the ωB97X-D/aug-cc-pVTZ level of theory. **Table 6.3** shows the molecular dipole and quadrupole moments

calculated by each electrostatic model and QM methods as well as the $RRMS_V$ of the A-T and G-C WC base pair dimers. It can be seen that the pGM-ind and pGM-perm models produce molecular dipole moments and quadrupole moments with better agreement with the QM moments than the additive model. However, nucleobases are also singular molecules in terms of the parameterizations of the pGM-perm model due to the existence of sp^2 carbons in all nucleobases (see **Appendix B**), which can explain the observation that the pGM-ind model gives slightly better agreements with the QM calculated electric moments than the pGM-perm model. On the other hand, the $RRMS_V$ consistently decreases with the order of the additive, pGM-ind, and pGM-perm models for both WC base pairs. For the A-T base pairs, the $RRMS_V$ of the pGM-ind (0.1250) and pGM-perm (0.0904) models are 86% and 62% of that of the additive model (0.1454); For the G-C base pairs, the $RRMS_V$ of the pGM-ind (0.1183) and pGM-perm (0.0766) models are 71% and 46% of that of the additive model (0.1657). Therefore, the pGM models outperform the additive model significantly in terms of transferability to WC base pairs with single-conformation fitting with A, T, G, and C monomers. Note that the G-C base pair (QM dipole = 6.0874 Debye) has much higher overall dipole moment than the A-T base pair (1.9010 Debye). The observation that the additive model gives higher $RRMS_V$ for the G-C base pair than for the A-T base pair, while the pGM models give lower $RRMS_V$ for the G-C base pair than for the A-T base pair indicates that the pGM models can better model the polarization effects in the highly polar G-C base pairs.

Figure 6.7A-C show the scatterplot of MM dipoles of the eight WC base pair tetramers from the BASE data set calculated by each electrostatic model versus those calculated at the ωB97X-D/aug-cc-pVTZ level of theory. It can be observed that the $RRMS_\mu$ of the pGM-ind (0.0141) and pGM-perm (0.0209) models are much lower than that of the additive model

(0.1293), as the $RRMS_\mu$ of the pGM-ind and pGM-perm models are only 11% and 16% of that of the additive model. The slightly better performance of the pGM-ind model than the pGM-perm model is consistent with the better electric moments agreement with QM results given by the pGM-ind model for WC base pair dimers, which might be caused by the singularity problem of the pGM-perm model. **Figure 6.7D** shows the boxplots of the $RRMS_V$ of the WC base pair tetramers of each electrostatic model, and we can see that the $ARRMS_V$ decreases in the order of the additive (0.2000), pGM-ind (0.1063), and pGM-perm (0.0737) models, as the $ARRMS_V$ of the pGM-ind and pGM-perm models are 53% and 37% of that of the additive model. To further explore the transferability difference among different models, the scatterplots of MM ESPs versus the QM ESPs for the G-C/G-C tetramer with the highest QM overall dipole (dipole = 10.5748 Debye) and the A-T/T-A tetramer with the lowest QM overall dipole (dipole = 2.1904 Debye) are shown in **Figure 6.8**. Once again, for both WC base pair tetramers, the $RRMS_V$ of the pGM-perm model are the lowest, and that of the pGM-ind model are the second lowest. For the G-C/G-C tetramer, the $RRMS_V$ of the additive, pGM-ind, and pGM-perm models are 0.1804, 0.1016, and 0.0678, respectively. For the A-T/T-A tetramer, the $RRMS_V$ of the additive, pGM-ind, and pGM-perm models are 0.2301, 0.1092, and 0.0781, respectively.

Table 6.3. Molecular Dipole/Quadrupole Moments and $RRMS_V$ of the A-T and G-C WC Base Pair Dimers Fitted with A, T, G, and C Monomers with the Additive, pGM-ind, and pGM-perm Models

WC Base Pair		Additive	pGM-ind	pGM-perm	QM
Dipole Moments/Debye[a]					
A-T		2.3174	1.8483	1.9134	1.9010
G-C		4.6236	5.9753	5.9603	6.0874
Quadrupole Moments/Debye Angstroms[b]					
A-T	Q_{xx}	46.5515	41.0910	40.7533	43.5328
	Q_{yy}	-19.7216	-17.9977	-17.5097	-18.6448
	Q_{zz}	-26.8299	-23.0933	-23.2436	-24.8879
G-C	Q_{xx}	46.5542	43.6740	43.6416	46.3355
	Q_{yy}	-20.9126	-19.4755	-19.1479	-20.4689
	Q_{zz}	-25.6416	-24.1985	-24.4937	-25.8666
$RRMS_V$					
A-T		0.1454	0.1250	0.0904	
G-C		0.1657	0.1183	0.0766	

[a] Dipole moment relative to center of mass. [b] Quadrupole moments along the principal axes.

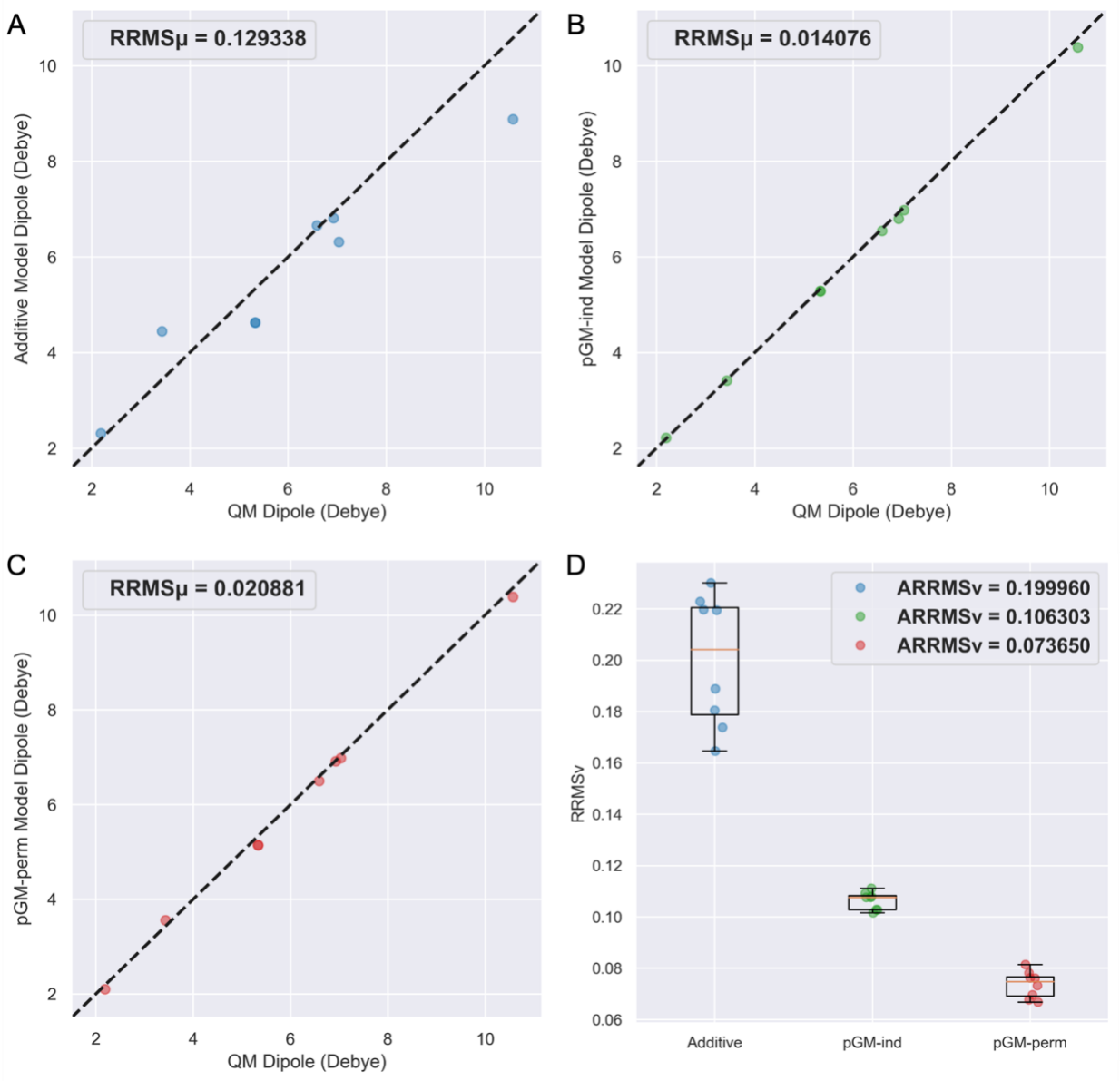


Figure 6.7. The transferability tests of the additive, pGM-ind, and pGM-perm models from A, T, G, and C monomers to WC base pair tetramers. **A-C.** Scatterplots of MM dipoles of each electrostatic model versus QM dipoles. **D.** Boxplots of $RRMS_V$ of each electrostatic model with QM results. Each scatterplot or boxplot shows a total of 8 data points, with each point representing a WC base pair tetramer.

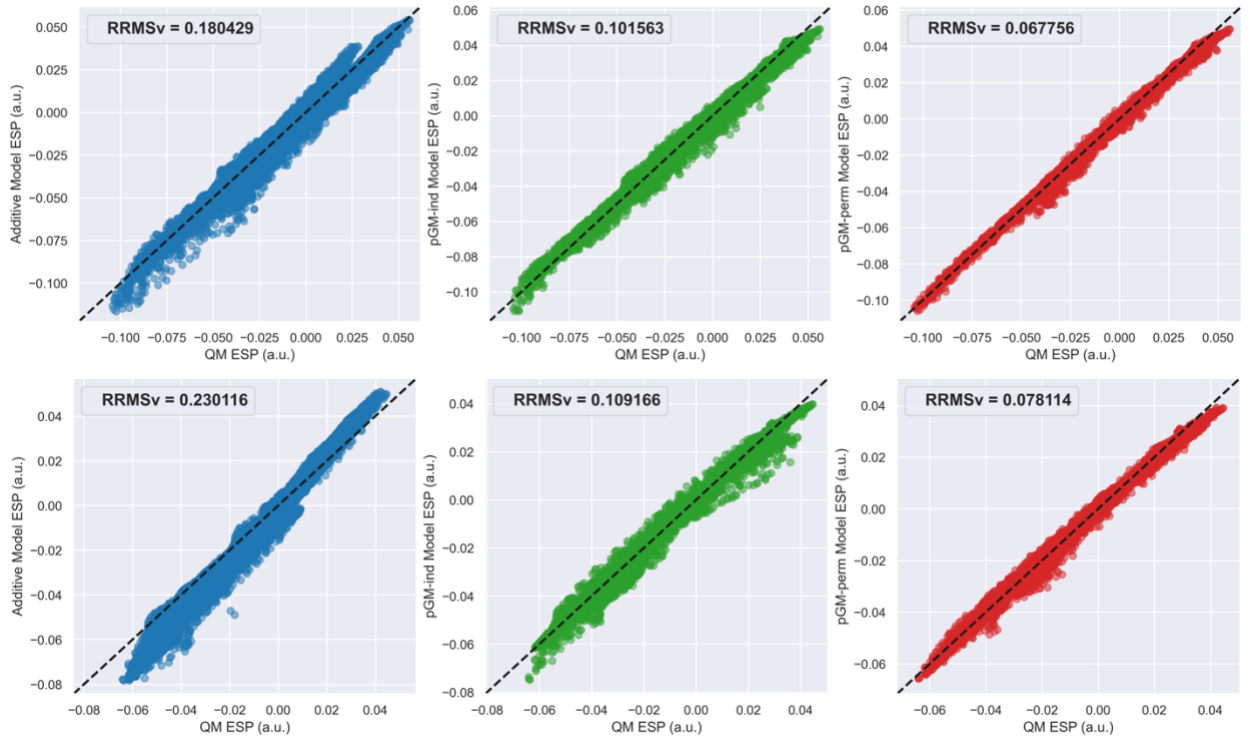


Figure 6.8. Scatterplots of MM ESPs of the additive, pGM-ind, and pGM-perm models versus QM ESPs for representative WC base pair tetramers. The upper panel is for the G-C/G-C tetramer with the highest QM dipole (dipole = 10.5748 Debye). Each plot shows a total of 14015 data points, with each point representing an ESP point. The lower panel is for the A-T/T-A tetramer with the lowest QM dipole (dipole = 2.1904 Debye). Each plot shows a total of 14196 data points, with each point representing an ESP point.

6.4. Discussion and Conclusions

Two desirable properties of molecular mechanical force fields are accuracy and transferability. Various previous works have demonstrated the accuracy of the pGM models.[50, 53-54] In this chapter, we assessed the transferability of the electrostatic parameters of the

pGM-ind and pGM-perm models by exploring whether the pGM models can accurately reproduce the electrostatic properties of larger molecular systems or different molecular conformations other than the molecules or conformations used for parametrizations. Encouragingly, as measured by $RRMS_{\mu}$ and $ARRMS_{V}$, both the pGM-ind and pGM-perm models show significantly better transferability than the point charge additive model. This has been demonstrated in the transferability tests (1) from water monomer to water oligomer clusters with various sizes; (2) across different conformations of amino acid dipeptides or tetrapeptides with widespread distributions of molecular dipole moments; (3) from amino acid tetrapeptides to longer polypeptides with up to 20 amino acid residues; and (4) from nucleobase monomers to WC base pair dimers and tetramers, which play key roles in the formation of double-stranded helical structures of DNA molecules. This and previous assessments together show that the accurate and transferable pGM models have the potential to serve as foundations for developing the next-generation polarizable force fields for modeling various biological processes that are sensitive to the polarization effects.

Another focus of this chapter is to identify the optimal parameterization strategy of amino acids for developing the next generation polarizable force fields based on the pGM models. Taking previous AMBER force fields as examples, the amino acid charge sets of the ff94 additive force field[5-6] and the ff02 polarizable force field[33] were both derived with C5/αR double-conformation fittings using amino acid dipeptides, and that of the ff12pol polarizable force field[35-38] was derived with αR/β/pII triple-conformation fittings, also using amino acid dipeptides. The electrostatic terms of the ff94 force field were parameterized using the *RESP* program,[13-14] which have remain unchanged in various subsequent additive Amber force fields for almost 30 years.[6-7, 17-22] The electrostatic terms of the ff02 and ff12pol

force fields were parameterized using an iterative charge fitting program named *i_RESP*.[23] Recently, the *PyRESP* program that performs electrostatic parameterizations for the pGM models using a direct matrix form solvation approach has been implemented.[54] Therefore, we aim to identify the amino acid conformations and the number of conformations for parameterizing the pGM models that lead to the optimal transferability. We first tested parametrizations using dipeptides in 1-5 conformations. However, although the electrostatic parameters derived by fitting dipeptides transfer well across the 14 different dipeptide conformations, the transferability from dipeptides to tetrapeptides is unsatisfactory. Therefore, we moved on to test parametrizations using tetrapeptides directly. Encouragingly, the αR/β double-conformation fitting with tetrapeptides shows great transferability not only across different tetrapeptide conformations, but also from tetrapeptides to longer polypeptides with lengths ranging from 1 to 20 repetitive amino acid residues for both the pGM-ind and pGM-perm models. In the future development of the pGM force fields for proteins, the αR/β double-conformation fittings with tetrapeptides are expected to be applied to derive the electrostatic parameters of all amino acids systematically.

An important question is: between the pGM-ind and pGM-perm models, which one has better transferability? In theory, the more elaborate pGM-perm model with atomic permanent dipoles has higher degree of freedom for parametrization, which can better reproduce the ESPs used for fitting and give better description for molecular electrostatic properties such as electric moments, leading to better transferability. This is indeed the case for water molecules as shown in **Figure 6.1-6.3**, where the pGM-perm and pGM-perm-v models yield much lower $ARRMS_V$ than the pGM-ind model, regardless of the water

oligomer cluster size. Additionally, all pGM models give similar $RRMS_{\mu}$ for each water oligomer cluster data set. However, as discussed in the **Appendix B**, the parameterization of the pGM-perm model suffers from the singularity problem for most biomolecules, due to the use of the permanent dipole local frame formed by CBVs. In contrast, the pGM-ind model does not have this problem since it does not take atomic permanent dipoles into account. In theory, the singularity problem can be addressed by the restrained fitting strategy as well as the multiple-conformation fitting strategy implemented in the *PyRESP* program. As shown in **Figure 6.4-6.7**, for single-conformation fittings of alanine dipeptides, alanine tetrapeptides, and nucleobases, which are all singular molecules, the pGM-ind model consistently shows better transferability than the pGM-perm model, as measured by both $RRMS_{\mu}$ and $ARRMS_{V}$. With multiple-conformation fittings, the pGM-perm model generally outperforms the pGM-ind model, especially in the transferability from amino acid tetrapeptides to longer amino acid polypeptides. Therefore, we conclude that the pGM-perm model can be expected to give better transferability than the pGM-ind model for nonsingular molecules such as water. For singular molecules such as amino acids and nucleotides, if there are more than one conformation available for multiple-conformation fittings, the pGM-perm model is expected to give better transferability; otherwise, the pGM-ind model is expected to give better transferability for single-conformation fittings.

Another important question for future users that wish to parameterize non-standard molecules (such as small molecule ligands) is: what types of conformations should be used for parameterizing the pGM models in general? For molecules that have rigid conformations such as nucleobases, there is probably not too many choices. However, the transferability tests on amino acids provide some insights for the parameterizations of flexible molecules.

For the parameterizations of both the alanine dipeptides and alanine tetrapeptides, we tested the single-conformation fittings and double-conformation fittings using conformations with the highest (αR for both dipeptide and tetrapeptide), lowest (β for both dipeptide and tetrapeptide), and intermediate ($C7_{eq}$ for dipeptide and pII for tetrapeptide) molecular dipole moments. Among all single-conformation fittings, the conformations with intermediate dipole moments ($C7_{eq}$ or pII) consistently give the best overall performance for the transferability of the pGM-ind and pGM-perm models. In contrast, among all double-conformation fittings, the best overall performance is consistently given by the combination of the conformations with the highest (αR) and lowest (β) dipole moments. Therefore, for selecting conformations for the parameterizations of flexible molecules, conformations with intermediate molecular dipole moments are recommended for single-conformation fittings, while the combination of conformations with widespread molecular dipole moments (such as conformations with the highest and lowest dipoles from all available conformations) are recommended for multiple-conformation fittings.

Our goal is to develop applicable and accessible pGM force fields for the molecular modeling community to perform simulation works on biomolecular systems that are sensitive to polarization effects. In future works, the electrostatic parameters of all amino acids and nucleotides for the pGM models will be derived using the strategy of restrained fitting in combination with multiple-conformation fitting provided by the *PyRESP* program.[54] A polarizable water model based on the pGM models will also be developed and analyzed. In addition, the van der Waals parameters for the pGM models need to be reoptimized using a similar strategy as was used in the development of the ff12pol force field.[38]

6.5. Supporting Information

Table S6.1. The QM Molecular Dipole Moments (Debye) of Alanine Polypeptides (ACE-ALA_n-NME) from the ALA-poly Data Set[a]

n	aβ	αR	β
1	2.2120	6.3642	0.7558
2	5.0324	9.2828	3.8851
3	4.6954	12.8546	1.4167
4	6.9891	16.5098	4.1018
5	7.2697	20.1370	2.1320
6	9.2825	23.9253	4.4437
7	9.8749	27.8535	2.8647
8	11.7179	31.7575	4.8784
9	12.4931	35.6524	3.6046
10	14.2228	39.6364	5.3824
11	15.1179	43.6645	4.3481
12	16.7661	47.6550	5.9376

13	17.7462	51.6496	5.0937
14	19.3327	55.7060	6.5310
15	20.3767	59.7685	5.8404
16	21.9144	63.7981	7.1528
17	23.0086	67.8427	6.5878
18	24.5065	71.9273	7.7963
19	25.6412	76.0020	7.3357
20	27.1060	80.0517	8.4565

[a] The QM molecular dipole moments are calculated at the ωB97X-D/aug-cc-pVTZ level of theory.

Table S6.2. The QM Molecular Dipole Moments (Debye) of Glycine Polypeptides (ACE-GLY_n-NME) from the GLY-poly Data Set[a]

n	aβ	αR	β
1	2.2053	6.4344	0.7992
2	5.0095	9.4508	3.8592
3	4.6377	12.9440	1.2402
4	6.9072	16.4893	3.9665

5	7.1531	20.0470	1.7384
6	9.1329	23.7208	4.1480
7	9.6946	27.4790	2.2585
8	11.4985	31.2236	4.3893
9	12.2468	34.9693	2.7887
10	13.9330	38.7742	4.6800
11	14.8043	42.6044	3.3240
12	16.4059	46.4132	5.0110
13	17.3647	50.2254	3.8623
14	18.9021	54.0766	5.3749
15	19.9268	57.9298	4.4023
16	21.4135	61.7641	5.7654
17	22.4900	65.6093	4.9436
18	23.9355	69.4781	6.1776
19	25.0540	73.3424	5.4858
20	26.4651	77.1917	6.6076

[a] The QM molecular dipole moments are calculated at the ωB97X-D/aug-cc-pVTZ level of theory.

Table S6.3. The QM Molecular Dipole Moments of WC Base Pair Tetramers from the BASE Data Set[a]

Tetramers	μ/Debye	Tetramers	μ/Debye
A-T/A-T	3.4309	G-C/A-T	7.0369
A-T/T-A	2.1904	G-C/T-A	5.3337
A-T/C-G	5.3321	G-C/C-G	6.5897
A-T/G-C	6.9265	G-C/G-C	10.5748

[a] The QM molecular dipole moments are calculated at the ωB97X-D/aug-cc-pVTZ level of theory.

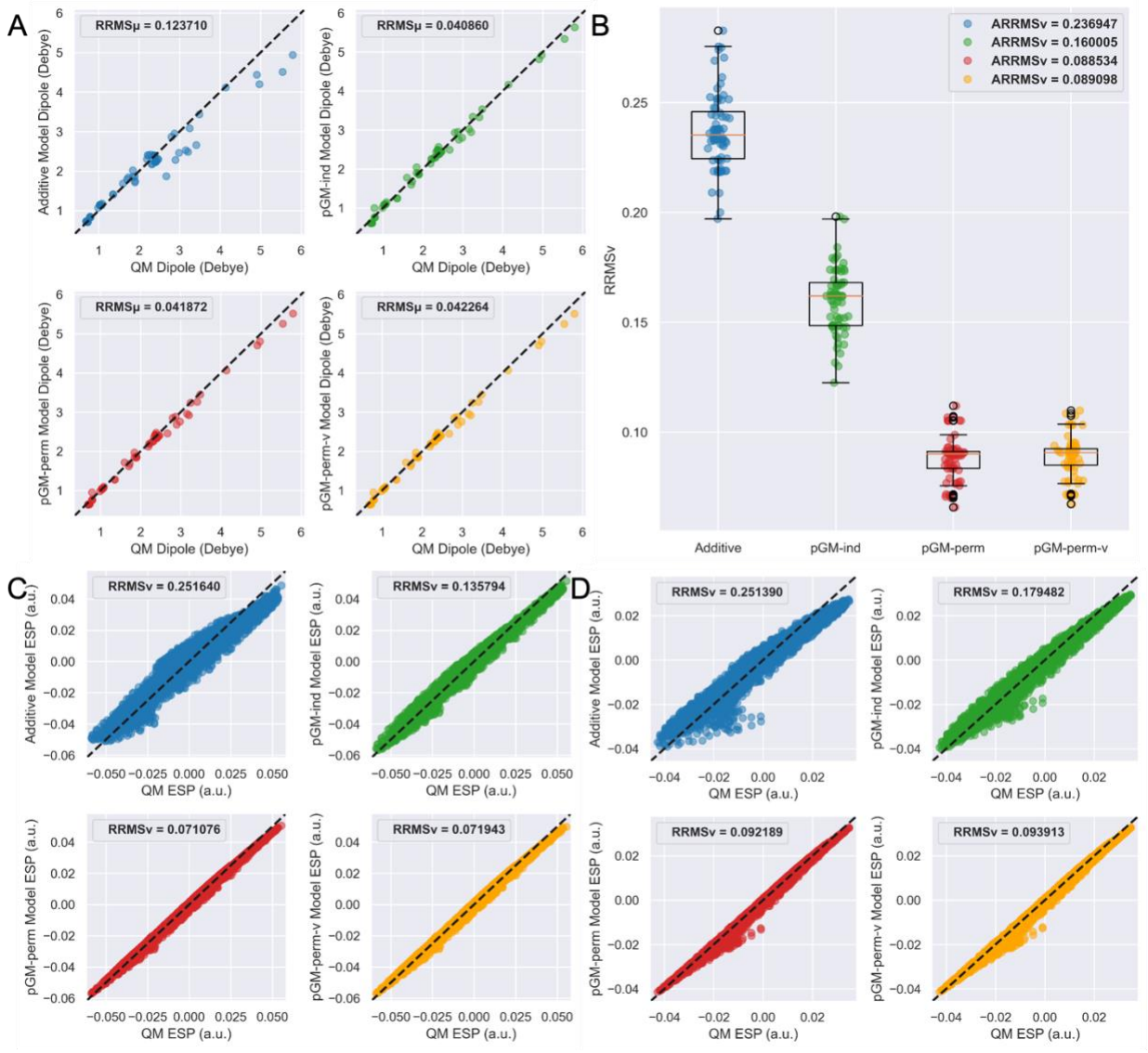


Figure S6.1. The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water hexamer clusters. **A.** Scatterplots of MM dipoles of each electrostatic model versus QM dipoles. Each plot shows a total of 72 data points, with each point representing a water hexamer. **B.** Boxplots of the $RRMS_V$ of each electrostatic model with QM results. Each plot shows a total of 72 data points, with each point representing a water hexamer. **C.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water hexamer with the highest QM dipole (dipole = 5.7951 Debye). Each

plot shows a total of 5907 data points, with each point representing an ESP point. **D.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water hexamer with the lowest QM dipole (dipole = 0.6914 Debye). Each plot shows a total of 5992 data points, with each point representing an ESP point. For **A**, **C**, and **D**, the dashed lines correspond to perfect matching.

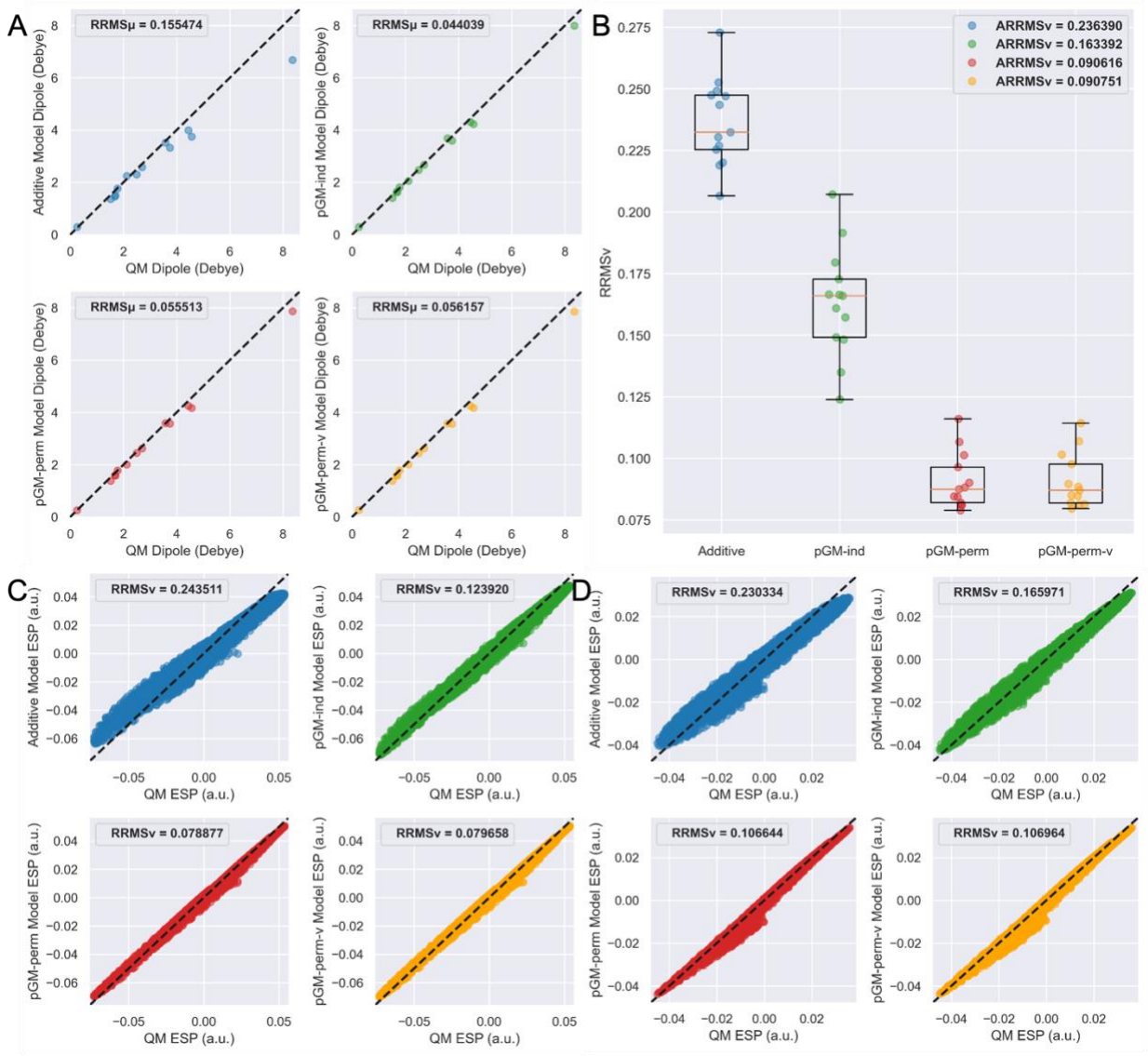


Figure S6.2. The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water octamer clusters. **A.** Scatterplots of MM dipoles of each electrostatic model versus QM dipoles. Each plot shows a total of 13 data points, with each point representing a water octamer. **B.** Boxplots of the $RRMS_V$ of each electrostatic model with QM results. Each plot shows a total of 13 data points, with each point representing a water octamer. **C.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water octamer with the highest QM dipole (dipole = 8.3545 Debye). Each plot shows a total of 7025 data points, with each point representing an ESP point. **D.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water octamer with the lowest QM dipole (dipole = 0.2514 Debye). Each plot shows a total of 7139 data points, with each point representing an ESP point. For **A**, **C**, and **D**, the dashed lines correspond to perfect matching.

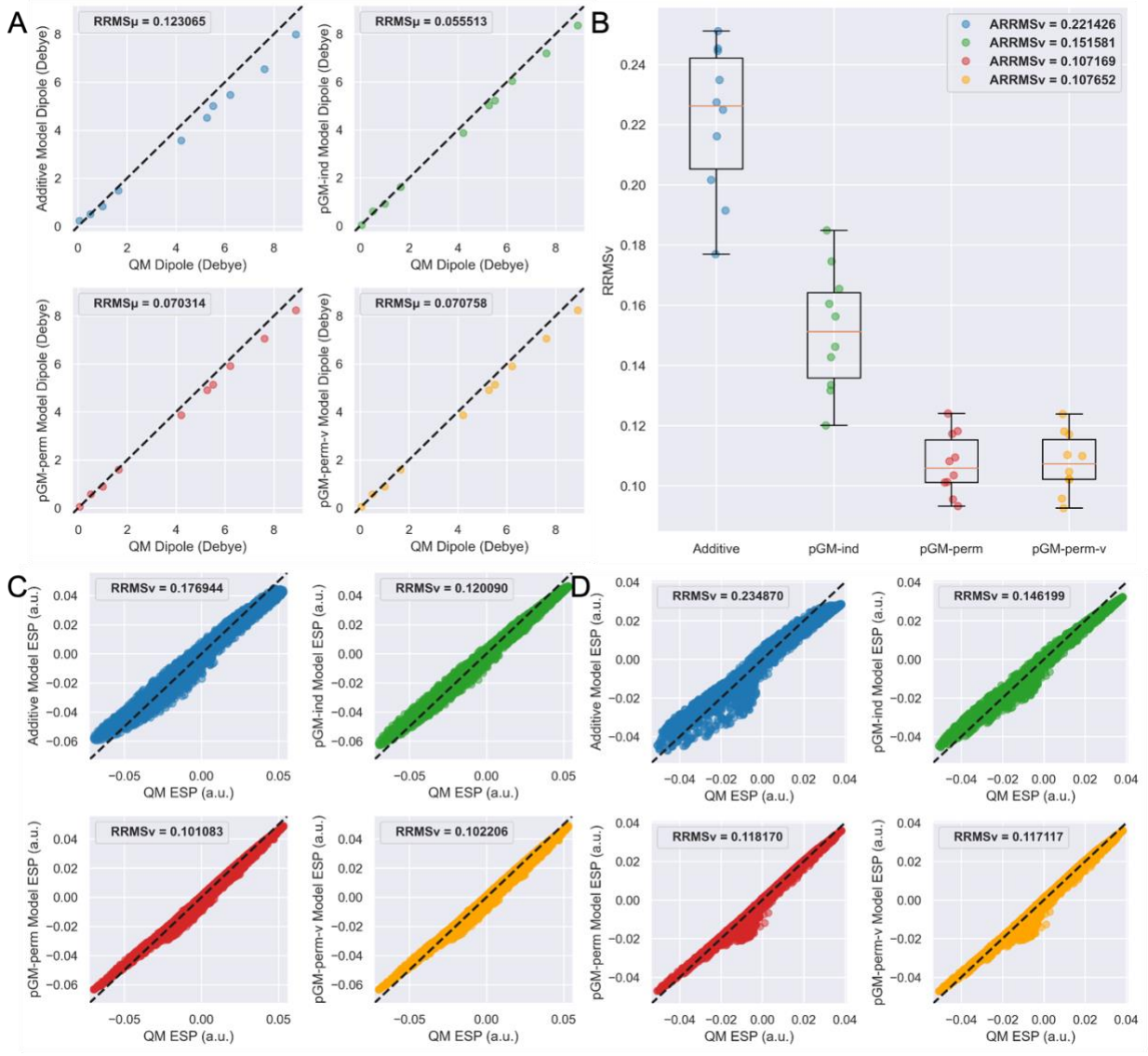


Figure S6.3. The transferability tests of the additive, pGM-ind, pGM-perm, and pGM-perm-v models from water monomer to water decamer clusters. **A.** Scatterplots of MM dipoles of each electrostatic model versus QM dipoles. Each plot shows a total of 10 data points, with each point representing a water decamer. **B.** Boxplots of the $RRMS_V$ of each electrostatic model with QM results. Each plot shows a total of 10 data points, with each point representing a water decamer. **C.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water decamer with the highest QM dipole (dipole = 8.9013 Debye). Each

plot shows a total of 7887 data points, with each point representing an ESP point. **D.** Scatterplots of MM ESPs of each electrostatic model versus QM ESPs for the water decamer with the lowest QM dipole (dipole = 0.0604 Debye). Each plot shows a total of 7669 data points, with each point representing an ESP point. For **A**, **C**, and **D**, the dashed lines correspond to perfect matching.

References

1. Leach, A. R., *Molecular modelling: principles and applications*. 2nd ed.; Pearson education: 2001.
2. Vitalis, A.; Pappu, R. V., Methods for Monte Carlo simulations of biomacromolecules. *Annual reports in computational chemistry* **2009,** *5*, 49-76.
3. Monticelli, L.; Tieleman, D. P., Force fields for classical molecular dynamics. *Biomolecular simulations* **2013**, 197-213.
4. Salomon‐Ferrer, R.; Case, D. A.; Walker, R. C., An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013,** *3* (2), 198-210.
5. Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A., Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *Journal of computational chemistry* **1995,** *16* (11), 1357-1377.
6. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society* **1995,** *117* (19), 5179-5197.
7. Wang, J.; Cieplak, P.; Kollman, P. A., How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of computational chemistry* **2000,** *21* (12), 1049-1074.
8. Momany, F. A., Determination of partial atomic charges from ab initio molecular electrostatic potentials. Application to formamide, methanol, and formic acid. *The Journal of Physical Chemistry* **1978,** *82* (5), 592-601.
9. Cox, S.; Williams, D., Representation of the molecular electrostatic potential by a net atomic charge model. *Journal of computational chemistry* **1981,** *2* (3), 304-323.
10. Chirlian, L. E.; Francl, M. M., Atomic charges derived from electrostatic potentials: A detailed study. *Journal of computational chemistry* **1987,** *8* (6), 894-905.
11. Besler, B. H.; Merz Jr, K. M.; Kollman, P. A., Atomic charges derived from semiempirical methods. *Journal of computational chemistry* **1990,** *11* (4), 431-439.
12. Breneman, C. M.; Wiberg, K. B., Determining atom‐centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of computational chemistry* **1990,** *11* (3), 361-373.

13. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993,** *97* (40), 10269-10280.
14. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A., Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* **1993,** *115* (21), 9620-9631.
15. Reynolds, C. A.; Essex, J. W.; Richards, W. G., Atomic charges for variable molecular conformations. *Journal of the American Chemical Society* **1992,** *114* (23), 9075-9079.
16. Stouch, T.; Williams, D. E., Conformational dependence of electrostatic potential derived charges of a lipid headgroup: Glycerylphosphorylcholine. *Journal of computational chemistry* **1992,** *13* (5), 622-632.
17. Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q., ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *Journal of chemical theory and computation* **2019,** *16* (1), 528-552.
18. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of chemical theory and computation* **2015,** *11* (8), 3696-3713.
19. Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **2006,** *65* (3), 712-725.
20. Zgarbová, M.; Sponer, J.; Jurečka, P., Z-DNA as a Touchstone for Additive Empirical Force Fields and a Refinement of the Alpha/Gamma DNA torsions for AMBER. *Journal of Chemical Theory and Computation* **2021,** *17* (10), 6292-6301.
21. Zgarbová, M.; Sponer, J.; Otyepka, M.; Cheatham III, T. E.; Galindo-Murillo, R.; Jurecka, P., Refinement of the sugar–phosphate backbone torsion beta for AMBER force fields improves the description of Z-and B-DNA. *Journal of chemical theory and computation* **2015,** *11* (12), 5723-5736.
22. Zgarbová, M.; Otyepka, M.; Šponer, J. i.; Mládek, A. t.; Banáš, P.; Cheatham III, T. E.; Jurecka, P., Refinement of the Cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *Journal of chemical theory and computation* **2011,** *7* (9), 2886-2902.
23. Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J., Polarization effects in molecular mechanical force fields. *Journal of Physics: Condensed Matter* **2009,** *21* (33), 333102.
24. Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P., Anisotropic, polarizable molecular mechanics studies of inter-and intramolecular interactions and ligand− macromolecule complexes. A bottom-up strategy. *Journal of chemical theory and computation* **2007,** *3* (6), 1960-1986.
25. Zhao, S.; Schaub, A. J.; Tsai, S.-C.; Luo, R., Development of a Pantetheine Force Field Library for Molecular Modeling. *Journal of chemical information and modeling* **2021,** *61* (2), 856-868.
26. King, E.; Qi, R.; Li, H.; Luo, R.; Aitchison, E., Estimating the roles of protonation and electronic polarization in absolute binding affinity simulations. *Journal of chemical theory and computation* **2021,** *17* (4), 2541-2555.
27. Draper, D. E.; Grilley, D.; Soto, A. M., Ions and RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **2005,** *34*, 221-243.
28. Gkionis, K.; Kruse, H.; Platts, J. A.; Mladek, A.; Koca, J.; Sponer, J., Ion binding to quadruplex DNA stems. Comparison of MM and QM descriptions reveals sizable polarization effects not included in contemporary simulations. *Journal of chemical theory and computation* **2014,** *10* (3), 1326-1340.
29. Fitch, C. A.; Karp, D. A.; Lee, K. K.; Stites, W. E.; Lattman, E. E.; García-Moreno, E. B., Experimental pKa values of buried residues: analysis with continuum methods and role of water penetration. *Biophysical journal* **2002,** *82* (6), 3289-3304.

30. Lin, Z.; van Gunsteren, W. F., Effects of Polarizable Solvent Models upon the Relative Stability of an α-Helical and a β-Hairpin Structure of an Alanine Decapeptide. *Journal of Chemical Theory and Computation* **2015,** *11* (5), 1983-1986.
31. Peng, X.; Zhang, Y.; Chu, H.; Li, Y.; Zhang, D.; Cao, L.; Li, G., Accurate evaluation of ion conductivity of the gramicidin a channel using a polarizable force field without any corrections. *Journal of chemical theory and computation* **2016,** *12* (6), 2973-2982.
32. Sun, R.-N.; Gong, H., Simulating the activation of voltage sensing domain for a voltage-gated sodium channel using polarizable force field. *The Journal of Physical Chemistry Letters* **2017,** *8* (5), 901-908.
33. Cieplak, P.; Caldwell, J.; Kollman, P., Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases. *Journal of computational chemistry* **2001,** *22* (10), 1048-1057.
34. Wang, Z. X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y., Strike a balance: optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides. *Journal of computational chemistry* **2006,** *27* (6), 781-790.
35. Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations I: parameterization of atomic polarizability. *The journal of physical chemistry B* **2011,** *115* (12), 3091-3099.
36. Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations II: induced dipole models significantly improve accuracy of intermolecular interaction energies. *The journal of physical chemistry B* **2011,** *115* (12), 3100-3111.
37. Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Luo, R., Development of polarizable models for molecular mechanical calculations. 3. Polarizable water models conforming to Thole polarization screening schemes. *The Journal of Physical Chemistry B* **2012,** *116* (28), 7999-8008.
38. Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.-J.; Luo, R.; Duan, Y., Development of polarizable models for molecular mechanical calculations. 4. van der Waals parametrization. *The Journal of Physical Chemistry B* **2012,** *116* (24), 7088-7101.
39. Banks, J. L.; Kaminski, G. A.; Zhou, R.; Mainz, D. T.; Berne, B.; Friesner, R. A., Parametrizing a polarizable force field from ab initio data. I. The fluctuating point charge model. *The Journal of chemical physics* **1999,** *110* (2), 741-754.
40. Patel, S.; Brooks III, C. L., CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *Journal of computational chemistry* **2004,** *25* (1), 1-16.
41. Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell Jr, A. D., A polarizable model of water for molecular dynamics simulations of biomolecules. *Chemical Physics Letters* **2006,** *418* (1-3), 245-249.
42. Lopes, P. E.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Polarizable empirical force field for aromatic compounds based on the classical drude oscillator. *The Journal of Physical Chemistry B* **2007,** *111* (11), 2873-2885.
43. Tan, Y.-H.; Luo, R., Continuum treatment of electronic polarization effect. *J. Chem. Phys.* **2007,** *126* (9), 094103.
44. Tan, Y.-H.; Tan, C.; Wang, J.; Luo, R., Continuum polarizable force field within the Poisson-Boltzmann framework. *J. Phys. Chem. B* **2008,** *112* (25), 7675-7688.
45. Applequist, J.; Carl, J. R.; Fung, K.-K., Atom dipole interaction model for molecular polarizability. Application to polyatomic molecules and determination of atom polarizabilities. *Journal of the American Chemical Society* **1972,** *94* (9), 2952-2960.

46. Thole, B. T., Molecular polarizabilities calculated with a modified dipole interaction. *Chemical Physics* **1981,** *59* (3), 341-350.
47. Elking, D.; Darden, T.; Woods, R. J., Gaussian induced dipole polarization model. *Journal of computational chemistry* **2007,** *28* (7), 1261-1274.
48. Elking, D. M.; Cisneros, G. A.; Piquemal, J.-P.; Darden, T. A.; Pedersen, L. G., Gaussian multipole model (GMM). *Journal of chemical theory and computation* **2010,** *6* (1), 190-202.
49. Elking, D. M.; Perera, L.; Duke, R.; Darden, T.; Pedersen, L. G., Atomic forces for geometry-dependent point multipole and Gaussian multipole models. *Journal of computational chemistry* **2010,** *31* (15), 2702-2713.
50. Wang, J.; Cieplak, P.; Luo, R.; Duan, Y., Development of Polarizable Gaussian Model for Molecular Mechanical Calculations I: Atomic Polarizability Parameterization To Reproduce ab Initio Anisotropy. *J. Chem. Theory Comput.* **2019,** *15* (2), 1146-1158.
51. Wei, H.; Qi, R.; Wang, J.; Cieplak, P.; Duan, Y.; Luo, R., Efficient formulation of polarizable Gaussian multipole electrostatics for biomolecular simulations. *The Journal of chemical physics* **2020,** *153* (11), 114116.
52. Wei, H.; Cieplak, P.; Duan, Y.; Luo, R., Stress tensor and constant pressure simulation for polarizable Gaussian multipole model. *The Journal of chemical physics* **2022,** *156* (11), 114114.
53. Zhao, S.; Wei, H.; Cieplak, P.; Duan, Y.; Luo, R., Accurate Reproduction of Quantum Mechanical Many-Body Interactions in Peptide Main-Chain Hydrogen-Bonding Oligomers by the Polarizable Gaussian Multipole Model. *Journal of Chemical Theory and Computation* **2022,** *18* (10), 6172-6188.
54. Zhao, S.; Wei, H.; Cieplak, P.; Duan, Y.; Luo, R., PyRESP: A Program for Electrostatic Parameterizations of Additive and Induced Dipole Polarizable Force Fields. *Journal of Chemical Theory and Computation* **2022,** *18* (6), 3654-3670.
55. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983,** *79* (2), 926-935.
56. Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; Cheatham III, T. E.; Cisneros, G. A.; Cruzeiro, V. W. D.; Darden, T. A.; Duke, R. E.; Giambasu, G.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, R.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shajan, A.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiong, Y.; Xue, Y.; York, D. M.; Zhao, S.; Kollman, P. A., *Amber 2022*. University of California, San Francisco: 2022.
57. Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A., Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields. *Journal of the American Chemical Society* **1997,** *119* (25), 5908-5920.
58. Bansal, M.; Bhattacharyya, D.; Ravi, B., NUPARM and NUCGEN: software for analysis and generation of sequence dependent nucleic acid structures. *Bioinformatics* **1995,** *11* (3), 281-287.
59. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.;

Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J., Gaussian 16 Revision A. 03. 2016; Gaussian Inc. *Wallingford CT* **2016,** *2* (4).
60. Connolly, M. L., Analytical molecular surface calculation. *Journal of applied crystallography* **1983,** *16* (5), 548-558.
61. Singh, U. C.; Kollman, P. A., An approach to computing electrostatic charges for molecules. *Journal of computational chemistry* **1984,** *5* (2), 129-145.
62. Xie, W.; Pu, J.; Gao, J., A coupled polarization-matrix inversion and iteration approach for accelerating the dipole convergence in a polarizable potential function. *The Journal of Physical Chemistry A* **2009,** *113* (10), 2109-2116.
63. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E., UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* **2004,** *25* (13), 1605-1612.

Appendix A: Proof of Many-Body Interaction Energies Decomposition

In Chapter 5, we claim that the many-body interaction energies $\mathrm{ME}(\mathrm{Gly_m}:\mathrm{Gly_n})$ (**eq 5.22**) can be decomposed into the non-additive contributions $\mathrm{ME_{NA}}(\mathrm{Gly_m}:\mathrm{Gly_n})$ and the additive contributions $\mathrm{ME_A}(\mathrm{Gly_m}:\mathrm{Gly_n})$, whose formulas are given in **eq 5.24** and **eq 5.25**, respectively. We first prove the formula of the non-additive contributions in subsection **A1**, and then prove the formula of the additive contributions in subsection **A2**.

A1. Non-additive Contributions

The non-additive effect refers to that, for a molecular system with more than two atoms involved, any two atoms will interact differently compared with the situation where other atoms were not present.[1] For additive force fields, the non-additive effect does not exist, i.e., $\mathrm{ME_{NA}}(\mathrm{Gly_m}:\mathrm{Gly_n})$ defined in **eq 5.24** is always zero. Therefore, the interaction energy of the two middle peptides $\mathrm{IE_{mid}}(\mathrm{Gly_m}:\mathrm{Gly_n})$ in the presence of the neighbor peptides $\mathrm{Gly_{m-1}}$ and $\mathrm{Gly_{n-1}}$ defined in **eq 5.23** should be the same as the interaction energy of the two middle peptides IE(Gly: Gly) in the absence of the neighbor peptides. The key to prove this is that, for additive force fields, the interaction energy IE(A, B: C, D) of a four-body system A,B:C,D can be decomposed into

$$\mathrm{IE}(\mathrm{A},\mathrm{B}:\mathrm{C},\mathrm{D}) = \mathrm{IE}(\mathrm{A}:\mathrm{C}) + \mathrm{IE}(\mathrm{B}:\mathrm{C}) + \mathrm{IE}(\mathrm{A}:\mathrm{D}) + \mathrm{IE}(\mathrm{B}:\mathrm{D}) \tag{A1}$$

Therefore, the first three terms in **eq 5.23** can be decomposed to

$$\begin{aligned}\mathrm{IE}(\mathrm{Gly_m}:\mathrm{Gly_n}) = {} & \mathrm{IE}(\mathrm{Gly_{m-1}X}:\mathrm{XGly_{n-1}}) + \mathrm{IE}(\mathrm{Gly_{m-1}X}:\mathrm{Gly}) \\ & + \mathrm{IE}(\mathrm{Gly}:\mathrm{XGly_{n-1}}) + \mathrm{IE}(\mathrm{Gly}:\mathrm{Gly})\end{aligned} \tag{A2}$$

$$\mathrm{IE}(\mathrm{Gly_m}\colon \mathrm{XGly_{n-1}}) = \mathrm{IE}(\mathrm{Gly_{m-1}X}\colon \mathrm{XGly_{n-1}}) + \mathrm{IE}(\mathrm{Gly}\colon \mathrm{XGly_{n-1}}) \quad \text{(A3)}$$

$$\mathrm{IE}(\mathrm{Gly_{m-1}X}\colon \mathrm{Gly_n}) = \mathrm{IE}(\mathrm{Gly_{m-1}X}\colon \mathrm{XGly_{n-1}}) + \mathrm{IE}(\mathrm{Gly_{m-1}X}\colon \mathrm{Gly}) \quad \text{(A4)}$$

Substitute **eq A2-A4** into **eq 5.23** gives

$$\mathrm{IE_{mid}}(\mathrm{Gly_m}\colon \mathrm{Gly_n}) = \mathrm{IE}(\mathrm{Gly}\colon \mathrm{Gly}) \quad \text{(A5)}$$

Therefore, for additive force fields, the non-additive contribution $\mathrm{ME_{NA}}(\mathrm{Gly_m}\colon \mathrm{Gly_n})$ in **eq 5.24** becomes zero. For polarizable force fields, the difference between $\mathrm{IE_N}(\mathrm{Gly_m}\colon \mathrm{Gly_n})$ and $\mathrm{IE}(\mathrm{Gly}\colon \mathrm{Gly})$ is naturally the non-additive contribution $\mathrm{ME_{NA}}(\mathrm{Gly_m}\colon \mathrm{Gly_n})$, which is a non-zero value.

A2. Additive Contributions

For either additive or polarizable force fields, the additive contribution can be expressed in the following alternative formula by substituting **eq 5.22-5.24** into **eq 5.25**

$$\mathrm{ME_A}(\mathrm{Gly_m}\colon \mathrm{Gly_n}) = \mathrm{IE}(\mathrm{Gly_m}\colon \mathrm{XGly_{n-1}}) + \mathrm{IE}(\mathrm{Gly_{m-1}X}\colon \mathrm{Gly_n}) - \mathrm{IE}(\mathrm{Gly_{m-1}X}\colon \mathrm{XGly_{n-1}}) \quad \text{(A6)}$$

For additive force fields, we need to show that the many-body interaction energies $\mathrm{ME}(\mathrm{Gly_m}\colon \mathrm{Gly_n})$ only have the additive contribution $\mathrm{ME_A}(\mathrm{Gly_m}\colon \mathrm{Gly_n})$. This can be done easily by substituting **eq A2** into the formula of the many-body interaction energy in **eq 5.22**. Therefore, we have proved that for additive force fields, we have

$$\mathrm{ME}(\mathrm{Gly_m}\colon \mathrm{Gly_n}) = \mathrm{ME_A}(\mathrm{Gly_m}\colon \mathrm{Gly_n}) \quad \text{(A7)}$$

For polarizable force fields, the difference between the many-body interaction energy $ME(Gly_m:Gly_n)$ and the non-zero non-additive contribution $ME_{NA}(Gly_m:Gly_n)$ naturally gives the additive contribution $ME_A(Gly_m:Gly_n)$.

Appendix B: The Singularity Problem of the pGM-perm and pGM-perm-v Models and Solutions

The parameterizations of the pGM-perm and pGM-perm-v models suffer from the singularity problem that originates from the use of the permanent dipole local frame formed by covalent basis vectors (CBVs). Since CBVs are along the direction of covalent bonds (and virtual bonds for pGM-perm-v), some molecules are "singular molecules" due to the existence of "singular atoms". Taking carbon dioxide (CO_2) as an example, the two covalent bonds associated with the central carbon atom are colinear, so that the two permanent C-O dipoles oriented in opposite directions can be assigned any value to give zero net dipole to the carbon atom. Therefore, the carbon atom in CO_2 is a singular atom, and the CO_2 molecule is a singular molecule. **Figure A1** gives several examples of singular and nonsingular molecules. The water (H_2O) molecule is nonsingular. Similar to the case of CO_2, there are two covalent bonds associated with the central oxygen atom of water. However, the permanent O-H dipoles are not colinear, so that there only exists one solution for the value of the O-H dipole to give the correct atomic dipole for oxygen. The carbon atom of the ethene (C_2H_4) molecule and the nitrogen atom of the ammonia (NH_3) molecule both have three covalent bonds associated. However, ethene is singular but ammonia is nonsingular. The two C-H dipoles and the C-C dipole of each carbon atom in the ethene molecule are coplanar, so that

the net atomic dipole of the ethene molecule can be produced by infinitely many linear combinations of the three dipoles. In contrast, the three N-H dipoles of the nitrogen atom in the ammonia molecule are not coplanar, so that there only exists one solution for the value of the N-H dipole to give the correct atomic dipole for nitrogen. For atoms associated with more than three covalent bonds (and virtual bonds), such as the central carbon of the methane (CH_4) molecule, no matter how these bonds are oriented, there will always be infinitely many linear combinations of the dipoles on these bonds that can produce the net atomic dipole for the atom. Therefore, any atoms associated with more than three bonds are singular atoms, and any molecules containing this type of atoms are singular molecules. Furthermore, the virtual dipoles of the pGM-perm-v model may cause additional singularity problems during parameterization. For example, the oxygen atoms in CO_2 are nonsingular atoms in the pGM-perm model but are singular atoms in the pGM-perm-v model, since the O-C covalent dipole and O-O virtual dipole are colinear.

The general rule for checking whether an atom is singular in the context of the pGM-perm and pGM-perm-v models is as follows: First, count the number of covalent bonds and virtual bonds associated with this atom. If there is only one bond, the atom is nonsingular; If there are more than three bonds, the atom is singular. In the case of two bonds, the atom is singular if the two bonds are colinear and nonsingular if the two bonds are not colinear. In the case of three bonds, the atom is singular if the three bonds are coplanar and nonsingular if the two bonds are not coplanar. In fact, most biomolecules are singular molecules, due to the widespread existence of sp^3 carbons, such as the alpha carbon in every amino acid backbone, and the five carbons in the sugar unit of every nucleotide. If there is at least one singular atom exist in the molecule, the molecule is a singular molecule.

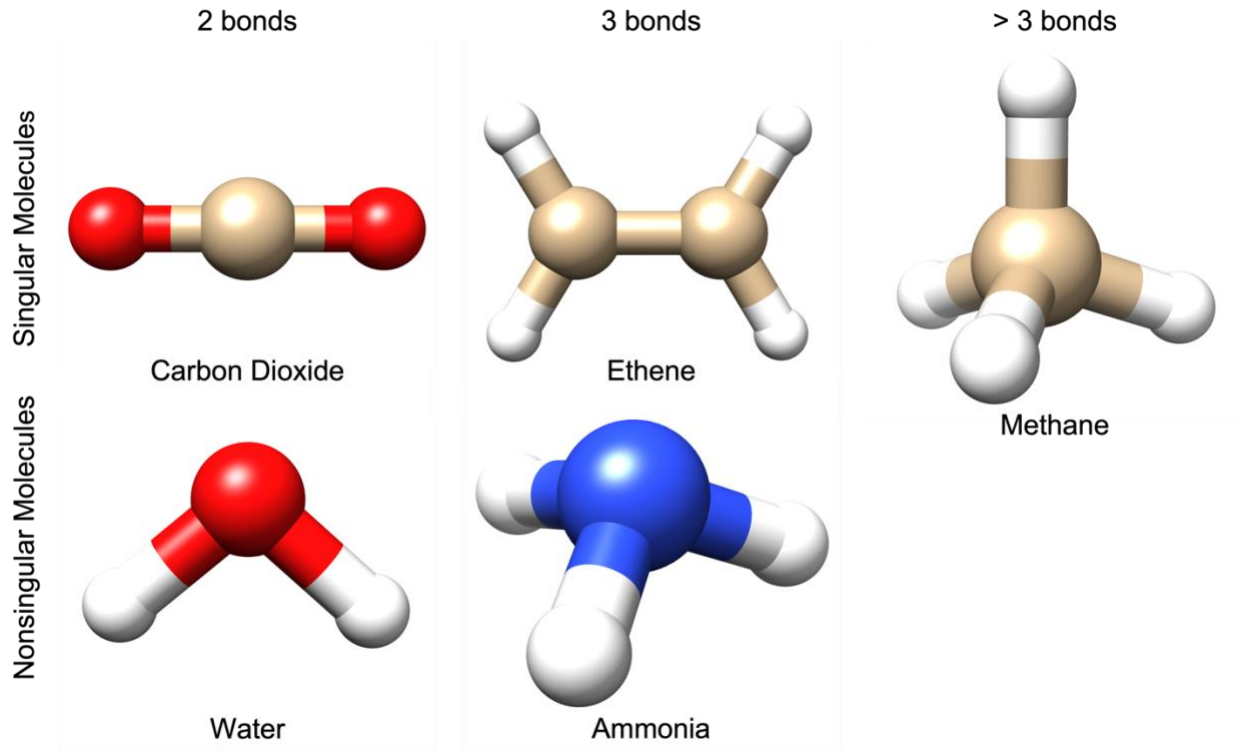


Figure A1. Several examples of singular and nonsingular molecules in the context of the parameterization of the pGM-perm model. The upper panel shows examples of singular molecules, and the lower panel shows examples of nonsingular molecules. In the left column, the singular carbon atom of the carbon dioxide (CO_2) molecule has 2 covalent bonds; In the middle column, the singular carbon atom of the ethene molecule has 3 covalent bonds; In the right column, the singular carbon atom of the methane molecule has 4 covalent bonds.

The mathematical explanation of the singularity problem is that the electrostatic parameterization of a molecule using the *PyRESP* program is essentially computing the least-squares solution of the following equation[2]

$$\boldsymbol{MQ} = \boldsymbol{V} \tag{A8}$$

where $\boldsymbol{Q}$ is a vector for all the point charges and permanent point dipoles of the molecule being parameterized, and the details of the equation can be found in our original *PyRESP* work.[2] The least-squares solution can be obtained by solving the following equation, the proof of which can be found in most linear algebra textbooks

$$\boldsymbol{M}^T\boldsymbol{MQ} = \boldsymbol{M}^T\boldsymbol{V} \tag{A9}$$

If **eq A9** has a unique solution, the square symmetric matrix $\boldsymbol{M}^T\boldsymbol{M}$ needs to be positive definite and invertible. However, for the parameterization of singular molecules such as methane with the pGM-perm or pGM-perm-v models, the matrix $\boldsymbol{M}$ contains linearly dependent columns, and the matrix $\boldsymbol{M}^T\boldsymbol{M}$ becomes a singular matrix, which is not invertible.

One solution to the singularity problem is the restrained fitting implemented the *PyRESP* program, which was originally implemented in its ancestor program *RESP*.[3-4] The *RESP* program applies the following hyperbolic restraining function χ to the least-squares fitting of additive models

$$\chi = a\sum_{i=1}^{n}\left(\sqrt{{q_i}^2 + b^2} - b\right) \tag{A10}$$

where q_i is the point charge of atom i; a is the scale factor that defines the restraining strength; b determines the "tightness" of the hyperbola around its minimum, which has been recommended to be set to 0.1 to make the restraint appropriately tight.[3] The *PyRESP* program extends the restraining functions of the *RESP* program by applying an additional penalty function with the same format as **eq A10** for restraining atomic permanent dipoles,

and allowing the users to choose different restraining strength a for point charges and permanent dipoles. In the restrained fitting process, the partial derivative of the penalty function χ to each electrostatic parameter is added to the diagonal terms of the matrix $\boldsymbol{M}^T\boldsymbol{M}$, introducing nonlinearity into the singular matrix. Therefore, the matrix $\boldsymbol{M}^T\boldsymbol{M}$ becomes invertible and **eq A9** has a unique solution.

Another solution to the singularity problem is the multiple-conformation fitting. By enforcing inter-molecular equivalences among multiple conformations of the same molecule, the rows and columns of the matrix $\boldsymbol{M}^T\boldsymbol{M}$ corresponding to equivalent permanent dipoles are added up to form a single row and column, giving rise to a smaller matrix $\boldsymbol{M}^T\boldsymbol{M}$. This operation essentially eliminates the linear dependence of the linearly dependent columns of the matrix $\boldsymbol{M}$, and the resulted smaller matrix $\boldsymbol{M}^T\boldsymbol{M}$ becomes invertible. However, the disadvantage of the multiple-conformation fitting strategy is that it may be difficult, if not impossible, to construct multiple optimized conformations for small rigid singular molecules such as CO_2, ethene, and methane. It is only an appropriate strategy for parameterizing large singular molecules such as amino acids and nucleotides.

References

1. Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J., Polarization effects in molecular mechanical force fields. *Journal of Physics: Condensed Matter* **2009,** *21* (33), 333102.
2. Zhao, S.; Wei, H.; Cieplak, P.; Duan, Y.; Luo, R., PyRESP: A Program for Electrostatic Parameterizations of Additive and Induced Dipole Polarizable Force Fields. *Journal of Chemical Theory and Computation* **2022,** *18* (6), 3654-3670.
3. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry* **1993,** *97* (40), 10269-10280.

4. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A., Application of RESP charges to calculate conformational energies, hydrogen bond energies, and free energies of solvation. *Journal of the American Chemical Society* **1993,** *115* (21), 9620-9631.