COMPUTATIONAL MOLECULAR DESIGN AND INNOVATION: FROM DRUG DISCOVERY TO EMERGING CONTAMINANTS

By

Yiğitcan Eken

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Chemistry—Doctor of Philosophy

2021

ABSTRACT

COMPUTATIONAL MOLECULAR DESIGN AND INNOVATION: FROM DRUG DISCOVERY TO EMERGING CONTAMINANTS

By

Yiğitcan Eken

Computational approaches have found great utility in areas including drug discovery and environmental contamination by investigating protein dynamics, binding and interaction patterns. For drug discovery, in silico biophysical methods serve an important role in reducing the cost of and accelerating the discovery process, as such methods aid in facilitating the identification, optimization and screening of potential drug candidates and in providing important understanding of drug mechanisms of actions and structure activity relationships at the atomic level. For computational drug discovery and protein modelling strategies, probable binding conformations of the ligand to its target can be predicted, and these conformations can be further evaluated by using scoring functions, molecular dynamics and free energy calculations to determine binding affinities and understand how a ligand recognizes its host.

Despite the utility of computational approaches in areas such as drug design and the study of protein functioning, the choice of methods is not straightforward. Because of this, a series of international blinded host-guest binding prediction challenges are available to identify the most effective approaches to predict a variety of properties. Some of the methods available for calculating free energies include free energy perturbation, replica exchange free energy perturbation and thermodynamic integration approaches, and end-state methods. The later are most promising due to their reduced computational cost and because there is no need for intermediate state simulation.

In this dissertation initially, performance of end-state approaches is considered. Then, computer simulations and modeling techniques combined with the optimal end-state parameter choices were used in application studies including; the ligand preference and biological function of three enzymes (Arthrobacter endoβ-N-acteylglucosaminidase, fibroblast growth factor-2 and heparanase), the effects of per- and polyfluoroalkyl substance binding on human pregnane X receptor and peroxisome proliferator-activated receptor gamma, and Ca^{2+} dependent activation of protein kinase C.

This dissertation is dedicated to family.

ACKNOWLEDGEMENTS

I am truly grateful for everyone who has supported me throughout my academic career. Firstly, I would like to thank my Ph.D. advisor and mentor, Professor Angela K. Wilson for her guidance and support over the years. I would also like to thank the past and current members of the Wilson group, for the insightful discussions, their support through the years and for amazing group activities including but not limited to Timothe, Lucas, Michael, Nuno, Semiha, Thanh, Zack, Hailey, Bradley, Guangyao, Sasha, Jared, Narasimhan, Prajay, Lenin, John D., John P., Zainab and Thomas. I would like to thank my committee members at Michigan State University; Professor Kenneth M. Merz Jr., Professor Xuefei Huang, Professor Edmund Ellsworth and my former committee member Professor Benjamin G. Levine.

I would like to thank all my friends. This includes but is certainly not limited to Hadi N., Christian S., Aslı Y., Erkan O., Kendal Ş., Şükrü A., Kaya V., Ufuk D., Yavuz B., Refik B., Emily C., Chelsea V., Oleksii K., Arial F., for the encouragement and the time, we have spent from dancing classes, exercises, gamming tournaments to kayaking. I also I want to thank my dancing teachers Richard and Alejandro for their positive and motivating classes during my time in Michigan.

Thank you to my large extended family and all their support from across the U.S. regardless of the distance your presence and support were always with me. To keep this list succinct: Erhan S., Ayhan S., Murat S., Müge S., Sabiha E., Gürol Ç., Yasemin Ç., Pelin A., Eric A., Tuba S.G., Aslı K., Burcu K. I would like to thank to my cousins; Neslihan Ç., Doruk S., Alper Ç., Ozan B. I am grateful for the time spent and memories. I also want to specifically acknowledge my cousin Ugur B. and my aunt Gülşen B. who passed away in 2018.

Special thanks to Lucila Garcia Lopez for her unconditional love, support. Your presence motivated me during my long nights of study and thank you for making sure I stayed focus and healthy during my doctoral candidacy exams. I also want to thank to the Garcia family for accepting/treating me as a family member.

Finally, I thank to my father, Muhammet E., sister Özgücan E., brother-in-law Burç T. and special thanks to my mother, Didem E. whom I lost recently in 2021: With our weekly and sometimes daily conversations you always kept me inspired, focused, and admired me throughout my life. While there is not enough words to describe how much love and support you have given to me my whole life, I am grateful for everything you have done for me forever.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction

1.1 Introduction

The growth in computational capabilities over the past decade has enabled computational biochemistry to be used on large biological systems such as proteins, viruses[1], and even whole cells.[2] The dynamic nature of proteins are linked to their function and small structural perturbations in an enzyme can affect proteins' activities across several orders of magnitude.[3] These structural perturbations, also referred to as conformational changes, occur on a timescale ranging from microseconds to seconds. Molecular dynamics simulations (MD) can be used to provide insight about the dynamics of proteins in order to understand specific biological phenomena. MD simulations can also be followed by binding free energy calculations to generate mechanistic hypotheses or activity predictions which can be further validated by laboratory experiments.[4] In the second chapter of this dissertation, computer-simulations and modeling methodologies used in the work presented in later chapter are overviewed.

In the third and fourth chapters of this dissertation, computer simulations and modeling were used to investigate host-guest binding on Statistical Assessment of the Modeling of Proteins and Ligands challenges (SAMPL6 and SAMPL7, respectively).[5,6] Host-guest structures are smaller in size and structurally less complex compared to ligand bound proteins. Due to their simplicity, host-guest systems are utilized in Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) challenges where the physicochemical properties predicted computationally are compared with experimental data to assess the reliability of different methods. During SAMPL challenges, a number of parameter choices have been considered including charge and solvation schemes and the reliability of Molecular Mechanics Poisson-Boltzmann Surface Area (MMPBSA) method in the prediction of binding free energies has been assessed. Due to the success of the methodology, it has been utilized in multiple studies here.

The research included in the fifth and sixth chapters of this dissertation was performed in collaboration with Professor Xuefei Huang and his research group from MSU. Chapter 5 describes computational investigation of the substrate binding preference between *Arthrobacter* endo-β-N-acteylglucosaminidase (Endo-A) enzyme and rare N-glycans synthesized by the Huang group. The computational results showed that inactive glycan hinders the gate amino acids in Endo-A and prohibits active site formation which is consistent with the compound's low glycosylation yield.[7]

In Chapter 6, the biological activity of human syndecan-4 glycopeptide bearing O-, N-sulfation and multiple aspartic acids upon heparanase and Fibroblast Growth Factor-2 (FGF-2) binding was studied computationally. Heparan sulfates (HS) are sulfated polysaccharides that have a range of biological functions including blood clothing prevention, growth factor and chemokine binding and controlling activity levels of various enzymes. *In vivo*, HS exists as a heterogenous mixtures where the length of their backbone and location of sulfates varies. Additionally, they can form proteoglycans where HS is covalently linked to a core protein or a core peptide. Originally, the core peptides are considered as do not possess any biological function. However, experiments performed by our collaborators and the modelling results shows that free HS and HS proteoglycan can poses different biological function.[8]

PFASs are man-made chemicals that are widely used in industrial products for food wrappers, fire-fighting foams, carpets, furniture, boots, clothes, non-stick cookware to name only a few. Regardless of extensive usage of PFAS for more than 50 years, recently environmental and health concerns related with PFAS exposure is recognized. As the EPA has banned some of the most common long-chain PFASs such as PFOA and PFOS alternatives such as ADONA, GenX and PFBS are now commonly used. In chapter 7 and chapter 8 research performed on per-

and polyfluoroalkyl substances (PFASs) are included. PFAS exposure has been linked to a number of serious health problems ranging from cancer to thyroid disease. In chapter 7, human pregnane X receptor (hPXR), a known PFAS targets that is important for sensing toxic substances within body, is studied for PFASs binding.[9] In Chapter 8 the same range of PFASs were studied along with recently discovered alternatives for peroxisome proliferator activated receptor γ (PPARγ) binding, a type II nuclear receptor fundamental in the regulation of genes, glucose metabolism, and insulin sensitization.[10] The models explain PFASs recognition on hPXR and PPARγ and potential effects of alternative PFASs on these targets.

In the nineth chapter of this dissertation, Protein Kinase C (PKC), a family of serine/threonine kinases involve in controlling various signaling pathways that regulate cell proliferation, survival, apoptosis, migration, invasion, differentiation, angiogenesis, and drug resistance is studied. PKC is known to be regulated by Ca^{2+} ions. By modelling PKC within various ion and Ca^{2+} concentrations successive binding of Ca^{2+} ions are displayed, and conformational changes related to this process is explained.

And, finally, the last chapter of this dissertation ends with concluding remarks and possible future directions stemming from the work described herein.

REFERENCES

REFERENCES

(1) Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. Mature HIV-1 Capsid Structure by Cryo-Electron Microscopy and All-Atom Molecular Dynamics. *Nature* **2013**, *497* (7451), 643–646. https://doi.org/10.1038/nature12162.

(2) Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular Dynamics Simulations of Large Macromolecular Complexes. *Curr. Opin. Struct. Biol.* **2015**, *31*, 64–74. https://doi.org/10.1016/j.sbi.2015.03.007.

(3) Mesecar, A. D.; Stoddard, B. L.; Koshland, D. E. Orbital Steering in the Catalytic Power of Enzymes: Small Structural Changes with Large Catalytic Consequences. *Sci. (80-.).* **1997**, *277* (5323), 202–206. https://doi.org/10.1126/science.277.5323.202.

(4) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **2009**, *19* (2), 120–127. https://doi.org/10.1016/j.sbi.2009.03.004.

(5) Eken, Y.; Patel, P.; Díaz, T.; Jones, M. R.; Wilson, A. K. SAMPL6 Host–Guest Challenge: Binding Free Energies via a Multistep Approach. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1097–1115. https://doi.org/10.1007/s10822-018-0159-1.

(6) Eken, Y.; Almeida, N. M. S.; Wang, C.; Wilson, A. K. SAMPL7: Host–Guest Binding Prediction by Molecular Dynamics and Quantum Mechanics. *J. Comput. Aided. Mol. Des.* **2021**, *35* (1), 63–77. https://doi.org/10.1007/s10822-020-00357-3.

(7) Yang, W.; Ramadan, S.; Orwenyo, J.; Kakeshpour, T.; Diaz, T.; Eken, Y.; Sanda, M.; Jackson, J. E.; Wilson, A. K.; Huang, X. Chemoenzymatic Synthesis of Glycopeptides Bearing Rare N-Glycan Sequences with or without Bisecting GlcNAc. *Chem. Sci.* **2018**, *9* (43), 8194–8206. https://doi.org/10.1039/c8sc02457j.

(8) Yang, W.; Eken, Y.; Zhang, J.; Cole, L. E.; Ramadan, S.; Xu, Y.; Zhang, Z.; Liu, J.; Wilson, A. K.; Huang, X. Chemical Synthesis of Human Syndecan-4 Glycopeptide Bearing O-, N-Sulfation and Multiple Aspartic Acids for Probing Impacts of the Glycan Chain and the Core Peptide on Biological Functions. *Chem. Sci.* **2020**, *11* (25), 6393–6404. https://doi.org/10.1039/d0sc01140a.

(9) Lai, T. T.; Eken, Y.; Wilson, A. K. Binding of Per- and Polyfluoroalkyl Substances to the Human Pregnane X Receptor. *Environ. Sci. Technol.* **2020**, *54* (24), 15986–15995. https://doi.org/10.1021/acs.est.0c04651.

(10) Nuno, A.; Eken, Y.; Wilson, A. K. Binding of Per- and Polyfluoro-Alkyl Substances (PFASs) to Peroxisome Proliferator-Activated Receptor Gamma (PPARγ). *ACS Omega* **2021**, *6 (23)*, 15103-15114. https://doi.org/10.1021/acsomega.1c01304.

CHAPTER TWO

Theory and Methods in Molecular Modeling

2.1 Theoretical Background

Computational chemistry has broad reach, from the description of detailed electronic manifolds of the smallest of molecules, to the modeling of biological systems. There are a number of branches of computational chemistry, and two primary areas that are used in the computational study of biological systems are molecular dynamics (MD) and quantum mechanics (QM). In QM, the electronic structure of the systems is solved using Schrödinger equation and electron wavefunctions to insight events like bond breakage or forming. However, as the system size increases, the computational cost of the calculations with respect to memory and time becomes impractical. Due to the enormous size of biological macromolecules, QM can only be applied to a limited number of conformers or treat only part of large systems.[1] Therefore, many studies of biological macromolecules use classical molecular dynamics (MD) to investigate their systems, and is used in this work. With this focus, the current chapter addresses the theory behind MD, statistics, force field parameters, bioinformatics and binding free energy calculations.

2.2 Molecular Dynamics

Molecular dynamics (MD) is commonly used to simulate macromolecular structures and dynamics. Biological and chemical systems at the atomistic level on timescales ranging from femtoseconds to milliseconds can be studied.[2] In classical MD, Newtonian mechanics are used to study the motions and interactions of atoms and molecules within the system.

$$F_i(t) = m_i \mathrm{a}_i(t) = m_i \ddot{\boldsymbol{r}}_i(t) = -\frac{\partial \mathrm{V}(\mathbf{r}(t))}{\partial \boldsymbol{r}_i(t)} \qquad (2.1)$$

Here, F_i (t) represents the total force on particle i at time t, $\ddot{\boldsymbol{r}}_i(t)$ as the second derivative of the position represents the corresponding acceleration a_i(t), m_i the particle's mass, r_i(t) the

position vector of the particle *i*. In equation (2.1) V(r(t)) describes the potential energy of an entire system of N-particles. The initial velocity of the atoms in the system is assigned randomly according to the Boltzmann distribution function and the accelerations are calculated by the forces acting on each atom. Versatile second order algorithms are developed to solve Newton's equations of motion such as Verlet,[3] velocity Verlet[4] and the "leapfrog"[5] algorithms. Among those the "leapfrog" algorithm is particularly suitable for solving Newton's equations because of its simplicity and stability. Additionally, the "leapfrog" algorithm preserves the time reversibility.

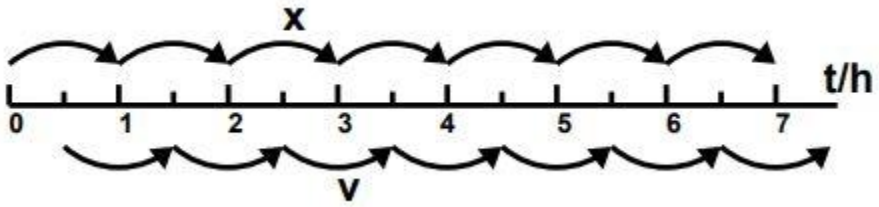


Figure 2. 1 Leapfrog algorithm steps. In this algorithm velocities are calculated on the midpoints of Δt, whereas positions are calculated explicitly at each Δt.

$$\dot{\mathbf{r}}\left(t+\frac{\Delta t}{2}\right)=\dot{\mathbf{r}}\left(t-\frac{\Delta t}{2}\right)+\ddot{\mathbf{r}}\left(t-\frac{\Delta t}{2}\right)\Delta t \qquad (2.2)$$

$$\mathbf{r}(t+\Delta t)=\mathbf{r}(t)+\dot{\mathbf{r}}\left(t+\frac{\Delta t}{2}\right)\Delta t+\ddot{\mathbf{r}}\left(t-\frac{\Delta t}{2}\right)\Delta t \qquad (2.3)$$

Equation (2.2) defines the velocity calculations in the "leapfrog" algorithm. In this equation, $\dot{\mathbf{r}}\left(t-\frac{\Delta t}{2}\right)$ and $\dot{\mathbf{r}}\left(t+\frac{\Delta t}{2}\right)$ represents the velocities after and before time step t of the propagation, $\ddot{\mathbf{r}}\left(t-\frac{\Delta t}{2}\right)$ represents the acceleration at time $(t-\frac{\Delta t}{2})$ and Δt is the selected time step. In the next time step, velocities are calculated. Then, velocities and accelerations are used to solve equation (2.3) and find the positions of each particle. In this equation, **r**(t) and **r**(t +Δt) stand for the positions before and after time-step, respectively. In the "leapfrog" algorithm, velocities are not calculated at the same time as the positions as shown in Figure 2.1. Instead, velocities are

calculated at the midpoints of Δt using accelerations, determined by the force as shown in equation (2.3) and those velocities are used to find the positions at t + Δt. In other words, velocities at each Δt are not explicitly calculated in this method but velocities at each Δt can be found by averaging the velocities at $(t - \frac{\Delta t}{2})$ and $(t + \frac{\Delta t}{2})$. Numerical integration of equations (2.2) and (2.3) generates the simulation trajectories in which the position of each particle in the system is evolving in time.

In MD, the system is represented by the "Ball and Stick Model" where the nuclei are shown as balls and the bonds between them are represented with springs. Forces are calculated using classical dynamics, which is only applicable to nuclei. Additionally, MD calculations utilize force fields (FF), where representative models of empirical potential energy function are used to estimate interactions and the total energy. The FFs utilized in MD and solving only for nuclei decreases the required computation time significantly when compared with the time that would be required with QM. Even though electrons are not represented in MD, their effect is included in the FF because FFs are parameterized from quantum mechanical calculations and experimental data such as crystal structures, vibrational frequencies, and molecular geometries. The typical mathematical expression of force field can be described as follows;[6,7]

$$V(r^N) = \sum_{bonds} k_b(l - l_0)^2 + \sum_{angles} k_a(\theta - \theta_0)^2 + \sum_{torsions} \sum_n \frac{V_n}{2}[1 + \cos n\varphi - \delta_n] + \sum_{ij}^{pairs} \varepsilon_{ij} \left[\left(\frac{r_{0,ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0,ij}}{r_{ij}} \right)^{6} \right] + \sum_{ij}^{pairs} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \qquad (2.4)$$

The first three terms in the equation (2.4) are known as the bonded terms. Their values depend on the intramolecular interactions between the atoms such as distances, bending and torsions. The final two terms represent the potential energy between non-bonded atoms; this includes electrostatic and van der Waals (vdw) interactions. The most common way of treating

vdw interactions is by using a 12-6 Lennard-Jones potential and the electrostatic energy described with a Coulomb potential.

In molecular systems each atom is bonded to only a few other atoms; bonded terms can be calculated entirely, but there are N^N non-bonded interactions for the N-particle system. The non-bonded terms are commonly treated with cutoff schemes to preserve the cost effectiveness, which can be done in number of ways. One way is using a truncated non-bonded potentials where the contribution is set to zero when the distance between two particle is higher than a designated cutoff distance. Another way is via the switching form, where a second distance is set to gradually alter the potential and smoothly decrease it to zero on the cut off distance. Yet, in large systems these approximations may lead to poor results because they often lead to artificial minima and potential energy of the particles can change suddenly. The particle mesh Ewald (PME) method is an efficient alternative and is described in the following manner:

$$E_{total}(t) = \sum_{ij} \varphi(\mathbf{r}_j - \mathbf{r}_i) \tag{2.5}$$

$$\varphi(\mathbf{r}) \stackrel{\text{def}}{=} \varphi_{sr}(\boldsymbol{r}) + \varphi_{lr}(\boldsymbol{r}) \tag{2.6}$$

$$E_{total} = E_{sr} + E_{lr} = \sum_{ij} \varphi_{sr}\left(\mathbf{r}_j - \mathbf{r}_i\right) + \sum_{k} \widetilde{\boldsymbol{\Phi}}_{lr}(\mathbf{k})|\tilde{p}(\mathbf{k})|^2 \tag{2.7}$$

In the PME method the total energy is calculated by the sum of interactions between all atom pairs as shown in equation (2.5). However, PME methods divide these interactions into long range (E_{lr}) and short range (E_{sr}) interactions and treat them differently (2.6). The last equation of PME (2.9) shows that short range interactions $\varphi_{sr}(\boldsymbol{r})$ are treated in the direct space sum. Whereas, long range interactions $\varphi_{lr}(\boldsymbol{r})$ are Fourier transformed and included in the frequency space sum that leads to the $\sum_{k} \widetilde{\boldsymbol{\Phi}}_{lr}(\mathbf{k})|\tilde{p}(\mathbf{k})|^2$ term in equation (2.7).[8] vdw interactions as calculated with a 12-6 Lennard-Jones potential in the equation (2.4) vanish quickly as the distance between pairs increases, as a result vdw interactions are commonly calculated till a

cutoff distance (mostly around 10 Å). However, electrostatic interactions are known as long range interactions, and do not vanish quickly with respect to distance. The PME method is commonly used for calculating long range interactions and it is particularly useful for lattice structures with periodic boundary conditions (PBC). In PBC, the system is defined as a unit cell and replicated infinitely many times in 3D space. In periodic calculations, an atom that crosses one boundary enters again from the other side of the cell to preserve the unit cell charge and atom numbers.

2.3 Binding Free Energy Calculations

Free energy calculations are useful in multiple areas of computational biology such as drug design, determination of ligand binding energies, and protein structure determination.[5] There are several methods available such as Free Energy Perturbation (FEP),[9] Replica exchange Free Energy Perturbation(REMD)[10] and Thermodynamic Integration (TI).[11] However, these methods are computationally demanding with their cost swiftly increasing with respect to system size. Another effective route via end-state free energy methods, which have reduced computation cost compared to FEP, REMD and TI.[12] Molecular mechanics combined with Poisson–Boltzmann or generalized Born surface area solvation (MMPBSA/MMGBSA) approaches are arguably the most popular end state free energy methods, and are frequently used to determine binding free energies in non-covalently bound receptor-ligand complexes.[12,13] MMPBSA/MMGBSA approaches are also commonly used during our studies to predict the binding energies of protein-ligand and host-guest systems.[14–18] These binding free energies are calculated by subtracting the free energies of the unbound receptor and ligand from the bound complex as shown below where solvation free energies are approximated through implicit solvation models;

$$\Delta G_{Binding,Solvated} = \Delta G_{Complex,Solvated} - \left[\Delta G_{Receptor,Solvated} + \Delta G_{Ligand,Solvated}\right] \quad (2.8)$$

More details about MMPBSA/MMGBSA approaches and on how various parameter choices affect the prediction accuracy are included and discussed in Chapters **3** and **4** where methodologies are evaluated during SAMPL (Statistical Assessment of Modeling Proteins and Ligands) challenges.

2.4 Sequence Alignment

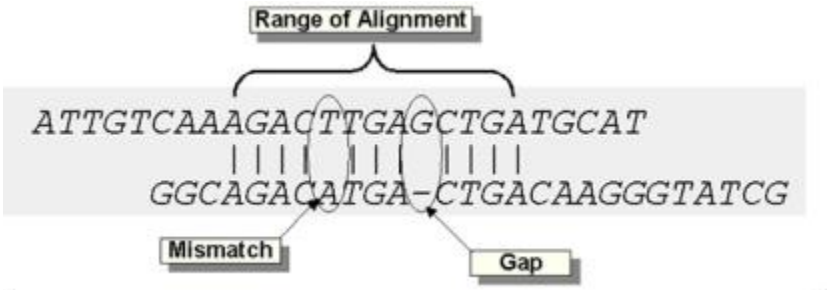


Figure 2. 2 Alignment of two short protein sequences.

Sequence alignment of proteins is among the most useful computational-based approaches applied in protein studies. In Figure 2.2 two short protein sequences are aligned to one another. The structure of a protein comes from the amino acid sequence and these alignments can be used to extract important insights about genes, or the protein's function. Additionally, these methods can be used to compare and evaluate similarities between multiple proteins and detect the domains which are less conserved within the family of proteins. Sequence alignment approaches are designed based on probability and statistics. Conserved regions of the protein families with similar structure and same function are used to calculate the frequency of changing amino acid a to amino acid b.[19] After, the frequencies are determined, these values are converted into the scoring matrix where the scores for each pair are summed to find the total score. A high score indicates there is a considerable similarity between the sequences that are compared, and that

these proteins can be functionally or evolutionarily related. A low score indicates that the sequences or proteins are different.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	-1	5																		
Asn	-2	0	6																	
Asp	-2	-2	1	6																
Cys	0	-3	-3	-3	9															
Gln	-1	1	0	0	-3	5														
Glu	-1	0	0	2	-4	2	5													
Gly	0	-2	0	-1	-3	-2	-2	6												
His	-2	0	1	-1	-3	0	0	-2	8											
Ile	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 2. 3 Blossom 62 matrix is a commonly used substitution matrix. In this matrix arginine to arginine substitutions scores +5 and arginine to lysine substitution scores +2, indicating substitution of these two positively charged amino acids frequently occurs within the functionally related proteins. Whereas arginine to aspartic acid substitution scores -2, meaning this substitution is not frequent among functionally related proteins.

$$S = \sum(identities, mismatches) - \sum(gap\ penalties) \qquad (2.9)$$

$$Score = Max(S) \qquad (2.10)$$

The score of an alignment S is calculated as the sum of substitutions determined according to the substitution matrix used (Figure 2.3) minus the sum of the gap penalty as shown in equation (2.9). Gap scores typically have different values for opening and extension and the highest score is considered as the result. Using these alignment scores, the similarity of the protein sequences

can be explored. However, the scoring does not include information about the structural arrangements in the protein.

REFERENCES

REFERENCES

(1) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C.; Brook, S.; Brook, S.; Brook, S. Comparison of Multiple AMBER Force Fields and Development of Improved Protien Backbone Parameters. *Proteins* **2006**, *65* (3), 712–725. https://doi.org/10.1002/prot.21123.Comparison.

(2) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3* (2), 198–210. https://doi.org/10.1002/wcms.1121.

(3) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

(4) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **1982**, *76* (1), 637–649. https://doi.org/10.1063/1.442716.

(5) Ganesan, A.; Coote, M. L.; Barakat, K. Molecular Dynamics-Driven Drug Discovery: Leaping Forward with Confidence. *Drug Discov. Today* **2017**, *22* (2), 249–269. https://doi.org/10.1016/j.drudis.2016.11.001.

(6) Hansson, T.; Oostenbrink, C.; Van Gunsteren, W. F. Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2002**, *12* (2), 190–196. https://doi.org/10.1016/S0959-440X(02)00308-1.

(7) Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nat. Struct. Biol.* **2002**, *9* (9), 646–652. https://doi.org/10.1038/nsb0902-646.

(8) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(9) Zwanzig, R. W. High-temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426. https://doi.org/10.1063/1.1740409.

(10) Meng, F.; Liu, L.; Chin, P. C.; D'Mello, S. R. Akt Is a Downstream Target of NF-Kappa B. *J. Biol. Chem.* **2002**, *277* (33), 29674–29680. https://doi.org/10.1074/jbc.M112464200.

(11) Kubo, R.; Ichimura, H.; Usui, T.; Hashizume, N. *Statistical Mechanics*; Harper & Row, Publishers, Inc.: New York, NY, 1965.

(12) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem.*

Theory Comput. **2012**, *8* (9), 3314–3321. https://doi.org/10.1021/ct300418h.

(13) Homeyer, N.; Gohlke, H. Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method. *Mol. Inform.* **2012**, *31* (2), 114–122. https://doi.org/10.1002/minf.201100135.

(14) Yang, W.; Ramadan, S.; Orwenyo, J.; Kakeshpour, T.; Diaz, T.; Eken, Y.; Sanda, M.; Jackson, J. E.; Wilson, A. K.; Huang, X. Chemoenzymatic Synthesis of Glycopeptides Bearing Rare N-Glycan Sequences with or without Bisecting GlcNAc. *Chem. Sci.* **2018**, 8194–8206. https://doi.org/10.1039/c8sc02457j.

(15) Eken, Y.; Patel, P.; Díaz, T.; Jones, M. R.; Wilson, A. K. SAMPL6 Host-Guest Challenge: Binding Free Energies Via a Multistep Approach. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1097–1115.

(16) Yang, W.; Eken, Y.; Zhang, J.; Cole, L. E.; Ramadan, S.; Xu, Y.; Zhang, Z.; Liu, J.; Wilson, A. K.; Huang, X. Chemical Synthesis of Human Syndecan-4 Glycopeptide Bearing O-, N-Sulfation and Multiple Aspartic Acids for Probing Impacts of the Glycan Chain and the Core Peptide on Biological Functions. *Chem. Sci.* **2020**, *11* (25), 6393–6404. https://doi.org/10.1039/d0sc01140a.

(17) Lai, T. T.; Eken, Y.; Wilson, A. K. Binding of Per- and Polyfluoroalkyl Substances to the Human Pregnane X Receptor. *Environ. Sci. Technol.* **2020**, *54* (24), 15986–15995. https://doi.org/10.1021/acs.est.0c04651.

(18) Eken, Y.; Almeida, N. M. S.; Wang, C.; Wilson, A. K. SAMPL7: Host–Guest Binding Prediction by Molecular Dynamics and Quantum Mechanics. *J. Comput. Aided. Mol. Des.* **2021**, *35* (1), 63–77. https://doi.org/10.1007/s10822-020-00357-3.

(19) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci.* **1992**, *89* (22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915.

CHAPTER THREE

SAMPL6 Host–Guest Challenge: Binding Free Energies via a Multistep Approach

About this chapter: This chapter is reprinted from Eken, Y.; Patel, P.; Díaz, T.; Jones, M. R.; Wilson, A. K. SAMPL6 Host – Guest Challenge: Binding Free Energies via a Multistep Approach. *J. Comput. Aided. Mol. Des*. **2018**, *32 (10),* 1097–1115. with permission of the Springer Nature. The docking, molecular dynamics simulations and MMPBSA calculations mentioned in this chapter are performed by Yiğitcan Eken, clustering analysis is done by Thomas Diaz and quantum mechanical calculations are done by co-authors Prajay Patel, Michael Jones and Thomas Diaz.

3.1 Introduction

Tremendous advances in technological capabilities have enabled computational approaches to be applied to discern a broad range of physical, chemical, and biological phenomena across scales in molecular science.[1–6] With emphasis on molecular design, computational approaches have found great utility towards innovation in drug discovery. Considering the time and cost of the drug pipeline, from the discovery process to market, in silico biophysical methods serve an important role in expediting and reducing the cost of the discovery process, facilitating the identification, optimization, and refinement of potential drug candidates and providing comprehensive insight into the mechanism of action and structure–property relationships at the atomic level that are ultimately critical to a drug's efficacy.[7–12]

In computational strategies towards structure-based design, an important step is the prediction of probable conformations of a ligand bound to the host. To identify better possible candidate binding modes, they can be ranked via scoring functions and further evaluated via molecular simulation and free energy calculations. From free energy calculations, selectivity profiles may be constructed not only to determine binding affinities but also to provide understanding into how the ligand recognizes its host.

Because of the complexity that occurs in ligand-bound protein systems, relatively smaller representative models such as polymer-based host–guest systems are used to assess free energy methods.[13–18] Although host structures selected to represent proteins are typically much smaller than proteins, they are large enough to possess a cavity or binding pocket that allows non-covalent binding of multiple guest molecules. The advantage of using host–guest systems for assessing free energy methods is that they tend to be more rigid and symmetric than proteins, which results in fewer conformations that need to be sampled.[19–23] Even in the representation of proteins by more simplistic models, modeling binding free energies for these smaller models is challenging since no clear "best" computational chemistry approach has been identified; efforts are needed to better resolve strategies towards predictions of binding free energies. Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) blind challenges provide a unique platform to validate available methods and stimulate the development of new methods for quantitative predictions.[13,16,18,24–26] In these challenges, binding affinities and other physicochemical properties are predicted, using computational models without the benefit of insight from experiment; they are then later compared to unpublished experimental measurements that allow the comparison of different computational prediction methods.

For the prediction of free energies, there are many methods available such as free energy perturbation[27], replica exchange free energy perturbation[28], and thermodynamic integration[29]. However, these methods are computationally demanding (memory, disk space, CPU time) as they converge poorly for large systems and require dividing the model into multiple intermediate steps or multiple configurations for accurate predictions.[30–32] In contrast, end state free energy methods are path-independent and do not require sampling of multiple configurations. These methods offer a simpler, computationally less costly approach to predict the binding free

energy.[33–38] Moreover, implicit solvation models can be used to further reduce the computational cost while predicting binding free energies similar to these more sophisticated and computationally demanding methods.

While classical molecular dynamics (MD) methods are commonly used to investigate host–guest interactions, molecular mechanics (MM) force fields result in a limited treatment of effects resulting from polarization, charge transfer, and many body effects which can impact the description of properties such as binding free energies.[9,39–43] To better account for these effects, quantum mechanical (QM) approaches, which are more costly, are commonly used in drug discovery research[9,44], and have been used in previous SAMPL competitions.[17,45–47] For example, in the SAMPL5 competition for host–guest binding, Caldarau et al.[45] used density functional theory (DFT) with an added dispersion correction (DFT-D3) and the wavefunction-based domain-based local pair-natural orbital coupled cluster (DLPNO-CCSD(T)) method to predict the binding energies for octa-acid (OA) host–guest systems. In this approach, they used TPSS-D3/def2-SVP optimized structures and host structures are constrained during MD simulations to reduce the flexibility of the host and limit the structural distortions resulting from the repulsion between the negative charge of the ligands and the large negative charge of the OA hosts. This approach yielded binding energies approximately 12.0 kcal mol^{-1} greater than the experimental binding affinities, with a low correlation coefficient ($r^2 \approx 0$), and a statistically insignificant Kendall's rank correlation coefficient ($\tau \leq 0.20$) for all attempts for the host–guest systems in the SAMPL5 blind challenge due to incorrect representative structures, not sampling enough conformational binding positions for ligands, and thermochemical corrections that yielded up to a 7.2 kcal mol^{-1} difference depending on the method of choice. This performance demonstrates

the limited sampling capabilities of current QM methods compared to MD methods, obtained representative structures, as well as thermodynamic and solvation corrections.

Contrary to this, in the SAMPL4 competition for host–guest binding, Mikulskis et al.[47] were successful with both MM- and QM-based approaches for OA hosts with mean absolute deviations (MADs) less than 2.0 kcal mol^{-1}. Their MM approach, which utilized free energy perturbation (FEP) calculations, yielded MADs of approximately 1.0 kcal mol^{-1} while their QM approaches with DFTD3 optimized structures yielded MADs of approximately 1.0–2.0 kcal mol^{-1} depending on the implementation of a solvent in the calculations, i.e. no solvent, implicit solvent, or a combined implicit-explicit solvent. However, the combination of FEP and DFT-D3 did not yield favorable results due to the large difference between the MM and DFT potential energy functions. Sure et al.[46] provided another successful attempt at using DFT-D3 for the SAMPL4 competition for host–guest binding of a macrocyclic cucurbit[7]uril host by optimizing the geometry at the TPSS-D3/def2-TZVP level of theory after pre-optimizing possible binding scenarios with the HF-3c semiempirical method. These optimizations were followed by single point calculations using PW6B95-D3/def2-QZVP with the g- and f-functions for non-hydrogen and hydrogen atoms removed, respectively, with the COSMO-RS implicit solvent model, which yielded an MAD of 2.0±0.5 kcal mol^{-1}. These two studies highlight that for the SAMPL4 competition, host–guest structure optimization and higher-level MM-based approaches like FEP can be vital in characterizing correct binding interactions at the QM level.

In this work, efforts in MD and QM methods are combined to predict binding affinities for fourteen ligands to a macrocyclic cucurbit[8]uril host[19,21,22,48] and eight ligands to two variants of the OA deep-cavity cavitands.[20,23] Using MD simulations to obtain representative structures, MM- and QM-based methods are utilized to predict binding free energies. Within the QM

methods, the use of a resolution-of-the-identity (RI) approximation designed for larger molecules[49], Grimme's D3 atom-pairwise dispersion corrections with Becke-Johnson damping[50], and truncated correlation consistent basis sets for the hydrogen atoms[51] are evaluated to probe how different electronic structure approaches that reduce the computational cost contribute to predicting binding affinities. Insights into what strategies are more favorable for host guest-binding will help to build a framework for predicting host–guest binding affinities using QM approaches.

3.2 Methods

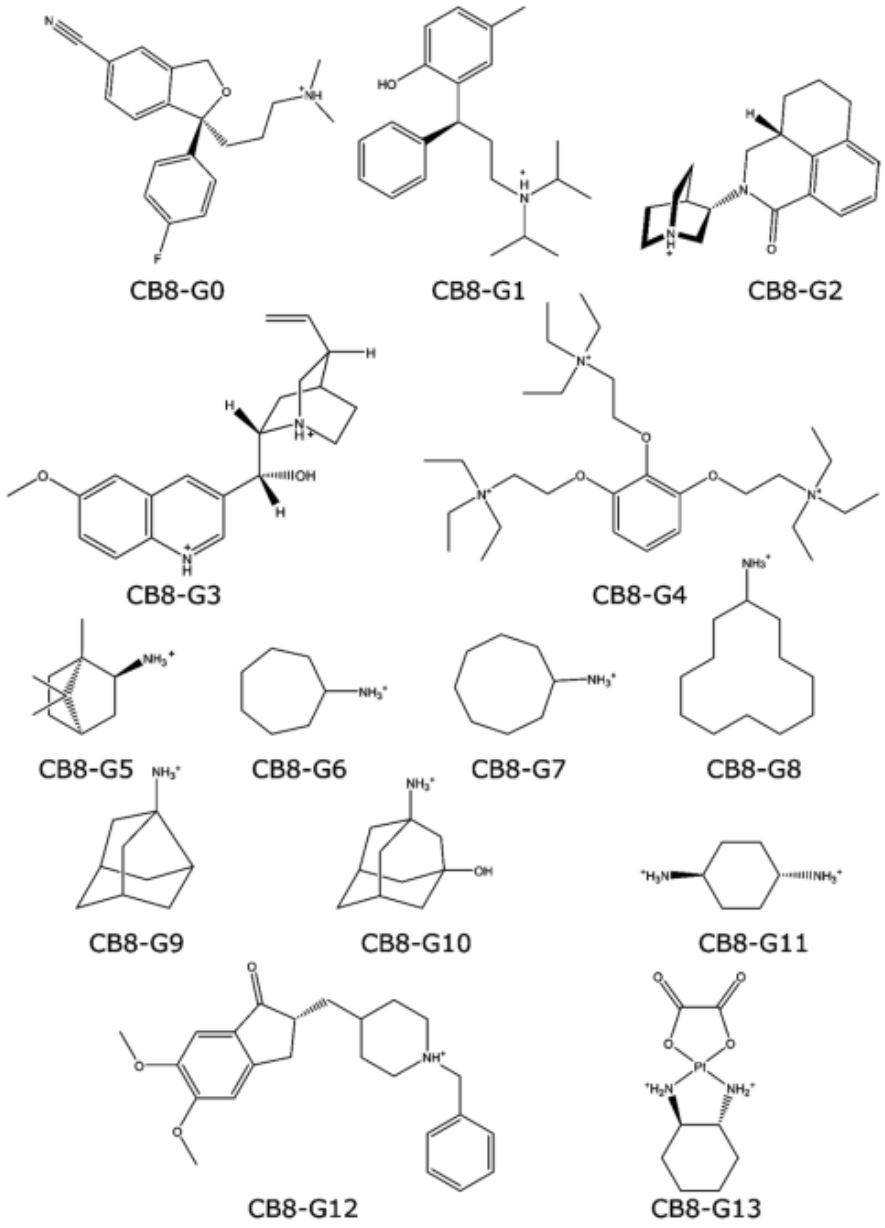


Figure 3. 1 The guest molecules for the cucurbit[8]uril (CB8).

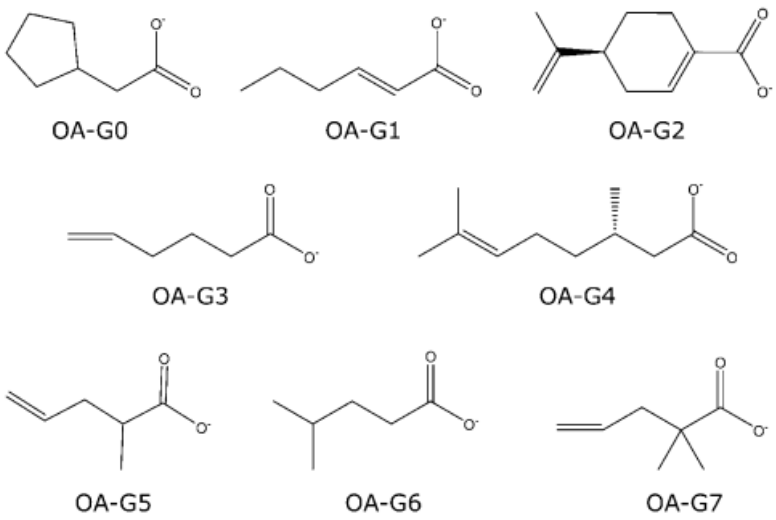


Figure 3. 2 The guest molecules for the octa-acid (OA) and tetra methyl octa-acid (TEMOA) hosts.

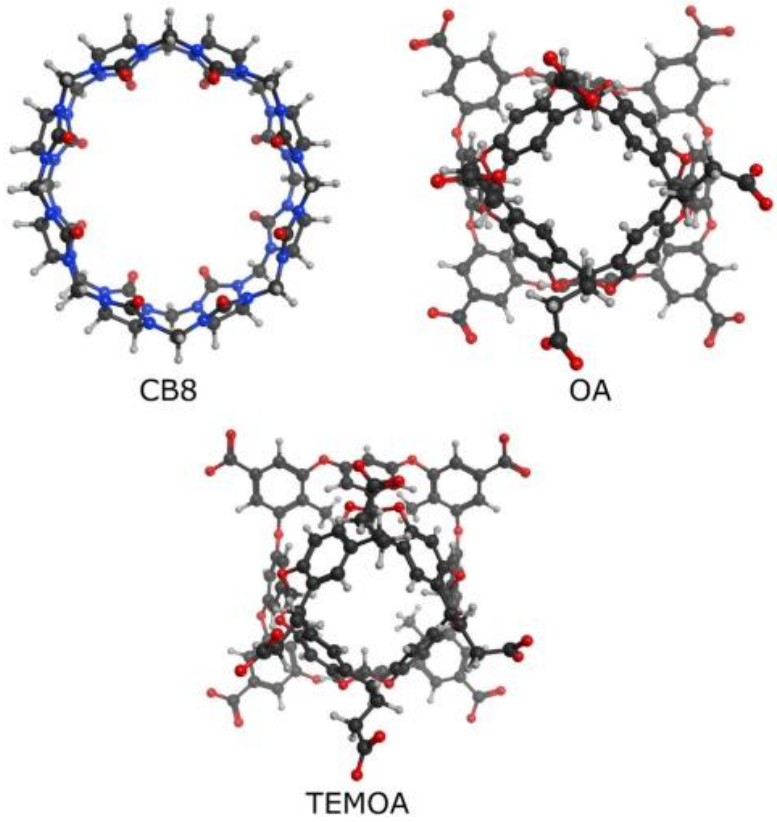


Figure 3. 3 The host molecules: cucurbit[8]uril (CB8), octa-acid (OA), and tetramethyl octa-acid (TEMOA).

3.2.1 System Preparation

The initial structures for the guest molecules, shown in Figures. 3.1 and 3.2, and the three host molecules, shown in Figure 3.3, cucurbit[8]uril (CB8), octa-acid (OA), and tetramethyl octaacid (TEMOA), that were issued with the SAMPL6 challenge dataset were used to generate the host–guest systems. The CB8 molecule has no formal charge whereas the octaacids (OA/TEMOA) have eight deprotonated carboxylic acid groups and thus a formal charge of −8. Even though OA and TEMOA are water-soluble structurally similar deep-cavity cavitands, the TEMOA host has four methyl groups in place of four hydrogen atoms present in the OA host located on the upper rim of the cavitand that enclose the hydrophobic binding pocket.

Initial binding poses of guest molecules binding to the host were generated using the docking feature implemented in Molecular Operating Environment (MOE) v2016.08[52]. The London ΔG scoring function[53] was used to estimate ligand placement in the pocket. The top 100 poses given by the London ΔG scoring function for each host–guest complex were refined to a list of ten poses by rescoring the flexible receptor and ligand conformation using the GBVI/WSA ΔG scoring algorithm.[52,53] Among these ten poses, those with minor structural differences were discarded from the list of ten poses and the chemically relevant poses with the highest GBVI/WSA ΔG scores were selected for further investigation. The chosen host–guest poses were minimized under the AMBER10: Extended Hückel Theory (EHT) potential implemented in MOE, which employs Amber ff10 and EHT bonded parameters.[54–56]

To generate force field parameters, the AM1-BCC scheme[57] was used for generating partial charges for the guest and host molecules using the Antechamber suite. As the guest molecule CB8-G13 contains a platinum atom, the Mulliken charges were calculated from a geometry optimization using B3LYP[58,59] in conjunction with the 6-31G(d) basis set[60] and the effective core

potential basis set Lan2L2DZ[61] for the platinum atom. All electronic structure calculations were performed in Gaussian 16[62] The host–guest systems were further prepared for simulation using the Leap module of AmberTools[63] under the General Amber Force Field (GAFF).[56] Each system was neutralized with counterions using parameters from Joung and Cheatham[64], and solvated in a 14.0 Å cube of TIP4P-Ew water[65] beyond the solute. To mimic the ionic strength of the experimental buffers, additional counter ions were added to create a buffer of 150 mM sodium chloride for the CB8 complexes and 60 mM sodium chloride for the OA/TEMOA complexes.

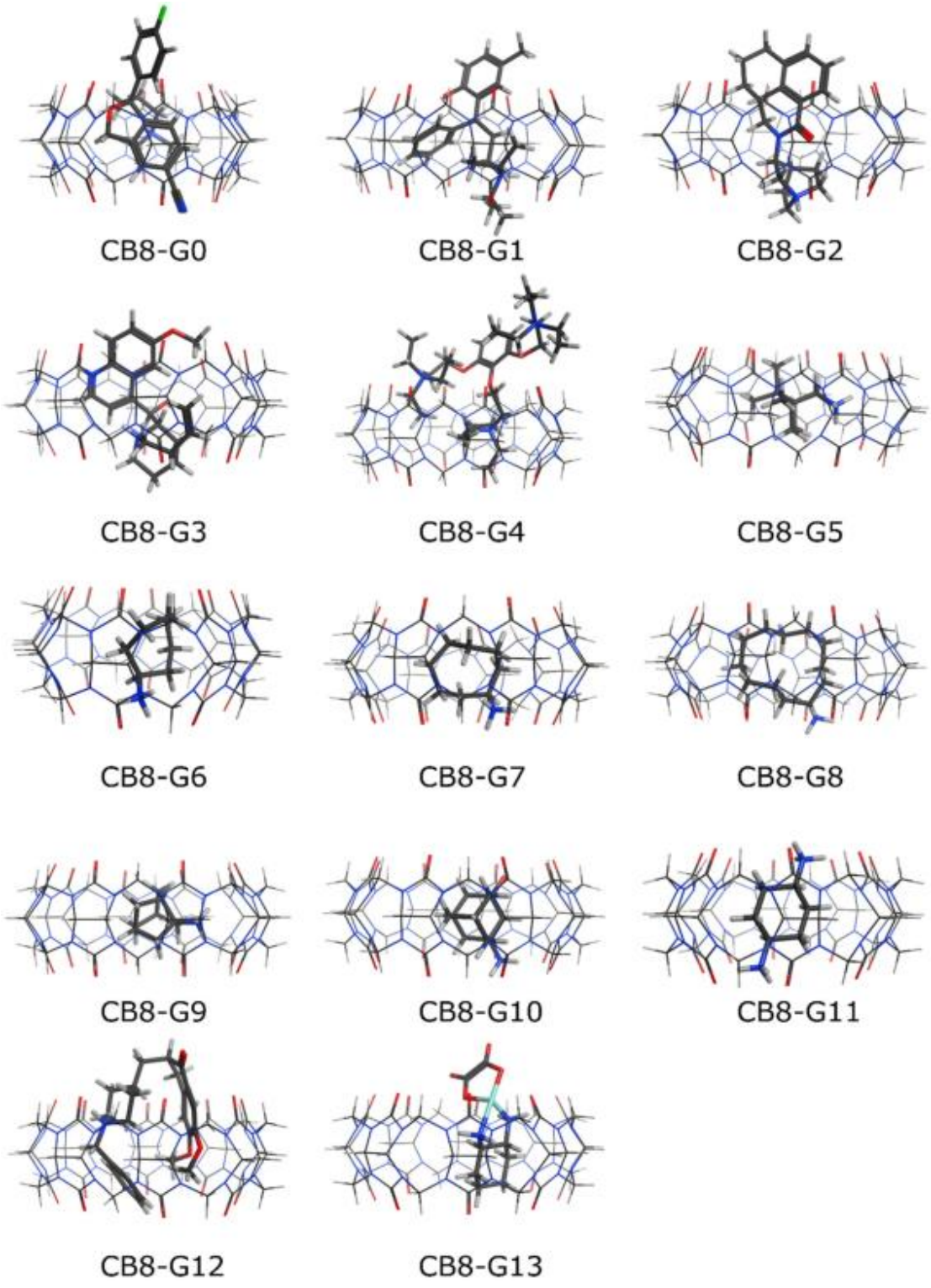


Figure 3. 4 The structures of the CB8 guest molecules inside the binding pocket. These structures are generated from the clustering analysis.

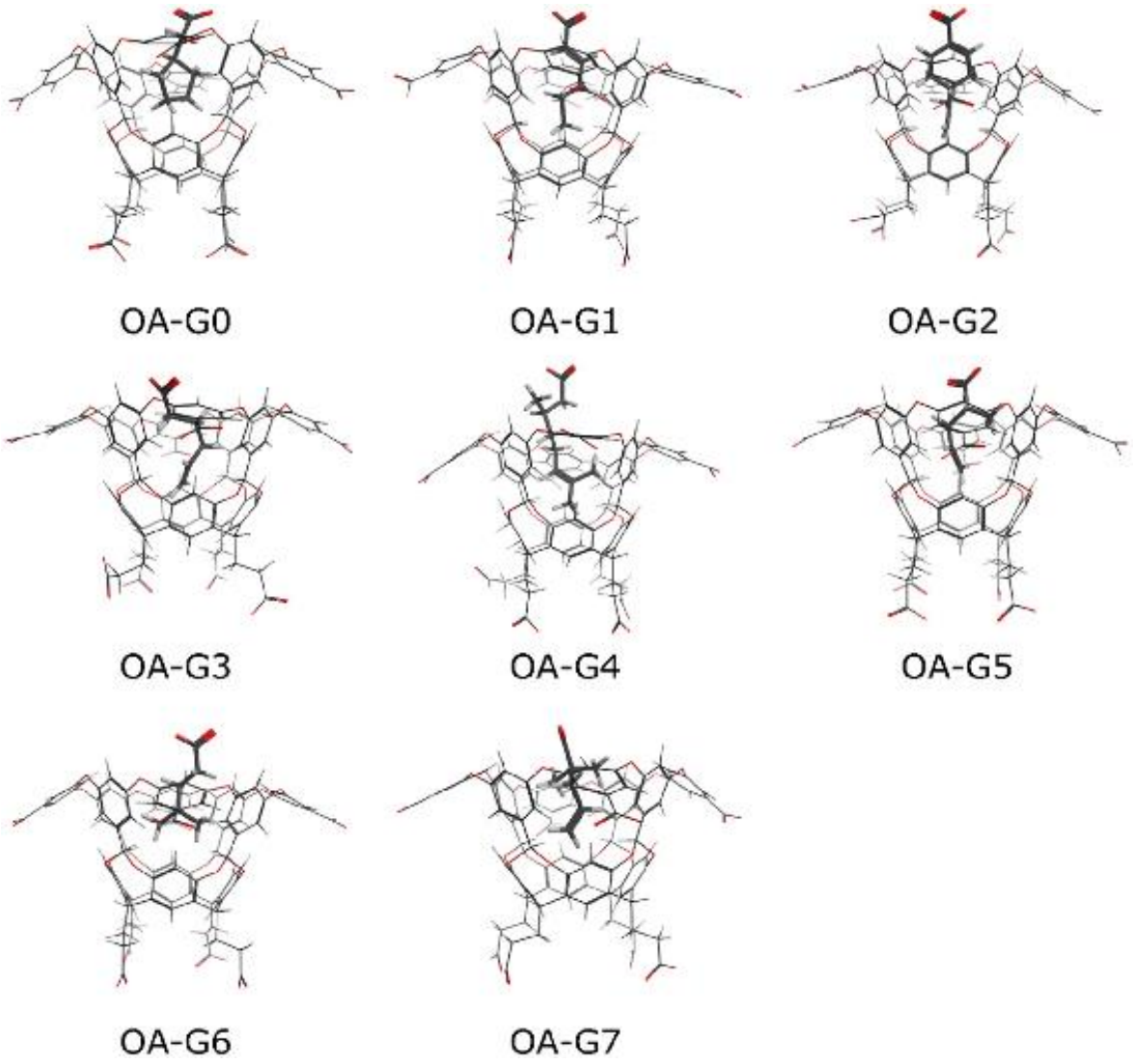


Figure 3. 5 The structures of the OA guest molecules inside the binding pocket. These structures are generated from the clustering analysis.

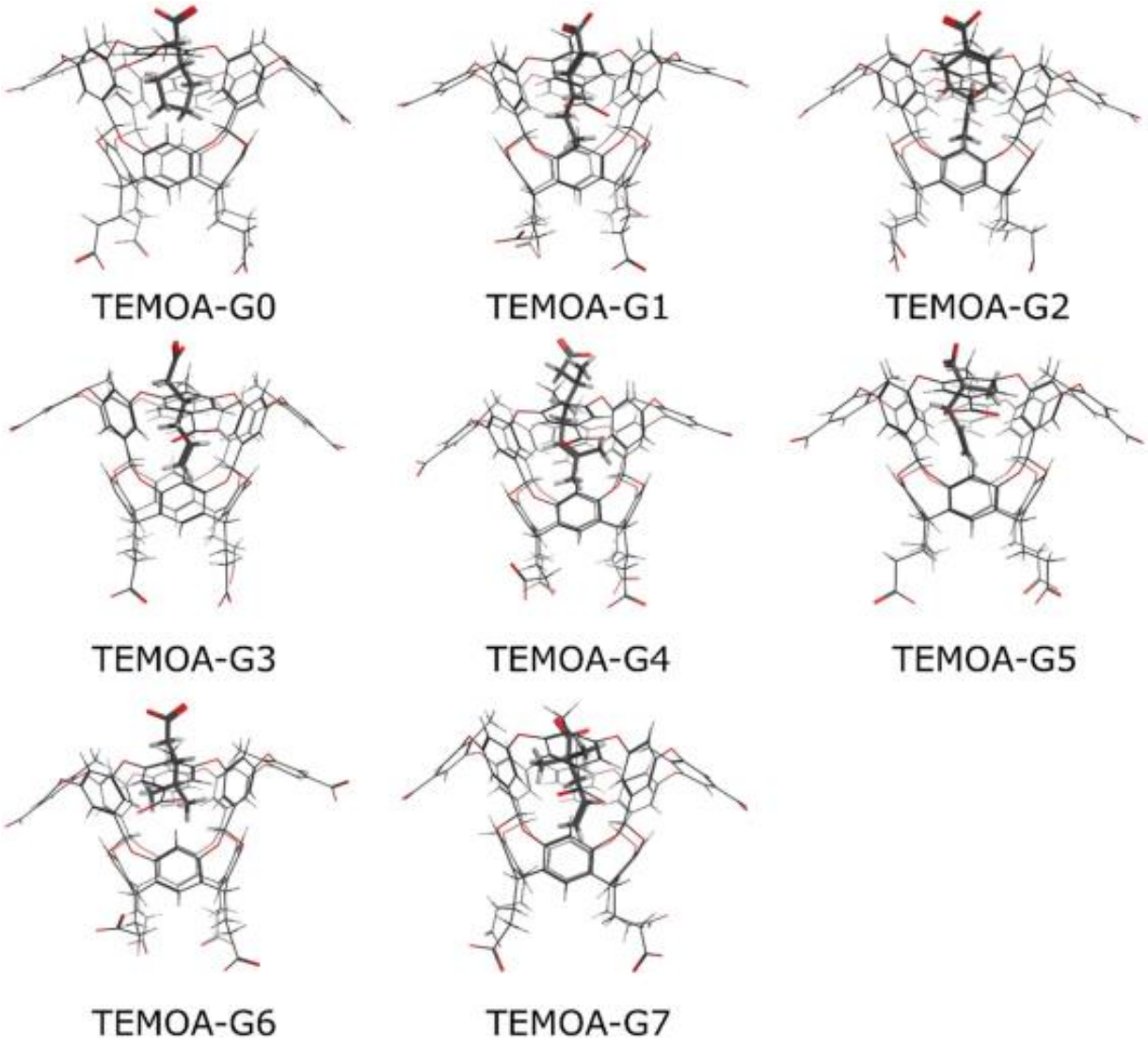


Figure 3. 6 The structures of the TEMOA guest molecules inside the binding pocket. These structures are generated from the clustering analysis.

3.2.2 Simulation Protocol

The host–guest systems were relaxed using NVT ensembles over six minimization procedures with decreasing restraints on the host of 500.0, 200.0, 20.0, 10.0, 5.0 kcal/ mol ($Å^2$), and then were heated to 300 K over 30 ps. The temperature was maintained at 300 K using Langevin dynamics and the pressure was coupled to 1 atm using isotropic position scaling. Atomistic molecular dynamics simulations were performed for 10 ns in triplicate to account for

randomized parameters that affect the MD trajectories. Nonbonded interactions were truncated with a 10.0 Å cutoff, whereas long-range electrostatics were handled with the particle-mesh Ewald (PME) method. Bonds involving hydrogen were constrained using SHAKE, and the simulation time step was set to 2 fs. All simulations were performed with AMBER16.7.[63]

The binding free energies were calculated with MMPBSA approach using the built-in PBSA-solver.[66] The internal and external dielectric constants were set to 1.0 and 80.0, respectively. The solvent accessible surface area (SASA) was determined with the default LCPO method using the modified Bondi atomic radii. Calculations for solute entropic contributions were not considered. For each system, the binding free energy was determined using the final 100 frames from the simulation.

Clusters were formed using the density-based spatial clustering of applications with noise (DBSCAN) algorithm based on two parameters, which are epsilon (Eps) and the minimum number of points in an Eps-neighborhood (MinPts). MinPts was set to 4 and the Eps value for DBSCAN was determined from the threshold point of a sorted 4-dist graph.[67] The cluster conformation representing the greatest number of frames from the MD simulations was used for further analyses. Additional QSAR (quantitative structure–activity relationship) calculations were performed on each guest molecule to determine the van der Waals volume each molecule occupies by using the connection table approximation descriptor in MOE (Tables 3.9 and 3.10).

3.2.3 Quantum Mechanical Methods

The individual structures generated from the clustering of MD trajectories, shown in Figures. 3.4, 3.5 and 3.6, for each host–guest complex were used for all quantum chemical calculations. The host and guest molecules were analyzed with the same geometry as from the complex. The thermal corrections for all molecules were calculated at the HF/6- 31G(d) level of theory in

Gaussian 16 and the vibrational contributions were scaled by 0.8953.[68] Single point energies were obtained using ORCA 4.0[69] with the B3PW91 density functional[58,70,71] since B3PW91 has been shown to properly treat long-range covalent interactions. In the treatment of the exact exchange in the functional, the RIJCOSX approximation[49] was used with the def2 auxiliary basis set[72] to reduce the computational cost associated with the number of atoms in the host–guest complex since the RIJCOSX approximation has been shown to be five times as efficient for molecules of similar size to the host–guest systems. To mimic the aqueous solution, the SMD implicit solvation model[73] was used with water (ε = 78.4) as the implicit solvent. Grimme's D3 dispersion correction with Becke-Johnson damping was used to investigate long-range covalent interactions as the inclusion of D3 dispersion improves intermolecular interaction energies predicted with DFT.[46,50,74,75]

The correlation consistent basis set family (cc-pVnZ)[76] was used for all single point calculations since these basis sets were developed to exhibit convergence behavior to the complete basis set (CBS) limit for wavefunction-based methods through extrapolation.[77–80] Knowing the CBS limit, which removes basis set incompleteness error, the error for the property of interest, i.e. binding free energy, only corresponds to the intrinsic error of the chosen QM method. Therefore, to extrapolate to the Kohn–Sham limit for DFT methods, analogous to the CBS limit for wavefunction-based methods, the cc-pVnZ basis sets were used (n=D, T) with the following extrapolation scheme proposed by Jensen

$$E(l_{max}) = E_{CBS} + A(l_{max} + 1)e^{-B\sqrt{\pi_s}} \qquad (3.1)$$

where $l_{\max}$ is the maximum angular momentum function in the basis set and ns is the number of s functions in the basis set.[81] The B-parameter was set to 5.5 in agreement with Jensen for use as a two-point extrapolation scheme. Due to the abundance of weak molecular interactions in

biomolecules, the calculated binding energies were counterpoise corrected before the extrapolations were performed on each host, guest, and host–guest complex.[82,83]

Additional electronic structure modeling techniques were applied to the CB8 host–guest systems to examine the impact of various approximations on the binding free energy. Targeting reduction in computational time, the correlation consistent basis sets were truncated via the removal of higher angular momentum basis functions for hydrogen atoms. This has been shown to reduce the computational time by approximately 42.9% and 57.8% when removing 1 d function from the cc-pVTZ basis set, denoted as cc-pVTZ(−1d), and 2 d functions and 1 f function from the cc-pVQZ basis set, denoted as cc-pVQZ(−1f2d), respectively, and yielded the results closest to the atomization energies generated with the full basis sets at the complete basis set limit.[51]

Binding free energies calculated with and without the use of the resolution-of-the-identity (RI) approximation were examined to gauge how the RI approximation, which leads to a reduction in CPU time, affects the accuracy. To characterize the ionic strength of the solution used in experiment, the dielectric constant for the implicit water solvent was also altered from 78.4 for pure water to 76.4 given the concentration of the sodium chloride solution used in the MD simulations and the experimentally determined relation between the concentration of an ionic solution and the dielectric constant.[84]

3.3 Results

The binding free energies submitted as part of the SAMPL6 competition are shown in Tables 3.1, 3.2 and 3.3 for CB8, OA, and TEMOA host–guest systems, respectively. For each host–guest complex, statistical measurements were used to gauge the effectiveness of each of the three methods, which are MMPBSA, RI-B3PW91-D3, and RI-B3PW91, in predicting experimental

binding free energies. These include the mean absolute error (MAE), the root mean square error (RMSE), Kendall's Tau (τ) rank correlation coefficient, which measures how well a method ranked calculated binding free energies relative to experimental binding free energies where τ values closer to one correspond to increased qualitative accuracy of the prediction, and the correlation coefficient (r^2). To demonstrate there is no correlation in ranking between the calculated binding free energies and the experimental binding free energies, τ values are compared against τ_{crit}, a cutoff value obtained through a table of critical values generated by Monte Carlo simulations of a τ distribution, which is similar to the normal Z distribution, used to reject the null hypothesis.[85,86]

3.3.1 Cucurbit[8]uril (CB8)

Table 3. 1 The binding free energies in kcal mol^{-1} for the CB8 host–guest systems.

Complex	Exp	MMPBSA	RIB3PW91-D3	RI-B3PW91
CB8-G0	−6.69±0.05	−29.4±0.3	−49.89	6.75
CB8-G1	−7.65±0.04	−31.5±0.3	−57.22	12.7
CB8-G2	−7.66±0.05	−25.6±0.3	−36.86	10.34
CB8-G3	−6.45±0.06	−34.2±0.5	−44.53	26.61
CB8-G4	−7.80±0.04	−30.8±0.3	−68.09	−11.11
CB8-G5	−8.18±0.05	−18.6±0.3	−35.92	2.39
CB8-G6	−8.34±0.05	−19.8±0.2	−31.95	1.26
CB8-G7	−10.00±0.10	−17.6±0.4	−14.90	18.09
CB8-G8	−13.50±0.04	−30.4±0.2	−50.34	4.49
CB8-G9	−8.68±0.08	−19.9±0.5	−37.07	−2.46
CB8-G10	−8.22±0.07	−19.6±0.3	−39.30	0.61
CB8-G11	−7.77±0.05	−17.5±0.4	−25.75	−1.07
CB8-G12	−9.86±0.03	−31.5±0.4	−62.05	15
CB8-G13	−7.11±0.03	−25.4±0.3	−44.04	0.17
MAE		16.7±0.3[a]	34.29	14.88
RMSE		17.8±0.8[b]	36.99	17.26
τ		−0.19	−0.14	0.05
r^2		0.00	0.00	0.00

The mean absolute error (MAE), root mean square error (RMSE), Kendall's Tau (τ), and r^2 are shown. These results correspond to those submitted for the competition.
[a]The uncertainty reported for MAE is the average of the absolute uncertainties.
[b]The uncertainty reported for RMSE is the uncertainty of the RMSE with the experimental and calculated uncertainties.

The binding free energy predictions for the CB8 host with the three methods submitted were compared to experiment (Table 3.1). The predicted values were significantly more negative than experimental binding free energies with an MAE of 16.69, 33.58, and 15.54 kcal mol^{-1} for MMPBSA, RI-B3PW91-D3, and RI-B3PW91, respectively.

When the binding affinities of the guests to CB8 are ranked from the lowest to the highest binding affinity, MMPBSA did not correctly rank any of the systems but predicted CB8-G12 to have a stronger binding affinity relative to the other complexes, which correlates to experiment

well. RI-B3PW91-D3 correctly ranked CB8-G2 as the tenth strongest bound host–guest complex and predicted that CB8-G12 was more tightly bound relative to the other CB8 host–guest systems. RI-B3PW91 correctly ranked CB8-G6, CB8-G2, CB8-G1, and CB8-G3 as fifth, tenth, eleventh, and fourteenth, respectively, while the remaining systems were ranked incorrectly. Unlike both MMPBSA and RI-B3PW91- D3, RI-B3PW91 predicted CB8-G12 to have a lower binding affinity relative to the other CB8 host–guest systems.

3.3.2 Octa acid (OA)

Table 3. 2 The binding free energies in kcal mol^{-1} for the OA host–guest systems.

Complex	Exp	MMPBSA	RI-B3PW91-D3	RI-B3PW91
OA-G0	−5.68±0.03	−12.6±0.2	−41.36	−16.57
OA-G1	−4.65±0.02	−11.6±0.1	−40.67	−17.15
OA-G2	−8.38±0.02	−18.2±0.2	6.54	44.53
OA-G3	−5.18±0.02	−10.0±0.2	−47.94	−17.62
OA-G4	−7.11±0.02	−17.0±0.2	−48.19	−13.49
OA-G5	−4.59±0.02	−9.1±0.2	−38.40	−16.42
OA-G6	−4.97±0.02	−11.3±0.2	−43.19	−23.31
OA-G7	−6.22±0.02	−11.4±0.1	−47.37	−23.78
MAE		6.8±0.2[a]	35.46[38.39]	17.86[12.85]
RMSE		7.1±0.4[b]	36.41[38.52]	22.51[13.39]
τ		0.64	0.29[0.71]	−0.21[0.05]
r^2		0.84	0.44[0.52]	0.6[0.03]

The mean absolute error (MAE), root mean square error (RMSE), Kendall's Tau (τ), and r^2 are shown. Bracketed values indicate the values after the removal of the statistical outlier (OA-G2). These results correspond to those submitted for the competition.
[a]The uncertainty reported for MAE is the average of the absolute uncertainties.
[b]The uncertainty reported for RMSE is the uncertainty of the RMSE with the experimental and calculated uncertainties.

The three sets of submitted binding free energy predictions for OA are reported in Table 3.2. All values predicted using MMPBSA were significantly more negative than experimental measurements with an MAE of 6.8±0.2 kcal mol^{-1}. When ranking the binding affinities of the

guest to the host from lowest to highest binding affinity, MMPBSA correctly placed OA-G2, OA-G4, OA-G6, OA-G5 as first, second, sixth, and eighth, respectively. The other systems were not ranked correctly; OA-G0, OA-G1, OA-G7 and OA-G3 ranked third, fourth, fifth, and seventh, respectively, whereas experimentally ranked fourth, seventh, third, and fifth, respectively.

For RI-B3PW91-D3 and RI-B3PW91, the binding free energy predicted for OA-G2 was determined as a statistical outlier with 99% confidence, visualized in Figure 3.8, using Dixon's Q-Test.[87] When the statistical outlier (OA-G2) was excluded from the RI-B3PW91-D3 set, the MAE, RMSE, Kendall's Tau (τ), and the correlation coefficient (r^2) increased from 35.46 to 38.39 kcal mol^{-1}, 36.41 to 38.52 kcal mol^{-1}, 0.29 to 0.71, and 0.44 to 0.52, respectively. When the binding free energy for OA-G2 was excluded from the set of binding free energies obtained with RI-B3PW91, the MAE, RMSE, and r^2 decreased from 17.87 to 12.85 kcal mol^{-1}, 22.51 to 13.39 kcal mol^{-1}, and 0.60 to 0.03, respectively, as shown in Table 3.2. In Figure 3.7b, the statistical outlier was removed, which improved and worsened the linear regression model comparing experiment to RI-B3PW91-D3 and RI-B3PW91, respectively. With the exclusion of OA-G2, ranking the binding affinities from lowest to highest, RI-B3PW91-D3 correctly ranked OA-G4, OA-G1, and OA-G5, as first, sixth, and seventh, respectively, while RI-B3PW91 did not correctly ranked any of the systems.

3.3.3 Tetramethyl octa acid (TEMOA)

TEMOA is structurally different from OA because of the substitution of four hydrogens around the portal to the binding pocket of OA with four methyl groups. While the same guests bound to TEMOA and OA with similar binding energies, G7 weakly binds to TEMOA relative to the other guests whereas it binds stronger to OA experimentally.

Table 3. 3 The binding free energies in kcal mol^{-1} for the TEMOA host– guest systems.

Complex	Exp	MMPBSA	RI-B3PW91-D3	RI-B3PW91
TEMOA-G0	−6.06±0.02	−12.0±0.2	−43.75	−12.80
TEMOA-G1	−5.97±0.04	−11.3±0.2	−41.98	−10.18
TEMOA-G2	−6.81±0.02	−19.3±0.2	−51.23	−7.22
TEMOA-G3	−5.60±0.04	−8.3±0.2	−43.56	−15.29
TEMOA-G4	−7.79±0.02	−19.2±0.3	−51.98	−12.39
TEMOA-G5	−4.16±0.02	−6.1±0.2	−37.04	−10.66
TEMOA-G6	−5.40±0.03	−10.4±0.2	−41.05	−16.94
TEMOA-G7	−4.13±0.02	−6.8±0.3	−45.98	−10.29
MAE		5.9±0.2[a]	38.83	6.23
RMSE		7.0±0.5[b]	39.03	7.00
τ		0.79	0.57	−0.14
r^2		0.86	0.55	0.00

The mean absolute error (MAE), root mean square error (RMSE), Kendall's Tau (τ), and r 2 are shown. These results correspond to those submitted to the competition.
[a]The uncertainty reported for MAE is the average of the absolute uncertainties.
[b]The uncertainty reported for RMSE is the uncertainty of the RMSE with the experimental and calculated uncertainties.

Binding free energy predictions using the submitted methods for the TEMOA host are reported in Table 3. Similar to OA, all three methods overestimated the binding free energies relative to experiment. RI-B3PW91-D3 overestimated the binding free energies with an MAE of 38.83 kcal mol^{-1}. Of the three methods considered, the MMPBSA method yielded better binding free energies, both quantitatively (MAE of 5.9±0.2 kcal mol^{-1}) and qualitatively ($\tau = 0.79$), than

the QM-based calculations. MMPBSA ranked TEMOA-G0 and TEMOA-G1 as the third and fourth strongest bound complexes, respectively. Additionally, MMPBSA predicted that TEMOA-G4 and TEMOA-G2 were the most tightly bound complexes while TEMOA-G7 and TEMOA-G5 were the most loosely bound complexes. RI-B3PW91-D3 correctly predicted that TEMOA-G4, TEMOA-G2, and TEMOA-G3 were the first, second, and fifth most tightly bound complexes, respectively. Like MMPBSA, RI-B3PW91-D3 predicted that TEMOA-G5 was a weakly bound host–guest complex relative to the other TEMOA host–guest systems. RI-B3PW91 correctly predicted TEMOA-G0 as the third strongest bound host–guest complex and yielded the lowest deviation from experiment (0.41 kcal mol^{-1}) for TEMOA-G2.

3.3.4 Quantum Mechanical Calculations

The CB8 host–guest systems were used to probe approaches for improving the binding free energy prediction. Specifically, the effects of (1) utilizing truncated correlation consistent basis sets as opposed to standard correlation consistent basis sets; (2) utilizing traditional DFT calculations (neglecting the RI approximation); and (3) modifying the dielectric constant used in the continuum solvation model to reflect the ionic strength of the solution used in experiment were examined.

As shown in Tables 3.1, 3.2 and 3.3, for CB8, OA without the statistical outlier (OA-G2), and TEMOA, the MAE, and RMSE increased by approximately 19.4, 25.5, and 32.6 kcal/ mol when using Grimme's D3 dispersion with RI-B3PW91, respectively, away from experiment. However, when using Grimme's D3 dispersion, the τ value decreases from 0.05 to −0.14 for CB8 but increases from −0.05 to 0.71 when the statistical outlier is removed for OA and increases from −0.14 to 0.57 for TEMOA. This shows the importance of using a dispersion correction for qualitative ranking of binding affinities.

Table 3. 4 The binding free energies for CB8 complexes in kcal mol^{-1} with various schemes involving not using the RI approximation, changing the dielectric constant of the implicit solvent with the truncated correlation consistent basis sets for hydrogen.

Complex	Exp	B3PW91-D3 (SMD, ε=78.4)			RI-B3PW91-D3 (SMD, ε=78.4)			RI-B3PW91-D3 (SMD, ε=76.4)		
		TZ (−1d)	TZ	QZ (−1f2d)	TZ (−1d)	TZ	QZ (−1f2d)	TZ (−1d)	TZ	QZ (−1f2d)
CB8-G0	−6.69±0.05	−49.85	−49.91	−49.27	−49.84	−49.89	−49.25	−49.84	−49.82	−36.26
CB8-G1	−7.65±0.04	−54.54	−57.22	−56.61	−57.21	−57.22	−56.61	−57.21	−57.24	−56.62
CB8-G2	−7.66±0.05	−37.32	−36.86	−36.39	−36.82	−36.86	−36.39	−36.82	−36.87	−36.40
CB8-G3	−6.45±0.06	−45.01	−44.54	−44.38	−44.51	−44.53	−44.38	−44.51	−44.55	−44.40
CB8-G4	−7.80±0.04	−69.19	−68.10	−67.50	−68.07	−68.09	−67.49	−68.07	−68.12	−67.52
CB8-G5	−8.18±0.05	−36.17	−16.10	−35.53	−35.89	−35.92	−35.52	−35.89	−35.95	−35.54
CB8-G6	−8.34±0.05	−31.95	−31.96	−31.63	−31.93	−31.95	−31.62	−31.95	−31.97	−31.64
CB8-G7	−10.00±0.10	−14.92	−14.95	−12.89	−14.88	−14.90	−12.89	−14.91	−14.92	−12.91
CB8-G8	−13.50±0.04	−50.61	−27.26	−49.89	−50.30	−50.34	−49.90	−50.30	−50.36	−49.92
CB8-G9	−8.68±0.08	−37.31	−19.22	−36.73	−37.05	−37.07	−36.71	−37.05	−37.09	−36.74
CB8-G10	−8.22±0.07	−42.27	−15.29	−38.92	−39.28	−39.30	−38.90	−39.28	−39.32	−38.91
CB8-G11	−7.77±0.05	−28.63	−10.21	−25.37	−25.74	−25.75	−25.36	−25.74	−25.80	−25.41
CB8-G12	−9.86±0.03	−62.53	−62.08	−61.43	−61.99	−62.05	−61.40	−61.99	−62.07	−61.41
CB8-G13	−7.11±0.03	−52.30	−51.72	−50.03	−51.73	−44.04	−50.00	−51.74	−51.75	−50.04
MAE		35.33	27.68	34.19	34.81	34.29	34.18	34.81	34.85	33.27
RMSE		37.96	33.79	37.03	37.56	36.99	37.02	37.56	37.60	36.13
τ		−0.14	−0.21	−0.12	−0.12	−0.14	−0.12	−0.12	−0.12	−0.08
r^2		0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The mean absolute error (MAE), root mean square error (RMSE), Kendall's Tau (τ), and r^2 are shown.

The binding free energies as a result of utilizing truncated basis sets individually and extrapolated to the Kohn–Sham limit with a two-point extrapolation using cc-pVDZ and cc-pVTZ (cc-pV∞Z[D,T]) and a three-point extrapolation using cc-pVDZ and truncated triple and quadruple correlation consistent basis sets, cc-pVTZ(− 1d) and ccpVQZ(−1f2d), denoted as cc(0,−1,−2), are reported in Tables 4 and 5, respectively.

Table 3. 5 The binding free energies for the CB8 complexes in kcal mol^{-1} with various schemes involving not using the RI approximation, changing the dielectric constant of the implicit solvent, and two options for basis set choice when extrapolating to the Kohn–Sham limit.

Complex	Exp	B3PW91-D3 (SMD, ε=78.4)		RI-B3PW91-D3 (SMD, ε=78.4)		RI-B3PW91-D3 (SMD, ε=76.4)	
		cc-pV∞Z [D, T]	cc (0,−1,−2)	cc-pV∞Z [D, T]	Cc (0,−1,−2)	cc-pV∞Z [D, T]	cc (0,−1,−2)
CB8-G0	−6.69±0.05	−49.91	−47.62	−49.89	−47.58	−49.82	−16.15
CB8-G1	−7.65±0.04	−57.22	−60.08	−57.22	−55.85	−57.24	−55.88
CB8-G2	−7.66±0.05	−36.86	−35.25	−36.86	−36.00	−36.87	−36.04
CB8-G3	−6.45±0.06	−44.54	−43.50	−44.53	−44.20	−44.55	−44.26
CB8-G4	−7.80±0.04	−68.10	−64.83	−68.09	−66.23	−68.12	−66.27
CB8-G5	−8.18±0.05	−16.10	−34.89	−35.92	−35.28	−35.95	−35.32
CB8-G6	−8.34±0.05	−31.96	−31.33	−31.95	−31.32	−31.96	−31.34
CB8-G7	−10.00±0.10	−14.95	−11.27	−14.90	−11.30	−14.92	−11.31
CB8-G8	−13.50±0.04	−27.26	−24.47	−50.34	−49.31	−50.36	−49.38
CB8-G9	−8.68±0.08	−19.22	−36.14	−37.07	−36.50	−37.09	−36.58
CB8-G10	−8.22±0.07	−15.29	−33.97	−39.30	−38.63	−39.32	−38.67
CB8-G11	−7.77±0.05	−10.21	−20.40	−25.75	−24.88	−25.80	−25.02
CB8-G12	−9.86±0.03	−62.08	−60.01	−62.05	−60.67	−62.07	−60.70
CB8-G13	−7.11±0.03	−51.72	−47.18	−44.04	−37.82	−51.75	−40.12
MAE		27.68	30.93	34.29	32.69	34.85	30.65
RMSE		33.79	34.71	36.99	35.56	37.60	34.12
τ		−0.21	−0.34	−0.14	−0.12	−0.12	−0.01
r^2		0.07	0.15	0.00	0.00	0.00	0.02

These options are cc-pV∞Z [D, T], which use cc-pVDZ and cc-pVTZ to extrapolate to the Kohn–Sham limit, and cc(0,−1, −2), which uses cc-pVDZ, cc-pVTZ(−1d), and cc-pVQZ(−1f2d) to extrapolate to the Kohn–Sham limit. The binding energies obtained with RI-B3PW91-D3 (SMD, ε=78.4)/cc-pV∞Z [D, T] were submitted. The mean absolute error (MAE), root mean square error (RMSE), Kendall's Tau (τ), and r^2 are shown.

For the CB8 complexes in Table 3.4, using standard DFT (B3PW91-D3) yielded a MAE of 35.33 kcal/ mol and 34.19 kcal mol^{-1} with cc-pVTZ(− 1d) and ccpVQZ(−1f2d), respectively, while RI-DFT (RI-B3PW91- D3) yielded a MAE of 34.81 and 34.18 kcal mol^{-1} for ccpVTZ(− 1d) and cc-pVQZ(− 1f2d), respectively. When changing ε from 78.4 for pure water to 76.4 to account for the ionic strength of the solution (RI-B3PW91-D3 (ε = 76.4)), all metrics (MAE, RMSE, τ, and r^2) used to gauge the method's predictive qualities for the binding free energies did not significantly change with respect to the binding free energies predicted in pure water (RIB3PW91-D3 (ε = 78.4)).

Table 3.5 shows the predicted binding free energies for B3PW91-D3 (ε = 78.4), RI-B3PW91-D3 (ε = 78.4), and RI-B3PW91-D3 (ε=76.4) at the Kohn–Sham limit using cc-pV∞Z[D,T], a two-point extrapolation using cc-pVDZ and cc-pVTZ, and cc(0,−1,−2), a three-point extrapolation using cc-pVDZ, cc-pVTZ(−1d) and cc-pVQZ(−1f2d) for the CB8 complexes. Using the cc(0,−1,−2) basis set choice for extrapolation, the binding free energies predicted by RIB3PW91-D3 (ε=78.4) and RI-B3PW91-D3 (ε=76.4) lowered the MAE by approximately 1.6 kcal mol^{-1} and 4.2 kcal/ mol, respectively, in regards to using the cc-pV∞Z[D,T] scheme.

3.4 Discussion

Calculating end-state binding free energies with MMPBSA is relatively fast and simple but results of a loss in accuracy and reliance compared to other free energy methods. It has been known that various factors affect the performance of the MMPBSA method such as the force field, solute dielectric constant, as well as sampling.[33] In our model, we employed the AM1-BCC partial charge scheme for the guest and host molecules for use with the GAFF force field to increase computational efficiency. GAFF was designed to use partial charges calculated from the restrained electrostatic potential fit (RESP) method.[56,88] Although, the AM1-BCC scheme was

parameterized to reproduce RESP charges, this may only be appropriate for the guest molecules rather than the larger host molecules. The interactions between the host and guest molecules may have been overestimated or underestimated as a result using the AM1-BCC charge scheme, hence the binding affinity predictions may be improved by using the RESP charge model.

3.4.1 Submission Analysis

For the methods submitted to the SAMPL6 competition, using RI-B3PW91-D3 yielded higher τ values for OA and TEMOA than using RI-B3PW91 for predicting binding free energies. Since there are eight guests that are bound to OA and TEMOA, τ_{crit} for α=0.05 is 0.57 for 8 data points. Only MMPBSA correlates with experiment ($|\tau| > \tau_{crit}$), as the τ values are 0.64, 0.29, and −0.21 for MMPBSA, RI-B3PW91-D3, and RI-B3PW91, respectively. However, after removing the statistical outlier, OA-G2, from the dataset, τ increases from 0.29 to 0.71, which implies that RIB3PW91-D3 also correlates with experiment. As shown in Table 3.2, RI-B3PW91-D3 ranked the binding free energies more correctly than MMPBSA when the outlier is excluded. For TEMOA, both MMPBSA and RI-B3PW91-D3 correlate with experiment with τ values of 0.79 and 0.57, respectively, which are greater than τ_{crit}.

As shown in Figure 3.7a, there is no correlation between experimental and predicted binding free energies for the CB8 host–guest systems. This is supported by $r^2 \approx 0$ and τ values of − 0.19, − 0.14, 0.12 for MMPBSA, RI-B3PW91-D3, and RI-B3PW91, respectively, which are smaller in magnitude than τ_{crit} for α=0.05 for 14 data points, which is 0.36. This also shows an inconsistency when using Grimme's dispersion correction, which may be due to the abundance of N and O atoms present in the CB8 host and empirical descriptors for those atoms. For all sets of the host–guest systems, RI-B3PW91 had a lower MAE and RMSE than RIB3PW91-D3 by approximately 19.4–32.6 kcal mol^{-1}, but as a tradeoff, resulted in qualitatively better predictions

of the binding affinities (Figure 3.8). This implies that using a dispersion correction overbinds the guest to the host but is needed for proper ranking.

To estimate the relative performance of the methods, the mean signed error (MSE) was used to offset the calculated binding free energies. After the removal of MSE from the MMPBSA and RI-B3PW91-D3 predicted binding free energies for OA and TEMOA, the MAE and the RMSE values are recalculated to estimate the performance of methods in relative terms as shown in Table 3.6. This correction improved the MAE and RMSE for MMPBSA by 6.8 and 5.9 kcal/ mol for OA and TEMOA, respectively. The correction improved the RI-B3PW91-D3 MAE and RMSE by 38.39 and 38.83 kcal mol^{-1} for OA without the OA-G2 outlier and TEMOA, respectively.

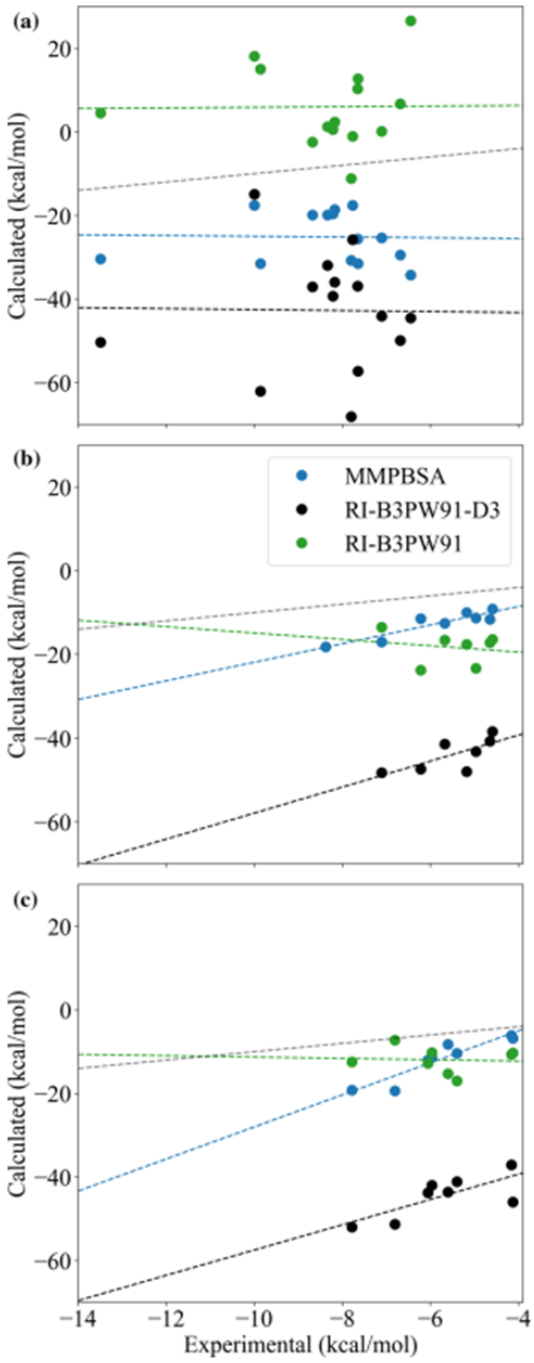


Figure 3. 7 Plots for calculated results in Tables 3.1, 3.2 and 3.3 versus experimental results in kcal mol^{-1} for (a) CB8, (b) OA, and (c) TEMOA for MMPBSA (blue), RI-B3PW91-D3 (black), and RI-B3PW91 (green). The dashed lines in each corresponding color refers to the best fit line where the statistical outlier (OA-G2) for RI-B3PW91 and RI B3PW91-D3 is removed for b and c. The dashed gray line is the y=x line.

3.4.2 Impact of Truncated Basis Sets

For the QM calculations, the subset of the CB8 host–guest systems was chosen because the size of these systems is smaller compared to the octa-acid host–guest systems investigated. While using the RI approximation, lowering ε from 78.4 for pure water to 76.4 to account for the ionic strength of the solution increased the MAE by 0.56 kcal mol^{-1}. However, altering the dielectric constant from 78.4 to 76.4 to account for the ionic strength of the solution lowered the MAE from 34.85 to 30.65 kcal mol^{-1} for the three-point extrapolation with truncated triple-ζ and quadruple-ζ correlation consistent basis sets, yet for RI-B3PW91-D3 (ε=78.4), the MAE only decreased from 34.29 to 32.69 kcal mol^{-1} (Table 3.5). Therefore, factors that can change the dielectric constant should be considered when using implicit solvent models for binding free energy predictions.

The use of the cc(0,−1,−2) basis set scheme lowered the MAE for CB8 complexes by 1.60 kcal mol^{-1} relative to using cc-pV∞Z[D,T] (Table 3.5) for RI-B3PW91-D3 (ε=78.4). In contrast, when using truncated basis sets and standard basis sets for binding free energies (Table 3.4), the MAE decreased by 0.51 kcal mol^{-1} for the CB8 complexes when using cc-pVTZ as opposed to cc-pVTZ(−1d) for RIB3PW91-D3 (ε=78.4). The MAE decreased by 0.31 kcal/ mol when increasing the basis set quality of truncated basis sets for RI-B3PW91-D3 (ε=78.4). Therefore, within the RI approximation, the decrease in MAE when using ccpVQZ(−1f2d) highlights the importance of using higher quality basis sets when extrapolating to the Kohn–Sham limit.

For predictions without the RI approximation, the binding free energies determined using B3PW91-D3/cc-pVTZ yielded a decrease in the MAE by 7.65 kcal mol^{-1} relative to B3PW91-D3/cc-pVTZ(−1d) as shown in Table 3.4. This is believed to be a result from including the four-center two-electron electron repulsion integrals removed via the RI approximation and the need

for additional polarization when describing interactions with hydrogens between the host and the guest. This effect also contributes to the increase of 3.25 kcal mol^{-1} in the MAE between B3PW91-D3/ cc-pV∞Z[D,T] and B3PW91-D3/cc(0,−1,−2). However, as shown in Table 3.5, when employing truncated basis sets (cc(0,−1,−2)), binding free energy predictions when using RI-B3PW91-D3 (ε=76.4) are more positive and yield a MAE of 0.28 kcal mol^{-1} lower than B3PW91-D3 (ε=78.4). This illustrates that within the RI approximation, changing the dielectric constant is as beneficial to predicting binding free energies as utilizing standard DFT, which is more computationally costly than RI-DFT.

For the CB8-G6 host–guest complex, which was one of the smaller systems in the set of host–guest systems, the number of basis functions decreased from 4016 to 3696 with the truncation of 1 d basis function from the cc-pVTZ basis set for hydrogen and decreased from 7640 to 6872 with the truncation of 1 f and 2 d basis functions from the cc-pVQZ basis set for hydrogen. Since DFT scales approximately N3 to N5 depending on the complexity of the functional where N is the number of basis functions, truncated basis sets become a practical option for further decreasing the computational cost while improving the quantitative prediction of binding free energies for these host–guest systems as truncating 1 d basis function from cc-pVTZ only affected the binding energy predicted with cc-pVTZ by ≤0.06 kcal mol^{-1} as shown in Table 3.4 for RI-B3PW91-D3.

3.4.3 Impact of the Extrapolation Scheme *B*-parameter

Another factor that can account for the large deviations between host–guest binding energies is the parameter used to fit Equation 3.1 for two-point extrapolations. The value of 5.5 proposed by Jensen for the *B*-parameter, which was used for atoms and diatomics, caused the extrapolation curve to converge at a very rapid rate and is reflected in the predictions for the CB8 complexes,

as the binding affinities in Table 3.1 are identical to those predicted with the ccpVTZ basis set with the respective method in Table 3.4. Also, when using the three-point extrapolations with truncated basis sets for the CB8 complexes, the *B*-parameter yielded an average value of 0.37 (Table 3.10). Therefore, the value of 0.37 for the B-parameter was applied to two-point extrapolations with cc-pVDZ and cc-pVTZ to gauge how changing the *B*-parameter affects the extrapolated binding free energies (Table 3.7). The results from using 0.37 as the *B*-parameter in a two-point extrapolation show that the MAE decreased by 0.84 and 0.42 kcal mol^{-1} for the CB8 and TEMOA complexes, respectively. The MAE did not change for the OA complexes. Setting the *B*-parameter to 0.37 did not change the τ values for CB8 and OA complexes, however, did increase the τ value from 0.57 to 0.71 for TEMOA.

In addition to applying 0.37 for the *B*-parameter to predict binding free energies for all host–guest systems using two-point extrapolations with cc-pVDZ and ccpVTZ, the value of the *B*-parameter was optimized to the value of 0.12 via minimizing the MAE and was applied (Table 3.7). For the CB8 host–guest systems, shifting the B-parameter from 5.5 to 0.12 had a noticeable impact on the MAE, which decreased from 34.29 to 29.84 kcal/ mol for RI-BWPW91-D3. A similar effect was observed for TEMOA with a decrease in the MAE of 5.07 kcal/ mol. There is no notable change in MAE, RMSE, or τ for the OA complexes with the change in the B-parameter. Furthermore, τ increases from 0.57 to 0.93 when the *B*-parameter is changed from 5.5 to 0.12 for TEMOA with RI-B3PW91-D3, which provides more evidence that dispersion-corrected functionals should be used for qualitative predictions of binding free energies since $|\tau| > \tau_{crit}$. The observed trends imply that the value of the *B*-parameter should be reoptimized when using Equation 3.1 for macromolecules.

Table 3. 6 The predicted binding energies for OA and TEMOA using MMPBSA and RI-B3PW91 after the removal of mean signed error (MSE)

	OA		TEMOA	
	MMPBSA	**RI-B3PW91-D3**	**MMPBSA**	**RI-B3PW91-D3**
MAE	1.6±0.2a	11.66 [2.81]	3.0±0.2[a]	3.49
RMSE	1.9±0.4b	17.87 [3.12]	3.7±0.5[b]	3.95
τ	0.64	0.29 [0.71]	0.79	0.57
r^2	0.84	0.44 [0.52]	0.86	0.55

Bracketed values indicate the values after the removal of the statistical outlier (OA-G2). The mean absolute error (MAE) in kcal mol^{-1}, root mean square error (RMSE) in kcal mol^{-1}, Kendall's Tau (τ), and r^2 are shown.
[a]The uncertainty reported for MAE is the average of the absolute uncertainties.
[b]The uncertainty reported for RMSE is the uncertainty of the RMSE with the experimental and calculated uncertainties.

Compared to other submissions employing QM methods in the SAMPL6 host–guest binding challenge, our approach yielded quantitatively poorer predictions that may have resulted from the approximations considered in this work. In our approach, only a single conformational state of the guest binding to the host system was considered. Additionally, the representative structures of the individual host–guest systems obtained from clustering the MD trajectories were not optimized with QM methods and is reflected in our model chemistries.

Table 3. 7 The predicted binding energies when using different values for B in Eq. 1 for two-point extrapolations using cc-pVDZ and cc-pVTZ with RI-B3PW91-D3.

	B=5.5	**B=0.37**	**B=0.12**
CB8			
MAE	34.29	33.45	29.84
RMSE	36.99	36.33	33.34
τ	−0.14	−0.14	−0.03
r^2	0	0	0
OA			
MAE	35.46[38.39]	35.46[38.42]	35.43[38.74]
RMSE	36.41[38.52]	36.43[38.54]	36.70[38.86]
τ	0.29[0.71]	0.29[0.71]	0.29[0.71]
r^2	0.44[0.52]	0.43[0.52]	0.43[0.54]
TEMOA			
MAE	38.83	38.41	33.76
RMSE	39.03	38.6	36.3
τ	0.57	0.71	0.93
r^2	0.55	0.75	0.58

Bracketed values indicate the values after the removal of the statistical outlier (OA-G2). The mean absolute error (MAE) in kcal mol^{-1}, the root mean square error (RMSE) in kcal mol^{-1}, the Kendall's Tau (τ), and r^2 are shown.

3.4.4 Impact of representative geometries

The cause of OA-G2 being a statistical outlier is suspected to be from the orientation of the substituted cyclohexene ring relative to the OA host (Figure 3.5). Comparing OA-G2 and TEMOA-G2 in Figures 3.5 and 3.6, where the only difference is the four methyl groups on the host, the structure of the OA-G2 complex has a smaller binding pocket than the TEMOA-G2 complex. While the experimental data suggests that G2 has a stronger binding affinity towards OA than TEMOA, MMPBSA suggests the opposite. More sampling of representative structures would aid in depicting whether the anomalous binding behavior of OA-G2 correlates with the positive binding free energies predicted with DFT.

Although the only difference between CB8-G6 and CB8- G7 was the expansion of the ring for the guest by one CH2 group, the predicted binding affinities for the CB8-G6 and CB8-G7 complexes differed by approximately 17.0 kcal/ mol. This may be due to the binding poses of CB8-G6 and CB8-G7 complexes, as G6 bound in a perpendicular fashion inside the binding pocket relative to the host whereas G7 bound in a parallel fashion inside the binding pocket. This would affect nearby electrostatic interactions and why for B3PW91-D3 (ε=78.4), RI-B3PW91-D3 (ε=78.4), and RI-B3PW91-D3 (ε=76.4), there was a 3.00 kcal mol^{-1} difference in the change of binding energies between CB8-G6 and CB8-G7 when improving basis set quality via the basis set scheme used for extrapolation (Table 3.5). Ergo, more sampling of chemically relevant structures or enhanced sampling methods can provide a more robust depiction of the host–guest binding environment.

The volumes of guest molecules for OA and CB8 molecules were compared to each other. The volumes of the guests bound to CB8 are larger than those bound to OA and TEMOA as shown in Tables 3.9 and 3.10. The guests CB8-G0, CB8-G1, CB8-G2, CB8-G3, CB8-G4, and CB8- G12 are among the largest ligands for this year's competition with volumes of 462, 518, 432, 468, 817, and 553 $Å^3$, respectively. These values are more than twice the average volume of OA guests and the absolute error between the experimental and the predicted binding free energies for the larger CB8 guests are among the highest for all our methods (MMPBSA, RI-B3W91-D3 and RI-B3PW91) as shown in Figure 3.8. The MMPBSA and RI-B3PW91-D3 methods have a definite correlation with the experiment based on ranking the binding affinities of the octa-acid guest molecules, which were smaller in volume on average compared to the CB8 guests. This correlation is evident from the τ values of 0.64, 0.79, 0.71, 0.57 for MMPBSA (OA),

MMPBSA (TEMOA), RI-B3PW91-D3 (OA without OA-G2 outlier), and RI-B3PW91-D3 (TEMOA), respectively.

However, these two methods do not correlate to the CB8 binding free energies since the τ values are −0.19 and −0.14 for MMPBSA and RI-B3PW91-D3, respectively. This may result from insufficient sampling as the CB8 guests are larger molecules with higher conformational flexibility. For example, the size of CB8-G4 does not allow the guest to fit entirely into the binding cavity. As a result, most of the CB8-G4 molecule is weakly bound to the host from outside of the binding pocket and only one of the three triethyl amines within the guest can fit into the pocket as shown in Figure 3.4. Each triethyl amine group could bind to the host from inside the binding cavity, which would result in alternative binding conformations and affect the overall binding free energy. To better understand binding free energies of these large structures, more sampling of the different binding modes is needed to generate weighted averages based on the thermodynamic stability of predicted poses.

The results for OA and TEMOA systems illustrate that MMPBSA and RI-B3PW91-D3 methods can be used to qualitatively rank binding energies of small molecules. Among those two methods, MMPBSA is computationally less expensive, but RI-B3PW91-D3 predicted the relative binding affinities better for OA and TEMOA host–guest systems. However, the MAE and the corresponding error plots (Figure 3.8) indicate that both methods overestimated the binding free energies. The MAE reported for the OA and TEMOA complexes state that MMPBSA and RI-B3PW91-D3 predict overbinding by 6.8 and 35.5 kcal mol^{-1}, respectively, for OA complexes and 5.9 and 38.8 kcal mol^{-1}, respectively, for TEMOA complexes. For all systems, the MMPBSA method was the best approach overall in terms of quantitative predictions.

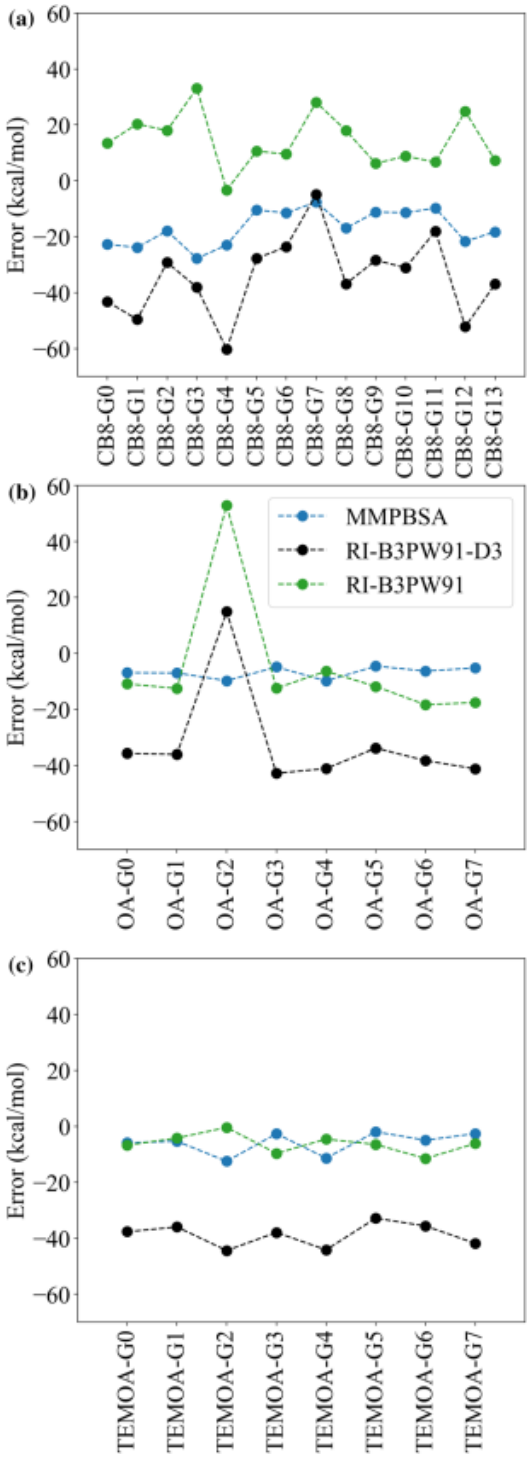


Figure 3. 8 Error plots from experimental results in kcal mol^{-1} for (a) CB8 (b) OA, and (c) TEMOA for MMPBSA (blue), RI-B3PW91- D3 (black), and RI-B3PW91 (green) for the submitted results from Tables 3.1, 3.2 and 3.3.

3.5 Conclusions

When implementing DFT for predicting host–guest binding affinities, the use of Grimme's D3 dispersion correction was essential for qualitatively predicting the binding free energies for the OA and TEMOA systems even though the MAE exceeded 35.0 kcal mol^{-1} for both the OA and TEMOA systems. When using implicit solvent models, factors that can change the dielectric constant, such as the ionic strength of the solution, are relevant for predicting binding free energies, as lowering the dielectric constant lowered the MAE. While RI-B3PW91-D3 reduced the computational cost relative to B3PW91-D3, B3PW91-D3 yielded a lower MAE. To attain more quantitatively favorable results, using cc-pVQZ(−1f2d) for hydrogen atoms reduces the computational cost relative to using cc-pVQZ while simultaneously providing a better standard for extrapolating to the Kohn–Sham limit than only utilizing cc-pVDZ and cc-pVTZ for extrapolations. Also, truncating 1 d basis function for hydrogen atoms had a very small effect on predicted binding free energies obtained with cc-pVTZ, indicating that truncated basis sets are a viable option to reduce the computational cost while yielding near-identical binding free energies. With the extrapolation scheme utilized, the B-parameter should be revised for macromolecules since reducing the value of the B-parameter from the proposed 5.5 to 0.12 reduced the MAE while providing extrapolated binding energies that were in alignment with those predicted using quadruple-ζ level basis sets. Sampling of different binding poses becomes pertinent for future investigations as binding orientation in the pocket affected the predicted binding free energies by approximately 17.0 kcal mol^{-1} when using RI-B3PW91-D3 for guests that only differed by one CH2 group.

All methods presented predict over binding character for these host–guest systems except for RI-B3PW91 for CB8 host–guest systems. MMPBSA and RI-B3PW91- D3 worked well at

ranking binding affinities for smaller guests regardless of the size of the host. The CB8 guest molecules with a larger van der Waals volume yielded poor prediction of binding free energy due to their higher conformational flexibility, which can complicate predicting binding poses. To better understand binding free energies of these large structures, enhanced sampling methods can be used, and multiple host–guest binding poses can be sampled.

APPENDIX

Table 3. 8 Van der Waals volumes in $Å^3$ of CB8 guest molecules are calculated using connection table approximation.

Guest	Volume
CB8-G0	462
CB8-G1	518
CB8-G2	432
CB8-G3	468
CB8-G4	817
CB8-G5	249
CB8-G6	190
CB8-G7	214
CB8-G8	312
CB8-G9	211
CB8-G10	244
CB8-G11	184
CB8-G12	553
CB8-G13	265
Average	366

Table 3. 9 Van der Waals volumes in $Å^3$ of OA and TEMOA guest molecules are calculated using connection table approximation.

Guest	Volume
OA-G0	176
OA-G1	160
OA-G2	238
OA-G3	160
OA-G4	258
OA-G5	160
OA-G6	166
OA-G7	184
Average	188

Table 3. 10 Fitting parameter values obtained when using Jensen's extrapolation scheme for each component in calculating the binding energy (Equation 1). The host and guest are counterpoise-corrected before the extrapolation was performed.

Complex	Complex	Host	Guest
CB8-G0	0.37	0.36	0.41
CB8-G1	0.36	0.35	0.37
CB8-G2	0.36	0.36	0.37
CB8-G3	0.36	0.36	0.37
CB8-G4	0.32	0.32	0.34
CB8-G5	0.38	0.38	0.39
CB8-G6	0.39	0.39	0.40
CB8-G7	0.38	0.38	0.40
CB8-G8	0.37	0.37	0.39
CB8-G9	0.39	0.39	0.39
CB8-G10	0.38	0.38	0.38
CB8-G11	0.39	0.39	0.40
CB8-G12	0.36	0.35	0.37
CB8-G13	0.39	0.38	0.40
Average	0.37	0.37	0.38

REFERENCES

REFERENCES

(1) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr. Opin. Struct. Biol.* **2009**, *19* (2), 120–127. https://doi.org/10.1016/j.sbi.2009.03.004.

(2) Shan, Y.; Seeliger, M. A.; Eastwood, M. P.; Frank, F.; Xu, H.; Jensen, M. O.; Dror, R. O.; Kuriyan, J.; Shaw, D. E. A Conserved Protonation-Dependent Switch Controls Drug Binding in the Abl Kinase. *Proc. Natl. Acad. Sci.* **2009**, *106* (1), 139–144. https://doi.org/10.1073/pnas.0811223106.

(3) Zhao, G.; Perilla, J. R.; Yufenyuy, E. L.; Meng, X.; Chen, B.; Ning, J.; Ahn, J.; Gronenborn, A. M.; Schulten, K.; Aiken, C.; Zhang, P. Mature HIV-1 Capsid Structure by Cryo-Electron Microscopy and All-Atom Molecular Dynamics. *Nature* **2013**, *497* (7451), 643–646. https://doi.org/10.1038/nature12162.

(4) Perilla, J. R.; Goh, B. C.; Cassidy, C. K.; Liu, B.; Bernardi, R. C.; Rudack, T.; Yu, H.; Wu, Z.; Schulten, K. Molecular Dynamics Simulations of Large Macromolecular Complexes. *Curr. Opin. Struct. Biol.* **2015**, *31*, 64–74. https://doi.org/10.1016/j.sbi.2015.03.007.

(5) Walkowicz, W. E.; Fernández-Tejada, A.; George, C.; Corzana, F.; Jiménez-Barbero, J.; Ragupathi, G.; Tan, D. S.; Gin, D. Y. Quillaja Saponin Variants with Central Glycosidic Linkage Modifications Exhibit Distinct Conformations and Adjuvant Activities. *Chem. Sci.* **2016**, *7* (3), 2371–2380. https://doi.org/10.1039/C5SC02978C.

(6) Hadden, J. A.; Perilla, J. R.; Schlicksup, C. J.; Venkatakrishnan, B.; Zlotnick, A.; Schulten, K. All-Atom Molecular Dynamics of the HBV Capsid Reveals Insights into Biological Function and Cryo-EM Resolution Limits. *Elife* **2018**, *7*, e32478. https://doi.org/10.7554/eLife.32478.

(7) García, M. A.; Meurs, E. F.; Esteban, M. The DsRNA Protein Kinase PKR: Virus and Cell Control. *Biochimie* **2007**, *89* (6–7), 799–811. https://doi.org/10.1016/j.biochi.2007.03.001.

(8) Tripathi, R. B.; Pande, M.; Garg, G.; Sharma, D. In-Silico Expectations of Pharmaceutical Industry to Design of New Drug Molecules. *J. Innov. Pharm. Biol. Sci.* **2016**, *3* (3), 95–103.

(9) Ryde, U.; Söderhjelm, P. Ligand-Binding Affinity Estimates Supported by Quantum-Mechanical Methods. *Chem. Rev.* **2016**, *116* (9), 5520–5566. https://doi.org/10.1021/acs.chemrev.5b00630.

(10) Ganesan, A.; Coote, M. L.; Barakat, K. Molecular Dynamics-Driven Drug Discovery: Leaping Forward with Confidence. *Drug Discov. Today* **2017**, *22* (2), 249–269. https://doi.org/10.1016/j.drudis.2016.11.001.

(11) Mobley, D. L.; Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annu. Rev. Biophys.* **2017**, *46* (1), 531–558. https://doi.org/10.1146/annurev-biophys-070816-033654.

(12) Huggins, D. J.; Sherman, W.; Tidor, B. Rational Approaches to Improving Selectivity in Drug Design. *J. Med. Chem.* **2012**, *55* (4), 1424–1444. https://doi.org/10.1021/jm2010332.

(13) Muddana, H. S.; Daniel Varnado, C.; Bielawski, C. W.; Urbach, A. R.; Isaacs, L.; Geballe, M. T.; Gilson, M. K. Blind Prediction of Host–Guest Binding Affinities: A New SAMPL3 Challenge. *J. Comput. Aided. Mol. Des.* **2012**, *26* (5), 475–487. https://doi.org/10.1007/s10822-012-9554-1.

(14) Rogers, K. E.; Ortiz-Sánchez, J. M.; Baron, R.; Fajer, M.; De Oliveira, C. A. F.; McCammon, J. A. On the Role of Dewetting Transitions in Host-Guest Binding Free Energy Calculations. *J. Chem. Theory Comput.* **2013**, *9* (1), 46–53. https://doi.org/10.1021/ct300515n.

(15) Yang, H.; Yuan, B.; Zhang, X.; Scherman, O. A. Supramolecular Chemistry at Interfaces: Host-Guest Interactions for Fabricating Multifunctional Biointerfaces. *Acc. Chem. Res.* **2014**, *47* (7), 2106–2115. https://doi.org/10.1021/ar500105t.

(16) Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. The SAMPL4 Host–Guest Blind Prediction Challenge: An Overview. *J. Comput. Aided. Mol. Des.* **2014**, *28* (4), 305–317. https://doi.org/10.1007/s10822-014-9735-1.

(17) Gallicchio, E.; Chen, H.; Chen, H.; Fitzgerald, M.; Gao, Y.; He, P.; Kalyanikar, M.; Kao, C.; Lu, B.; Niu, Y.; Pethe, M.; Zhu, J.; Levy, R. M. BEDAM Binding Free Energy Predictions for the SAMPL4 Octa-Acid Host Challenge. *J. Comput. Aided. Mol. Des.* **2015**, *29* (4), 315–325. https://doi.org/10.1007/s10822-014-9795-2.

(18) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 Host–Guest Challenge: Are We Doing Better? *J. Comput. Aided. Mol. Des.* **2017**, *31* (1), 1–19. https://doi.org/10.1007/s10822-016-9974-4.

(19) Liu, S.; Ruspic, C.; Mukhopadhyay, P.; Chakrabarti, S.; Zavalij, P. Y.; Isaacs, L. The Cucurbit[n]Uril Family: Prime Components for Self-Sorting Systems. *J. Am. Chem. Soc.* **2005**, *127* (45), 15959–15967. https://doi.org/10.1021/ja055013x.

(20) Gan, H.; Benjamin, C. J.; Gibb, B. C. Nonmonotonic Assembly of a Deep-Cavity Cavitand. *J. Am. Chem. Soc.* **2011**, *133* (13), 4770–4773. https://doi.org/10.1021/ja200633d.

(21) Biedermann, F.; Scherman, O. A. Cucurbit[8]Uril Mediated Donor–Acceptor Ternary Complexes: A Model System for Studying Charge-Transfer Interactions. *J. Phys. Chem. B* **2012**, *116* (9), 2842–2849. https://doi.org/10.1021/jp2110067.

(22) Vázquez, J.; Remón, P.; Dsouza, R. N.; Lazar, A. I.; Arteaga, J. F.; Nau, W. M.; Pischel,

U. A Simple Assay for Quality Binders to Cucurbiturils. *Chem. - A Eur. J.* **2014**, *20* (32), 9897–9901. https://doi.org/10.1002/chem.201403405.

(23) Gibb, C. L. D.; Gibb, B. C. Binding of Cyclic Carboxylates to Octa-Acid Deep-Cavity Cavitand. *J. Comput. Aided. Mol. Des.* **2014**, *28* (4), 319–325. https://doi.org/10.1007/s10822-013-9690-2.

(24) Nicholls, A.; Wlodek, S.; Grant, J. A. The SAMP1 Solvation Challenge: Further Lessons Regarding the Pitfalls of Parametrization. *J. Phys. Chem. B* **2009**, *113* (14), 4521–4532. https://doi.org/10.1021/jp806855q.

(25) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations †. *J. Phys. Chem. B* **2009**, *113* (14), 4533–4537. https://doi.org/10.1021/jp806838b.

(26) Geballe, M. T.; Skillman, a. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 Blind Prediction Challenge: Introduction and Overview. *J. Comput. Aided. Mol. Des.* **2010**, *24* (4), 259–279. https://doi.org/10.1007/s10822-010-9350-8.

(27) Zwanzig, R. W. High-temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426. https://doi.org/10.1063/1.1740409.

(28) Jiang, W.; Hodoscek, M.; Roux, B. Computation of Absolute Hydration and Binding Free Energy with Free Energy Perturbation Distributed Replica-Exchange Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5* (10), 2583–2588. https://doi.org/10.1021/ct900223z.

(29) Mitchell, M. J.; McCammon, J. A. Free Energy Difference Calculations by Thermodynamic Integration: Difficulties in Obtaining a Precise Value. *J. Comput. Chem.* **1991**, *12* (2), 271–275. https://doi.org/10.1002/jcc.540120218.

(30) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical Free Energy Methods for Drug Discovery: Progress and Challenges. *Curr. Opin. Struct. Biol.* **2011**, *21* (2), 150–160. https://doi.org/10.1016/j.sbi.2011.01.011.

(31) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10* (7), 2632–2647. https://doi.org/10.1021/ct500161f.

(32) Williams-Noonan, B. J.; Yuriev, E.; Chalmers, D. K. Free Energy Methods in Drug Design: Prospects of "Alchemical Perturbation" in Medicinal Chemistry. *J. Med. Chem.* **2018**, *61* (3), 638–649. https://doi.org/10.1021/acs.jmedchem.7b00681.

(33) 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. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82. https://doi.org/10.1021/ci100275a.

(34) Homeyer, N.; Gohlke, H. Free Energy Calculations by the Molecular Mechanics Poisson−Boltzmann Surface Area Method. *Mol. Inform.* **2012**, *31* (2), 114–122. https://doi.org/10.1002/minf.201100135.

(35) Genheden, S.; Ryde, U. The MM / PBSA and MM / GBSA Methods to Estimate Ligand-Binding Affinities. **2015**.

(36) Wang, C.; Greene, D.; Xiao, L.; Qi, R.; Luo, R. Recent Developments and Applications of the MMPBSA Method. *Front. Mol. Biosci.* **2018**, *4*. https://doi.org/10.3389/fmolb.2017.00087.

(37) Genheden, S.; Ryde, U. Comparison of the Efficiency of the LIE and MM/GBSA Methods to Calculate Ligand-Binding Energies. *J. Chem. Theory Comput.* **2011**, *7* (11), 3768–3778. https://doi.org/10.1021/ct200163c.

(38) Hansson, T.; Marelius, J.; Aqvist, J. Ligand Binding Affinity Prediction by Linear Interaction Energy Methods. *J. Comput. Aided. Mol. Des.* **1998**, *12* (1), 27–35. https://doi.org/10.1023/A:1007930623000.

(39) Steinmann, C.; Olsson, M. A.; Ryde, U. Relative Ligand-Binding Free Energies Calculated from Multiple Short QM/MM MD Simulations. *J. Chem. Theory Comput.* **2018**, Article ASAP. https://doi.org/10.1021/acs.jctc.8b00081.

(40) Curutchet, C.; Cupellini, L.; Kongsted, J.; Corni, S.; Frediani, L.; Steindal, A. H.; Guido, C. A.; Scalmani, G.; Mennucci, B. Density-Dependent Formulation of Dispersion-Repulsion Interactions in Hybrid Multiscale Quantum/Molecular Mechanics (QM/MM) Models. *J. Chem. Theory Comput.* **2018**, *14* (3), 1671–1681. https://doi.org/10.1021/acs.jctc.7b00912.

(41) Sellers, B. D.; James, N. C.; Gobbi, A. A Comparison of Quantum and Molecular Mechanical Methods to Estimate Strain Energy in Druglike Fragments. *J. Chem. Inf. Model.* **2017**, *57* (6), 1265–1275. https://doi.org/10.1021/acs.jcim.6b00614.

(42) Lu, Y.; Yang, C. Y.; Wang, S. Binding Free Energy Contributions of Interfacial Waters in HIV-1 Protease/Inhibitor Complexes. *J. Am. Chem. Soc.* **2006**, *128* (36), 11830–11839. https://doi.org/10.1021/ja058042g.

(43) Bonnet, P.; Bryce, R. A. Molecular Dynamics and Free Energy Analysis of Neuraminidase – Ligand Interactions. *Protein Sci.* **2004**, *13*, 946–957. https://doi.org/10.1110/ps.03129704.four-hydroxyl.

(44) Kitamura, K.; Tamura, Y.; Ueki, T.; Ogata, K.; Noda, S.; Himeno, R.; Chuman, H. Binding Free-Energy Calculation Is a Powerful Tool for Drug Optimization: Calculation and Measurement of Binding Free Energy for 7-Azaindole Derivatives to Glycogen Synthase Kinase-3β. *J. Chem. Inf. Model.* **2014**, *54* (6), 1653–1660. https://doi.org/10.1021/ci400719v.

(45) Caldararu, O.; Olsson, M. A.; Riplinger, C.; Neese, F.; Ryde, U. Binding Free Energies in

the SAMPL5 Octa-Acid Host–Guest Challenge Calculated with DFT-D3 and CCSD(T). *J. Comput. Aided. Mol. Des.* **2017**, *31* (1), 87–106. https://doi.org/10.1007/s10822-016-9957-5.

(46) Sure, R.; Antony, J.; Grimme, S. Blind Prediction of Binding Affinities for Charged Supramolecular Host-Guest Systems: Achievements and Shortcomings of DFT-D3. *J. Phys. Chem. B* **2014**, *118* (12), 3431–3440. https://doi.org/10.1021/jp411616b.

(47) Mikulskis, P.; Cioloboc, D.; Andrejić, M.; Khare, S.; Brorsson, J.; Genheden, S.; Mata, R. A.; Söderhjelm, P.; Ryde, U. Free-Energy Perturbation and Quantum Mechanical Study of SAMPL4 Octa-Acid Host-Guest Binding Energies. *J. Comput. Aided. Mol. Des.* **2014**, *28* (4), 375–400. https://doi.org/10.1007/s10822-014-9739-x.

(48) Murkli, S.; McNeil, J.; Isaacs, L. CB[8]-Guest Binding Affinities: A Blinded Dataset for the SAMPL6 Challenge. *Supramol. Chem.* **2018**, (Submitted).

(49) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, Approximate and Parallel Hartree–Fock and Hybrid DFT Calculations. A 'Chain-of-Spheres' Algorithm for the Hartree–Fock Exchange. *Chem. Phys.* **2009**, *356* (1–3), 98–109. https://doi.org/10.1016/j.chemphys.2008.10.036.

(50) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. https://doi.org/10.1063/1.3382344.

(51) Mintz, B.; Lennox, K. P.; Wilson, A. K. Truncation of the Correlation Consistent Basis Sets: An Effective Approach to the Reduction of Computational Cost? *J. Chem. Phys.* **2004**, *121* (12), 5629–5634. https://doi.org/10.1063/1.1785145.

(52) Chemical Computing Group Inc. Molecular Operating Environment (MOE). Montreal 2016.

(53) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in Docking Success Rates Due to Dataset Preparation. *J. Comput. Aided. Mol. Des.* **2012**, *26* (6), 775–786. https://doi.org/10.1007/s10822-012-9570-1.

(54) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39* (6), 1397–1412. https://doi.org/10.1063/1.1734456.

(55) 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* **2006**, *65* (3), 712–725. https://doi.org/10.1002/prot.21123.

(56) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. a; Case, D. a. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(57) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641. https://doi.org/10.1002/jcc.10128.

(58) Becke, A. D. Density-functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648–5652. https://doi.org/10.1063/1.464913.

(59) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37* (2), 785–789. https://doi.org/10.1103/PhysRevB.37.785.

(60) Hariharan, P. C.; Pople, J. A. *The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies*; Springer-Verlag, 1973; Vol. 28. https://doi.org/10.1007/BF00533485.

(61) Hay, P. J.; Wadt, W. R. Ab Initio Effective Core Potentials for Molecular Calculations. Potentials for the Transition Metal Atoms Sc to Hg. *J. Chem. Phys.* **1985**, *82* (1), 270–283. https://doi.org/10.1063/1.448799.

(62) 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.

(63) Case, D. A.; Betz, R. M.; Botello-Smith, W.; Cerutti, D. S.; Cheatham III, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; York, D. M.; Kollman, P. A. Amber 16. 2016. https://doi.org/10.1002/jcc.23031.

(64) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(65) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120* (20), 9665–9678. https://doi.org/10.1063/1.1683075.

(66) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8* (9), 3314–3321. https://doi.org/10.1021/ct300418h.

(67) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD-96*; 1996; pp 226–231.

(68) Merrick, J. P.; Moran, D.; Radom, L. An Evaluation of Harmonic Vibrational Frequency Scale Factors. *J. Phys. Chem. A* **2007**, *111* (45), 11683–11700. https://doi.org/10.1021/jp073974n.

(69) Neese, F. Software Update: The ORCA Program System, Version 4.0. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8* (1), e1327. https://doi.org/10.1002/wcms.1327.

(70) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy. *Phys. Rev. B* **1992**, *45* (23), 13244–13249. https://doi.org/10.1103/PhysRevB.45.13244.

(71) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46* (11), 6671–6687. https://doi.org/10.1103/PhysRevB.46.6671.

(72) Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary Basis Sets to Approximate Coulomb Potentials. *Chem. Phys. Lett.* **1995**, *240* (4), 283–290. https://doi.org/10.1016/0009-2614(95)00621-A.

(73) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378–6396. https://doi.org/10.1021/jp810292n.

(74) Goerigk, L.; Grimme, S. A General Database for Main Group Thermochemistry, Kinetics, and Noncovalent Interactions - Assessment of Common and Reparameterized (Meta-)GGA Density Functionals. *J. Chem. Theory Comput.* **2010**. https://doi.org/10.1021/ct900489g.

(75) Goerigk, L.; Grimme, S. A Thorough Benchmark of Density Functional Methods for General Main Group Thermochemistry, Kinetics, and Noncovalent Interactions. *Phys. Chem. Chem. Phys.* **2011**, *13* (14), 6670. https://doi.org/10.1039/c0cp02984j.

(76) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90* (2), 1007–1023. https://doi.org/10.1063/1.456153.

(77) Feller, D. Application of Systematic Sequences of Wave Functions to the Water Dimer. *J. Chem. Phys.* **1992**, *96* (8), 6104–6114. https://doi.org/10.1063/1.462652.

(78) Martin, J. M. L. Ab Initio Total Atomization Energies of Small Molecules — towards the Basis Set Limit. *Chem. Phys. Lett.* **1996**, *259* (5–6), 669–678. https://doi.org/10.1016/0009-2614(96)00898-6.

(79) Wilson, A. K.; Dunning Jr., T. H. Benchmark Calculations with Correlated Molecular Wave Functions. X. Comparison with "Exact" MP2 Calculations on Ne, HF, H2O, and N2. *J. Chem. Phys.* **1997**, *106* (21), 8718–8726. https://doi.org/10.1063/1.473932.

(80) Feller, D.; Peterson, K. A.; Crawford, T. D. Sources of Error in Electronic Structure Calculations on Small Chemical Systems. *J. Chem. Phys.* **2006**, *124* (5), 54107. https://doi.org/10.1063/1.2137323.

(81) Jensen, F. Polarization Consistent Basis Sets. II. Estimating the Kohn–Sham Basis Set Limit. *J. Chem. Phys.* **2002**, *116* (17), 7372–7379. https://doi.org/10.1063/1.1465405.

(82) Faver, J. C.; Zheng, Z.; Merz, K. M. Model for the Fast Estimation of Basis Set Superposition Error in Biomolecular Systems. *J. Chem. Phys.* **2011**, *135* (14). https://doi.org/10.1063/1.3641894.

(83) Boys, S. F.; Bernardi, F. The Calculation of Small Molecular Interactions by the Differences of Separate Total Energies. Some Procedures with Reduced Errors. *Mol. Phys.* **1970**, *19* (4), 553–566. https://doi.org/10.1080/00268977000101561.

(84) Gavish, N.; Promislow, K. Dependence of the Dielectric Constant of Electrolyte Solutions on Ionic Concentration: A Microfield Approach. *Phys. Rev. E* **2016**, *94* (1), 012611. https://doi.org/10.1103/PhysRevE.94.012611.

(85) Kendall, M. G. A New Measure of Rank Correlation. *Biometrika* **1938**, *30* (1/2), 81. https://doi.org/10.2307/2332226.

(86) Berry, K. J.; Johnston, J. E.; Zahran, S.; Mielke, P. W. Stuart ' s Tau Measure of Effect Size for Ordinal Variables : Some Methodological Considerations. **2009**, *41* (4), 1144–1148. https://doi.org/10.3758/BRM.41.4.1144.

(87) Dean, R. B.; Dixon, W. J. Simplified Statistics for Small Numbers of Observations. *Anal. Chem.* **1951**, *23* (4), 636–638. https://doi.org/10.1021/ac60052a025.

(88) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269–10280. https://doi.org/10.1021/j100142a004.

CHAPTER FOUR

SAMPL7: Host–Guest Binding Prediction by Molecular Dynamics and Quantum Mechanics

About this chapter: This chapter is reprinted from Eken, Y.; Almeida, N. M. S.; Wang, C.; Wilson, A. K. SAMPL7: Host–Guest Binding Prediction by Molecular Dynamics and Quantum Mechanics. J. Comput. Aided. Mol. Des. **2021**, *35 (1)*, 63–77 with permission of the Springer Natures. The docking, molecular dynamics simulations and MMPBSA/MMGBSA calculations mentioned in this chapter are performed by Yiğitcan Eken and quantum mechanics calculations are done by co-authors Nuno M.S. Almeida and Cong Wang.

4.1 Introduction

Computer-aided drug design (CADD) has become a fundamental approach for the pharmaceutical industry and medicinal community.[1–4] Even though CADD methods are used in various phases of drug design to help enhance, accelerate, and reduce the cost of the discovery of pharmaceuticals, the balance between accuracy and computational cost of the prediction method with respect to computing time and memory is still being scrutinized.

SAMPL challenges provide opportunities for large-scale investigations of computational strategies for the prediction of physicochemical properties (i.e. solvation free energies, distribution coefficient, pKa values, and binding free energies) of a series of compounds.[5–12] In SAMPL challenges, the physicochemical properties of the compounds are measured experimentally prior to the challenge, and the results from experiment are not provided until the computational predictions have been submitted. Then, the predicted properties are compared with the experimental measurements to assess the submissions. One of the central challenges of CADD is the accurate prediction of ligand binding to a target protein, which can significantly reduce the number of compounds synthesized and considered experimentally. The biggest barrier in developing and applying such methods are large protein structures. The proteins have flexible configurations along with multiple binding sites and conformations, which make the study

computationally time consuming and difficult to sample all binding configurations. On the other hand, host molecules, which have uses such as reaction vessels, separation devices, and modulators for redox-active or fluorescent guests, have smaller structures and less conformations in comparison to proteins.[13] Due to their low complexity, host–guest systems are commonly used for the study of ligand binding predictions.[5–8,10]

In 2019, the SAMPL7 Gibb Deep Cavity Cavitand (GDCC) challenge prompted the study of the binding of eight guest molecules on two GDCC hosts, which are the Octa Acid (OA) and exo-Octa Acid (exoOA) hosts. The OA host and guests 1 to 6 (G1–G6) also were used in the previous SAMPL competitions.[6,7,10] The previously studied OA systems were considered as reference, two new guests were included, and studied for the OA, and a total of eight guests were studied on the new exoOA host. For the prediction of binding free energies there are numerous computational methods available, and examples include empirical ligand docking scoring functions, and methods that use molecular dynamics (MD) trajectories, such as free-energy perturbation[14], replica exchange free energy perturbation [15], thermodynamic integration [16] and end-state free energy methods [17]. During previous SAMPL challenges, predictions based on MD simulations are commonly adopted.[7,10,18–25] QM approaches have also been valuable, but the size of the system can make computational calculations impractical. However, methods such as DFT have been useful, because their computational cost is lower relative to ab initio quantum mechanical methods. MMPBSA or MMGBSA methods [17] are among the most popular end-state free energy methods, which can be used to calculate the free energy change between two states (such as bound and a free state of host–guest or protein–ligand). As these end-state methods do not require simulation of the intermediate states, they are among the computationally least demanding of the free energy methods that can be used without the loss of accuracy as compared

to more rigorous techniques.[26] The MMPBSA/MMGBSA methods are built upon a series of calculations: free energies of the solvated unbound receptor, unbound ligand, and the ligand bound protein complex systems (Eq. 4.1). The free energy of each system defined in Eq. 4.2 is calculated by approximating gas-phase energies (E_{gas}) from molecular mechanics, and the solvation free energies ($\Delta G_{solvation}$) by PBSA, or GBSA. The solute entropies (S_{solute}) are determined from N-mode frequencies, or via a quasi-harmonic approximation, or are sometimes neglected (Eq. 4.2).[17] Then, the calculated free energies of the unbound receptor and ligand are subtracted from the free energy of the ligand bound protein complex to determine the free energy of binding as described on Eq. 4.2. In the MD study described herein, the effect of different solvation models, entropy calculations, and charge models on accuracy and efficiency is evaluated.

$$\Delta G_{Binding} = G_{complex} - \left(G_{Protein} + G_{Ligand}\right) \qquad (4.1)$$

$$G_{System} = E_{gas} + \Delta G_{Solvation} - TS_{Solute} \qquad (4.2)$$

In terms of the SAMPL competitions, since SAMPL3 [27], quantum mechanical calculations have been used as valuable prediction methods for host–guest binding energies. QM methods have advantages in describing stationary systems with high accuracy, but typically address far fewer configurations than is possible with MD related approaches. From initial inputs which are often based on crystal structures, chemical experience, or MD poses, complexes are optimized to stationary points on potential energy surfaces, and calculation are done to obtain frequencies. The QM approaches involved in the prior SAMPL competitions, to a certain extent, reflect the development of QM methods throughout the years. In the SAMPL3 competition, the Lee–Yang–Parr three-parameter hybrid functional (B3LYP)[28–30] was employed in single-point calculations. Structure optimization and frequency calculations were performed with the MMFF94 force

field.[27] In SAMPL4[31], dispersion corrections[32–34] were included in the DFT calculations. The structures were optimized with density functional methods (TPSS-D3, PW6B95-D3) and the frequency calculations were carried out using a semi-empirical HF3c level.[35] In another submission of SAMPL4, local correlated coupled-cluster singles and doubles with perturbatively corrected triples (LCCSD(T)) was adopted on dispersion corrected functional, TPSS-D3, optimized structures. The frequency calculation calculations were carried out at a force field level.[36] In SAMPL5[37], the domain-based local pair-natural orbital coupled cluster singles and doubles with perturbatively corrected triples (DLPNO-CCSD(T))[38] method was used while the frequency calculations remain at HF3c level. In SAMPL6, our research group performed dispersion corrected DFT calculations (B3PW91-D3[28,33,39–41]) with the solvation model based on the density (SMD) model, with structures generated from MD simulations.[23,42] In the QM effort described here, a double-hybrid functional (B2PLYP-D3)[43,44] was considered. For the exoOA-G2 complex, the effect of geometry optimization and thermochemistry corrections in solvent was evaluated. Single-point calculations were performed on the resulting structure using the recently developed functional ωB97M-V, which has resulted in excellent predictions of binding energies in earlier studies.[45,46]

4.2 Methods

The initial structures of the eight guest molecules are shown in Figure. 4.1 and the two host molecules OA and exoOA are shown in Figure 4.2 (SAMPL7 GDCC dataset). The protonation state of both hosts and guests were determined by the Protonate3D module implemented in Molecular Operating Environment version 2019.01 (MOE).[47,48] Host–guest binding complexes were generated by the docking feature of MOE. The placement step of the docking was performed by the triangle matcher algorithm, and the refinement step performed by the induced

fit method which also accounts for the changes in the host structure upon binding.[48,49] The resulting host–guest binding poses with the highest generalized-Born volume integral/weighted surface area score (GBVI/WSA ΔG) were minimized with molecular mechanics using AMBER10: Extended Hückel Theory (EHT) force field implemented in MOE, which employs Amber ff10 and EHT bond parameters.[50–53] The minimized structures (Figures. 4.4, 4.5) are further investigated using MD and QM. More details about the MD and QM calculations are explained in the "Molecular dynamics protocol" and "Quantum mechanical methods" sections, respectively. The QM optimized structures are shown in Figure. 4.8.

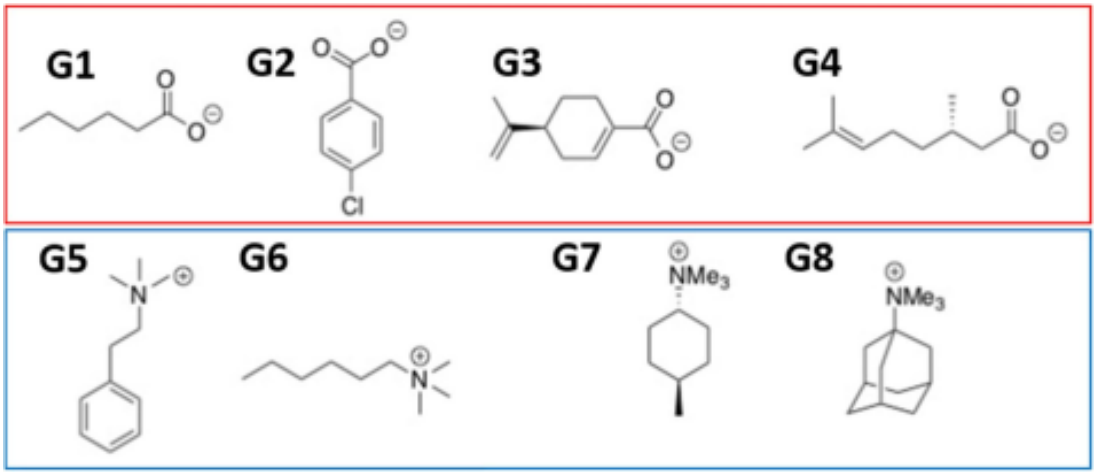


Figure 4. 1 Guest molecules in the SAMPL7 GDCC host–guest binding challenge. The binding of these eight guest molecules is considered for both OA and exoOA hosts.

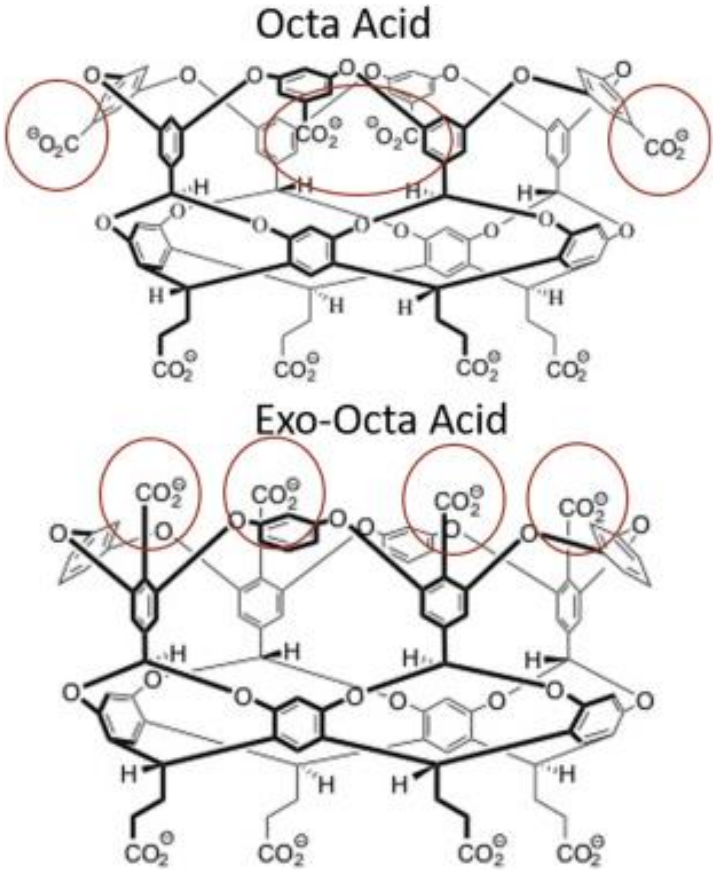


Figure 4. 2 The guest molecules for the octa-acid (OA) and tetra methyl octa-acid (TEMOA) hosts.

4.2.1 Molecular dynamics protocol

In order to assess the effect of the charge scheme on the accuracy of the binding predictions, partial charges of the host and guest atoms were calculated with two different methods: AM1-BCC and HF/6-31G* restrained electrostatic potential charges (RESP) and the resulting binding free energies are compared. The AM1-BCC charges are generated using the Antechamber module of Amber.[54] The RESP charges are generated using RED server.[55,56] The calculated partial charges were fitted using the general Amber force field (GAFF) to generate MD parameters for the host and the guest molecules.[52] The simulation box for each host–guest system was generated using the "leap" module of Amber tools.[57] Then, each system was

neutralized with sodium counterions with parameters from Joung and Cheatham.[58] After this step, the host–guest systems were solvated in a 14.0 Å cube of TIP4P-Ew water, which has been previously shown to be in good agreement with experiment, in terms of the ion hydration free energy, hydration radius, and coordination numbers.[59,60] Finally, additional sodium chloride ions were added to the simulation box in order to mimic the experimental ionic strength of the 10 mM sodium phosphate buffer.

The host–guest systems were minimized with decreasing energy restraints on the host molecules (500.0, 200.0, 20.0, 10.0, 5.0, 0.0 kcal mol^{-1}). Then, the systems were heated gradually to 300 K over 30 ps. After heating, 10 ns production simulations at 300 K and 1 atm pressure were performed. The production simulations were done in triplicate to account for randomized parameters that affect the MD trajectories such as initial velocities. During all simulations, the temperature was controlled by Langevin dynamics and the pressure was controlled by isotropic position scaling.[57,61,62] Nonbonded interactions were truncated with a 10.0 Å cutoff, whereas long-range electrostatics were handled with the particle-mesh Ewald (PME) method.[63] Bonds involving hydrogen were constrained using SHAKE, and the simulation time step was set to 2 fs.[64] All simulations were performed with AMBER18 and 500 snapshots are extracted from each of the production runs for further use in MMPBSA/MMGBSA calculations. The RMSD plots for the MD simulations can be found in Figures. 4.9–4.16.

4.2.2 MMPBSA/MMGBSA calculations

The binding free energies of the host-guest complexes were calculated using both MMPBSA and MMGBSA methods with a modified General Born solvation model by Onufriev et al.[65] to consider the effect of solvation models on the accuracy. MMPBSA and MMGBSA approaches are implemented in the Amber PBSA-solver. For all calculations, the default internal and

external dielectric constants were used (1.0 and 80.0, respectively), the solvent accessible surface area (SASA) was determined with the default Linear Combinations of Pairwise Overlaps (LCPO) method using modified Bondi atomic radii. For both MMPBSA and MMGBSA, the initial 500 snapshots of the MD simulations were used to calculate the binding energies. It has been shown by Hou et al., that such simulations are useful, and longer timeframes do not necessarily correspond to better accuracy in the calculated binding energies relative to the experimental binding energies.[66]

To consider how solute entropies affect the accuracy of the calculations, the solute entropies were determined using N-mode approximation and were compared to the neglected solute entropy results. To correct the calculated binding energies, the OA host–guest binding dataset from SAMPL6 was used. The binding energies of the SAMPL6 guests to OA were predicted with the RESP-MMPBSA method with the neglected solute entropies and the results are provided in Table 4.3. The SAMPL6 predictions were plotted against their experimental values to create a linear fitting curve (Figure 4.17). The fitting curve equation was used to correct SAMPL7 RESP-MMPBSA results.

4.2.3 Quantum Mechanical Methods

For QM calculations, docking poses were used as initial guess structures and the geometries of the host–guest systems were optimized using the Gaussian 16 software package.[67] Due to the constraints associated with the size of the system, DFT was employed for quantum mechanical calculations. The B3PW91 functional was chosen along with GD3BJ to describe the dispersion forces.[28,33,39–41] This method was also considered for SAMPL6.[23] For exoOA-G3, there were difficulties reaching convergence with B3PW91 in the time constraints of the competition. Thus, B97D was considered partnered with SMD, which did reach convergence.[39] cc-pVDZ basis sets

were used for all atoms in each of the complexes.[68] Frequencies were calculated for all geometry optimization steps, guaranteeing they were at a minimum on the potential energy surface. Note that while a double-ζ level basis set is not ideal for small molecules, because of the size of the host–guest systems, it is used here.

After this step, single-point energy calculations were carried out using B2PLYP-D3[43,44], which includes GD3BJ dispersion.[33] B2PLYP is a double-hybrid functional that includes Hartree–Fock exchange and MP2-like correlation, and has been shown previously to provide lower overall errors as compared with other DFT functionals in terms of long-, short-range, and side chain-side chain interactions.[69] Due to the size of the system, an MP2 correction was not considered. The non-double-hybrid part of B2PLYP includes gradient approximations of GGA methods with Becke exchange, Lee, Yang and Parr correlation, along with Hartree–Fock exchange.[43] To account for the role of the solvent (water), the SMD solvation model was employed.[42] The inclusion of an implicit solvation model was deemed essential to mimic the stabilization that water molecules have on the system and produce reliable binding energies.

Regarding basis sets, a set of double-, triple-, and quadruple-ζ correlation consistent basis sets were used for single-point calculations.[68,70,71] For oxygen and chlorine atoms, the augmented form of the basis sets was important, due to the negative charges located on the oxygens and electronegative nature of the chlorine.[70,71] As a modified version of the correlation consistent basis was recommended to replace the original correlation consistent basis sets for second-row atoms, for species that included chlorine, the tight-d forms of the basis sets (cc-pV(D+d)Z and cc-pV(T+d)Z[72]) were considered for OA-G2 and exoOA-G2 and the predictions have been included in Table 4.4.

Due to inaccuracies associated with the G2 guest binding predictions (described in the next section), an investigation of the influence of the solvent on molecular complexes was performed for the exoOA-G2 complex. Geometry optimizations and frequency calculations were conducted with B3LYP-D3 (GD3BJ)/6-31G*[28–30,73,74] in combination with the integral equation formalism variant of polarizable continuum model (IEF-PCM)[75] solvation model using Gaussian 16 for the complex and monomers. The reason for adopting the IEF-PCM model is that a converged structure was not obtained in geometry optimization that employed SMD. Thermochemistry corrections were carried out at 298.15 K and scaled by 0.96 for anharmonicity.[76] The standard state corrections were applied.[42,77] The SMD solvation method was combined with the conductor like PCM (C-PCM)[75,78] for single-point calculations with the ωB97M-V functional[45,46] in conjunction with a range of correlation consistent basis sets using ORCA 4.2.1.[79] An exponential form of a basis set extrapolation scheme to the complete basis set (CBS) limit—Kohn–Sham (KS) limit for DFT—was adopted[80] and the extrapolation exponent (5.46) was considered from Neese et al..[80]

In this combined approach, the Gibbs free energy is calculated as;

$$\Delta E = E_{complex} - E_{host} - E_{guest} \tag{4.3}$$

$$\Delta G = \Delta E^{wB97M-V/CBS/SMD} + \Delta G^{B3LYP-D3(GD3BJ)/6-31G*/IEF-PCM}_{scaled\ RRHO} + \Delta G^{0}_{gas/solute} \tag{4.4}$$

$$E^{X}_{SCF} = E^{CBS}_{SCF} + \mathrm{Aexp}(-\alpha\sqrt{x}) \tag{4.5}$$

Here RRHO represents the rigid–rotor harmonic-oscillator. In Eq. (4.3), ΔE stands for the difference of electronic energies between the complex, guest, and host. In Eq. (4.4), ΔG is the difference of Gibbs free energy corrections; $\Delta G^{B3LYP-D3(GD3BJ)/6-31G*/IEF-PCM}_{scaled\ RRHO}$ represents the thermochemistry corrections from B3LYP-D3 with a smaller basis set, 6-31G*, along the solvent correction, IEF-PCM. Since the electronic contribution is the leading term in molecular energy, it

is typically adopted to calculate the single-point energy with a higher-level method and the thermochemical corrections using a lower-level method $\Delta G^{o}_{gas\rightarrow solute}$ represents the -1.89 kcal mol^{-1} correction due to difference in the standard state in gas phase and solvent. Eq. (4.5) has been adopted to extrapolate to the CBS limit of HF energies. Similar convergence patterns were found for DFT energies.[81]

In prior work, it has been suggested that the solvation energy should be calculated at the level where the solvation model was parameterized.[77] This method led to less accurate energies than using the ωB97M-V functional[45,46] with SMD for the single-point calculations in Eq. (4). For instance, the binding energy from gas phase ωB97M-V/cc-pVDZ combined with the solvation energy from B3LYP/6- 31G* led to − 17.77 kcal mol^{-1}, which is ~ 5 kcal mol^{-1} lower than for ωB97M-V/cc-pVDZ with SMD directly, −12.69 kcal mol^{-1} in Table 4.2.

Since the spin-contamination from an unrestricted Hartree–Fock (UHF) wavefunction may indicate the inappropriateness of the ground state description[82], a stability analysis was performed with Gaussian 16 on the host exoOA.[74,83] To further consider possible multireference character, the complete active space self-consistent field (CASSCF)[84] method was used to calculate the partial occupation numbers in natural orbitals and conduct a T1 diagnosis[38,85,86] with ORCA 4.2.1.

4.3 Results

4.3.1 OA and exoOA Binding Cavities

The structures and binding cavities of OA and exoOA hosts for the SAMPL7 GDCC challenge are provided in Figures. 4.2 and 4.3, respectively. The hosts differ by four carboxylic acid groups which are located on the rim of the binding cavity. On the OA host the carboxylic

acid groups are placed further away from the cavity opening whereas on the exoOA structure, the carboxylic acids are located next to the opening.

4.3.2 Host Guest Binding Poses

The structures of the SAMPL7 GDCC guests, binding poses of guests on the OA host, and binding poses of guests on the exoOA host predicted by docking are provided in Figures. 4.1, 4.4, and 4.5, respectively. During the SAMPL7 GDCC challenge four negatively charged guest molecules (G1–G4) with carboxylic acid functional groups and four positively charged guest molecules with amino groups (G5–G8) were investigated. The docking results show that carboxylic acid and amino groups prefer to orient toward the opening of the cavity rather than deeper in the cavity.

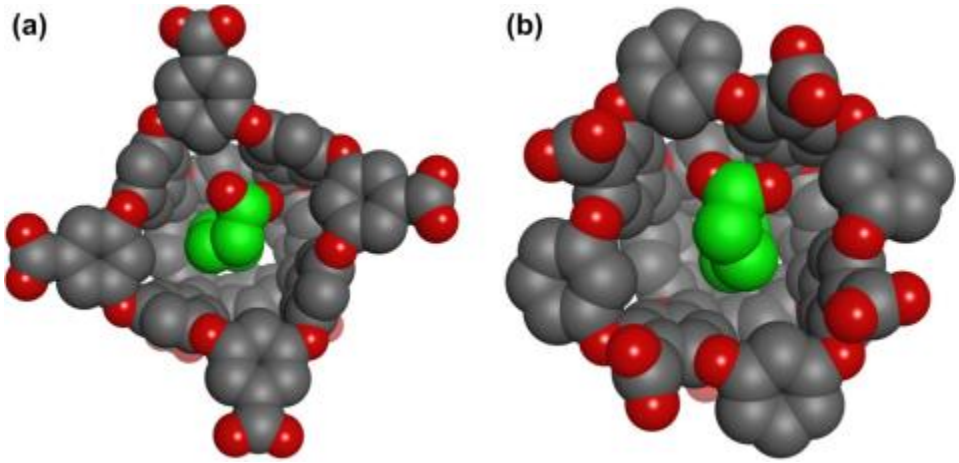


Figure 4. 3 **a** Binding cavity of OA together with G1 (shown in green). **b** Binding cavity of exoOA together with G1 (shown in green).

The binding free energies studied with different models and levels of theory as a part of SAMPL7-GDCC challenge are given in Tables 4.1 and 4.2. Table 4.1 includes the results obtained from MMPBSA and MMGBSA binding free energy calculations using the MD simulation frames. Table 4.2 contains predictions made by B2PLYP-D3, and ωB97M-V. To

assess the binding energies determined using each method with respect to the experimental values; root mean square errors (RMSE), mean absolute errors (MAE), mean errors (ME), r^2 correlation coefficients, slope of the correlation plots (m), and Kendall's Tau (τ) rank correlation coefficients, which measures how well the method ranks the binding free energy of the guest compounds with respect to experiment, are also included in Tables 4.1 and 4.2.

Table 4.1 shows MMPBSA/MMGBSA binding free energy predictions. In addition, RESP and AM1 partial charges were evaluated, and the influence of adding N-mode solute entropies are compared to the experimental values. For the ranked submission, SAMPL6 OA systems are used to perform a linear correction of the binding free energies calculated with RESP charges, the Poison–Boltzmann (PB) solvation model, and neglected entropies (Table 4.1, RESP-MMPBSA-Cor). The PB solvation model leads to smaller errors and better correlation when compared to the Generalized-Born (GB) solvation model with a RMSE of 8.66 and 11.43 kcal mol^{-1}, and r^2 of 0.70 and 0.51, respectively. When RESP predictions are considered, the binding free energy predictions have smaller errors and slightly better correlation as compared to AM1 charges with a RMSE of 8.66 and 10.67 kcal mol^{-1}, and r^2 of 0.70 and 0.63, respectively. Additionally, the effect of the N-mode solute entropy corrections on the binding free energy predictions are also assessed. In all other MMPBSA/MMGBSA calculations, the solute entropies are neglected, with the exception of the RESP-MMPBSA-Nmode calculations (Table 4.1). Within the RESP-MMPBSA-Nmode method, the solute entropies are calculated with an N-mode analysis of the harmonic frequencies. The binding energy prediction with the RESP-MMPBSA-Nmode of G5 to the OA and G1-G5 to the exoOA is positive, which suggests that these guests do not bind to their hosts.

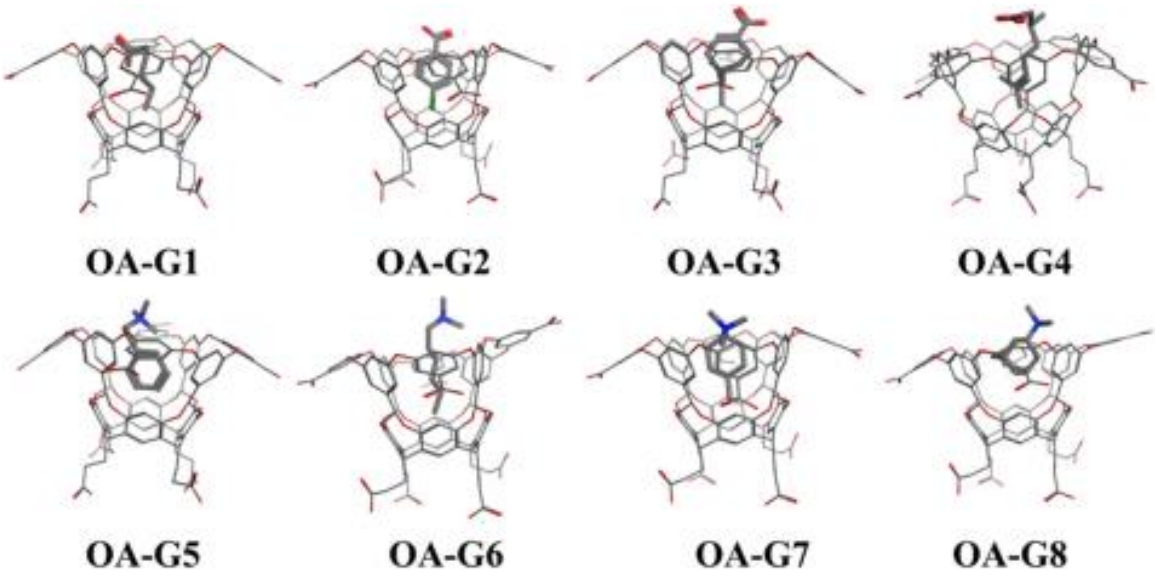


Figure 4. 4 Binding modes of guest to OA host generated with docking.

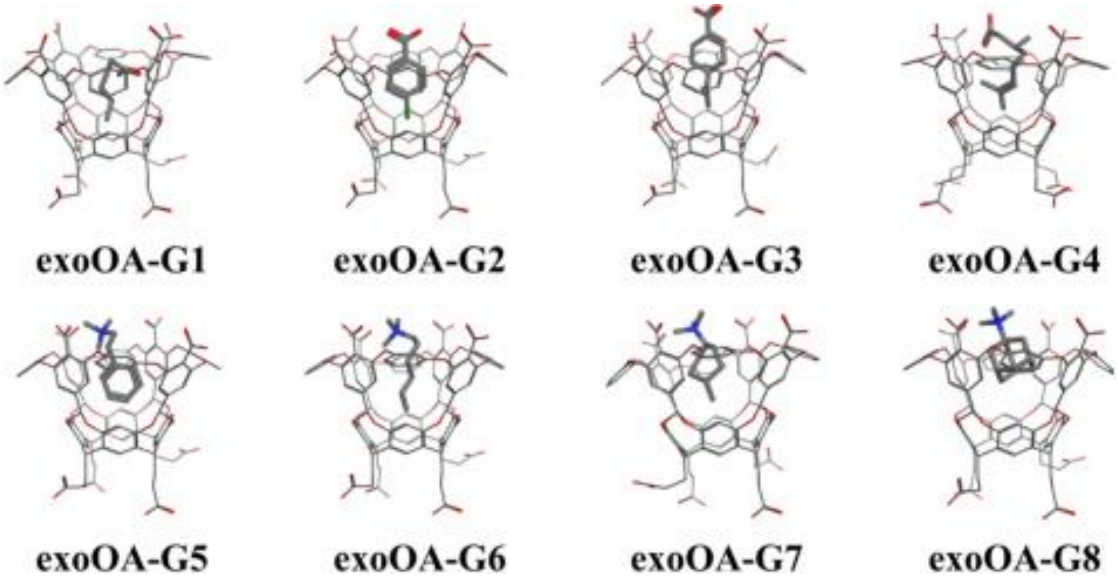


Figure 4. 5 Binding modes of guest to exoOA host generated with docking.

The experimental and predicted binding free energies, and the plot used during correction can be found in Figure 4.17. When the results obtained after linear correction are compared to the results without correction, both results have the same correlation with the experiment (r^2 =0.70, Figure 4.6). However, the linear correction shifts the predicted values and puts them closer to

experimental values, resulted in a decrease in the RMSE from 8.66 to 1.45 kcal mol^{-1} (Table 4.1).

Table 4.2 shows calculated binding energies using B2PLYP-D3. In addition, comparison of binding energies determined when the double-, triple-, and quadruple-ζ levels of basis sets were used, and a structural optimization in the solvent was considered. When the DZ, TZ and QZ predictions are compared to the experimental binding energies, little correlation is found (r^2 values are 0.25, 0.30 and 0.29, respectively). However, when the r^2 values of OA and exoOA systems are calculated separately, better correlation is obtained for exoOA predictions (see "Quantum mechanics" section in discussion). A gradual improvement is observed in the RMSE and MAE occurs as the basis set size is increased from DZ to QZ. The resulting RMSEs are 7.11, 6.70 and 3.92 kcal mol^{-1} and the MAE are 6.16, 4.84 and 3.92 kcal mol^{-1} for DZ, TZ and QZ respectively. The smallest deviation from the experimental binding energies was observed with QZ basis set evidenced by its lower RMSE and MAE compared to the others.

4.4 Discussion

4.4.1 Molecular Dynamics

Of the wide variety of available molecular dynamics methods, the MMPBSA and MMGBSA approaches are considered to be an intermediate option between semi-empirical docking scoring approaches and computationally more rigorous methods (i.e. free energy perturbation, replica exchange and thermodynamic integration). However, there are number of factors that can impact the performance of the MMPBSA/MMGBSA methods including atomic partial charge method, dielectric constant, force field, solvation model, and solute entropy correction method. In the current study, the impact of the solvation model, partial charges, and the N-mode solute entropy correction upon the utility of the method for predicting binding energies was investigated.

When comparing the ranked submissions of OA and exoOA host–guest complexes (RESP-MMPBSA-Cor) several trends can be noted. The exoOA host–guest systems have better correlation with experiment then the OS systems (r 2=0.95 vs 0.26 respectively). A similar correlation arises for QM calculations (see “QM discussion”). In terms of error analysis, for RESP-MMPBSA-Cor, OA and exoOA do not have significant differences. Considering the OA host–guest systems, the RMSE value is 1.55 and MAE is 1.28 kcal mol^{-1} respectively. For exoOA, RMSE is 1.32 and MAE 1.03 kcal mol^{-1}.

Table 4. 1 The binding free energies in kcal mol^{-1} for the OA and exoOA host–guest systems predicted from MMPBSA/MMGBSA.

Complex	Exp	RESP-MMPBSA	RESP-MMGBSA	AM1-MMPBSA	RESP-MMPBSA-Nmode	RESP-MMPBSA-Cor
OA-G1	− 4.97±0.02	− 13.01±0.04	− 14.26±0.04	− 12.55±0.05	− 1.41±0.04	− 6.18±0.04
OA-G2	− 6.91±0.02	− 12.38±0.04	− 12.90±0.04	− 11.66±0.04	− 1.49±0.04	− 5.90±0.04
OA-G3	− 8.10±0.05	− 17.34±0.05	− 17.09±0.04	− 18.34±0.05	− 2.79±0.05	− 8.13±0.05
OA-G4	− 6.76±0.05	− 18.49±0.05	− 18.81±0.04	− 18.55±0.05	− 3.80±0.05	− 8.65±0.05
OA-G5	− 4.73±0.02	− 14.03±0.06	− 17.15±0.07	− 15.84±0.06	0.28±0.06	− 6.64±0.06
OA-G6	− 4.97±0.02	− 16.56±0.05	− 18.91±0.07	− 18.36±0.05	− 3.21±0.05	− 7.78±0.05
OA-G7	− 6.07±0.05	− 15.57±0.05	− 21.55±0.08	− 19.94±0.05	− 1.81±0.05	− 7.33±0.05
OA-G8	− 8.25±0.02	− 17.89±0.04	− 20.91±0.05	− 21.70±0.05	− 4.71±0.04	− 8.36±0.04
exoOA-G1	0.00±0.00	− 7.84±0.08	− 11.54±0.06	− 8.80±0.06	4.37±0.08	− 3.84±0.08
exoOA-G2	− 1.31±0.02	− 7.60±0.07	− 10.25±0.06	− 7.68±0.06	3.48±0.07	− 3.73±0.07
exoOA-G3	− 3.37±0.05	− 10.62±0.08	− 12.65±0.07	− 12.66±0.10	4.57±0.08	− 5.10±0.08
exoOA-G4	− 3.61±0.05	− 10.32±0.12	− 13.91±0.08	− 10.21±0.11	4.52±0.12	− 4.96±0.12
exoOA-G5	− 5.57±0.02	− 12.41±0.07	− 16.26±0.06	− 14.40±0.07	1.78±0.07	− 5.91±0.07
exoOA-G6	− 5.83±0.02	− 14.88±0.07	− 17.91±0.07	− 18.06±0.06	− 1.14±0.07	− 7.02±0.07
exoOA-G7	− 6.98±0.05	− 14.64±0.05	− 20.53±0.07	− 18.44±0.06	− 1.33±0.05	− 6.91±0.05
exoOA-G8	− 7.67±0.02	− 16.54±0.05	− 19.82±0.05	− 21.02±0.06	− 3.61±0.05	− 7.77±0.05
RMSE		8.66	11.43	10.67	5.26	1.45
MAE		8.48	11.19	10.29	4.96	1.16
ME		8.48	11.19	10.29	− 4.96	1.02
r^2		0.7	0.51	0.63	0.68	0.7
m		1.36	1.27	1.74	1.3	0.61
τ		0.57	0.52	0.57	0.61	0.57

4.4.2 Comparison of Poisson Boltzmann and Generalized Born Solvation Models

The Poisson-Boltzmann (PB) model is a detailed description of the electrostatic environment of a solute in an ion containing solvent. On the other hand, the Generalized-Born model is built upon approximating the linearized PB model, to achieve a computationally less demanding solution for the solvation.[17,87] However, the predictions arising from MMPBSA and MMGBSA methods are system dependent, when compared to the experimental binding energies.[66] In this section, a comparison between RESP-MMPBSA and RESP-MMGBSA is performed (Table 4.1). Binding energies predicted using the PB solvation model were closer to the experimental values, and led to smaller RMSE, MAE and ME as compared to energies predicted using the GB model. Moreover, PB binding energies also showed better correlation with experimental energies, as demonstrated by higher r^2 as compared to GB binding energies (r^2 =0.70 and 0.51, respectively for PB and GB). Finally, the PB solvation model performed slightly better in the correct ranking of host–guest systems relative to their experimental binding energies as compared to the ranking provided by GB, as evidenced by the higher τ of the PB model (τ=0.57 and 0.52, respectively for PB and GB). Overall, the results demonstrated the superiority of the PB model relative to the GB model in the predictions of the binding energies of the SAMPL7-GDCC dataset.

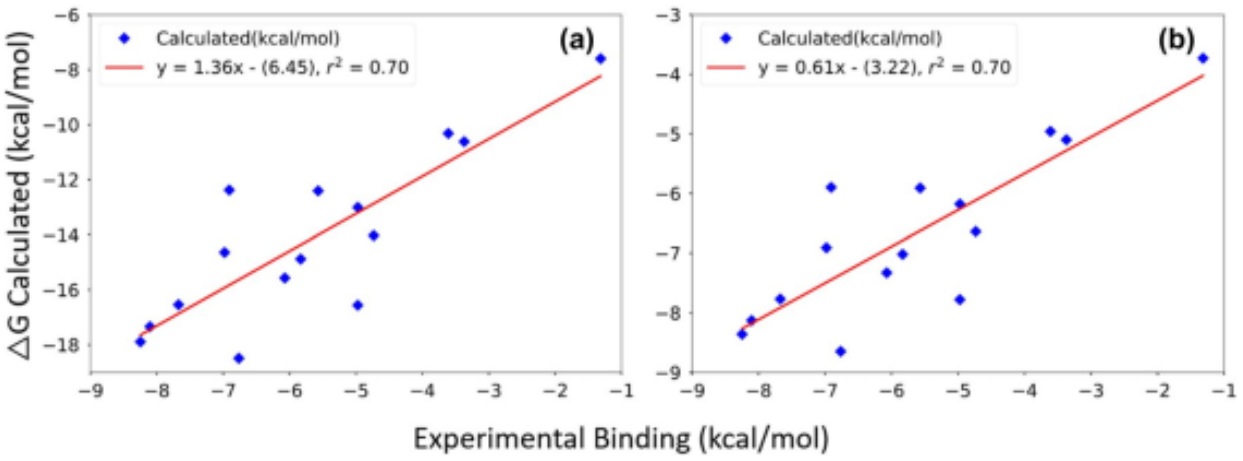


Figure 4. 6 **a** MMPBSA-RESP correlation with experiment. **b** MMPBSA-RESP correlation with experiment after linear correction. The linear correction shifted the y-values (△G Calculated) closer to the x-values (experimental) without changing the correlation coefficient (r^2).

4.4.3 Comparison of RESP and AM1 charges

Both RESP and AM1-bcc charges were used during MMPBSA/MMGBSA calculations. Among the two, AM1-bcc is parameterized to generate atomic charges efficiently that emulate the HF/6-31G* electrostatic potential (RESP), and the charge generation is fully automatized on Amber tools.[54,57] However, the calculation of RESP charges requires additional steps, including the extraction of electrostatic potential from GAMESS or Gaussian output files, though it results in more accurate charges. To understand the impact of the RESP and AM1 charge models, both methods were examined in this SAMPL7-GDCC challenge, and the binding energy predictions are provided in the RESP-MMPBSA and AM1-MMPBSA columns in Table 4.1. In general, using RESP charges resulted in the prediction of binding energies that are closer to the experimental values and resulted in smaller RMSE, MAE and ME as compared to those arising from the use of AM1- bcc charges. Additionally, binding energies predicted with RESP-charges

resulted in slightly better correlation with experimental values compared to the AM1-bcc prediction (r^2=0.70 and 0.63, respectively for RESP and AM1-bcc). However, the two methods showed the same performance with respect to ranking the binding energies of host–guest systems (τ = 0.57 for both RESP and AM1-bcc results). Overall, RESP charges quantitatively worked better, but qualitatively, the two charge methods resulted in similar predictions. The complexity and computational demand of obtaining RESP charges are higher as compared to obtaining AM1-bcc charges, so using the latter during MMPBSA/MMGBSA calculations might be advantageous.

4.4.4 Solute Entropies

For the prediction of absolute binding energies using MMPBSA/MMGBSA methods, solute entropies of the ligand and the target in the bound and unbound states were calculated. Among the methods available, a normal mode analysis of harmonic frequencies (N-mode) from minimized snapshots of MD frames is commonly used. However, N-mode calculations are demanding with respect to computing time and memory. Due to these constraints, N-mode calculations can only be performed on a few snapshots of the MD simulation, which limits the possible conformational space that can be studied. On similar systems with respect to size and complexity (i.e. binding to the same protein), the solute entropy contribution to the binding is considered to be similar. For this reason, solute entropies are commonly neglected when relative binding energies to the same, or similar targets are studied. To understand the difference between binding energy predictions both methods were considered: When the solute entropies were neglected or calculated through N-mode (RESP-MMPBSA and RESP-MMPBSA-Nmode columns of Table 4.1, respectively). The results showed that N-mode analysis overestimated the

solute entropy difference between the bound and unbound systems and led to unfeasible binding energies for G5 binding to the OA and G1–G5 binding energies for the exoOA.

Due to approximations used within the MMPBSA/MMGBSA methodology, and the lack of a fast method of calculating accurate solute entropies, MMPBSA/MMGBSA commonly overestimate binding energies, even though the predictions are qualitatively correct. In order to improve the quantitative predictions from MMPBSA/ MMGBSA, linear corrections from similar systems are typically used (Figure 4.6b). In our ranked submission for the SAMPL7-GDCC challenge, a linear correction was also beneficial. Due to the structural resemblance between OA and exoOA, the linear fitting curve obtained from SAMPL6 OA systems also improved the exoOA predictions. Correlation plots with and without correction are provided in Figure 4.6. The linear correction shifted the predicted results closer to the experimental values without changing the correlation, or the binding affinity ranking of the host–guest systems. In other words, even though the correction improved MMPBSA/MMGBSA results quantitatively, and it brought the predictions closer to their absolute values, it did not change the quality of the predictions.

4.4.5 Quantum Mechanics

The binding energies submitted for SAMPL7 using QM methods are shown in Table 4.2 for the host–guest systems. As mentioned previously in the “Results” section, the correlation in the binding energies for exoOA in comparison with experiment was better than for OA. In addition, for OA and exoOA, the binding energy of the anions (G1–G4 ligands) was nearer that of experiment than for the cations (G5–G8 ligands). The only exception was for the binding of the G2 ligand. Quadruple-level basis sets improved the accuracy for most guest–host systems, but for the G2 ligand there were Journal of Computer-Aided Molecular Design 1 3 still some discrepancy with respect to experimental results. To address this issue, a number of methods

were considered for both the structural and energetic predictions (see "Comparison of gas phase and solvated structures").

Table 4. 2 Calculated binding energies using B2PLYP-D3 vs experimental binding energies, using a range of basis sets. The geometry was optimized in the gas phase. Values shown are in kcal mol^{-1}.

Complex	Exp	B2PLYP-D3		
		DZ	TZ	QZ
OA-G1	− 4.97±0.02	0.4	3.58	4.16
OA-G2	− 6.91±0.02	− 7.29	− 25.24	− 26.83[− 7.29][a]
OA-G3	− 8.10±0.05	− 11.39	− 7.95	− 7.21
OA-G4	− 6.76±0.05	− 11.33	− 7.92	− 7.20
OA-G5	− 4.73±0.02	− 10.56	− 6.34	− 5.52
OA-G6	− 4.97±0.02	− 13.34	− 9.44	− 8.86
OA-G7	− 6.07±0.05	− 15.57	− 11.61	− 11.04
OA-G8	− 8.25±0.02	− 11.29	− 7.82	− 7.19
exoOA-G1	0.00±0.00	2.66	5.79	6.41
exoOA-G2	− 1.31±0.02	− 2.16	1.4	1.94
		− 12.69[b]	− 5.47[b]	− 2.01[b][− 0.96][c]
exoOA-G3	− 3.37±0.05	− 10.37	− 6.23	− 5.40
exoOA-G4	− 3.61±0.05	− 6.97	− 2.89	− 2.15
exoOA-G5	− 5.57±0.02	− 14.89	− 11.13	− 10.32
exoOA-G6	− 5.83±0.02	− 17.87	− 14.08	− 13.49
exoOA-G7	− 6.98±0.05	− 19.43	− 15.65	− 15.04
exoOA-G8	− 7.67±0.02	− 14.52	− 10.90	− 10.23
RMSE		7.11	6.7	3.92
MAE		6.16	4.84	3
ME		5.44	3.09	1.84
r^2		0.25	0.3	0.29
m		1.41	2	1.17
τ		0.33	0.38	0.35

[a] Parenthesis value indicate calculation done with DZ basis set and used as ranked submission.
[b] Value indicate calculated binding energies using ωB97M-V. The geometry optimized in a solvated environment.
[c] Parenthesis value indicate calculated binding energies using ωB97M-V extrapolated to CBS.

4.4.6 OA Discussion of Results

The ranked submission results were performed using a quadruple-ζ level basis set (cc-pVQZ (C, N, H) and aug-cc-pVQZ (O, Cl). The only exception was for OA-G2, where a double-ζ level basis set was used. The quadruple-ζ value for the binding energy was − 26.83 kcal mol^{-1}, which deviates from experiment. For OA-G2, the chlorine atom seems to present a challenge, as the binding energy predictions deviate significantly from experiment when considering any level basis set. Though the tight-d basis sets are the recommended sets for chlorine, the cc-pV(D+d)Z and cc-pV(T+d) Z sets also resulted in similar deviations, and did not resolve the differences (Table 4.4).

In addition, to the quadruple-ζ level binding energies, double, and triple-ζ quality results were also submitted as non-ranked. Last year's submission from our research group included single-point calculations using B3PW91-D3, which resulted in a large overestimation of the binding energy. Single-point calculations with B3PW91-D3 were performed for SAMPL7 as a check, and the same outcome was observed, so these results were not included in the non-ranked submission. The RMSE and MAE for the octa-acid are 2.99 kcal mol^{-1} and 2.22 kcal mol^{-1} respectively, which means binding energies close to experiment were obtained (Table 4.5).

The G1 guest performs differently for OA and exoOA. B2PLYP-D3 predicts that the complex OA-G1 does not form. Increasing the basis set size from double-ζ to quadruple-ζ, the bonding energy increases from 0.40 to 4.16 kcal mol^{-1}. The difference in chemical structure between G1 and G4 is small, however, they have starkly different binding patterns. For OA-G4, the best submitted results are 0.44 kcal mol^{-1} different from experiment, while for OA-G1, B2PLYP-D3 predicts a non-binding interaction.

From OA-G3 to OA-G8, B2PLYP-D3/QZ the binding energy predictions are within~1 kcal mol^{-1} agreement from experimental measurements, with the exception of OA-G6 and OA-G7. The guests G5 to G8 are positively charged and affect the binding energies differently. OA-G5 and OA-G8 are very close to experiment, but OA-G6 and OA-G7 overestimate the binding energy by~4–5 kcal mol^{-1}.

4.4.7 exoOA Discussion of Results

For the ranked submission which entailed the use of quadruple-ζ level basis sets (cc-pVQZ (C, N, H) and aug-cc-pVQZ (O, Cl)), exoOA complexes have a higher RMSD and MAE than for the OA complexes (4.76 and 3.90 kcal mol^{-1} compared to 2.99 and 2.22 kcal mol^{-1}, respectively (Table 4.5)). In terms of correlation (r^2 and Kendall's Tau) the values are quite different for OA and exoOA. For exoOA, the r^2 and Kendall's Tau values are 0.72 and 0.58. For OA, the correlation is much less significant (r^2 =0.09 and Kendall's Tau=0.076).

For exoOA-G1, B2PLYP-D3/cc-pVXZ correctly predicted a positive binding energy, indicating that the complex does not form. It is interesting to note the difference in binding energies for OA-G2 and exoOA-G2. The binding energy prediction for OA-G2 led to large negative values with a quadruple-ζ basis set, but the exoOA-G2 prediction indicates that the complex does not form. For exoOA-G3 and exoOA-G4, the difference between the calculated results at the QZ level is~2 kcal mol^{-1}. Similar to OA-G5 to G8, B2PLYP-D3 overestimates the predicted binding energy compared to the experiment; guests that are positively charged overshoot the binding energy (~6–7 kcal mol^{-1}).

Since the predicted binding energies of OA-G2 and exoOA-G2 present large deviations from the experimental values, further investigations were performed. The following results were not submitted to the competition, but they are included in the present work to provide additional

insight about what is (and is not) needed to describe these systems, or similar systems, in the future. ExoOA-G2 was optimized at the B3LYP-D3 (GD3BJ)/6-31G* level with the IEF-PCM solvent model for water. The performance of B3PW91-D3 and B3LYP-D3 led to similar binding energies as compared to the reference values provided by CCSD(T)/complete basis set (CBS) limit from Mardirossian et al..[46]

4.4.8 Comparison of Gas Phase and Solvated Structures

The comparison between the structure in the gas phase and solvent is shown in Figure 4.7. In the solvent, the host slightly bends to the guest. This difference may come from a competition between electrostatic repulsion and dispersion interaction. The electrostatic repulsion between the negative charges in the host and guest may be screened by the dielectric constant of the solvent. The dispersion attraction may be less influenced by the solvent, since it does not come from the net charge. The G2 guest in the gas phase has C2v symmetry. In the solvent, the optimization of the geometry within the same symmetry leads to an imaginary frequency of 45i cm^{-1}. The local minimum is found to be of C2 symmetry. For the exoOA host, it was noticed that both in gas phase and solvent (water), the C1 structure has a lower Gibbs free energy than the C4 isomer (-0.9 and -2.3 kcal mol^{-1}, respectively). This difference may arise because it is a large system that can have near-fat potential surfaces. In addition, in the solvent optimized hosts, both C4 and C1 have small. imaginary frequencies (7i and 15i for C4 and C1, respectively). Similar results have been reported by Grimme et al. in host–guest complexes.[88]

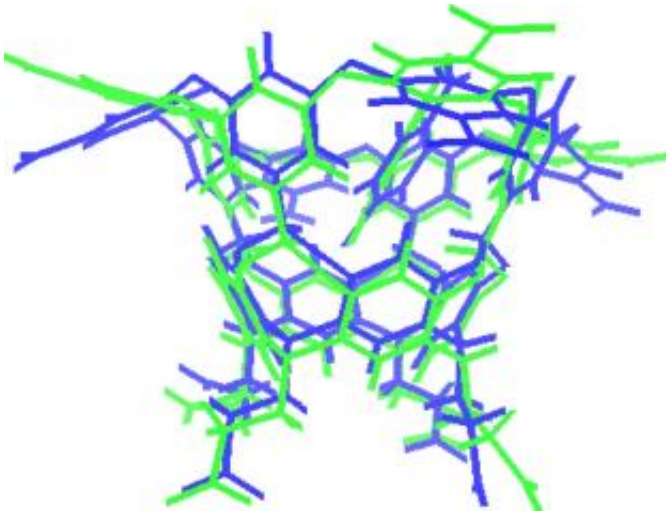


Figure 4. 7 Comparison between gas-phase (green) and solvent (blue) optimized structures of exoOA-G2.

Additionally, the exoOA host presented symmetry breaking solutions at HF levels, though not with the present functional B3LYP. An instability analysis at the HF/ cc-pVDZ level on the gas-phase optimized structure leads to an ⟨ S2 ⟩ =5.18. This large value of spin contamination may indicate the need to account for non-dynamical correlation.[82] To further consider possible multireference character, a CASSCF(6,6)/STO-3G calculation for the occupation number and DLPNO-CCSD/def2-SVP for the T1 diagnostic value were done.[89] (Though the STO-3G basis set is far too small for reasonable calculations, it is used here just to provide a quick, approximate assessment in regards to potential need to account for non-dynamical correlation.) The active space in the CASSCF calculation included the long pairs of negatively charged oxygen atoms and anti-bonding orbitals of benzene rings as a starting point, along with 6 electrons (6,6). The same active space has been adopted from symmetry reasons in a fullerene system. The occupation numbers (1.94 and 0.05), and T1 value (0.014) point to less multireference nature than suggested by instability analysis (⟨S2 ⟩ =5.18). This may be an example of the artificial symmetry breaking of spin state proposed by Head-Gordon et al..[90,91] It

has been found for this type of system, single-reference methods can provide reasonable energies as compared to those from experimental data and from multi-reference calculations. Hence, it is expected that the present density functional approaches provide a reasonable choice for the present systems.

4.5 Conclusion

In the SAMPL7 competition, MD and QM simulations were performed to predict the binding free energies of host–guest systems. In this MD study, MMPBSA/MMGBSA approaches were used, and the effects of PB and GB solvation models, RESP, AM1-bcc partial charges, and N-mode solute entropy were considered to determine the best route for the prediction of binding energies. Simulations with the PB solvation models led to better agreement with experiment than GB solvation models, which resulted in lower RMSE values and higher correlation coefficients. The comparison between the two charge methods showed that RESP charges led to quantitatively slightly better results with a lower RMSE value. However, r^2 and τ values for the predictions made with RESP and AM1-bcc charges were similar. As the complexity and the cost required for obtaining RESP charges were also considered, using AM1-bcc charges may be advantageous for systems of increased size. Comparison of the binding energy predictions with and without N-mode solute entropies showed that, N-mode calculations overestimate the solute entropy difference, and may led to unfeasible binding energies. In contrast, qualitatively and quantitatively better results can be obtained by using neglected solute entropies with a correction on the predicted results using a similar dataset.

For QM simulations, two strategies were adopted to compute the binding free energy:

(i) Using the non-double-hybrid part of the functional B2PLYP-D3 with the SMD implicit solvent model, with single-point energy differences (complex–host– guest) as

the final values. Our predictions yield substantially higher accuracy than SAMPL6 QM predictions.

(ii) For the exoOA-G2 system, we performed a combined approach including geometry optimizations, frequency calculations within solvent models for the guest, host, and complex, scaled thermochemistry corrections, the standard state correction, and basis set extrapolation with the recently developed functional ωB97M-V. The prediction agrees well with the experimental value.

DFT studies were performed on SAMPL7 guest–host binding systems. Since the host system presents instabilities for a spin restricted Hartree–Fock wavefunction, determining binding energies from orbital optimizations in correlated approaches may be considered[92,93], especially as local correlated coupled cluster methods have been adopted in SAMPL4 and 5 host-guest systems[36,37], for which chemical accuracy (1–2 kcal mol^{-1}) has not been reached. Considering the thermochemical correction, it has been suggested to adopt rotation-type formalism for low vibrational frequency contributions.[94,95] Li et al., showed that low-vibration corrections led to better agreement with experimental data.[95] However, scaling factors have not been explored or suggested on this type of approaches which may be a future interest. Moreover, the MP2 part of B2PLYP-D3 approach may be evaluated with density-fitting, incorporating additional correlation, in order to improve predictions.

In summary, the routes investigated in this study provided better results than in our previous SAMPL6 efforts for both MD and QM simulations. In general, the exoOA host–guest systems correlated better with experiment in comparison to OA. Considering the MD study, using the RESP partial charges, which were not used for SAMPL6, along with a linear correction led to better correlation and accuracy for the SAMPL7 approach. In addition, the inclusion of a linear ft

correction, yielded very accurate predictions, however the approach is limited to the availability of similar types of structures.

For QM predictions, higher accuracy for the binding energies can be achieved with a number of approaches. A geometry optimization was initially performed from generated poses, which guaranteed a minimum at the potential energy surface. The additional single-point calculations at B2PLYP-D3 level rendered more accurate binding energies than our SAMPL6 approach, and did not overestimate the binding energies. On the other hand, low correlation was obtained for the OA-systems. The binding energy obtained for the G2 ligand for OA and exoOA can be attributed the susceptibility of the DFT functional B2PLYP to strongly electronegative atoms, such as chlorine. The utility of the newly considered approaches (B2PLYP-D3 and combined method), should be examined for a broader range of systems. Vibrational corrections and explicit solvation models could also be considered.

APPENDIX

Table 4. 3 SAMPL6-OA host guest binding data used during linear correction. Units are in in kcal mol^{-1}.

	Guest	Predicted Binding	Experimental Binding
SAMPL 6	**G1**	-12.60	-5.68
	G2	-10.98	-4.65
	G3	-17.18	-8.38
	G4	-9.93	-5.18
	G5	-17.32	-7.11
	G6	-8.67	-4.59
	G7	-10.49	-4.97
	G8	-11.02	-6.22

Table 4. 4 Calculated binding energies using B2PLYP-D3 vs experimental binding energies, using cc-pV(D+*d*)Z and cc-pV(T+*d*)Z. The geometry was optimized in the gas phase. Values shown are in kcal mol^{-1}.

Complex	**Experimental Binding**	**B2PLYP-D3**	
		DZ	**TZ**
OA-G2	-4.97 ± 0.02	-7.47	-25.33
exoOA-G2	-1.31 ± 0.02	-2.21	1.16

Table 4. 5 Root mean square errors (RMSE), mean absolute errors (MAE), mean errors (ME), r^2 correlation coefficients, slope of the correlation plots (m), and Kendall's Tau (τ) rank correlation coefficients for OA and exoOA for the ranked submission. Values shown are in kcal mol^{-1}.

	OA	**exoOA**
RMSE	2.99	4.76
MAE	2.22	3.90
ME	0.39	3.50
$\boldsymbol{r^2}$	0.093	0.72
m	0.72	1.97
τ	0.077	0.59

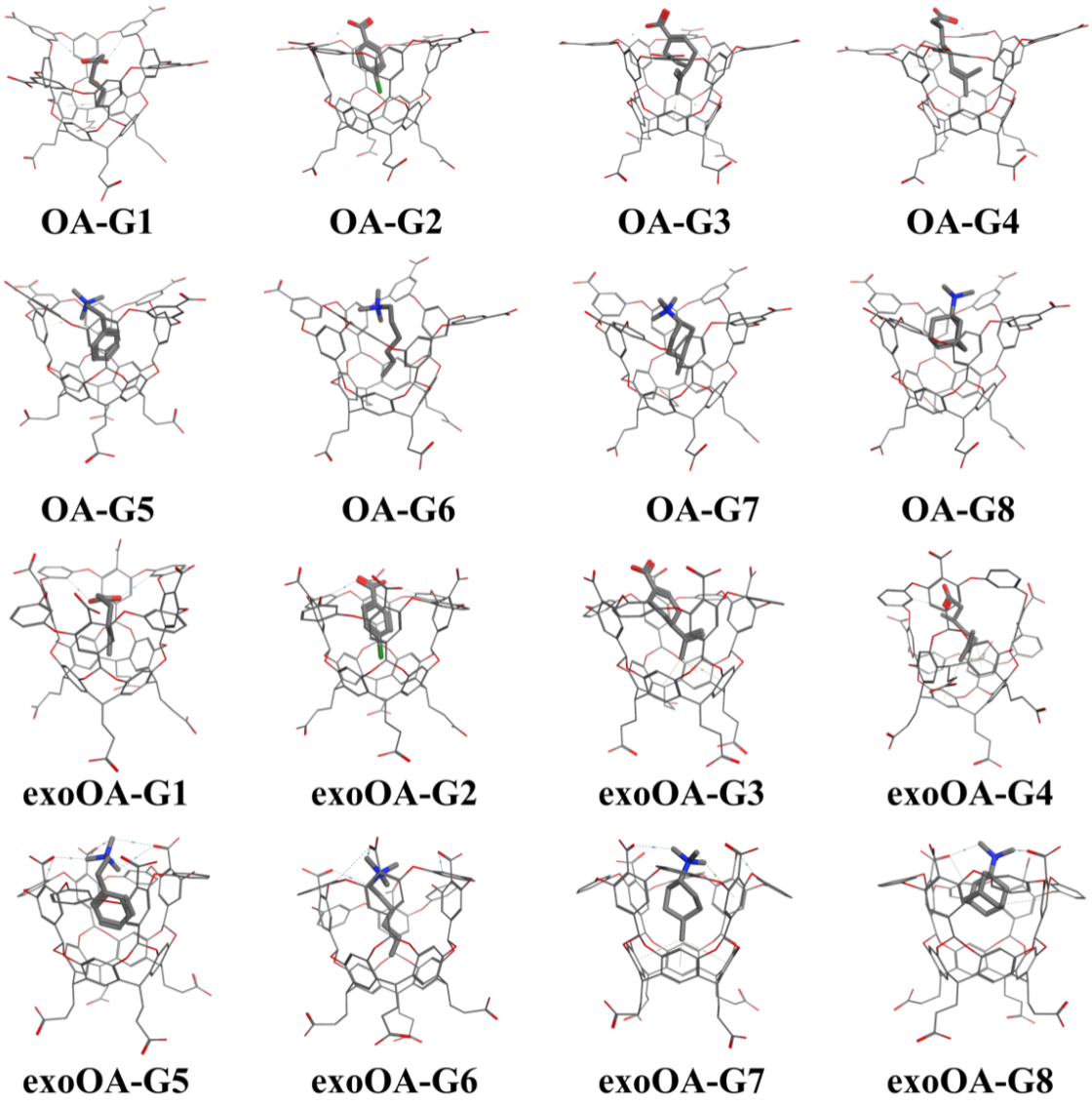


Figure 4. 8 Geometry optimized structures of OA and exoOA host/guess with B3PW91-D3/cc-pVDZ.

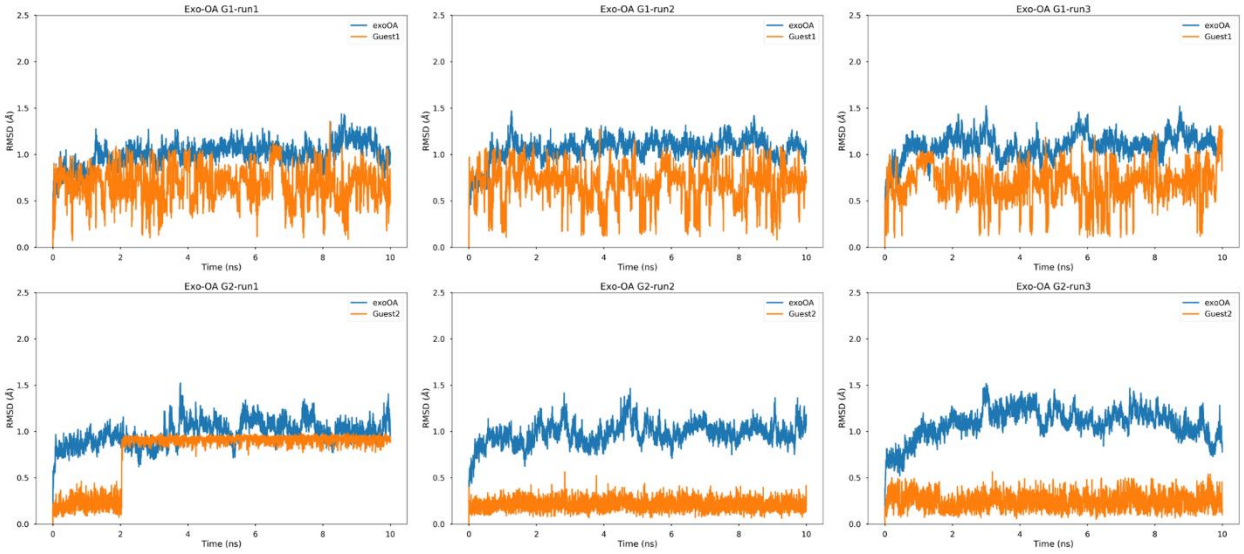


Figure 4. 9 RMSD plots of exoOA-G1 and exoOA-G2 MD simulations.

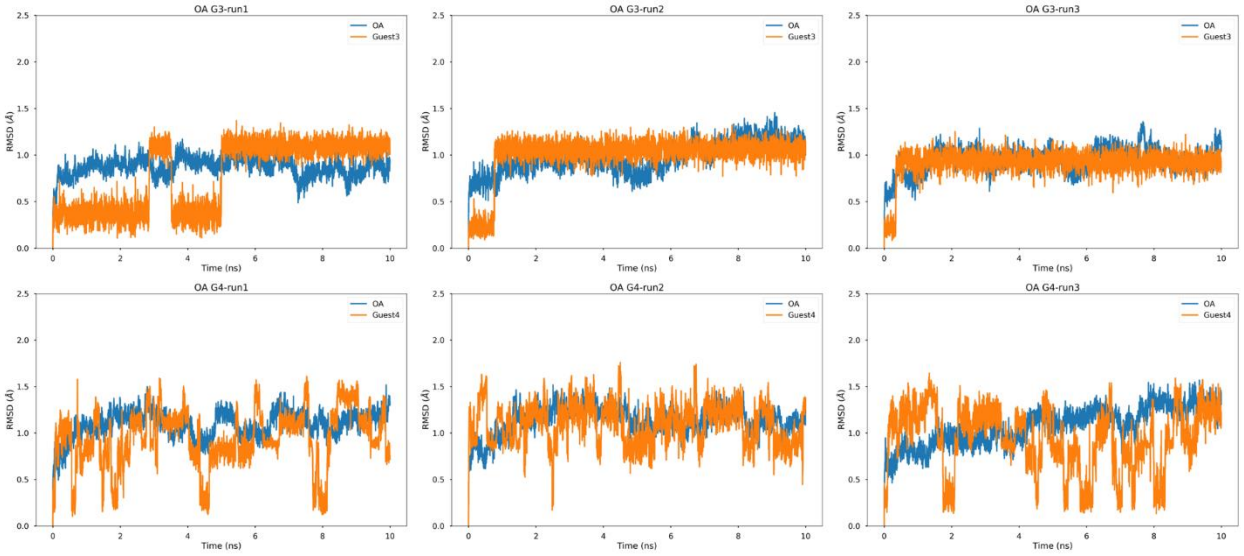


Figure 4. 10 RMSD plots of exoOA-G3 and exoOA-G4 MD simulations.

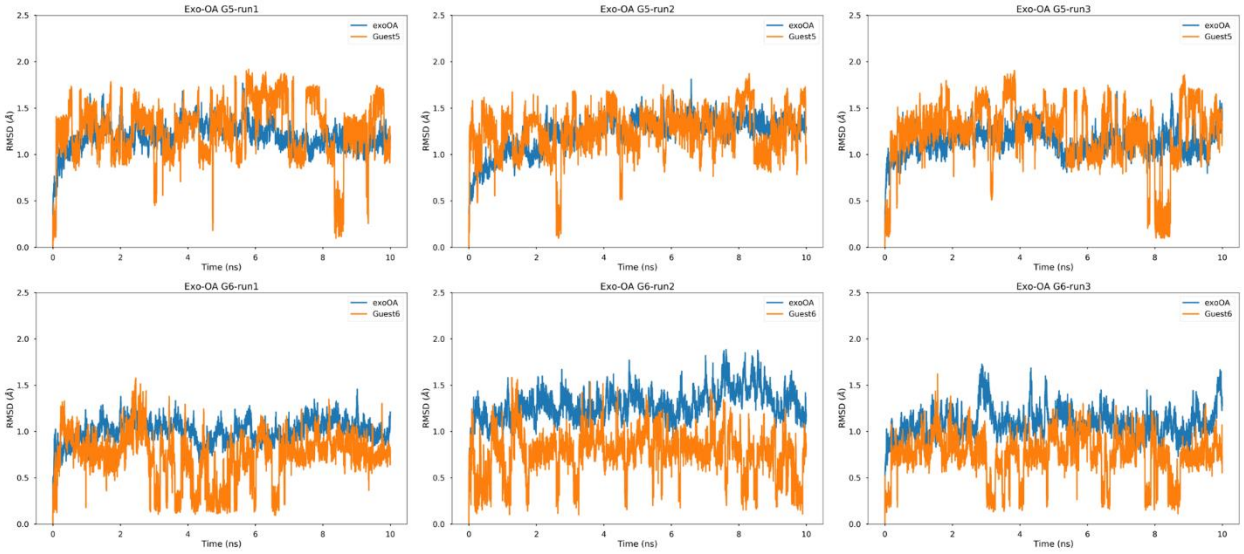


Figure 4. 11 RMSD plots of exoOA-G5 and exoOA-G6 MD simulations.

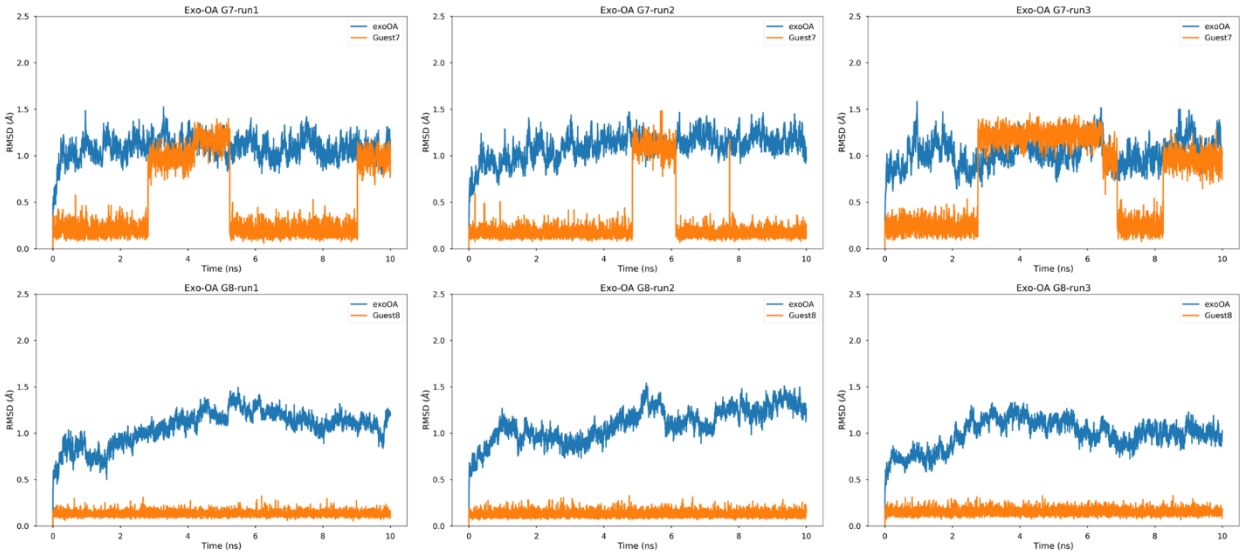


Figure 4. 12 RMSD plots of exoOA-G7 and exoOA-G8 MD simulations.

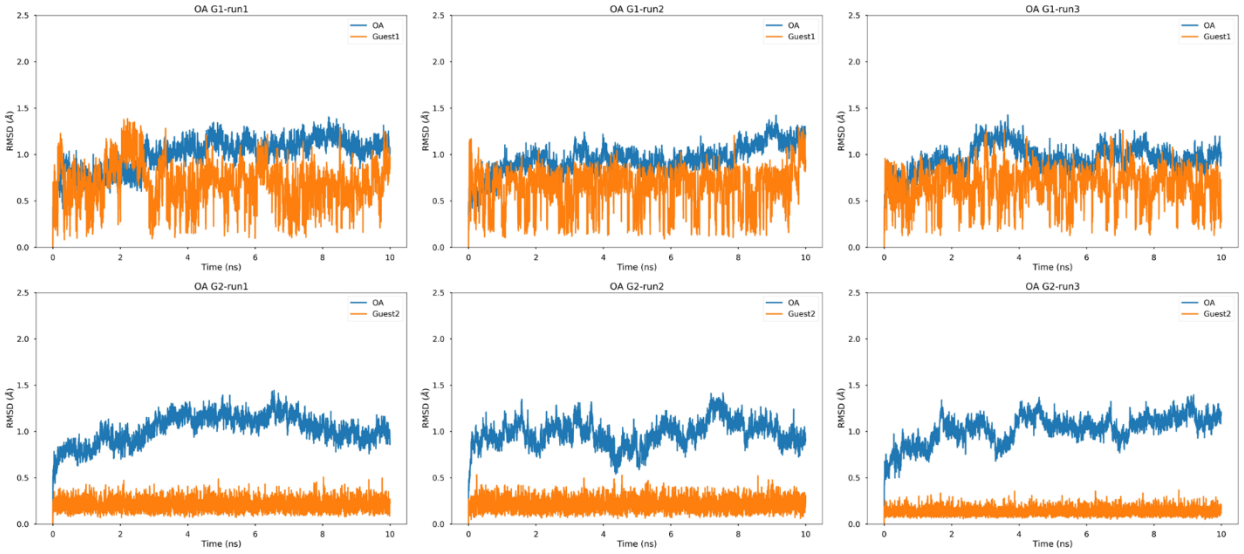


Figure 4. 13 RMSD plots of OA-G1 and OA-G2 MD simulations.

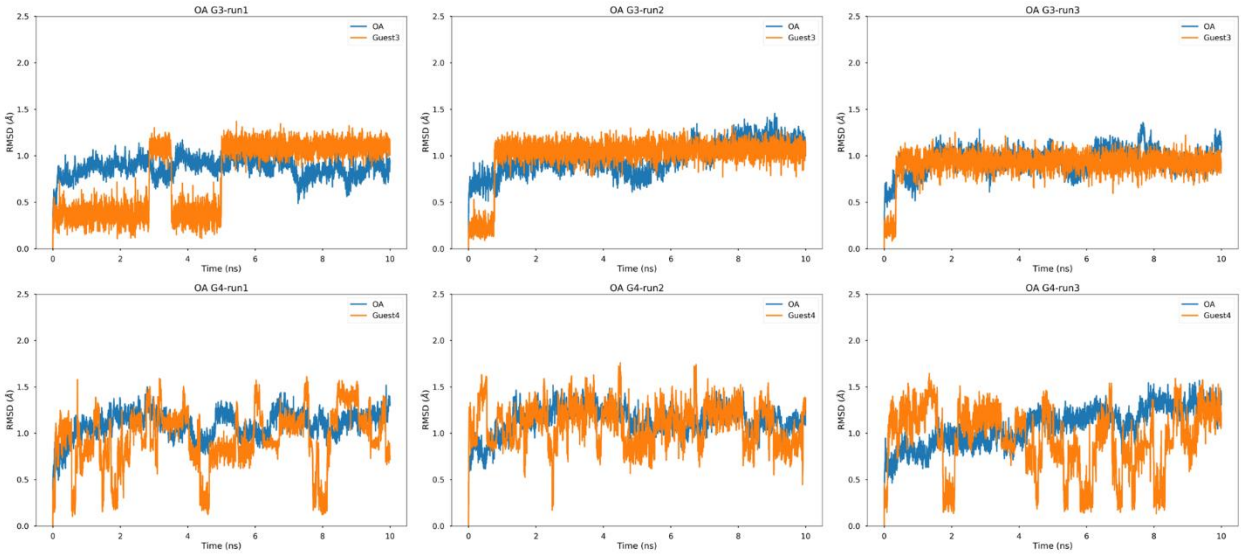


Figure 4. 14 RMSD plots of OA-G3 and OA-G4 MD simulations.

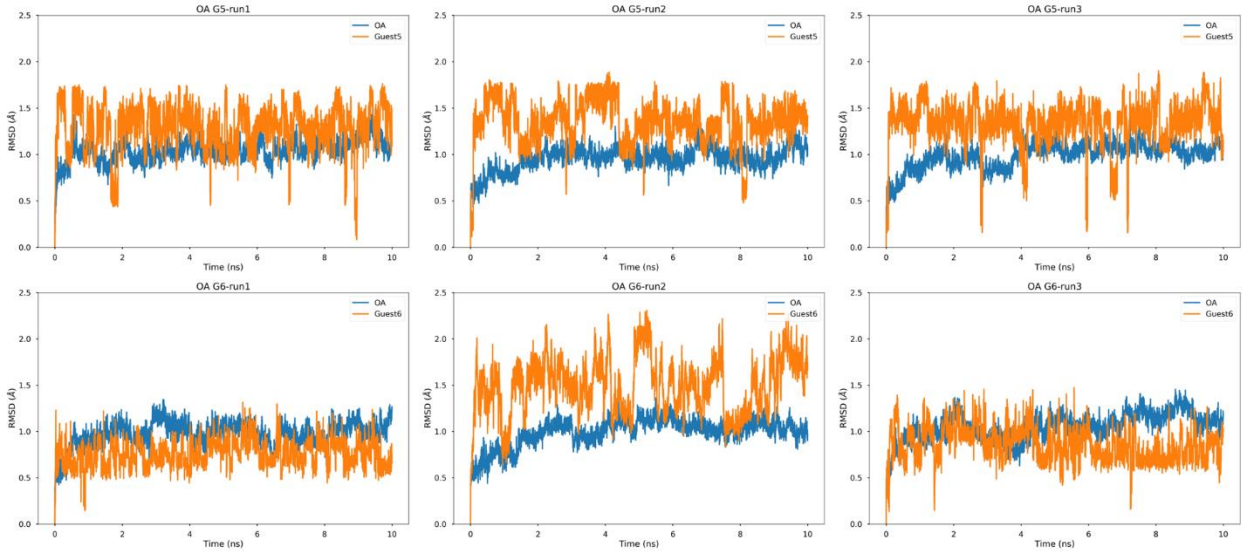


Figure 4. 15 RMSD plots of OA-G5 and OA-G6 MD simulations.

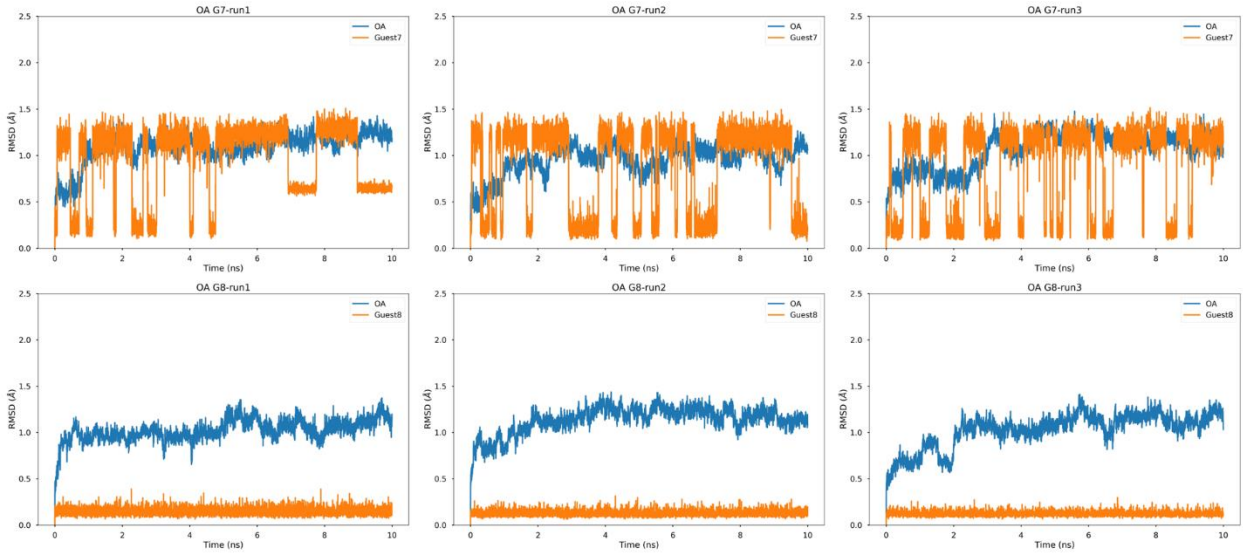


Figure 4. 16 RMSD plots of OA-G7 and OA-G8 MD simulations.

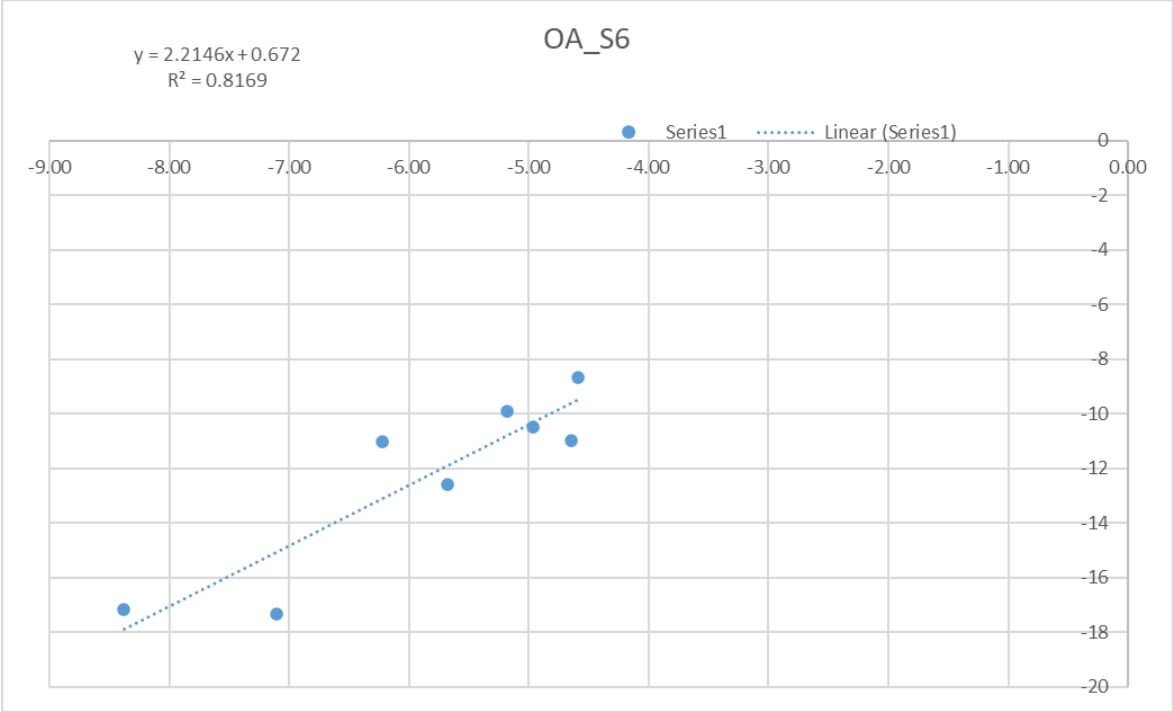


Figure 4. 17 Correlation plot of SAMPL6-OA host-guest binding. The x-axis provides the experimental binding energies and the y-axis contains binding energies predicted by RESP-MMPBSA method without solute entropies. A trendline equation is used to correct the predicted SAMPL7 binding energies.

REFERENCES

REFERENCES

(1) Alonso, H.; Bliznyuk, A. A.; Gready, J. E. Combining Docking and Molecular Dynamic Simulations in Drug Design. *Med. Res. Rev.* **2006**, *26* (5), 531–568. https://doi.org/10.1002/med.20067.

(2) Brown, F. K.; Sherer, E. C.; Johnson, S. A.; Holloway, M. K.; Sherborne, B. S. The Evolution of Drug Design at Merck Research Laboratories. *J. Comput. Aided. Mol. Des.* **2017**, *31* (3), 255–266. https://doi.org/10.1007/s10822-016-9993-1.

(3) Cerchietti, L. C.; Ghetu, A. F.; Zhu, X.; Da Silva, G. F.; Zhong, S.; Matthews, M.; Bunting, K. L.; Polo, J. M.; Farès, C.; Arrowsmith, C. H.; Yang, S. N.; Garcia, M.; Coop, A.; MacKerell, A. D.; Privé, G. G.; Melnick, A. A Small-Molecule Inhibitor of BCL6 Kills DLBCL Cells In Vitro and In Vivo. *Cancer Cell* **2010**, *17* (4), 400–411. https://doi.org/10.1016/j.ccr.2009.12.050.

(4) Jiang, X.; Dulubova, I.; Reisman, S. A.; Hotema, M.; Lee, C. Y. I.; Liu, L.; McCauley, L.; Trevino, I.; Ferguson, D. A.; Eken, Y.; Wilson, A. K.; Wigley, W. C.; Visnick, M. A Novel Series of Cysteine-Dependent, Allosteric Inverse Agonists of the Nuclear Receptor RORγt. *Bioorganic Med. Chem. Lett.* **2020**, *30* (6), 126967. https://doi.org/10.1016/j.bmcl.2020.126967.

(5) Muddana, H. S.; Daniel Varnado, C.; Bielawski, C. W.; Urbach, A. R.; Isaacs, L.; Geballe, M. T.; Gilson, M. K. Blind Prediction of Host–Guest Binding Affinities: A New SAMPL3 Challenge. *J. Comput. Aided. Mol. Des.* **2012**, *26* (5), 475–487. https://doi.org/10.1007/s10822-012-9554-1.

(6) Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. The SAMPL4 Host–Guest Blind Prediction Challenge: An Overview. *J. Comput. Aided. Mol. Des.* **2014**, *28* (4), 305–317. https://doi.org/10.1007/s10822-014-9735-1.

(7) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 Host–Guest Challenge: Are We Doing Better? *J. Comput. Aided. Mol. Des.* **2017**, *31* (1), 1–19. https://doi.org/10.1007/s10822-016-9974-4.

(8) Nicholls, A.; Wlodek, S.; Grant, J. A. The SAMP1 Solvation Challenge: Further Lessons Regarding the Pitfalls of Parametrization. *J. Phys. Chem. B* **2009**, *113* (14), 4521–4532. https://doi.org/10.1021/jp806855q.

(9) Geballe, M. T.; Skillman, a. G.; Nicholls, A.; Guthrie, J. P.; Taylor, P. J. The SAMPL2 Blind Prediction Challenge: Introduction and Overview. *J. Comput. Aided. Mol. Des.* **2010**, *24* (4), 259–279. https://doi.org/10.1007/s10822-010-9350-8.

(10) Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 Host-Guest Binding Affinity Prediction Challenge. **2018**. https://doi.org/10.1101/371724.

(11) Işık, M.; Bergazin, T. D.; Fox, T.; Rizzi, A.; Chodera, J. D.; Mobley, D. L. *Assessing the Accuracy of Octanol–Water Partition Coefficient Predictions in the SAMPL6 Part II Log P Challenge*; Springer International Publishing, 2020; Vol. 34. https://doi.org/10.1007/s10822-020-00295-0.

(12) Işık, M.; Levorse, D.; Rustenburg, A. S.; Ndukwe, I. E.; Wang, H.; Wang, X.; Reibarkh, M.; Martin, G. E.; Makarov, A. A.; Mobley, D. L.; Rhodes, T.; Chodera, J. D. PK a Measurements for the SAMPL6 Prediction Challenge for a Set of Kinase Inhibitor-like Fragments. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1117–1138. https://doi.org/10.1007/s10822-018-0168-0.

(13) Gibb, C. L. D.; Gibb, B. C. Binding of Cyclic Carboxylates to Octa-Acid Deep-Cavity Cavitand. *J. Comput. Aided. Mol. Des.* **2014**, *28* (4), 319–325. https://doi.org/10.1007/s10822-013-9690-2.

(14) Zwanzig, R. W. High-temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22* (8), 1420–1426. https://doi.org/10.1063/1.1740409.

(15) Jiang, W.; Hodoscek, M.; Roux, B. Computation of Absolute Hydration and Binding Free Energy with Free Energy Perturbation Distributed Replica-Exchange Molecular Dynamics. *J. Chem. Theory Comput.* **2009**, *5* (10), 2583–2588. https://doi.org/10.1021/ct900223z.

(16) Mitchell, M. J.; McCammon, J. A. Free Energy Difference Calculations by Thermodynamic Integration: Difficulties in Obtaining a Precise Value. *J. Comput. Chem.* **1991**, *12* (2), 271–275. https://doi.org/10.1002/jcc.540120218.

(17) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8* (9), 3314–3321. https://doi.org/10.1021/ct300418h.

(18) Procacci, P.; Guarrasi, M.; Guarnieri, G. SAMPL6 Host–Guest Blind Predictions Using a Non Equilibrium Alchemical Approach. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 965–982. https://doi.org/10.1007/s10822-018-0151-9.

(19) Frank, L.; Nupur, S.; Zheng, Z.; Merz, K. M. Detailed Potential of Mean Force Studies on Host – Guest Systems from the SAMPL6 Challenge. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1013–1026. https://doi.org/10.1007/s10822-018-0153-7.

(20) Caldararu, O.; Olsson, M. A.; Misini Ignjatović, M.; Wang, M.; Ryde, U. Binding Free Energies in the SAMPL6 Octa-Acid Host–Guest Challenge Calculated with MM and QM Methods. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1027–1046. https://doi.org/10.1007/s10822-018-0158-2.

(21) Nishikawa, N.; Han, K.; Wu, X.; Tofoleanu, F.; Brooks, B. R. Comparison of the Umbrella Sampling and the Double Decoupling Method in Binding Free Energy Predictions for SAMPL6 Octa-Acid Host–Guest Challenges. *J. Comput. Aided. Mol. Des.*

2018, *32* (10), 1075–1086. https://doi.org/10.1007/s10822-018-0166-2.

(22) Laury, M. L.; Wang, Z.; Gordon, A. S.; Ponder, J. W. Absolute Binding Free Energies for the SAMPL6 Cucurbit[8]Uril Host–Guest Challenge via the AMOEBA Polarizable Force Field. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1087–1095. https://doi.org/10.1007/s10822-018-0147-5.

(23) Eken, Y.; Patel, P.; Díaz, T.; Jones, M. R.; Wilson, A. K. SAMPL6 Host–Guest Challenge: Binding Free Energies via a Multistep Approach. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1097–1115. https://doi.org/10.1007/s10822-018-0159-1.

(24) Dixon, T.; Lotz, S. D.; Dickson, A. Predicting Ligand Binding Affinity Using On- and off-Rates for the SAMPL6 SAMPLing Challenge. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1001–1012. https://doi.org/10.1007/s10822-018-0149-3.

(25) Hudson, P. S.; Han, K.; Woodcock, H. L.; Brooks, B. R. Force Matching as a Stepping Stone to QM/MM CB[8] Host/Guest Binding Free Energies: A SAMPL6 Cautionary Tale. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 983–999. https://doi.org/10.1007/s10822-018-0165-3.

(26) Sun, H.; Duan, L.; Chen, F.; Liu, H.; Wang, Z.; Pan, P.; Zhu, F.; Zhang, J. Z. H.; Hou, T. Assessing the Performance of MM/PBSA and MM/GBSA Methods. 7. Entropy Effects on the Performance of End-Point Binding Free Energy Calculation Approaches. **2018**, 14450–14460. https://doi.org/10.1039/c7cp07623a.

(27) Hamaguchi, N.; Fusti-Molnar, L.; Wlodek, S. Force-Field and Quantum-Mechanical Binding Study of Selected SAMPL3 Host-Guest Complexes. *J. Comput. Aided. Mol. Des.* **2012**, *26* (5), 577–582. https://doi.org/10.1007/s10822-012-9553-2.

(28) Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648. https://doi.org/10.1063/1.464913.

(29) Lee, C.; Yang, eitao; Parr, R. G. *Development of the Colic-Salvetti Correlation-Energy Formula into a Functional of the Electron Density*; Vol. 37.

(30) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98* (45), 11623–11627. https://doi.org/10.1021/j100096a001.

(31) Sure, R.; Antony, J.; Grimme, S. Blind Prediction of Binding Affinities for Charged Supramolecular Host-Guest Systems: Achievements and Shortcomings of DFT-D3. *J. Phys. Chem. B* **2014**, *118* (12), 3431–3440. https://doi.org/10.1021/jp411616b.

(32) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132* (15), 154104. https://doi.org/10.1063/1.3382344.

(33) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2011**, *32* (7), 1456–1465. https://doi.org/10.1002/jcc.21759.

(34) Becke, A. D.; Johnson, E. R. A Density-Functional Model of the Dispersion Interaction. *J. Chem. Phys.* **2005**, *123* (15). https://doi.org/10.1063/1.2065267.

(35) Sure, R.; Grimme, S. Corrected Small Basis Set Hartree-Fock Method for Large Systems. *J. Comput. Chem.* **2013**, *34* (19), 1672–1685. https://doi.org/10.1002/jcc.23317.

(36) Mikulskis, P.; Cioloboc, D.; Andrejić, M.; Khare, S.; Brorsson, J.; Genheden, S.; Mata, R. A.; Söderhjelm, P.; Ryde, U. Free-Energy Perturbation and Quantum Mechanical Study of SAMPL4 Octa-Acid Host-Guest Binding Energies. *J. Comput. Aided. Mol. Des.* **2014**, *28* (4), 375–400. https://doi.org/10.1007/s10822-014-9739-x.

(37) Caldararu, O.; Olsson, M. A.; Riplinger, C.; Neese, F.; Ryde, U. Binding Free Energies in the SAMPL5 Octa-Acid Host–Guest Challenge Calculated with DFT-D3 and CCSD(T). *J. Comput. Aided. Mol. Des.* **2017**, *31* (1), 87–106. https://doi.org/10.1007/s10822-016-9957-5.

(38) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse Maps - A Systematic Infrastructure for Reduced-Scaling Electronic Structure Methods. II. Linear Scaling Domain Based Pair Natural Orbital Coupled Cluster Theory. *J. Chem. Phys.* **2016**, *144* (2). https://doi.org/10.1063/1.4939030.

(39) Grimme, S. Semiempirical GGA-Type Density Functional Constructed With A Long-Range Dispersion Correction. *J. Comput. Chem.* **2006**, *27* (15), 1787–1799. https://doi.org/10.1002/jcc.20495.

(40) Perdew, J. P.; Wang, Y. Accurate and Simple Analytic Representation of the Electron-Gas Correlation Energy. *Phys. Rev. B* **1992**, *45* (23), 13244–13249. https://doi.org/10.1103/PhysRevB.45.13244.

(41) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46* (11), 6671–6687. https://doi.org/10.1103/PhysRevB.46.6671.

(42) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378–6396. https://doi.org/10.1021/jp810292n.

(43) Grimme, S. Semiempirical Hybrid Density Functional with Perturbative Second-Order Correlation. *J. Chem. Phys.* **2006**, *124* (3). https://doi.org/10.1063/1.2148954.

(44) Goerigk, L.; Grimme, S. Efficient and Accurate Double-Hybrid-Meta-GGA Density Functionals- Evaluation with the Extended GMTKN30 Database for General Main Group

Thermochemistry, Kinetics, and Noncovalent Interactions. *J. Chem. Theory Comput.* **2011**, *7* (2), 291–309. https://doi.org/10.1021/ct100466k.

(45) Mardirossian, N.; Head-Gordon, M. ω B97M-V: A Combinatorially Optimized, Range-Separated Hybrid, Meta-GGA Density Functional with VV10 Nonlocal Correlation. *J. Chem. Phys.* **2016**, *144* (21). https://doi.org/10.1063/1.4952647.

(46) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115* (19), 2315–2372. https://doi.org/10.1080/00268976.2017.1333644.

(47) Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins Struct. Funct. Bioinforma.* **2009**, *75* (1), 187–205. https://doi.org/10.1002/prot.22234.

(48) Chemical Computing Group Inc. Molecular Operating Environment (MOE). Montreal 2016.

(49) Anthony, W. J.; Bender, A.; Kaya, T.; Clemons, P. A. Alpha Shapes Applied to Molecular Shape Characterization Exhibit Novel Properties Compared to Established Shape Descriptors. *J. Chem. Inf. Model.* **2009**, *49* (10), 2231–2241. https://doi.org/10.1021/ci900190z.

(50) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39* (6), 1397–1412. https://doi.org/10.1063/1.1734456.

(51) 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* **2006**, *65* (3), 712–725. https://doi.org/10.1002/prot.21123.

(52) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(53) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in Docking Success Rates Due to Dataset Preparation. *J. Comput. Aided. Mol. Des.* **2012**, *26* (6), 775–786. https://doi.org/10.1007/s10822-012-9570-1.

(54) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641. https://doi.org/10.1002/jcc.10128.

(55) Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F. Y. R.E.D. Server: A Web Service for Deriving RESP and ESP Charges and Building Force Field Libraries for New Molecules and Molecular Fragments. *Nucleic Acids Res.* **2011**, *39* (SUPPL. 2), 511–517. https://doi.org/10.1093/nar/gkr288.

(56) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97* (40), 10269–10280. https://doi.org/10.1021/j100142a004.

(57) Case, D. A.; Betz, R. M.; Botello-Smith, W.; Cerutti, D. S.; Cheatham III, T. E.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; York, D. M.; Kollman, P. A. Amber 18. **2018**. https://doi.org/10.1002/jcc.23031.

(58) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(59) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. Development of an Improved Four-Site Water Model for Biomolecular Simulations: TIP4P-Ew. *J. Chem. Phys.* **2004**, *120* (20), 9665–9678. https://doi.org/10.1063/1.1683075.

(60) Döpke, M. F.; Moultos, O. A.; Hartkamp, R. On the Transferability of Ion Parameters to the TIP4P/2005 Water Model Using Molecular Dynamics Simulations. *J. Chem. Phys.* **2020**, *152* (2). https://doi.org/10.1063/1.5124448.

(61) 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* **1992**, *32* (5), 523–535. https://doi.org/10.1002/bip.360320508.

(62) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690. https://doi.org/10.1063/1.448118.

(63) Cerutti, D. S.; Case, D. A. Multi-Level Ewald: A Hybrid Multigrid/Fast Fourier Transform Approach to the Electrostatic Particle-Mesh Problem. *J. Chem. Theory Comput.* **2010**, *6* (2), 443–458. https://doi.org/10.1021/ct900522g.

(64) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(65) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Genet.* **2004**, *55* (2), 383–394. https://doi.org/10.1002/prot.20033.

(66) 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. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82. https://doi.org/10.1021/ci100275a.

(67) 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.

(68) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90* (2), 1007–1023. https://doi.org/10.1063/1.456153.

(69) Smith, D. G. A.; Burns, L. A.; Patkowski, K.; Sherrill, C. D. Revised Damping Parameters for the D3 Dispersion Correction to Density Functional Theory. *J. Phys. Chem. Lett.* **2016**, *7* (12), 2197–2203. https://doi.org/10.1021/acs.jpclett.6b00780.

(70) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96* (9), 6796–6806. https://doi.org/10.1063/1.462569.

(71) Woon, D. E.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum through Argon. *J. Chem. Phys.* **1993**, *98* (2), 1358–1371. https://doi.org/10.1063/1.464303.

(72) Dunning, J.; Peterson, K. A.; Wilson, A. K. Gaussian Basis Sets for Use in Correlated Molecular Calculations. X. The Atoms Aluminum through Argon Revisited. *J. Chem. Phys.* **2001**, *114* (21), 9244–9253. https://doi.org/10.1063/1.1367373.

(73) Hariharan, P. C.; Pople, J. A. The Influence of Polarization Functions on Molecular Orbital Hydrogenation Energies. *Theor. Chim. Acta* **1973**, *28* (3), 213–222. https://doi.org/10.1007/BF00533485.

(74) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self — Consistent Molecular Orbital Methods . XII . Further Extensions of Gaussian — Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules Published by the AIP Publishing Articles You May Be Interested in Self - Consistent Molecular Or. *J.Chem. Phys.* **1972**, *56* (5), 2257–2261. https://doi.org/10.1063/1.1677527.

(75) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105* (8), 2999–3093. https://doi.org/10.1021/cr9904009.

(76) Merrick, J. P.; Moran, D.; Radom, L. An Evaluation of Harmonic Vibrational Frequency

Scale Factors. *J. Phys. Chem. A* **2007**, *111* (45), 11683–11700. https://doi.org/10.1021/jp073974n.

(77) Jensen, J. H. Predicting Accurate Absolute Binding Energies in Aqueous Solution: Thermodynamic Considerations for Electronic Structure Methods. *Phys. Chem. Chem. Phys.* **2015**, *17* (19), 12441–12451. https://doi.org/10.1039/c5cp00628g.

(78) Klamt, A.; Moya, C.; Palomar, J. A Comprehensive Comparison of the IEFPCM and SS(V)PE Continuum Solvation Methods with the COSMO Approach. *J. Chem. Theory Comput.* **2015**, *11* (9), 4220–4225. https://doi.org/10.1021/acs.jctc.5b00601.

(79) Riplinger, C.; Neese, F. An Efficient and near Linear Scaling Pair Natural Orbital Based Local Coupled Cluster Method. *J. Chem. Phys.* **2013**, *138* (3). https://doi.org/10.1063/1.4773581.

(80) Neese, F.; Valeev, E. F. Revisiting the Atomic Natural Orbital Approach for Basis Sets: Robust Systematic Basis Sets for Explicitly Correlated and Conventional Correlated Ab Initio Methods? *J. Chem. Theory Comput.* **2011**, *7* (1), 33–43. https://doi.org/10.1021/ct100396y.

(81) Jensen, F. Polarization Consistent Basis Sets. II. Estimating the Kohn–Sham Basis Set Limit. *J. Chem. Phys.* **2002**, *116* (17), 7372–7379. https://doi.org/10.1063/1.1465405.

(82) Stück, D.; Baker, T. A.; Zimmerman, P.; Kurlancheek, W.; Head-Gordon, M. On the Nature of Electron Correlation in C60. *J. Chem. Phys.* **2011**, *135* (19). https://doi.org/10.1063/1.3661158.

(83) Bauernschmitt, R.; Ahlrichs, R. Stability Analysis for Solutions of the Closed Shell Kohn-Sham Equation. *J. Chem. Phys.* **1996**, *104* (22), 9047–9052. https://doi.org/10.1063/1.471637.

(84) Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. M. A Complete Active Space SCF Method (CASSCF) Using a Density Matrix Formulated Super-CI Approach. *Chem. Phys.* **1980**, *48* (2), 157–173. https://doi.org/10.1016/0301-0104(80)80045-0.

(85) Lee, T. J.; Taylor, P. R. A Diagnostic for Determining the Quality of Single-reference Electron Correlation Methods. *Int. J. Quantum Chem.* **1989**, *36* (23 S), 199–207. https://doi.org/10.1002/qua.560360824.

(86) Lee, T. J.; Rice, J. E.; Scuseria, G. E.; Schaefer, H. F. Theoretical Investigations of Molecules Composed Only of Fluorine, Oxygen and Nitrogen: Determination of the Equilibrium Structures of FOOF, (NO)2 and FNNF and the Transition State Structure for FNNF Cis-Trans Isomerization. *Theor. Chim. Acta* **1989**, *75* (2), 81–98. https://doi.org/10.1007/BF00527711.

(87) Lee, M. C.; Yang, R.; Duan, Y. Comparison between Generalized-Born and Poisson-Boltzmann Methods in Physics-Based Scoring Functions for Protein Structure Prediction. *J. Mol. Model.* **2005**, *12* (1), 101–110. https://doi.org/10.1007/s00894-005-0013-y.

(88) Sure, R.; Grimme, S. Comprehensive Benchmark of Association (Free) Energies of Realistic Host−Guest Complexes. **2015**. https://doi.org/10.1021/acs.jctc.5b00296.

(89) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7* (18), 3297–3305. https://doi.org/10.1039/b508541a.

(90) Lee, J.; Head-Gordon, M. Distinguishing Artificial and Essential Symmetry Breaking in a Single Determinant: Approach and Application to the C60, C36, and C20 Fullerenes. *Phys. Chem. Chem. Phys.* **2019**, *21* (9), 4763–4778. https://doi.org/10.1039/c8cp07613h.

(91) Sherrill, C. D.; Lee, M. S.; Head-Gordon, M. On the Performance of Density Functional Theory for Symmetry-Breaking Problems. *Chem. Phys. Lett.* **1999**, *302* (5–6), 425–430. https://doi.org/10.1016/S0009-2614(99)00206-7.

(92) Sherrill, C. D.; Krylov, A. I.; Byrd, E. F. C.; Head-Gordon, M. Energies and Analytic Gradients for a Coupled-Cluster Doubles Model Using Variational Brueckner Orbitals: Application to Symmetry Breaking in O4+. *J. Chem. Phys.* **1998**, *109* (11), 4171–4181. https://doi.org/10.1063/1.477023.

(93) Lee, J.; Head-Gordon, M. Regularized Orbital-Optimized Second-Order Møller-Plesset Perturbation Theory: A Reliable Fifth-Order-Scaling Electron Correlation Model with Orbital Energy Dependent Regularizers. *J. Chem. Theory Comput.* **2018**, *14* (10), 5203–5219. https://doi.org/10.1021/acs.jctc.8b00731.

(94) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. - A Eur. J.* **2012**, *18* (32), 9955–9964. https://doi.org/10.1002/chem.201200497.

(95) Li, Y. P.; Gomes, J.; Sharada, S. M.; Bell, A. T.; Head-Gordon, M. Improved Force-Field Parameters for QM/MM Simulations of the Energies of Adsorption for Molecules in Zeolites and a Free Rotor Correction to the Rigid Rotor Harmonic Oscillator Model for Adsorption Enthalpies. *J. Phys. Chem. C* **2015**, *119* (4), 1840–1850. https://doi.org/10.1021/jp509921r.

CHAPTER FIVE

Chemoenzymatic Synthesis of Glycopeptides Bearing Rare N-Glycan Sequences with or without Bisecting GlcNAc

About this chapter: This chapter is reprinted from Yang, W.; Ramadan, S.; Orwenyo, J.; Kakeshpour, T.; Diaz, T.; Eken, Y.; Sanda, M.; Jackson, J. E.; Wilson, A. K.; Huang, X. Chemoenzymatic Synthesis of Glycopeptides Bearing Rare N-Glycan Sequences with or without Bisecting GlcNAc. Chem. Sci. **2018**, *9 (43)*, 8194–8206 with the permission of the Royal Society of Chemistry. The experiments mentioned in this chapter are performed by our collaborators from Huang Group, simulations on Glycan **39** are performed by Thomas Diaz, and simulations on Glycan **41** are performed by Yiğitcan Eken.

5.1 Introduction

The glycosylation of proteins is one of the most common post translational modifications and results in the great diversity of glycoprotein structures. Various attached carbohydrate groups play critical roles in directing biological functions, structure stability, and conformations. However, natural glycopeptides typically exist as a mixture of glycoforms with different oligosaccharide groups attached. This makes it difficult to isolate pure individual forms of glycopeptide. Due to that, synthetic methods are developed to produce pure glycopeptide forms. The chemoenzymatic approach using the *Arthrobacter* endoβ-N-acteylglucosaminidase (Endo-A) enzyme that has transglycosylation activity and can transfer free N-glycans to N-acetyl glucosamine (GlcNAC) bearing acceptors is a method widely used and shown in Figure 5.1.[1]

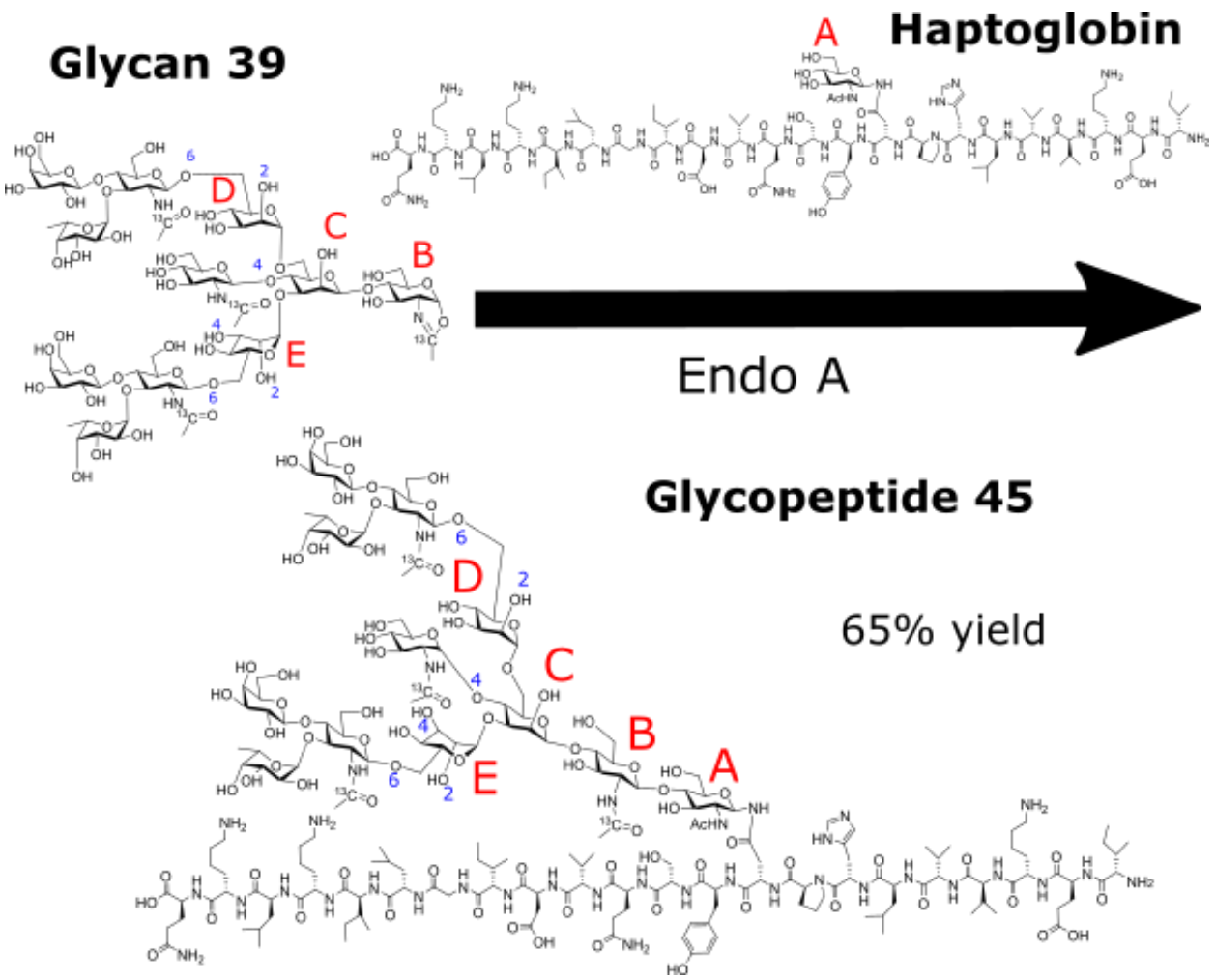


Figure 5. 1 Glycan **39** treated with Endo-A enzyme and GlcNAc (unit A) bearing haptoglobin as the acceptor glycopeptide leads to a reaction yield of 65% glycopeptide **45**. Potential branching sites are indicated on the figure with corresponding carbon numbers they associate within the saccharide unit shown through letters A, B, C, D, E. [2]

This transformation results in N-linked glycoproteins, where the carbohydrate binds to the protein backbone by using an asparagine residue. During the free N-glycan synthesis for transglycosylation, most common points of attachments appear in the OH groups of C2/C6 carbons of mannose D and C2/C4 carbons of mannose E or introducing a GlcNAC structure on the hydroxyl group attached to the C4 carbon of mannose C. Our collaborators from Huang

group synthesized glycan **39** and glycan **41**[2] oxazolines, which bear a branch at 6-OH of mannose E. The structures of glycan **39** and glycan **41** are shown in Figure 5.2.

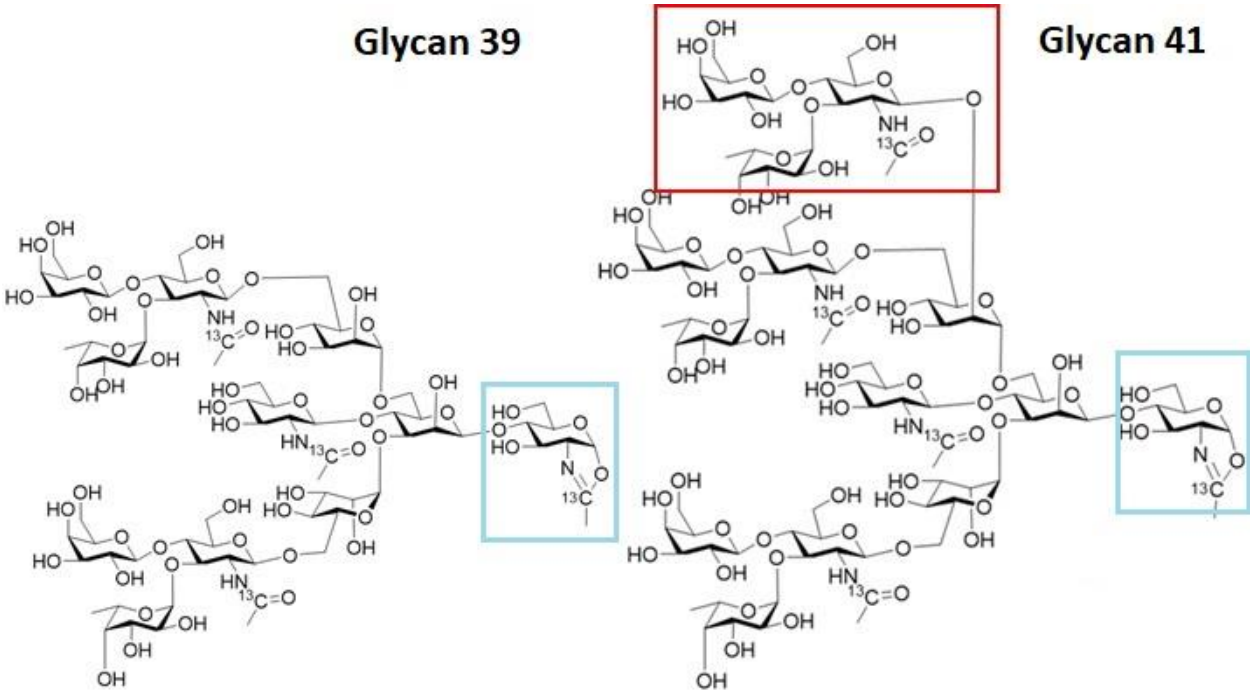


Figure 5. 2 Structures of two glycan substrates. Glycan 39 is shown on the left and glycan 41 is shown on the right.The additional LewisX trisaccharide thioglycosyl donor group is marked with red and the oxazoline ring, where the transglycosylation occurs, is marked with blue.

The synthesis was performed by the Huang group and it is beyond the scope of this dissertation; the focus here will be on the computational work performed. The only difference between these two glycans is the additional LewisX trisaccharide thioglycosyl group present at the C2 carbon of mannose D on glycan **41** as indicated in Figure 5.2. These two rare GlcNAc containing oxazolines were tested for transglycosylation reaction using Endo-A enzyme and GlcNAC bearing haptoglobin glycopeptide as an acceptor.[2] The experimental results show that when glycan **39** is used as a donor, the reaction results in glycopeptide **45** with 65% yield (Figure 5.1). In contrast, when glycan **41** is used, it does not participate in this reaction and the reaction

does not lead to the desired product. The expected reaction site for transglycosylation is the oxazoline rings present in B saccharides of both glycans (highlighted with blue in Figure 5.2). Considering the similarity between the two glycans with the additional branching being far from the reaction site, this divergent behavior of the transglycosylation reaction was unexpected. To better understand this behavior and detect potential sources for low transglycosylation of glycan **41** between two free oxazolines that are docked to the Endo-A enzyme active site, molecular dynamics simulations are performed in primary poses. The binding energies are calculated via end-state free energy calculations and used to assess glycan **39** and glycan **41**'s preference to the Endo-A binding site, which might be the potential cause of reduced glycosylation yield in glycan **41**.

5.2 Computational Methodology

Initial coordinates of Endo-A were obtained from the Protein Data Bank[3] (PDB ID: 3FHA)[4]. As the focus of the study was on pocket residue-ligand interaction, missing segments and residues outside the pocket region were capped using Molecular Operating Environment v.2016.08 (MOE).[5] Gate-keeper residues, W216 and W244 are positioned parallel to one another during transglycoslyation.[4] W244 was rotated from its original perpendicular orientation to parallel with W216. The protein structure was initially minimized in MOE under the AMBER ff10 force field[6] and Extended Hückel Theory. The compounds were then non-covalently docked with the docking program in MOE. Binding poses were refined using an induced fit refinement method. The geometries of the N-glycan oxazoline compounds were optimized using the Gaussian 16 program package[7]. The optimizations were performed using the AM1 method.[8] The obtained Mulliken charges were used with the antechamber of Amber 16 in the generation of parameters for the N-glycan compounds. The systems were prepared using the Leap module of

AmberTools16[9] under the AMBER ff14SB[10] and GAFF force fields. Each enzyme complex was solvated in a 14 Å cube of TIP4P-Ew water beyond the solute and 100 mM sodium chloride. The systems were relaxed under NVT conditions over six minimization procedures with decreasing restraints on the protein of 500.0, 200.0, 20.0, 10.0, 5.0 kcal/(mol $Å^2$) to no restraints. The systems were then heated to 300 K over 30 ps. Atomistic molecular dynamics simulations were performed for 30 ns at 300 K and 1 atm using AMBER 16. The SHAKE algorithm constrained bonds involving hydrogen.[11] The trajectories were produced using Langevin dynamics and the pressure of the system was regulated with isotropic position scaling. Long-range electrostatic effects were modeled using the particle-mesh Ewald method with a cutoff of 10 Å. The resulting trajectories were analyzed using AMBER 16 and visualized with MOE and the UCSF Chimera package. Free energy of binding was calculated for every picosecond using the Poisson Boltzmann model form the MMPBSA.py module of AmberTools and AMBER 16.[12] The relative free energy trends between models were compared, so solute entropy was neglected.

5.3 Results and Discussion

Endo-A catalyzed transglycosylation reaction occurs in the binding site where active residues W93, N171, E173, Y205, F125, W216, F243, W244,Y299 are present. During the reaction, these critical residues surround the substrate and stimulate the oxazoline ion intermediate formation and nucleophilic attack on this intermediate. The reaction mechanism requires a strong interaction between the free oxazolines and the Endo-A enzyme.[4] The experimental results showed that when oxazoline **39** is treated with Endo-A enzyme and GlcNAc bearing haptoglobin glycopeptide as a the acceptor, 65% transglycosylation yield is obtained. However, when the experiment repeated with oxazoline **41**, no desired glycopeptides where produced. To better understand this differing behavior of **39** and **41** docking, MD and free energy calculations were

performed. First, the two glycans docked into Endo-A (representative poses are shown in Figure 5.3). Poses with the oxazoline ring position within the active site are simulated using atomistic MD and the binding energy of each pose is calculated from free energy calculations.

Table 5. 1 Endo-A Binding energies of various binding poses of **39** and **41**

Compound	**Binding Poses**	**Binding energy (kcal mol^{-1})**
39	1	−72.97 ± 6.04
	2	−94.00 ± 9.15
	3	−77.36 ± 7.96
Average	−81.44 ± 11.10	
41	1	−52.08 ± 11.26
	2	−60.17 ± 11.56
	3	−55.26 ± 7.95
	4	−58.80 ± 7.83
	5	−54.86 ± 11.17
Average	−56.24 ± 9.95	

From the docking of glycan **39** to Endo-A, three different poses that are in the range of the active site were obtained, whereas glycan **41** poses exhibit more diversity which might be due to the lack of strong interactions with the protein. As a result, five poses are analyzed in order to investigate the binding of each conformer. Next, 30 ns MD simulations were performed on these systems and free energy calculations are performed. The results show an average binding energy of **39** with Endo-A of −81.44 ± 11.10 kcal mol^{-1}. Yet, binding of glycan **41** and Endo-A is

significantly weaker at −56.24 ± 9.95 kcal mol^{-1} which might account for the lack of transglycosylation.

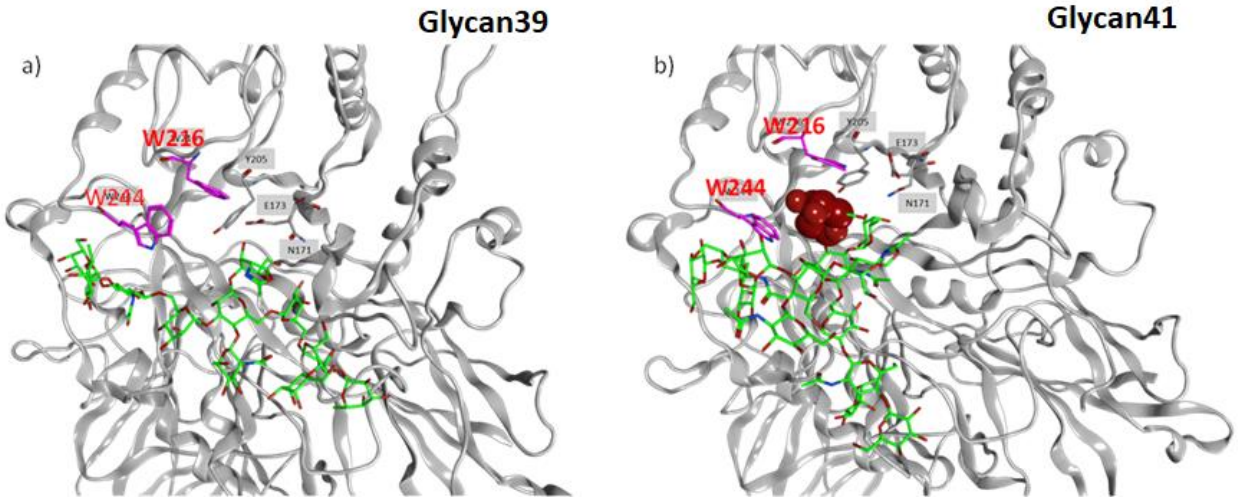


Figure 5. 3 Binding pose representations for the two glycans investigated. The figure on the left is a snapshot taken from the MD simulation of glycan 39 with Endo-A and the indole rings of W216 and W244 are in the perpendicular position. Snapshot taken from the MD simulation of glycan 41 with Endo-A is shown on the right, indole rings of W216 and W244 are in the parallel position because of the hindrance caused by the additional antenna.

In the crystal structure of Endo-A and Endo-A complexed with tetrasaccharide oxazoline substrate, the indole rings of W244 and W216 are perpendicular to each other. It is known that the indole rings of W244 and W216 should be in parallel position and act as a gate to allow substrate entry to the active site.[4] These rings were moved to be parallel in order to allow substrate entry to the active site. In all simulations with oxazoline **39**, these rings turned back into their original perpendicular orientation (Figure 5.3a). However, in the complex with additional tri-antennae bearing donor **41**, the additional antenna stays between W244 and W216 and hinders the rotation of the indole rings. This may prohibit the closed active site formation and account for the reducing yield of glycosylation.

5.4 Conclusions

In this study molecular modeling is used to study interactions between Endo-A enzyme and glycans **39**, **41** that are synthesized and experimentally evaluated for trans glycosylation reaction yields. Experimentally, Endo-A enzyme shows substrate preference towards glycan **39**. Supportive of experiment, when the Endo-A binding energy predicted for glycan **39** and **41**; glycan **41** showed significantly weaker binding compared to glycan **39**. This indicates glycan **39** have higher affinity towards the active site of Endo-A. Additionally, simulations showed active site gate residues W244 and W216 are hindered when the glycan **41** binds to the Endo-A active site which can mechanistically explain why transglycosylation reaction did not occur on glycan **41**.

REFERENCES

REFERENCES

(1) Li, B.; Zeng, Y.; Hauser, S.; Song, H.; Wang, L. X. Highly Efficient Endoglycosidase-Catalyzed Synthesis of Glycopeptides Using Oligosaccharide Oxazolines as Donor Substrates. *J. Am. Chem. Soc.* **2005**, *127* (27), 9692–9693. https://doi.org/10.1021/ja051715a.

(2) Yang, W.; Ramadan, S.; Orwenyo, J.; Kakeshpour, T.; Diaz, T.; Eken, Y.; Sanda, M.; Jackson, J. E.; Wilson, A. K.; Huang, X. Chemoenzymatic Synthesis of Glycopeptides Bearing Rare N-Glycan Sequences with or without Bisecting GlcNAc. *Chem. Sci.* **2018**, 8194–8206. https://doi.org/10.1039/c8sc02457j.

(3) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(4) Yin, J.; Li, L.; Shaw, N.; Li., Y.; Song, J. K.; Zhang, W.; Xia, C.; Zhang, R.; Joachimiak, A.; Zhang, H. C.; Wang, L. X.; Liu, Z. J.; Wang, P. Structural Basis and Catalytic Mechanism for the Dual Functional Endo-$β$-N-Acetylglucosaminidase A. *PLoS One* **2009**, *4* (3). https://doi.org/10.1371/journal.pone.0004658.

(5) Chemical Computing Group Inc. Molecular Operating Environment (MOE). Montreal 2016.

(6) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, *26* (16), 1668–1688. https://doi.org/10.1002/jcc.20290.

(7) 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.

(8) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107* (13), 3902–3909. https://doi.org/10.1021/ja00299a024.

(9) Case, D. A.; Cerutti, D. S.; T.E. Cheatham, I. I. I.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A. E.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. Amber 2016. **2016**, No. April. https://doi.org/10.13140/RG.2.2.36172.41606.

(10) 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. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(11) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(12) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8* (9), 3314–3321. https://doi.org/10.1021/ct300418h.

CHAPTER SIX

Chemical Synthesis of Human Syndecan-4 Glycopeptide Bearing O-, N- Sulfation and Multiple Aspartic Acids for Probing Impacts of the Glycan Chain and the Core Peptide on Biological Functions

About this chapter: This chapter is reprinted from Yang, W.; Eken, Y.; Zhang, J.; Cole, L. E.; Ramadan, S.; Xu, Y.; Zhang, Z.; Liu, J.; Wilson, A. K.; Huang, X. Chemical Synthesis of Human Syndecan-4 Glycopeptide Bearing O-, N-Sulfation and Multiple Aspartic Acids for Probing Impacts of the Glycan Chain and the Core Peptide on Biological Functions. *Chem. Sci*. **2020***, 11 (25),* 6393–6404 with the permission of the Royal Society of Chemistry. The experiments mentioned in this chapter are performed by our collaborator Huang Group and all the theoretical work is performed by Yiğitcan Eken.

6.1 Introduction

Heparan sulfates (HS) are linear sulfated polysaccharides who are found in all animal tissues. They have a range of biological functions including blood clothing prevention, growth factor and chemokine binding and controlling activity levels of various enzymes.[1–3] In nature, HS exists as a heterogenous mixtures where the length of their backbone and location of sulfates varies. In order to produce pure forms of HS for therapeutic purposes or to study HS structure activity relationship, synthetic routes are commonly adopted.

Within the tissues HS are originally exist as a proteoglycan where HS is covalently linked to a core protein or a core peptide and form heparan sulfate proteoglycan (HSPG).[4,5] These core proteins originally thought as carriers and do not possess any biological activity. However, recent studies suggest that the core protein itself can also be biologically active.[6–8] To further understand the role of the core protein on the biological function of HSPG glycopeptides, synthesis and study of well-defined homogenous glycans and glycopeptides are vital.

During this study glycan **28** bearing an N- and O-sulfated glycan chain and glycopeptide **2,** where glycan **28** covalently linked to a human syndecan-4[9] (amino acids 60–71) with four

aspartic acids in the peptide backbone was synthesized by our collaborators from Huang group (Figure 6.1).[10]

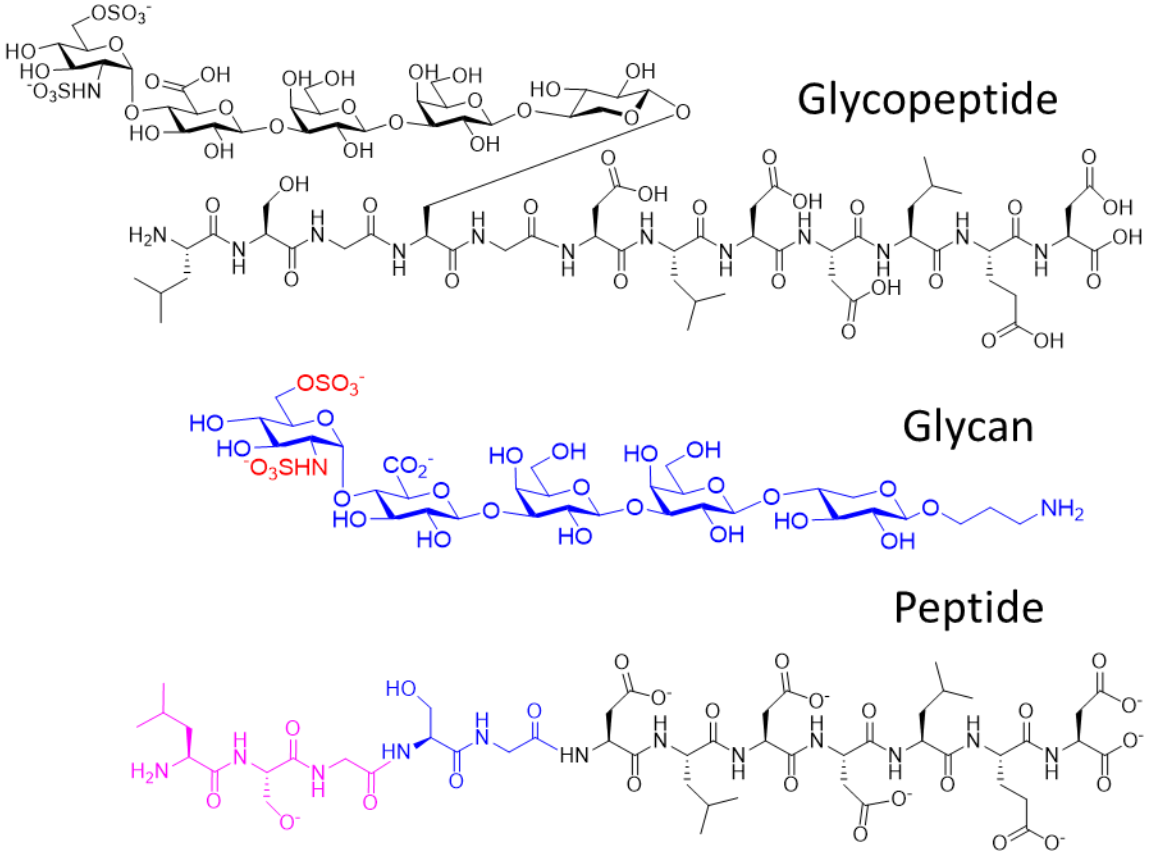


Figure 6. 1 Chemical structures of glycopeptide, glycan and peptide synthesized by our collaborators.

Table 6. 1 Inhibitory activities of glycopeptide, glycan and peptide towards heparanase (5 nM) and their dissociation constant respect to FGF-2 binding measure through biolayer interferometry.

Compound	**Heparanase % Inhibition**			**FGF-2 K_D (nm)**
	3.3μM	**10μM**	**33μM**	
Glycopeptide	NA	NA	NA	5
Glycan	NA	32%	61%	14.5
Peptide	NA	NA	NA	17

After the synthesis, glycopeptide **2** and glycan **28** were experimentally tested for heparanase and FGF-2 biological activity along with the peptide **29** backbone by itself. The activity data showed that function of the glycan chain is affected by the peptide. During the heparanase study, having glycan by itself showed inhibitory activity while glycopeptide and peptide showed no inhibition (Table 6.1). However, the FGF-2 dissociation constants showed glycopeptide 3-fold enhanced binding compared to glycan and peptide (Table 6.1). To insight activity data and understand how peptide backbone impacts HS functions molecular modeling is used.[10]

6.2 Computational Methodology

FGF-2 modeling studies were performed on the FGF-2 complexes with the glycopeptide **2**, peptide **29** and glycan **28** respectively, using crystal structure of the FGF2 protein (PDB[11] ID: 4OEE).[12] The potential ligand binding sites on the protein were detected by the Site Finder program implemented in Molecular Operating Environment (MOE).[13,14] The results showed three potential ligand binding sites on FGF-2 with a positive Propensity of Ligand Binding (PLB) score (Figure 6.2). Glycopeptide **2**, peptide **29** and glycan **28** structures were docked

individually into each of these potential binding sites. Molecular dynamics (MD) simulations and binding free energy calculations were performed on the distinct binding poses with highest GBVI/WSA △G scores.

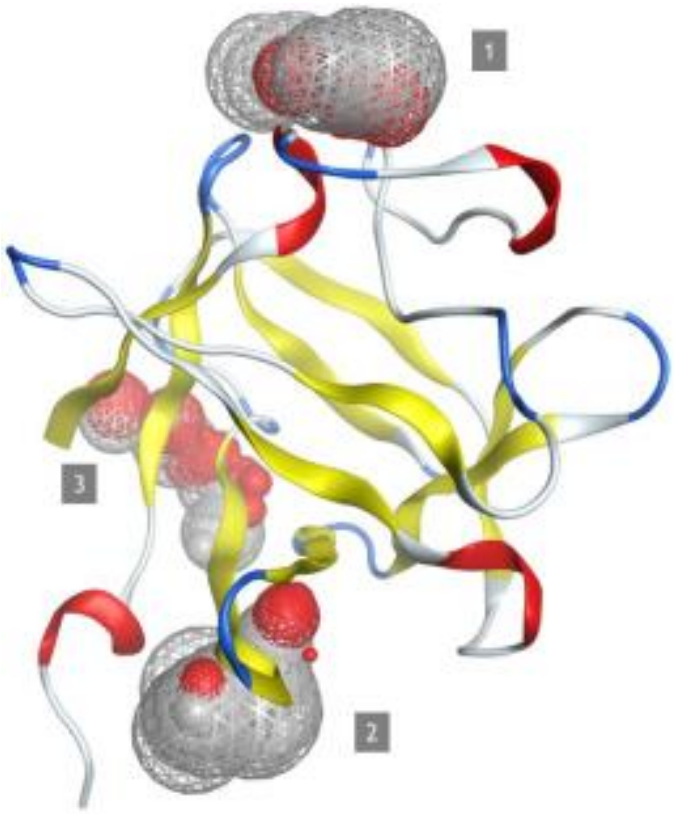


Figure 6. 2 Potential binding Sites on the FGF2 structure.

Similar to FGF-2 study, the binding behavior of glycopeptide **2**, peptide **29** and glycan **28** on heparanase has also been investigated computationally. For this purpose, molecules and the biotin tag used during experiments were docked into the heparin binding site of the heparanase (PDB ID: 5E9C)[15] using MOE. The distinct poses with highest GBVI/WSA △G scores were further studied with molecular dynamics and binding free energy calculations. The average binding energies and energies calculated from individual poses can be found in Table 6.2. The experiments were performed on the glycopeptide **2**, peptide **29** and glycan **28** structures with a biotin tag. Due to that, the biotin tag was also included in this study to assess its contribution to

the binding. The biotin tag gave little binding energy with heparanase, indicating that majority of the binding energy results from interactions between glycan and heparanase.

6.3 Computational Results and Analyses of the Interactions

6.3.1 FGF-2 Binding

The scan of FGF-2 structure led to identification of 3 potential ligand binding sites. Each of these binding sites are evaluated for glycopeptide **2**, peptide **29** and glycan **28** binding through MD simulations and binding free energy calculations. The average binding energy results of glycopeptide **2**, peptide **29** and glycan **28** for each site can be found in Table 6.4. The results showed site 1 had the highest affinity for both glycopeptide, glycan and peptide**.** The X-ray crystal structure of complexes of FGF-2 and heparin oligosaccharides from literature showed that the glycans reside in the site 1,[12,16] which is consistent with our computation results. The average binding energies and their experimental K_D values for FGF2 are listed in Table 6.1, and energies calculated from individual poses can be found in Table 6.2.

Table 6. 2 Binding free energy for glycopeptide, glycan and peptide with FGF-2 calculated for various poses.

FGF-2 Site 1 Glycan Binding Comparison (kcal mol^{-1})				
Compound	**Pose**	**ΔG Binding**	**STD**	**Average ΔG**
Glycan	1	-34.37	± 8.60	-35.09 ± 8.01
	2	-35.25	± 6.59	
	3	-36.40	± 8.96	
	4	-31.72	± 6.45	
	5	-37.74	± 9.45	
Peptide	1	-30.92	± 10.78	-30.40 ± 10.55
	2	-25.55	± 8.97	
	3	-26.84	± 7.88	
	4	-35.85	± 10.16	
	5	-32.86	± 14.94	
Glycopeptide	1	-53.51	± 14.79	-60.04 ± 13.65
	2	-77.67	± 16.44	
	3	-69.51	± 17.63	
	4	-47.82	± 12.29	
	5	-53.81	± 5.36	
	6	-51.88	± 18.17	
	7	-66.07	± 10.85	

Binding site 1 of FGF-2 is lined with many basic residues including Asn27, Arg44, Lys 119, Arg120, Lys125, Lys129, Gln134 and Lys135 (Fig. 4). MD simulations of FGF-2 complex with glycopeptide showed that these residues formed hydrogen bonds with glycopeptide. The distances between the side chains of Lys125 and Lys119 are within 5 Å from the sulfates on the glycan, indicating potential electrostatic interactions. In all glycopeptide **2** binding poses, the glycan is located within binding site 1 while the peptide extends out of the pocket and towards the protein's surface. Meanwhile, glycan **28** binds site 1 with an analogous conformation as that of glycopeptide (Fig. 6.3). The binding poses and the MD simulations showed the peptide portion of glycopeptide extends out of site 1 towards the surface of FGF-2. This leads to the formation of additional salt bridges with the basic residues outside of binding site 1 including Arg22 and Lys21 (Figure 6.3). These additional salt bridges are presumably responsible for improved binding to FGF-2 as observed in glycopeptide as compared to glycan.

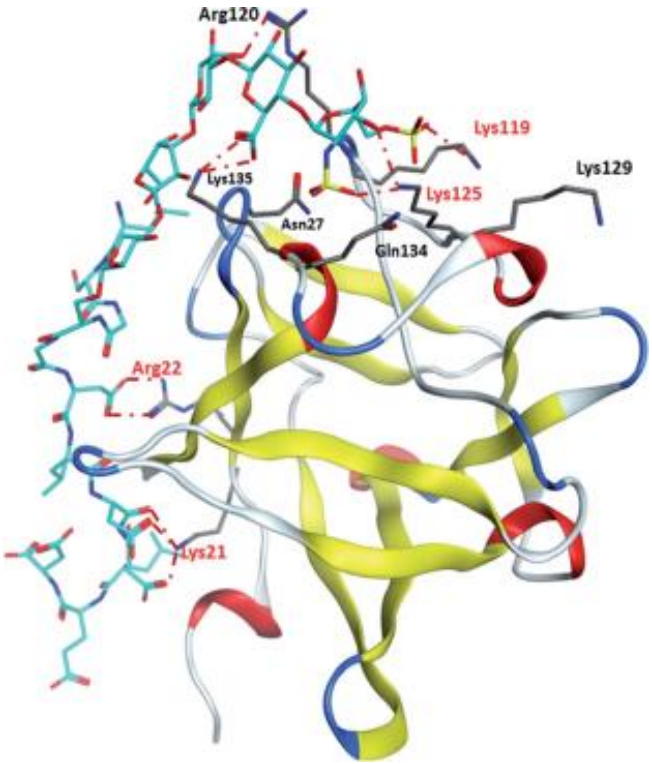


Figure 6. 3 Representative binding pose of glycopeptide to FGF2.

6.3.2 Heparanase Binding

Table 6. 3 Binding free energy for glycopeptide **2**, peptide **29** and glycan **28** with heparanase calculated for various poses.

Heparanase Binding (kcal mol^{-1})				
Compound	**Pose**	**ΔG Binding**	**STD**	**Average ΔG**
Glycan **28**	1	-59.97	14.44	-57.36 ± 12.19
	2	-53.89	10.00	
	3	-49.84	12.50	
	4	-72.58	10.83	
	5	-50.52	13.20	
Peptide **29**	1	-34.62	12.47	-43.14 ± 14.45
	2	-61.90	20.10	
	3	-32.25	10.36	
	4	-35.28	12.51	
	5	-51.66	16.80	
Glycopeptide **2**	1	-45.65	13.18	-50.55 ± 15.67
	2	-55.65	15.23	
	3	-49.41	15.53	
	4	-46.77	15.42	
	5	-55.26	18.99	

The average binding energies to heparanase and energies calculated from individual poses are included in Table 6.3. The binding energy results show that glycan **28** has a higher affinity to heparanase compared to peptide **29** and glycopeptide **2**, respectively. The glycan **28**'s higher binding affinity towards heparanase compared to glycopeptide **2** is contrary to the results observed with FGF-2. When inhibitory activities of glycan, peptide and glycopeptide toward heparanase is considered (Table 6.1), glycan showed 32% inhibitory activity in 10μM concentration and 61% inhibition on 33μM concentration. Peptide and glycopeptide showed no inhibitory activity towards heparanase.

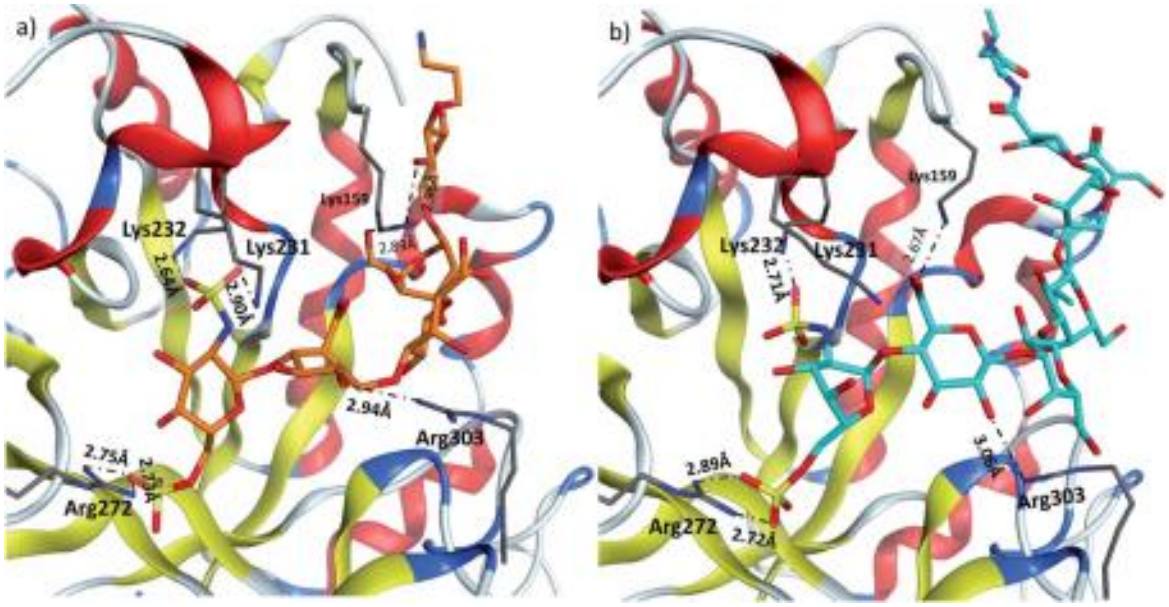


Figure 6. 4 Comparison of (a) glycan **28** and (b) glycopeptide **2** binding to the site 1 of heparanase (heparin binding site).

Heparanase binding site consists of many basic residues including Lys159, Arg272, Lys231, Lys232, Arg303. Glycan **28** is oriented within the binding site by interacting with these basic residues through hydrogen bonds and ionic bonds (Fig. 6.4a). In glycopeptide **2** complex with heparanase, the glycan is situated within the binding site, while the peptide backbone extends toward the solvent (Figure 6.4b). The comparison of glycan **28** and glycopeptide **2** binding

shows that core H-bonds and ionic interactions in the binding pocket are weakened in the glycopeptide complex. For example, the interaction between Lys231 and N-sulfate group observed in glycan **28**/heparanase is lost in the glycopeptide **2**/heparanase complex. Furthermore, in glycan **28**/heparanase complex vs. glycopeptide **2**/heparanase, the distance between Lys232 and N-sulfate group increased from 2.64 Å to 2.71 Å, the distance between Arg272 and O-sulfate group increased from 2.75 Å to 2.89 Å, and H-bond distance between Arg303 and a hydroxyl group increased from 2.94 Å to 3.06 Å (Fig. 5). This weakening of glycan/protein interactions can be explained by the peptide backbone of glycopeptide not fitting in the pocket, thus disrupting the glycan interactions with heparanase, which presumably leads to reduced affinity and inhibitory activity of glycopeptide **2** on heparanase.

6.4 Conclusion

With this study for the first time, HSPG glycopeptides bearing multiple Asp residues in the peptide backbone and O- and N-sulfation on the glycan chain have been successfully synthesized and tested for biological functions by the Huang group. The results showed the glycan inhibited the activities of heparanase, while the glycopeptide did not alter the heparanase activity. Additionally, the glycopeptide showed enhanced binding comparison to glycan and peptide by itself in FGF-2 systems. The molecular dynamics simulations are used to insight functioning of these ligands with respect to heparanase and FGF-2 binding. The simulations showed the peptide portion of the glycopeptide **2** can led to additional salt bridges in FGF-2 systems, whereas in heparanase it tends to pull the glycan core towards solvent which may explain opposite effect of peptide attachment in activity. The experimental results combined with the structural insights gained from molecular modeling, suggests that transferring HS to a core protein as in proteoglycans may be used to modulate HS functions.

APPENDIX

Table 6. 4 Average binding free energies and standard deviations calculated for glycan **28**, peptide **29** and glycopeptide **2** on 3 potential binding sites.

Compound	**Site**	**Average Binding Energies (kcal mol^{-1})**	
Glycan 28	1	-35.09	± 8.01
	2	-26.75	± 10.55
	3	-20.04	± 7.17
Peptide 29	1	-30.40	± 10.55
	2	-25.59	± 10.24
	3	-26.26	± 11.07
Glycopeptide 2	1	-60.04	± 13.65
	2	-37.40	± 13.88
	3	-41.48	± 12.03

REFERENCES

REFERENCES

(1) Petitou, M.; van Boeckel, C. A. A. A Synthetic Antithrombin III Binding Pentasaccharide Is Now a Drug! What Comes Next? *Angew. Chemie Int. Ed.* **2004**, *43* (24), 3118–3133. https://doi.org/10.1002/anie.200300640.

(2) Bernfield, M.; Götte, M.; Park, P. W.; Reizes, O.; Fitzgerald, M. L.; Lincecum, J.; Zako, M. Functions of Cell Surface Heparan Sulfate Proteoglycans. *Annu. Rev. Biochem.* **1999**, *68* (1), 729–777. https://doi.org/10.1146/annurev.biochem.68.1.729.

(3) Lindahl, U.; Kusche-Gullberg, M.; Kjellén, L. Regulated Diversity of Heparan Sulfate. *J. Biol. Chem.* **1998**, *273* (39), 24979–24982. https://doi.org/10.1074/jbc.273.39.24979.

(4) Häcker, U.; Nybakken, K.; Perrimon, N. Heparan Sulphate Proteoglycans: The Sweet Side of Development. *Nat. Rev. Mol. Cell Biol.* **2005**, *6* (7), 530–541. https://doi.org/10.1038/nrm1681.

(5) Bishop, J. R.; Schuksz, M.; Esko, J. D. Heparan Sulphate Proteoglycans Fine-Tune Mammalian Physiology. *Nature* **2007**, *446* (7139), 1030–1037. https://doi.org/10.1038/nature05817.

(6) Choi, S.-J.; Lee, H.-W.; Choi, J.-R.; Oh, E.-S. Shedding; towards a New Paradigm of Syndecan Function in Cancer. *BMB Rep.* **2010**, *43* (5), 305–310. https://doi.org/10.5483/BMBRep.2010.43.5.305.

(7) Morgan, M. R.; Humphries, M. J.; Bass, M. D. Synergistic Control of Cell Adhesion by Integrins and Syndecans. *Nat. Rev. Mol. Cell Biol.* **2007**, *8* (12), 957–969. https://doi.org/10.1038/nrm2289.

(8) Iozzo, R. V. Series Introduction: Heparan Sulfate Proteoglycans: Intricate Molecules with Intriguing Functions. *J. Clin. Invest.* **2001**, *108* (2), 165–167. https://doi.org/10.1172/JCI13560.

(9) David, G.; van der Schueren, B.; Marynen, P.; Cassiman, J. J.; van den Berghe, H. Molecular Cloning of Amphiglycan, a Novel Integral Membrane Heparan Sulfate Proteoglycan Expressed by Epithelial and Fibroblastic Cells. *J. Cell Biol.* **1992**, *118* (4), 961–969. https://doi.org/10.1083/jcb.118.4.961.

(10) Yang, W.; Eken, Y.; Zhang, J.; Cole, L. E.; Ramadan, S.; Xu, Y.; Zhang, Z.; Liu, J.; Wilson, A. K.; Huang, X. Chemical Synthesis of Human Syndecan-4 Glycopeptide Bearing O-, N-Sulfation and Multiple Aspartic Acids for Probing Impacts of the Glycan Chain and the Core Peptide on Biological Functions. *Chem. Sci.* **2020**, *11* (25), 6393–6404. https://doi.org/10.1039/d0sc01140a.

(11) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(12) Li, Y.; Ho, I.; Ku, C.; Zhong, Y.; Hu, Y.; Chen, Z.; Wang, C.; Hsiao, C. Interactions That In Fl Uence the Binding of Synthetic Heparan Sulfate Based Disaccharides to Fibroblast Growth Factor - 2. **2014**, 6–11.

(13) Chemical Computing Group Inc. Molecular Operating Environment (MOE). Montreal 2016.

(14) Labute, P.; Santavy, M. SiteFinder-Locating Binding Sites in Protein Structures http://www.chempcomp.com/journal/sitefind.htm%5Cnhttps://www.chemcomp.com/journal/sitefind.htm.

(15) Wu, L.; Viola, C. M.; Brzozowski, A. M.; Davies, G. J. Structural Characterization of Human Heparanase Reveals Insights into Substrate Recognition. *Nat. Struct. Mol. Biol.* **2015**, *22* (12), 1016–1022. https://doi.org/10.1038/nsmb.3136.

(16) Faham, S.; Hileman, R. E.; Fromm, J. R.; Linhardt, R. J.; Rees, D. C. Heparin Structure and Interactions with Basic Fibroblast Growth Factor. **1996**, *271* (5252), 1116–1120.

CHAPTER SEVEN

Binding of Per- and Polyfluoroalkyl Substances to the Human Pregnane X Receptor

About this chapter: This chapter is reprinted from Lai, T. T.; Eken, Y.; Wilson, A. K. Binding of Per- and Polyfluoroalkyl Substances to the Human Pregnane X Receptor. *Environ. Sci. Technol*. **2020**, *54 (24)*, 15986–15995 with the permission of American Chemical Society. Lai Thanh and Yiğitcan Eken contributed equally to this research by investigating interactions of half of the PFASs for hPXR.

7.1 Introduction

The production of perfluoroalkyl substances (PFASs) in the 1940s and 1950s is credited as an industrial breakthrough due to the unique properties of PFASs including water and oil repellency, high surface activity, and durability.[1] The use of these compounds has been widespread for food packaging, fire-fighting foams, carpet, furniture, boots, clothes, nonstick cookware, to name only a few.[2–4] PFASs are synthetic organofluorine compounds that have most (poly-) or all (per-) of their carbon-bonded hydrogens replaced with fluorine. PFASs are colloquially referred to as "zombie chemicals" or "forever chemicals," for their known resistance to degradation, which is caused by the strong electronegativity difference between their carbon−fluorine bonds.[5–7] Environmental and health concerns over the past two decades[8–15] have led to actions such as 3 M's voluntary perfluorooctane sulfonic acid (PFOS) phase-out in 2000[16] and EPA's 2006 perfluorooctanoic acid (PFOA) Stewardship Program.[17] "Long-chain" PFASs, defined as perfluoroalkyl carboxylic acids (PFCAs) with seven or more carbons and perfluorosulfonic acids (PFSAs) with six or more carbons forming their carbon backbone, have been slowly replaced with alternative PFASs, both "short-chain" variants and fluorinated alternatives, which typically have different functionalities.[18] Most common replacements are ADONA (trade name for 4,8-dioxa-3Hperfluorononanoic acid)[19] and Gen-X (trade name for 2,3,3,3- tetrafluoro-2-heptafluoropropoxy propanoic acid), which are used as alternatives to

PFOA.[20] 6:2 Fluorotelomer carboxylic acid (6:2 FTCA) is considered to be another alternative to PFOA, even though there has been no reported large-scale usage of the compound.[19,21–24] A limited number of studies have been done on the potential impact of alternative PFASs on the environment and to human health (see refs[19,20,25–27]). These studies suggest that alternative PFASs may exhibit comparable or even greater adverse health effects than their counterparts. The adverse effects of PFASs are believed to be chain length and functional group dependent, such that shorter PFASs or differently functionalized PFASs (such as ether groups in place of a number of fluorinated carbons) may be less toxic.[18,28,29] Thus, it is crucial to study molecular recognition of PFASs together with alternative PFASs in a fast and efficient manner. Yet, few studies have been performed on the structural differences of various PFASs and how the structure correlates to their binding.

PFASs are shown to interact with various human proteins such as thyroid hormone transport proteins, plasma proteins, liver fatty acid-binding protein, and also nuclear receptors such as pregnane X receptor (PXR), peroxisome proliferator activated receptors (PPARs), etc.[30–37] A number of epidemiological studies have suggested links between PFASs and adverse health effects such as adult thyroid problems, early childhood immunosuppression, as well as non-high-density lipoprotein (HDL)/total cholesterol.[38–40] Furthermore, PFASs such as PFOA have been shown to induce hepatic toxicity in mice as well as liver cancer in rodents.[33,41] As the PFAS chemical space (>4000 compounds) and the number of proteins that might interact with PFASs are considered, computational approaches are important.

Both the interaction and binding of PFASs to proteins play essential roles on their toxicity and bioaccumulation potential, and the prediction of their binding and interaction can be used as a proximity for their bioaccumulation and toxicity assessment.[42] In this study, we utilized *in*

silico methods based on molecular dynamics (MD) to investigate protein−PFAS interactions. To calculate binding affinities between PFAS and proteins, end-state approaches are selected due to their good balance between computational cost and accuracy.[43,44] To be more specific, molecular mechanics combined with Poisson−Boltzmann surface area (MM-PBSA) and molecular mechanics combined with Generalized Born surface area (MM-GBSA) are used in this investigation to predict relative binding energies.

hPXR is involved in a variety of biological and clinical functions such as xenobiotic and bile acid metabolism, steroid hormone homeostasis, and mediation of various drug−drug interactions.[45–47] Due to its large (1150 A^3) and flexible ligand binding cavity present on its ligand binding domain,[48] the hPXR is able to bind to a variety of ligands including naturally occurring steroids such as progestins, glucocorticoids, bile acids, and estrogens.[49] The binding of ligands to this domain is associated with an increased stability of the receptor, which mediates coactivator binding to the ligand-dependent activation function 2 (AF-2) surface and ultimately leads to the induction of hPXR. However, the exact molecular mechanisms are still elusive.[50,51] The induction of hPXR has been associated with hepatic steatosis, atherosclerosis, oxidative stress, lipid homeostasis, endocrine disruption effects, carcinogenesis, and adverse drug interactions.[46,52–56]

In this study, the molecular basis of the PFAS-induced activation on hPXR as well as the differences and similarities between how legacy and alternative (replacement) PFASs interact with hPXR are studied computationally (Table 7.1). Particularly useful for this study is the availability of both the crystal structure for the ligand binding domain (LBD) and experimental bioactivity data for a number of the PFASs investigated here for human pregnane X receptor (hPXR).[57,58] Molecular dynamics simulations (MD), residue−ligand interaction energy

calculations, alanine mutation studies, free energy of binding calculations, and hydrogen bond (H-bond) analysis are used to investigate relative binding energies of PFAS−hPXR complexes, hydrogen bond frequencies, and key residue−ligand interactions to produce a quantitative molecular-level description of PFAS−hPXR interactions. The various interaction patterns of PFAS−hPXR are compared, focusing on structural differences. Additionally, several PFOA alternatives, ADONA, Gen-X, 6:2 FTCA, and a short-chain PFSA variant, PFBS, are also included in this study to consider interactions with hPXR, as the agonistic activity of these species on hPXR was not previously determined.

7.2 Materials and Methods

Table 7. 1 Nomenclature for Perfluoroalkyl Substances (PFASs) Studied[a]

Type	Acronym	Perfluorinated Carbon	Name	Chemical Formula
PFCA	PFBA	3	perfluorobutanoic acid	CF_3-$(CF_2)_2$-COOH
PFCA	PFPA	4	perfluoropentanoic acid	CF_3-$(CF_2)_3$-COOH
PFCA	PFHxA	5	perfluorohexanoic acid	CF_3-$(CF_2)_4$-COOH
PFCA	PFHpA	6	perfluoroheptanoic acid	CF_3-$(CF_2)_5$-COOH
PFCA	PFOA	7	perfluorooctanoic acid	CF_3-$(CF_2)_6$-COOH
PFCA	PFNA	8	perfluorononanoic acid	CF_3-$(CF_2)_7$-COOH
PFCA	PFDA	9	perfluorodecanoic acid	CF_3-$(CF_2)_8$-COOH
PFCA	PFDoA	11	perfluorododecanoic acid	CF_3-$(CF_2)_{10}$-COOH
PFSA	PFBS	4	perfluorobutane sulfonic acid	CF_3-$(CF_2)_3$-SO_3H
PFSA	PFOS	8	perfluorooctane sulfonic acid	CF_3-$(CF_2)_7$-SO_3H
FTOH	6:2 FTOH	6	6:2 fluorotelomer alcohol	CF_3-$(CF_2)_5$-CH_2-OH
FTCA	6:2 FTCA	6	6:2 fluorotelomer carboxylic acid	CF_3-$(CF_2)_5$-CH_2-COOH
Alternative	Gen-X	5	2,3,3,3-tetrafluoro-2-heptafluoropropoxy propanoic acid	CF_3-$(CF_2)_2$-O-(CF_3)CF-COOH
Alternative	ADONA	6	4,8-dioxa-3H-perfluorononanoic acid	CF_3-O-$(CF_2)_3$-O-CHF-CF_2-COOH
PFSA = CF_3-$(CF_2)_n$-SO_3H				
PFCA = CF_3-$(CF_2)_n$-COOH				
FTOH = CF_3-$(CF_2)_n$-$(CH_2)_m$-OH				
FTCA = CF_3-$(CF_2)_n$-$(CH_2)_m$-COOH				

Note: The chemical structures of the compounds are provided in Table 7.3

7.2.1 Site Analysis and Molecular Docking

The hPXR protein structure was taken from the RSCB Protein Data Bank (PDB ID: 6DUP).[58] The Molecular Operating Environment's (MOE) Site Finder program was used to detect potential binding sites in the hPXR structure.[59] The site finder method detects α shapes on the protein structure and evaluates them according to their propensity of ligand binding (PLB).[60] The site that had the highest PLB score - a proven binding site for T1317 and rifampicin ligands - was used as the PFAS binding site.[61,62] Starting PFAS structures were obtained from PubChem.[63] Protonation states of the PFASs under the physiological conditions are determined by Protonate3D module implemented in MOE.[59,64] The resulting PFAS structures were minimized in MOE with the AMBER10: Extended Hückel Theory (EHT) force field, which uses Amber ff10 for macromolecules and Extended Hückel Theory for the ligands.[65–67] Ligand binding poses were determined by docking PFASs to the binding site using MOE. The London ΔG scoring function[68] was used to evaluate 100 initial ligand placements. Then, the initial placements were further refined to 10 poses via the Generalized Born Volume Integral/Weighted Surface area scoring function (GBVI/WSA) ΔG with induced-fit protein settings.[59,68] From these 10 refined poses, structurally distinct ones with the highest (GBVI/WSA) ΔG scores were selected for further studies.

7.2.2 Simulation Protocol

The selected complex structures were minimized via AMBER10:EHT in MOE. The topologies and the parameters for the minimized structures were created using the Leap module of Amber Tools[69] under the General Amber Force Field (GAFF), AMBER ff14sb force fields.[70] The AM1-BCC charge scheme[71] was used to calculate partial charges of the ligand atoms, and these partial charges were fit to GAFF using the Antechamber[69] suite to generate ligand

parameters. The protein−ligand complex structures were placed in a 14 Å3 beyond the solute box, neutralized and ionized with 100 mM NaCl ions using the parameters from Joung and Cheatham.[72]

The systems were minimized with decreasing energy restraints on the protein (500.0, 200.0, 20.0, 10.0, 5.0, 0.0 kcal mol^{-1}). Then, the systems were heated from 100 to 300 K in 30 ps MD simulation and equilibrated for 100 ps at 300 K. After equilibration, 30 ns MD simulations were performed to ensure the convergence of the system at 300 K and 1 atm pressure. During all simulations, the pressure was controlled by isotropic position scaling, the temperature was controlled by Langevin dynamics, and the time step was set to 2 fs. Furthermore, SHAKE algorithm[73] was used to constrain hydrogen bonds to allow the use of the 2 fs time step. Nonbonded interactions were truncated to 10 Å, while the particle-mesh Ewald (PME) method was used to efficiently approximate long-range electrostatic interactions.

7.2.3 Binding Energy Calculations

MM-PBSA and MM-GBSA methods are used for predicting the binding energies between PFASs and hPXR. These methods are based on subtracting the free energies of the unbound receptor and the ligand from the free energy of the ligand bound protein complex using the structures generated during MD simulations.[74]

$$\Delta G_{Bind} = G_{Complex} - G_{Protein} - G_{Ligand} \quad (1)$$

Many studies have demonstrated the success of these methods for finding relative binding affinities and ranking binding energies of molecules,[75,76] though very few of the studies have focused on PFASs.[37] While methods such as MM-PBSA and MM-GBSA have been useful, the methods are built upon different thermochemical approximations, and, thus, the predictions arising from these methods can be system dependent. For example, when the MM-GBSA and

MM-PBSA binding energies for six different protein−ligand systems including α-thrombin (7 ligands), avidin (7 ligands), cytochrome C peroxidase (18 ligands), neuraminidase (8 ligands), P450cam (12 ligands), and penicillopepsin (7 ligands) are compared, MM-GBSA results in: better correlation with experiments for α-thrombin, penicillopepsin, neuraminidase, similar correlation for avidin, and poorer correlation for cytochrome C peroxidase and P450cam in comparison to MM-PBSA.[76] Therefore, since the performances of MM-PBSA and MM-GBSA cannot be determined a priori, it is necessary to consider both methods for the PFAS− hPXR system and compare the results with the experiment, when available.

In this study, the binding free energies of the ligand−protein complexes were calculated using both MM-PBSA and MM-GBSA with a modified General Born solvation model by Onufriev et al.,[77] approaches implemented in the Amber PBSA-solver.[74] Default internal and external dielectric constants were used (1.0 and 80.0, respectively). The solvent-accessible surface area (SASA) was determined with the default linear combinations of pairwise overlap (LCPO) method using modified Bondi atomic radii. For both MM-PBSA and MM-GBSA, the frames from the first nanosecond of the MD simulations were used to calculate binding energies since it has been shown that such simulations can be useful, and that longer simulations do not necessarily correspond to a better accuracy.[76] Solute entropies were neglected because the primary focus of this effort was on the relative binding energies of PFASs on hPXR. Binding contributions of the residues at the binding site were calculated by per-residue decomposition and the energy contribution for each residue averaged from all poses tested.[69] Additionally, mutagenesis studies were performed by replacing target residues with the alanine from the complex structure, followed by MD and MM-GBSA efforts. The MM-GBSA electrostatic energies of these mutant complexes were compared with their wild-type counterparts.

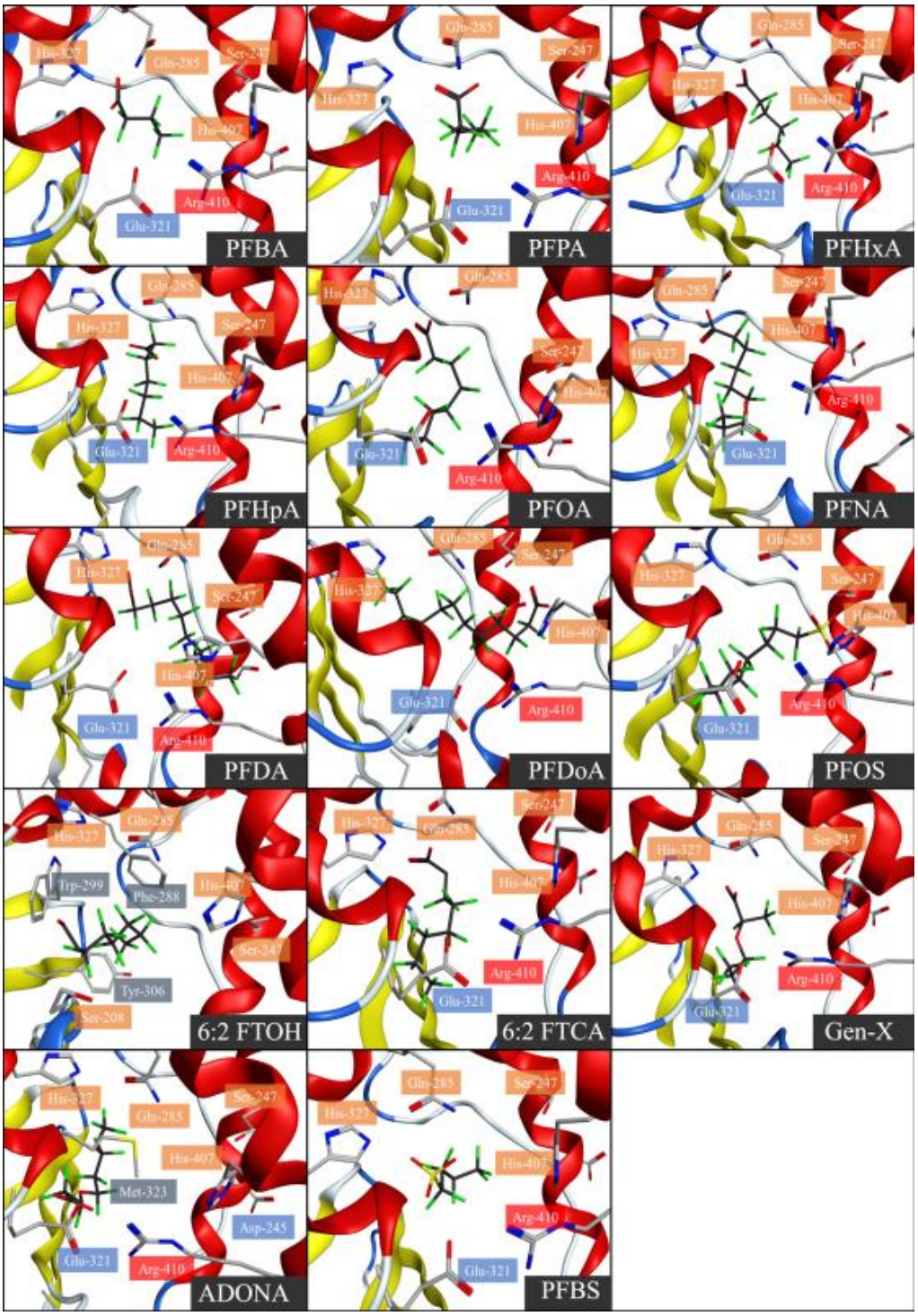


Figure 7. 1 Binding modes of PFASs to the hPXR ligand binding pocket.

7.2.4 Hydrogen Bond Analysis

Hydrogen bond lifetime analyses were performed via CPPTRAJ for every PFAS ligand.[78] For each PFAS, the PFAS−hPXR complex with the lowest MM-GBSA relative binding energy was selected for analysis. Ser-247, Gln-285, His-327, and His-407 were analyzed for hydrogen bond lifetimes.

7.3 Results and Discussion

7.3.1 Molecular Docking and MD Simulations

The binding poses of 14 PFASs that have the highest affinity to hPXR LBD, as determined by MM-GBSA free energy results, are provided in Figure 7.1. To account for the changes that occur in the binding domain upon PFAS binding, induced-fit docking is used for the generation of the binding poses. The docking algorithm allows for movement of the protein side chains together with the ligand in its refinement step, which ensures that the protein side chains are adjusted in accordance to the ligand structure. This type of approach is commonly used in computer-aided drug design with success,[68,79–85] especially for protein targets with a flexible binding domain such as the flexible binding domain encountered in this study for hPXR. Most of the PFAS binding modes have the carboxylate/ sulfonate group hydrogen bonding with Gln-285 and His-327, or Ser-247. It should be noted that Zhang et al. reported PFAS binding modes that hydrogen bond to Ser-247.57 In the current effort, 30 ns MD simulations are adopted to ensure system convergence. MD simulations show that poses that hydrogen bond with Gln-285 and His-327 are still able to hydrogen bond to Ser-247 with minor movements to the carboxylate/sulfonate group. For the most part, docking poses are preserved in MD simulations, and any ligand movements are often attributed to changes in hydrogen bonding partners of the carboxylate/sulfonate functional group.

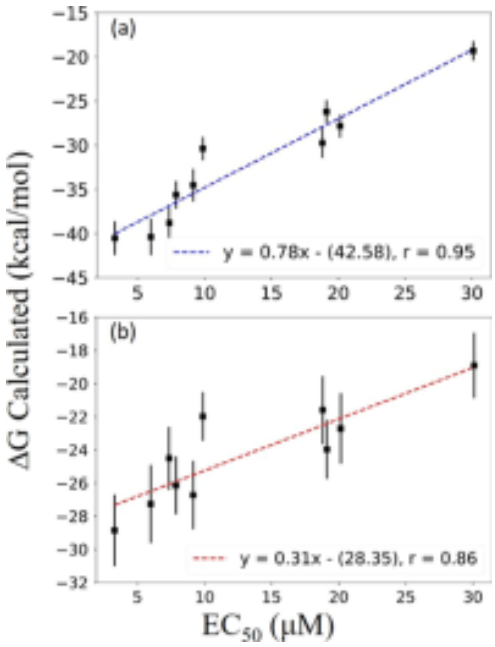


Figure 7. 2 (a) Correlation observed between experimental EC_{50} values from Zhang et al. and predicted binding free energies from MM-GBSA. (b) Correlation observed between EC_{50} values from Zhang et al. and predicted binding free energies from MM-PBSA. Error bars indicate standard deviations.

7.3.2 Binding Free Energy Calculations

The utilities of both MM-GBSA and MM-PBSA for PFAS−hPXR systems are first evaluated by comparing the predicted binding energies of PFBA, FPPA, PFHxA, PFHpA, PFOA, PFNA, PFDA, PFDoA, PFOS, and 6:2 FTOH with available experimental half maximal effective concentration (EC_{50}) data from Zhang et al. (Figure 7.2a,b, respectively).[57] Strong correlation between experimental EC_{50} values and predicted binding free energies is observed with both MM-GBSA and MM-PBSA methods with correlation coefficients of 0.95 and 0.86, respectively, and Kendal's Tau values of 0.96 and 0.69, respectively. Yet, MM-GBSA performed better on the PFAS−hPXR systems, proven by its slightly higher correlation coefficient and Kendal's Tau results compared to those of MM-PBSA. To assess the affinity and potential impact of ADONA,

6:2 FTCA, Gen-X, and PFBS upon hPXR binding, MM-GBSA calculations were expanded to include these alternative PFASs whose agonistic activity on hPXR has not been reported previously (Figure 7.3).

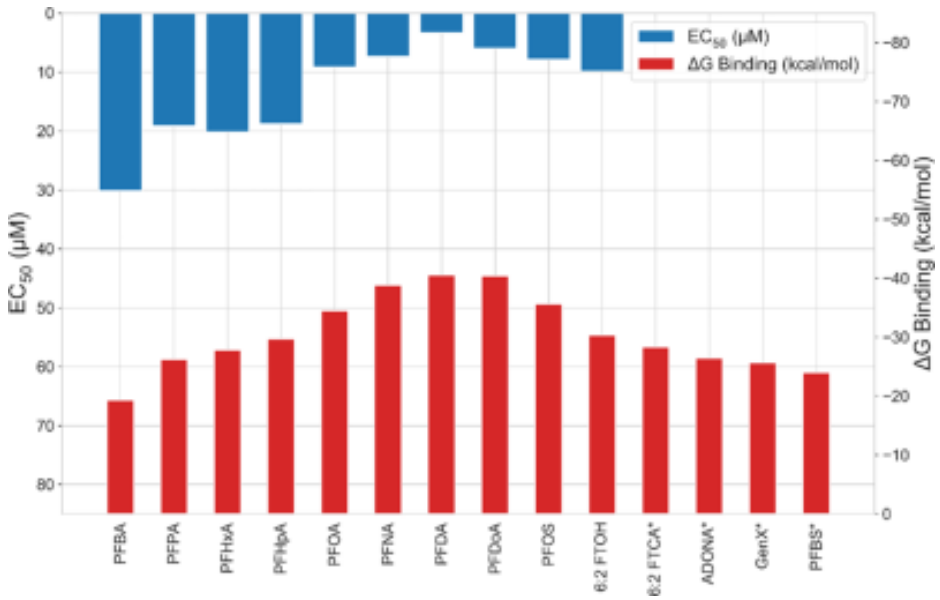


Figure 7. 3 Binding energies of PFASs to hPXR calculated with MM-GBSA in comparison to EC_{50} values measured by Zhang et al. (the predicted binding energies are listed in Table 7.4).

Predicted ΔGs of PFCAs (PFBA, PFPA, PFHxA, PFHpA, PFOA, PFNA PFDA, and PFDoA) suggest that as the perfluorinated carbon number increases, the affinity of the PFCAs to hPXR LBD also increases, explaining the relationship between the increased agonistic activity (EC_{50} values) measured with respect to increased perfluorinated carbon chain length (Figure 7. 3). When comparing PFSAs, increasing carbon chain length also leads to decreased affinity such that PFBS (four carbons) is higher in relative binding energy (+11.6 kcal mol^{-1}) than PFOS (8 carbons). Finally, binding energies of ADONA, Gen-X, and 6:2 FTOH show a lower affinity to hPXR compared to PFOA. However, ADONA, Gen-X, and 6:2 FTOH's binding to hPXR are predicted as similar to PFPA and PFHxA, indicating even though the binding energies are lower than PFOA, they still exhibit binding and may show agonistic activity.

Table 7. 2 hPXR Residues Interact with PFASs Upon Binding

Ligand	**H-Bonding Residues**	**Largest Energy Contributors**
PFBA	Ser-247, His-407	Arg-410, Lys-210, Lys-226, His-407, Ser-247
PFPA	Gln-285, His-327, His-407	Arg-410, Lys-210, Lys-226, His-407, Gln-285
PFHxA	Ser-247, Gln-285	Arg-410, Lys-210, Lys-226, His-407, Gln-285
PFHpA	Ser-247, His-407	Arg-410, Lys-210, Lys-226, His-407, Ser-247
PFOA	Ser-247, His-327	Arg-410, Lys-210, Lys-226, His-407, Ser-247
PFNA	Gln-285, His-327	Arg-410, Lys-210, Lys-226, His-407, Gln-285
PFDA	Gln-285, His-327, His-407	Arg-410, Lys-210, Lys-226, His-407, Gln-285
PFDoA	Ser-247, His-407	Arg-410, Lys-210, Lys-226, His-407, Gln-285
PFBS	Ser-247, His-407	Arg-410, Lys-210, Lys-226, His-407, Gln-285
PFOS	Ser-247, Gln-285, His-407	Arg-410, Lys-210, Lys-226, His-407, Gln-285
6:2 FTOH	—	Trp-299, Ser-208, Phe-288, Tyr-306, Gln-285
6:2 FTCA	Ser-247	Arg-410, Lys-210, Lys-226, His-407, Ser-247
Gen-X	Ser-247, His-407	Arg-410, Lys-210, Lys-226, His-407, Ser-247
ADONA	—	Arg-410, Lys-210, Lys-226, Met-323, His-327

7.3.3 PFAS Recognition on hPXR

Residue decomposition is employed to understand molecular recognition of PFASs on hPXR and can also be used to provide insight about the activity of untested PFASs on hPXR. Residue decomposition shows that Lys-210, Lys-226, Ser-247, Gln-285, His-327, His-407, and Arg-410 are among the largest energy contributors for PFAS−hPXR binding for all PFASs tested, except 6:2 FTOH, which does not possess an acidic functional group (Table 7.2 and Figure 7.7).

The binding energies for the top three residues are quite similar between the short/alternative PFASs and long-chain PFASs. Among the binding site residues, Arg-410 has the lowest

interaction energy for both PFASs at ~−40 kcal mol^{-1}, followed by Lys-210 at ~−25 kcal mol^{-1}, and Lys-226 at ~−17 kcal mol^{-1}, with the exception of 6:2 FTOH, where the functional group is an alcohol rather than an acid. On the contrary, the contribution to the binding from Ser-247, Gln285, His-327, and His-407 varies according to the ligand with a range from −5 to −12 kcal mol^{-1}. The energy contributions of Lys-210 and Arg-410 tend to increase as the carbon chain length increases. Since both Lys-210 and Arg-410 are located near the entrance of the cavity (Figure 7.1 and Figure 7.8), their interaction primarily arises from long-range electrostatic forces, rather than from short-range hydrogen bonding or van der Waals interaction. Unlike Lys-210, Lys-226, and Arg-410, which interact strongly with almost every PFAS studied, binding energy contribution of Ser-247, Gln-285, His-327, and His-407 is alternating for different PFASs. To better understand the H-bonding behavior of hPXR residues interact with the PFASs, insight is gained about the hydrogen bond lifetimes using MD trajectories and the results showed that Ser-247, Gln-285, His-327, and His-407 commonly make hydrogen bonds with PFASs.

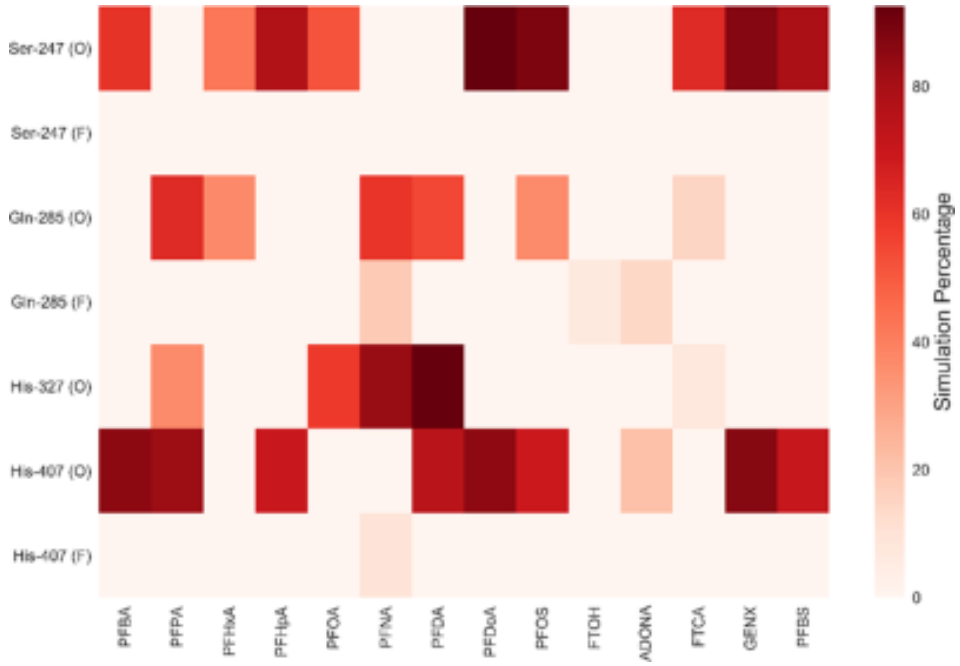


Figure 7. 4 Hydrogen bond lifetimes observed during MD simulations.

Hydrogen bond lifetime analysis shows that the stability of the H-bond between Gln-285, Ser-247, His-327, His-407, and PFASs is ligand dependent as the H-bonding lifetimes vary (Figure 7.4). The PFASs' carboxylic acid and sulfonic acid functional groups often engage in hydrogen bonding with the hydroxyl, amide, and imidazole groups in Gln-285, Ser-247, His-327, and His-407, respectively (Figure 7.5). On the other hand, the fluorine atoms present on the PFAS's perfluorinated carbon chains do not form any significant hydrogen bonds during simulations (Figure 7.4).

Finally, despite residue decomposition which shows that Lys-210, Lys-226, and Arg-410 contribute significantly to the binding of PFASs to hPXR, Lys-210, Lys-226, and Arg-410 do not form hydrogen bond with PFASs. This further supports that PFASs interact mainly through long-range electrostatics rather than short-range interactions such as through hydrogen bonding with hPXR. The lack of hydrogen bonding for the Lys-210, Lys-226, and Arg-410 may be attributed to the orientation of PFASs within the hPXR binding site. Carboxylic acid and sulfonic acid functional groups facing inside of the binding cavity is commonly observed upon PFAS−hPXR binding, allowing PFASs to hydrogen bond with cavity residues such as Ser-247, Gln-285, His-327, and His-407. In contrast, Lys-210, Lys-226, and Arg-410 are located near the entrance of the binding cavity and could not form hydrogen bonds. 6:2 FTOH, which contains an alcohol rather than an acidic group, does not form any significant hydrogen bond throughout the simulations (Figure 7. 4).

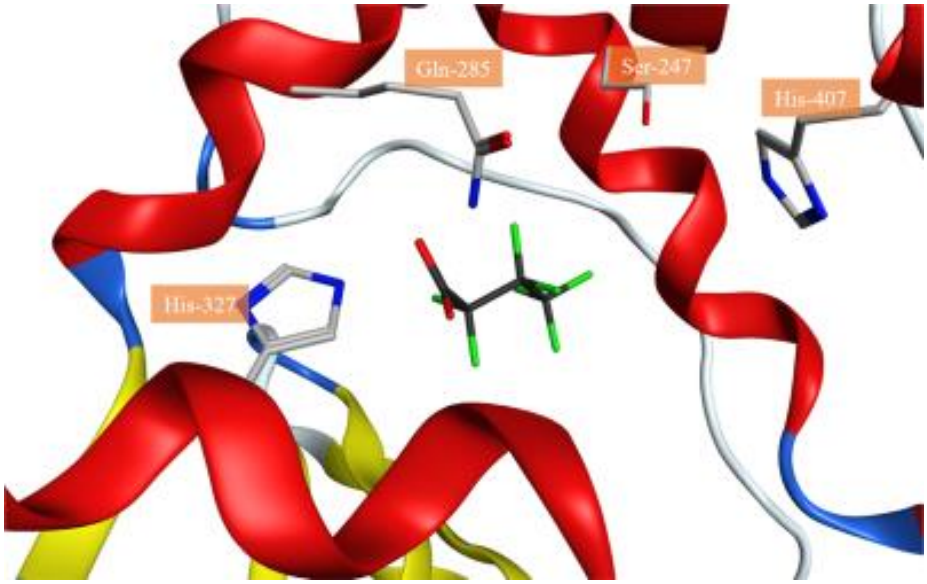


Figure 7. 5 Important residues that mediate ligand stability through hydrogen bonding.

Residue decomposition results showed that Asp-205, Asp-245, and Glu-321 destabilize the binding of PFASs to the hPXR LBD. The destabilizations most likely arise from the repulsion between the negative charge of aspartic acid, glutamic acid residues, and the negative charge of PFASs present on the carboxylic acid or sulfonic acid functional groups. Mutagenesis of Asp-205, Asp-245, and Glu-321 to alanine in selected ligand−protein complexes (PFOS, PFOA, ADONA, Gen-X, 6:2 FTCA, and PFBS) showed an overall decrease in total electrostatic energy (EEL) contribution for every ligand mutant complex (Figure 7. 6).

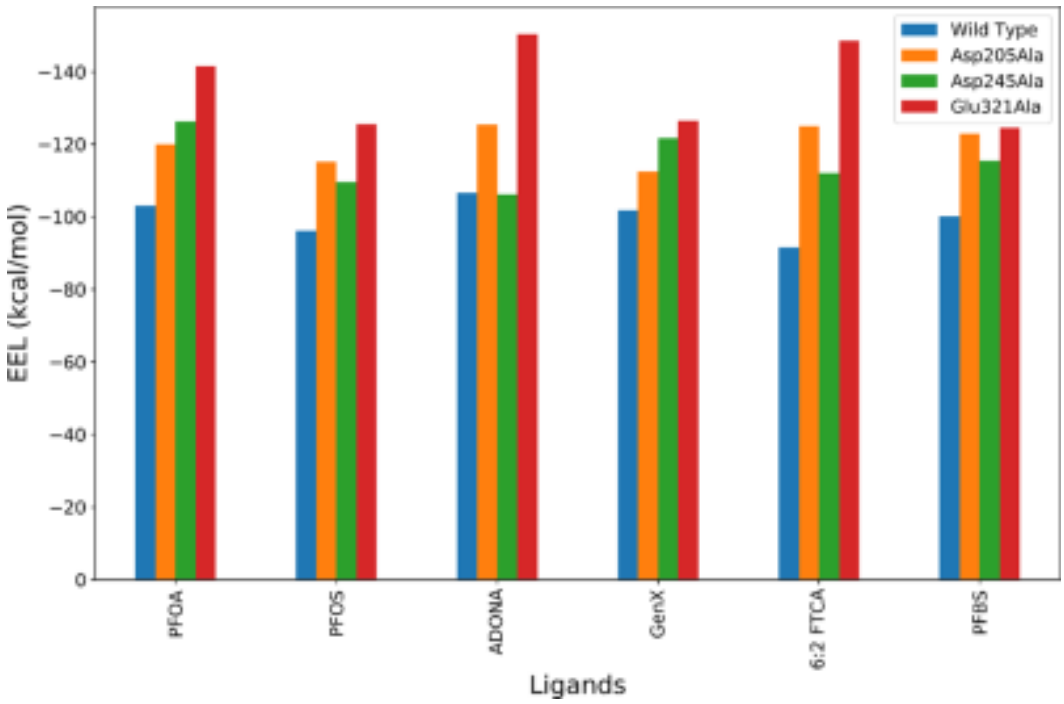


Figure 7. 6 Total electrostatic energy (EEL) contribution of various PFASs on binding to mutant hPXR complexes.

When compared to the EEL energies of the wild-type ligand−protein complexes, the presence of Glu-321 reduces the favorable EEL contribution by an average of −39.51 kcal/ mol, Asp-245 reduces EEL contribution by an average of −16.73 kcal mol^{-1}, and Asp-205 reduces it by an average of −25.53 kcal mol^{-1}. This implies that the net negative charges of Asp-205, Asp-245, and Glu-321 destabilize the binding of PFASs to hPXR, and proteins with more acidic residues in their binding pockets are less likely to be PFAS targets, which has implications on the evaluation of potential PFAS protein targets.

Residue decomposition and hydrogen bond analysis provide an understanding about how the chemical structure of PFASs affects their binding behavior. The results indicate that carboxylate/sulfonate functional groups on the PFAS's structure contribute strongly to its hPXR binding through long-range electrostatic interactions with Arg-410, Lys-210, and Lys-226 and

H-bonding with Gln-285, Ser-247, His-327, and His-407, and that ADONA, Gen-X, PFBS, and 6:2 FTCA are potential hPXR agonists. Thus, at least for hPXR, these efforts suggest that further insight about the impact of PFAS without a carboxylic acid or sulfonic acid functional group should be garnered to identify alternative PFASs that are less potent to hPXR and other proteins.

APPENDIX

Table 7. 3 All PFAS ligands tested.

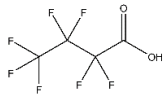
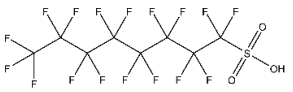
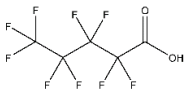
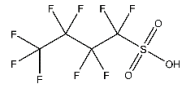
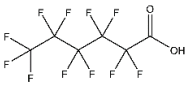
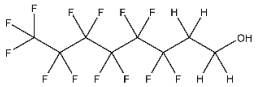
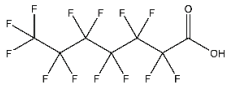
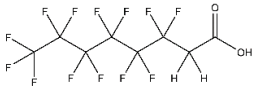
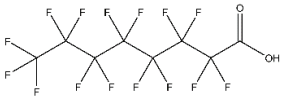
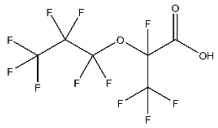
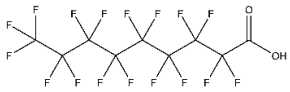
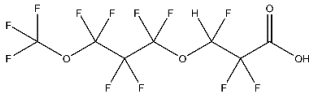
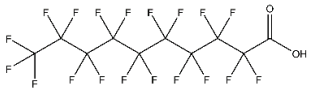
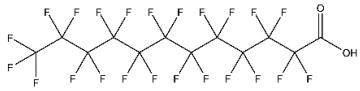
Structure	Name	Structure	Name
	Perfluorobutanoic Acid (PFBA, CAS No. 375-22-4)		Perfluorooctanesulfonic Acid (PFOS, CAS No. 1763-23-1)
	Perfluoropentanoic Acid (PFPA, CAS No. 2706-90-3)		Perfluorobutanesulfonic Acid (PFBS, CAS No. 375-73-5)
	Perfluorohexanoic Acid (PFHxA CAS No. 307-24-4)		6:2 Fluorotelomer Alcohol (6:2 FTOH, CAS No. 647-42-7)
	Perfluoroheptanoic Acid (PFHpA, CAS No. 375-85-9)		6:2 Fluorotelomer Carboxylic Acid (6:2 FTCA)
	Perfluorooctanoic Acid (PFOA, CAS No. 335-67-1)		2,3,3,3-tetrafluoro-2-heptafluoropropoxypropanoic acid (GenX CAS No. 62037-80-3)
	Perfluorononanoic Acid (PFNA, CAS No. 375-95-1)		ADONA (CAS No. 958445-448)
	Perfluorodecanoic Acid (PFDA, CAS No. 335-76-2)		
	Perfluorododecanoic Acid (PFDoA, CAS No. 307-55-1)		

Table 7. 4 MMPBSA and MMGBSA relative binding energies of every PFAS tested.

Ligands	MMPBSA	MMGBSA
PFBA	-18.91±7.9	-19.31±4.4
PFPA	-23.97±6.2	-26.19±4.6
PFHxA	-22.71±11.4	-27.84±7.0
PFHpA	-21.60±8.4	-29.74±6.9
PFOA	-26.72±7.3	-34.51±6.7
PFNA	-24.51±6.7	-38.81±6.2
PFDA	-28.85±11.6	-40.52±10.3
PFDoA	-27.27±10.2	-40.38±8.3
PFOS	-26.14±6.1	-35.61±5.4
PFBS	-22.99±6.9	-23.96±5.2
6:2 FTOH	-21.98±5.39	-30.37±4.59
6:2 FTCA	-19.72±7.9	-28.31±6.4
GEN X	-23.27±7.6	-25.64±4.7
ADONA	-22.68±10.7	-26.45±8.0

Table 7. 5 Long-chain PFAS average *per-residue decomposition* energies (kcal mol^{-1}).

PFOA		PFNA		PFDA		PFDoA	
Residue	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**
Lys-210	-20.97	**Lys-210**	-23.94	**Lys-210**	-21.59	**Lys-210**	-24.85
Lys-226	-16.86	**Lys-226**	-15.60	**Lys-226**	-15.94	**Lys-226**	-16.54
Ser-247	-13.59	**Ser-247**	-6.55	**Ser-247**	-7.56	**Ser-247**	-5.22
Gln-285	-7.61	**Gln-285**	-9.21	**Gln-285**	-8.40	**Gln-285**	-7.63
His-327	-4.05	**His-327**	-8.23	**His-327**	-6.40	**His-327**	-4.82
His-407	-15.32	**His-407**	-9.73	**His-407**	-14.20	**His-407**	-12.75
Arg-410	-33.53	**Arg-410**	-36.67	**Arg-410**	-42.49	**Arg-410**	-48.64
PFOS		**6:2 FTOH**		**Average PFAS Binding**			
Residue	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**		
Lys-210	-22.75	**Ser-208**	-4.75	**Lys-210**	-18.96		
Lys-226	-17.12	**Leu-209**	-2.01	**Lys-226**	-13.65		
Ser-247	-7.57	**Gln-285**	-2.14	**Ser-247**	-6.81		
Gln-285	-9.30	**Phe-288**	-3.46	**Gln-285**	-7.38		
His-327	-4.56	**Trp-299**	-4.86	**His-327**	-4.82		
His-407	-11.45	**Tyr-306**	-2.99	**His-407**	-10.90		
Arg-410	-46.82	**Met-323**	-2.07	**Arg-410**	-34.65		

Residues with interactions lower than -5 kcal mol^{-1} are shown. The major residues (Lys-210, Lys-226, Ser-247, Gln-285, His-327, and Arg-410) are listed regardless of their interaction energy.

Table 7. 6 Short-chain/alternative PFAS average *per-residue decomposition* energies.

PFBA		**PFPA**		**PFHxA**		**PFHpA**	
Residue	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**
Lys-210	-21.38	**Lys-210**	-21.42	**Lys-210**	-21.29	**Lys-210**	-22.54
Lys-226	-15.78	**Lys-226**	-16.51	**Lys-226**	-17.49	**Lys-226**	-17.66
Ser-247	-9.12	**Ser-247**	-2.26	**Ser-247**	-8.271	**Ser-247**	-8.65
Gln-285	-7.91	**Gln-285**	-10.53	**Gln-285**	-9.68	**Gln-285**	-7.61
His-327	-4.19	**His-327**	-8.22	**His-327**	-4.98	**His-327**	-4.85
His-407	-15.26	**His-407**	-14.12	**His-407**	-11.90	**His-407**	-10.48
Arg-410	-35.20	**Arg-410**	-38.03	**Arg-410**	-37.49	**Arg-410**	-36.69
ADONA		**GEN X**		**6:2 FTCA**		**PFBS**	
Residue	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**	**Residue**	**Average ΔG Bind**
Lys-210	-29.70	**Lys-210**	-23.19	**Lys-210**	-22.69	**Lys-210**	-22.26
Lys-226	-19.51	**Lys-226**	-17.87	**Lys-226**	-17.46	**Lys-226**	-16.24
Ser-247	-0.89	**Ser-247**	-8.00	**Ser-247**	-7.95	**Ser-247**	-12.17
Gln-285	-3.64	**Gln-285**	-5.63	**Gln-285**	-6.67	**Gln-285**	-12.62
His-327	-5.386	**His-327**	-4.53	**His-327**	-4.79	**His-327**	-3.60
His-407	-3.92	**His-407**	-15.52	**His-407**	-11.08	**His-407**	-14.38
Arg-410	-55.74	**Arg-410**	-42.20	**Arg-410**	-44.30	**Arg-410**	-34.15
Average PFAS Binding							
Residue	**Lys-210**	**Lys-226**	**Ser-247**	**Gln-285**	**His-327**	**His-407**	**Arg-410**
Average ΔG Bind	-22.81	-17.3141	-7.16	-8.04	-5.07	-12.08	-40.48

ADONA, GEN X, and 6:2 FTCA are alternatives of PFOA. PFBS is a short chain variant of PFOS. PFBA, PFPA, PFHxA, and PFHpA are short chain variants of the long chain perfluoroalkyl carboxylic acids (Table 1). Residues with interactions lower than -5 kcal mol^{-1} are shown. The major residues (Lys-210, Lys-226, Ser-247, Gln-285, His-327, and Arg-410) are listed regardless of their interaction energy.

Table 7. 7 Total electrostatic energies of various mutant PFAS-hPXR complexes.

Mutagenesis of Asp-250, Asp-245, Glu-321 MMGBSA Total Electrostatic Energies				
	Wild Type	**Asp205Ala**	**Asp245Ala**	**Glu321Ala**
PFOA	-102.96	-119.91	-126.25	-141.48
PFOS	-96.23	-115.21	-109.64	-125.50
PFBS	-100.09	-122.87	-115.48	-124.49
ADONA	-106.59	-125.45	-106.25	-150.38
6:2 FTCA	-91.58	-124.97	-112.12	-148.52

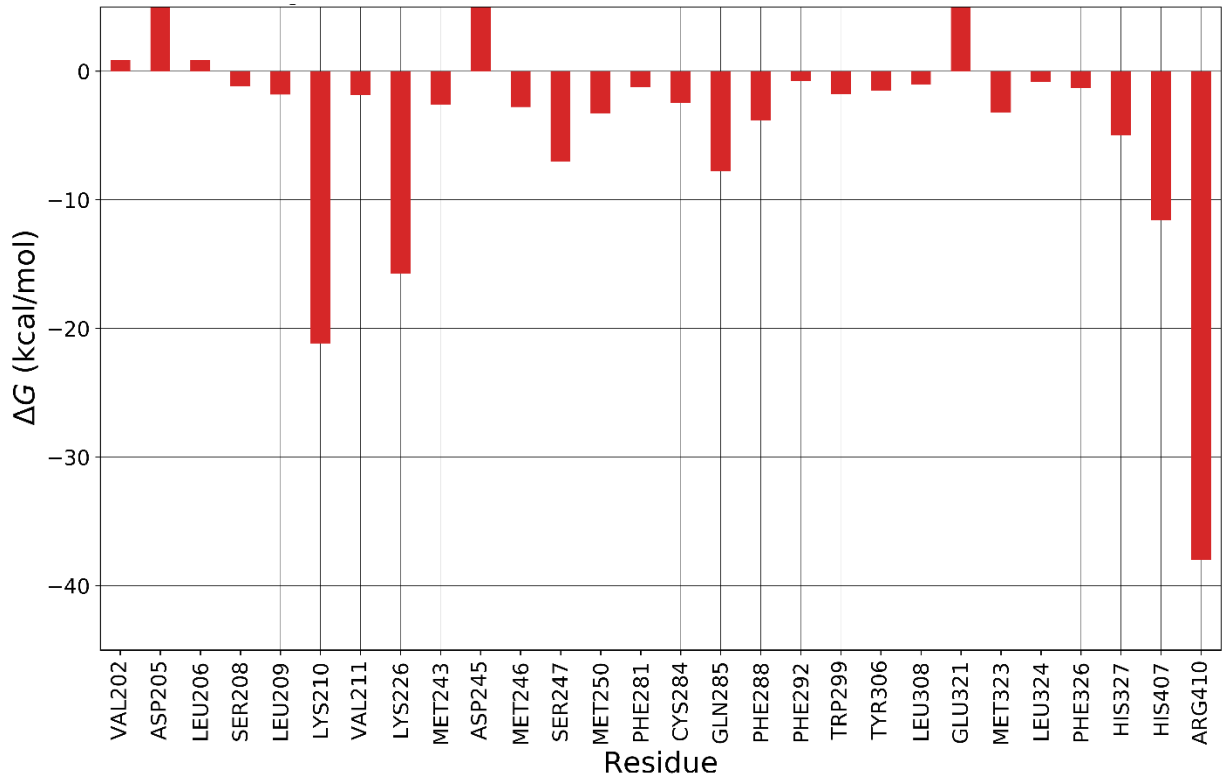


Figure 7. 7 Average residue contributions to the PFAS binding to hPXR calculated from residue decomposition.

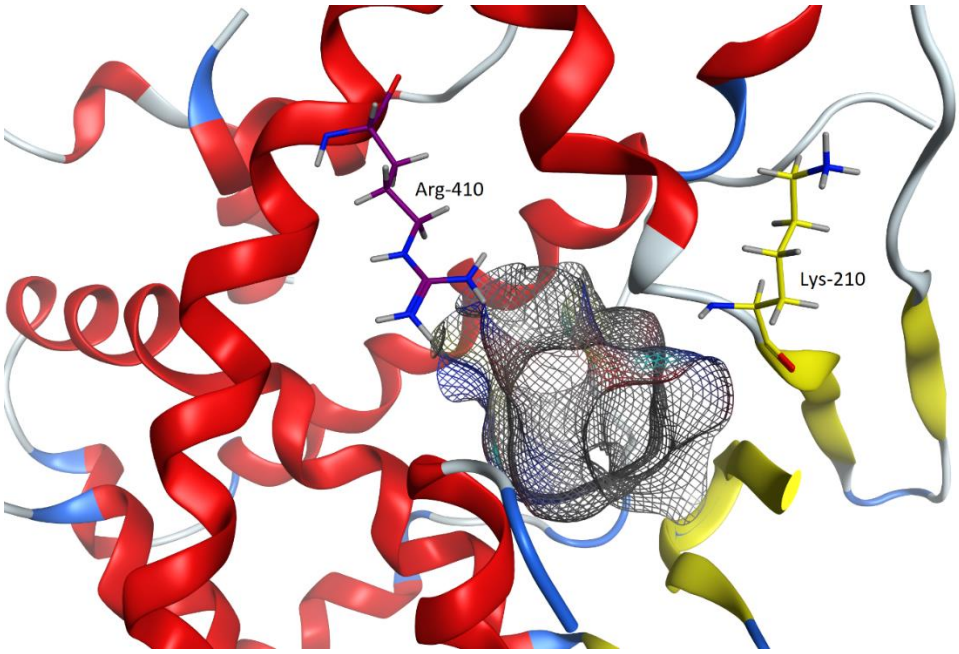


Figure 7. 8 Arg-410 and Lys-210 positioned outside of the binding cavity.

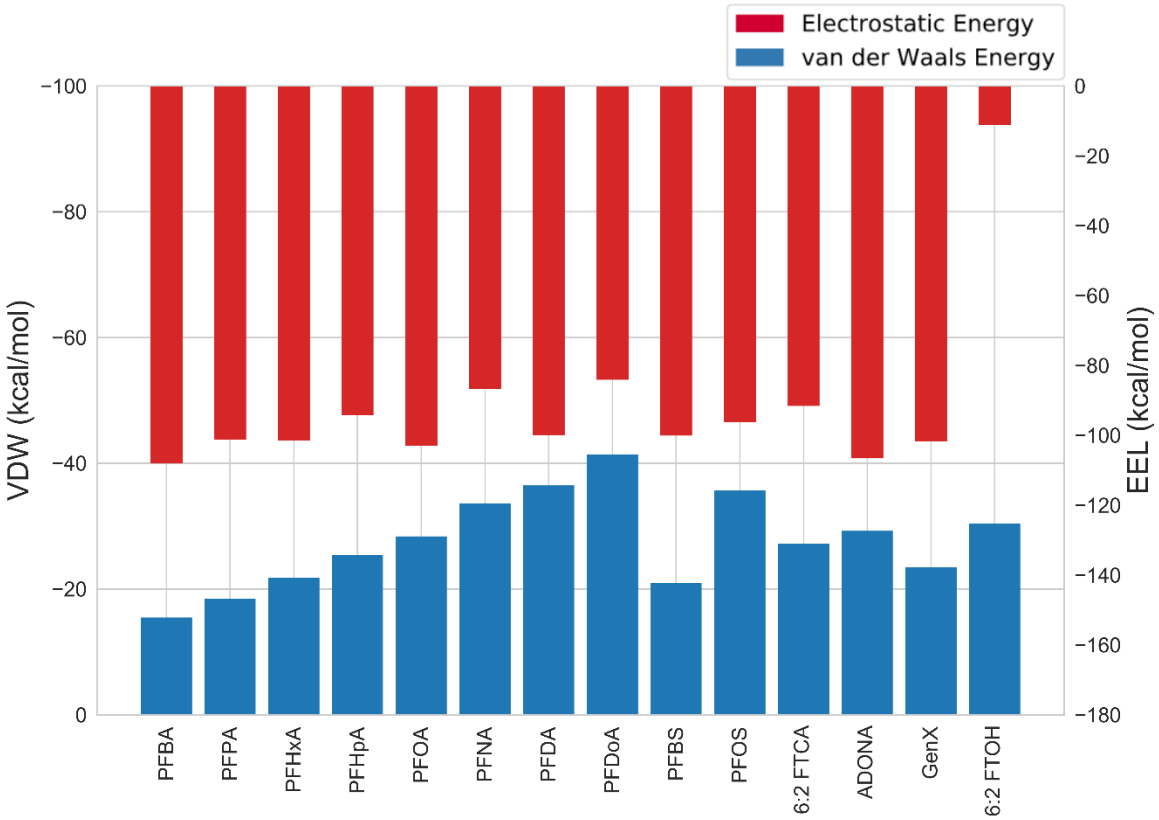


Figure 7. 9 Comparison of VDW and electrostatic energies of every tested ligand.

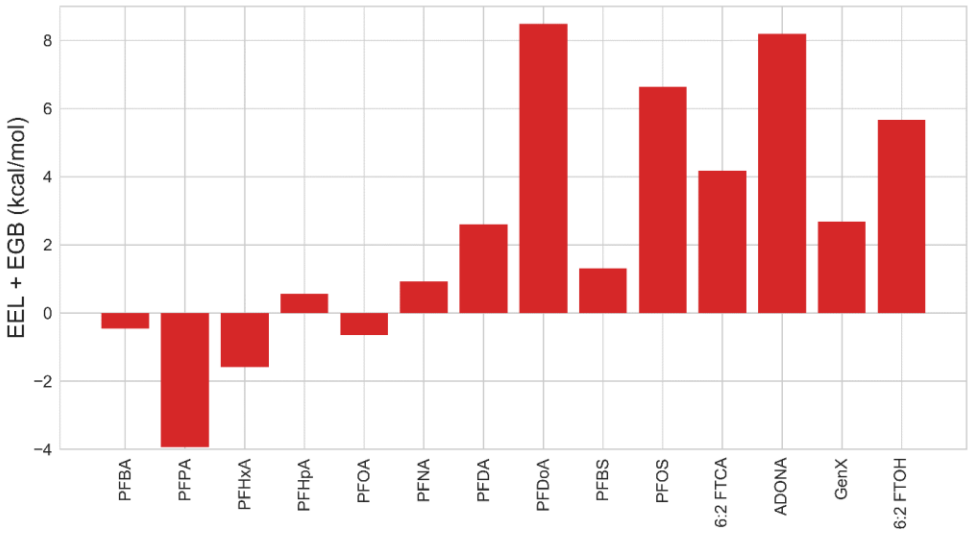


Figure 7. 10 Electrostatic energies + energy of solvation calculated by MMGBSA for every tested ligand.

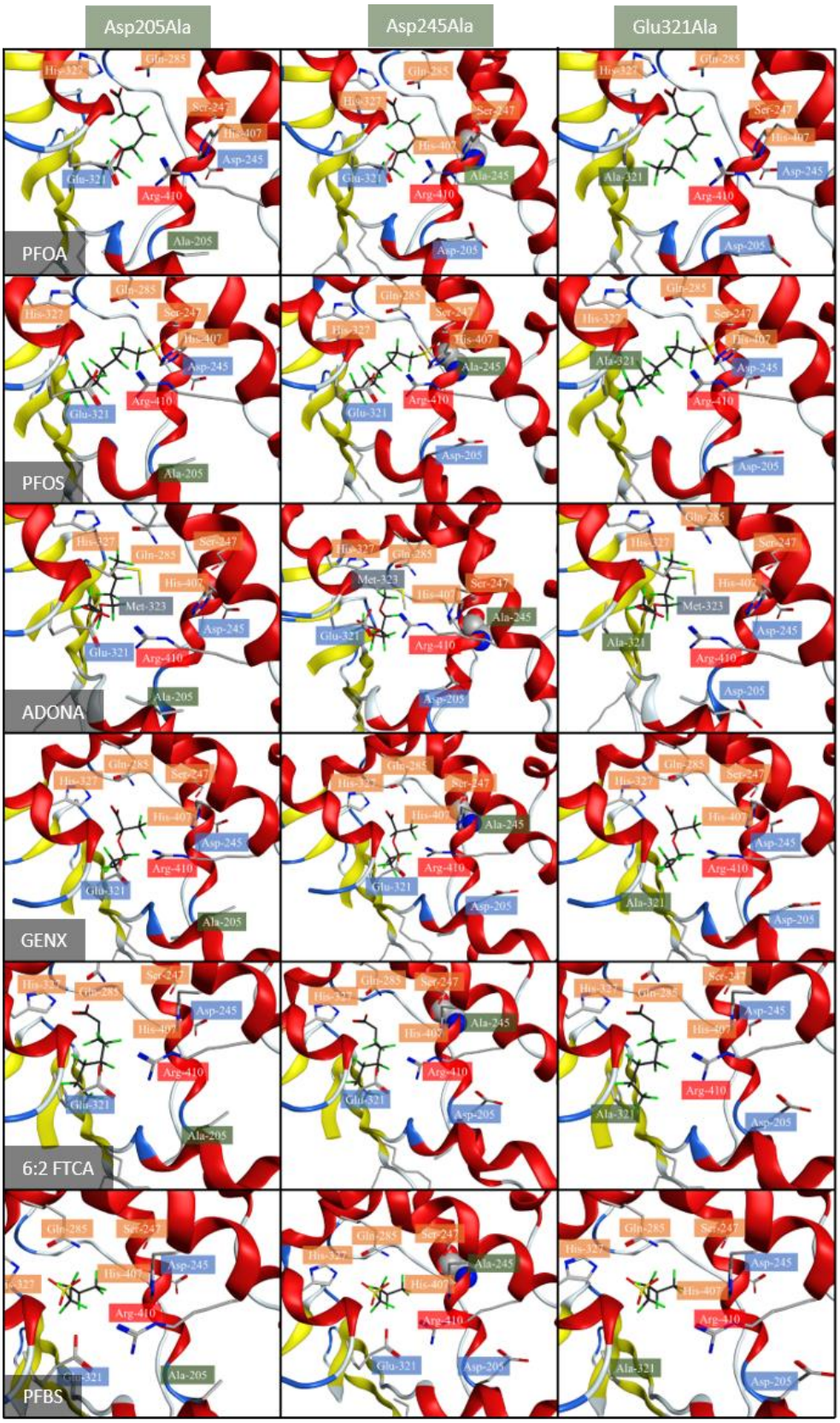


Figure 7. 11 Binding modes of PFASs to mutant hPXR ligand binding pocket.

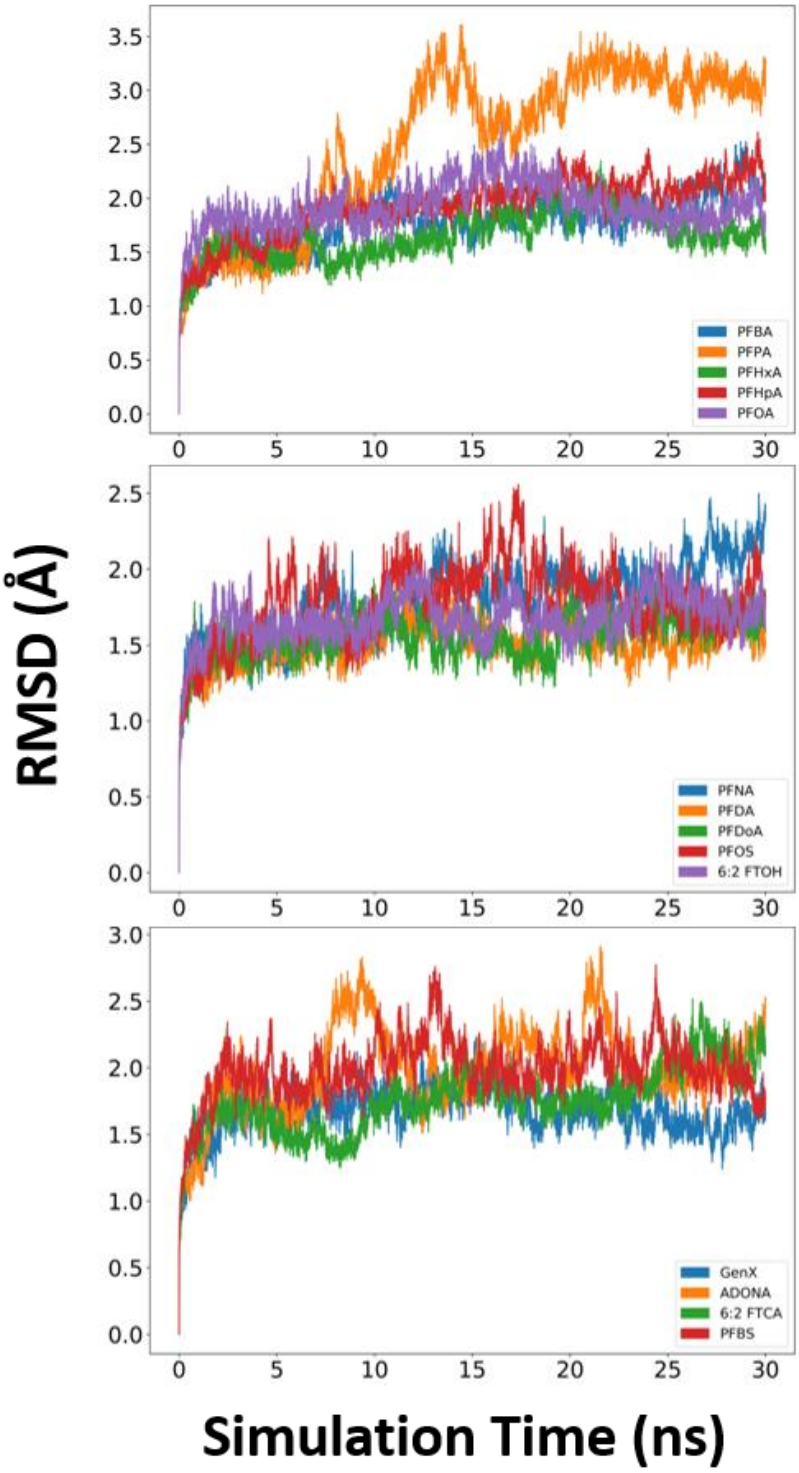


Figure 7. 12 Root mean square deviation (RMSD) plots of the highest affinity PFAS poses from 30ns MD simulations.

REFERENCES

REFERENCES

(1) Kissa, E. *Fluorinated Surfactants and Repellents*, 2nd ed.; Schick, M., Hubbard, A., Eds.; Marcel Dekker: New York, 2001; Vol. 97.

(2) Schaider, L. A.; Balan, S. A.; Blum, A.; Andrews, D. Q.; Strynar, M. J.; Dickinson, M. E.; Lunderberg, D. M.; Lang, J. R.; Peaslee, G. F. Fluorinated Compounds in U.S. Fast Food Packaging. *Environ. Sci. Technol. Lett.* **2017**, *4* (3), 105–111. https://doi.org/10.1021/acs.estlett.6b00435.

(3) Rao, N. S.; Baker, B. E. Textile Finishes and Fluorosurfactants. In *Organofluorine Chemistry*; Banks, R. E., Smart, B. E., Tatlow, J. C., Eds.; Springer US: Boston, MA, 1994; pp 321–338. https://doi.org/10.1007/978-1-4899-1202-2_15.

(4) Sajid, M.; Ilyas, M. PTFE-Coated Non-Stick Cookware and Toxicity Concerns: A Perspective. *Environ. Sci. Pollut. Res.* **2017**, *24* (30), 23436–23440. https://doi.org/10.1007/s11356-017-0095-y.

(5) Matheny, K. PFAS contamination is Michigan's biggest environmental crisis in 40 years.

(6) Gardner, P.; Ellison, G. Michigan's next water crisis is PFAS - and you may already be affected.

(7) O'Hagan, D. Understanding Organofluorine Chemistry. An Introduction to the C–F Bond. *Chem. Soc. Rev.* **2008**, *37* (2), 308–319. https://doi.org/10.1039/B711844A.

(8) Conder, J. M.; Hoke, R. A.; Wolf, W. de; Russell, M. H.; Buck, R. C. Are PFCAs Bioaccumulative? A Critical Review and Comparison with Regulatory Criteria and Persistent Lipophilic Compounds. *Environ. Sci. Technol.* **2008**, *42* (4), 995–1003. https://doi.org/10.1021/es070895g.

(9) Biege, L. B.; Hurtt, M. E.; Frame, S. R.; O'Connor, J. C.; Cook, J. C. Mechanisms of Extrahepatic Tumor Induction by Peroxisome Proliferators in Male CD Rats. *Toxicol. Sci.* **2001**, *60* (1), 44–55. https://doi.org/10.1093/toxsci/60.1.44.

(10) Yang, Q.; Xie, Y.; Depierre, J. W. Effects of Peroxisome Proliferators on the Thymus and Spleen of Mice. *Clin. Exp. Immunol.* **2000**, *122* (2), 219–226. https://doi.org/10.1046/j.1365-2249.2000.01367.x.

(11) Yang, Q.; Xie, Y.; Eriksson, A. M.; Nelson, B. D.; DePierre, J. W. Further Evidence for the Involvement of Inhibition of Cell Proliferation and Development in Thymic and Splenic Atrophy Induced by the Peroxisome Proliferator Perfluorooctanoic Acid in Mice. *Biochem. Pharmacol.* **2001**, *62* (8), 1133–1140. https://doi.org/10.1016/S0006-2952(01)00752-3.

(12) Yang, Q.; Abedi-Valugerdi, M.; Xie, Y.; Zhao, X.-Y.; Möller, G.; Dean Nelson, B.; DePierre, J. W. Potent Suppression of the Adaptive Immune Response in Mice upon

Dietary Exposure to the Potent Peroxisome Proliferator, Perfluorooctanoic Acid. *Int. Immunopharmacol.* **2002**, *2* (2–3), 389–397. https://doi.org/10.1016/S1567-5769(01)00164-3.

(13) Yang, Q.; Xie, Y.; Alexson, S. E. H.; Dean Nelson, B.; DePierre, J. W. Involvement of the Peroxisome Proliferator-Activated Receptor Alpha in the Immunomodulation Caused by Peroxisome Proliferators in Mice. *Biochem. Pharmacol.* **2002**, *63* (10), 1893–1900. https://doi.org/10.1016/S0006-2952(02)00923-1.

(14) Giesy, J. P.; Kannan, K. Global Distribution of Perfluorooctane Sulfonate in Wildlife. *Environ. Sci. Technol.* **2001**, *35* (7), 1339–1342. https://doi.org/10.1021/es001834k.

(15) Langley, A. E.; Pilcher, G. D. Thyroid, Bradycardic and Hypothermic Effects of Perfluoro-n-decanoic Acid in Rats. *J. Toxicol. Environ. Health* **1985**, *15* (3–4), 485–491. https://doi.org/10.1080/15287398509530675.

(16) US EPA. EPA and 3M announce phase out of PFOS.

(17) US EPA. Fact Sheet: 2010/2015 PFOA Stewardship Program https://www.epa.gov/assessing-and-managing-chemicals-under-tsca/%0Afact-sheet-20102015-pfoa-stewardship-program (accessed Apr 29, 2019).

(18) Buck, R. C.; Franklin, J.; Berger, U.; Conder, J. M.; Cousins, I. T.; Voogt, P. De; Jensen, A. A.; Kannan, K.; Mabury, S. A.; van Leeuwen, S. P. J. Perfluoroalkyl and Polyfluoroalkyl Substances in the Environment: Terminology, Classification, and Origins. *Integr. Environ. Assess. Manag.* **2011**, *7* (4), 513–541. https://doi.org/10.1002/ieam.258.

(19) Wang, Y.; Chang, W.; Wang, L.; Zhang, Y.; Zhang, Y.; Wang, M.; Wang, Y.; Li, P. A Review of Sources, Multimedia Distribution and Health Risks of Novel Fluorinated Alternatives. *Ecotoxicol. Environ. Saf.* **2019**, *182*, 109402. https://doi.org/10.1016/j.ecoenv.2019.109402.

(20) Ahearn, A. A Regrettable Substitute: The Story of GenX. *Pod. Res. Perspect.* **2019**, *2019* (1), EHP5134. https://doi.org/10.1289/EHP5134.

(21) Poulsen, P. B.; Jensen, A. A.; Wallström, E.; Aps, E. *More Environmentally Friendly Alternatives to PFOS-Compounds and PFOA*; 2005.

(22) Wang, Z.; Cousins, I. T.; Scheringer, M.; Hungerbühler, K. Fluorinated Alternatives to Long-Chain Perfluoroalkyl Carboxylic Acids (PFCAs), Perfluoroalkane Sulfonic Acids (PFSAs) and Their Potential Precursors. *Environ. Int.* **2013**, *60*, 242–248. https://doi.org/10.1016/j.envint.2013.08.021.

(23) Sagisaka, M.; Ito, A.; Kondo, Y.; Yoshino, N.; Ok Kwon, K.; Sakai, H.; Abe, M. Effects of Fluoroalkyl Chain Length and Added Moles of Oxyethylene on Aggregate Formation of Branched-Tail Fluorinated Anionic Surfactants. *Colloids Surfaces A Physicochem. Eng. Asp.* **2001**, *183–185*, 749–755. https://doi.org/10.1016/S0927-7757(01)00501-5.

(24) Buck, R. C.; Murphy, P. M.; Pabon, M. Chemistry, Properties, and Uses of Commercial Fluorinated Surfactants; 2012; pp 1–24. https://doi.org/10.1007/978-3-642-21872-9_1.

(25) Sunderland, E. M.; Hu, X. C.; Dassuncao, C.; Tokranov, A. K.; Wagner, C. C.; Allen, J. G. A Review of the Pathways of Human Exposure to Poly- and Perfluoroalkyl Substances (PFASs) and Present Understanding of Health Effects. *J. Expo. Sci. Environ. Epidemiol.* **2019**, *29* (2), 131–147. https://doi.org/10.1038/s41370-018-0094-1.

(26) Gomis, M. I.; Vestergren, R.; Borg, D.; Cousins, I. T. Comparing the Toxic Potency in Vivo of Long-Chain Perfluoroalkyl Acids and Fluorinated Alternatives. *Environ. Int.* **2018**, *113*, 1–9. https://doi.org/10.1016/j.envint.2018.01.011.

(27) Conley, J. M.; Lambright, C. S.; Evans, N.; Strynar, M. J.; McCord, J.; McIntyre, B. S.; Travlos, G. S.; Cardon, M. C.; Medlock-Kakaley, E.; Hartig, P. C.; Wilson, V. S.; Gray, L. E. Adverse Maternal, Fetal, and Postnatal Effects of Hexafluoropropylene Oxide Dimer Acid (GenX) from Oral Gestational Exposure in Sprague-Dawley Rats. *Environ. Health Perspect.* **2019**, *127* (3), 37008. https://doi.org/10.1289/EHP4372.

(28) Qin, P.; Liu, R.; Pan, X.; Fang, X.; Mou, Y. Impact of Carbon Chain Length on Binding of Perfluoroalkyl Acids to Bovine Serum Albumin Determined by Spectroscopic Methods. *J. Agric. Food Chem.* **2010**, *58* (9), 5561–5567. https://doi.org/10.1021/jf100412q.

(29) Kudo, N.; Suzuki-Nakajima, E.; Mitsumoto, A.; Kawashima, Y. Responses of the Liver to Perfluorinated Fatty Acids with Different Carbon Chain Length in Male and Female Mice: In Relation to Induction of Hepatomegaly, Peroxisomal β-Oxidation and Microsomal 1-Acylglycerophosphocholine Acyltransferase. *Biol. Pharm. Bull.* **2006**, *29* (9), 1952–1957. https://doi.org/10.1248/bpb.29.1952.

(30) Takacs, M. L.; Abbott, B. D. Activation of Mouse and Human Peroxisome Proliferator–Activated Receptors (α, β/δ, γ) by Perfluorooctanoic Acid and Perfluorooctane Sulfonate. *Toxicol. Sci.* **2007**, *95* (1), 108–117. https://doi.org/10.1093/toxsci/kfl135.

(31) Ikeda, T.; Aiba, K.; Fukuda, K.; Tanaka, M. The Induction of Peroxisome Proliferation in Rat Liver by Perfluorinated Fatty Acids, Metabolically Inert Derivatives of Fatty Acids. *J. Biochem.* **1985**, *98* (2), 475–482.

(32) Pastoor, T. P.; Lee, K. P.; Perri, M. A.; Gillies, P. J. Biochemical and Morphological Studies of Ammonium Perfluorooctanoate-Induced Hepatomegaly and Peroxisome Proliferation. *Exp. Mol. Pathol.* **1987**, *47* (1), 98–109. https://doi.org/10.1016/0014-4800(87)90011-6.

(33) Abdellatif, A.; Preat, V.; Taper, H. S.; Roberfroid, M. The Modulation of Rat Liver Carcinogenesis by Perfluorooctanoic Acid, a Peroxisome Proliferator. *Toxicol. Appl. Pharmacol.* **1991**, *111* (3), 530–537. https://doi.org/10.1016/0041-008X(91)90257-F.

(34) Ren, X. M.; Qin, W. P.; Cao, L. Y.; Zhang, J.; Yang, Y.; Wan, B.; Guo, L. H. Binding Interactions of Perfluoroalkyl Substances with Thyroid Hormone Transport Proteins and

Potential Toxicological Implications. *Toxicology* **2016**, *366–367*, 32–42. https://doi.org/10.1016/j.tox.2016.08.011.

(35) Zhang, L.; Ren, X. M.; Guo, L. H. Structure-Based Investigation on the Interaction of Perfluorinated Compounds with Human Liver Fatty Acid Binding Protein. *Environ. Sci. Technol.* **2013**, *47* (19), 11293–11301. https://doi.org/10.1021/es4026722.

(36) Han, X.; Snow, T. A.; Kemper, R. A.; Jepson, G. W. Binding of Perfluorooctanoic Acid to Rat and Human Plasma Proteins. *Chem. Res. Toxicol.* **2003**, *16* (6), 775–781. https://doi.org/10.1021/tx034005w.

(37) Cheng, W.; Ng, C. A. Predicting Relative Protein Affinity of Novel Per- and Polyfluoroalkyl Substances (PFASs) by An Efficient Molecular Dynamics Approach. *Environ. Sci. Technol.* **2018**, *52* (14), 7972–7980. https://doi.org/10.1021/acs.est.8b01268.

(38) Shrestha, S.; Bloom, M. S.; Yucel, R.; Seegal, R. F.; Wu, Q.; Kannan, K.; Rej, R.; Fitzgerald, E. F. Perfluoroalkyl Substances and Thyroid Function in Older Adults. *Environ. Int.* **2015**, *75*, 206–214. https://doi.org/10.1016/j.envint.2014.11.018.

(39) Nelson, J. W.; Hatch, E. E.; Webster, T. F. Exposure to Polyfluoroalkyl Chemicals and Cholesterol, Body Weight, and Insulin Resistance in the General U.S. Population. *Environ. Health Perspect.* **2010**, *118* (2), 197–202. https://doi.org/10.1289/ehp.0901165.

(40) Granum, B.; Haug, L. S.; Namork, E.; Stølevik, S. B.; Thomsen, C.; Aaberge, I. S.; Van Loveren, H.; Løvik, M.; Nygaard, U. C. Pre-Natal Exposure to Perfluoroalkyl Substances May Be Associated with Altered Vaccine Antibody Levels and Immune-Related Health Outcomes in Early Childhood. *J. Immunotoxicol.* **2013**, *10* (4), 373–379. https://doi.org/10.3109/1547691X.2012.755580.

(41) Son, H.-Y.; Kim, S.-H.; Shin, H.-I.; Bae, H. I.; Yang, J.-H. Perfluorooctanoic Acid-Induced Hepatic Toxicity Following 21-Day Oral Exposure in Mice. *Arch. Toxicol.* **2008**, *82* (4), 239–246. https://doi.org/10.1007/s00204-007-0246-x.

(42) Ng, C. A.; Hungerbühler, K. Bioconcentration of Perfluorinated Alkyl Acids: How Important Is Specific Binding? *Environ. Sci. Technol.* **2013**, *47* (13), 7214–7223. https://doi.org/10.1021/es400981a.

(43) Rastelli, G.; Rio, D. A.; Degliesposti, G.; Sgobba, M. Fast and Accurate Predictions of Binding Free Energies Using MM-PBSA and MM-GBSA. *J. Comput. Chem.* **2009**, NA--NA. https://doi.org/10.1002/jcc.21372.

(44) de Ruiter, A.; Oostenbrink, C. Free Energy Calculations of Protein–Ligand Interactions. *Curr. Opin. Chem. Biol.* **2011**, *15* (4), 547–552. https://doi.org/10.1016/j.cbpa.2011.05.021.

(45) Kliewer, S. A.; Willson, T. M. Regulation of Xenobiotic and Bile Acid Metabolism by the Nuclear Pregnane X Receptor. *J. Lipid Res.* **2002**, *43* (3), 359–364.

(46) Ma, X.; Idle, J. R.; Gonzalez, F. J. The Pregnane X Receptor: From Bench to Bedside. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4* (7), 895–908. https://doi.org/10.1517/17425255.4.7.895.

(47) Blumberg, B.; Sabbagh, W.; Juguilon, H.; Bolado, J.; Van Meter, C. M.; Ong, E. S.; Evans, R. M. SXR, a Novel Steroid and Xenobiotic-Sensing Nuclear Receptor. *Genes Dev.* **1998**, *12* (20), 3195–3205. https://doi.org/10.1101/gad.12.20.3195.

(48) Watkins, R. E. The Human Nuclear Xenobiotic Receptor PXR: Structural Determinants of Directed Promiscuity. *Science (80-.).* **2001**, *292* (5525), 2329–2333. https://doi.org/10.1126/science.1060762.

(49) Kliewer, S. A.; Moore, J. T.; Wade, L.; Staudinger, J. L.; Watson, M. A.; Jones, S. A.; McKee, D. D.; Oliver, B. B.; Willson, T. M.; Zetterström, R. H.; Perlmann, T.; Lehmann, J. M. An Orphan Nuclear Receptor Activated by Pregnanes Defines a Novel Steroid Signaling Pathway. *Cell* **1998**, *92* (1), 73–82. https://doi.org/10.1016/S0092-8674(00)80900-9.

(50) Mani, S.; Dou, W.; Redinbo, M. R. PXR Antagonists and Implication in Drug Metabolism. *Drug Metab. Rev.* **2013**, *45* (1), 60–72. https://doi.org/10.3109/03602532.2012.746363.

(51) Navaratnarajah, P.; Steele, B. L.; Redinbo, M. R.; Thompson, N. L. Rifampicin-Independent Interactions between the Pregnane X Receptor Ligand Binding Domain and Peptide Fragments of Coactivator and Corepressor Proteins. *Biochemistry* **2012**, *51* (1), 19–31. https://doi.org/10.1021/bi2011674.

(52) Zhai, Y.; Pai, H. V.; Zhou, J.; Amico, J. A.; Vollmer, R. R.; Xie, W. Activation of Pregnane X Receptor Disrupts Glucocorticoid and Mineralocorticoid Homeostasis. *Mol. Endocrinol.* **2007**, *21* (1), 138–147. https://doi.org/10.1210/me.2006-0291.

(53) Zhou, J.; Febbraio, M.; Wada, T.; Zhai, Y.; Kuruba, R.; He, J.; Lee, J. H.; Khadem, S.; Ren, S.; Li, S.; Silverstein, R. L.; Xie, W. Hepatic Fatty Acid Transporter Cd36 Is a Common Target of LXR, PXR, and PPARγ in Promoting Steatosis. *Gastroenterology* **2008**, *134* (2), 556–567. https://doi.org/10.1053/j.gastro.2007.11.037.

(54) Gong, H.; Singh, S. V.; Singh, S. P.; Mu, Y.; Lee, J. H.; Saini, S. P. S.; Toma, D.; Ren, S.; Kagan, V. E.; Day, B. W.; Zimniak, P.; Xie, W. Orphan Nuclear Receptor Pregnane X Receptor Sensitizes Oxidative Stress Responses in Transgenic Mice and Cancerous Cells. *Mol. Endocrinol.* **2006**, *20* (2), 279–290. https://doi.org/10.1210/me.2005-0205.

(55) Lehmann, J. M.; McKee, D. D.; Watson, M. A.; Willson, T. M.; Moore, J. T.; Kliewer, S. A. The Human Orphan Nuclear Receptor PXR Is Activated by Compounds That Regulate CYP3A4 Gene Expression and Cause Drug Interactions. *J. Clin. Invest.* **1998**, *102* (5), 1016–1023. https://doi.org/10.1172/JCI3703.

(56) Zhang, Y.-M.; Wang, T.; Yang, X.-S. An in Vitro and in Silico Investigation of Human Pregnane X Receptor Agonistic Activity of Poly- and Perfluorinated Compounds Using

the Heuristic Method–Best Subset and Comparative Similarity Indices Analysis. *Chemosphere* **2020**, *240*, 124789. https://doi.org/10.1016/j.chemosphere.2019.124789.

(57) Zhang, Y. M.; Dong, X. Y.; Fan, L. J.; Zhang, Z. L.; Wang, Q.; Jiang, N.; Yang, X. S. Poly- and Perfluorinated Compounds Activate Human Pregnane X Receptor. *Toxicology* **2017**, *380*, 23–29. https://doi.org/10.1016/j.tox.2017.01.012.

(58) Vaz, R. J.; Li, Y.; Chellaraj, V.; Reiling, S.; Kuntzweiler, T.; Yang, D.; Shen, H.; Batchelor, J. D.; Zhang, Y.; Chen, X.; McLean, L. R.; Kosley Jr., R. Amelioration of PXR-Mediated CYP3A4 Induction by MGluR2 Modulators. *Bioorg. Med. Chem. Lett.* **2018**, *28*, 3194–3196. https://doi.org/10.2210/PDB6DUP/PDB.

(59) Chemical Computing Group Inc. Molecular Operating Environment (MOE). **2016**.

(60) Labute, P.; Santavy, M. SiteFinder-Locating Binding Sites in Protein Structures http://www.chempcomp.com/journal/sitefind.htm%5Cnhttps://www.chemcomp.com/journal/sitefind.htm.

(61) Xue, Y.; Chao, E.; Zuercher, W. J.; Willson, T. M.; Collins, J. L.; Redinbo, M. R. Crystal Structure of the PXR–T1317 Complex Provides a Scaffold to Examine the Potential for Receptor Antagonism. *Bioorg. Med. Chem.* **2007**, *15* (5), 2156–2166. https://doi.org/10.1016/j.bmc.2006.12.026.

(62) Chrencik, J. E.; Orans, J.; Moore, L. B.; Xue, Y.; Peng, L.; Collins, J. L.; Wisely, G. B.; Lambert, M. H.; Kliewer, S. A.; Redinbo, M. R. Structural Disorder in the Complex of Human Pregnane X Receptor and the Macrolide Antibiotic Rifampicin. *Mol. Endocrinol.* **2005**, *19* (5), 1125–1134. https://doi.org/10.1210/me.2004-0346.

(63) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2018**, *47* (D1), D1102--D1109. https://doi.org/10.1093/nar/gky1033.

(64) Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins Struct. Funct. Bioinforma.* **2009**, *75* (1), 187–205. https://doi.org/10.1002/prot.22234.

(65) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39* (6), 1397–1412. https://doi.org/10.1063/1.1734456.

(66) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C.; Brook, S.; Brook, S.; Brook, S. Comparison of Multiple AMBER Force Fields and Development of Improved Protien Backbone Parameters. *Proteins* **2006**, *65* (3), 712–725. https://doi.org/10.1002/prot.21123.Comparison.

(67) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(68) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in Docking Success Rates Due to Dataset Preparation. *J. Comput. Aided. Mol. Des.* **2012**, *26* (6), 775–786. https://doi.org/10.1007/s10822-012-9570-1.

(69) Case, D. A.; Cerutti, D. S.; T.E. Cheatham, I. I. I.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A. E.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. Amber17. **2017**, No. April. https://doi.org/10.13140/RG.2.2.36172.41606.

(70) 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. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(71) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641. https://doi.org/10.1002/jcc.10128.

(72) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(73) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(74) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8* (9), 3314–3321. https://doi.org/10.1021/ct300418h.

(75) Eken, Y.; Patel, P.; Díaz, T.; Jones, M. R.; Wilson, A. K. SAMPL6 Host–Guest Challenge: Binding Free Energies via a Multistep Approach. *J. Comput. Aided. Mol. Des.* **2018**, *32* (10), 1097–1115. https://doi.org/10.1007/s10822-018-0159-1.

(76) 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. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82. https://doi.org/10.1021/ci100275a.

(77) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Genet.* **2004**, *55* (2), 383–394. https://doi.org/10.1002/prot.20033.

(78) Roe, D. R. Introduction to hydrogen bond analysis https://amber.utah.edu/AMBER-workshop/London-2015/Hbond/ (accessed Apr 19, 2019).

(79) Durdagi, S.; Şentürk, M.; Ekinci, D.; Balaydın, H. T.; Göksu, S.; Küfrevioğlu, Ö. İ.; Innocenti, A.; Scozzafava, A.; Supuran, C. T. Kinetic and Docking Studies of Phenol-Based Inhibitors of Carbonic Anhydrase Isoforms I, II, IX and XII Evidence a New Binding Mode within the Enzyme Active Site. *Bioorg. Med. Chem.* **2011**, *19* (4), 1381–1389. https://doi.org/10.1016/j.bmc.2011.01.016.

(80) Singh, N.; Tiwari, S.; Srivastava, K. K.; Siddiqi, M. I. Identification of Novel Inhibitors of Mycobacterium Tuberculosis PknG Using Pharmacophore Based Virtual Screening, Docking, Molecular Dynamics Simulation, and Their Biological Evaluation. *J. Chem. Inf. Model.* **2015**, *55* (6), 1120–1129. https://doi.org/10.1021/acs.jcim.5b00150.

(81) Khan, K. M.; Rahim, F.; Halim, S. A.; Taha, M.; Khan, M.; Perveen, S.; Zaheer-ul-Haq; Mesaik, M. A.; Iqbal Choudhary, M. Synthesis of Novel Inhibitors of β-Glucuronidase Based on Benzothiazole Skeleton and Study of Their Binding Affinity by Molecular Docking. *Bioorg. Med. Chem.* **2011**, *19* (14), 4286–4294. https://doi.org/10.1016/j.bmc.2011.05.052.

(82) Salam, N. K.; Huang, T. H.-W.; Kota, B. P.; Kim, M. S.; Li, Y.; Hibbs, D. E. Novel PPAR-Gamma Agonists Identified from a Natural Product Library: A Virtual Screening, Induced-Fit Docking and Biological Assay Study. *Chem. Biol. Drug Des.* **2007**, *71* (1), 57–70. https://doi.org/10.1111/j.1747-0285.2007.00606.x.

(83) Jiang, X.; Dulubova, I.; Reisman, S. A.; Hotema, M.; Lee, C. Y. I.; Liu, L.; McCauley, L.; Trevino, I.; Ferguson, D. A.; Eken, Y.; Wilson, A. K.; Wigley, W. C.; Visnick, M. A Novel Series of Cysteine-Dependent, Allosteric Inverse Agonists of the Nuclear Receptor RORγt. *Bioorganic Med. Chem. Lett.* **2020**, *30* (6), 126967. https://doi.org/10.1016/j.bmcl.2020.126967.

(84) Yang, W.; Eken, Y.; Zhang, J.; Cole, L. E.; Ramadan, S.; Xu, Y.; Zhang, Z.; Liu, J.; Wilson, A. K.; Huang, X. Chemical Synthesis of Human Syndecan-4 Glycopeptide Bearing O-, N-Sulfation and Multiple Aspartic Acids for Probing Impacts of the Glycan Chain and the Core Peptide on Biological Functions. *Chem. Sci.* **2020**, *11* (25), 6393–6404. https://doi.org/10.1039/d0sc01140a.

(85) Yang, W.; Ramadan, S.; Orwenyo, J.; Kakeshpour, T.; Diaz, T.; Eken, Y.; Sanda, M.; Jackson, J. E.; Wilson, A. K.; Huang, X. Chemoenzymatic Synthesis of Glycopeptides Bearing Rare N-Glycan Sequences with or without Bisecting GlcNAc. *Chem. Sci.* **2018**, *9* (43), 8194–8206. https://doi.org/10.1039/c8sc02457j.

CHAPTER EIGHT

Binding of Per- and Polyfluoro-Alkyl Substances (PFASs) to Peroxisome Proliferator-Activated Receptor Gamma (PPARγ)

About this chapter: This chapter is reprinted from Nuno, A.; Eken, Y.; Wilson, A. K. Binding of Per- and Polyfluoro-Alkyl Substances (PFASs) to Peroxisome Proliferator-Activated Receptor Gamma (PPARγ). *ACS Omega* **2021**, *6 (23),* 15103-15114 with the permission of American Chemical Society. Both, Nuno M.S. Almeida and Yiğitcan Eken investigated interactions of half of the compounds with PPARγ included in this chapter.

8.1 Introduction

Per- and poly-fluoroalkyl substances (PFASs) are "forever chemicals", a number of which have been implicated with long lasting effects on humans, animals and the environment.[1] The first report of PFASs dates back to 1940.[2] Due to their oil and fat repellent properties along with their resilient nature, these chemicals were initially used for military purposes. Later, they were applied to industrial products, such as coating agents, oil repellents, and firefighting foam.[3–5]

Perfluorooctane sulfonate acid (PFOS) and perfluorooctanoic acid (PFOA) are the two most well-known PFASs. PFOA was initially used in commercial products to produce polytetrafluoroethylene (PTFE), for non-stick coatings.[3] Several studies in the 1990s confirmed the presence of PFOS in blood serum. Eight chemical companies agreed to stop the production of PFOA and PFOS in 2006.[6] In 2015, the production of PFOS, PFOA, perfluorosulfonic acids with six or more carbon atoms, and perfluorocarboxylic acids with eight, or more carbon atoms in the United States ended.[6,7] Despite safety concerns, which has stopped U.S. production and use, the manufacturing of these chemicals has continued in other countries.[8]

Recently, concerns have been raised about the possible levels of PFAS compounds in water sources, and, mitigation efforts are underway in many states.[9] In 2016, the EPA released a health advisory recommending that the combined concentration of PFOS and PFOA in water should be less than 70 ng/L.[10] Despite the health advisory, there are no mandatory federal standards, and

each state in the U.S. has its own regulations, or guidelines for the safety of drinking water, ranging from 11 to 1000 ng/L.[10]

Assessing the impact of PFASs on organisms at the molecular level is fundamental to understanding their possible effects and identifying routes to mitigate them. The hepatotoxicity, neurotoxicity, reproductive toxicity, immunotoxicity, thyroid disruption, and cardiovascular toxicity of PFOS has been discussed by Zeng *et. al.*[11] For a number of affected proteins linked to such toxicological impacts, there is crystal structure data available, facilitating molecular level studies. In addition, recent *in-vivo* and *in-vitro* studies have been conducted to study the interactions between human and animal proteins with PFASs (see, e.g., Ref [12–26]).

In recent studies, PFOS was implicated in renal fibrosis.[27,28] The mechanism by which PFOS can cause renal injury, involves the deacetylation and inactivation of PPARg, playing a very important role in cell signaling processes. Liu *et. al.* studied the associations of different PFASs and serum biochemical markers for uremic patients under hemodialysis.[29] They found that the effects of PFOS and PFOA on the kidneys are long-lasting, and provided an explanation for the long half-life that PFASs have in humans.

PPARg functions as a regulator for fatty acid storage and glucose metabolism by binding to DNA and acting as a transcription factor. The homodimerization of PPARγ and its biological relevance have been discussed in the literature.[30–35] Fulton *et. al.* provides direct evidence that PPARγ homodimerizes by using yeast two-hybrid experiments, where the physical interaction between the two PPARγ monomers, and formation of homodimers, has been shown by reporter activation.[30] Todorov *et. al.* studied nuclear receptor proteins from CaLu-6 cells probed with ^{33}P-labeled human renin Pal3 sequence using electrophoretic mobility-shift assay.[31] The addition of anti-PPARγ antibody in these assays resulted in retardation of two separate protein complex

bands. In other words, the anti-PPARγ antibody bound and slowed down two different PPARγ containing protein complexes present in the cells. Since RXRα is the standard interaction partner for PPARγ, Todorov *et. al.* suggested that these two bands might correspond to PPARγ/RXRα heterodimer and PPARγ/PPARγ homodimer.[31] Estany *et. al.* found two inverted half site DNA motifs which may allow two PPARγ proteins to bind to each half site as a homodimer.[32] Okuno *et. al.* utilized gel shift analysis showing that PPARγ might bind to the Pal3 DNA motif as a homodimer, in comparison to the DR1 motif, which is a commonly known PPARγ/RXR heterodimer binding site.[33] Many PPARγ crystal structures including the one reported by Nolte *et. al.* and the one studied here (PDB ID:3ADV) by Waku *et. al.* shows that PPARγ has a homodimer interface and can form a homodimer complex similar to other nuclear receptors (i.e. estrogen receptor-α and RXR-α).[34,35] Due, to the possible biological relevance of the PPARγ homodimer, the homodimer was considered in this study.

The activation of PPARg causes insulin sensitization and regulates glucose metabolism, and, the intake of any kinds of sugar is a fundamental process for the body to regulate. Chou *et. al.* investigated how L-carnitine plays an essential role in attenuating the effects of PFOS in the kidneys via PPARg and Sirt1 mechanisms.[27] Additionally, L-carnitine can be synthetized on a cellular level by methionine and lysine, and in prior studies, it is shown to diminish the effects of gentamicin-induced apoptosis in PPARa.[27,28]

To better understand PFAS structure/protein activity relationships, computational studies are important, although they are scarce. One of the first such studies was performed by Salvalaglio *et. al.*[36] They examined the binding energies and binding sites in human serum albumin, describing how PFOS and PFOA bind to this protein. The authors utilized molecular dynamics simulations along with molecular mechanics generalized Born solvation area (MM-GBSA)

calculations to predict free binding energies[36], and describe guidelines for PFASs with lower bio accumulative potential. Other studies have utilized computation to investigate the interaction of different PFASs with human or animal proteins and analyze possible binding sites and poses.[37–40]

Takacs *et. al.* investigated the interaction between PPARg and PFOS and PFOA.[12] They observed that there was no PPARg activity alteration in both mice and humans in the presence of these PFASs. Zhang *et. al.* determined half maximum inhibition concentrations (IC_{50}) for twelve PFASs with PPARg, providing docking and activity studies, and concluded that hydrogen bonding of the ligands to Tyr 473, and interactions with His 323 and His 449 were deemed essential for PPARg activation. Additionally, the authors identified key residues and important hydrogen bond pairs on PPARg for the ligand binding pocket (LBP) using molecular docking.[17] For PPARg, different studies identify His 323, His 499 and Tyr 473 as key for PPARg's activity, along with the size and length of the carbon chain (see example references [41] and [42]). In terms of structural properties, the importance of helixes AF-2, 3, 7 and 10 has been documented prior for PPARg. The position of PFASs within the ligand binding pocket and AF-2 helix, along with key residue interactions are of paramount importance for PPARg's activity. [17,43]

Activity and docking studies were also performed on PPARb/d using a range of PFASs by Li *et. al.*[44] The authors found that the binding geometries of selected PFASs were similar to those of fatty acids, fitting in the ligand binding pocket of PPARβ/δ. Furthermore, Li *et. al.* found that both isoforms of PPAR are activated by PFASs, and that the transcriptional activity was associated with the carbon length.[44] Recently, Behr *et. al.* probed the activation of nuclear receptors with PFAS.[18] Although PPARα could activate several PFASs, PPARγ was shown to only be activated by perfluoro-2-methyl-3-oxahexanoic acid (PMOH) and 3H-perfluoro-3-[(3-methoxypropoxy) propanoic acid (PMPP). In comparison with *in vitro* experimental results by

Zhang *et. al*., *Behr et. al.* reported much different PPARγ activity. These inconsistencies were attributed to the selected PPARγ constructs and different cell lines used in the experiment. [17,18] Due to the conflicting conclusions from the prior studies, a better understanding of how PPARg interacts with different residues at a molecular level is needed.

In this study, different binding pockets are investigated, as well as the interactions between PPARg and 27 widely used PFASs. Herein, in addition to the orthosteric binding pocket present in the PPARg ligand binding domain (LBD), a new binding site present in the PPARg homodimer is identified: dimer pocket and studied as a potential bio accumulative target. The dimer pocket is situated between the two PPARg LBD monomers, and computational predictions showed binding to a variety of PFASs.

The PFASs investigated here represent a variety of carbon chain lengths and functional groups (amines, carboxylic groups, alcohols, and sulfonic groups) to provide insight about how structural modifications affect the binding of PFAS species to the receptor. A number of "short chain" PFAS alternatives are considered including 2,3,3,3-tetrafluoro-2-heptafluoropropoxy propanoic acid (GenX), 4,8-dioxa-3H-perfluorononanoic acid (ADONA), 6:2 fluorotelomer carboxylic acid (6:2 FTCA), and 6:2 fluorotelomer alcohol (6:2 FTOH). "Short chain" alternatives to PFOS and PFOA are perfluoroalkyl carboxylic acids (PFCAs) with six or less fluorinated carbons and perfluorosulfonic acids (PFSAs) with five or less fluorinated carbons. "Short chain" PFASs are generally thought to be less harmful; however, their effects on the human body and environment are less understood.[45–47] The influence of basic and acidic residues upon the interactions has been investigated, as has the impact of L-carnitine and its interaction with different binding pockets.

8.2 Computational Methods

8.2.1 Site Analysis and Molecular Docking

The PPARγ dimer structure was taken from the RSCB Protein Data Bank (PDB ID: 3ADV[35]), and was protonated using the Protonate 3D[48] program from the Molecular Operating Environment's (MOE).[49] 3ADV structure is a PPARγ homodimer, which has seen less attention in the literature and allowed us to identify a new binding site for PFASs (dimer pocket). Additionally, 3ADV has a fatty acid metabolite, which has an amphiphilic nature similar to PFASs and also has good X-ray resolution (2.27 Å), which allows for detecting positions of the side chain atoms confidently.[35] The protonated PPARγ dimer was scanned for potential binding pockets using MOE's "site finder program". The site finder program detects alpha shapes on the protein surface and evaluates them according to their propensity of ligand binding (PLB) score.[50]

The initial structures of the PFASs and L-carnitine were obtained from PubChem.[51] The chemical formulas and acronyms for the PFASs can be found on Table 8.1 and the chemical structures of the compounds are included in Table 8.2. The protonation states of the PFASs and L-carnitine under physiological conditions (pH 7, 300K and 1 atm) were determined using the Protonate3D module and the structures were minimized in MOE with the AMBER10: Extended Hückel Theory (EHT) force field, which uses Amber ff10 for macromolecules and Extended Hückel Theory for the ligands.[52–54] PFASs' and L-carnitine binding modes to the dimer pocket and LBP were determined by docking to the binding sites using MOE.[49] During the generation of L-carnitine binding poses to the LBP, hydrogen bond to the Tyr 473 was implemented as a query for a pharmacophore approach, which is associated with PPARγ activity.

The London ΔG scoring function was used to evaluate 100 initial ligand placements.[55] Then, these initial 100 placements were further refined to ten poses via the Generalized-Born Volume

Integral/Weighted Surface area scoring function (GBVI/WSA) ΔG with induced fit protein settings. The structurally distinct refined poses with the highest (GBVI/WSA) ΔG scores were selected for further studies.

8.2.2 Simulation Protocol

The selected complex structures were minimized using molecular mechanics (MM) with the AMBER10:EHT forcefield in MOE.[52–54] The topologies and the parameters for the minimized structures were created using the Leap module of Amber Tools[56] by using General Amber Force Field (GAFF), AMBER ff14sb force fields.[57] The AM1-BCC charge scheme[58] was used to calculate partial charges of the ligand atoms, and these partial charges were fit to GAFF by using the Antechamber[56] suite to generate ligand parameters. The protein-ligand complex structures were placed in a 14 Å cube beyond the solute box, neutralized and ionized with 100mM NaCl ions using parameters from Joung and Cheatham in order to replicate a biological ionic environment.[59]

In the minimization protocol, a series of harmonic potentials (500.0, 200.0, 20.0, 10.0, 5.0, 0.0 kcal mol^{-1}) were used, which restrain the protein structure, and allow water molecules, ions and the ligand to relax. Then, the systems were heated from 100 K to 300 K in 30 picosecond MD simulations. After heating, 30 ns, MD simulations were performed to ensure the convergence of the system at 300 K and 1 atm pressure (see example RMSD plots Figures 8.16-8.19). During all simulations, the pressure and temperature were controlled by isotropic position scaling and Langevin dynamics, respectively. Furthermore, the SHAKE algorithm[60] was used to constrain hydrogen bonds which allowed the use of a 2-femtosecond time step. Non-bonded interactions were truncated to 10 Å, while the particle-mesh Ewald (PME) method was used to

efficiently approximate long-range electrostatic interactions. The minimization protocol and MD simulations were performed with Amber.[56]

8.2.3 Binding Energy Calculations

The binding free energies of the ligand-protein complexes were calculated using both Molecular Mechanics Poisson–Boltzmann Surface Area (MM-PBSA) and Molecular Mechanics General Born Surface Area (MM-GBSA) with a modified General Born solvation model[61] implemented in the Amber PBSA-solver.[62] The default internal and external dielectric constants were used (1.0 and 80.0, respectively). The solvent accessible surface area (SASA) was determined with the default Linear Combinations of Pairwise Overlaps (LCPO) method using modified Bondi atomic radii. Due to the high computational cost of the methodology, initial 500 frames of the simulation were used for the MM-GBSA and MM-PBSA calculations. As shown in Figures 8.16-8.19, the overall protein RMSD has reached stability by this point, so longer simulations are not necessary. A prior study has demonstrated, that choice of different/longer time frames will have little impact on the binding energy predictions.[63] The solute entropies were not considered, because the primary focus of this effort was on the relative binding energies of the ligands on PPARγ. The binding contributions of the residues were calculated by per-residue decomposition[56] and the energy contribution for each acidic and basic residues were averaged from all of the poses tested. The residue decomposition was performed using CPPTRAJ from Amber was used and the full length of the simulation was considered.[56,64] This step is important to understand specific interactions, selectivity and recognition in PPARγ.

8.2.4 Hydrogen Bond Analysis

Hydrogen bond lifetime analyses were performed via CPPTRAJ for every ligand tested.[64] The ligand-PPARγ complex with the strongest MM-PBSA relative binding energy was selected for analysis.

8.3 Results and Discussion

8.3.1 Binding pockets on PPARγ

The two potential binding sites with the highest PLB scores, referred to here as the dimer pocket and the Ligand Binding Pocket (LBP), were investigated and are shown in Figure 8.1. The dimer pocket, not previously studied, has the highest PLB score in comparison to other pockets. It is located between the two PPARγ dimer structures and is ~1900 $Å^3$ in size. This is in contrast to the LBP, which is ~ 1300 $Å^3$ in size. The LBP is known to bind to a variety of ligands (i.e. medium chain fatty acids, thiazolidinediones, phenyl acetic acids and phenyl propanoic acids).[65–67] In this study, both the dimer pocket and the LBP were considered as potential binding sites for the PFASs (Table 8.1) and L-carnitine.

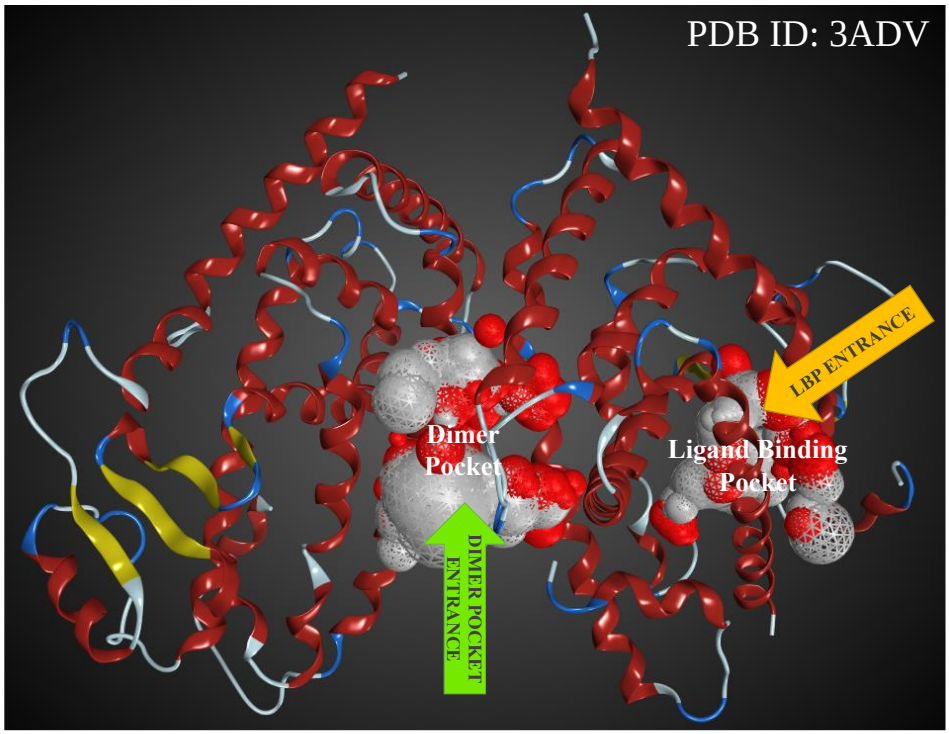


Figure 8. 1 Binding pockets detected on the PPARγ dimer structure (PDB ID: 3ADV) using MOE's Site Finder. Two potential binding sites are identified and their entrances are shown. The surface and area of the binding sites are depicted. The red spheres indicate a hydrophilic, while silver depicts hydrophobic surfaces.

8.3.2 Binding Poses of PFASs

To determine how PFASs orient within the potential binding sites, molecular docking was used. The ligand binding to PPARγ is a complex process. The PPARγ receptor contains flexible binding cavities and can host a variety of structurally distinct ligands.[68] Due to the complexity of binding, induced-fit docking is used during the pose generation. Induced-fit docking accounts for the movements in the protein structure upon ligand binding and multiple binding possess generated during this step are further evaluated through MD and binding free energy calculations. The binding poses with highest affinity are evaluated through the residue decomposition schemes and hydrogen bond analysis. The highest affinity binding poses of the

ligands into the LBP and the dimer pocket are shown in Figures 8.2 and 8.9, respectively. PFASs which have more than six, and less than 14 per-fluorinated carbon orient their functional groups towards Tyr 473, His 449 and His 323, which have previously been proposed as important residues for PPARγ activity. [17]

8.3.3 Binding Free Energy Calculations (MM-GBSA/MM-PBSA) and Correlation Plots

The binding modes of PFASs and L-carnitine to the LBP and dimer pocket were studied using MM-GBSA and MM-PBSA and the resulting binding energies are depicted in Figures 8.3 and 8.10, respectively. The binding energies were determined by averaging the results for different PPARγ binding poses for each compound. In comparing the experimental IC_{50} values by Zhang *et. al.* (see, Ref [17]) to our predicted PFASs to LBP binding energies, better correlation was obtained using MM-PBSA rather than MM-GBSA.

The binding energy values correlate directly with the carbon chain length; however, the effects of the carbon chain length differ for the dimer pocket and the LBP. On average, the binding energies for the dimer pocket were lower than for the LBP. Et-PFOSA-AcOH and Me-PFOSA-AcOH showed high affinity towards the dimer pocket. Their chain lengths in addition to their sulfonic and carboxylic functional groups enabled very strong interactions (~25 kcal mol^{-1}). L-Carnitine also showed strong binding to the dimer pocket and strong residue interactions (see Section 8.3.4).

The PFASs showed stronger binding to the LBP than to the dimer pocket while L-carnitine showed similar binding to both pockets according to MM-PBSA. This indicates that PFASs are prone to bind more strongly to the LBP, although the dimer pocket can still have a role on the accumulation of PFASs. Ligand binding to LBP is important for the activity of PPARγ (see, e.g., Ref. [17]). In order to assess how the calculated binding energies for LBP correlate to the PPARγ

activity, IC_{50} values of PFDA, PFNA, PFHxS, PFOA, PFOS, PFHxDA, PFOcDa, PFTeDA, and PFDoA determined by Zhang *et. al.* are used for comparison, as shown in Figure 8.4. The binding energies of PFOcDA and PFHxS were calculated only for the LBP to compare with respective experimental IC_{50} values by Zhang *et. al.* [17] The predicted binding energies of L-carnitine show that it can compete to replace PFASs from both binding sites.

On average, the affinity of PFASs to LBP increased with the size of the carbon chain length. There is a rise in binding energy from PFBA to PFOcDA, which is consistent with the increasing size of the carbon chain length. The LBP is approximately three times larger than other nuclear receptors' ligand pockets, which allows for compounds as large as PFOcDA to bind strongly.[65] PFASs with sulfonic acid groups (PFSAs) showed higher affinity to the LBP in comparison to the carboxylic acids, fluoro telomer alcohols (FTOHs), and fluoro telomer carboxylic acids (FTCAs), with the same number of per-fluorinated carbons. The PFASs that have a 6-8 per-fluorinated carbons along with both sulfonic acid and carboxylic acid groups (Et-PFOSA-AcOH and Me-PFOSA-AcOH) showed strong binding to LBP and to the dimer pocket.

In recent work, MM-GBSA and MM-PBSA binding energy predictions were evaluated for PFASs and the hPXR protein.[69] In this prior study, both MM-PBSA and MM-GBSA correlate well with the experimental EC_{50}, though the MM-GBSA correlation was slightly better.[69] However large PFAS molecules such as PFTeDA, PFHxDA and PFOcDA were not studied for the hPXR receptor and for these larger molecules, MM-GBSA and MM-PBSA differ. As shown previously, the utility of MM-GBSA and MM-PBSA can vary with respect to the studied system.[70] Factors such as hydrophobicity, lipophilicity, and electrostatics of the ligand and choice of binding site, all play an important role on the performance of the theoretical methods, directly influencing computed predictions. For the large PFASs (PFTeDA, PFHxDA and

PFOcDA), the tail portion of the compound is more solvent exposed and MM-PBSA provides a more rigorous treatment of these solvent effects, thus, MM-PBSA results in better correlation with experimental IC_{50} values. For this reason, only the MM-PBSA correlation plot (Figure 8.4) has been included. MM-GBSA correlation is shown in Figure 8.11. The r^2 between calculated binding energies and experimental IC_{50} values is 0.6, which indicates that the calculated binding energies for LBP correlate with the activity data, although some variance is observed. This variance is associated with both experimental and calculated standard deviations. Another element that contributes to lower correlation is the fact that experimental IC_{50} values relate to the structure activity data, which is not the case for MM-GBSA or MM-PBSA. For example, for 6:2 FTOH, or 8:2 FTOH, Zhang *et. al.* does not detect any activity experimentally, however, in the current study, these species do bind, though they do not contribute to the receptor's activity. PFHpA is an outlier and has not been included in Figure 8.4, due to its large IC_{50} value and large experimental uncertainty for PPARγ activation (192.4 ± 17.2).

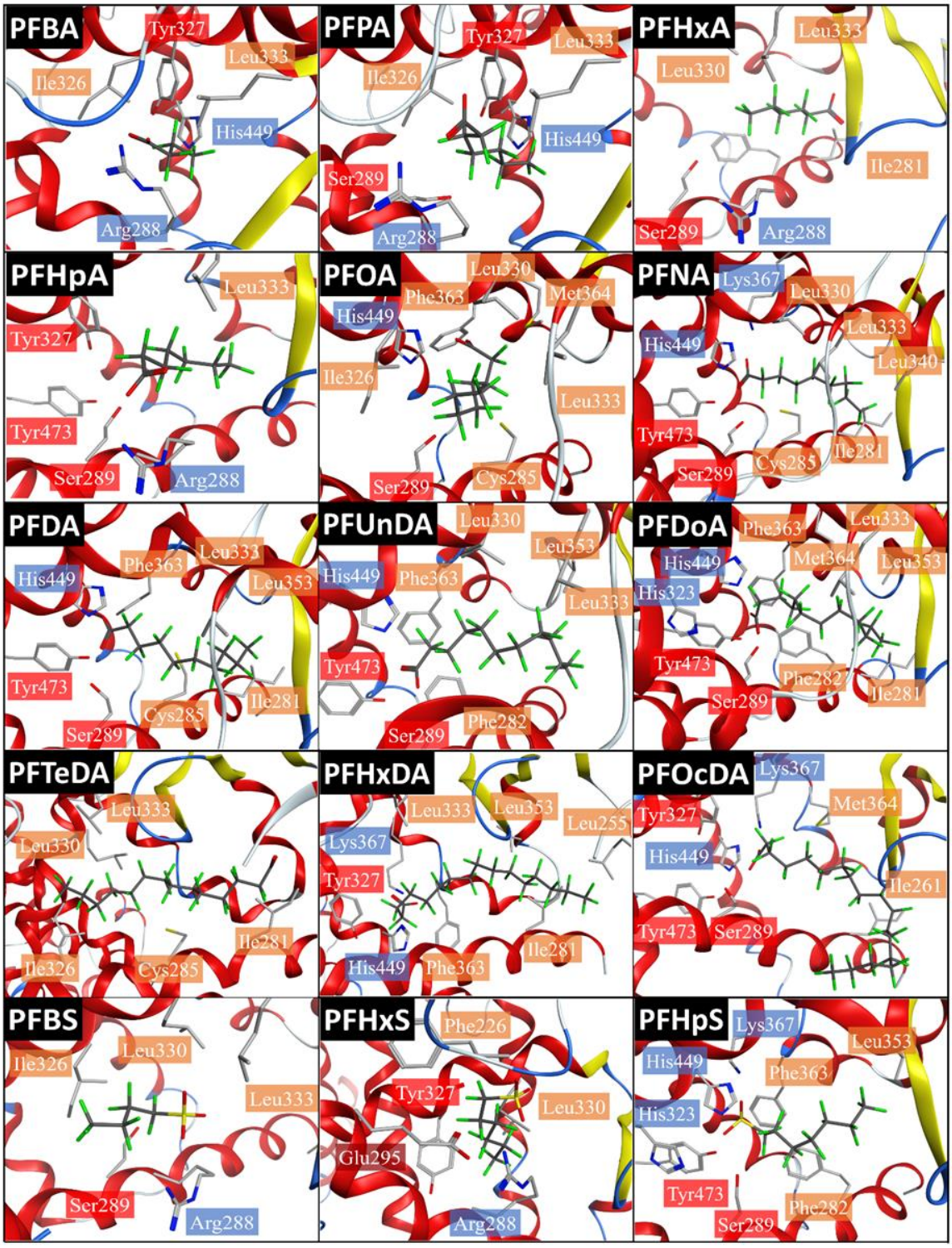
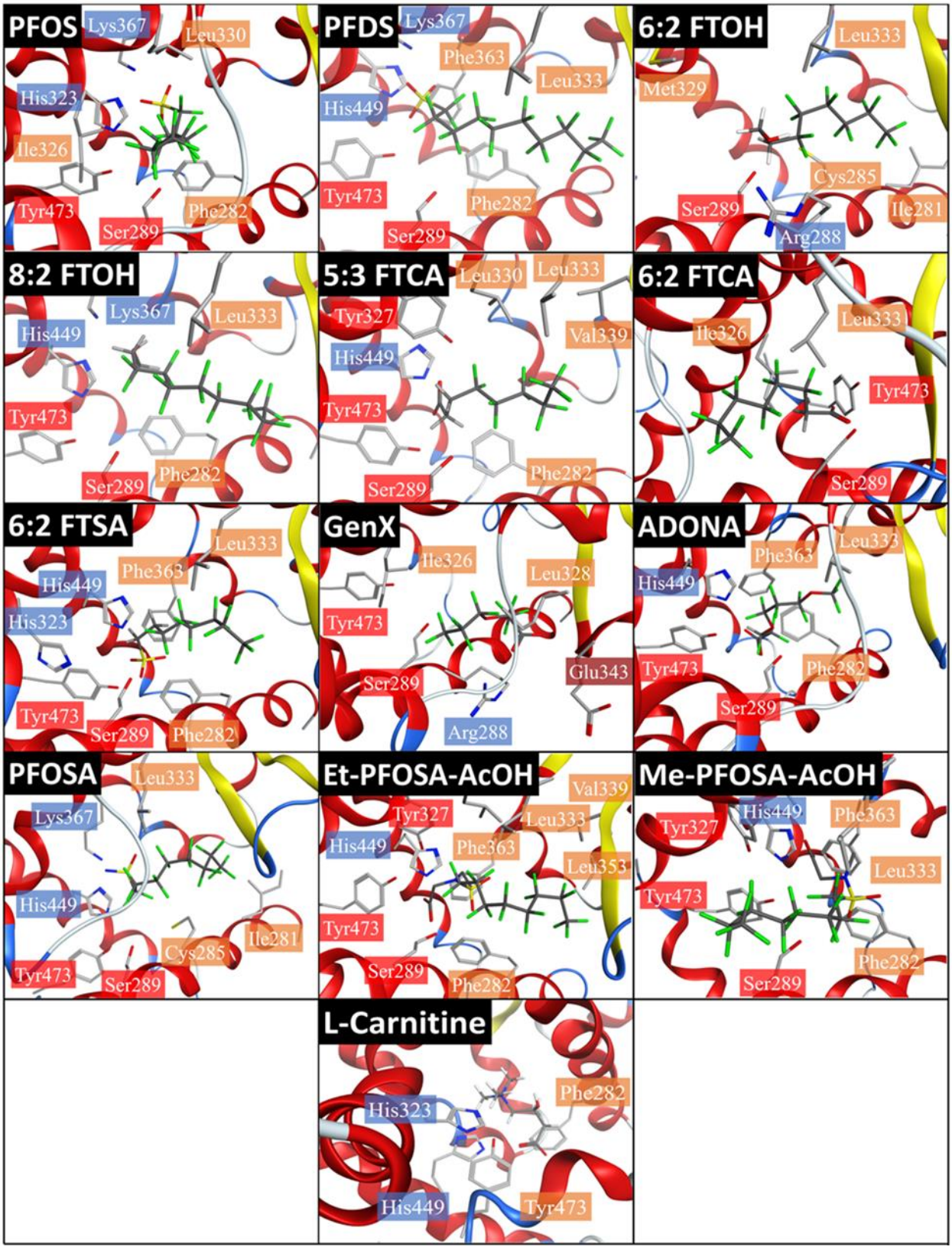


Figure 8. 2 Binding poses of PFASs and L-carnitine on PPARγ. The binding modes that have the highest binding affinity determined from MM-PBSA are shown. Residues depicted belong to Chain A.

Figure 8. 2 (cont'd)



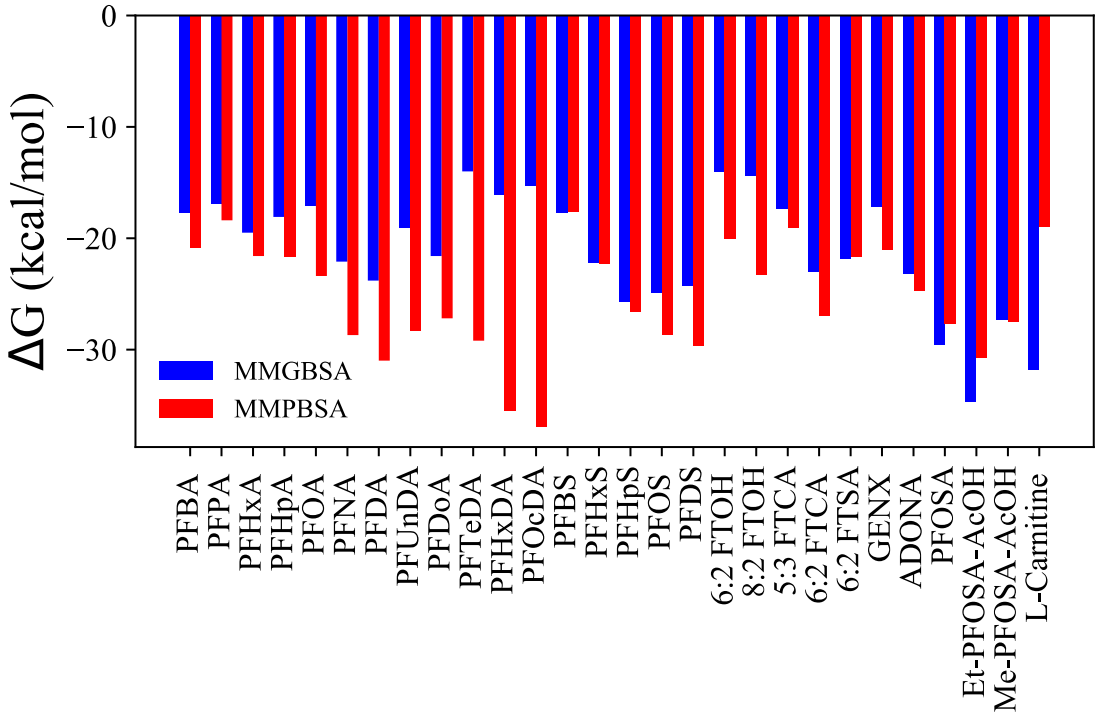


Figure 8. 3 Average binding energies of PFASs and L-carnitine calculated with MM-GBSA and MM-PBSA for the LBP. PFASs are divided into subgroups: perfluoroalkyl carboxylic acids (PFCAs), followed by perfluoroalkyl sulfonic acids (PFSAs), fluorotelomer alcohols (FTOHs), fluorotelomer carboxylic acids (FTCAs), fluorotelomer sulfonic acids (FTSAs) and then alternatives. Each subgroup was listed from shortest chain length to longest (Tables 8.1 and 8.2 for acronyms and structures).

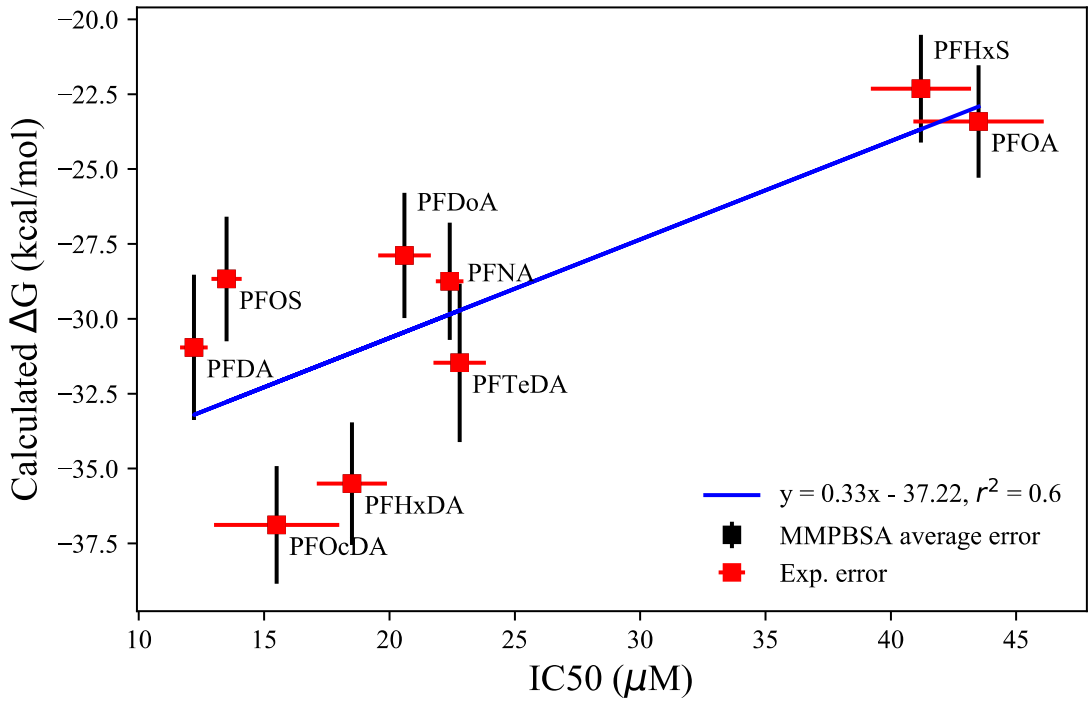


Figure 8. 4 Average calculated binding energies of PFASs with MM-PBSA in comparison with IC_{50} values determined experimentally by Zhang *et. al*. On the y-axis, the average calculated binding energies are plotted, and along the x-axis, the experimental IC_{50} values are provided. Error bars are depicted in black (MM-PBSA) and red (experimental).

8.3.4 Residue decomposition analysis

8.3.4.1 Binding contribution from nearby residues to PFASs and L-carnitine

To evaluate the contribution of nearby residues to the Gibbs free energy of binding, a space of 5-6 Å around PFASs and L-carnitine was selected. The binding energy contribution within this space was determined via a per-residue decomposition, which accounts for electrostatic and van der Waals contributions to the binding. The average residue contributions for PFASs (red) and L-carnitine (green) were determined from the highest affinity poses for the LBP and dimer pocket, and are compared in Figures 8.5 and 8.12, respectively. At pH 7, L-carnitine is neutral,

but it has two charged groups. One side of the molecule is positively charged ($N^+C_3H_9$) and the other side has a deprotonated carboxylic group (COO^-). It also has an OH group which can serve as a hydrogen donor (Section 8.3.5). As discussed in Section 8.3.3, L-carnitine shows similar binding energies to the dimer pocket and LBP, with average binding energies of -19.0 kcal mol^{-1} from MMPBSA (Tables 8.3 and 8.4).

For the dimer pocket, the acidic residues such as Glu 324, Asp 396, Glu 407 and Asp 441 repel PFASs derivatives very strongly, as demonstrated by the average binding contributions of ~ 30 kcal mol^{-1} (Figure 8.12). For L-carnitine, the acidic residues contribute positively, or negatively to the overall energy depending on their orientation towards the NH_3^+ and COO^- groups in the molecule. For example, in the dimer pocket, L-carnitine is repelled by Glu 324 (15 kcal mol^{-1}), whereas Glu 407 has a negative contribution to the binding energy (-15 kcal mol^{-1}). The interaction energy of L-carnitine with basic residues, especially arginines and lysines is significant, but not as strong as for PFASs.

Figure 8.5 shows the interaction energy of PFASs' with close residues within the LBP. As shown, Arg 288 and Lys 367 have the strongest contributions to the binding, whereas Glu 295 and Glu 343 repel PFASs from binding to the LBP. In contrast, L-carnitine is not repelled by Glu 295 and Glu 343, and additionally showed strong interaction energy with Lys 367. Tyr 473 contributes slightly to the binding of PFASs and L-carnitine to the LBP, due to the hydrogen bonding observed with the long carbon chain molecules. (Zhang *et. al.* proposed hydrogen bonding to Tyr 473 as key to the PPARγ activity.[17] The hydrogen bonding interaction is discussed in Section 8.3.5). L-Carnitine has a -6.6 kcal mol^{-1} interaction energy to Tyr 473, compared to a slightly lower value of average PFASs. PFASs that are shorter in length such as PFBA and PFPA did not form a hydrogen bond with Tyr 473 (Figure 8.8).

As the importance of His 449 and His 323 PPARγ activity has been reported[41,65], the role of these residues is examined. His 449 has an interaction energy of ~ -5 kcal mol^{-1}, with the PFASs and L-carnitine. For His 323 the calculated interaction energy was -5.3 kcal mol^{-1} for L-carnitine, but positive for PFASs.

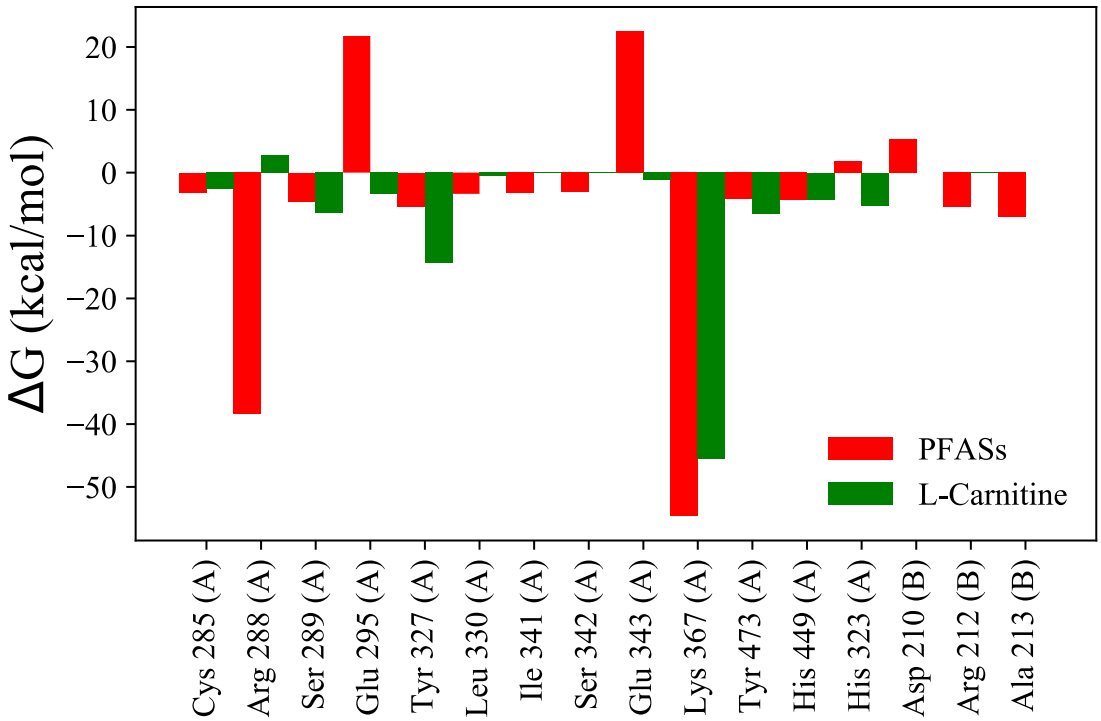


Figure 8. 5 Binding contribution of each nearby residue for PFASs and L-carnitine (LBP). For PFASs, highest affinity poses are averaged and for L-carnitine the highest affinity pose is used.

8.3.4.2 Binding energy contribution from acidic and basic residues to PFASs and L-carnitine

A residue decomposition of PPARg in terms of long-range electrostatic interaction was done. To date, there is no such study done for PPAR receptors. Here, we consider two questions: How

are ligands affected by long range interactions? How is the LBP affected by residues on the other side of the protein?

To investigate these questions, basic residues (arginines, lysines, histidines) and acidic (glutamate, aspartate) residues within the PPARγ dimer were studied from the A and B chains. All ligand poses were considered for the dimer pocket and LBP. Average interaction energies for all of the PFASs investigated were compared with the L-carnitine interaction energy. In Figures 8.6 and 8.7, the average interaction energies for LBP are shown for PFASs and L-carnitine, respectively.

The average interaction energies for the dimer pocket can be found on Figures 8.13 and 8.14. As the dimer pocket is situated between the two monomers (Figure 8.1), it is able to interact with both chains of the protein (almost symmetrically, when comparing the energies of Chain A and Chain B). For basic residues, the strongest interactions are observed with Arg 397, Arg 443, Lys 373, Lys 434 and Lys 438, and for acidic residues the strongest repulsion is observed with Asp 396, Glu 324, Glu 407 and Asp 441 ($> \pm 25$ kcal mol^{-1}). The short-range electrostatic interactions within the chains of the protein, can stabilize the ligand, or repel it. When comparing PFASs with L-carnitine, the average interaction energies for the PFASs with Asp 396, Glu 324, Glu 407 and Asp 441 reveal a different trend than for L-carnitine. PFASs are strongly repelled by these residues, while L-Carnitine is only slightly repelled (~ 5 kcal mol^{-1}) by Glu 324 but attracted by the other ones.

Considering the LBP, the strongest interactions correspond to residues in Chain A (Arg 288, Lys 367, Glu 291, Glu 295 and Glu 343), which are situated mainly in the LBP (Figures 8.6 and 8.7). There are large contributions from the residues on the other chain, that range from -5 to -15 kcal mol^{-1} for the basic residues and 5 to 15 kcal mol^{-1} for the acidic residues.

For L-carnitine, considering the acidic residues' interaction energy, there is a different trend compared to PFASs (Figure 8.7). The acidic residue energies vary from positive to negative, which shows that not all are repulsive towards L-carnitine. Regarding basic residues, Lys 367 is the major contributor towards its affinity in the pocket and contributes strongly to the LBP binding.

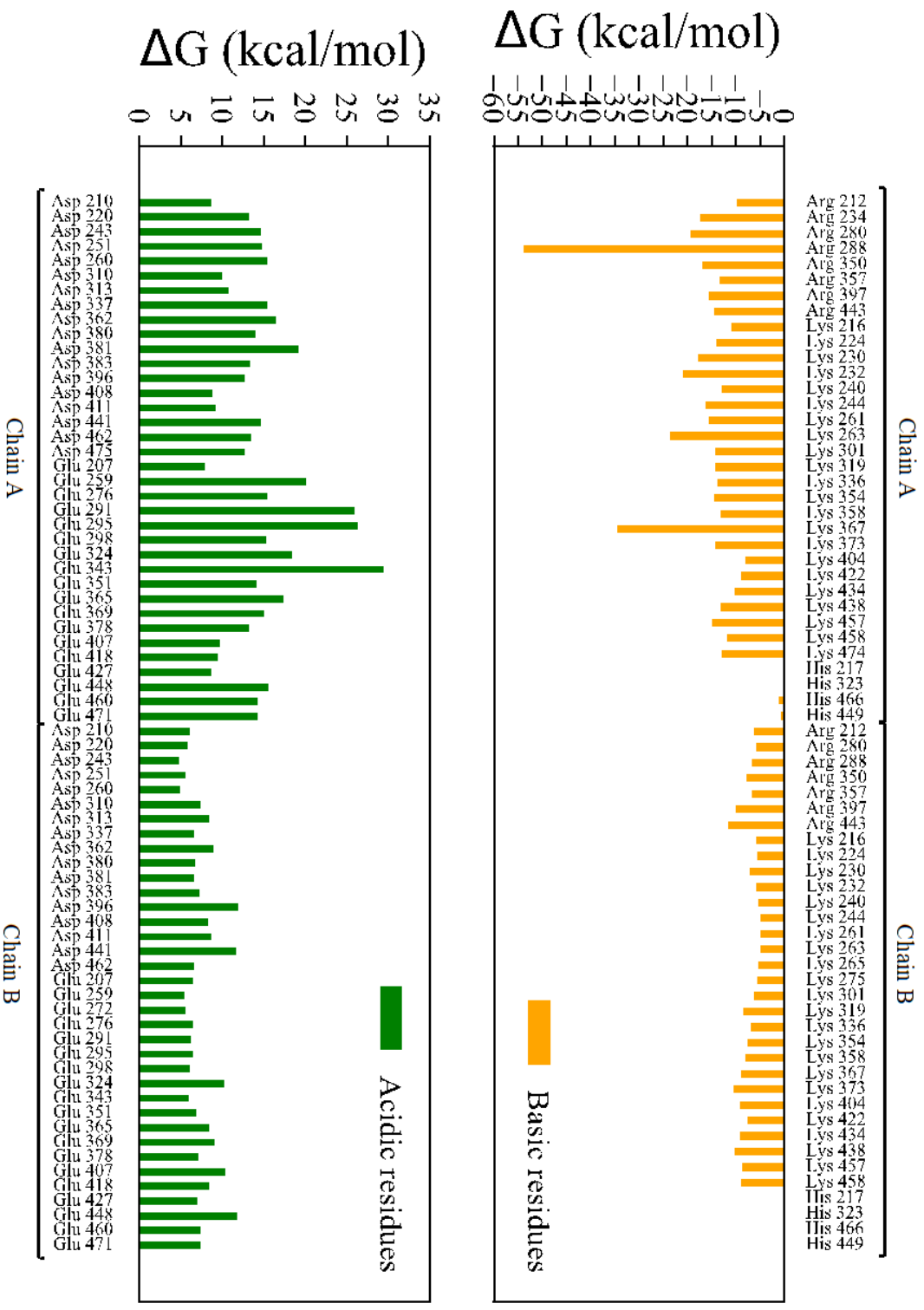


Figure 8. 6 Binding contributions of the acidic and basic residues for PFASs (LBP) in Chain A and Chain B.

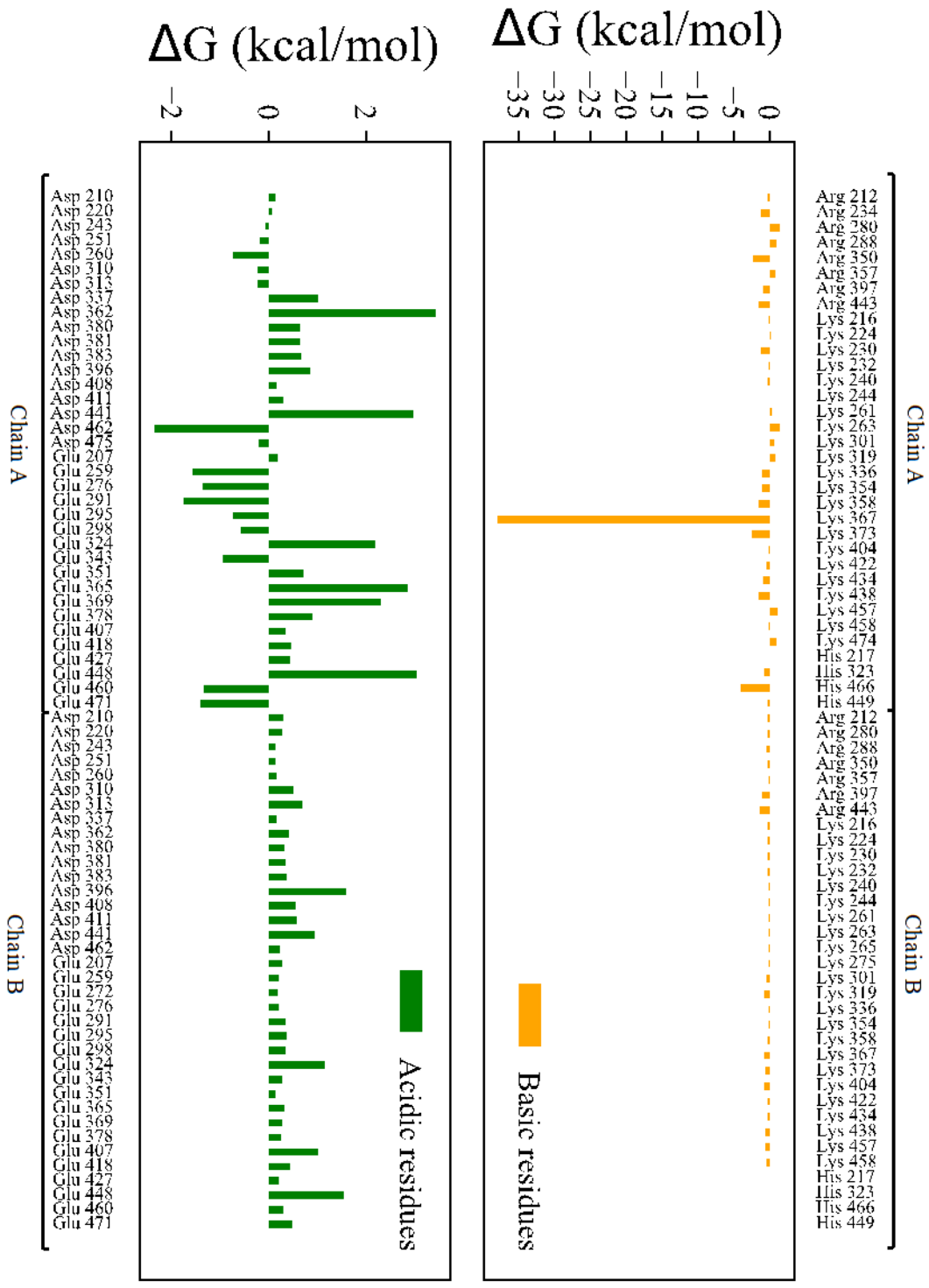


Figure 8. 7 Binding contributions of the acidic and basic residues for L-carnitine (LBP) in Chain A and Chain B.

8.3.5 Hydrogen bonding

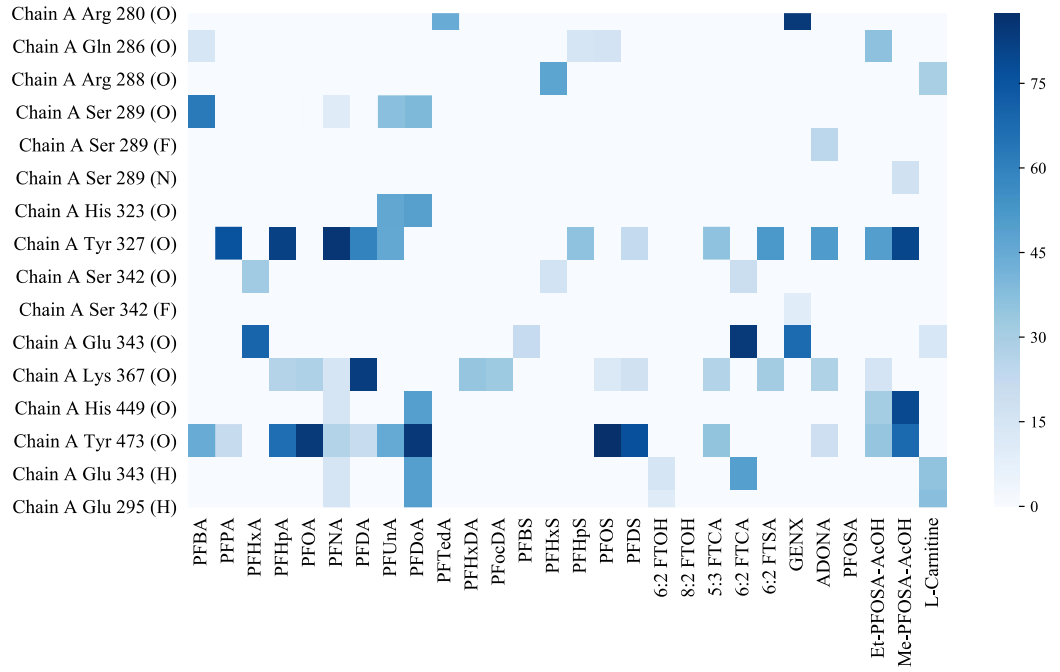


Figure 8. 8 Hydrogen bond lifetimes for the LBP. The y-axis depicts the chain and residue number from the receptor, and in brackets, the atom from the ligand performing the hydrogen bonding is shown. Acceptors are portrayed by "(O), (F), (N)", and donors by "(H)". In the x-axis the different PFASs and L-carnitine are shown.

A detailed analysis of the propensity of the dimer pocket and LBP to hydrogen bond is fundamental for understanding the intermolecular interactions between ligands and residues. By using MD trajectories, it is possible to understand fundamental binding properties, and the activity of the receptor/protein. Herein, some of the ligands; 6:2 FTOH, 8:2 FTOH, L-carnitine, Et-PFOSA and Met-PFOSA can be hydrogen donors or acceptors (Figures 8.8 and 8.15).

In Figure 8.15, the hydrogen bonding percentage is shown for the dimer pocket. Lys 438, Arg 443 and Arg 397 have the highest percentage of hydrogen bonding. These residues were noted earlier (Section 3.4.1) as being in close proximity to the ligands in the binding cavity. L-

Carnitine is stabilized in this pocket by three hydrogen bonds with Gln 437, Arg 443 and Ser 394. L-Carnitine's positive and negative charged groups allow for different bonding with residues in the dimer pocket. Et-PFOSA-AcOH and Met-PFOSA-AcOH have very strong affinity to the dimer pocket and form strong hydrogen bonding with Arg 443. The sulfonic and carboxylic functional groups interact strongly with nearby residues. In addition, Et-PFOSA-AcOH and Met-PFOSA-AcOH are also stabilized by the interaction with Asp 396 and Gln 444. In the dimer pocket, hydrogen bonding from fluorines can occur, though it is minimal.

In Figure 8.8, the LBP hydrogen bonding is described for PFASs and L-carnitine. As mentioned earlier, hydrogen bonding to Tyr 473 is directly associated to the activity of the receptor. PFASs with 7-12 perfluorinated carbons such as PFHpA, PFOA, PFNA, PFDA, PFDoA, PFOS, Et-PFOSA-AcOH, Met-PFOSA-AcOH show high affinity to this residue. PFOS, Et-PFOSA-AcOH, Met-PFOSA-AcOH and PFDS have a sulfonic group, which enables them to undergo strong hydrogen bonding, occurring for nearly the entire simulation. From the literature, 6:2 FTOH, 8:2 FTOH, 6:2 FTCA, PFBS and PFBA show no activity against PPARg, which is corroborated in Figure 8.8, there is no hydrogen bonding to Tyr 473.[17] Even though PFTeDA, PFHxDA, and PFOcDA, show activity experimentally, the MD simulations do not show hydrogen bond formation with Tyr 473. There are examples of PPARg agonists that do not form H-bonds with Tyr 473 but are still able to activate a receptor through immobilization of the H12 helix.[17,43] Due to the size of these larger PFASs, the binding poses obtained for them were more distant from Tyr 473 and more solvent exposed and thus the hydrogen bonding with Tyr 473 is not demonstrated. Also, the scope of this study was to compare relative binding energies of various PFASs and understand the molecular interactions behind the PPARg recognition. For this purpose, 30ns MD simulations were performed, allowing more PFAS molecules and poses

to be considered. PFASs alternatives such as ADONA, GenX, 6:2 FTOH, 6:2 FTCA, Et-PFOSA-AcOH and Met-PFOSA-AcOH have large binding energies, but not all of them showed hydrogen bonding with Tyr 473 during MD simulations. Short-chain PFASs exhibit binding towards PPARg, yet they show limited hydrogen bonding with Tyr 473. PFASs that have between six and twelve carbons form strong hydrogen bonds with Tyr 473 and alter PPARγ's activi8ty. L-Carnitine forms strong hydrogen bonds as an acceptor with Tyr 327, Lys 367, His 449 and Tyr 473 (Figure 8.8). As a donor, it also interacts with Ser 289. ADONA is a proposed alternative to PFASs and also forms a hydrogen bond with Tyr 473, which shows its ability to activate PPARγ. Tyr 327 and Lys 367 form a hydrogen bond with a range of PFASs.

8.4 Conclusions

The interactions of twenty-seven PFAS molecules and one of its natural ligands, L-carnitine with two potential binding pockets on the PPARγ dimer were investigated. Possible poses for the PFASs and L-carnitine, their binding energies, and important residue interactions, including hydrogen bond analysis were evaluated. The role of the dimer pocket is discussed and shown to be important for binding PFASs and L-carnitine. The PFASs' binding energies predicted for the dimer pocket show evidence for potential bioaccumulation of PFASs at this site. Significant correlation is observed between the predicted binding energies for the LBP and experimental IC_{50} values of PFASs in PPARγ, which allowed the activity of the remaining PFASs to be estimated.

Shorter-chain PFASs, such as PFBA, PFPA, 6:2 FTCA, Met-PFOSA-AcOH and Et-PFOSA-AcOH bind strongly to the dimer pocket, which indicates their potential bioaccumulation at this site. The PFASs in this study that have between six and twelve carbons form strong hydrogen bonds with Tyr 473 and alter the activity of PPARγ. PFAS alternatives such as ADONA, GENX,

6:2 FTOH, 6:2 FTCA, Et-PFOSA-AcOH and Met-PFOSA-AcOH also have large binding energies, but not all of them showed hydrogen bonding with Tyr 473 during MD simulations, which is deemed essential for PPARγ activation. L-Carnitine also showed hydrogen bonding with Tyr 473.

The affinity of L-carnitine to LBP determined by MMPBSA is -19.0 kcal mol^{-1}, which shows similar binding in comparison to most of the PFASs. In addition, acid/base, and short distance residue interactions contribute more towards the L-carnitine binding affinity than towards the studied PFASs. For the dimer pocket the binding affinity of L-carnitine is one of the largest binding energies. The high affinity of L-carnitine to both pockets, demonstrates that it could viably be used to compete/replace PFASs from the binding sites. The important interactions detailed here can provide useful insight about how these species may interact with other proteins, and about traits that may be important in building an inhibitor that can help to alleviate the effects of these “forever chemicals” on PPARγ.

APPENDIX

Table 8. 1 The PFASs used in this study are listed and are categorized based on their structural families: perfluoroalkyl carboxylic acids (PFCAs), perfluorosulfonic acids (PFSAs), fluoro telomer alcohols (FTOH), fluoro telomer sulfonic acids (FTSA), fluoro telomer carboxylic acids (FTCA).

Type	Acronym	Perfluorinated Carbon	Name	Chemical Formula
PFCA	PFBA	3	perfluorobutanoic acid	CF_3-$(CF_2)_2$-COOH
PFCA	PFPA	4	perfluoropentanoic acid	CF_3-$(CF_2)_3$-COOH
PFCA	PFHxA	5	perfluorohexanoic acid	CF_3-$(CF_2)_4$-COOH
PFCA	PFHpA	6	perfluoroheptanoic acid	CF_3-$(CF_2)_5$-COOH
PFCA	PFOA	7	perfluorooctanoic acid	CF_3-$(CF_2)_6$-COOH
PFCA	PFNA	8	perfluorononanoic acid	CF_3-$(CF_2)_7$-COOH
PFCA	PFDA	9	perfluorodecanoic acid	CF_3-$(CF_2)_8$-COOH
PFCA	PFUnDA	10	perfluoroundecanoic acid	CF_3-$(CF_2)_9$-COOH
PFCA	PFDoA	11	perfluorododecanoic acid	CF_3-$(CF_2)_{10}$-COOH
PFCA	PFTeDA	13	perfluorotetradecanoic acid	CF_3-$(CF_2)_{12}$-COOH
PFCA	PFHxDA	15	perfluorohexadecanoic acid	CF_3-$(CF_2)_{14}$-COOH
PFCA	PFOcDA	17	perfluorooctadecanoic acid	CF_3-$(CF_2)_{16}$-COOH
PFSA	PFBS	4	perfluorobutane sulfonic acid	CF_3-$(CF_2)_3$-SO_3H
PFSA	PFHxS	6	perfluorohexa sulfonic acid	CF_3-$(CF_2)_5$-SO_3H
PFSA	PFHpS	7	perfluoroheptane sulfonic acid	CF_3-$(CF_2)_6$-SO_3H
PFSA	PFOS	8	perfluorooctane sulfonic acid	CF_3-$(CF_2)_7$-SO_3H
PFSA	PFDS	10	perfluorodecane sulfonic acid	CF_3-$(CF_2)_9$-SO_3H
FTOH	6:2 FTOH	6	6:2 fluorotelomer alcohol	CF_3-$(CF_2)_5$-$(CH_2)_2$-OH
FTOH	8:2 FTOH	8	8:2 fluorotelomer alcohol	CF_3-$(CF_2)_7$-$(CH_2)_2$-OH
FTCA	5:3 FTCA	5	5:3 Fluorotelomer Carboxylic Acid	CF_3-$(CF_2)_4$-$(CH_2)_2$-COOH
FTCA	6:2 FTCA	6	6:2 Fluorotelomer Carboxylic Acid	CF_3-$(CF_2)_5$-CH_2-COOH
FTSA	6:2 FTSA	6	6:2 Fluorotelomer Sulfonic Acid	CF_3-$(CF_2)_5$-$(CH_2)_2$- SO_3H

Table 8. 1 (cont'd)

Alternative	GenX	5	2,3,3,3-tetrafluoro-2-heptafluoropropoxy Propanoic Acid	CF_3-$(CF_2)_2$-O-(CF_3)CF-COOH
Alternative	ADONA	6	4,8-dioxa-3H-perfluorononanoic acid	CF_3-O-$(CF_2)_3$-O-CHF-CF_2-COOH
-	PFOSA	8	Perfluorooctane Sulfanamido	CF_3-$(CF_2)_7$-SO_2NH_2
-	Et-PFOSA-AcOH	8	2-(N-Ethylperfluorooctanesulfoamido) Acetic Acid	CF_3-$(CF_2)_7$-$SO_2N(C_2H_5)$-CH_2-COOH
-	Me-PFOSA-AcOH	6	2-(N-Methylperfluorooctanesulfoamido) acetic acid	CF_3-$(CF_2)_7$-$SO_2N(CH_3)$-CH_2-COOH
PFSA = CF_3-$(CF_2)_n$-SO_3H				
PFCA = CF_3-$(CF_2)_n$-COOH				
FTOH = CF_3-$(CF_2)_n$-$(CH_2)_m$-OH				
FTSA = CF_3-$(CF_2)_n$-$(CH_2)_m$- SO_3H				
FTCA = CF_3-$(CF_2)_n$-$(CH_2)_m$-COOH				

Table 8. 2 PFASs chemical structures used in this study.

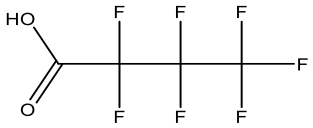
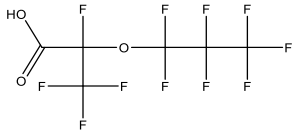
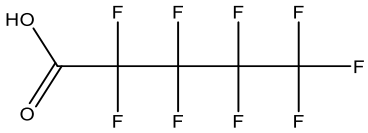
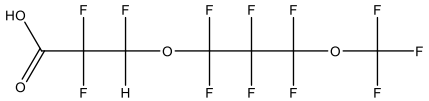
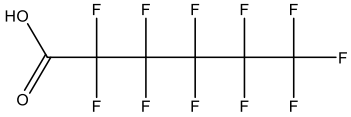
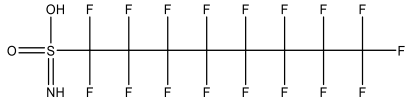
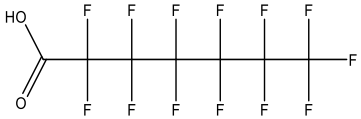
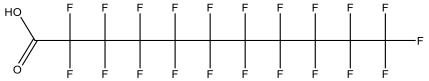
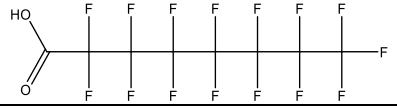
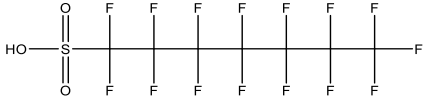
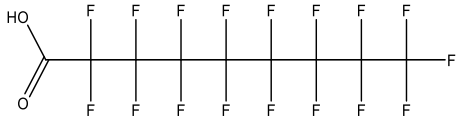
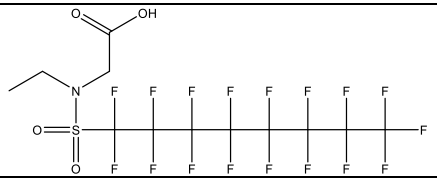
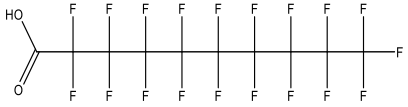
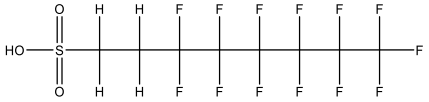
Structure	Name	Structure	Name
	Perfluorobutanoic Acid (PFBA, CAS No. 375-22-4)		2,3,3,3-tetrafluoro-2-heptafluoropropoxypropanoic acid (GenX, CAS No. 62037-80-3)
	Perfluoropentanoic Acid (PFPA, CAS No. 2706-90-3)		(ADONA, CAS No. 958445-448)
	Perfluorohexanoic Acid (PFHxA, CAS No. 307-24-4)		Perfluorooctane Sulfanamido (PFOSA, CAS No. 754-91-6)
	Perfluoroheptanoic Acid (PFHpA, CAS No. 375-85-9)		Perfluoroundecanoic Acid (PFUnDA CAS No. 2058-94-8)
	Perfluorooctanoic Acid (PFOA, CAS No. 335-67-1)		Perfluoroheptanesulfonic Acid (PFHpS, CAS No. 375-92-8)
	Perfluorononanoic Acid (PFNA, CAS No. 375-95-1)		2-N-Ethylperfluoroocatensulfanomido-Aceticacid (Et-PFOSA-AcOH, CAS No. 2991-50-6)
	Perfluorodecanoic Acid (PFDA, CAS No. 335-76-2)		6:2 FluorotelomerSulfonic Acid (6:2 FTSA, CAS No. 27619-97-2)

Table 8. 2 (cont'd)

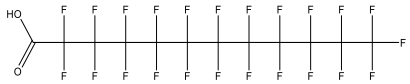
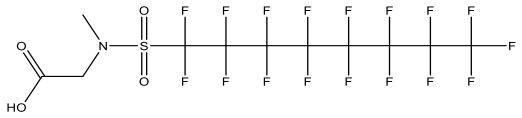
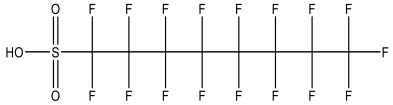
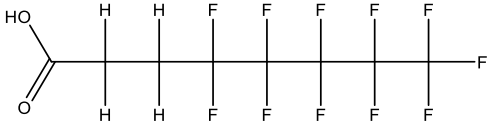
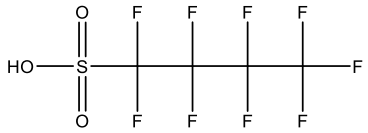
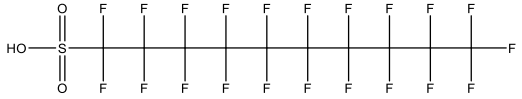
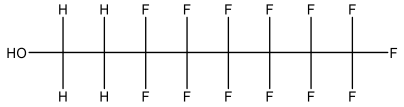
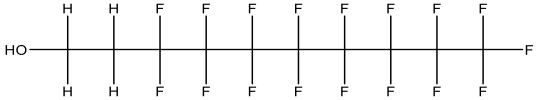
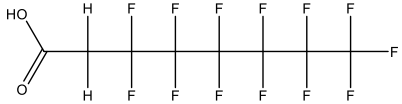
	Perfluorododecanoic Acid (PFDoA, CAS No. 307-55-1)		2NMethylperfluorooctanesulfonamido Aceticacid (Me-PFOSA-AcOH, CAS No. 2355-31-9)
	Perfluorooctanesulfonic Acid (PFOS, CAS No. 1763-23-1)		2H,2H,3H,3H-Perfluorooctanoic Acid (5:3 FTCA, CAS No. 914637-49-3)
	Perfluorobutanesulfonic Acid (PFBS, CAS No. 375-73-5)		PerfluorodecanesulfonicAcid (PFDS, CAS No. 335-77-3)
	6:2 Fluorotelomer Alcohol (6:2 FTOH, CAS No. 647-42-7)		8:2 FluorotelomerAlcohol (8:2 FTOH, CAS No. 678-39-7)
	6:2 Fluorotelomer Carboxylic Acid (6:2 FTCA, CAS No. 647-42-7)		

Table 8. 3 Binding energies for the dimer pocket and standard deviations in kcal mol^{-1} for all PFASs and L-carnitine.

Compound name	Average MMGBSA binding energy	STD MMGBSA	Average MMPBSA binding energy	STD MMPBSA
PFBA	-17.8	3.2	-15.2	4.0
PFPA	-12.8	3.9	-11.3	4.7
PFHxA	-7.4	3.5	-9.3	4.4
PFHpA	-8.9	3.4	-14.6	4.1
PFOA	-6.6	3.6	-13.5	4.2
PFNA	-8.2	4.0	-16.9	4.3
PFDA	-1.3	4.2	-12.8	3.7
PFUnDA	-8.2	3.7	-17.8	4.3
PFDoA	3.5	4.4	-12.4	4.7
PFTeDA	-2.5	4.2	-19.7	4.7
PFHxDA	-0.8	4.4	-21.8	4.6
PFBS	-18.1	3.7	-15.7	3.7
PFHpS	-9.1	3.3	-14.0	3.9
PFOS	-9.9	4.0	-16.8	4.3
PFDS	-8.8	4.0	-16.9	4.0
6:2 FTOH	-9.2	3.6	-17.5	3.7
8:2 FTOH	-2.2	4.2	-10.9	4.1
5:3 FTCA	-18.0	5.0	-16.7	5.1
6:2 FTCA	-13.3	4.3	-19.3	5.7
6:2 FTSA	-27.8	5.3	-18.2	5.1
GenX	-18.8	3.6	-19.6	4.3
ADONA	-14.2	4.2	-11.8	5.4
PFOSA	-16.7	4.7	-15.0	5.2
Et-PFOSA-AcOH	-31.1	5.5	-26.9	5.2
Me-PFOSA-AcOH	-25.3	5.0	-25.9	5.4
L-carnitine	-19.0	5.4	-19.0	5.6

Table 8. 4 Binding energies for the ligand binding pocket (LBP) and standard deviations in kcal mol-1 for all PFASs and L-carnitine.

Compound name	Average MMGBSA binding energy	STD MMGBSA	Average MMPBSA binding energy	STD MMPBSA
PFBA	-17.7	2.5	-20.9	4.7
PFPA	-16.9	2.9	-18.4	4.0
PFHxA	-19.5	2.7	-21.6	4.4
PFHpA	-18.1	2.6	-21.7	4.2
PFOA	-17.1	3.0	-23.4	3.8
PFNA	-22.1	3.2	-28.7	3.9
PFDA	-23.8	3.8	-31.0	4.8
PFUnDA	-19.4	3.2	-28.3	3.7
PFDoA	-21.6	3.9	-27.9	4.2
PFTeDA	-14.0	3.4	-29.2	3.5
PFHxDA	-16.1	4.1	-35.5	4.1
PFOcDA	-15.3	4.1	-36.9	3.9
PFBS	-17.7	2.8	-17.6	4.2
PFHxS	-22.4	3.2	-21.6	4.4
PFHpS	-25.7	3.4	-26.6	4.2
PFOS	-24.9	3.9	-28.7	4.2
PFDS	-24.32	3.8	-29.7	3.8
6:2 FTOH	-14.1	2.7	-20.1	3.1
8:2 FTOH	-14.4	3.2	-23.3	2.6
5:3 FTCA	-17.4	3.2	-19.1	4.3
6:2 FTCA	-23.0	3.3	-27.0	4.5
6:2 FTSA	-21.9	3.7	-21.7	4.5
GenX	-17.2	2.7	-21.1	4.2
ADONA	-23.2	2.6	-24.7	3.9
PFOSA	-29.6	4.6	-27.7	4.9
Et-PFOSA-AcOH	-34.7	3.6	-30.8	4.2
Me-PFOSA-AcOH	-27.3	3.6	-27.6	4.2
L-carnitine	-31.8	3.3	-19.0	4.1

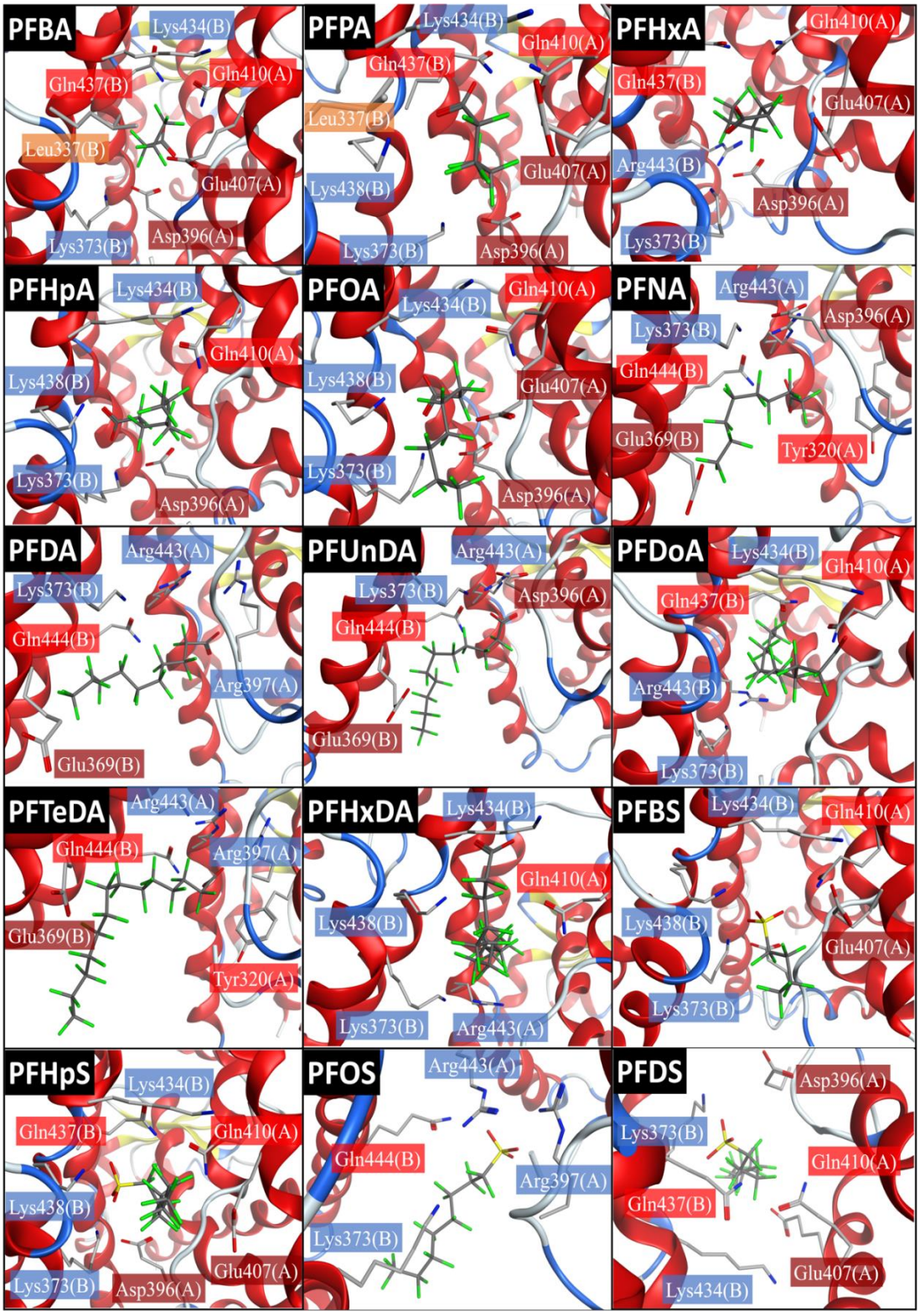
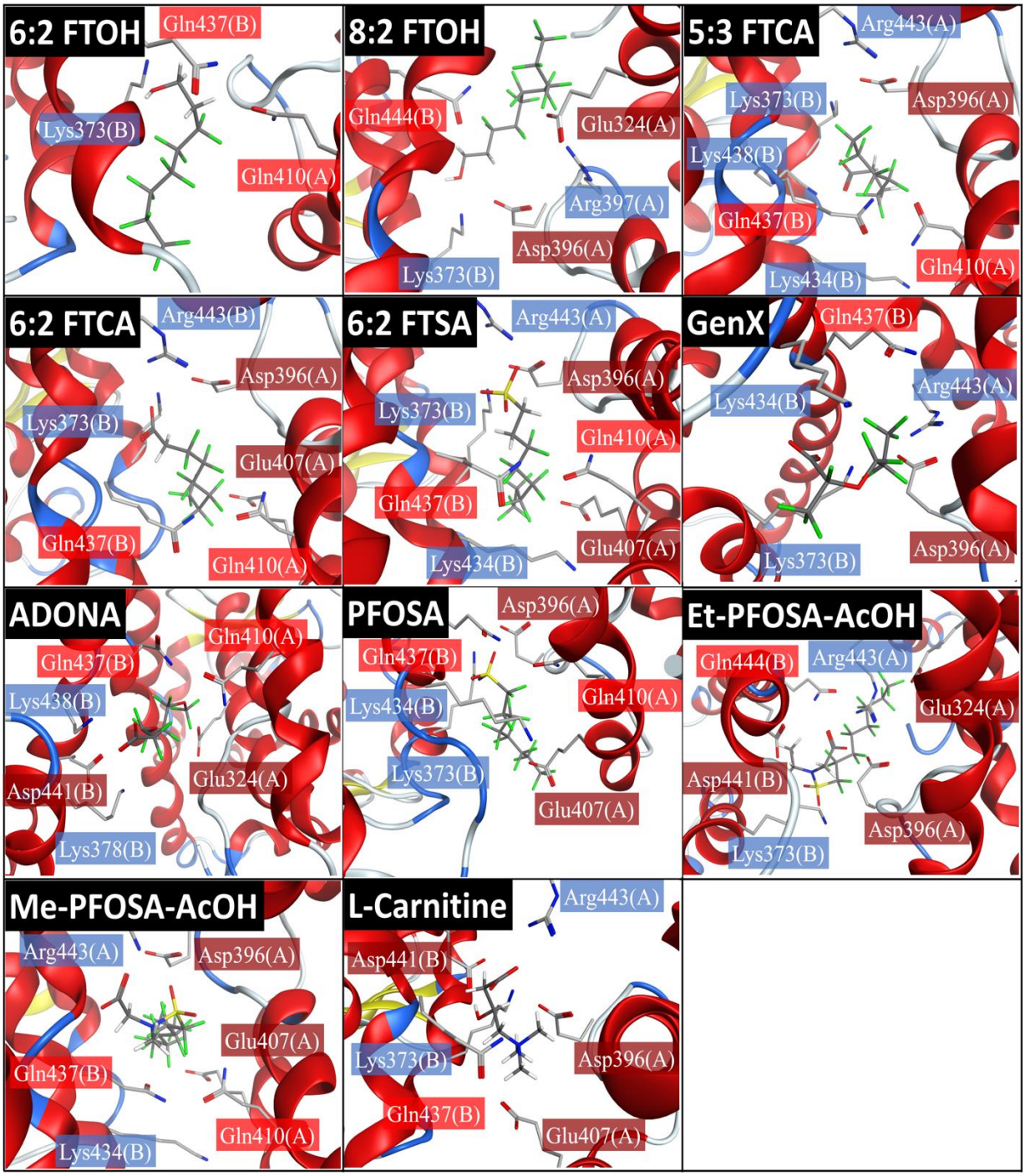


Figure 8. 9 Binding poses of PFASs and L-carnitine on the PPARγ dimer pocket. The binding modes that have the highest binding affinity determined from MM-PBSA are shown.

Figure 8. 9 (cont'd)



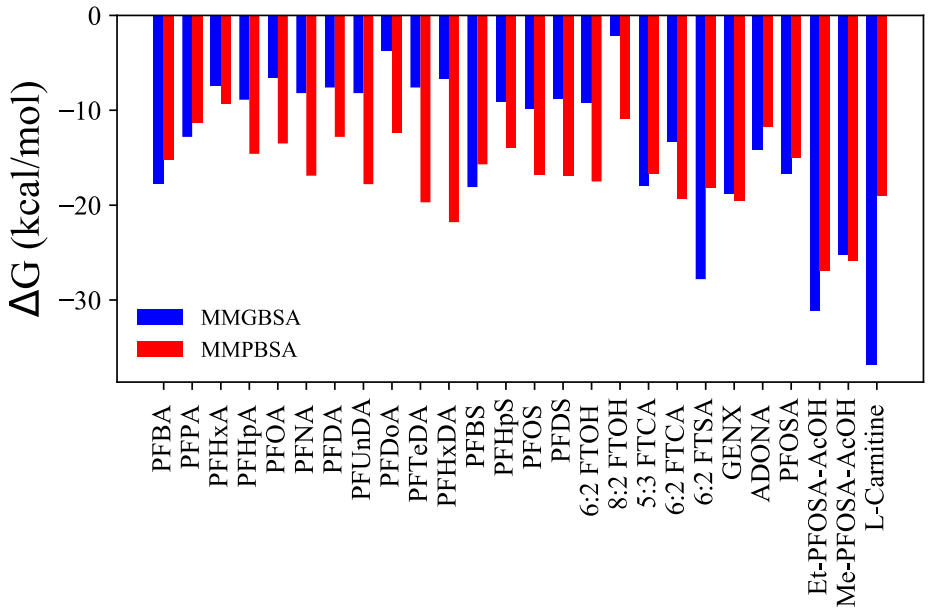


Figure 8. 10 Average binding energies of PFASs and L-carnitine calculated with MM-GBSA and MM-PBSA for the dimer pocket.

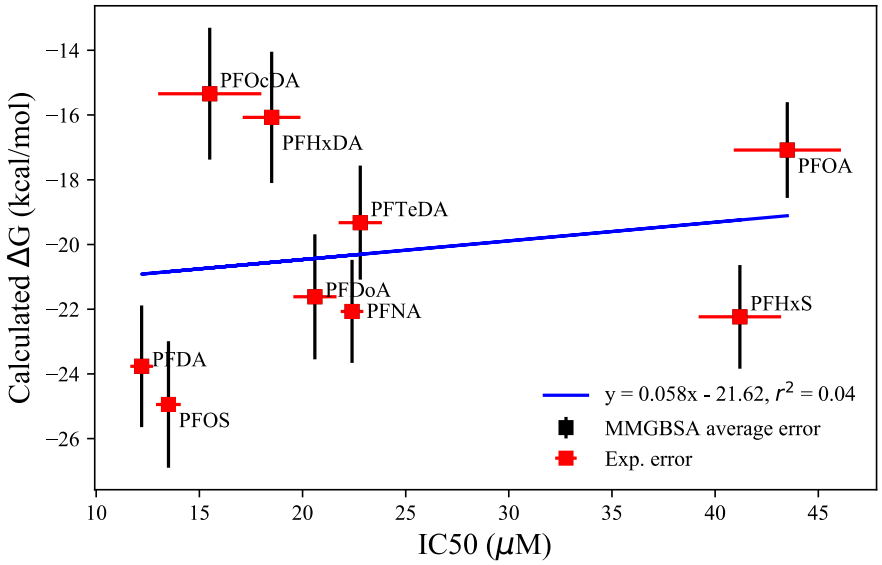


Figure 8. 11 MM-GBSA in comparison with IC50 values measured experimentally by Zhang et. al. for the LBP.[17] On the y-axis, average calculated binding energies are plotted, and along the x-axis, the experimental IC50 values are provided. Error bars are depicted in black (MM-GBSA) and red (experimental).

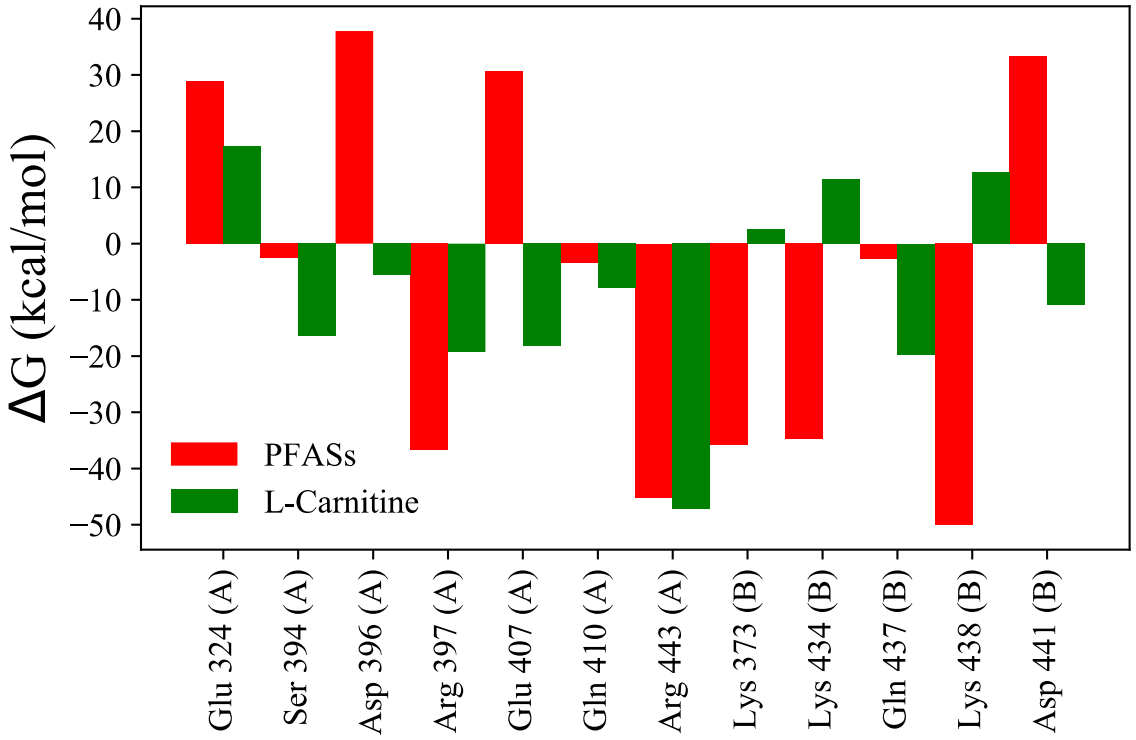


Figure 8. 12 Binding contribution of each nearby residue for PFASs and L-carnitine (dimer pocket).

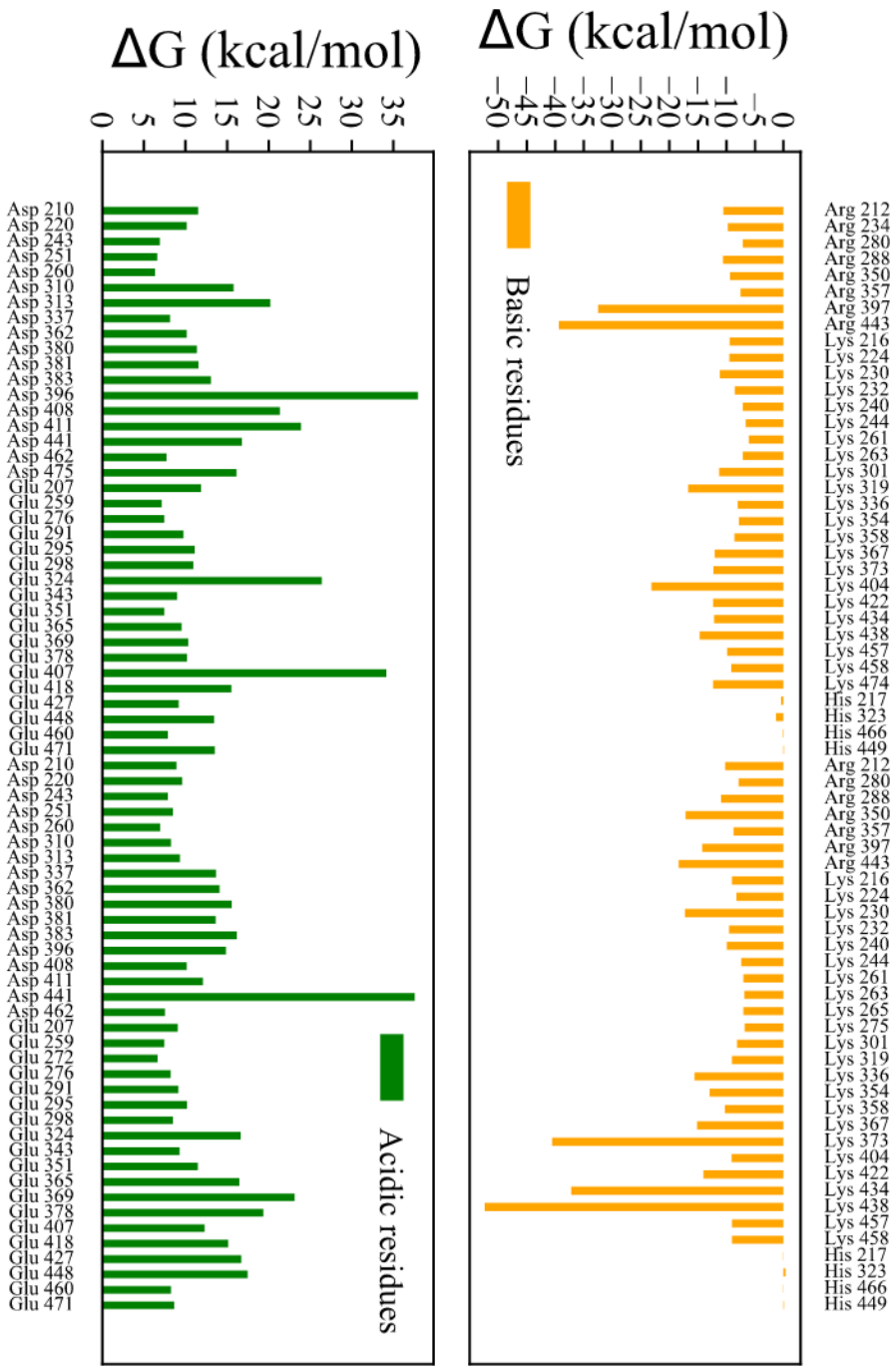


Figure 8. 13 Binding contributions of the acidic and basic residues for PFASs (dimer pocket) in Chain A and Chain B.

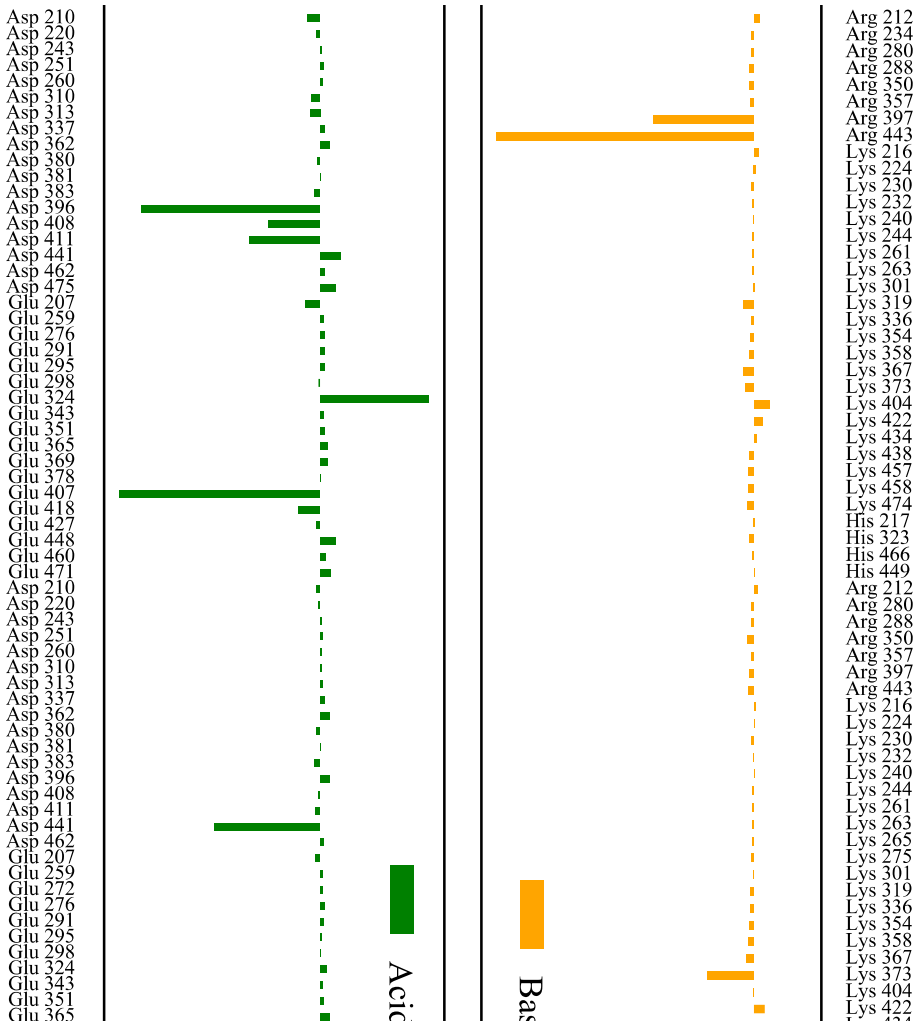


Figure 8. 14 Binding contributions of the acidic and basic residues for L-carnitine (dimer pocket) in Chain A and Chain B.

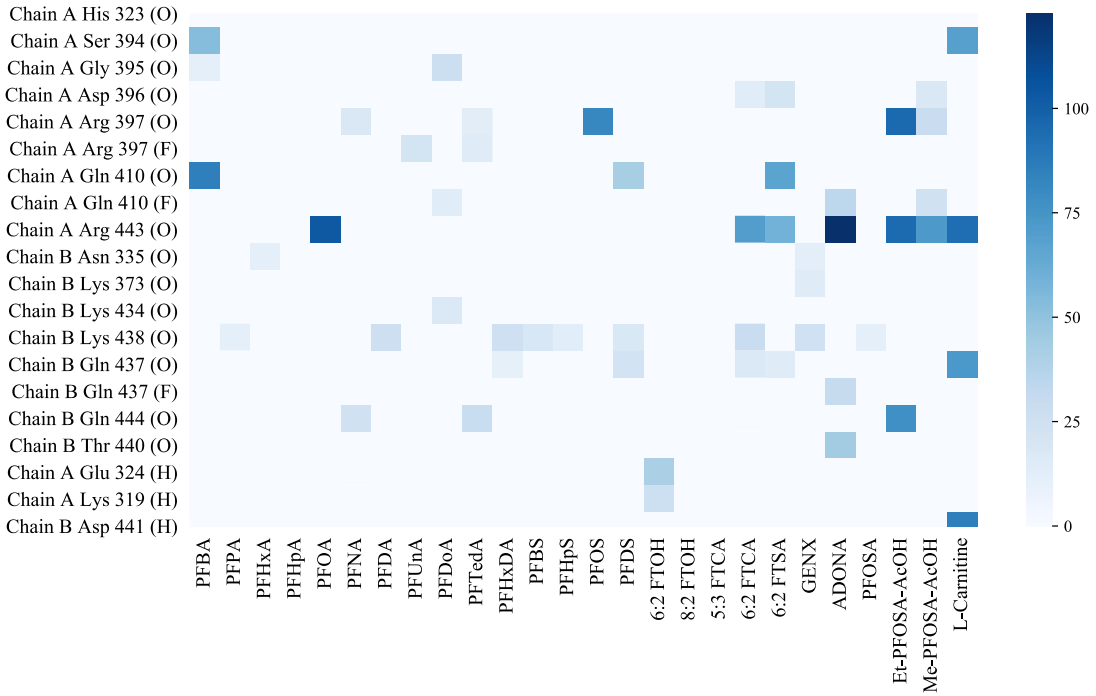


Figure 8. 15 Hydrogen bond lifetimes for the dimer pocket. The y-axis depicts the chain and residue number from the receptor, and in brackets, the atom from the ligand performing the hydrogen bonding is shown. Acceptors are portrayed by “(O), (F), (N)”, and donors by “(H)”. In the x-axis the different PFASs and L-Carnitine are shown

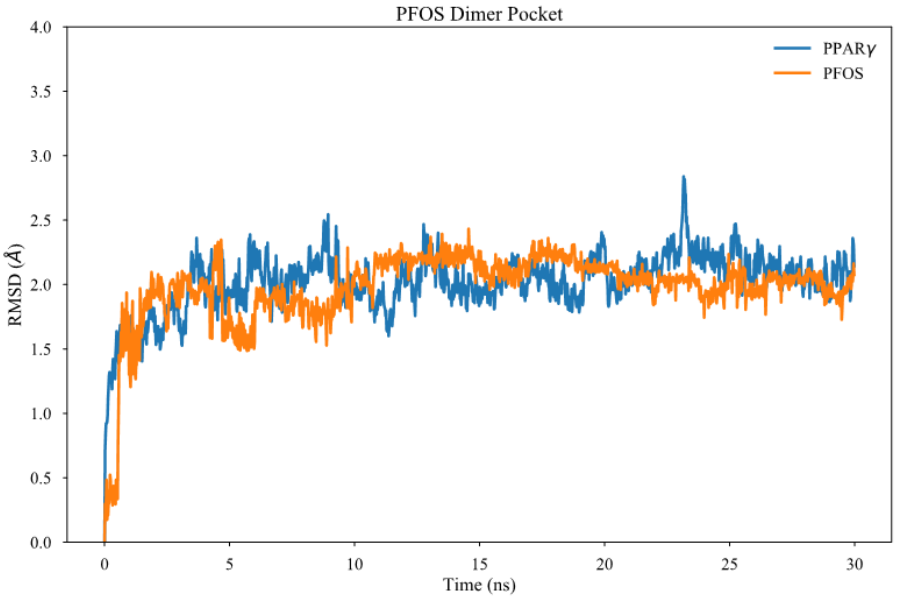


Figure 8. 16 PFOS RMSD plots for the dimer pocket.

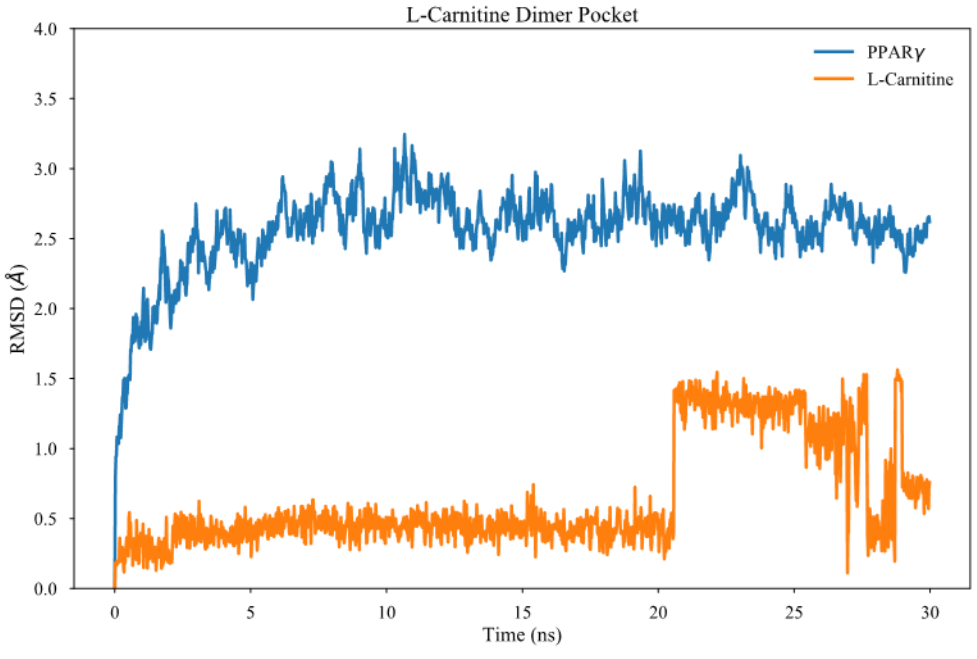


Figure 8. 17 L-Carnitine RMSD plots for the dimer pocket.

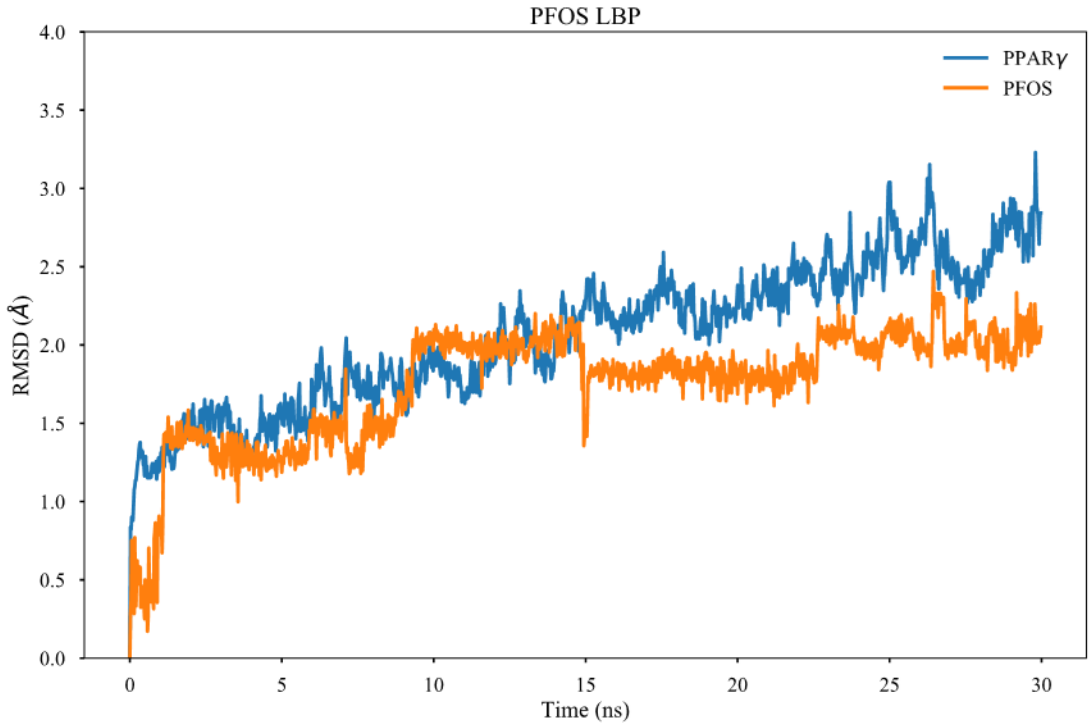


Figure 8. 18 PFOS RMSD plots for the LBP pocket.

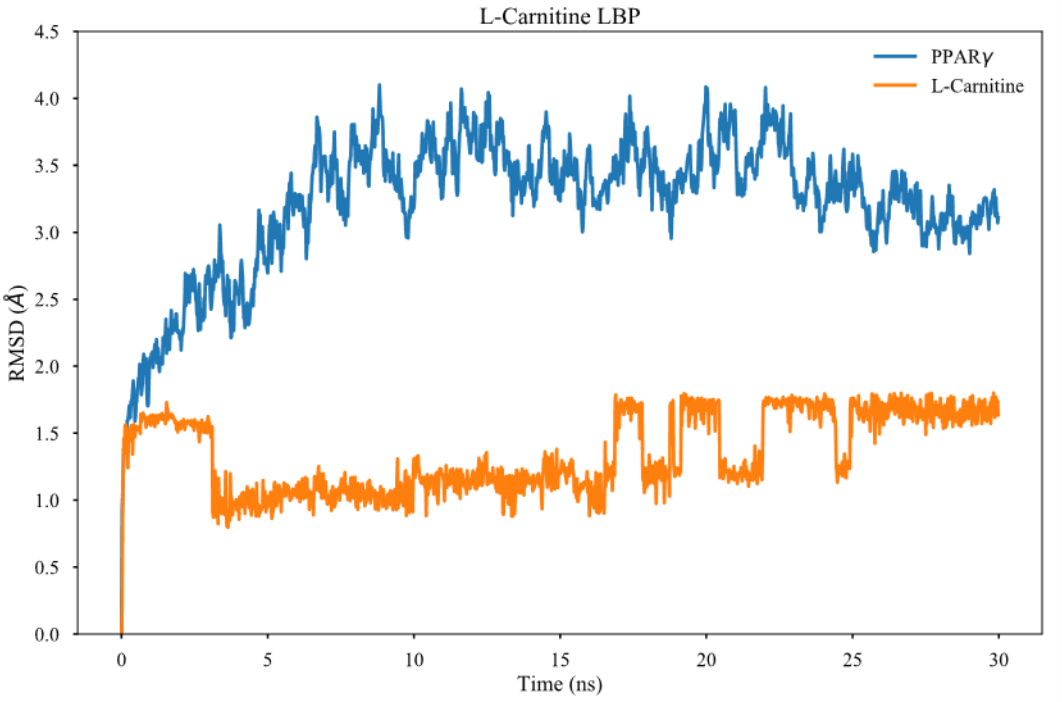


Figure 8. 19 L-Carnitine RMSD plots for the LBP pocket.

REFERENCES

REFERENCES

(1) Sinclair, G. M.; Long, S. M.; Jones, O. A. H. What Are the Effects of PFAS Exposure at Environmentally Relevant Concentrations? *Chemosphere* **2020**, *258*, 127340. https://doi.org/10.1016/j.chemosphere.2020.127340.

(2) Paul, A. G.; Jones, K. C.; Sweetman, A. J. A First Global Production, Emission, and Environmental Inventory for Perfluorooctane Sulfonate. *Environ. Sci. Technol.* **2009**, *43* (2), 386–392. https://doi.org/10.1021/es802216n.

(3) Sajid, M.; Ilyas, M. PTFE-Coated Non-Stick Cookware and Toxicity Concerns: A Perspective. *Environ. Sci. Pollut. Res.* **2017**, *24* (30), 23436–23440. https://doi.org/10.1007/s11356-017-0095-y.

(4) Rao, N. S.; Baker, B. E. Textile Finishes and Fluorosurfactants. In *Organofluorine Chemistry*; Banks, R. E., Smart, B. E., Tatlow, J. C., Eds.; Springer US: Boston, MA, 1994; pp 321–338. https://doi.org/10.1007/978-1-4899-1202-2_15.

(5) Schaider, L. A.; Balan, S. A.; Blum, A.; Andrews, D. Q.; Strynar, M. J.; Dickinson, M. E.; Lunderberg, D. M.; Lang, J. R.; Peaslee, G. F. Fluorinated Compounds in U.S. Fast Food Packaging. *Environ. Sci. Technol. Lett.* **2017**, *4* (3), 105–111. https://doi.org/10.1021/acs.estlett.6b00435.

(6) *State of Minnesota. Civil Action No. 27-CV-10-28862, State of Minnesota, et Al. v. 3M Company. Expert Report of Philippe Grandjean, MD, DMSc. Prepared on Behalf of Plaintiff State of Minnesota; State of Minnesota District Court for the County of Hennepin*; 2017.

(7) US EPA. EPA and 3M announce phase out of PFOS https://yosemite.epa.gov/opa/admpress.nsf/0/33aa946e6cb11f35852568e1005246b4.

(8) Wang, Z.; Dewitt, J. C.; Higgins, C. P.; Cousins, I. T. A Never-Ending Story of Per- and Polyfluoroalkyl Substances (PFASs)? *Environ. Sci. Technol.* **2017**, *51* (5), 2508–2518. https://doi.org/10.1021/acs.est.6b04806.

(9) Post, G. B.; Gleason, J. A.; Cooper, K. R. Key Scientific Issues in Developing Drinking Water Guidelines for Perfluoroalkyl Acids: Contaminants of Emerging Concern. *PLoS Biol.* **2017**, *15* (12), e2002855. https://doi.org/10.1371/journal.pbio.2002855.

(10) Cordner, A.; De La Rosa, V. Y.; Schaider, L. A.; Rudel, R. A.; Richter, L.; Brown, P. Guideline Levels for PFOA and PFOS in Drinking Water: The Role of Scientific Uncertainty, Risk Assessment Decisions, and Social Factors. *J. Expo. Sci. Environ. Epidemiol.* **2019**, *29* (2), 157–171. https://doi.org/10.1038/s41370-018-0099-9.

(11) Zeng, Z.; Song, B.; Xiao, R.; Zeng, G.; Gong, J.; Chen, M.; Xu, P.; Zhang, P.; Shen, M.; Yi, H. Assessing the Human Health Risks of Perfluorooctane Sulfonate by in Vivo and in Vitro Studies. *Environ. Int.* **2019**, *126*, 598–610.

https://doi.org/10.1016/j.envint.2019.03.002.

(12) Takacs, M. L.; Abbott, B. D. Activation of Mouse and Human Peroxisome Proliferator–Activated Receptors (α, β/δ, γ) by Perfluorooctanoic Acid and Perfluorooctane Sulfonate. *Toxicol. Sci.* **2007**, *95* (1), 108–117. https://doi.org/10.1093/toxsci/kfl135.

(13) Ikeda, T.; Aiba, K.; Fukuda, K.; Tanaka, M. The Induction of Peroxisome Proliferation in Rat Liver by Perfluorinated Fatty Acids, Metabolically Inert Derivatives of Fatty Acids. *J. Biochem.* **1985**, *98* (2), 475–482.

(14) Butenhoff, J. L.; Pieterman, E.; Ehresman, D. J.; Gorman, G. S.; Olsen, G. W.; Chang, S. C.; Princen, H. M. G. Distribution of Perfluorooctanesulfonate and Perfluorooctanoate into Human Plasma Lipoprotein Fractions. *Toxicol. Lett.* **2012**, *210* (3), 360–365. https://doi.org/10.1016/j.toxlet.2012.02.013.

(15) MacManus-Spencer, L. A.; Tse, M. L.; Hebert, P. C.; Bischel, H. N.; Luthy, R. G. Binding of Perfluorocarboxylates to Serum Albumin: A Comparison of Analytical Methods. *Anal. Chem.* **2010**, *82* (3), 974–981. https://doi.org/10.1021/ac902238u.

(16) Zhang, X.; Chen, L.; Fei, X. C.; Ma, Y. S.; Gao, H. W. Binding of PFOS to Serum Albumin and DNA: Insight into the Molecular Toxicity of Perfluorochemicals. *BMC Mol. Biol.* **2009**, *10* (1), 16. https://doi.org/10.1186/1471-2199-10-16.

(17) Zhang, L.; Ren, X. M.; Wan, B.; Guo, L. H. Structure-Dependent Binding and Activation of Perfluorinated Compounds on Human Peroxisome Proliferator-Activated Receptor γ. *Toxicol. Appl. Pharmacol.* **2014**, *279* (3), 275–283. https://doi.org/10.1016/j.taap.2014.06.020.

(18) Behr, A. C.; Plinsch, C.; Braeuning, A.; Buhrke, T. Activation of Human Nuclear Receptors by Perfluoroalkylated Substances (PFAS). *Toxicol. Vitr.* **2020**. https://doi.org/10.1016/j.tiv.2019.104700.

(19) Pastoor, T. P.; Lee, K. P.; Perri, M. A.; Gillies, P. J. Biochemical and Morphological Studies of Ammonium Perfluorooctanoate-Induced Hepatomegaly and Peroxisome Proliferation. *Exp. Mol. Pathol.* **1987**, *47* (1), 98–109. https://doi.org/10.1016/0014-4800(87)90011-6.

(20) Abdellatif, A.; Preat, V.; Taper, H. S.; Roberfroid, M. The Modulation of Rat Liver Carcinogenesis by Perfluorooctanoic Acid, a Peroxisome Proliferator. *Toxicol. Appl. Pharmacol.* **1991**, *111* (3), 530–537. https://doi.org/10.1016/0041-008X(91)90257-F.

(21) Ren, X. M.; Qin, W. P.; Cao, L. Y.; Zhang, J.; Yang, Y.; Wan, B.; Guo, L. H. Binding Interactions of Perfluoroalkyl Substances with Thyroid Hormone Transport Proteins and Potential Toxicological Implications. *Toxicology* **2016**, *366–367*, 32–42. https://doi.org/10.1016/j.tox.2016.08.011.

(22) Zhang, L.; Ren, X. M.; Guo, L. H. Structure-Based Investigation on the Interaction of Perfluorinated Compounds with Human Liver Fatty Acid Binding Protein. *Environ. Sci.*

Technol. **2013**, *47* (19), 11293–11301. https://doi.org/10.1021/es4026722.

(23) Han, X.; Snow, T. A.; Kemper, R. A.; Jepson, G. W. Binding of Perfluorooctanoic Acid to Rat and Human Plasma Proteins. *Chem. Res. Toxicol.* **2003**, *16* (6), 775–781. https://doi.org/10.1021/tx034005w.

(24) Wang, Y.; Zhang, H.; Kang, Y.; Cao, J. Effects of Perfluorooctane Sulfonate on the Conformation and Activity of Bovine Serum Albumin. *J. Photochem. Photobiol. B Biol.* **2016**, *159*, 66–73. https://doi.org/10.1016/j.jphotobiol.2016.03.024.

(25) Beesoon, S.; Martin, J. W. Isomer-Specific Binding Affinity of Perfluorooctanesulfonate (PFOS) and Perfluorooctanoate (PFOA) to Serum Proteins. *Environ. Sci. Technol.* **2015**, *49* (9), 5722–5731. https://doi.org/10.1021/es505399w.

(26) Honda, M.; Muta, A.; Akasaka, T.; Inoue, Y.; Shimasaki, Y.; Kannan, K.; Okino, N.; Oshima, Y. Identification of Perfluorooctane Sulfonate Binding Protein in the Plasma of Tiger Pufferfish Takifugu Rubripes. *Ecotoxicol. Environ. Saf.* **2014**, *104* (1), 409–413. https://doi.org/10.1016/j.ecoenv.2013.11.010.

(27) Chou, H. C.; Wen, L. L.; Chang, C. C.; Lin, C. Y.; Jin, L.; Juan, S. H. L-Carnitine via PPARγ- and Sirt1-Dependent Mechanisms Attenuates Epithelial-Mesenchymal Transition and Renal Fibrosis Caused by Perfluorooctanesulfonate. *Toxicol. Sci.* **2017**, *160* (2), 217–229. https://doi.org/10.1093/toxsci/kfx183.

(28) Wen, L. L.; Lin, C. Y.; Chou, H. C.; Chang, C. C.; Lo, H. Y.; Juan, S. H. Perfluorooctanesulfonate Mediates Renal Tubular Cell Apoptosis through PPARgamma Inactivation. *PLoS One* **2016**, *11* (5), e0155190. https://doi.org/10.1371/journal.pone.0155190.

(29) Liu, W. S.; Lai, Y. T.; Chan, H. L.; Li, S. Y.; Lin, C. C.; Liu, C. K.; Tsou, H. H.; Liu, T. Y. Associations between Perfluorinated Chemicals and Serum Biochemical Markers and Performance Status in Uremic Patients under Hemodialysis. *PLoS One* **2018**, *13* (7), e0200271. https://doi.org/10.1371/journal.pone.0200271.

(30) Fulton, J.; Mazumder, B.; Whitchurch, J. B.; Monteiro, C. J.; Collins, H. M.; Chan, C. M.; Clemente, M. P.; Hernandez-Quiles, M.; Stewart, E. A.; Amoaku, W. M.; Moran, P. M.; Mongan, N. P.; Persson, J. L.; Ali, S.; Heery, D. M. Heterodimers of Photoreceptor-Specific Nuclear Receptor (PNR/NR2E3) and Peroxisome Proliferator-Activated Receptor-γ (PPARγ) Are Disrupted by Retinal Disease-Associated Mutations. *Cell Death Dis.* **2017**, *8* (3), e2677–e2677. https://doi.org/10.1038/cddis.2017.98.

(31) Todorov, V. T.; Desch, M.; Schmitt-Nilson, N.; Todorova, A.; Kurtz, A. Peroxisome Proliferator-Activated Receptor-γ Is Involved in the Control of Renin Gene Expression. *Hypertension* **2007**, *50* (5), 939–944. https://doi.org/10.1161/hypertensionaha.107.092817.

(32) Estany, J.; Ros-Freixedes, R.; Tor, M.; Pena, R. N. A Functional Variant in the Stearoyl-CoA Desaturase Gene Promoter Enhances Fatty Acid Desaturation in Pork. *PLoS One* **2014**, *9* (1), e86177. https://doi.org/10.1371/journal.pone.0086177.

(33) Okuno, M.; Arimoto, E.; Ikenobu, Y.; Nishihara, T.; Imagawa, M. Dual DNA-Binding Specificity of Peroxisome-Proliferator-Activated Receptor γ Controlled by Heterodimer Formation with Retinoid X Receptor α. *Biochem. J.* **2001**, *353* (2), 193–198. https://doi.org/10.1042/bj3530193.

(34) Nolte, R. T.; Wisely, G. B.; Westin, S.; Cobb, J. E.; Lambert, M. H.; Kurokawa, R.; Rosenfeld, M. G.; Willson, T. M.; Glass, C. K.; Milburn, M. V. Ligand Binding and Co-Activator Assembly of the Peroxisome Proliferator-Activated Receptor-γ. *Nature* **1998**, *395* (6698), 137–143. https://doi.org/10.1038/25931.

(35) Waku, T.; Shiraki, T.; Oyama, T.; Maebara, K.; Nakamori, R.; Morikawa, K. The Nuclear Receptor PPARγ Individually Responds to Serotonin-and Fatty Acid-Metabolites. *EMBO J.* **2010**, *29* (19), 3395–3407. https://doi.org/10.1038/emboj.2010.197.

(36) Salvalaglio, M.; Muscionico, I.; Cavallotti, C. Determination of Energies and Sites of Binding of PFOA and PFOS to Human Serum Albumin. *J. Phys. Chem. B* **2010**, *114* (46), 14860–14874. https://doi.org/10.1021/jp106584b.

(37) Ng, C. A.; Hungerbuehler, K. Exploring the Use of Molecular Docking to Identify Bioaccumulative Perfluorinated Alkyl Acids (PFAAs). *Environ. Sci. Technol.* **2015**, *49* (20), 12306–12314. https://doi.org/10.1021/acs.est.5b03000.

(38) Chen, H.; He, P.; Rao, H.; Wang, F.; Liu, H.; Yao, J. Systematic Investigation of the Toxic Mechanism of PFOA and PFOS on Bovine Serum Albumin by Spectroscopic and Molecular Modeling. *Chemosphere* **2015**, *129*, 217–224. https://doi.org/10.1016/j.chemosphere.2014.11.040.

(39) Zhang, W.; Xiong, X.; Wang, F.; Ge, Y.; Liu, Y. Studies of the Interaction between Ronidazole and Human Serum Albumin by Spectroscopic and Molecular Docking Methods. *J. Solution Chem.* **2013**, *42* (6), 1194–1206. https://doi.org/10.1007/s10953-013-0027-5.

(40) Cheng, W.; Ng, C. A. Predicting Relative Protein Affinity of Novel Per- and Polyfluoroalkyl Substances (PFASs) by An Efficient Molecular Dynamics Approach. *Environ. Sci. Technol.* **2018**, *52* (14), 7972–7980. https://doi.org/10.1021/acs.est.8b01268.

(41) Tsukahara, T.; Tsukahara, R.; Yasuda, S.; Makarova, N.; Valentine, W. J.; Allison, P.; Yuan, H.; Baker, D. L.; Li, Z.; Bittman, R.; Parrill, A.; Tigyi, G. Different Residues Mediate Recognition of 1-O-Oleyl-Lysophosphatidic Acid and Rosiglitazone in the Ligand Binding Domain of Peroxisome Proliferator-Activated Receptor. *J. Biol. Chem.* **2006**, *281* (6), 3398–3407. https://doi.org/10.1074/jbc.M510843200.

(42) Uppenberg, J.; Svensson, C.; Jaki, M.; Bertilsson, G.; Jendeberg, L.; Berkenstam, A. Crystal Structure of the Ligand Binding Domain of the Human Nuclear Receptor PPARgamma. *J. Biol. Chem.* **1998**, *273* (47), 31108–31112. https://doi.org/10.1074/jbc.273.47.31108.

(43) Zoete, V.; Grosdidier, A.; Michelin, O. Peroxisome Proliferator-Activated Receptor

Structures: Ligand Specificity, Molecular Switch and Interactions with Regulators. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **2007**, *1771* (8), 915–925. https://doi.org/10.1016/j.bbalip.2007.01.007.

(44) Li, C. H.; Ren, X. M.; Cao, L. Y.; Qin, W. P.; Guo, L. H. Investigation of Binding and Activity of Perfluoroalkyl Substances to the Human Peroxisome Proliferator-Activated Receptor β/δ. *Environ. Sci. Process. Impacts* **2019**, *21* (11), 1908–1914. https://doi.org/10.1039/c9em00218a.

(45) Wang, Z.; Cousins, I. T.; Scheringer, M.; Hungerbühler, K. Fluorinated Alternatives to Long-Chain Perfluoroalkyl Carboxylic Acids (PFCAs), Perfluoroalkane Sulfonic Acids (PFSAs) and Their Potential Precursors. *Environ. Int.* **2013**, *60*, 242–248. https://doi.org/10.1016/j.envint.2013.08.021.

(46) Poulsen, P. B.; Jensen, A. A.; Wallström, E.; Aps, E. *More Environmentally Friendly Alternatives to PFOS-Compounds and PFOA*; 2005.

(47) Wang, Y.; Chang, W.; Wang, L.; Zhang, Y.; Zhang, Y.; Wang, M.; Wang, Y.; Li, P. A Review of Sources, Multimedia Distribution and Health Risks of Novel Fluorinated Alternatives. *Ecotoxicol. Environ. Saf.* **2019**, *182*, 109402. https://doi.org/10.1016/j.ecoenv.2019.109402.

(48) Labute, P. Protonate3D: Assignment of Ionization States and Hydrogen Coordinates to Macromolecular Structures. *Proteins Struct. Funct. Bioinforma.* **2009**, *75* (1), 187–205. https://doi.org/10.1002/prot.22234.

(49) Tobergte, D. R.; Curtis, S. MOE Molecular Operating Environment. *Journal of Chemical Information and Modeling*. Montreal 2013, pp 1689–1699. https://doi.org/10.1017/CBO9781107415324.004.

(50) Labute, P.; Santavy, M. SiteFinder-Locating Binding Sites in Protein Structures http://www.chempcomp.com/journal/sitefind.htm%5Cnhttps://www.chemcomp.com/journal/sitefind.htm.

(51) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(52) Hoffmann, R. An Extended Hückel Theory. I. Hydrocarbons. *J. Chem. Phys.* **1963**, *39* (6), 1397–1412. https://doi.org/10.1063/1.1734456.

(53) 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 Struct. Funct. Genet.* **2006**, *65* (3), 712–725. https://doi.org/10.1002/prot.21123.

(54) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and

Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157–1174. https://doi.org/10.1002/jcc.20035.

(55) Corbeil, C. R.; Williams, C. I.; Labute, P. Variability in Docking Success Rates Due to Dataset Preparation. *J. Comput. Aided. Mol. Des.* **2012**, *26* (6), 775–786. https://doi.org/10.1007/s10822-012-9570-1.

(56) D.A. Case, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, D.; Greene, N. Homeyer, S. Izadi, A. Kovalenko, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. L.; D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A.; Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X.; Wu, L. Xiao, D. M. Y. and P. A. K. Amber17. 2017. https://doi.org/10.13140/RG.2.2.36172.41606.

(57) 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. *J. Chem. Theory Comput.* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(58) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23* (16), 1623–1641. https://doi.org/10.1002/jcc.10128.

(59) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J. Phys. Chem. B* **2008**, *112* (30), 9020–9041. https://doi.org/10.1021/jp8001614.

(60) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(61) Onufriev, A.; Bashford, D.; Case, D. A. Exploring Protein Native States and Large-Scale Conformational Changes with a Modified Generalized Born Model. *Proteins Struct. Funct. Genet.* **2004**, *55* (2), 383–394. https://doi.org/10.1002/prot.20033.

(62) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA.Py: An Efficient Program for End-State Free Energy Calculations. *J. Chem. Theory Comput.* **2012**, *8* (9), 3314–3321. https://doi.org/10.1021/ct300418h.

(63) 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. *J. Chem. Inf. Model.* **2011**, *51* (1), 69–82. https://doi.org/10.1021/ci100275a.

(64) Roe, D. R. Introduction to hydrogen bond analysis https://amber.utah.edu/AMBER-workshop/London-2015/Hbond/ (accessed Apr 19, 2019).

(65) Liberato, M. V.; Nascimento, A. S.; Ayers, S. D.; Lin, J. Z.; Cvoro, A.; Silveira, R. L.; Martínez, L.; Souza, P. C. T.; Saidemberg, D.; Deng, T.; Amato, A. A.; Togashi, M.; Hsueh, W. A.; Phillips, K.; Palma, M. S.; Neves, F. A. R.; Skaf, M. S.; Webb, P.; Polikarpov, I. Medium Chain Fatty Acids Are Selective Peroxisome Proliferator Activated Receptor (PPAR) γ Activators and Pan-PPAR Partial Agonists. *PLoS One* **2012**, *7* (5), 1–10. https://doi.org/10.1371/journal.pone.0036297.

(66) Shi, G. Q.; Dropinski, J. F.; McKeever, B. M.; Xu, S.; Becker, J. W.; Berger, J. P.; MacNaul, K. L.; Eibrecht, A.; Zhou, G.; Doebber, T. W.; Wang, P.; Chao, Y. S.; Forrest, M.; Heck, J. V.; Moller, D. E.; Jones, A. B. Design and Synthesis of α-Aryloxyphenylacetic Acid Derivatives: A Novel Class of PPARα/γ Dual Agonists with Potent Antihyperglycemic and Lipid Modulating Activity. *J. Med. Chem.* **2005**, *48* (13), 4457–4468. https://doi.org/10.1021/jm0502135.

(67) Kuwabara, N.; Oyama, T.; Tomioka, D.; Ohashi, M.; Yanagisawa, J.; Shimizu, T.; Miyachi, H. Peroxisome Proliferator-Activated Receptors (PPARs) Have Multiple Binding Points That Accommodate Ligands in Various Conformations: Phenylpropanoic Acid-Type PPAR Ligands Bind to PPAR in Different Conformations, Depending on the Subtype. *J. Med. Chem.* **2012**, *55* (2), 893–902. https://doi.org/10.1021/jm2014293.

(68) Hughes, T. S.; Giri, P. K.; de Vera, I. M. S.; Marciano, D. P.; Kuruvilla, D. S.; Shin, Y.; Blayo, A.-L.; Kamenecka, T. M.; Burris, T. P.; Griffin, P. R.; Kojetin, D. J. An Alternate Binding Site for PPARγ Ligands. *Nat. Commun.* **2014**, *5* (1), 3571. https://doi.org/10.1038/ncomms4571.

(69) Lai, T. T.; Eken, Y.; Wilson, A. K. Binding of Per- and Polyfluoroalkyl Substances to the Human Pregnane X Receptor. *Environ. Sci. Technol.* **2020**, *54* (24), 15986–15995. https://doi.org/10.1021/acs.est.0c04651.

(70) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opinion on Drug Discovery*. 2015, pp 449–461. https://doi.org/10.1517/17460441.2015.1032936.

CHAPTER NINE

Mechanisms behind Protein Kinase C (PKC) Activation

9.1 Introduction

Protein kinase C (PKC) encompass a family of serine/threonine kinases involved in controlling various signaling pathways that regulate cell proliferation, survival, apoptosis, migration, invasion, differentiation, angiogenesis, and drug resistance.[1] PKC acts by changing the activities of other PKC family members and proteins within signaling pathways by phosphorylation of the hydroxyl groups of serine and threonine residues. Members of the PKC family are considered promising targets for several diseases including multiple types of cancer, cardiovascular diseases, immune and inflammatory diseases, neurological and metabolic disorders due to their essential role in the cell cycle.[1] PKCs are considered to be suitable therapeutic targets, as there are no mutations in PKC encoding genes, thus, eliminating failures anticipated due to mutations.[2] While it has been a goal of academic and industrial researchers to develop PKC-specific inhibitors, a major challenge is targeting a specific kinase resulting from the highly similar structures of different PKC isoforms.[3]

Early studies have shown that in the absence of Ca^{2+}, PKCα weakly interacts with lipid bilayer.[4] As shown in Figure 9.1, the first step of activation is dependent on intracellular Ca^{2+} binding to the PKCα-C2 domain which increases its affinity for membranes and causes the enzyme to drift to the cell membrane (even though this initial electrostatic interaction is still low in affinity. After PKCα-C2 is docked to the membrane, it moves deeper into the membrane and interacts with phosphatidylinositol 4,5-bisphosphate (PIP2) completing the second step of activation.[5]

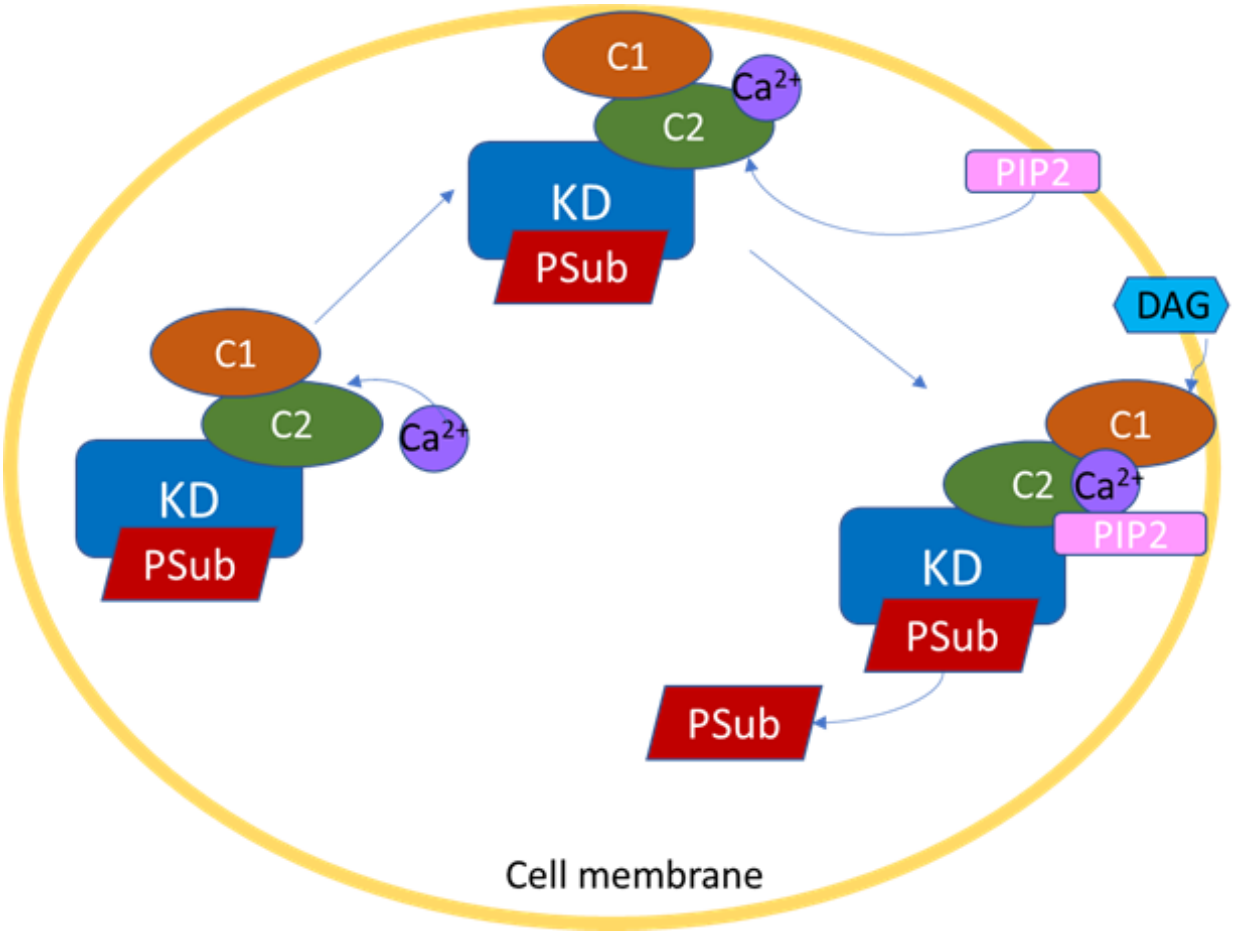


Figure 9. 1 A schematic of the PKC activation pathway. In the first activation step the Ca^{2+} binds to the C2 domain, increasing the membrane affinity of the enzyme and PKC drifts to the membrane. Next, PIP2 that is present in the membrane binds to the C2 domain and loosens the C1-C2 domain interaction causing the C1 domain to move inside the membrane where it can bind to DAG. After Ca^{2+}, PIP2 and DAG binding is established, the pseudo substrate domain leaves the active site in the kinase domain completing the activation of the enzyme.[4,6]

The third step occurs after this secondary interaction interrupts electrostatic C1/C2 inter domain binding and allows the C1 domain to penetrate the membrane and bind to the diacylglycerol (DAG). Even though both C1-DAG and C2-PIP2 interactions are relatively low in affinity, the combined energetics from the two leads to a strong binding to the membrane.

Establishing this strong binding is the key for the final activation step. The PKCα structure goes through a final conformational change, where the auto-inhibitory pseudo-substrate (PSub) domain is expelled from the kinase domain, leaving the active site of the enzyme available for substrate binding and thus completing the activation.[4]

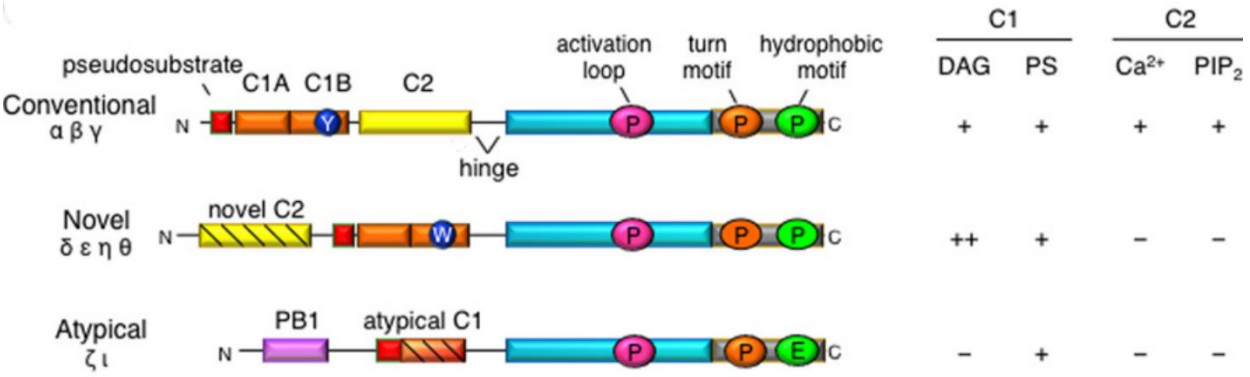


Figure 9. 2 PKC subgroups have slightly varying structures and regulators. All isoforms carry a kinase domain with an activation loop shown as blue. Both conventional and novel PKCs contain a C1 domain that can be regulated by DAG, PS as shown in orange, whereas atypical PKC C1 domain can only be regulated by PS. The C2 domain that can be regulated by Ca^{2+} and PIP2 is only present in the conventional subgroup, novel PKCs contain a modified C2 domain that lacks the necessary residues for binding. Atypical PKCs carry Phox and Bem 1 (PB1) domain instead of the C2 domain present in the other subgroups.[6]

Based on structure and cofactor regulation, PKC isozymes can be classified into three groups: conventional (cPKC α, β, γ), novel (nPKC ε, η, θ, δ), and atypical (aPKC ι, ζ).[2] As shown in Figure 9.2, all isoforms contain the kinase domain with an activation loop in the middle. This part is also known as the catalytic domain and contains the necessary motifs for ATP, substrate binding and also the residues that catalyze the kinase reaction.[4] In its inactive form the PSub domain blocks this active loop and prevents substrates to reach the active site.[7] Interaction between the PSub domain and the kinase domain must be interrupted by regulation of

C1 and C2 domains in order to have an active form of the enzyme. Both conventional, novel and atypical PKC structures contain a C1 domain regulated by the phosphatidylserine (PS), for cPKCs and nPKCs there is an additional DAG binding site present. C2 domain regulation is unique to cPKCs for two reasons. First, aPKCs do not have this domain, instead they carry the Phox and Bem 1 domain (PB1) where interactions between protein scaffolds are mediated. Second, nPKCs have C2 domains but their C2 domain lack the necessary amino acid residues that can stimulate Ca^{2+} or PIP2 binding that are essential for the activation of cPKCs. Thus, revealing the mechanisms behind C2 regulation covered, will bring out factors and potent target sites that can be used in the design of new therapeutics.

9.2 Methods

The domains of PKCs are compared using the NCBI Basic Local Alignment Search Tool (BLAST). Results are scored using the BLOSUM62 matrix with a gap cost 11 and extension 1.[8] Initial coordinates of PKCα-C2(PDB ID: 4DNL[9]) and PKCδ-C2(PDB ID: 1YRK[10]) are obtained from the Protein Data Bank (PDB ID: 4DNL); missing residues and hydrogens are added using Molecular Operating Environment v.2016.08 (MOE).[11] These structures are initially minimized in MOE with the AMBER ff10 force field. The systems were prepared using Gromacs-5.0.1[12] with the amber99sb[13] force field, and placed into a triclinic unit cell with a 1 nm solute box distance. The unit cell is solvated in SPC/E-type waters and ions corresponding to 150 mM NaCl, 150 mM $CaCl_2$, 100 mM $CaCl_2$, and 50 mM $CaCl_2$ solutions are explicitly replaced with random water molecules. First the systems are minimized using conjugate gradient algorithm for 10,000 steps, and the steepest descent method is used every tenth step. Then, water around the protein is equilibrated for 20 picoseconds by restraining protein atoms to their initial position. Next, the production simulations are performed by removing protein restraints. The trajectories

were produced using velocity rescale thermo couple to keep the temperature at 300 K, and Berendsen barostat to keep the pressure at 1 atm. The SHAKE algorithm was used to constrain bonds involving hydrogens,[14] vdw interactions treated with a 10 Å cutoff and long-range electrostatic interactions were modeled with PME also with a 10 Å cutoff.

9.2 Results and Discussion

9.2.1 Sequence Alignment

As noted earlier, based on structure and cofactor regulation, these isozymes can be classified into three groups: conventional (α, β, γ), novel (ε, η, θ, δ), and atypical (ι, ζ) PKCs. 7 In order to understand differences between these subgroups, sequence alignment was performed on different domains of PKCs. During these alignments, PKCα was used as a reference for alignment.

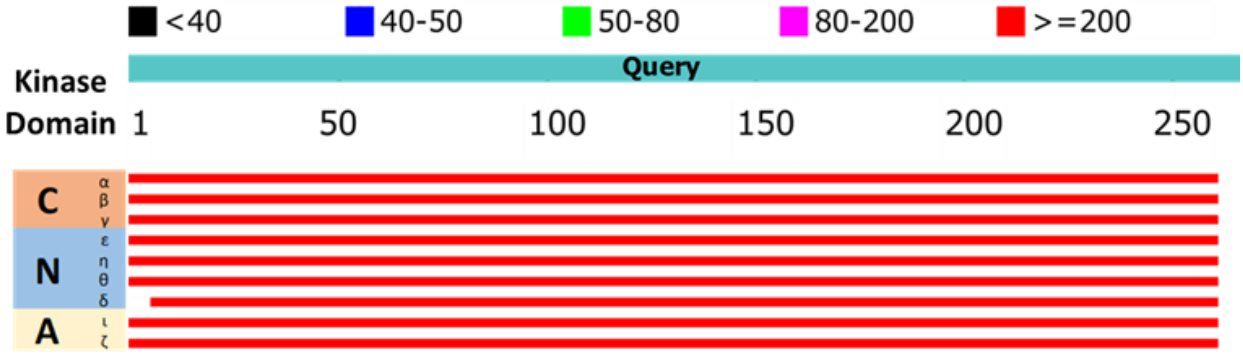


Figure 9. 3 Comparison of kinase domain of different PKC family isoforms with sequence alignment.

Sequence alignments of kinase domains of different PKCs resulted in scores higher than 200 for both isoforms in the same group and outside of the group which are compared (Figure 9.3). These results show that the kinase domain has the highest similarity among PKC isozymes. This domain carries an ATP binding domain and a kinase active site. Possible inhibitors of active sites have been developed but selective inhibition is extremely difficult because of this high similarity.

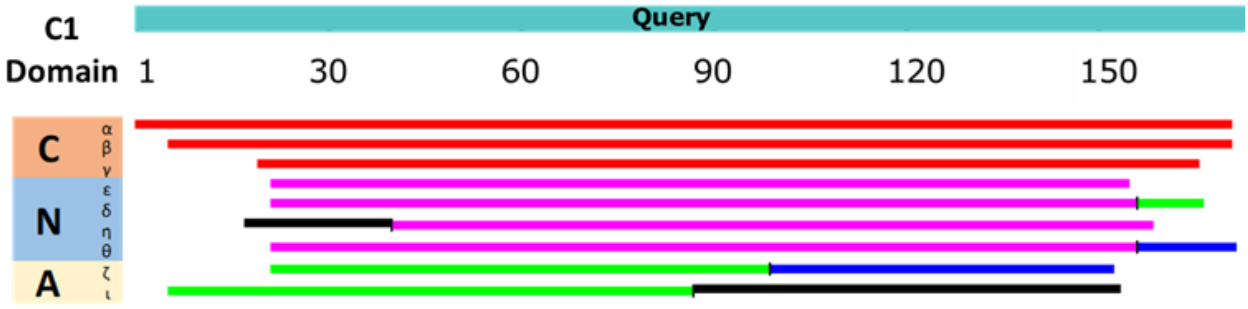


Figure 9. 4 PKC family C1 domain sequence alignment.

The C1 domain sequence alignment scores are higher than 200 for the conventional subgroup members, as shown in Fig. 9.4. This indicates that the C1 domain shows high similarity between the same members of the group. Whereas, when the PKCα-C1 sequence is aligned with members of the other groups, the score slightly decreases. This domain shows slight variation between members of different subgroups. This moderately similar part of the protein contains the potential binding site for DAG and phosphatidylserine.[4]

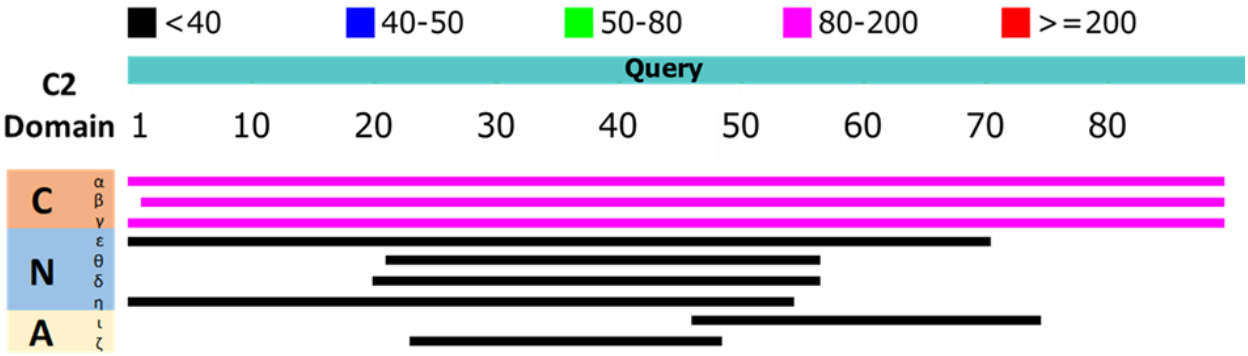


Figure 9. 5 PKC family C2 domain sequence alignment.

Sequence alignment scores of the C2 domain show that this domain contains slight differences among other members of the conventional PKC subgroup with a scores between 80-200 (Figure 9.5). However, the sequence alignment for others subgroup members results in very small scores, indicating that this domain which holds the Ca^{2+} and PIP2 binding site, is

significantly different among the different subgroups. The results of sequence alignment suggest that studying the C2 domain activation might hold a solution for the target specificity problem.

9.2.2 Binding Site Environment Comparison

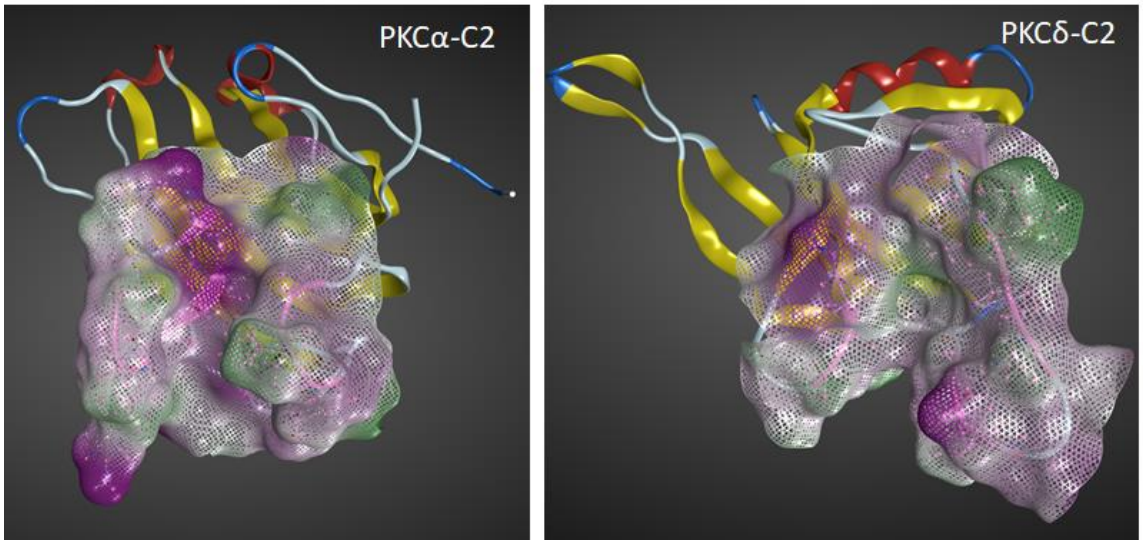


Figure 9. 6 PKCα-C2 and PKCδ-C2 binding site comparison. Potential sites for hydrogen bonding are in purple, hydrophobic regions in green, and neutral regions in white.

The sequence alignment results indicate that there is a significant number of differences between the C2 domains of the PKC subgroups. Our focus, thus, shifts to understanding how the C2 domain is regulated and what are the differences among these subgroups. Variations among PKCs C2 domain crystal structures of different subgroups are compared to one another (conventional PKCα and novel PKCδ. atypical PKCs lack this domain.) It is known that conventional PKCs hold a Ca^{2+} binding site at this domain (Figure 9.6 left). Investigation of potential binding sites using a geometrical approach[15,16] showed that novel PKCs also have a binding site similar in tertiary structure in place (Figure 9.6 right). The comparison of molecular surfaces are generated using a grid-based method[17] calculated in MOE program,[15] show that interior of the PKCα pocket is hydrophilic, whereas this site is hydrophobic in PKCδ.

Table 9. 1 Character of PKCα-C2 and PKCδ-C2 binding site residues as obtained from a comparison of potential binding site residues.

Site	Nonpolar (%)	Polar (%)	Positively charged (%)	Negatively charged (%)
PKCα	14	79	14	36
PKCδ	56	44	11	11

A comparison of residue contents of these two sites shows that the PKCα-C2 binding site residues are predominantly hydrophilic and make up 79% of the binding site residues, 50% of the residues are charged with 36% to 14% negative and positive, respectively and there is an overall negative charge due to the higher percentage of negatively charged residue (Table 9.1). In contrary, the PKCδ-C2 binding site consists of 56% hydrophobic residues, and there is no net charge within this site. These results show that even though PKCα-C2 and PKCδ-C2 are structurally similar by exhibiting a potential binding site in the same region, these binding sites have very different character.

9.2.3 Molecular Dynamics Simulations

To better understand the behavior of different PKCs in varying environments, both PKCα-C2 and PKCδ-C2 are placed in 150 mM NaCl, 100 mM $CaCl_2$, 50 mM $CaCl_2$, 150 mM $CaCl_2$ salt solutions and these structures are simulated for 100 ns using atomistic MD (Figures 9.7 and 9.8).

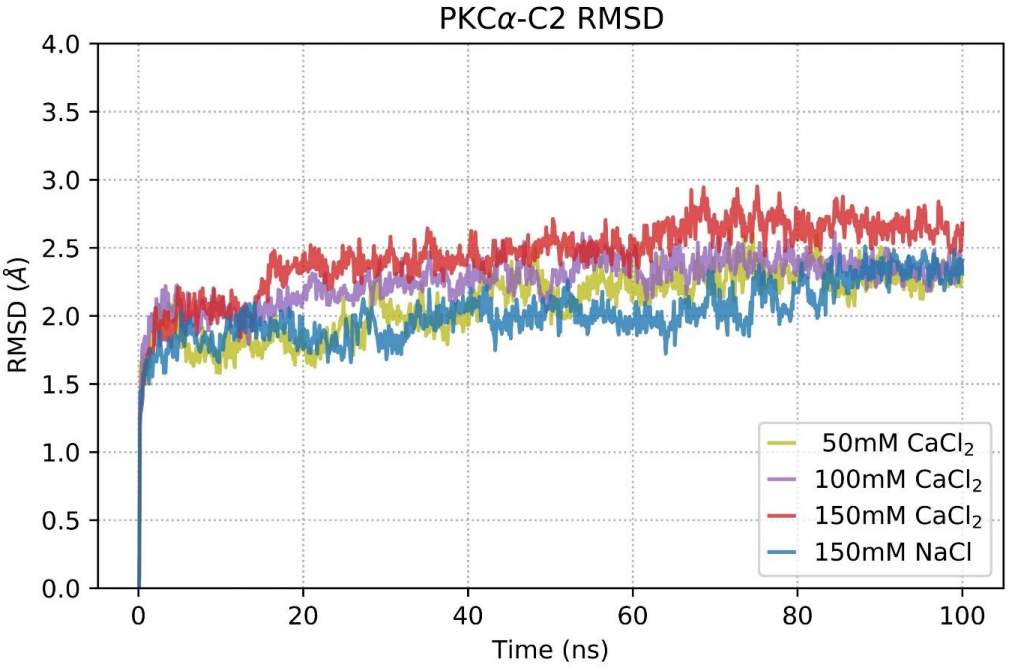


Figure 9. 7 PKCα-C2 domain RMSD for the systems in different salt concentrations.

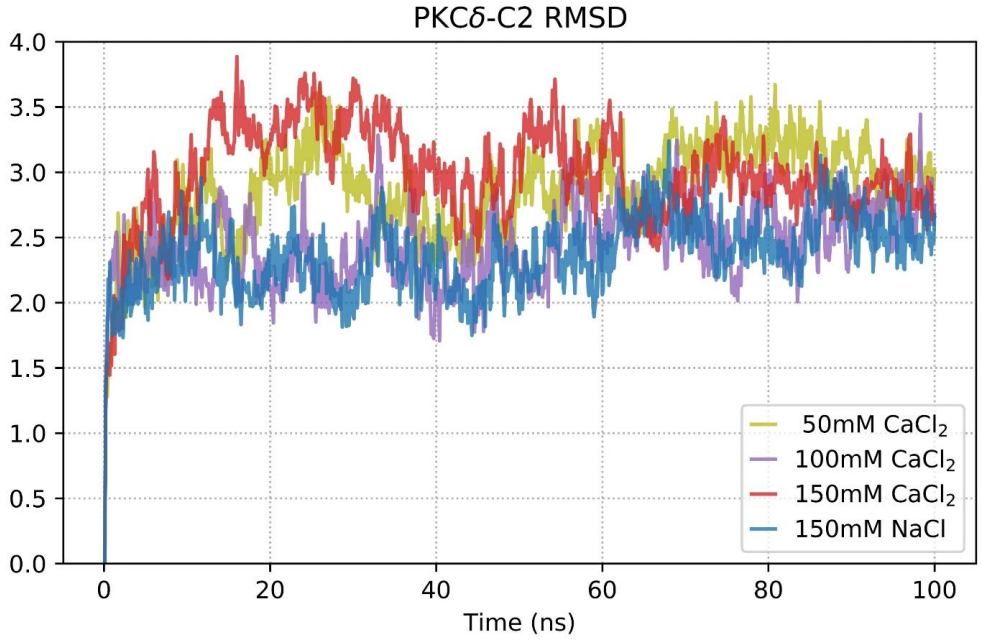


Figure 9. 8 PKCδ-C2 domain RMSD for the systems in different salt concentrations.

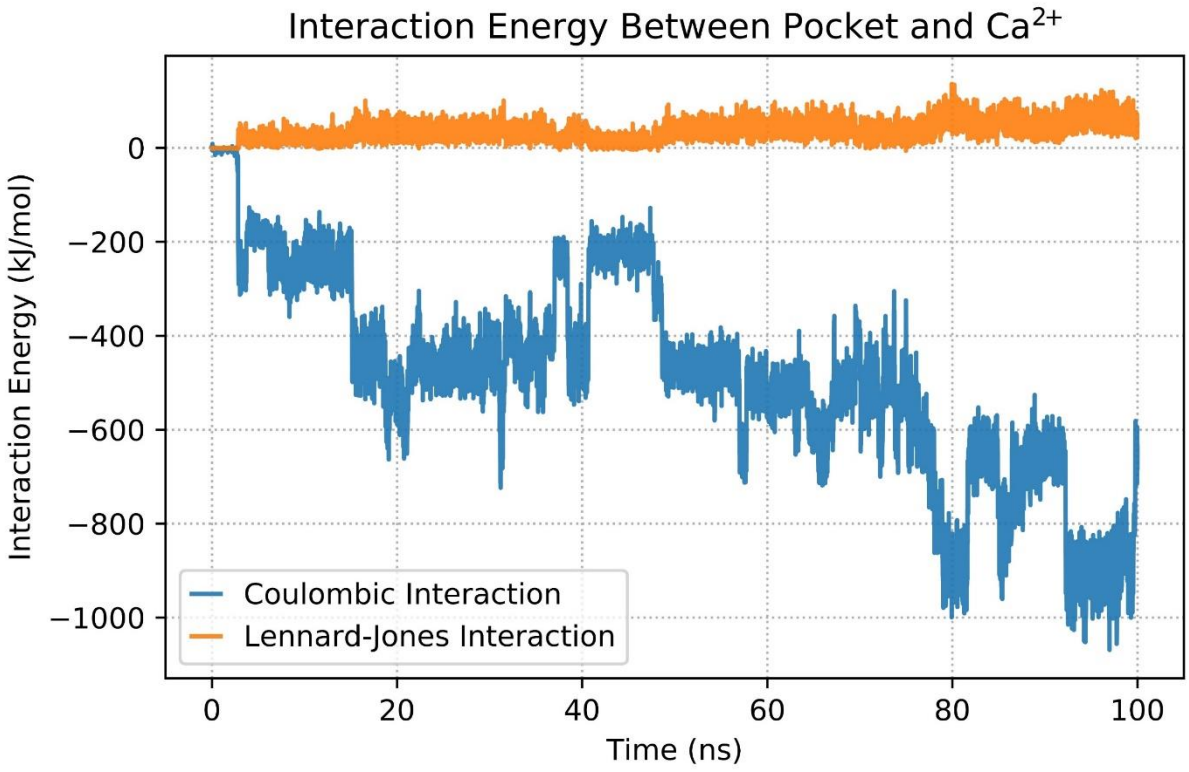


Figure 9. 9 Coulombic and Lennard-Jones interaction energy between PKCα-C2 binding site and Ca^{2+} ions in the system for extended simulation of PKCα-C2 in 150 mM $CaCl_2$.

In order to better understand the interactions between the two Ca^{2+} that are bound to the PKCα-C2, the MD simulations are extended to 100 ns for the PKCα-C2 in 150 mM $CaCl_2$, and Lennard Jones and Coulombic interaction energies between pocket residues and Ca^{2+} on the systems are analyzed. Figure 9.9 shows that electrostatic interaction is the dominant interaction type and even though the first Ca^{2+} binds to the system at 18 ns and the second Ca^{2+} enters the site at 58 ns, the electrostatic energy fluctuates during the entire simulation.

The electrostatic interaction with the first Ca^{2+} starts at 18 ns where the ion enters the pocket; at 50 ns its interaction energy increases corresponding to the Ca^{2+} moving deeper into the pocket. The second Ca^{2+} enters the binding site at 58 ns after the relocation of first Ca^{2+}. During the later

stages of simulation, it moves to a place where it can establish stronger binding as indicated by the increase of second Ca^{2+} interaction energy at 78 ns.

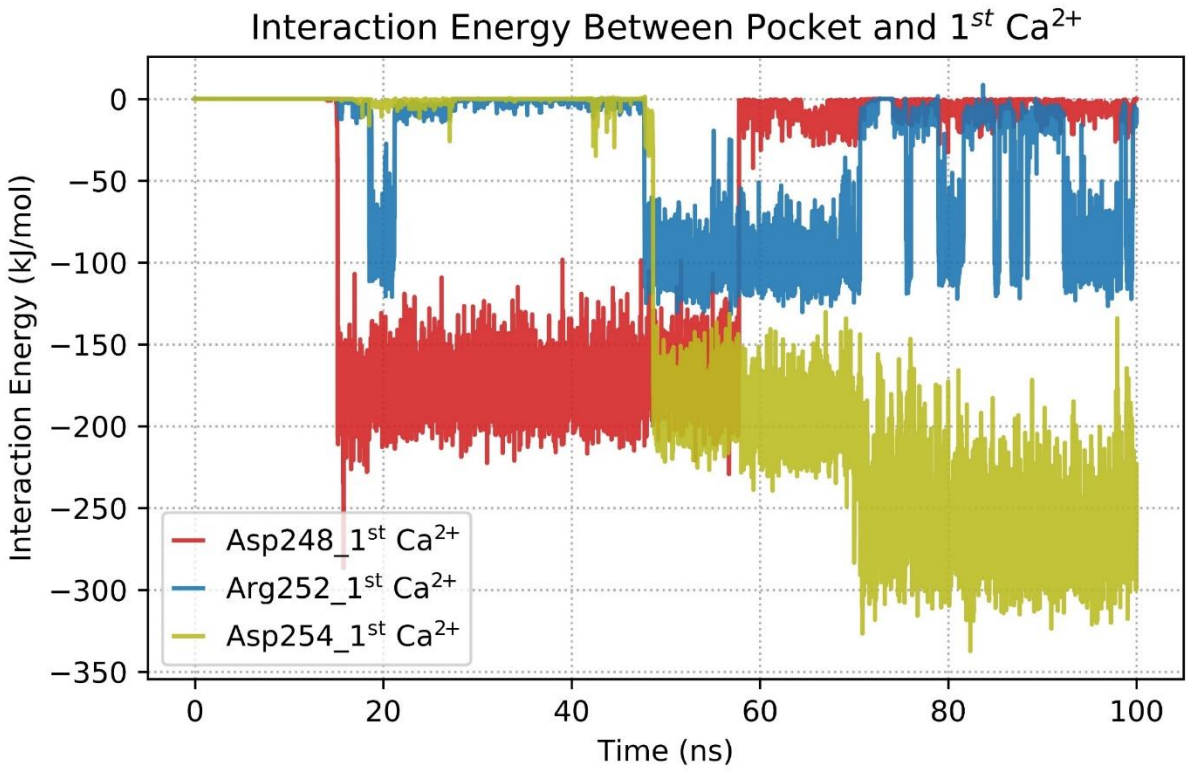


Figure 9. 10 Interaction energy between PKCα-C2 binding site residues and the first Ca^{2+} entering the site for extended simulation of PKCα-C2 in 150 mM $CaCl_2$.

Figure 9.10 shows the important interactions between the first Ca^{2+} and the residues in the binding site. The first interaction that is established occurs between Asp248 and the Ca^{2+} as shown in red in Figure 9.10. During the Ca^{2+} relocation at 50 ns the ion starts to interact with two other residues (Arg252 and Asp254) as shown in blue and yellow, in the Figure 9.10. The initial interaction with Asp248 is lost when the second Ca^{2+} enters the site (Figure 9.11).

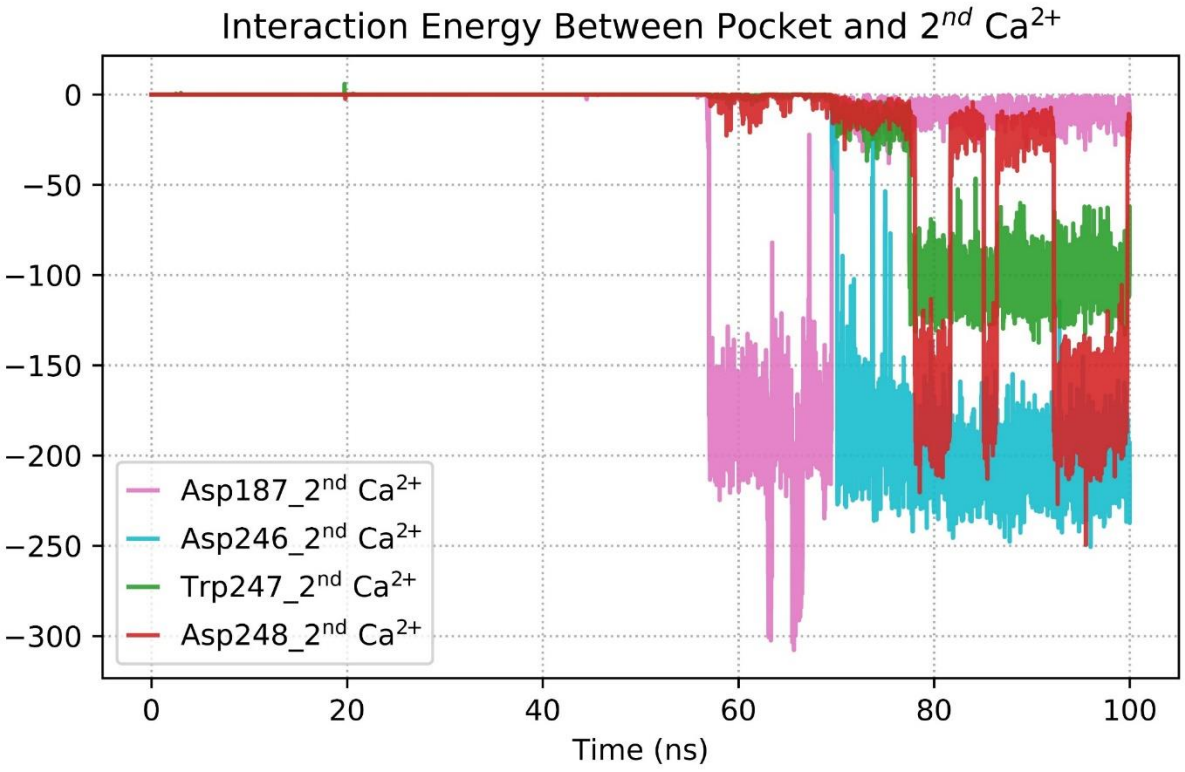


Figure 9. 11 Interaction energy between PKCα-C2 binding site residues and the second Ca^{2+} entering the site for extended simulation of PKCα-C2 in 150 mM $CaCl_2$.

Figure 9.11 shows the important interactions between the second Ca^{2+} and the binding site. When the second Ca^{2+} first entered, it mostly interacts with Asp187 which is lost at later stages of the simulation to Asp246. It is also important to note that Asp248 stimulates first Ca^{2+} binding during the early stages of simulations then starts switches to interacting with second Ca^{2+}, suggesting the two ions are competing for this interaction.

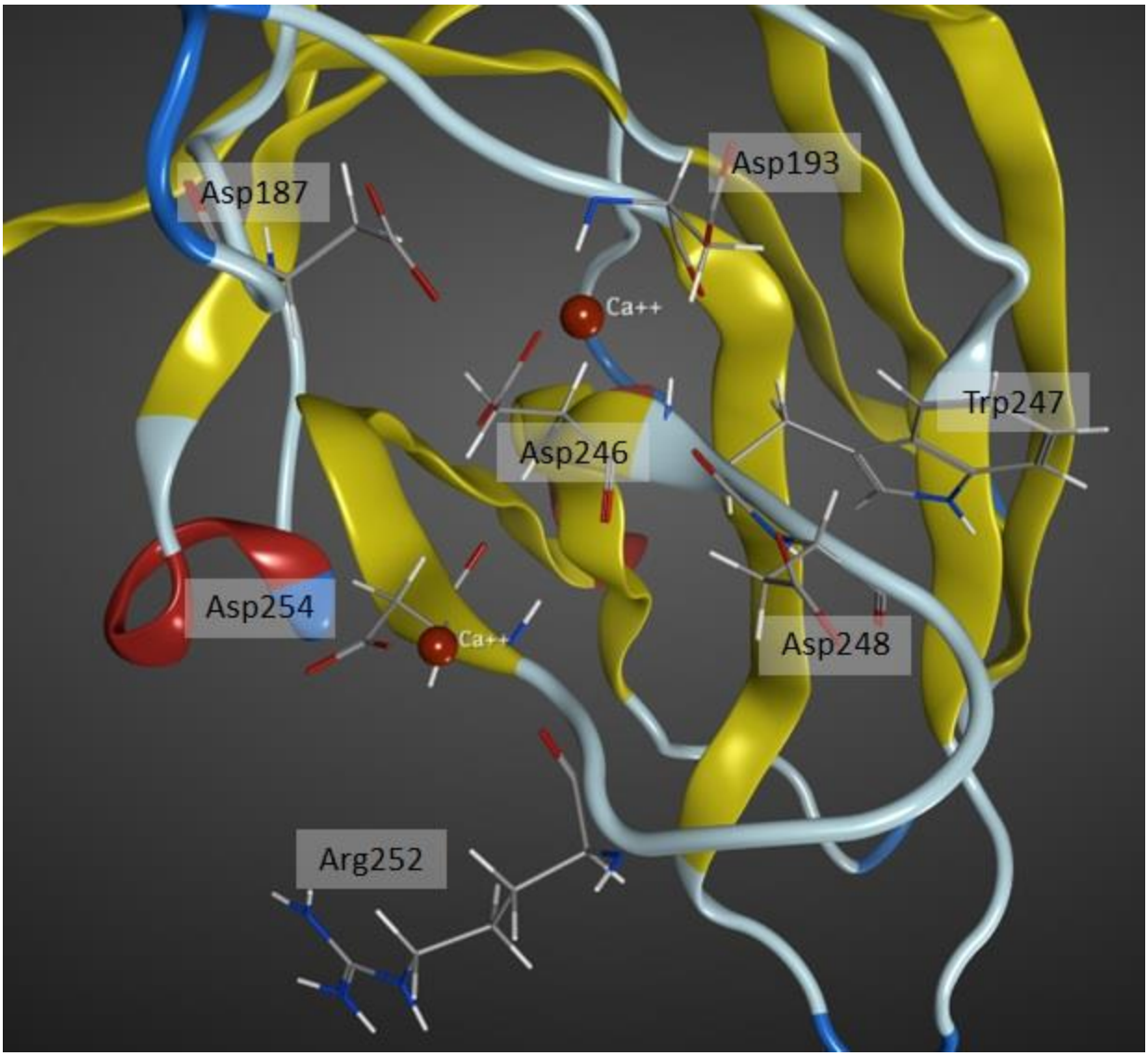


Figure 9. 12 Minimum energy frame of PKCα-C2 in 150mM $CaCl_2$. Two Ca^{2+} and the important residues are also shown.

The present study on highly conserved PKC family of enzymes suggests that the C2 domain has the most significant difference among different isoforms sequentially. The C2 domain might also embrace a solution for the target specificity problem that occurs in therapeutic applications targeting these enzymes. The binding site comparison of PKCα-C2 PKCδ-C2 shows that the two binding sites exhibit a very different environment especially on the site where the Ca^{2+} binding

occurs in the conventional PKCs. The Ca^{2+} binding site of PKCα takes an overall negative charge and five aspartic acid residues present in this site (Asp248, Asp254, Asp246, Asp193, Asp187, Figure 9.12) that are involved in the Ca^{2+} binding activation mechanism. Not having these residues might result a lack of Ca^{2+} regulation in PKCδ-C2. This is the first time that the PKC Ca^{2+} binding activation mechanism has been investigated using molecular dynamics.

REFERENCES

REFERENCES

(1) Garg, R.; Benedetti, L. G.; Abera, M. B.; Wang, H.; Abba, M.; Kazanietz, M. G. Protein Kinase C and Cancer: What We Know and What We Do Not. *Oncogene* **2014**, *33* (45), 5225–5237. https://doi.org/10.1038/onc.2013.524.

(2) Koivunen, J.; Aaltonen, V.; Peltonen, J. Protein Kinase C (PKC) Family in Cancer Progression. *Cancer Lett.* 2006, pp 1–10. https://doi.org/10.1016/j.canlet.2005.03.033.

(3) Mochly-Rosen, D.; Das, K.; Grimes, K. V. Protein Kinase C, an Elusive Therapeutic Target? *Nat. Rev. Drug Discov.* **2012**, *11* (12), 937–957. https://doi.org/10.1038/nrd3871.

(4) Steinberg, S. F. Structural Basis of Protein Kinase C Isoform Function. *Physiol. Rev.* **2008**, *88* (4), 1341–1378. https://doi.org/10.1152/physrev.00034.2007.

(5) Alwarawrah, M.; Wereszczynski, J. Investigation of the Effect of Bilayer Composition on PKC$α$-C2 Domain Docking Using Molecular Dynamics Simulations. *J. Phys. Chem. B* **2017**, *121* (1), 78–88. https://doi.org/10.1021/acs.jpcb.6b10188.

(6) Newton, A. C.; Antal, C. E.; Steinberg, S. F. Protein Kinase C Mechanisms That Contribute to Cardiac Remodelling. *Clin. Sci.* **2016**, *130* (17), 1499–1510. https://doi.org/10.1042/CS20160036.

(7) Newton, A. C. Protein Kinase C: Poised to Signal. *AJP Endocrinol. Metab.* **2010**, *298* (3), E395--E402. https://doi.org/10.1152/ajpendo.00477.2009.

(8) Ramsay, L.; Macaulay, M.; Degli Ivanissevich, S.; MacLean, K.; Cardle, L.; Fuller, J.; Edwards, K. J.; Tuvesson, S.; Morgante, M.; Massari, A.; Maestri, E.; Marmiroli, N.; Sjakste, T.; Ganal, M.; Powell, W.; Waugh, R. A Simple Sequence Repeat-Based Linkage Map of Barley. *Genetics* **2000**, *156* (4), 1997–2005. https://doi.org/10.1093/nar/25.17.3389.

(9) (TCELL) Joint Center For Structural Genomics (JCSG), P. F. T.-C. B. X-Ray Diffraction Data for the Crystal Structure of a C2 Domain of a Protein Kinase C Alpha (PRKCA) from Homo Sapiens at 1.90 A Resolution (4DNL). **2012**. https://doi.org/10.18430/M34DNL.

(10) Funabiki, H. Two Birds with One Stone - Dealing with Nuclear Transport and Spindle Assembly. *Cell* **2005**, *121* (2), 157–158. https://doi.org/10.1016/j.cell.2005.04.003.

(11) Chemical Computing Group Inc. Molecular Operating Environment (MOE). Montreal 2016.

(12) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001.

(13) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C.; Brook, S.; Brook, S.; Brook, S. Comparison of Multiple AMBER Force Fields and Development of Improved Protien Backbone Parameters. *Proteins* **2006**, *65* (3), 712–725. https://doi.org/10.1002/prot.21123.Comparison.

(14) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23* (3), 327–341. https://doi.org/10.1016/0021-9991(77)90098-5.

(15) Labute, P.; Santavy, M. SiteFinder-Locating Binding Sites in Protein Structures http://www.chempcomp.com/journal/sitefind.htm%5Cnhttps://www.chemcomp.com/journal/sitefind.htm.

(16) Anthony, W. J.; Bender, A.; Kaya, T.; Clemons, P. A. Alpha Shapes Applied to Molecular Shape Characterization Exhibit Novel Properties Compared to Established Shape Descriptors. *J. Chem. Inf. Model.* **2009**, *49* (10), 2231–2241. https://doi.org/10.1021/ci900190z.

(17) Sethian, J. Advancing Interfaces: Level Set and Fast Marching Methods. **1999**, 12.

CHAPTER TEN

Conclusions and Future Directions

Conclusions and Future Directions

With ever-evolving technological developments and advancement of computational modelling techniques, computational biochemistry can be used to study the dynamics of large systems. Proteins are dynamic systems in nature and their activity depends on their conformational states (i.e active/inactive), which may be affected by ligand binding. Understanding ligand binding phenomena, protein dynamics, and structural perturbations triggered by the binding is critical to understand biology. Computational modelling allows the study of these dynamics and simulation/analysis of ligand binding at a molecular level. For this dissertation, molecular dynamics, binding free energy calculations and bioinformatic tools were used to study binding and dynamics of a number of host-guest, protein-ligand and protein-ion systems.

Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) blind challenges provide a platform to validate/gauge current modelling techniques to predict physicochemical properties. In chapters 3 and 4 molecular dynamics and quantum dynamics were used to predict binding energies between the host-guest molecules. The results showed that MD followed by MMPBSA/MMGBSA calculations can be used to qualitatively rank binding energies of small molecules with low computational cost and memory, even though the predictions result in systematically higher binding energies than experiments. Due to its success, the MD simulations followed by MMPBSA/MMGBSA calculations can be applied to various applications where relative binding energies are important. To predict absolute binding energies without corrections based on similar systems, implementation of a better solute entropy model with respect to performance and accuracy on MMPBSA/MMGBSA solver should be considered in the future.

In Chapter 5, molecular modeling was used to study interactions between the Endo-A enzyme and glycans **39**, **41** which were synthesized and experimentally studied by Huang group. Experimentally, Endo-A enzyme shows substrate preference towards glycan **39**. The simulations showed significantly weaker binding of glycan **41** toward Endo-A which can explain the lack of glycosylation. In addition, the simulations also pointed out a mechanistic explanation on the Endo-A substrate preference: In all glycan **41**- Endo-A simulations, active site gate residues W244 and W216 are prohibited from closing, which can account for the reduced yield from glycosylation reaction by preventing the formation of the closed active site.

In another collaborative effort, the Huang group synthesized HS glycopeptide and HS glycan, discussed in chapter 6, by using total synthesis and experimentally measured FGF-2 dissociation constants and heparanase inhibition percentages of these compounds along with the peptide backbone. The experimental studies showed that only the glycan showed inhibitory activity against heparanase, while the glycopeptide showed a three-fold enhanced binding in comparison to the glycan binding to FGF-2. To understand different biological functions of glycan and glycopeptide on the FGF-2 and heparanase systems, molecular modeling was used. HS glycan, HS glycopeptide and peptide binding to the FGF-2 and heparanase enzymes were studied through molecular dynamics and free energy calculations. The simulations showed the peptide portion of the glycopeptide can lead to additional salt bridges in FGF-2 systems, whereas in heparanase, the glycopeptide tends to pull the glycan core towards solvent and loosen the hydrogen bonds. Both experiments and simulations showed that HS and HS proteoglycan can possess different biological functions. As highlighted through simulations depending on the target, peptide backbone can loosen binding of the core or result in additional interactions with the targets. For future directions, these interactions can be further analyzed by simulation of

mutant heparanase, and FGF-2 complexes bound to HS and HS proteoglycan to quantitively assess each interaction and its contribution to the binding.

In chapter 7 and 8, the MD/MMPBSA approach was used in SAMPL challenges to investigate the binding of per and poly fluoroalkyl substances (PFASs) to a number of human receptors. PFASs are emerging contaminants with a large and quickly growing chemical space. Human Pregnane X receptor (hPXR) and Peroxisome Proliferator Activated Receptor γ (PPARγ) are known targets for legacy PFASs with available toxicity data. However, toxicity of recently emerged PFAS alternatives is not measured for these systems. Molecular modeling was used to predict alternative PFASs' toxicity on hPXR and PPARγ and showed they still exhibit binding and may show toxicity. Additionally, long- and short-range interactions between the amino acids within the binding sites and PFASs were investigated to understand how PFAS is recognized on these receptors. The results outlined key residues that contribute strongly to the binding. The pioneer studies detailed in chapters in 7 and 8 show how biophysical tools can be a fundamental part of understanding PFASs at a molecular level, and guide scientist to find solutions for PFASs related environmental issues. The methodologies discussed can be further used to investigate other known PFAS targets and recently developed PFASs alternatives, which can also have damaging effects on the environment. Moreover, the molecular recognition patterns identified through models, can be used to develop environment friendly PFASs or PFAS inhibitors, such as L-carnitine has proven to be for PPARγ.

In chapter 9, Protein Kinase Cs (PKCs), a family of serine/threonine kinases, have been studied. The Ca^{2+} binding induced activation of conventional and novel PKCs were studied through bioinformatics and molecular dynamics simulations. The simulations displayed successive binding of multiple Ca^{2+} to the C2 domain of PKCα. Additionally, interaction

energies identified five aspartic acid residues important to attract and hold calcium ions in this domain. As shown with the sequence alignments and bioinformatic results the C2 domain of PKC is a promising drug target. Future investigations should include further understanding of the C2 domain's role on the remaining steps of PKC activation such as phosphatidylinositol 4,5-bisphosphate (PIP_2) binding to the C2 domain. With sufficient understanding of the activity and potential binding sites, the therapeutic view of this protein can be used to develop new therapies for several diseases that are known to be affected by active PKC levels, including multiple types of cancer, cardiovascular diseases, immune and inflammatory diseases, neurological and metabolic disorders.