INSIGHTS INTO HUMAN HEALTH THROUGH COMPUTATIONAL MODELING: TARGETING TUBERCULOSIS AND UNDERSTANDING PFAS TOXICITY

By

Semiha Kevser Bali

A DISSERTATION

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

Chemistry—Doctor of Philosophy

2024

ABSTRACT

Computational modeling approaches have been instrumental to biological and biochemical research for over 50 years, providing insight that can impact advances in human, animal, and environmental health. The use of in silico methods in drug discovery can reduce financial cost and time required for the development of new drugs by providing molecular-level insight and information about structure-activity relationships. Computational chemistry approaches can provide compound identification, optimization and screening towards the development of more potent and novel molecules. Computational biophysical methods are not only useful in drug discovery, but they are also useful in environmental science as well, in areas such as determining the toxicity of pollutants.

The validation of computational protocols is an important step in computational modeling. SAMPL blind challenges provide host-guest systems with known binding affinities and physical properties to benchmark developed methods and protocols against experiment. A benchmark for the protocols used throughout the dissertation has been provided.

Per- and polyfluoroalkyl substances (PFAS) are man-made molecules that have very interesting chemical features. These compounds are both water and oil-resistant therefore, they have been utilized in many industrial processes and household products, including fast food packaging, fire-fighting foams, dental floss, water-resistant garments, batteries, and non-stick cookware. However, research conducted in recent years has shown that some PFAS can cause significant health problems thyroid problems, cholesterol and lipid issues, and cancer in living organisms upon consistent exposure and bioaccumulation. Several of the chapters in this dissertation highlight investigations of three different protein targets of PFAS:

- human Peroxisome proliferator receptor gamma - retinoid x receptor alpha /DNA (PPARγ-RXRα/DNA) is an important protein for regulation of glucose metabolism and fat cell differentiation,
- human thyroglobulin protein (hTG) which is responsible for producing the thyroid hormones,
- rainbow trout estrogen receptor α and β (ERα, ERβ) controlling the reproduction.

The goal of each of the studies was to gain a molecular-level understanding of the impact of selected common and alternative PFAS on the proteins.

In considering specific disease, tuberculosis (TB) is a persistent disease largely observed in rural parts of the world. While there are available treatment regimens for both drug-susceptible as well as drug-resistant TB, these treatment protocols require use of many different drugs and take from six months to up to two years. Therefore, there is a need for the development of better treatment strategies. A chapter in this dissertation highlights how computational chemistry approaches have been used in studies to develop compounds targeting the treatment of Tuberculosis via two different protein targets, mycobacterium membrane protein large 3 (mmpL3) and DosS in collaboration with medicinal chemists. Homology modeling, binding energy estimations, as well as conformational dynamics of these proteins were investigated.

ACKNOWLEDGEMENTS

I would like to extend my thanks to my advisor, Dr. Angela K. Wilson, allowing me to be part of her group and engaging me in thrilling research projects. I also would like to thank my committee members Dr. Kenneth Merz, Dr. Katharine Hunt, and Dr. Edmund Ellsworth for their invaluable feedback and support. To the past and present members of the Wilson group: your friendship and stimulating discussions have been truly appreciated. I also would like to thank the Chemistry department members and staff for their assistance. A special thank you goes out to all my friends who have stood by me throughout this degree.

I am deeply appreciative of our collaborators, Dr. Edmund Ellsworth, Dr. Robert Abramovitch, and Dr. Xuefei Huang, along with their respective groups, as well as the computational and medicinal chemistry team at Reata Pharmaceuticals, and our experimental collaborators at MSU PFAS Center, for their invaluable insights and discussions.

I also would like to thank Dr. Viktorya Aviyente, for her continuous support and mentorship throughout my career. She is a great source of inspiration, and I consider myself fortunate to have her guidance.

Lastly, I will be forever grateful for:

- my husband, Hoa, you carried the second biggest burden of my Ph.D., and your presence and support made this possible. Thank you for always being there.
- our cats, Effie and Babycakes, you made the long Michigan winters more entertaining.
- my parents, your endless support and encouragement in my pursuit of science have been the foundation of my journey.
- myself, for the resilience, determination, and perseverance that have brought me to this moment. *Tôi là tôi*.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION TO MOLECULAR MODELING

During the past 50 years, computational modeling has become an important component of bio- chemical research. from drug discovery programs to toxicology studies. The improvements and the growth of computational capabilities now allow the study of large biological complexes. As the dynamics of the system is related to its function, the conformational changes observed in these biomolecules are crucial to understand processes such as ligand binding and cell signaling. Molecular dynamics simulations (MD) provide a useful route to study these phenomena in biological systems. With MD simulations, the conformational changes associated with protein activity and ligand binding can be investigated, and furthermore, the binding affinities can be predicted.

This dissertation covers various applications of computational modeling for different biological systems. In Chapter 3, a class of environmental pollutants (PFAS) and their toxic effects on PPARγ-RXRα/DNA complex is investigated. The allosteric mechanism in which PFAS can activate this complex was explained. In the fourth Chapter, PFAS binding to ERα and ERβ from rainbow trout was investigated to elucidate the impact of PFAS toxicity in aquatic species. The different binding modes of PFAS in ERα and ERβ were discovered and the important residues contributing to their binding affinities were investigated. Chapter 5 highlights how PFAS can interfere with the thyroid hormone synthesis by binding to hTG protein in humans. Due to the conformational rigidity caused by the PFAS binding, we showed that the specific types of PFAS can have more impact on the thyroid hormone synthesis.

The research shown in Chapters 6 and 7 highlights the efforts for the development against two different targets, mmpL3 and DosS, are shown. In addition to elucidating the binding modes and estimated binding affinities of these compounds, the activation mechanism of DosS protein was also investigated.

In Chapter 8, the utility of the modeling procedures used in these studies was evaluated using the host-guest dataset from the Statistical Assessment of Modeling of Proteins and Ligands9 (SAMPL9) challenge. The detailed protocols for estimating binding free energies and their statistical performance were evaluated against the experimental binding affinities.

CHAPTER 2

MOLECULAR MODELING: THEORY AND METHODOLOGY

2.1 Molecular Mechanics and the Concept of Force Field

Molecular Mechanics (MM) is an approach that uses classical mechanics and is based on Newtonian dynamics. In this method, the atoms are represented as spheres and the bonds are defined as strings. MM is suitable for systems that can be difficult to model using quantum mechanics due to its substantial computational cost - such as for systems with more than 1,000 atoms. For this reason, a classical approach is utilized to study proteins.

The combination of the parameters which are used for MM to describe a system and its total energy is called force field. A force field has two main components: bonded and non-bonded terms. Bonded terms are bond stretching, angle bending, torsion (bond rotation), while non-bonded interactions include electrostatic and van der Waals terms. These parameters are derived from empirical data to predict the particular parameters of the molecules.

Two of the bonded terms, the bond stretching and angle bending can be expresses using simple harmonic motion:

$$V_{bonded} = \sum_{bond} \frac{1}{2} k_{bond} (l_0 - l)^2 \tag{2.1}$$

$$V_{angle} = \sum_{angle} \frac{1}{2} k_{angle} (\theta_0 - \theta)^2 \tag{2.2}$$

where k_{bond} and k_{angle} are force constants for bonds and angles, respectively. l_0 and θ_0 are the reference values for the bond length and angle, respectively. The force needed to change the bond length between two bonded atoms is large, hence any significant changes in both bond lengths and bond angles are prevented.

The torsional term provides the largest contribution to the total energy, and is described as following:

$$V_{torsion} = \sum_{n=0}^{N} \frac{1}{2} V_n [1 + cos(n\omega - \gamma)] \tag{2.3}$$

The term V_n describes the depth of the potential energy surface of rotations about ω, over the periodicity of n, with the minimum angle of γ.

Improper torsions are used to define the out-of-plane bending motions, and the following equation shows its functional form:

$$V_{improper} = \frac{1}{2} k_{\omega} [1 - cos2\omega] \tag{2.4}$$

in which ω is used to define the angle between four atoms that are not bonded together in sequential order.

The charges on the atoms are considered as point charges, therefore, the electrostatic interactions can be defined using the Coulomb's Law:

$$V_{el} = \sum_{i} \sum_{j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{2.5}$$

where q_i and q_j define the point charges on the atoms, and r_{ij} is the distance between them.

There are different approaches used to calculate the point charges, including quantum mechanical methods and experimental approaches. AMBER force field, which is used in this study, uses the point charges derived from electrostatic potentials [1].

The final term, the van der Waals interaction energy, is defined as the sum of all interactions of the molecules while considering their positions as well as the relative orientations. Lennard-Jones function is the most widely used representations to define vdW interactions:

$$V_{vdw} = 4\epsilon_{AB} \left[\left(\frac{\sigma_{AB}}{r} \right)^{12} - \left(2\frac{\sigma_{AB}}{r} \right)^{6} \right] \tag{2.6}$$

where ϵ_{AB} is the amplitude of the potential, and σ_{AB} called as collision diameter, which is the arithmetic mean of individual diameters of pure species (σ_{A} and σ_{B}).

2.2 Molecular Dynamics

Molecular Dynamics (MD) uses the force field terms to calculate forces for each particle in the system, and with the help of statistical mechanical approach it models the dynamics of particles, hence gives a prediction for the position of the particles within a given system. This prediction is obtained by numerically solving Newton's equations of motion.

Classical Hamiltonian ($H(\mathbf{p}_i(t),\mathbf{r}_i(t))$) can be used to describe the time evaluated motion of a system with N particles:

$$H(\mathbf{p}_i(t), \mathbf{r}_i(t)) = \sum_{i=1}^{N} \frac{1}{2m_i}\mathbf{p}_i^2 + V(\mathbf{r}_i) \tag{2.7}$$

in this expression $\mathbf{p}_i(t)$ is the momentum vector and $V(\mathbf{r}_i)$ is the potential energy. The Hamiltonian is the sum of potential and kinetic energies of all particles in the system, and its partial derivation yields the equations of motion, i.e. velocity of the particle and the force acting on the particle.

$$\frac{\partial \mathbf{r}_i}{\partial t} = \frac{\partial H}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} = \mathbf{v}_i \tag{2.8}$$

$$\frac{\partial \mathbf{p}_i}{\partial t} = -\frac{\partial H}{\partial \mathbf{r}_i} = -\frac{\partial V}{\partial \mathbf{r}_i} = F \tag{2.9}$$

which leads to the Newton's second law:

$$\frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \frac{F}{m_i} \tag{2.10}$$

for a particle with mass m_i to move along coordinate $\mathbf{r}_i$ under the influence of an external force F. The last equation above is used to obtain the coordinates of the particles. When the position of a particle changes, the force acting on it also changes. $F(t)$, the total force at time t, is obtained by the vector sum of all interactions between the individual particles, and if the time step, ∂t, is small, it is assumed to be constant. There are many methods implemented in MD software to integrate the equations of motion, and *Velocity Verlet* is one of them[2]. In the *Velocity Verlet* algorithm, at time $(t + \partial t)$, the velocity of a particle (v) and its position (r) are defined as follows:

$$\mathbf{r}(t + \partial t) = \mathbf{r}(t) + \partial t \mathbf{v} t + \frac{1}{2}\partial t^2 m^{-1} F(t) \tag{2.11}$$

$$\mathbf{v}(t + \partial t) = \mathbf{v}(t) + \frac{1}{2} + \partial t m^{-1}(F(t) + F(t + \partial t)) \tag{2.12}$$

To initiate an MD simulation, an initial set of coordinates are required. These coordinates can be obtained from various sources. For instance, the coordinates of biological systems can be obtained via X-Ray crystallography studies or NMR structures. Maxwell-Boltzmann distribution is used to derive the initial velocities of the particles, v(0), and it is adjusted so that the total momentum would be zero for the whole system. Initial forces at t=0 are obtained using the Eqn. 9.

For a system that has all of the requirements (initial coordinates, r(0); initial velocities, v(0); initial forces, F(0)) the simulation cycle follows these steps:

1. Using Eqn.11, the displacement of coordinates are calculated with respect to the initial positions within a time interval ∂t.

2. With the help of Eqn.9, the forces on the particles are calculated using the positions from step(i).

3. Using Eqn.12, new velocities are calculated with the initial force and the new force that is obtained in step(ii).

4. Steps above are repeated until reaching a specified amount of simulation time.

According to *ergodic hypothesis*, within an infinite time of simulation, all possible states of a system can be obtained; however, an infinitely long simulation is impossible, and with the current computational limitations, there is a trade-off between long simulation time and the cost, which needs to be considered.[3]To be able to reach a sufficient simulation time without increasing the computational cost too much, the time step (∂t) should be taken as high as possible (usually (∂t is around 1fs to 2fs for classical MD simulations). However, ∂t is also limited by the integration algorithm that is used by the software.

When considering the MD simulations of biological systems, one should bear in mind that the proteins usually exist in a continuous solvent environment. However, the dimensions of the solvent boxes in MD simulations are limited. Therefore, to mimic the bulk effect *Periodic Boundary Conditions*(PBC) is employed. In PBC, the simulation box is copied infinitely in each direction, and only the coordinates of the original simulation box are followed throughout the simulation. If an atom in the original box moves to the outside of the box, its periodic image from the other box moves in the same direction, allowing the total number of particles in a box at a given time is constant always. In PBC, to treat the electrostatic interaction between particles, generally the Ewald sum method is employed[4].

Statistical ensembles are used to obtain the thermodynamic properties of the system of interest. These ensembles are constructed based on the temperature(T), volume(V), number of particles(N) and pressure(P). The most common ones are:

- NVE: microcanonical ensemble - has constant N, V and E;
- NVT: canonical ensemble - has constant N, V and T;
- NPT: isothermal isobaric ensemble - has constant N, P and T

2.3 Thermostats

For biological system simulations, the most popular choice is the NVT ensemble due to its computational efficiency. When using the canonical ensemble in simulations, the exchange of energy is contained with different thermostat models by adjusting the temperature to the desired value. The classical definition of average kinetic energy in NVT ensemble is follows:

$$< E_k >_{NVT} = \frac{1}{2} \sum_{i=1}^{N} m_i v_i^2 \tag{2.13}$$

and the average kinetic energy also can be rewritten such that it is related to the temperature using the classical equipartition theory:

$$< E_k >_{NVT} = \frac{3N - 6}{2} k_\beta T \tag{2.14}$$

in which k_β is the Boltzmann constant, and T corresponds to the temperature. Velocities of each step are scaled as $v_{new}=\lambda v_i$, then the temperature change ($\Delta T = Ti\text{-}T(t)$) can be calculated using the correlation between v_i and T from Eqn 13 and 14 as follows:

$$\Delta T = \frac{1}{2} \sum_{i=1}^{N} \frac{2}{3} \frac{m_i (\lambda v_i)^2}{N k_\beta} - \frac{1}{2} \sum_{i=1}^{N} \frac{2}{3} \frac{m_i v_i^2}{N k_\beta} \tag{2.15}$$

which gives

$$\Delta T = (\lambda^2 - 1) T(t) \tag{2.16}$$

To obtain the value of λ with respect to the target temperature T_{max}, and the instantaneous temperature $T(t)$:

$$\lambda = \sqrt{T_{new}/T(t)} \tag{2.17}$$

This is the procedure of the simplest thermostat[5]:, in which the T_{max} can be obtained by multiplying the velocity with λ and using the temperature obtained from the kinetic energy, *T (t)*. However, the drawback of this method is that along the simulation, the temperature difference between the solute and solvent may occur.

To solve the problems with the previous thermostat model, Andersen thermostat, which is based on the stochastic collision model, was developed[6]. In the Andersen thermostat, the system is immersed in a heat bath, and the velocity of the particles in a random time interval is assigned using Maxwell-Boltzmann distribution at temperature *T(t)*. At each step, the simulation is performed with constant energy so no thermal difference within the system occurs. In addition, the calculated velocities follow the Gaussian distribution.

Langevin thermostat is another model in which all particles obtain a random force at each time step, and their velocities are lowered by using a constant friction. To obey the "fluctuation-dissipation" theorem, average strength of the random forces and the friction are related. The equations of motion are modified as:

$$\frac{\partial \mathbf{p_i}}{\partial t} = -\frac{\partial H}{\partial \mathbf{q_i}} - \gamma \mathbf{p_i} + \sigma \epsilon_i \tag{2.18}$$

$$\sigma^2 = 2\gamma m_i kT \tag{2.19}$$

in which γ is used to create a damping force to the momenta, and σ and γ particles have a relation that is defined by the fluctuation-dissipation relation to recover the canonical ensemble distribution (Eqn. 19).

In the Langevin method, it is assumed that big particles (solute) exist in a pool of smaller particles (solvent), and the smaller particles usually randomly collide with the solute molecules and influence their dynamics. Moreover, solvent molecules also have dampening effect on the solute molecules described as a fictional drag force. Langevin thermostat incorporates these two factors.

2.4 Barostats

The isothermal-isobaric ensemble is also frequently used in MD simulations, and the methods that are used to control temperature can be adapted for pressure control as well. One method among

those is Berendsen barostat. The pressure tensor can be calculated as follows:

$$\mathbf{P} = \frac{2}{V}(\mathbf{E}_t kin - \Xi) \tag{2.20}$$

in which V is the box volume, E_{kin} is the kinetic energy and Ξ is the inner virial tensor, which is used to describe the behavior of diluted gases and it is described as follows:

$$\Xi = -\frac{1}{2}\sum_{i<j} \mathbf{r}_{ij}\mathbf{F}_{ij} \tag{2.21}$$

In the Berendsen barostat, the system is coupled weakly to an external bath. In the equations of motion, an extra term is needed for the pressure change:

$$\left(\frac{\partial p}{\partial t}\right)_{bath} = \frac{p_0 - p}{\tau_p} \tag{2.22}$$

where τ_p is the time constant for the coupling. The pressure change is proportional to the isothermal compressibility β:

$$\frac{\partial P}{\partial t} = -\frac{1}{\beta V}\frac{dV}{dt} = -\frac{3\alpha}{\beta} \tag{2.23}$$

And α can be calculated as:

$$\alpha = -\frac{\beta(p_0 - p)}{3\tau_p} \tag{2.24}$$

Hence, the equation of the motion is:

$$\frac{\partial r(t)}{\partial t} = v - \frac{\beta(p_0) - p}{3\tau_p} r \tag{2.25}$$

that represents the proportional scaling of coordinates.

2.5 Solvation Models

To be able to mimic the natural conditions of biological systems, simulation systems need to be immersed in suitable solvent environments. In MD simulations, the solvent is usually explicit water molecules that are defined by certain water models. The most common solvation model is the rigid 3-site TIP3P water model, in which Coulomb's law and Lennard-Jones potentials are used to describe the electrostatic interactions. On the other hand, 4-site TIP4P-EW model is found to be better at describing the bulk properties of water as maintaining the geometric parameters from TIP4P. In TIP4P-EW, the long-range interactions (Coulomb and LJ) are incorporated[7,8].

2.6 Homology Modeling

Homology modeling essentially targets building a three-dimensional structure for proteins by using the available structures of closely related proteins. Since not all proteins have their 3-D structures experimentally determined, being able to predict them successfully with in silico methods is extremely useful in drug discovery studies. There are currently many available tools for homology modeling, and each of them uses a different approach. One of the most successful one is I-TASSER by Zhang Lab[9–11]. The amino acid sequence is first matched with the sequence of available crystal structures in Protein Data Bank, producing fragments. Then, these fragments obtained from PDB templates are combined to form full-length structure models with Monte Carlo simulations, and the clustering is used to obtain a model. In the final step, this model is used to re-assemble the structures to obtain the final model with the lowest energy. Given the success of this approach in CASP (Critical Assessment of Techniques for Protein Structure Prediction) competitions, it was used to model *M. tuberculosis* MmpL3 protein structure and Estrogen receptors from rainbow trout (rERs).

2.7 Molecular Docking

Another approach that is commonly used in drug discovery research is molecular docking. It provides a relatively low-cost virtual screening for determining potential drug molecules. The docking approaches that are currently used can be classified into two main groups: (i) ligand-based, (ii) structure-based. While the former is used when there is no structural information regarding the target system/protein, latter is used when the 3-D structure of the protein is available. Ligand-based methods include pharmacophore modeling and QSAR (quantitative structure-activity relationship). Molecular docking falls into the structure-based docking, and there are many different open-source (i.e. AutoDock Vina) and commercial tools (i.e. MOE, Maestro) available. A docking process generally follows a two-step approach. First, the ligand conformation as well as the orientation and position are determined, then, the binding affinity for a specific orientation (pose) is calculated after further refinement. The success of docking depends on the scoring algorithms and the sampling methods used in these steps. In addition, the docking methodologies include rigid docking where

the ligand/protein is treated as rigid body, and induced fit docking in which the flexibility of ligand/protein are taken into account.

MOE (Molecular Operating Environment) is a commercial software that can be used to study small molecules and proteins. The docking suite of MOE is capable of performing induced fit docking with a user-friendly GUI. The algorithm used for ligand placement is Triangle Matcher which uses alpha-spheres to define the binding site[12]. The ligand is positioned so that the triplets of ligand atoms are superimposed on alpha spheres, and if there is a clash with protein atoms, that pose is removed. The scoring function for placement step is called London dG that includes the terms for ligand flexibility, hydrogen bonds and desolvation[13,14]. In Eq. 26, E_{flex} corresponds to the estimated ligand entropy; c, c_{hb} and c_m are constants trained on more than 400 proteins. f_{hb} and f_m used to account for the geometric imperfections for ligand-protein and metal-ligand interactions, respectively. And the last term indicates the approximated desolvation energy. After the placement step, specified number of poses are refined for final ranking. For the refinement step, force-field based GBVI/WSA dG (Generalized-Born Volume Integral/Weighted Surface area) scoring function is used (Eq.27)[14]. GBVI/WSA dG scoring function was trained using MMFF94x and AMBER99 force-fields on 99 protein-ligand complexes training set. α and β are the constants that were determined during the training, E_{sol} is the solvated electrostatic term, and $SA_{weighted}$ corresponds to the weighted solvent-accessible area scaled with β[14].

$$\Delta G_{LdG} = c + E_{flex} + \sum_{h-bonds} c_{hb} f_{hb} + \sum_{metal-lig} c_m f_m + \sum_{atomsi} \Delta D_i \tag{2.26}$$

$$\Delta G_{GBVI} = c + \alpha \left[\frac{2}{3} (\Delta E_{sol} + \Delta E_{coul}) + \Delta E_{vdw} + \beta \Delta SA_{weighted} \right] \tag{2.27}$$

2.8 Free Energy Calculations

Free energy drives all molecular processes, including the ligand-protein interactions, folding, and chemical reactions. Therefore, having in silico methods that describe the free energies as accurately as possible is crucial. In drug design, the binding free energy is considered as an indication of the binding strength between the ligand and protein. In a real system where ligand

binds to protein under NPT condition, the free energy is expressed as:

$$F = -k_b T ln Z \tag{2.28}$$

where Z is the partition function, and it is given by:

$$Z = \frac{1}{V_0 N! h^{3N}} \int exp\left(-\frac{PV}{k_b T}\right) dV \int \int \left(-\frac{H(p,r)}{k_b T}\right) dpdr \tag{2.29}$$

In the equation above, the Hamiltonian is the total energy of the system for particular position and momentum.

The computational free energy calculation methods such as thermodynamic integration (TI), free energy perturbation (FEP), and molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) and generalized Born surface are (MM-GBSA) are widely used in virtual screening and lead optimization steps during drug development[15]. The first two methods, TI and FEP, are called pathway methods, and they obtain the free energy by converting the system from an initial state to final state via very small changes of the energy function[15]. This yields an accurate result; however, those methods are also computationally expensive, and sometimes the convergence can be an issue as well. On the other hand, MM-GBSA/PBSA methods, also called end-point methods, are based on the sampling of the final states therefore they are less expensive[16,17]. MM-GBSA/PBSA has been used for the analysis of docking poses, estimation of binding affinities. In addition, it can also help analyzing the individual residue contributions or different energy terms[15,18]. However, there can be issues with this method. One major source of error for MM-PBSA/GBSA is the lack of a conformational entropy term. This term can be computed using an additional calculation called normal-mode analysis, which is computationally costly.

2.9 Constant pH Molecular Dynamics Simulations

As is known, the environment in which the protein exists significantly impacts the protonation states of the residues and hence, the activity of the protein. While classical MD methods assume only a single protonation state for a given residue, there are cases where more than one protonation states may be possible or need to be considered. With a method developed by Mongan et al. in 2004, that is currently available in Amber20, an implicit solvent (generalized Born) can be used to

perform along with periodic Monte Carlo (MC) sampling for the protonation states.[19] At each of the MC step, the protonation state for a given residue is randomly chosen, and then, the free energy associated with the transition to deprotonation or protonation is calculated:

$$dG = k_B T(pH - pK_{a,ref})ln10 + dG_{elec} - dG_{elec,ref} \tag{2.30}$$

where pH is the solvent pH that is specified, $pK_{a,ref}$ is the pK_a of the reference residue, dG_{elec} is the electrostatic portion of the calculated free energy for the residue, and lastly, $dG_{elec,ref}$ is the electrostatic portion of the calculated free energy for the reference residue. The non-electrostatic component of this equation includes all free energy components but the GB electrostatics, with an assumption that it would be very similar to the value obtained independently from electrostatic environment. The electrostatic component of the equation is calculated using the difference between the current and proposed protonation state in a single step as an implicit solvent is being used. The dG is used as a criterion whether to accept the transition or reject it. If accepted, the simulation will continue with the new protonation state, and if rejected, the protonation state will not be changed.

2.10 Steered Molecular Dynamics

Steered molecular dynamics (SMD) creates changes in coordinates within a given specific time by applying an external force onto the system[20]. The way it is implemented in Amber20 uses a constant velocity. It is an approach that is similar to the umbrella sampling, with a difference in which the center of the restraint is now time-dependent:

$$V_{rest}(t) = (1/2)k[x - x_0(t)]^2 \tag{2.31}$$

Here, x can be any quantity such as distance, angle or torsion. The generalized work can be computed by integrating the force over time, which then can be used to compute the free energy differences with Jarzynski equation[21]. If we have two states A and B, and their generalized coordinates differ in x:

$$exp(-\Delta G/k_B T) = \langle exp(-W/k_B T)\rangle_A \tag{2.32}$$

This indicates that the computing the work between states A and B, and averaging over the initial state (A), the equilibrium free energies can be calculated from the non-equilibrium calculations[22,23].

One method to apply forces to a system is to apply a harmonic restraint and shift it in a specific direction. If we assume a generalized reaction coordinate x again:

$$U = K(x - x_0)^2/2 \tag{2.33}$$

where K is the force constant determining how "stiff" the restraint is, and X_0 is the initial position of the restraint moving at a constant speed v. Then, the external force on the system can be expressed as:

$$F = K(x_0 + vt - x) \tag{2.34}$$

BIBLIOGRAPHY

[1] et al Case, D., J. Berryman, R. B. (2018). AMBER18.

[2] Verlet, L. (1967). Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review*, 159(1):98–103.

[3] Genheden, S. and Ryde, U. (2012). Will molecular dynamics simulations of proteins ever reach equilibrium? *Physical Chemistry Chemical Physics*, 14(24):8662–8677.

[4] Cerutti, D. S. and Case, D. A. (2010). Multi-level ewald: A hybrid multigrid/fast fourier transform approach to the electrostatic particle-mesh problem. *Journal of Chemical Theory and Computation*, 6(2):443–458.

[5] Woodcock, L. V. (1971). Isothermal molecular dynamics calculations for liquid salts. *Chemical Physics Letters*, 10(3):257–261.

[6] Andersen, H. C. (1980). Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of Chemical Physics*, 72(4):2384–2393.

[7] Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *Journal of Chemical Physics*, 120(20):9665–9678.

[8] Horn, H. W., Swope, W. C., and Pitera, J. W. (2005). Characterization of the TIP4P-Ew water model: Vapor pressure and boiling point. *Journal of Chemical Physics*, 123(19):194504.

[9] Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction.

[10] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, 12(1):7–8.

[11] Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738.

[12] Edelsbrunner, H. (1992). Weighted alpha shapes. Technical report, Technical paper of theDepartment of Computer Science of the University of Illinois atUrbana-Champaign, Urbana, Illinois.

[13] Corbeil, C. R., Williams, C. I., and Labute, P. (2012). Variability in docking success rates due to dataset preparation.

[14] Labute, P. (2008). The generalized born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. *Journal of Computational Chemistry*, 29(10):1693–1698.

[15] Wang, E., Sun, H., Wang, J., Wang, Z., Liu, H., Zhang, J. Z., and Hou, T. (2019). End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design.

[16] Kollman, P. A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O., Cieplak, P., Srinivasan, J., Case, D. A., and Cheatham, T. E. (2000). Calculating structures and free energies of complex molecules: Combining molecular mechanics and continuum models. *Accounts of Chemical Research*, 33(12):889–897.

[17] Massova, I. and Kollman, P. A. (2000). Combined molecular mechanical and continuum solvent approach (MM- PBSA/GBSA) to predict ligand binding.

[18] Hou, T., Wang, J., Li, Y., and Wang, W. (2011). Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *Journal of Chemical Information and Modeling*, 51(1):69–82.

[19] Mongan, J., Case, D. A., and McCammon, J. A. (2004). Constant ph molecular dynamics in generalized born implicit solvent. *Journal of computational chemistry*, 25:2038–2048.

[20] Hummer, G. and Szabo, A. (2001). Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 98:3658–3661.

[21] Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78:2690.

[22] Izrailev, S., Stepaniants, S., Isralewitz, B., Kosztin, D., Lu, H., Molnar, F., Wriggers, W., and Schulten, K. (1999). Steered molecular dynamics. pages 39–65.

[23] Jensen, M., Park, S., Tajkhorshid, E., and Schulten, K. (2002). Energetics of glycerol conduction through aquaglyceroporin glpf. *Proceedings of the National Academy of Sciences of the United States of America*, 99:6731–6736.

CHAPTER 3

BINDING OF PER-AND POLYFLUOROALKYL SUBSTANCES (PFAS) TO THE PPAR/RXRA–DNA COMPLEX

About this chapter: This chapter is reprinted from Almeida NMS; Bali, SK; James, D; Wang, C; Wilson, AK, J. Chem. Inf. Model. 2023, 63, 23, 7423–7443. with permission of the American Chemical Society. Only results pertaining to PPARγ presented here.

3.1 Introduction

Per- and polyfluoroalkyl substances (PFAS) are a group of chemical compounds used as fluoropolymers, which have industrial applications ranging from coatings, adhesives, firefighting foam, to oil repellants, due to their high heat resistance.[1–3] PFAS are considered "forever chemicals" due to their resistance to degradation[4,5] and their persistence in the environment (soil and water)[6,7], and in humans and animals (see e.g. Ref.[8,9]). To illustrate their prevalence, in the U.S., it is estimated that PFAS can be found in more than 99% of the population.[10] The persistence can be attributed at least in part to the strength of the carbon fluorine bond, one of nature's strongest bonds.

The number of PFAS compounds is quite large. The U.S. Environmental Protection Agency (EPA) has more than 14,000 PFAS listed in the PFASTRUCT database as of March 2023, and in a recent study, over 140 compounds were shown to be potentially harmful in in vitro assays.[EPA,12] Overall, the compounds can be classified into two groups: legacy and emerging PFAS compounds. The most common legacy PFAS are perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate acid (PFOS). These are two of the earliest known PFAS to be produced on a large scale.[1] The emerging PFAS, which are usually created to offer "better alternatives" to replace legacy PFAS, must be well understood, not only at the molecular, but also at a mechanistic level. However, recent investigations have linked not only legacy PFAS, but also emerging PFAS, with effects on the environment and living organisms.[7,8,13–15]

Nuclear receptors (NRs) are a superfamily of ligand-activated transcription factors and have been the focus of many drug discovery programs. One of the most studied NRs is peroxisome proliferator activator gamma (PPARγ) duetoitsroleinglucosemetabolism,regulation of adipogenesis, and lipid metabolism.[16–18] Even though the biological relevance of PPARγ in its homodimer form has been discussed in the literature[19–23], the known biologically relevant form of PPARγ that controls gene transcription is the heterodimer form, i.e, peroxisome proliferator-activator gamma/retinoic

X receptor (PPARγ/RXRα).[16,22,24]The PPARγ and RXRα proteins, similar to other NR proteins, consist of three main domains: a ligand-binding domain (LBD), a DNA-binding domain (DBD), and a hinge domain that connects the DNA binding domain (DBD) and the ligand binding domain (LBD). The ligand molecules can bind to the LBD, causing the reorientation of the Helix-12 and consequently aiding in the recruitment of coactivators.[20] Before undergoing ligand binding, PPARγ complexes with its corepressor peptides. Upon dimerization with RXRα, the ligands can bind to the LBD and initiate the dissociation of the corepressor by promoting a conformational change. Then, when coactivators are recruited, the transcriptional activities of the dimer can occur.[25]The DBD includes two 4-cysteine (Cys4) zinc-finger motifs that are vital to the sequence-specific binding to DNA.[26,27] The DBD domain, as well as the zinc-finger domain, are present among other heterodimers complexed with DNA.[28] It has also found that there are bridging water molecules that facilitate the interactions between DNA bases and DBD residues.[27] The solvent accessibility of the zinc-finger domains may enable the structure and the activity of zinc-finger domains to be impacted by the presence of compounds, such as PFAS in solvents.[29]The interaction of zinc-finger domains with various metals that cause metal toxicities, or with small molecules that are used for cancer treatments, including cisplatin, have been investigated using experimental and computational approaches.[29–36]Quantum chemistry calculations have shown that the zinc cation can assist in deprotonating cysteines in the zinc-fingers, and this deprotonation is also thought to play an important role to keep the zinc-finger domain in the functional folded conformation.[36]However, the stability of the zinc-finger domains is system-dependent and protein backbone motions can stabilize, or destabilize, the cysteine cores. To the best of our knowledge, there is little or no insight about how PFAS molecules can affect/interact with the zinc-finger, or how PFAS can stabilize/destabilize the interaction of zinc-finger domains with DNA.

To investigate nuclear receptors, it is important to understand how their activity can be affected by structural and conformational shifts that occur upon ligand binding. To fully understand the effect of PFAS on the PPARγ/RXRα complex, the structural route towards activation/inactivation, and the interactions between the two proteins as well as with the DNA need to be investigated.

The binding of agonist compounds to RXRα-LBD monomer can trigger structural motions for activating the receptor.[37] However, for the PPARγ/RXRα heterodimer, the activation of the RXRα nuclear receptor can also activate PPARγ, regardless of the occupation of the PPARγ-LBD.[17,38–40] It is known that the PPARγ-LBD can be activated via hydrogen bonding to Tyr473. Usually, this mechanism can occur for full agonists, which interact with the activation function 2 (AF-2) region and Helix-12.[23,41] Partial agonists have been postulated to have an activation mechanism through water bridging, i.e, they do not directly interact with Helix-12, and their transcriptional activity may not be entirely structural, or connected to movements of Helix-12.[42–48] For the PPARγ/RXRα heterodimer, allosteric pathways also have been found for their activation, which can elucidate ligand-dependent transcription factors.[49,50]

Several studies have linked the effects of PFAS in humans to several types of toxicities (i.e., hepa, neuro, reproductive, immuno, and cardiovascular) and thyroid disruption, among other health issues.[51–58] Recently, PPAR proteins have been investigated for potential binding to PFAS, which carries nefarious outcomes. For example, PPARγ has been shown to be affected by PFOS, leading to renal fibrosis.[59–61] Liu et al. reported that PFOA and PFOS exposure can cause long lasting effects on uremic patients.[62] The LBD of PPARγ was also investigated in vitro, and insight was gained about how 16 PFAS bind to this receptor.[63] Among the 16 PFAS compounds that were investigated, several of them bind to PPARγ, which results in activation of PPARγ. The aforementioned study reports the maximum inhibition concentrations, or IC_{50}s, obtained through in vitro experiments. The authors found that the size of the carbon chain and functional groups had a clear influence on how strongly PFAS bind to PPARγ.[63] More recently, Khazee et al. calculated dissociation constants (Kd) of short chain PFAS, and it was also the first time Kd for short chain PFAS were reported in sub-micromolar concentration.[64]

Interestingly, an investigation by Chou et al. showed that L-carnitine is able to attenuate the effects of PFOS on PPARγ via Sirt1 mechanisms.[59] L-carnitine is a molecule that is absorbed from diet and also is synthesized in the brain, kidney, and liver. It can also be easily purchased commercially.[65] In previous investigations, L-carnitine has been reported to decrease the level of

apoptosis in kidney cells through a PPARγ-dependent mechanism.[59,60,66]

Experimentally, there is little information about how PFAS bind to RXRs. Heuval et al. reported that RXRα can be activated by PFAS in mice.[67] In the same study, only mild activation was found for PPARγ. More recently, it was shown that PFAS can bind to RXRβ and target a particular agonistic bioactivity of this receptor.[12] Although there are many in vitro experimental studies on PPARγ, there is not much known about how PFAS interacts with PPARγ/RXRα complex and how this interaction can affect DNA binding, on a molecular level. In one of the first computational studies on PPARγ, the authors reported binding sites and binding energies for PFOA and PFOS.[68] Other efforts have focused on how PFAS bind to different human and animal proteins, and the prediction of binding pockets and poses.[9,69–72] Zhang et al., performed molecular docking simulations for PFAS on PPARγ. The authors showed that Tyr473, His323 and His449 were important residues towards binding in the PPARγ binding pocket.[63] Li et al. reported PPARβ/δ activities and performed docking studies for both receptors.[73] In a more recent study, Behr et al., concluded that PPARα could be activated by a range of PFAS. However, PPARγ was only activated by perfluoro-2-methyl-3-oxahexanoic acid and 3H-perfluoro-3-(3-methoxypropoxy) propanoic acid. In one of our recent studies, the interactions of 27 PFAS and L-carnitine with PPARγ, and the roles of the acidic and basic residues in two binding pockets were investigated. A new binding pocket (dimer pocket) was postulated for the PPARγ homodimer structure. L-carnitine was shown to have the potential to bioaccumulate in the dimer pocket as well as similar binding to most of the studied PFAS. The acid/base and residue decomposition indicated that interactions with PPARγ were more favorable towards L-carnitine than the PFAS, indicating that L-carnitine can competitively replace PFAS from both of the investigated binding pockets.

Because of the large number of PFAS compounds, recently, machine learning (ML) approaches have also been utilized to predict the binding between PFAS and nuclear receptors.[74] One of the most recent approaches considered the binding of 4,464 PFAS to PPARα and γ, and the thyroid hormone receptor.[75] The authors concluded that the binding energies of PFAS to thyroid hormone receptors are 2-3 kcal mol^{-1} stronger than to PPARγ. As well, a machine learning strategy was

utilized to identify novel PFAS compounds that may be less toxic than current PFAS such as GenX.[74]

Herein, a variety of PFAS were investigated to consider their effect on the activity of the PPARγ/RXRα-DNA complex. In total, nine PFAS with different chain lengths and functional groups were selected, along with L-carnitine. The characteristics of the PFAS and the specific species include those with: (a) one sulfonic group, perfluorooctane sulfonic acid (PFOS); (b) an amino group, perfluorooctane sulfonamido (PFOSA), and (c) acidic groups, PFOA and PFHxDA. For alternative PFAS, 2,3,3,3-tetrafluoro-2-heptafluoropropoxy propanoic acid (GenX) and 4,8-dioxa-3H-perfluorononanoic acid (ADONA) were considered. Furthermore, the alcohol and carboxylic acid fluorotelomers investigated herein were 6:2 fluorotelomer alcohol (6:2 FTOH) and 6:2 fluorotelomer sulfonic acid (6:2 FTSA). In addition, the PFAS which showed the largest binding affinity in our previous study, 2-(N-Ethyl-perfluorooctane sulfonamido) acetic acid, Et-PFOSAAcOH, was also included in this investigation.[2] Molecular dynamics simulations and binding free energy calculations have been fundamental approaches to study small molecule-protein interactions.[76–78] Molecular docking approaches, along with molecular dynamics simulations were used to investigate the effects of selected compounds on the PPARγ/RXRα-DNA complex. Further binding analysis using the Poisson - Boltzmann surface area (MM-PBSA) and molecular mechanics with a modified general born solvation model (MM-GBSA) methodologies were used to assess the binding strength of selected PFAS for the PPARγ and RXRα ligand binding domains. Structural changes upon PFAS binding were investigated as well. For the RXRα DNA binding domain located near the PPARγ-LBD, quantum mechanical calculations, using several different density functional approaches and DLPNO-CCSD(T) calculations were performed, providing a more robust assessment of the binding trends of the pocket.

3.2 Computational Protocols

3.2.1 System Preparation

The DNA-bound PPARγ/RXRα (PPARγ/RXRα-DNA) structure was obtained from the RSCB Protein Data Bank (PDB ID: 3DZU).[24] Before the docking procedure, the protein and DNA

structures were prepared with the Molecular Operating Environment (MOE) software[79] using the Protonate 3D at the physiological pH.[80,81] All solvent molecules, ions, and co-activator peptides were removed from the structure. For the apo simulation, the co-activator peptide (NCOA2) was included in the simulation, due to the lack of ligands in the binding pockets, and to maintain the stability of the secondary protein structure. With these modifications, the overall structure does not change from its activated state, nor does it significantly change the secondary structure of the heterodimer. To identify the possible binding pockets, MOE's "site finder" algorithm was employed.[82] Three different binding pockets were considered for docking and molecular dynamic (MD) simulations: PPARγ -LBD for Pocket 1, RXRα-LBD for Pocket 2, and one of the DBD for Pocket 3 (Figure S1). The protonation states of the different PFAS and L-carnitine were obtained under physiological conditions (pH=7, 300K and 1atm). For this step, the Protonate 3D module was utilized.81 For the generation of poses, the London Δ G scoring function was employed obtaining 100 initial placements.[83] The GBVI/WSA Δ G scoring function, with the induced fit protein method was utilized to refine the final ten poses. The poses with different functional group orientations and with the highest scoring functions were selected for molecular dynamics (MD) simulations. For Pocket 1, a pharmacophore approach was utilized, which features a hydrogen bond to Tyr473, consistent with PPARγ's activation.[41]

3.2.2 Molecular Docking Protocol and Pose Selection

For the PPARγ-LBD binding pocket, or Pocket 1, Tyr473 plays an important role in PFAS binding and towards the activation of the PPARγ protein.[63] A pharmacophore approach was used to place the functional groups of PFAS molecules near the -OH on the side chain of Tyr473 residue, and for each PFAS compound, two poses that are distinct from one another were selected for further analysis. The selected poses differ from each other by the orientation of their tail ends. In Figure S1-S2, the binding pocket locations, as well as the binding orientations of selected poses are shown. The docking scores of the selected poses are also reported in Table S1. Overall, the binding orientations of the selected poses are classified into two different categories based on the orientation of the tail end of PFAS, either pointing towards Tyr473 or in the opposite direction

towards Tyr473. For the RXRα-LBD, Pocket 2, there is no experimental evidence suggesting the importance of any residue interaction with the PFAS compounds, hence the docking was performed without a pharmacophore model. Instead, the Triangle Matcher algorithm was used at the pose generation step and the GBVI/WSA ΔG scoring function was used for ranking the poses.[80,83] Most of the poses obtained with this approach had PFAS head groups oriented towards Arg316. For each PFAS, two distinct poses were selected for further investigation (Figure S3). Regarding Pocket 3, two distinct poses were selected for molecular dynamics simulations. The binding pocket in the DBD was identified by site finder in MOE, and this pocket location is shown in Figure S1. This binding pocket is interesting as it is at the interface of the PPARγ-LBD and the RXRα-DBD. Due to the difference in charge between Zn^{2+} and the PFAS head groups, there is a strong electrostatic interaction which makes the binding possible. (It was later found that without a strong electrostatic interaction between an atom from the functional groups of the PFAS, and the zinc ion, the pose was not stable and moved away from the pocket.)

3.2.3 Molecular Dynamics Simulations and Binding Free Energy Calculations

To prepare the PFAS and L-carnitine for the MD simulations, restrained electrostatic potential charges (RESP) were calculated with the RED server.[84,85] For each compound, a short MD simulation was performed to sample conformations at 350 K with a 4-fs time step. The trajectories were clustered and the top three representative frames were used in the calculation of the partial charges using the RESP method. The simulation box for each complex was generated by using the leap module as featured in Amber Tools.[86] For the simulations, ff14SB, OL15, and gaff2 were used for the protein, DNA, and small molecules (i.e., PFAS), respectively.[87,88] Each system was neutralized in accordance with Joung and Cheatham parameters in 0.1M NaCl. In addition, the TIP4P-Ew water model was considered for all simulations.[89–91] Due to the presence of the zinc-finger motif in the DBD, a Leonard-Jones 12-6-4 potential was used to describe the Zn-Cys4 interactions.[92–94] In addition, the cysteines coordinated to Zn^{2+} were deprotonated. In total, ~130,000 water molecules were added for each PPARγ/RXRα-DNA and PFAS system.[92]

For the minimization, a series of harmonic potentials were selected (100.0, 50.0, 10.0 and 0.0

kcal mol^{-1} Å-2), which restrain all atoms with the exception of the water molecules and ions. Then, the PPARγ/RXRα-DNA and PFAS complex was heated from 0 K to 300 K in a stepwise manner. The systems were gradually heated with restraints that were released in gradually. After the heating step, a 500 ps equilibrium simulation was performed with a time step of 1 fs. For the production run, a 75 ns long MD simulation was performed for PPARγ/RXRα-DNA. For each PFAS and L-carnitine bound to Pockets 1 and 2, two poses were considered to sample various conformation of the ligands. In addition, for a given compound, the values for two poses were averaged for residue decomposition, binding free energy, and hydrogen bonding analyses. For pocket 3 (DBD), the one pose obtained from docking was submitted to optimization with DFT, and all of them converged to a minimum on the potential energy surface, with real frequencies. Furthermore, per the SI, after ~10 ns of every simulation, the root mean square deviations (RMSD) plateaued, which indicate the stabilization of the protein structure. For this study, 75 ns is enough to provide all the information needed to analyze the binding of different PFAS and L-carnitine to PPARγ/RXRα-DNA. A 1 fs time step was considered for all simulations, and 1000 frames per nanosecond were written out to disk. This frame collection allows for a large sampling of trajectories and an in-depth hydrogen bonding analysis. The SHAKE algorithm[93] was utilized for covalent bonds with hydrogen atoms. The particle-mesh Ewald approach was utilized to approximate long-range electrostatic interactions. The molecular dynamic simulations were performed with AMBER 2020 using the pmemd module with CUDA.[86]

Binding free binding energy calculations were performed for Pocket 1 and Pocket 2. These calculations were carried out using molecular mechanics with a Poisson - Boltzmann surface area (MM-PBSA) and molecular mechanics with a modified general born solvation model (MM-GBSA) as implemented in the Amber 2020/AmberTools21.[94,95,86] In prior work on a single NR (PPARγ) and 27 different PFAS, the applicability of MM-GBSA and MM-PBSA was demonstrated, and, thus, the approaches have been utilized for the current study. Since an energetic assessment of PFAS binding strengths is investigated, both MM-GBSA and MM-PBSA yield a relatively fast assessment of different points of the simulation, providing useful data sampling. While methods

such as free energy perturbation (FEP) or thermodynamic integration (TI) may provide an even more useful assessment, these methods are too demanding for the present study, due to the numbers and sizes of the systems.

To achieve a better sampling of the results, 7500 frames (equally spaced) from 75 ns long trajectories of the MM-GBSA and MM-GBSA simulations for each PFAS were selected for the binding free energy calculations. The simulations with the highest binding affinities for each PFAS were averaged. The residue decomposition analysis for each binding pocket was performed for the amino acids within ~10 Åof the PFAS.86 Moreover, root-mean-square-distance (RMSD), root-mean-square fluctuations (RMSF), residue decomposition analysis, and hydrogen bond analysis were performed with the CPPTRAJ module as implemented in AmberTools21 using the default settings.[96] All data were plotted using the matplotlib module and the figures were obtained using UCSF Chimera and MOE.[97,79]

3.2.4 QM-cluster Approach for Pocket 3

Due to the presence of metallic atoms in the DNA binding domain (DBD), an alternative to MM-PBSA and MM-GBSA methodologies is needed. To calculate the binding energies of PFAS coordinating to zinc in this pocket, a quantum mechanics-clustering (QM-clustering) approach was used for stable poses. Such an approach has been useful in the study of metal-protein coordination as well as enzymatic reactions, as shown in prior studies.[81,98–100] Because of the size of the system, density functional theory (DFT) approaches can provide useful insight while maintaining an affordable computational cost. However, a far more expansive ab initio method, the domain-based local pair natural orbital (DLPNO) form of coupled cluster single, double, and perturbative triple excitation (CCSD(T)) method (DLPNO-CCSD(T)) was also employed to predict binding energies.[83]

For the QM calculations, the investigated binding site consists of Zn-Cys4 along with the residues near the bound PFAS compounds (Figure 1). For the clustering of MD simulations, a hierarchical agglomerative algorithm with an epsilon value of 3.0 was chosen, and for each PFAS, ten clusters were calculated r. Each tenth frame of 75000 frames was considered for clustering

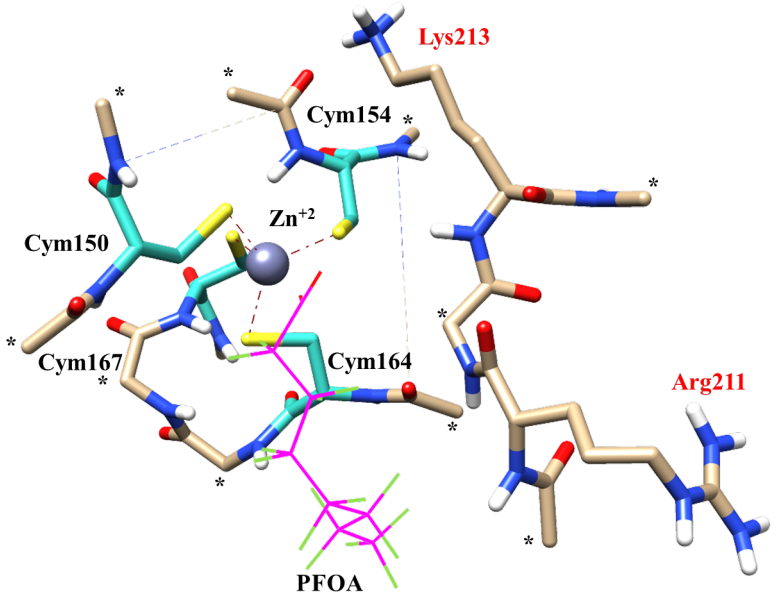


Figure 3.1 Example of the DBD binding site with the Zn-Cym4 motif and nearby residues with capped backbones (shown with an asterisk) and PFOA, which was utilized for the binding energy calculations. This structure was optimized with B3LYP-D3BJ/def2-SV(P) in a PCM water environment). The Cym residues are from PPARγ-DBD, the residues Arg211 to Lys213 are from the RXRα protein. Atoms with asterisks ($*$) are fixed in their original positions.

the trajectories, and the cpptraj module was used as implemented in AmberTools21.[101,96] For the selected PFAS (PFOS, PFOA, Et-PFOSAAcOH, 6:2 FTSA, GenX, and ADONA), the first rendered cluster was prevalent for almost 100% of the simulations, therefore this first cluster was selected for each of the DFT calculations. For the geometry optimization step, the dispersion corrected density functional, Becke, 3-parameter, Lee–Yang–Parr, B3LYP-D3(GD3BJ) in conjunction with def2-SV(P) basis sets were utilized.[102–106] In prior studies, the B3LYP-D3(GD3BJ) approach with the def2-SV(P) basis set has resulted in valid equilibrium structures for structures of this size.110 This basis set was also used previously for protein-ligand interactions.[106,107]The complex, the protein, and PFAS were each optimized separately. The corresponding structures are provided in Table S2. To simulate water solvation within a biological environment, the implicit-solvent polarizable continuum model (PCM) including non-electrostatic contributions (solute-solvent dispersion, solute-solvent repulsion, and solute cavitation) was considered.[108–111] To calculate binding

energies (Be), single point calculations were performed based on Equation (1), where Ecomplex, Eprotein, and EPFAS correspond to the energies of the complex, protein, and PFAS, respectively. The complex, the protein, and PFAS were each optimized separately. The corresponding structures are provided in Table S2. To simulate water solvation within a biological environment, the implicit-solvent polarizable continuum model (PCM) including non-electrostatic contributions (solute-solvent dispersion, solute-solvent repulsion, and solute cavitation) was considered.[108–111] To calculate binding energies (Be), single point calculations were performed based on Equation (1), where Ecomplex, Eprotein, and EPFAS correspond to the energies of the complex, protein, and PFAS, respectively:

$$B_e = E_{complex} - E_{protein} - E_{PFAS} \tag{3.1}$$

For the binding energy calculations (Equation (1)), B3LYP-D3BJ/def2-SV(P) and B3LYP-D3BJ/def2-TZVPP calculations were performed, incorporating PCM. To provide a comparison to B3LYP-D3BJ, the Minnesota 15 (MN15) functional was also considered. MN15 is known to be useful for noncovalent interactions and includes some level of parametrization for transition metals.[112] This functional was also partnered with the def2-TZVPP basis sets and the PCM implicit solvation model.

To probe the effect of the more electronegative atoms on the binding energies, the def2-TZVPPD basis sets, which include additional diffuse functions, were also employed for the B3LYP-D3BJ calculations. In addition, the SMD implicit model was utilized for comparison with PCM.[113] To probe the effect of the more electronegative atoms on the binding energies, the def2-TZVPPD basis sets, which include additional diffuse functions, was also employed for the B3LYP-D3BJ calculations. In addition, the SMD implicit model was utilized for comparison with PCM.119 To better account for electron correlation, DLPNO-CCSD(T) was considered, though at a triple-ζ basis set level (with def2-TZVP(-f)), due to its computational cost.79 For the DLPNO-CCSD(T) calculations, two implicit solvation environments were considered: the conductor-like polarizable continuum model (C-PCM) and SMD.[114–116] For these calculations, ORCA 5.0.3 was utilized.[117,115] Finally,

as energy convergence with respect to basis set is often not reached until the quadruple-ζ level for DFT methods for transition metals,123 B3LYP-D3BJ with PCM and SMD were utilized with def2-QZVPP for hydrogen, carbon, and zinc, and def2-QZVPPD for N, O, F and S to calculate the binding energies. To simplify the notation of the DFT calculations, D3BJ will be omitted when referring to a DFT functional. To simplify the calculations, the protein backbone in the complexes was replaced by -CH3 and -CH2 groups, reducing the size of the model systems (as shown in Figure 1). The selected residues coordinating to Zn cation were truncated from the Cα of the adjacent residues to preserve the peptide bonds.[99,117–119] For Cym164 and Cym167, only the side chains of the residues between them were removed and the peptide backbone was kept intact. The substituted functional groups were frozen during the constrained geometry optimizations. Namely, the positions of two types of atoms were fixed: (a) the external -C(α)H3 group and (b) the -C(α)H2 groups between two connected cysteines. All DFT calculations were performed using the Gaussian 16 software package, revision C01.

3.3 Results and Discussion

3.3.1 Structural convergence and fluctuations

The structural movement of proteins allows for correlation analysis between different structures, allowing for the analysis between the protein with no ligands bound (apo) and the protein bound with (a) its co-crystallized ligands, (b) PFAS, and (c) L-carnitine. Assessing the differences between the structures allows for insight into the movement of the secondary structure of proteins which is important, because the movement influences the activation, or inactivation of the protein. To check that structural convergence of the PPARγ/RXRα-DNA complex was achieved throughout the simulation, the RMSD was monitored for convergence. As 6:2 FTOH did not have a stable docking pose in Pocket 1, it was not simulated, or included in the analysis for this pocket. For Pocket 2, all of the PFAS poses remained in the pocket. For Pocket 3, L-carnitine was not stable in the pocket, and, hence, it was not included in the analysis. A structural comparison among PFAS-bound PPARγ/RXRα-DNA, co-crystallized ligands (2-[(2,4-dicholorobenzoyl) amino]-5-(pyrimidin-2-yloxy) benzoic acid for PPARγ, and (9cis)-retinoic acid for RXRα) bound complexes, and the apo

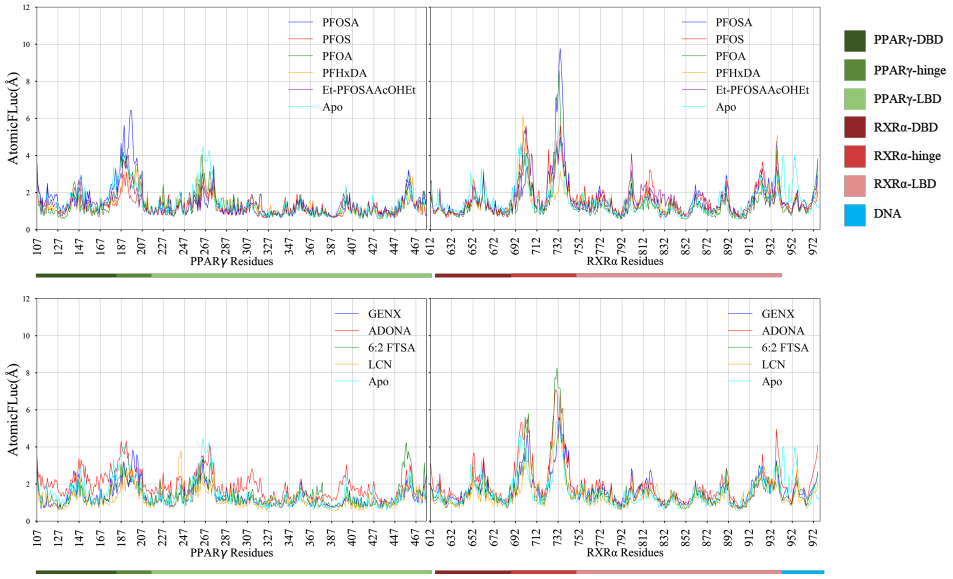


Figure 3.2 RMSF plot for all protein residues and DNA, for nine different PFAS and L-carnitine (LCN) for Pocket 1. The values are calculated for 75 ns MD simulations in Pocket 1.

structure was done.

3.3.1.1 Pocket 1 residue fluctuations and stability

The time-series RMSD plots for Pocket 1 (PFAS and L-carnitine) are shown in Figures S4-S13. Overall, all protein RMSDs converged within the 75 ns simulation time. When the LBDs, DBDs, and hinge domains are compared for both PPARγ and RXRα proteins separately, the hinge domains resulted in the largest overall RMSD. A large RMSD for the hinge domains is expected due to the lack of a definitive secondary structure. Furthermore, in all of the simulations, the DBDs of both proteins resulted in the lowest RMSDs in comparison to LDBs. On the other hand, LBDs resulted in a variety of conformational changes throughout the simulation time. The low RMSDs observed for the DBDs may indicate that strong interactions with DNA stabilize the domain movements.

For the majority of the simulations, the PFAS remained stable in the pocket (i.e. small RMSD); however, there were poses in which the PFAS was observed to change conformations within Pocket 1. In Figure S14, S15, S16, S17, S18 the small conformation changes of PFOA, PFHxDA, 6:2 FTSA, and Et-PFOSAAcOH are shown, respectively. Other PFAS including PFOS, PFOSA, GenX, and ADONA did not have any significant conformation changes and their RMSDs were stable throughout the simulation time.

With the presence of PFAS in Pocket 1, the RMSD time-series plots showed that the DNA oligomer reaches a stable RMS distance early in the simulations, with the exception of the PFOS-bound complex. The apo simulation (Figure S19) also has a stable RMSD for the first 75 ns of the simulation with an average RMSD of ~3 Å. In the presence of co-crystallized ligands (Figure S20), the RMSD of DNA is ~2 Åuntil 50 ns, and there is an increase observed after that. In all of the simulations, PFAS in Pocket 1 led to a very stable PPARγ-DBD; however, the hinge domain and PPARγ-LBD showed differences depending on which PFAS is bound to the binding pocket. Both the apo complex and simulation with co-crystallized ligands result in a very stable PPARγ-DBD which could indicate that the stability of PPARγ-DBD may not be directly influenced by the ligand binding to PPARγ-LBD, within the time frame considered. The presence of co-crystallized ligands led to more stable and lower RMSD hinge and LBD domains overall, with the highest RMSD being ~2 Å. On the contrary, the apo system displayed a high RMSD for the hinge (~ 3Å), while the PPARγ-LDB domain was ~2 Å. The L-carnitine compound also had various conformational changes throughout the trajectory; however, these conformational changes were relatively small and did not result in large motions of LCN.

In all of the PFAS simulations, the PPARγ hinge domain resulted in the largest RMSD within the PPARγ protein, with an average value of ~3 Å. The PPARγ-LDB domain, similarly, shows very little deviation in RMSDs and is quite stable in all of the Pocket 1 simulations. The RXRα protein had a very stable DBD in all simulations with a PFAS bound to Pocket 1, whereas the hinge and RXRα LBD had different convergence times. Both the apo complex and system with co-crystallized ligands have a stable RXRα-DBD domain with RMSD less than 1 Å. The presence of co-crystallized ligands, however, reduced the RMSD of the RXRα-LDB domain to ~1.5 Å, on average. The RMSF plots of Pocket 1 shown in Figure 2 illustrate the impact of the binding of PFAS on the overall protein and DNA motions. The apo simulation as well as the PFAS indicate a general trend in which the hinge domains always have a high fluctuation while the DBD and LBD domains have less. This observation is in parallel with the RMSD analysis, where the RMSD of the hinge domain was the largest among all investigated domains. An RMSF value between 4 to 8

Åwas observed for the RXRα hinge domain, which is higher than what is observed for the PPARγ hinge loop. Another region with high RMSF was the Ω loop on PPARγ (residues Lys261-Glu276). The Ω loop of PPARγ showed the largest fluctuations in apo simulations, and the presence of co-crystallized ligands was observed to lower the RMSF of the Ω loop. Of the PFAS, ADONA resulted in the highest RMSF for the Ω loop, and the PFOSA had the lowest. The Ω loop is thought to be important for the allosteric activation mechanism of PPARγ and affecting the conformation of H12 helix. 131 While these observations were also present in the RMSF plot for L-carnitine, in general, the RMSF values are lower than other simulated systems and are comparable to apo and co-crystallized ligand-bound systems. This would indicate that binding of a small compound like L-carnitine did not structurally affect the complex.

3.3.1.2 Pocket 3 residue fluctuations and stability

RMSD and RMSF plots for the investigated DBD pocket are shown in Figures S21-S29 and Figure 3, respectively. For all PFAS that coordinated to the zinc finger and stayed complexed with four cysteines and a zinc ion have a very stable RMSD (PFOSA and 6:2 FTOH did not show coordination to zinc). These poses are quite stable and did not shift away from the pocket. Throughout the 75 ns simulations, the RMSD of PPARγ was smaller than that of RXRα, when PFAS are bound to the DBD. This outcome is also consistent with observations made for Pocket 1 and Pocket 2. In addition, all complexes converged during the simulation. The hinge regions of both PPARγ and RXRα resulted in the largest RMSD values when PFAS are bound to the DBD (Pocket 3). For RMSFs, the hinge regions of both PPARγ-LBD and RXRα-LBD also had the highest RMSF values and were affected by a range of PFAS in this pocket. All PFAS, except PFOSA and 6:2 FTOH, do not coordinate to the zinc ion, and their constant movement within the pocket altered the RMSF at the hinge regions for both proteins. This is an important outcome, because the hinges connect the DNA binding domains to RXRα and PPARγ LBDs. Having higher fluctuations in the hinge domains affects the communication of the nuclear receptors and DNA.49 Herein, the area which showed the highest fluctuations when PFAS are bound to the DBD is the RXRα region, namely residues ranging from Glu233-Asp273. For example, the most extreme case

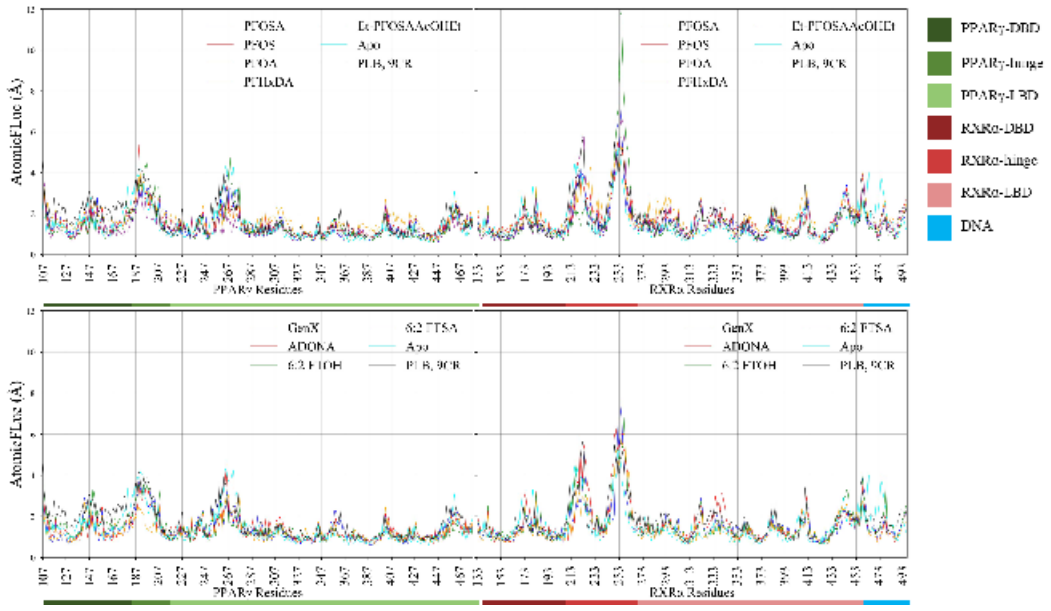


Figure 3.3 RMSF plot for all protein residues and DNA bases. The values are determined for 75ns MD simulations in the DBD pocket near the zinc finger domain.

of fluctuations for this region occurred for PFOA, with RMSF close to 12 Å, versus an average value of 7 Åfor other PFAS. The apo and the protein structures with native ligands had RMSFs of ~5 Å, i.e., showing less fluctuations/movement.

3.3.2 Binding free energy calculations

In order to understand the binding strengths of PFAS in each pocket, MM-GBSA/PBSA methodology was employed. Previously, this approach has shown good agreement with the experimental IC50 values for PFAS bound to Pocket 1.

3.3.2.1 PPARγ ligand binding pocket – Pocket 1

The MM-PBSA/GBSA methodologies were used to calculate the binding energies of the investigated compounds in the binding pockets. In Figure 4, the average binding energies of the compounds in Pocket 1 are shown. The MM-PBSA results resulted in a ranking of the compounds as follows (from highest binding energy to lowest): Et-PFOSAAcOH and PFHxDA (~-44 kcal mol^{-1}); PFOS (~-31 kcal mol^{-1}); PFOSA (~-25 kcal mol^{-1}); 6:2 FTSA, GenX and ADONA (~-22 kcal mol^{-1}); PFOA (~-20 kcal mol^{-1}); L-carnitine (~-15 kcal mol^{-1}). Among those, PFOSA, Et-PFOSAAcOH, and PFOS have eight perfluorinated carbons, PFOA has six perfluronated car-

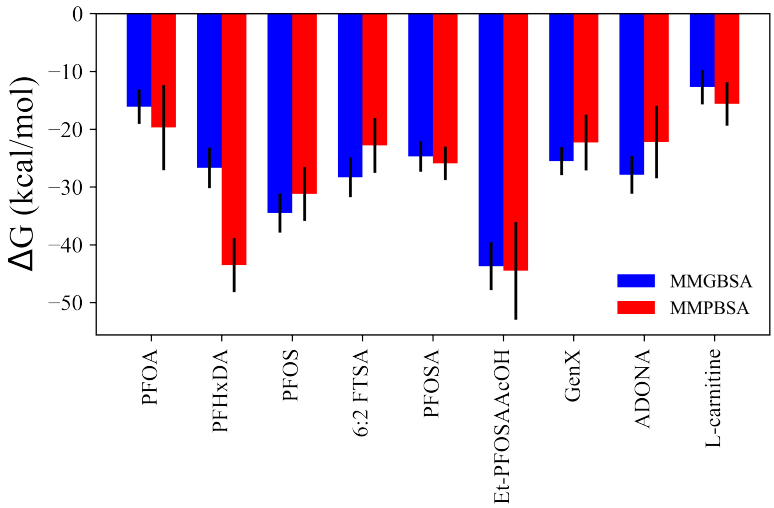


Figure 3.4 Average MM-PBSA/GBSA binding energies for Pocket 1. The binding energies were averaged over a 75 ns long MD simulation.

bons, while 6:2 FTSA and ADONA have six perfluorinated carbons. The comparison between the MM-PBSA binding affinities and the chain length of the perfluorinated carbons indicates that the longer chain PFAS binds more strongly than the shorter chain PFAS. PFHxDA (16 carbons) and Et-PFOSAAcOH (12 carbons) are the strongest binders, while PFOA and L-carnitine are the weakest. The alternative PFAS (GenX and ADONA) have binding strengths that are comparable to PFOSA and 6:2 FTSA.

3.3.2.2 DNA binding pocket – Pocket 3

The DBD binding energies calculated with DFT methodologies are included in Table 1. Of the ten MD simulations performed for Pocket 3, only seven PFAS stayed within the pocket. 6:2 FTOH and PFOSA moved out of the pocket, and L-carnitine travelled into the solvent, so it was not considered further towards analysis. For the seven PFAS which remained stable within the binding domain, the binding energies were calculated as per Equation 1. The cartesian coordinates of the final optimized structures are included in the SI (Table S2). When comparing the final optimized structures to the highest populated cluster from MD simulations, only PFOA and Et-PFOSAAcOH maintained coordination to the zinc ion. For these two structures, a five-ligand coordinated structure was formed with four cysteines and the zinc ion. For the PFOA structure, zinc's covalent bond length to the deprotonated cysteines (Cym) increased from ~2.2 Åto ~2.3-2.4 Å, and PFOA moved

from a distance of ~2.2 Åto 2.86 Å. PFOA was still coordinated to the zinc dication, even though it was repelled from the first coordination sphere. PFOA was also stabilized by interactions with Lys213 and Asp214 from the RXRα protein. Even though PFOA coordinated with the zinc and the four deprotonated cysteines, its binding energy was still positive for all DFT functionals and basis set combinations, with the exception of def2-SV(P). Et-PFOSAAcOH formed a hydrogen bond with Lys213, which maintained the PFAS coordination to the zinc dication. The distance between sulfonate oxygen to zinc is 2.53 Åand, the zinc-Cym4 bond length averages were ~2.4 Å. The binding energy for this complex was positive for all methodologies when considering triple-ζ and quadruple-ζ basis sets. For PFOA and Et-PFOSAAcOH, it is shown later (section 3.4.3) in the residue decomposition that most of the residues around these two PFAS contributed positively (repel). However, both of these PFAS formed stabilizing electrostatic interactions with the Zn^{2+} ion, through the negatively charged head group oxygens. Even though none of the other five PFAS maintained coordination to the zinc dication, electrostatic interactions with the Lys213, Asp214, Gly212, Arg211, and Gln210 residues allowed for these PFAS to stay in the pocket. For example, PFOS formed a strong hydrogen bond with Gln216 and Arg147. With a similar size to PFOS, 6:2 FTSA had the same orientation as PFOS within the pocket. In addition, it also bonded to Gln216 through hydrogen bonding, but not with Arg147. However, 6:2 FTSA forms a hydrogen bond with a cysteine (Cym162) coordinated to Zn^{2+}. The 6:2 FTSA structure has two carbons with four hydrogens, which allows for a hydrogen bond donation to this negatively charge cysteine. For these two PFAS, all of the DFT methods in Table 1 predicted a negative binding energy, for all but the prediction for PFOS utilizing SMD and the triple-ζ basis set. In addition, DLPNO-CCSD(T) predicted a positive binding energy for PFOS in both a C-PCM and a SMD environment. From a triple- to quadruple-ζ basis set, B3LYP-D3BJ/PCM dropped 0.9 kcal mol^{-1} in binding energy, though the energy was still negative. However, B3LYP/SMD predicted a positive binding energy. For 6:2 FTSA with DLPNO-CCSD(T) in a SMD and C-PCM environment, the binding energy was -1.3 kcal mol^{-1} and 0.4 kcal mol^{-1}, respectively. For the quadruple-ζ calculation in both solvation environments, slight binding was still predicted.

Table 3.1 Binding energies were calculated for the DNA binding pocket (DBD) using a range of DFT functionals and basis sets with PCM and SMD implicit solvation models. At the triple-ζ level, DLPNO-CCSD(T)/def2-TZVP(-f) was utilized with C-PCM and SMD. The geometry of each PFAS was optimized at the B3LYP-D3BJ/def2-SV(P) level, utilizing the PCM implicit solvation model. Units are in kcal mol^{-1}.

PFAS	B3LYP-D3BJ PCM[a]	B3LYP-D3BJ PCM[b]	MN15 PCM[b]	B3LYP-D3BJ PCM[c]	B3LYP-D3BJ SMD[c]	DLPNO-CCSD(T) C-PCM[d]	DLPNO-CCSD(T) SMD[d]	B3LYP-D3BJ PCM[e]	B3LYP-D3BJ SMD[e]
6:2 FTSA	-29.9	-4.6	-3.6	-2.2	-2.0	0.4	-1.3	-0.9	-0.6
ADONA	-38.6	-14.3	-8.2	-11.9	-6.2	-3.4	-3.0	-10.9	-5.2
PFHxDA	-30.4	-7.6	-4.6	-5.6	-6.6	-3.5	-5.7	-4.8	-5.7
EtPFOSAAcOH	-6.8	6.9	5.9	8.3	5.3	3.6	5.1	9.1	6.0
GenX	-12.8	4.3	3.4	6.0	3.5	5.1	1.9	6.8	4.4
PFOA	-7.4	9.3	6.4	11.0	2.5	2.8	-0.5	12.1	3.6
PFOS	-38.6	-12.7	-5.6	-10.9	0.5	2.1	2.5	-10.0	1.5

a-def2-SV(P)
b-def2-TZVPP
c-def2-TZVPP+def2-TZVPPD for N,O,F,S
d-def2-TZVP(-f)
e-def2-QZVPP+def2-QZVPPD for N,O,F,S

The largest PFAS studied, PFHxDA, formed a hydrogen bond interaction with Gly212 and Gln210, keeping this PFAS compound in the pocket. In addition, Cym162 interacted directly with the oxygen from the carboxylic acid functional group of PFHxDA. From the double- to triple-ζ basis set, the PCM binding energy predictions resulted a drop of 2 kcal mol^{-1}. The SMD solvation model and MN15-PCM at the triple-ζ level basis set level predicted negative binding energies, demonstrating affinity towards this pocket. DLPNO-CCSD(T) predicted a negative binding energy for this PFAS with each of the implicit solvation methods investigated. Negative binding energies were also predicted with B3LYP and quadruple-ζ basis sets. For the two alternative PFAS investigated, GenX and ADONA, two different poses were identified within the pocket. GenX was not as close to Gly212 and Gln210 as ADONA, so it did not interact as strongly with these residues. It was also not close enough to the zinc dication in order to coordinate to it or interact with any of the deprotonated cysteines. Furthermore, due to the loss of these two important interactions, the binding energies at triple- and quadruple-ζ levels are positive. On the other hand, ADONA was far more stable and interacted favorably with Gly212, Gln210, and Cym162, as demonstrated by its negative binding energy. B3LYP-PCM, using a combination of quadruple-ζ basis sets rendered the largest binding energy for the complexes at -10.9 kcal mol^{-1}. For DLPNO-CCSD(T), the binding

energy was -3.4 and 3.0 kcal mol^{-1} for C-PCM and SMD predictions. DLPNO is a powerful method for binding energy predictions but can only be paired with a smaller basis set, due to its computational cost. Even though explicit solvation is not possible to utilize due to the system size, implicit solvation is crucial to obtain valid and meaningful binding energies. In addition, DFT allows a great balance between computational cost and accuracy, estimating binding energies up to quadruple-ζ level. Regarding the different DFT methods, it should be noted that the basis set used for the geometry optimization step is not appropriate for the energetics (def2-SV(P)). B3LYP at a quadruple-ζ level is our most robust functional/basis set combination. When directly compared to DLPNO-CCSD(T) at a triple-ζ level, with different solvation methods, both methods demonstrate that PFHxDA and ADONA bind to this pocket. However, only B3LYP-D3BJ/PCM shows affinity for PFOS, but not B3LYP-D3BJ/PCM or DLPNO-CCSD(T)/SMD or C-CPCM.

Even though the binding energies for the DBD are less negative than for the other two binding pockets considered, PFAS can still bind in the DBD. One of the reasons for the lower binding energies relies on the fact that the zinc finger domain is coordinated by four deprotonated cysteines, however the docking algorithm places all the PFAS in the first coordination shell of the cysteines. Since the Zn^{2+} has a full 3d shell, it does not want to accept another ligand. After optimizing the geometries with DFT for the different fragments separately, most of the PFAS move further into the pocket, or stay in the second coordination shell of the Zn^{2+} atom.

3.3.3 Residue interactions and hydrogen bonding

Residue interaction analyses provide insight about how the binding pocket residues interact with PFAS and about the strengths of these interactions. Together with the hydrogen bonding patterns, these analyses provide insight into the role of different residues in the stabilization of PFAS in the investigated binding pockets.

3.3.3.1 PPARγ ligand binding domain – Pocket 1

The interaction patterns of PFAS with the surrounding residues in Pocket 1 provides critical insight about the binding patterns of these compounds. The interaction energies of the different PFAS versus L-carnitine in Pocket 1 with each residue were averaged and are plotted in Figure

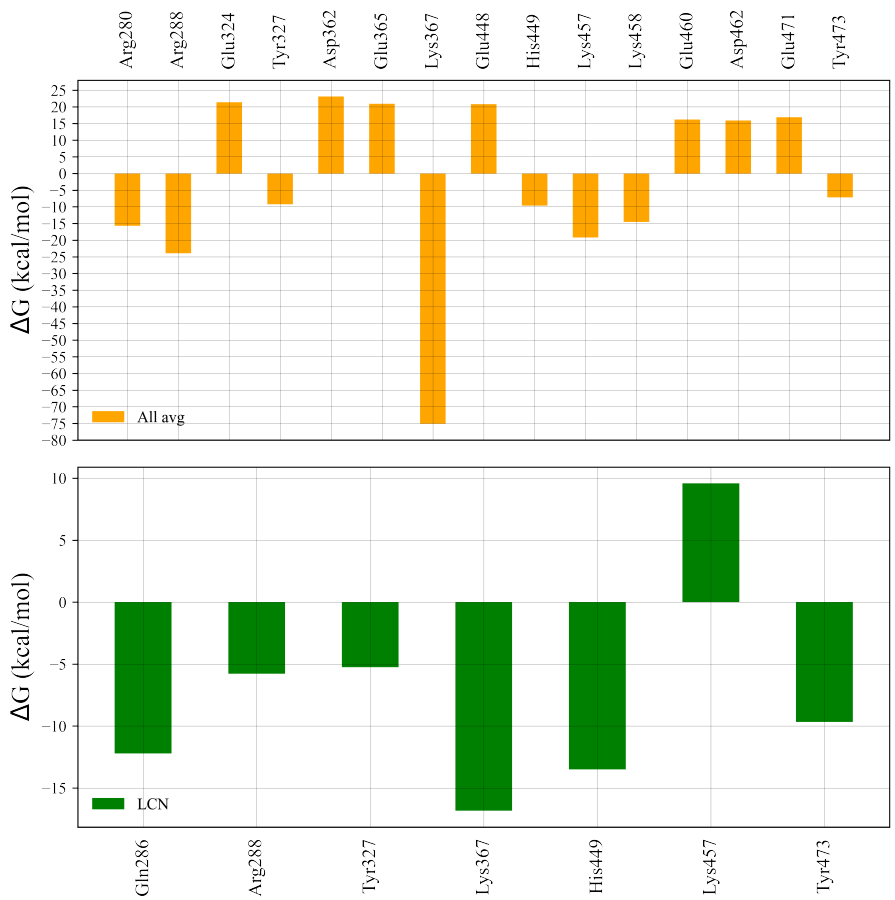


Figure 3.5 Average residue decomposition energies for Pocket 1. Averaged energies of PFAS (PFOS, PFOA, PFHxDA, ADONA, GenX and Et-PFOSAAcOH, and 6:2 FTSA) vs L-carnitine (LCN). Only the residues that have contributions above +5 kcal mol^{-1} and below -5 kcal mol^{-1} are shown.

5. The hydrogen bond percentages for each PFAS and L-carnitine in Pocket 1 are reported in Figure 8 (A). Even though the side chains of the basic residues do not directly interact with the PFAS compounds, Arg288, as an example, has stabilizing energetic contributions to the binding. Arginines and lysines have negative, i.e, stabilizing, effects on PFAS due to their positively charged side chains. On the other hand, the amino acids with acidic side chains have non-stabilizing effects on PFAS binding, due to negative-negative charge repulsions. This observation can be attributed to the total charge of the functional groups of the PFAS. These PFAS compounds, with the exception of 6:2 FTOH and PFOSA, have a net -1 charge, which enables salt bridges to be formed with nearby basic residues such as Lys367. These salt bridges are very strong and persistent, with large hydrogen

bond percentages (Figure 8 (A)). For instance, the Lys367 residue formed a strong hydrogen bond interaction with each PFAS that has a net charge but did not interact with PFOSA which is a neutral compound (Figure 5). Interestingly, the only hydrogen bond that PFOSA made was with His449, which has an interaction strength of ~ -10 kcal mol^{-1}. The largest negative electrostatic interaction came from Lys367, resulting in a -75 kcal mol^{-1}, on average. In addition, Lys367 forms the strongest hydrogen bonds to investigated compounds, with the exception of PFOSA and L-carnitine. Tyr327 also formed a hydrogen bond with most of the PFAS; although this residue did not have a very large negative electrostatic interaction on average (~ -10 kcal mol^{-1}), the hydrogen bonding with the PFAS species was quite strong. His449 also formed interactions with PFAS at ~1 kcal mol^{-1}, on average, while forming persistent hydrogen bonding with PFOSA and L-carnitine. An interesting observation for Pocket 1 was that L-carnitine only had one repulsive interaction with Lys457, while the rest of the PFAS have strong electrostatic interactions with this residue. Due to the zwitterionic nature of L-carnitine, the positively charged moiety orients towards Lys457 and results in repulsive interaction with Lys457.

3.3.3.2 DNA binding domain (DBD) – Pocket 3

For Pocket 3, the largest contributing residues towards binding were calculated and analyzed. With the exception of Et-PFOSAAcOHEt, no other PFAS formed hydrogen bonds with the surrounding residues. However, there were still strong electrostatic interactions with some of the residues, as detailed in Figure 7. The per-residue decomposition of the PFAS that coordinate to the zinc finger domain was plotted against PFAS that did not coordinate to zinc (Figure 7). The PFAS that kept their coordination to zinc were PFOS, PFOA, PFHxDA, GenX, ADONA, Et-PFOSAAcOHEt, and 6:2 FTSA. The other two PFAS (PFOSA and 6:2 FTOH), did not coordinate to zinc, but stayed in the pocket and in the vicinity of Zn^{2+}. Even though PFOSA and 6:2 FTOH did not coordinate to the Zn^{2+} and moved substantially within the binding pocket, they remained bound to the protein, albeit near the DNA instead of the zinc ion. The PFAS that did coordinate to Zn^{2+} were very stable within the binding pocket and did not show large conformational changes due to the strong interaction with the zinc ion. The average PFAS interaction with the zinc ion

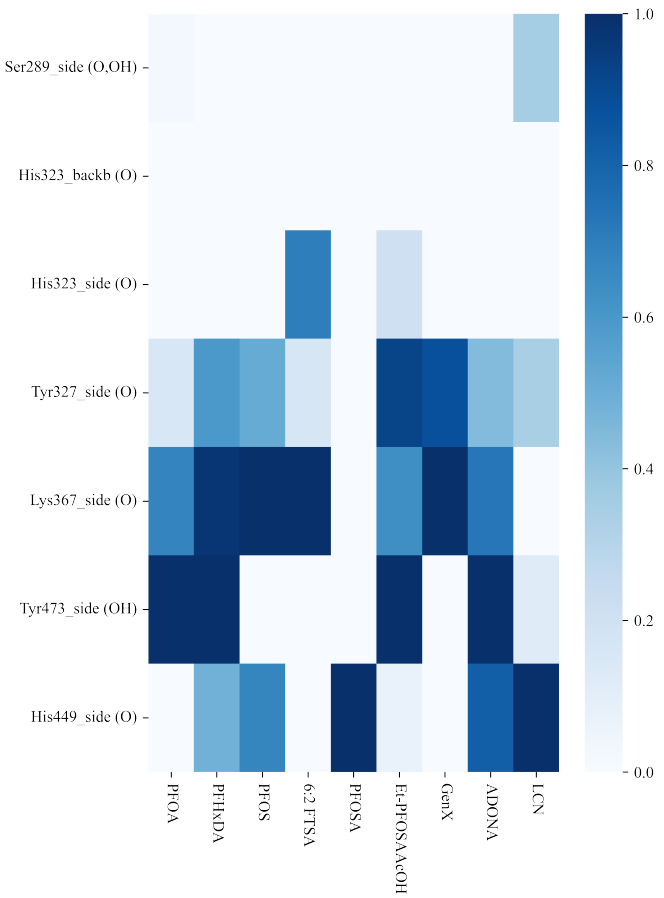


Figure 3.6 Hydrogen bond lifetimes of the PFAS and L-carnitine (LCN) in Pocket 1. The x-axis shows the simulated systems and y-axis shows the residue/atom information.

is -211 kcal mol^{-1}. Other residues that formed stabilizing interactions in this pocket are Arg147, Arg209, and Arg211, along with Lys161 and Lys213 from PPARγ. As per Figure S3, the other four residues that coordinate to zinc are deprotonated cysteines. These cysteines repel PFAS in the binding pocket, with average interaction energies of ~75-80 kcal mol^{-1} for Cym148, Cym152, and Cym162. However, for Cym165, the repulsion energy dropped to 54.5 kcal mol^{-1} versus the energy of other cysteines. The non-stabilizing contributions from these cysteine residues resulted in the largest contributions among the binding pocket residues. In general, aspartate residues, due to the negative charge on their side chain, also repelled the PFAS. It is interesting to note the role of DNA bases in binding. The energetic contribution from the DNA bases was always positive and ranged from 18.1 kcal mol-1 for DG471 to 24.2 kcal mol^{-1} for DT485. The PFAS that did not coordinate to the zinc dication remained in the proximity of the DNA bases during the simulations. The

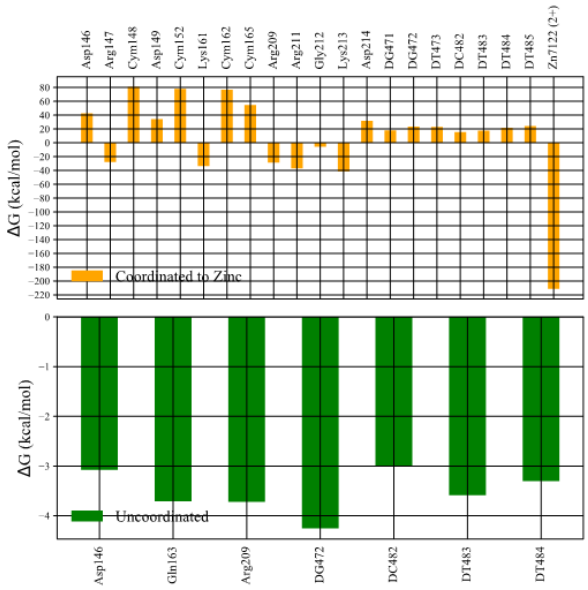


Figure 3.7 Average residue decomposition energies for the DBD pocket. Averaged energies of PFAS (PFOS, PFOA, PFHxDA, ADONA, GenX and Et-PFOSAAc-OHEt and 6:2 FTSA) coordinated to zinc versus energies of PFAS not coordinated to zinc (6:2 FTOH and PFOSA).

interactions with DG472, DC482, DT483 and DT484 base pairs were quite weak (all below -5 kcal mol^{-1}). In addition, Asp146 resulted in negative interactions with PFOSA and 6:2 FTOH relative to its interactions with the other seven PFAS that coordinate to Zn^{2+}. The other two residues that contribute towards binding were Gln163 and Arg209.

3.3.4 Interactions of DNA binding domains with the DNA molecule

The hydrogen bonding lifetimes of PPARγ-DNA and RXRα-DNA for the apo structures, in the presence of the co-crystallized ligands, bound PFAS, and L-carnitine in Pockets 1, 2, and 3 are depicted in Figures 8 and 9. The hydrogen bond network between the DBDs and DNA is crucial for the communication between the nuclear receptors and DNA. Comparing the interaction patterns in the apo structures and co-crystallized ligands against PFAS and L-carnitine bound complexes for the different pockets provides insight about the changes in the hydrogen bonding network with DNA. In addition, this analysis provides clarity about how PFAS binding affects the communication between PPARγ/RXRα receptors and the DNA. Most of the hydrogen bonds between the PPARγ-DBD and the DNA remained the same between the apo and co-crystallized ligand complexes (Figure 8). For these residue pairs, the most persistent interaction was DG486/Arg166 with a 100% lifetime,

followed by DT464/Tyr123 (~50%), DT464/Arg132 (~50%), and DG468/Arg159 (~50%). When the ligand binding domains did not include a ligand (apo), DG465 formed a hydrogen bond with Arg140 (~70% of the simulation time); however, in the presence of co-crystallized ligands (9CR and PLB), this interaction was not observed. Furthermore, in apo simulations, the DNA base DC488 forms a hydrogen bond with Glu129, which no longer occurs in the presence of co-crystallized ligands. On the contrary, the DG486/Arg137 interaction was only observed for the protein with the co-crystallized ligands, but not for the apo system. For the majority of the PFAS simulations in Pocket 1, the most striking differences were observed for the hydrogen bonding of DT485/Arg137 and DG486/Arg166 pairs, between the apo and co-crystalized ligand complexes simulation. While DT485/R137 had a low persistence in apo and co-crystallized ligand systems, in the majority of the PFAS simulations in Pocket 1, apart from PFOA, DT585 formed a strong hydrogen bond with Arg137. Another exception was observed for GenX where the DG486/Arg166 interaction was no longer present throughout the simulation. In addition, the interaction between the DG486/R159 occurred for a longer timeframe in the simulation. The hydrogen bonds in Pocket 2 showed similarities to what was observed in Pocket 1. DT485/R137 had a strong presence (~90% of simulation time) for all PFAS in Pocket 2 apart from PFOA, and the DG486/Arg159 and DG486/Arg166 interactions persisted for all PFAS in Pocket 2 with a higher hydrogen bonding percentage in L-carnitine. On the other hand, the DG464/Tyr123 and DG464/Lys132 interactions displayed an interaction strength similar to that of apo and co-crystallized ligand systems for PFOA, PFHxDA, PFOS, and 6:2 FTSA. For the rest of the compounds, the hydrogen bond lifetimes were very short. Another interesting interaction observed in Pocket 2 was the interaction between DG464 and Arg140 only for PFOA, PFHxDA, and PFOS, for almost ~80% of the simulation time. This interaction strength was not observed for the same PFAS in Pocket 1. And lastly, the presence of ADONA prompted the interaction between DT484 and Glu163 residue with a lifetime of ~90% of the simulation. The Pocket 3 hydrogen bond patterns between PPARγ-DBD and DNA bases present similar interactions for DG486/Arg159 and DG486/Arg166 pairs. In contrast to Pockets 1 and 2, compounds in Pocket 3 showed higher hydrogen bond percentages for 6:2 FTSA, PFOSA,

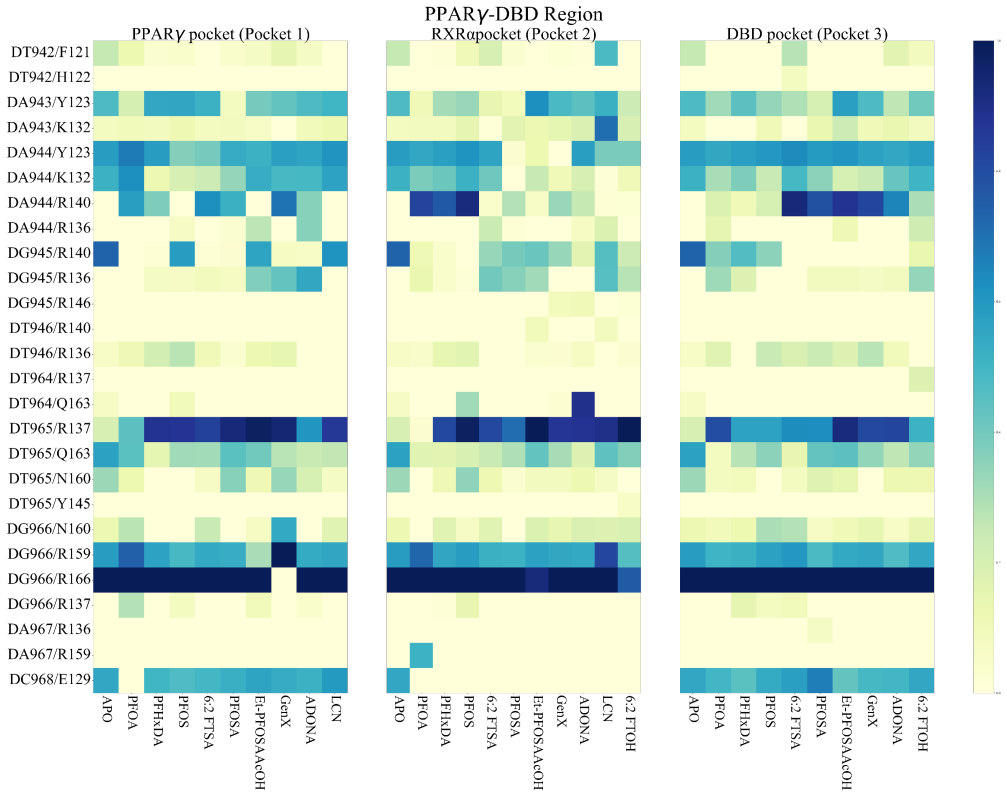


Figure 3.8 Hydrogen bond lifetimes of the PPARγ-DNA binding domain considering the protein's: apo structure, with its co-crystallized ligands (PLB, 9CR), PFAS, and L-carnitine (LCN) bound to Pockets 1, 2, and 3. The y-axis shows the hydrogen bond pairs between DNA and PPARγ residues involved in hydrogen bonding. The interactions that persist more than 10% of the simulation time are reported. DA, DC, DG, and DT represent the DNA bases. The notation on the y-axis represents the DNA base/Protein residue.

Et-PFOSAAcOH, GenX, and ADONA simulations. Another significant difference for Pocket 3 is that, overall, the hydrogen bonding persistence of DT485/Arg137 is lower than what was observed in Pocket 1 and Pocket 2.

When comparing the apo and the co-crystallized ligand complexes with PFAS bound structures, there are notable differences. The most persistent hydrogen bond of all simulations, DG486/Arg166, was hindered by the presence of GenX, which went from 100% persistence to 0% in Pocket 1. For Pockets 1 and 2, there was an increase in the hydrogen bond lifetime for DT485/Arg137 (doubled) for all PFAS in these pockets, except for PFOA. For Pocket 3, there was also an increase in the hydrogen bond lifetime of DT485/Arg137 pair. As mentioned previously, for the apo structure, the

base pair DC488 forms a hydrogen bond with Glu129. In addition, for Pocket 1 and 3, the DC488 and Glu129 bond also forms hydrogen bonding upon PFAS and L-carnitine binding, (except for PFOA in Pocket 1). However, for Pocket 2, the DC488 and Glu129 bond hydrogen bond either did not occur, or it occurred for less than 10% of the simulation time for all PFAS and L-carnitine. Furthermore, when the co-crystalized ligands are present, the hydrogen bond does not form for DC488/Glu129 pair. Another interesting feature when comparing the three pockets occurs for the DG464 and Arg140 interaction. For the simulation with co-crystallized ligands and the apo complex, there was a very small percentage of hydrogen bonding between the DNA base and Arg140. On the other hand, in Pockets 1, 2, and 3, there were numerous simulations that indicated an increased percentage of this hydrogen bonding, while L-carnitine showed a negligible hydrogen bonding percentage. Furthermore, an average hydrogen bond between DG465/Arg140 lasted for ~60% of the apo simulation, but for all of the PFAS simulations, it was a lot weaker or nonexistent in all pockets. The effects of the apo structure, co-crystallized ligands, PFAS and L-carnitine bound to the RXRα-LBD on the hydrogen bonding network with the DNA (DBD) are depicted in Figure 9. The interactions between the RXRα-DBD and the DNA molecule show that the DT478/Arg161 (~100% of time), DG479/Arg191 (~90%) and DG479/Arg184 (~80%), DG471/Tyr147(~60%) interactions persisted very strongly in both the apo and co-crystallized ligand complexes. On the other hand, DG472/Arg164 formed a hydrogen bond interaction only for the natural ligand system for ~80% of the simulation time. Similarly, DC481/Arg141 only occurred in the apo simulation for ~40% of the simulation time. DG471/Arg164 interaction resulted in a ~60% hydrogen bonding lifetime for apo, and 40% for the co-crystallized ligand simulations; however, there was a large increase for PFAS bound in Pockets 1, 2, and 3. For instance, PFOSA bound systems (for all pockets investigated) showed a higher hydrogen bonding percentage for the DG471/Arg164 interaction. Generally, PFAS bound to Pocket 3 had a higher persistence than the two other binding pockets for this residue pair. Another example of a large change in the hydrogen bonding network was observed for DC488/Glu129. When the co-crystallized ligands are bound to RXRα and PPARγ, there was no hydrogen bonding between this base pair and Glu129. For Pockets 1 and 3, this hydrogen bond

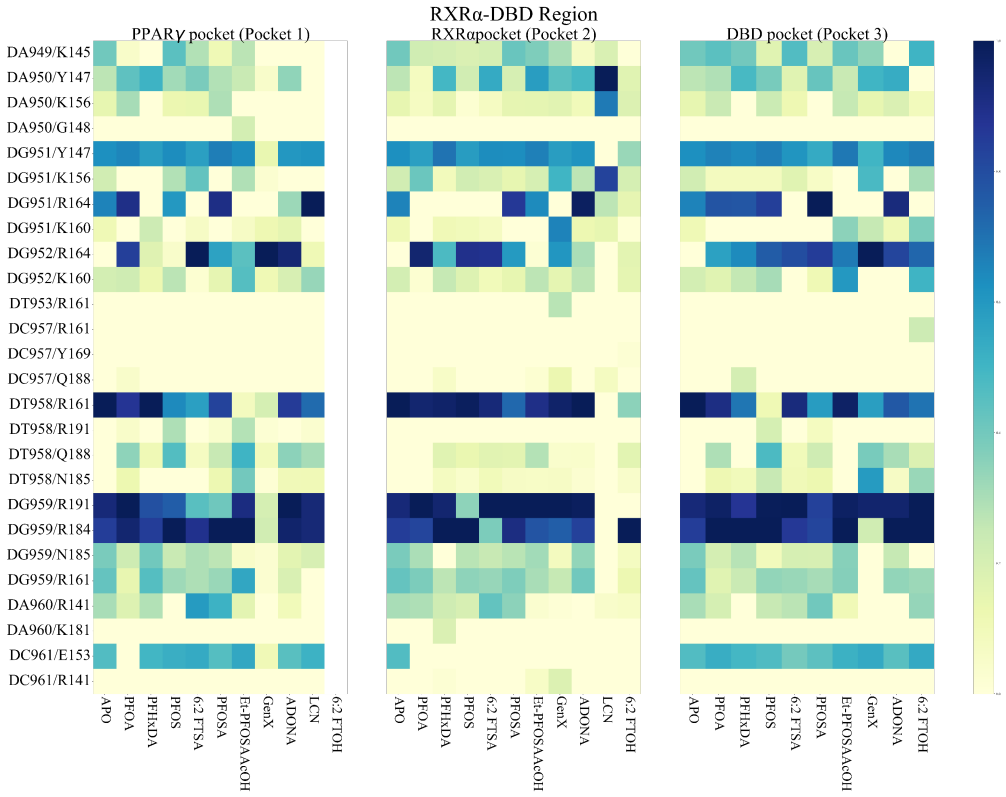


Figure 3.9 Hydrogen bond lifetimes of the RXRα-DNA binding domain considering the apo structure, with its co-crystallized ligands (PLB, 9CR), PFAS, and L-carnitine (LCN) bound to Pockets 1, 2, and 3. The y-axis shows the hydrogen bond pairs between RXRα residues, and in parenthesis, the residue involved in hydrogen bonding pertaining PFAS/L-carnitine. The interactions that persist more than 10% of the simulation time are reported.

was formed for all PFAS (except PFOA in Pocket 1) and the apo structure. However, for Pocket 2, this interaction was not formed for any PFAS bound to the RXRα binding domain.

3.3.5 Effect of PFAS binding on the DNA motion

In the previous section, how the binding of PFAS to Pocket 1, 2 and 3 has a direct impact on the interaction between proteins and the DNA molecule was discussed. To further understand the effects of the PFAS on DNA motions, the bending of the DNA was investigated for all simulations.49,132[49,120] Skaf et al. investigated the effect on an isolated DNA stretch and discovered that the apo structure is prone to bend up to 50°with a most dominant bending angle of ~15-20°.[49] In this work, a similar analysis was performed, i.e, the whole structure of the heterodimer, DNA, and the co-activator (NCOA2) were considered in its apo form. For this 200 ns simulation, the

DNA bending was calculated to be ~42°(Figures S30, S31). In addition, a comparison with the co-crystallized ligand simulation was also performed, without the presence of the co-activator as the structure was already in its activated form. For the latter, the average DNA bending was ~ 9°. The last analysis performed was the comparison of the DNA bending among the three pockets upon PFAS and L-carnitine binding. Overall, the bending increased between 1 to 2°with PFAS binding for all considered binding pockets with respect to the co-crystallized ligand complex. Compounds in Pocket 1 showed bending in a range of 9.0 to 9.5°in the presence of PFAS. ADONA had the smallest bending angle, and L-carnitine had the largest. For Pocket 2, the DNA bending upon PFAS binding was not as pronounced as Pocket 1. Et-PFOSAAcOH, ADONA, GenX and PFHxDA led to bending of the DNA molecule around 9 degrees, while PFOSA, PFOS, 6:2 FTSA, L-carnitine changed the angle to 9.5°. And finally, in Pocket 3, some of the PFAS had a more pronounced effect on the DNA bending compared to Pocket 1 and 2. Since Pocket 3 corresponds to one of the DNA binding domains, it is expected to have a stronger effect on the DNA interaction. Et-PFOSAAcOH had the most pronounced effect on the bending, very close to 10 degrees bending. The effect from other PFAS was 9 to 10°. It is important to note that for all investigated binding pockets, larger DNA bending was observed for PFAS and L-carnitine when compared to the co-crystallized ligand complex. The co-crystallized ligands are agonists for PPARγ and RXRα, therefore, observing similar DNA bending angles in the presence of PFAS implies that these molecules can replicate the downstream effects as agonist compounds. Furthermore, interestingly, for Pocket 3, the binding of PFAS results in similar behaviors as for Pockets 1 and 2, indicating that Pocket 3 could be another potential binding location for these PFAS compounds.

3.4 Conclusion

Herein, detailed structural analyses of the PPARγ/RXRα-DNA structures bound to PFAS, and L-carnitine were performed showing the potential of the selected PFAS as agonists. In addition, a comparison of the co-crystallized ligands with the apo structure was conducted. RMSF analysis indicated clearly that PFAS binding to the investigated binding pockets affects the movements of the LBDs and DBDs. More specifically, the hinge regions of RXRα and PPARγ have higher

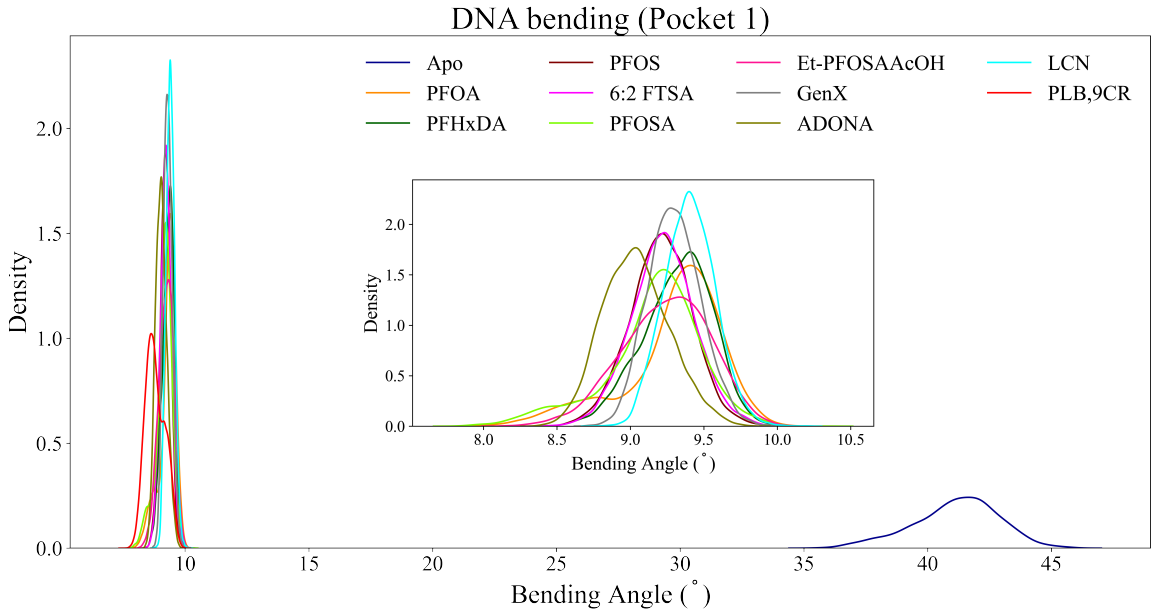


Figure 3.10 Distribution of the DNA bending angle in the Pocket 1 simulations with PFAS and L-carnitine. LCN: L-carnitine, PLB, 9CR: Natural ligands.

fluctuations upon PFAS binding, when compared to the apo and co-crystallized ligand simulations. A direct comparison of different PFAS and their binding energies indicated that the size of the carbon chain is proportional to the binding energies. Furthermore, the longer the carbon chain, the stronger the interaction energies, suggesting that the not only electrostatic interactions formed by the functional groups of PFAS, but the hydrophobic interactions with the PFAS tail are crucial for the strength of binding. In terms of the RXRα/PPARγ binding domains (Pockets 1 and 2), both PFHxDA and Et-PFOSAAcEtOH resulted in the highest binding affinities. Emergent PFAS such as ADONA and GenX can be competitively replaced by L-carnitine in Pocket 2. Considering the DNA binding domain, DFT and DLPNO calculations predicted that ADONA and PFOS are the strongest binding PFAS within the pocket. Moreover, some of the PFAS showed that by moving from the initial zinc and cysteine coordination, the PFAS can still be buried in the pocket and be stabilized by other key residues, with no disruption to the overall secondary structure packing of the proteins or to the interaction with the DNA oligomer. This is the first time in literature such a discovery has been made. It is known that the ligand binding domains of PPARγ and RXRα proteins are the primary binding sites for the investigated PFAS, with stronger preference for the RXRα LBD, based

on our overall binding energies. The third investigated site, near to a Zinc finger domain, can be a secondary or non-specific binding of PFAS to the PPARγ/RXRα-DNA complex. Furthermore, for all of the investigated binding pockets, the key residues have been identified, which are fundamental for developing compounds that can competitively replace PFAS from NRs. Finally, disruptions of the hydrogen bonding network of RXRα-DNA and PPARγ-DNA upon PFAS binding have been carried out for the three pockets. For DNA-residue pairs such as DC488/Glu129 (PPARγ-DNA) and DC481/Glu153 (RXRα-DNA), there is a decrease in hydrogen bonding when PFAS are bound to Pockets 1, 2 and 3. On the other hand, for residues such as DT485/Arg137 and DG472/Arg164, there is an increase in the hydrogen bond network of these DNA base pairs and protein residues for certain PFAS. DNA bending is associated with activation of PPARγ/RXRα complex. Upon PFAS and L-carnitine binding to the three different pockets, a bending angle of ~9°, similar to that shown in co-crystallized ligand bound simulations, was observed. However, with the removal of all ligands from the binding pockets, the bending angle of the DNA reaches the highest value at ~42°. It is important to note that, along with the interactions made with Helix 12, the observed bending angles provide evidence that PFAS acts as an agonist and may trigger the same downstream effects as natural ligands upon binding to the PPARγ/RXRα complex. The results presented here for PFAS binding to a biologically relevant nuclear receptor complex provides important insight towards establishing PFAS mitigation strategies and better understanding the health implications of PFAS exposure. Furthermore, by identifying where and how strong PFAS bind, and which residues are responsible for molecular recognition, insight can be gained towards potential in vivo mitigation strategies - for the rational design of a mitigator compound which could help to alleviate the effects of PFAS in humans.

BIBLIOGRAPHY

[1] Sajid, M. and Ilyas, M. (2017). Ptfe-coated non-stick cookware and toxicity concerns: A perspective. *Environmental Science and Pollution Research*, 24:23436–23440.

[2] Rao, N. S. and Baker, B. E. (1994). *Textile Finishes and Fluorosurfactants*, pages 321–338. Springer US.

[3] Schaider, L. A., Balan, S. A., Blum, A., Andrews, D. Q., Strynar, M. J., Dickinson, M. E., Lunderberg, D. M., Lang, J. R., and Peaslee, G. F. (2017). Fluorinated compounds in u.s. fast food packaging. *Environmental Science & Technology Letters*, 4:105–111. doi: 10.1021/acs.estlett.6b00435.

[4] Buck, R. C., Franklin, J., Berger, U., Conder, J. M., Cousins, I. T., de Voogt, P., Jensen, A. A., Kannan, K., Mabury, S. A., and van Leeuwen, S. P. (2011). Perfluoroalkyl and polyfluoroalkyl substances in the environment: Terminology, classification, and origins. *Integrated Environmental Assessment and Management*, 7:513–541.

[5] Gagliano, E., Sgroi, M., Falciglia, P. P., Vagliasindi, F. G., and Roccaro, P. (2020). Removal of poly- and perfluoroalkyl substances (pfas) from water by adsorption: Role of pfas chain length, effect of organic matter and challenges in adsorbent regeneration. *Water Research*, 171:115381.

[6] Schumm, C. E., Loganathan, N., and Wilson, A. K. (2023). Influence of soil minerals on the adsorption, structure, and dynamics of genx. *ACS ES&T Water*, 3:2659–2670. doi: 10.1021/acsestwater.3c00171.

[7] Loganathan, N. and Wilson, A. K. (2022). Adsorption, structure, and dynamics of short- and long-chain pfas molecules in kaolinite: Molecular-level insights. *Environmental Science & Technology*, 56:8043–8052. doi: 10.1021/acs.est.2c01054.

[8] Almeida, N. M. S., Eken, Y., and Wilson, A. K. (2021). Binding of per- and polyfluoro-alkyl substances to peroxisome proliferator-activated receptor gamma. *ACS Omega*, 6:15103–15114. doi: 10.1021/acsomega.1c01304.

[9] Lai, T. T., Eken, Y., and Wilson, A. K. (2020). Binding of per- and polyfluoroalkyl substances to the human pregnane x receptor. *Environmental Science & Technology*, 54:15986–15995.

[10] Yu, C. H., Riker, C. D., en Lu, S., and Fan, Z. T. (2020). Biomonitoring of emerging contaminants, perfluoroalkyl and polyfluoroalkyl substances (pfas), in new jersey adults in 2016–2018. *International Journal of Hygiene and Environmental Health*, 223:34–44.

[EPA] Us environmental protection agency epa's per- and polyfluoroalkyl substances (pfas) action plan 2019 no. february.

[12] Houck, K. A., Patlewicz, G., Richard, A. M., Williams, A. J., Shobair, M. A., Smeltz, M.,

Clifton, M. S., Wetmore, B., Medvedev, A., and Makarov, S. (2021). Bioactivity profiling of per- and polyfluoroalkyl substances (pfas) identifies potential toxicity pathways related to molecular structure. *Toxicology*, 457:152789.

[13] Munoz, G., Liu, J., Duy, S. V., and Sauvé, S. (2019). Analysis of f-53b, gen-x, adona, and emerging fluoroalkylether substances in environmental and biomonitoring samples: A review. *Trends in Environmental Analytical Chemistry*, 23:e00066.

[14] Guo, H., Chen, J., Zhang, H., Yao, J., Sheng, N., Li, Q., Guo, Y., Wu, C., Xie, W., and Dai, J. (2022). Exposure to genx and its novel analogs disrupts hepatic bile acid metabolism in male mice. *Environmental Science & Technology*, 56:6133–6143. doi: 10.1021/acs.est.1c02471.

[15] Robarts, D. R., Venneman, K. K., Gunewardena, S., and Apte, U. (2022). Genx induces fibroinflammatory gene expression in primary human hepatocytes. *Toxicology*, 477:153259.

[16] Weikum, E. R., Liu, X., and Ortlund, E. A. (2018). The nuclear receptor superfamily: A structural perspective. *Protein Science*, 27:1876–1892.

[17] Desvergne, B. and Wahli, W. (1999). Peroxisome proliferator-activated receptors: Nuclear control of metabolism. *Endocrine Reviews*, 20:649–688.

[18] Lemotte, P. K., Keidel, S., and Apfel, C. M. (1996). Phytanic acid is a retinoid x receptor ligand. *European Journal of Biochemistry*, 236:328–333. https://doi.org/10.1111/j.1432-1033.1996.00328.x.

[19] 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., and Heery, D. M. (2017). Heterodimers of photoreceptor-specific nuclear receptor (pnr/nr2e3) and peroxisome proliferator-activated receptor- γ (ppar γ) are disrupted by retinal disease-associated mutations. *Cell Death & Disease*, 8:e2677–e2677.

[20] Todorov, V. T., Desch, M., Schmitt-Nilson, N., Todorova, A., and Kurtz, A. (2007). Peroxisome proliferator-activated receptor- γ is involved in the control of renin gene expression. *Hypertension*, 50:939–944. doi: 10.1161/HYPERTENSIONAHA.107.092817.

[21] Estany, J., Ros-Freixedes, R., Tor, M., and Pena, R. N. (2014). A functional variant in the stearoyl-coa desaturase gene promoter enhances fatty acid desaturation in pork. *PLoS ONE*, 9:e86177.

[22] Okuno, M., Arimoto, E., Ikenobu, Y., Nishihara, T., and Imagawa, M. (2001). Dual dna-binding specificity of peroxisome-proliferator-activated receptor γ controlled by heterodimer formation with retinoid x rceptor α. *Biochemical Journal*, 353:193–198.

[23] 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., and Milburn, M. V. (1998). Ligand binding and co-activator

assembly of the peroxisome proliferator-activated receptor-γ. *Nature*, 395:137–143.

[24] Chandra, V., Huang, P., Hamuro, Y., Raghuram, S., Wang, Y., Burris, T. P., and Rastinejad, F. (2008). Structure of the intact ppar-γ–rxr-α nuclear receptor complex on dna. *Nature*, 456:350–356.

[25] Hernandez-Quiles, M., Broekema, M. F., and Kalkhoven, E. (2021). Ppargamma in metabolism, immunity, and cancer: Unified and diverse mechanisms of action. *Frontiers in Endocrinology*, 12:36.

[26] Bain, D. L., Heneghan, A. F., Connaghan-Jones, K. D., and Miura, M. T. (2007). Nuclear receptor structure: Implications for function. *Annual Review of Physiology*, 69:201–220.

[27] Khorasanizadeh, S. and Rastinejad, F. (2001). Nuclear-receptor interactions on dna-response elements. *Trends in Biochemical Sciences*, 26:384–390.

[28] Jeninga, E. H., Gurnell, M., and Kalkhoven, E. (2009). Functional implications of genetic variation in human pparγ. *Trends in Endocrinology & Metabolism*, 20:380–387.

[29] Krezel, A. and Maret, W. (2016). The biological inorganic chemistry of zinc ions. *Archives of Biochemistry and Biophysics*, 611:3–19.

[30] Harney, A. S., Lee, J., Manus, L. M., Wang, P., Ballweg, D. M., LaBonne, C., and Meade, T. J. (2009). Targeted inhibition of snail family zinc finger transcription factors by oligonucleotide-co(iii) schiff base conjugate. *Proceedings of the National Academy of Sciences*, 106:13667–13672. doi: 10.1073/pnas.0906423106.

[31] Yuan, S., Ding, X., Cui, Y., Wei, K., Zheng, Y., and Liu, Y. (2017). Cisplatin preferentially binds to zinc finger proteins containing c3h1 or c4 motifs. *European Journal of Inorganic Chemistry*, 2017:1778–1784. https://doi.org/10.1002/ejic.201601140.

[32] Sheng, Y., Cao, K., Li, J., Hou, Z., Yuan, S., Huang, G., Liu, H., and Liu, Y. (2018). Selective targeting of the zinc finger domain of hiv nucleocapsid protein ncp7 with ruthenium complexes. *Chemistry – A European Journal*, 24:19146–19151. https://doi.org/10.1002/chem.201803917.

[33] Kluska, K., Adamczyk, J., and Krezel, A. (2018). Metal binding properties, stability and reactivity of zinc fingers. *Coordination Chemistry Reviews*, 367:18–64.

[34] Quintal, S. M., DePaula, Q. A., and Farrell, N. P. (2011). Zinc finger proteins as templates for metal ion exchange and ligand reactivity. chemical and biological consequences. *Metallomics*, 3:121.

[35] Baglivo, I., Russo, L., Esposito, S., Malgieri, G., Renda, M., Salluzzo, A., Blasio, B. D., Isernia, C., Fattorusso, R., and Pedone, P. V. (2009). The structural role of the zinc ion can be dispensable in prokaryotic zinc-finger domains. *Proceedings of the National Academy of*

Sciences, 106:6933–6938.

[36] Dudev, T. and Lim, C. (2002). Factors governing the protonation state of cysteines in proteins: An ab initio/cdm study. *Journal of the American Chemical Society*, 124:6759–6766. doi: 10.1021/ja012620l.

[37] Issemann, I., Prince, R. A., Tugwood, J. D., and Green, S. (1993). The peroxisome proliferator-activated receptor: Retinoid x receptor heterodimer is activated by fatty acids and fibrate hypolipidaemic drugs. *Journal of Molecular Endocrinology*, 11:37–47.

[38] Mangelsdorf, D. J. and Evans, R. M. (1995). The rxr heterodimers and orphan receptors. *Cell*, 83:841–850.

[39] Kliewer, S. A., Umesono, K., Noonan, D. J., Heyman, R. A., and Evans, R. M. (1992). Convergence of 9-cis retinoic acid and peroxisome proliferator signalling pathways through heterodimer formation of their receptors. *Nature*, 358:771–774.

[40] Schulman, I. G., Shao, G., and Heyman, R. A. (1998). Transactivation by retinoid x receptor–peroxisome proliferator-activated receptor γ (pparγ) heterodimers: Intermolecular synergy requires only the ppar]γ hormone-dependent activation function. *Molecular and Cellular Biology*, 18:3483–3494.

[41] Xu, H., Lambert, M. H., Montana, V. G., Parks, D. J., Blanchard, S. G., Brown, P. J., Sternbach, D. D., Lehmann, J. M., Wisely, G., Willson, T. M., Kliewer, S. A., and Milburn, M. V. (1999). Molecular recognition of fatty acids by peroxisome proliferator–activated receptors. *Molecular Cell*, 3:397–403.

[42] Oberfield, J. L., Collins, J. L., Holmes, C. P., Goreham, D. M., Cooper, J. P., Cobb, J. E., Lenhard, J. M., Hull-Ryde, E. A., Mohr, C. P., Blanchard, S. G., Parks, D. J., Moore, L. B., Lehmann, J. M., Plunket, K., Miller, A. B., Milburn, M. V., Kliewer, S. A., and Willson, T. M. (1999). A peroxisome proliferator-activated receptor γ ligand inhibits adipocyte differentiation. *Proceedings of the National Academy of Sciences*, 96:6102–6106. doi: 10.1073/pnas.96.11.6102.

[43] Ostberg, T., Svensson, S., Selén, G., Uppenberg, J., Thor, M., Sundbom, M., Sydow-Bäckman, M., Gustavsson, A.-L., and Jendeberg, L. (2004). A new class of peroxisome proliferator-activated receptor agonists with a novel binding epitope shows antidiabetic effects. *Journal of Biological Chemistry*, 279:41124–41130.

[44] Burgermeister, E., Schnoebelen, A., Flament, A., Benz, J., Stihle, M., Gsell, B., Rufer, A., Ruf, A., Kuhn, B., Marki, H. P., Mizrahi, J., Sebokova, E., Niesor, E., and Meyer, M. (2006). A novel partial agonist of peroxisome proliferator-activated receptor-γ (pparγ) recruits pparγ-coactivator-1α, prevents triglyceride accumulation, and potentiates insulin signaling in vitro. *Molecular Endocrinology*, 20:809–830.

[45] Pochetti, G., Godio, C., Mitro, N., Caruso, D., Galmozzi, A., Scurati, S., Loiodice, F., Fracchiolla, G., Tortorella, P., Laghezza, A., Lavecchia, A., Novellino, E., Mazza, F., and Crestani, M. (2007). Insights into the mechanism of partial agonism: Crystal structures of the peroxisome proliferator-activated receptor gamma ligand-binding domain in the complex with two enantiomeric ligands. *The Journal of biological chemistry*, 282:17314–24.

[46] Li, Y., Wang, Z., Furukawa, N., Escaron, P., Weiszmann, J., Lee, G., Lindstrom, M., Liu, J., Liu, X., Xu, H., Plotnikova, O., Prasad, V., Walker, N., Learned, R. M., and Chen, J.-L. (2008). T2384, a novel antidiabetic agent with unique peroxisome proliferator-activated receptor γ binding properties. *Journal of Biological Chemistry*, 283:9168–9176.

[47] Motani, A., Wang, Z., Weiszmann, J., McGee, L. R., Lee, G., Liu, Q., Staunton, J., Fang, Z., Fuentes, H., Lindstrom, M., Liu, J., Biermann, D. H. T., Jaen, J., Walker, N. P. C., Learned, R. M., Chen, J.-L., and Li, Y. (2009). Int131: A selective modulator of ppar gamma. *Journal of molecular biology*, 386:1301–11.

[48] Bruning, J. B., Chalmers, M. J., Prasad, S., Busby, S. A., Kamenecka, T. M., He, Y., Nettles, K. W., and Griffin, P. R. (2007). Partial agonists activate ppargamma using a helix 12 independent mechanism. *Structure (London, England : 1993)*, 15:1258–71.

[49] Ricci, C. G., Silveira, R. L., Rivalta, I., Batista, V. S., and Skaf, M. S. (2016). Allosteric pathways in the pparγ-rxrα nuclear receptor complex. *Scientific Reports*, 6:19940.

[50] Levin, A. A., Sturzenbecker, L. J., Kazmer, S., Bosakowski, T., Huselton, C., Allenby, G., Speck, J., Ratzeisen, C., Rosenberger, M., Lovey, A., and Grippo, J. F. (1992). 9-cis retinoic acid stereoisomer binds and activates the nuclear receptor rxrα. *Nature*, 355:359–361.

[51] Zeng, Z., Song, B., Xiao, R., Zeng, G., Gong, J., Chen, M., Xu, P., Zhang, P., Shen, M., and Yi, H. (2019). Assessing the human health risks of perfluorooctane sulfonate by in vivo and in vitro studies. *Environment international*, 126:598–610.

[52] Sunderland, E. M., Hu, X. C., Dassuncao, C., Tokranov, A. K., Wagner, C. C., and Allen, J. G. (2019). A review of the pathways of human exposure to poly- and perfluoroalkyl substances (pfass) and present understanding of health effects. *Journal of Exposure Science & Environmental Epidemiology*, 29:131–147.

[53] Rappazzo, K., Coffman, E., and Hines, E. (2017). Exposure to perfluorinated alkyl substances and health outcomes in children: A systematic review of the epidemiologic literature. *International Journal of Environmental Research and Public Health*, 14:691.

[54] Szilagyi, J. T., Avula, V., and Fry, R. C. (2020). Perfluoroalkyl substances (pfas) and their effects on the placenta, pregnancy, and child development: a potential mechanistic role for placental peroxisome proliferator–activated receptors (ppars). *Current Environmental Health Reports*, 7:222–230.

[55] Anderko, L. and Pennea, E. (2020). Exposures to per-and polyfluoroalkyl substances (pfas): Potential risks to reproductive and children's health. *Current Problems in Pediatric and Adolescent Health Care*, 50:100760.

[56] Roth, J., Abusallout, I., Hill, T., Holton, C., Thapa, U., and Hanigan, D. (2020). Release of volatile per- and polyfluoroalkyl substances from aqueous film-forming foam. *Environmental Science & Technology Letters*, 7:164–170. doi: 10.1021/acs.estlett.0c00052.

[57] Xu, Y., Jurkovic-Mlakar, S., Li, Y., Wahlberg, K., Scott, K., Pineda, D., Lindh, C. H., Jakobsson, K., and Engström, K. (2020). Association between serum concentrations of perfluoroalkyl substances (pfas) and expression of serum micrornas in a cohort highly exposed to pfas from drinking water. *Environment International*, 136:105446.

[58] Hu, X. C., Andrews, D. Q., Lindstrom, A. B., Bruton, T. A., Schaider, L. A., Grandjean, P., Lohmann, R., Carignan, C. C., Blum, A., Balan, S. A., Higgins, C. P., and Sunderland, E. M. (2016). Detection of poly- and perfluoroalkyl substances (pfass) in u.s. drinking water linked to industrial sites, military fire training areas, and wastewater treatment plants. *Environmental Science & Technology Letters*, 3:344–350. doi: 10.1021/acs.estlett.6b00260.

[59] Chou, H.-C., Wen, L.-L., Chang, C.-C., Lin, C.-Y., Jin, L., and Juan, S.-H. (2017). From the cover: l-carnitine via pparγ- and sirt1-dependent mechanisms attenuates epithelial-mesenchymal transition and renal fibrosis caused by perfluorooctanesulfonate. *Toxicological Sciences*, 160:217–229.

[60] Wen, L.-L., Lin, C.-Y., Chou, H.-C., Chang, C.-C., Lo, H.-Y., and Juan, S.-H. (2016). Perfluorooctanesulfonate mediates renal tubular cell apoptosis through ppargamma inactivation. *PLOS ONE*, 11:e0155190.

[61] Duan, X., Sun, W., Sun, H., and Zhang, L. (2021). Perfluorooctane sulfonate continual exposure impairs glucose-stimulated insulin secretion via sirt1-induced upregulation of ucp2 expression. *Environmental Pollution*, 278:116840.

[62] Liu, W.-S., Lai, Y.-T., Chan, H.-L., Li, S.-Y., Lin, C.-C., Liu, C.-K., Tsou, H.-H., and Liu, T.-Y. (2018). Associations between perfluorinated chemicals and serum biochemical markers and performance status in uremic patients under hemodialysis. *PloS one*, 13:e0200271.

[63] Zhang, L., Ren, X.-M., Wan, B., and Guo, L.-H. (2014). Structure-dependent binding and activation of perfluorinated compounds on human peroxisome proliferator-activated receptor γ. *Toxicology and Applied Pharmacology*, 279:275–283.

[64] Khazaee, M., Christie, E., Cheng, W., Michalsen, M., Field, J., and Ng, C. (2021). Perfluoroalkyl acid binding with peroxisome proliferator-activated receptors α, γ, and δ, and fatty acid binding proteins by equilibrium dialysis with a comparison of methods. *Toxics*, 9:45.

[65] Soderstrom, S., Lille-Langoy, R., Yadetie, F., Rauch, M., Milinski, A., Dejaegere, A., Stote,

R. H., Goksoyr, A., and Karlsen, O. A. (2022). Agonistic and potentiating effects of perfluoroalkyl substances (pfas) on the atlantic cod (gadus morhua) peroxisome proliferator-activated receptors (ppars). *Environment International*, 163:107203.

[66] Döpke, M. F., Moultos, O. A., and Hartkamp, R. (2020). On the transferability of ion parameters to the tip4p/2005 water model using molecular dynamics simulations. *The Journal of Chemical Physics*, 152:024501. doi: 10.1063/1.5124448.

[67] Dale, K., Yadetie, F., Horvli, T., Zhang, X., Froysa, H. G., Karlsen, O. A., and Goksoyr, A. (2022). Single pfas and pfas mixtures affect nuclear receptor- and oxidative stress-related pathways in precision-cut liver slices of atlantic cod (gadus morhua). *Science of The Total Environment*, 814:152732.

[68] Sun, X., Xie, Y., Zhang, X., Song, J., and Wu, Y. (2023). Estimation of per- and polyfluorinated alkyl substance induction equivalency factors for humpback dolphins by transactivation potencies of peroxisome proliferator-activated receptors. *Environmental science & technology*, 57:3713–3721.

[69] Flanagan, J. L., Simmons, P. A., Vehige, J., Willcox, M. D., and Garrett, Q. (2010). Role of carnitine in disease. *Nutrition and Metabolism*, 7:30.

[70] Heuvel, J. P. V., Thompson, J. T., Frame, S. R., and Gillies, P. J. (2006). Differential activation of nuclear receptors by perfluorinated fatty acid analogs and natural fatty acids: A comparison of human, mouse, and rat peroxisome proliferator-activated receptor-α, -β, and -γ, liver x receptor-β, and retinoid x receptor-α. *Toxicological Sciences*, 92:476–489.

[71] Salvalaglio, M., Muscionico, I., and Cavallotti, C. (2010). Determination of energies and sites of binding of pfoa and pfos to human serum albumin. *Journal of Physical Chemistry B*, 114:14860–14874. doi: 10.1021/jp106584b.

[72] Ng, C. A. and Hungerbuehler, K. (2015). Exploring the use of molecular docking to identify bioaccumulative perfluorinated alkyl acids (pfaas). *Environmental Science & Technology*, 49:12306–12314. doi: 10.1021/acs.est.5b03000.

[73] Chen, H., He, P., Rao, H., Wang, F., Liu, H., and Yao, J. (2015). Systematic investigation of the toxic mechanism of pfoa and pfos on bovine serum albumin by spectroscopic and molecular modeling. *Chemosphere*, 129:217–224.

[74] Cheng, W. and Ng, C. A. (2018). Predicting relative protein affinity of novel per- and polyfluoroalkyl substances (pfass) by an efficient molecular dynamics approach. *Environmental Science & Technology*, 52:7972–7980. doi: 10.1021/acs.est.8b01268.

[75] Li, C.-H., Ren, X.-M., Cao, L.-Y., Qin, W.-P., and Guo, L.-H. (2019). Investigation of binding and activity of perfluoroalkyl substances to the human peroxisome proliferator-activated receptor β/δ. *Environmental Science: Processes & Impacts*, 21:1908–1914.

[76] Behr, A.-C., Plinsch, C., Braeuning, A., and Buhrke, T. (2020). Activation of human nuclear receptors by perfluoroalkylated substances (pfas). *Toxicology in Vitro*, 62:104700.

[77] Evans, N., Conley, J. M., Cardon, M., Hartig, P., Medlock-Kakaley, E., and Gray, L. E. (2022). In vitro activity of a panel of per- and polyfluoroalkyl substances (pfas), fatty acids, and pharmaceuticals in peroxisome proliferator-activated receptor (ppar) alpha, ppar gamma, and estrogen receptor assays. *Toxicology and Applied Pharmacology*, 449:116136.

[78] Lai, T. T., Kuntz, D., and Wilson, A. K. (2022). Molecular screening and toxicity estimation of 260,000 perfluoroalkyl and polyfluoroalkyl substances (pfass) through machine learning. *Journal of Chemical Information and Modeling*, 62:4569–4578.

[79] (2022). Molecular operating environment (moe), 2022.02 chemical computing group ulc, 1010 sherbooke st. west, suite 910, montreal, qc, canada, h3a 2r7.

[80] Eken, Y., Almeida, N. M., Wang, C., and Wilson, A. K. (2021). Sampl7: Host–guest binding prediction by molecular dynamics and quantum mechanics. *Journal of Computer-Aided Molecular Design*, 35:63–77.

[81] Bali, S. K., Marion, A., Ugur, I., Dikmenli, A. K., Catak, S., and Aviyente, V. (2018). Activity of topotecan toward the dna/topoisomerase i complex: A theoretical rationalization. *Biochemistry*, 57:1542–1551.

[82] Labute, P. and Santavy, M. (2010). Sitefinder - locating binding sites in protein structures.

[83] Riplinger, C., Pinski, P., Becker, U., Valeev, E. F., and Neese, F. (2016). Sparse maps—a systematic infrastructure for reduced-scaling electronic structure methods. ii. linear scaling domain based pair natural orbital coupled cluster theory. *The Journal of Chemical Physics*, 144:024109.

[84] Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P., and Dupradeau, F.-Y. (2011). 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 Research*, 39:W511–W517.

[85] Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, 97:10269–10280.

[86] York, D. and P.A. Kollman, D. e. a. (2020). Amber 2020.

[87] Galindo-Murillo, R., Robertson, J. C., Zgarbová, M., Šponer, J., Otyepka, M., Jurečka, P., and Cheatham, T. E. (2016). Assessing the current state of amber force field modifications for dna. *Journal of Chemical Theory and Computation*, 12:4114–4127. doi: 10.1021/acs.jctc.6b00186.

[88] He, X., Man, V. H., Yang, W., Lee, T.-S., and Wang, J. (2020). A fast and high-quality charge model for the next generation general amber force field. *The Journal of Chemical Physics*, 153:114502.

[89] Li, P., Song, L. F., and Merz, K. M. (2015). Parameterization of highly charged metal ions using the 12-6-4 lj-type nonbonded model in explicit water. *The Journal of Physical Chemistry B*, 119:883–895. doi: 10.1021/jp505875v.

[90] Li, P. and Merz, K. M. (2017). Metal ion modeling using classical mechanics. *Chemical Reviews*, 117:1564–1686. doi: 10.1021/acs.chemrev.6b00440.

[91] Li, P., Song, L. F., and Merz, K. M. (2015). Systematic parameterization of monovalent ions employing the nonbonded model. *Journal of Chemical Theory and Computation*, 11:1645–1657. doi: 10.1021/ct500918t.

[92] Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *The Journal of Chemical Physics*, 120:9665–9678.

[93] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23:327–341.

[94] Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function and Genetics*, 55:383–394.

[95] Miller, B. R., McGee, T. D., Swails, J. M., Homeyer, N., Gohlke, H., and Roitberg, A. E. (2012). Mmpbsa.py : An efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation*, 8:3314–3321.

[96] Roe, D. R. and Cheatham, T. E. (2013). Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9:3084–3095. doi: 10.1021/ct400341p.

[97] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). Ucsf chimera: A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25:1605–1612.

[98] Blomberg, M. R. A., Borowski, T., Himo, F., Liao, R.-Z., and Siegbahn, P. E. M. (2014). Quantum chemical studies of mechanisms for metalloenzymes. *Chemical Reviews*, 114:3601–3658. doi: 10.1021/cr400388t.

[99] Findik, B. K., Cilesiz, U., Bali, S. K., Atilgan, C., Aviyente, V., and Dedeoglu, B. (2022). Investigation of iron release from the n- and c-lobes of human serum transferrin by quantum

chemical calculations. *Organic & Biomolecular Chemistry*, 20:8766–8774.

[100] Tzeliou, C. E., Mermigki, M. A., and Tzeli, D. (2022). Review on the qm/mm methodologies and their application to metalloproteins. *Molecules*, 27:2660.

[101] Roe, D. R. (2015). Introduction to hydrogen bond analysis.

[102] Becke, A. D. (1993). A new mixing of hartree–fock and local density-functional theories. *The Journal of Chemical Physics*, 98:1372–1377.

[103] Perdew, J. P. (1986). Erratum: Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Physical Review B*, 34:7406–7406.

[104] Grimme, S., Ehrlich, S., and Goerigk, L. (2011). Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, 32:1456–1465.

[105] Weigend, F. and Ahlrichs, R. (2005). Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7:3297–3305.

[106] Weigend, F. (2006). Accurate coulomb-fitting basis sets for h to rn. *Physical Chemistry Chemical Physics*, 8:1057.

[107] Rydberg, P. and Olsen, L. (2009). The accuracy of geometries for iron porphyrin complexes from density functional theory. *The Journal of Physical Chemistry A*, 113:11949–11953. doi: 10.1021/jp9035716.

[108] Tomasi, J., Mennucci, B., and Cammi, R. (2005). Quantum mechanical continuum solvation models. *Chemical Reviews*, 105:2999–3094.

[109] Floris, F. and Tomasi, J. (1989). Evaluation of the dispersion contribution to the solvation energy. a simple computational model in the continuum approximation. *Journal of Computational Chemistry*, 10:616–627.

[110] Floris, F. M., Tomasi, J., and Ahuir, J. L. P. (1991). Dispersion and repulsion contributions to the solvation energy: Refinements to a simple computational model in the continuum approximation. *Journal of Computational Chemistry*, 12:784–791.

[111] Pierotti, R. A. (1976). A scaled particle theory of aqueous and nonaqueous solutions. *Chemical Reviews*, 76:717–726. doi: 10.1021/cr60304a002.

[112] Yu, H. S., He, X., Li, S. L., and Truhlar, D. G. (2016). Mn15: A kohn–sham global-hybrid exchange–correlation density functional with broad accuracy for multi-reference and single-reference systems and noncovalent interactions. *Chemical Science*, 7:5032–5051.

[113] Marenich, A. V., Cramer, C. J., and Truhlar, D. G. (2009). 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. *The Journal of Physical Chemistry B*, 113:6378–6396.

[114] Barone, V. and Cossi, M. (1998). Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. *The Journal of Physical Chemistry A*, 102:1995–2001. doi: 10.1021/jp9716997.

[115] Neese, F. (2022). Software update: The orca program system—version 5.0. *WIREs Computational Molecular Science*, 12:e1606.

[116] Tekarli, S. M., Drummond, M. L., Williams, T. G., Cundari, T. R., and Wilson, A. K. (2009). Performance of density functional theory for 3d transition metal-containing complexes: Utilization of the correlation consistent basis sets. *The Journal of Physical Chemistry A*, 113:8607–8614. doi: 10.1021/jp811503v.

[117] Himo, F. and de Visser, S. P. (2022). Status report on the quantum chemical cluster approach for modeling enzyme reactions. *Communications Chemistry*, 5:29.

[118] Himo, F. (2017). Recent trends in quantum chemical modeling of enzymatic reactions. *Journal of the American Chemical Society*, 139:6780–6786.

[119] Siegbahn, P. E. M. (2011). The effect of backbone constraints: The case of water oxidation by the oxygen-evolving complex in psii. *ChemPhysChem*, 12:3274–3280.

[120] Robinson, C. E., Wu, X., Morris, D. C., and Gimble, J. M. (1998). Dna bending is induced by binding of the peroxisome proliferator-activated receptor $\gamma 2$ heterodimer to its response element in the murine lipoprotein lipase promoter. *Biochemical and Biophysical Research Communications*, 244:671–677.

APPENDIX A

SUPPORTING TABLES

Table S3.1 Docking scores of selected poses of PFAS compounds using MOE (LCN: L-carnitine).

Compound Name	Docking Score (Pocket 1)	Docking Score (Pocket 2)	Docking Score (Pocket 3)
PFOA	-5.06	-6.68	-7.54
PFHxDA	-5.21	-8.39	-9.18
PFOS	-4.61	-6.82	-5.41
6:2 FTOH	-----	-6.14	-5.42
6:2 FTSA	-5.95	-6.61	-5.94
PFOSA	-4.32	-6.76	-5.76
Et-PFOSAAcOH	-6.06	-8.23	-8.30
GenX	-5.03	-5.78	-5.36
ADONA	-6.13	-5.75	-6.42
LCN	-5.05	-5.64	-8.45

APPENDIX B

SUPPORTING FIGURES

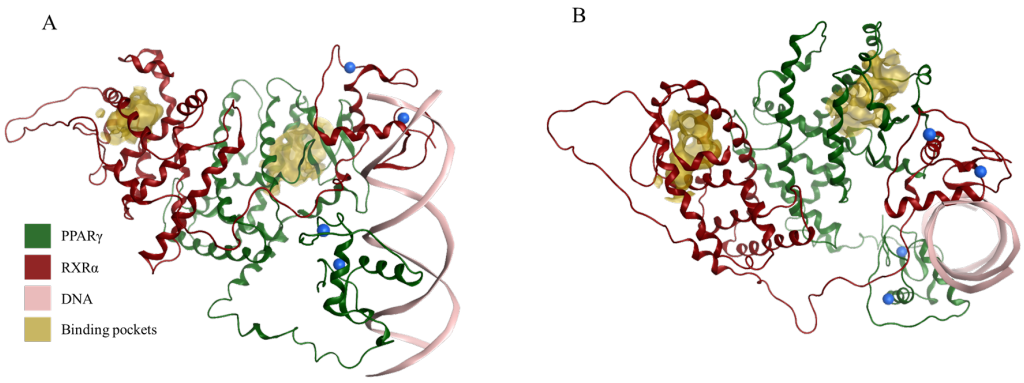


Figure S3.1 The docking pockets for PPARγ and RXRα LBD are shown with yellow surfaces. The other colors and their representations are as follows: Red: RXRα, green: PPARγ, tan: DNA, blue: Zn^{2+} ions. A: The view from the side; B; the view from the above.

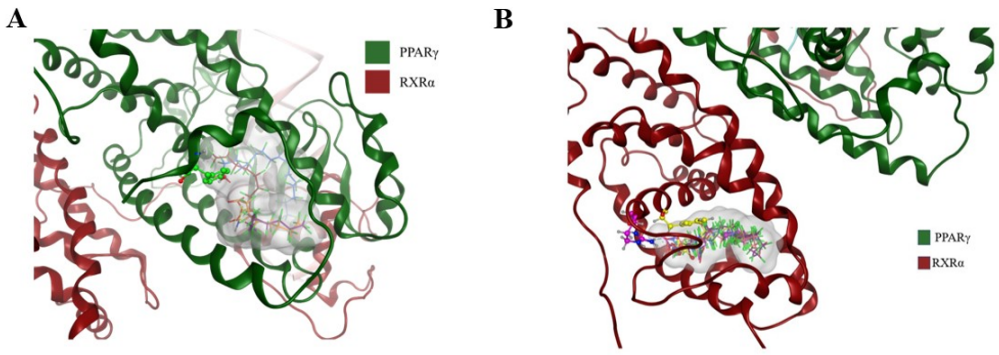


Figure S3.2 A: Overlap of the selected primary poses in the PPARγ-LBD pocket. The Tyr473 residue is shown in green ball-and-stick representation, and the PFAS compounds are shown in stick representation. B: The overlap of the selected primary poses in the RXR-LBD pocket. The Arg316 residue is shown in yellow ball-and-stick representation, The Phe 313 is shown in pink ball-and-stick representation and the PFAS compounds are shown in stick representation as well.

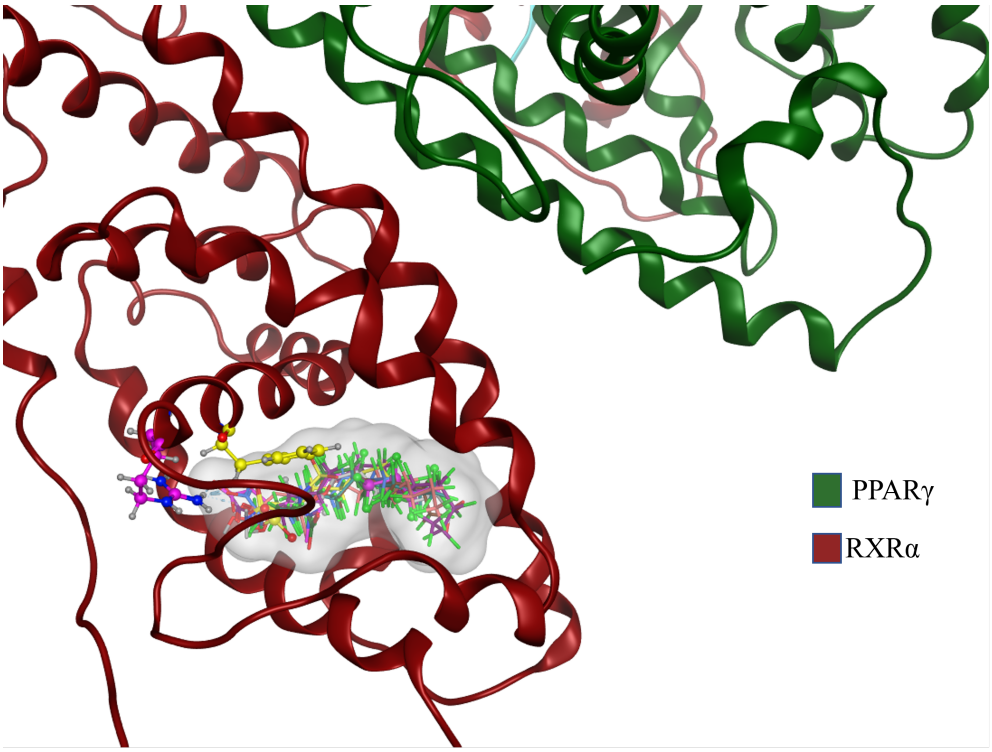


Figure S3.3 Overlap of the selected primary poses in the DBD pocket. Coordinating cysteine residues are shown in ball-and-stick representation, and the PFAS compounds are shown in stick representation. The zinc coordinating atom is shown in pink.

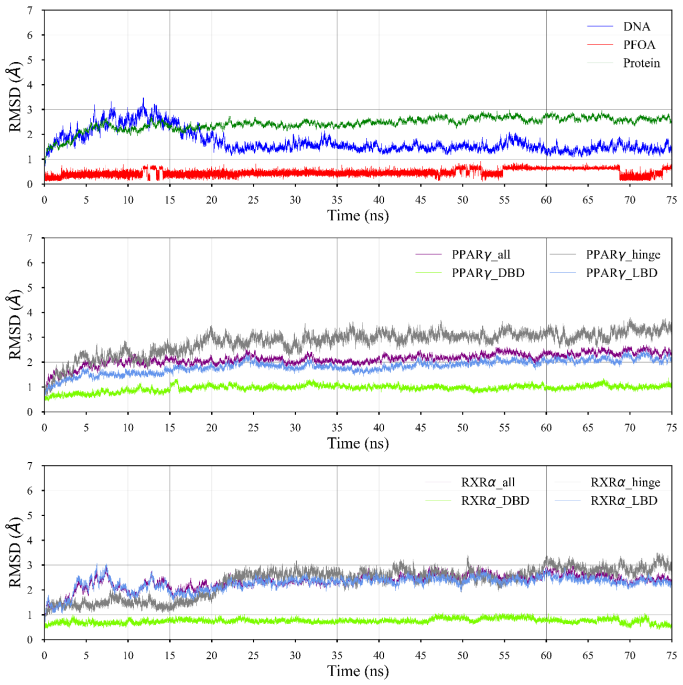


Figure S3.4 PFOA RMSD plots for Pocket 1.

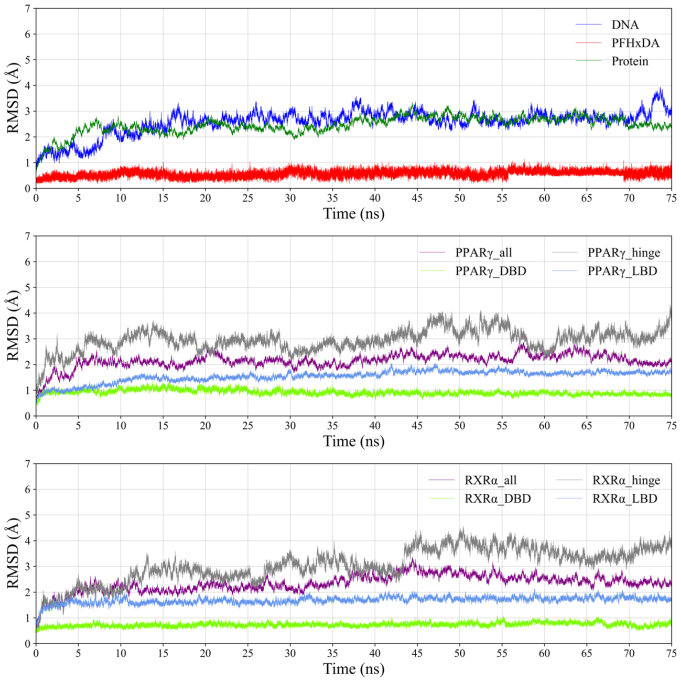


Figure S3.5 PFHxDA RMSD plots for Pocket 1.

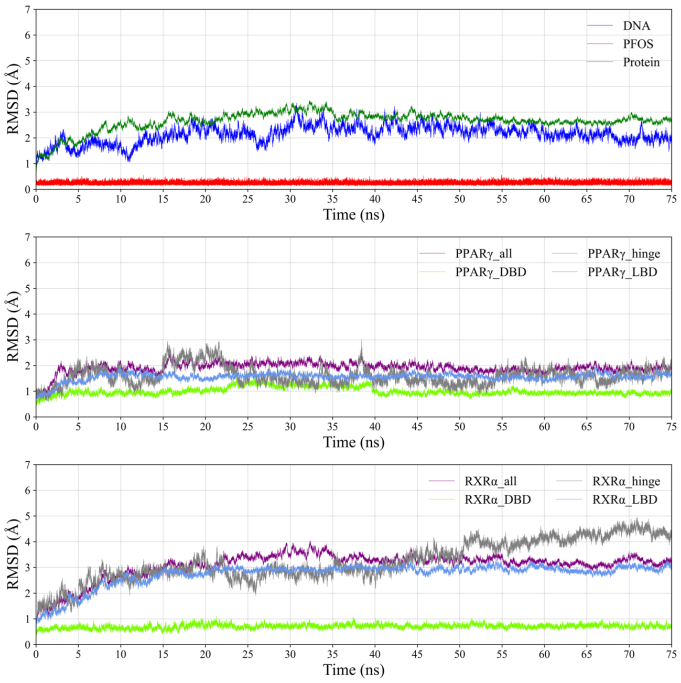


Figure S3.6 PFOS RMSD plots for Pocket 1.

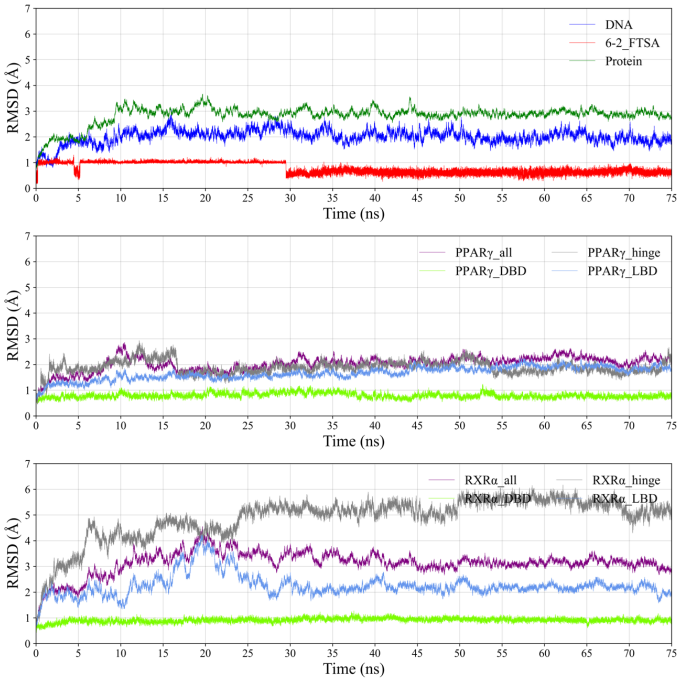


Figure S3.7 6:2 FTSA RMSD plots for Pocket 1.

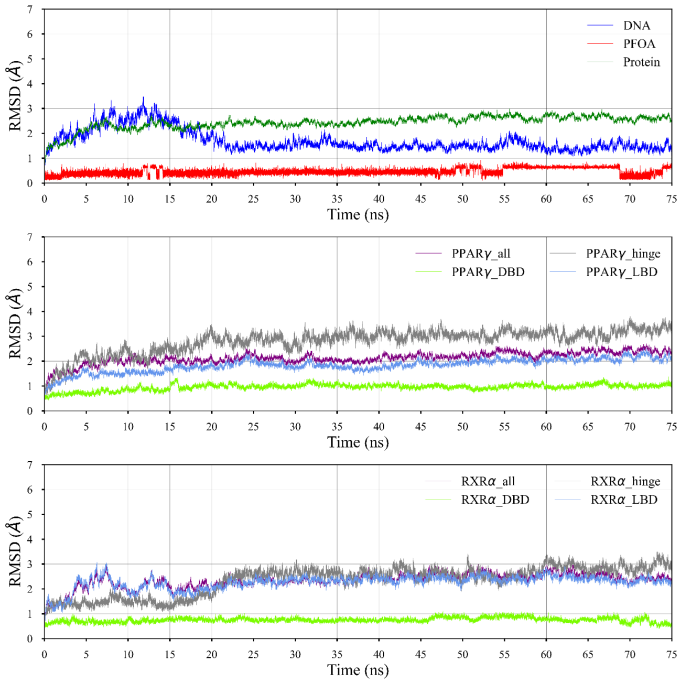


Figure S3.8 PFOA RMSD plots for Pocket 1.

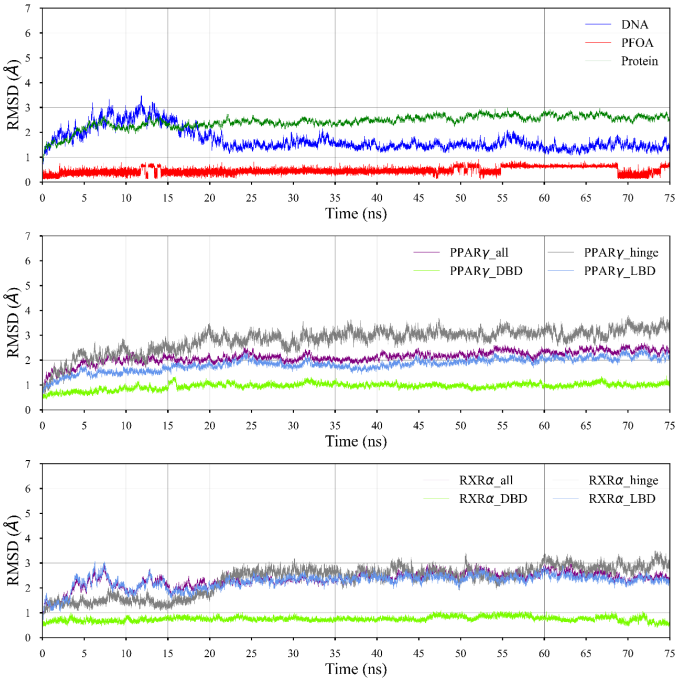


Figure S3.9 PFOA RMSD plots for Pocket 1.

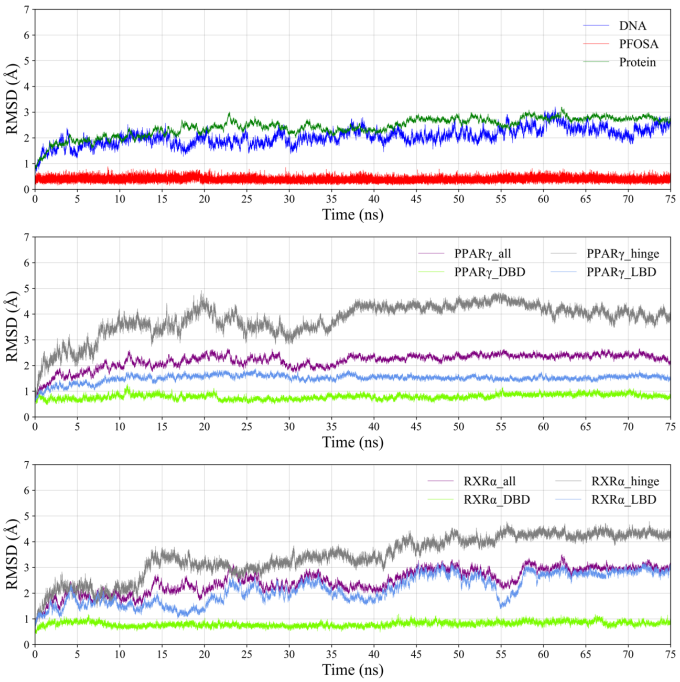


Figure S3.10 PFOSA RMSD plots for Pocket 1.

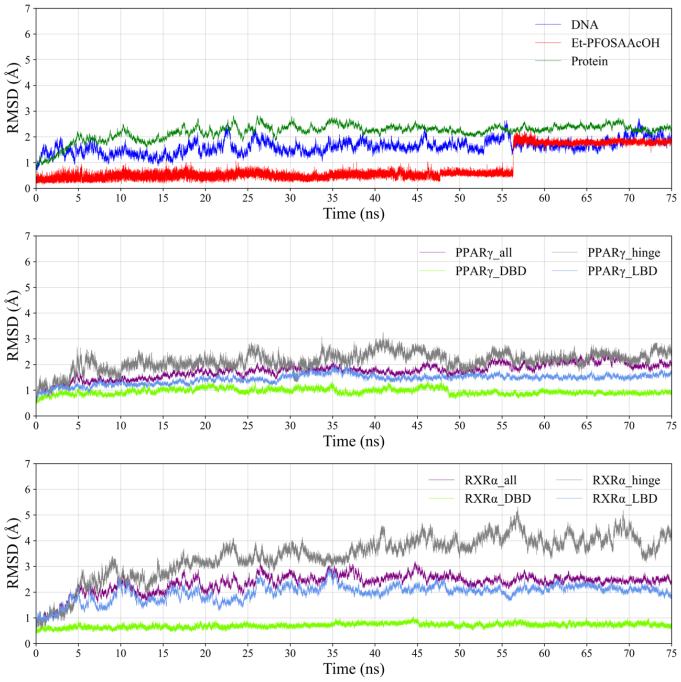


Figure S3.11 Et-PFOSAAcOH RMSD plots for Pocket 1.

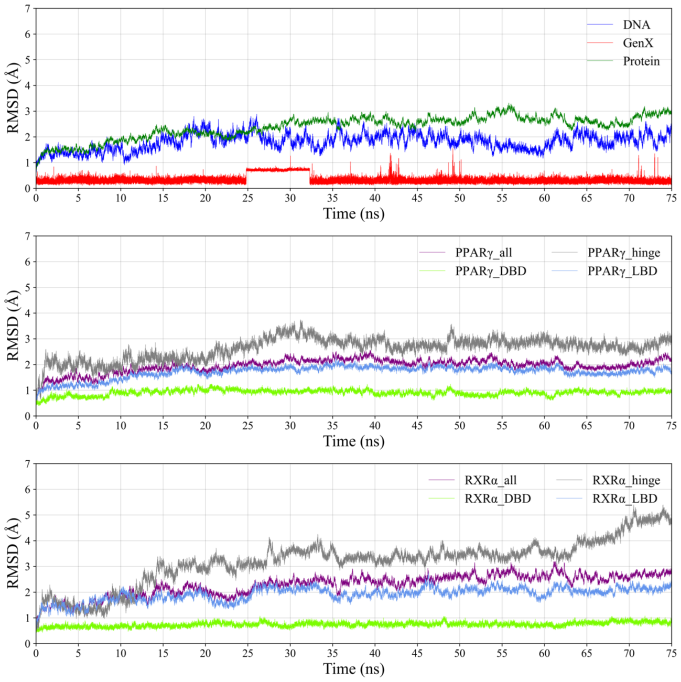


Figure S3.12 GenX RMSD plots for Pocket 1.

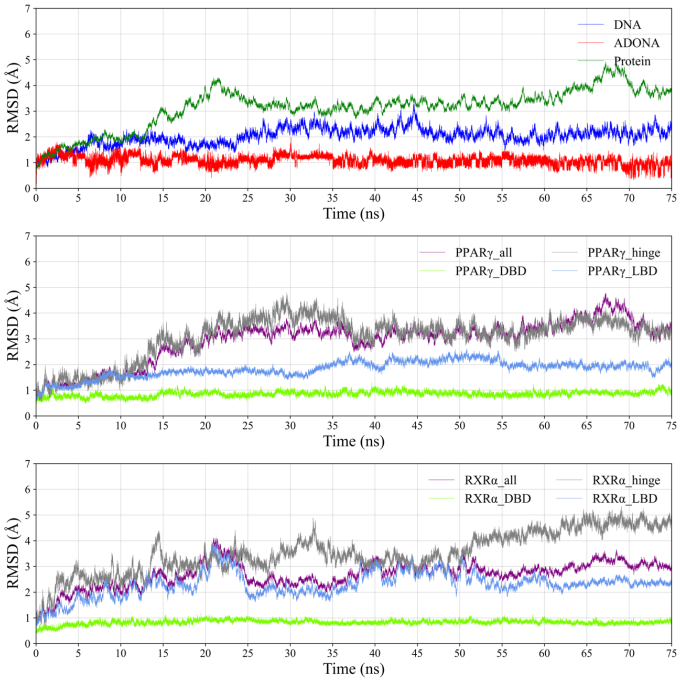


Figure S3.13 ADONA RMSD plots for Pocket 1.

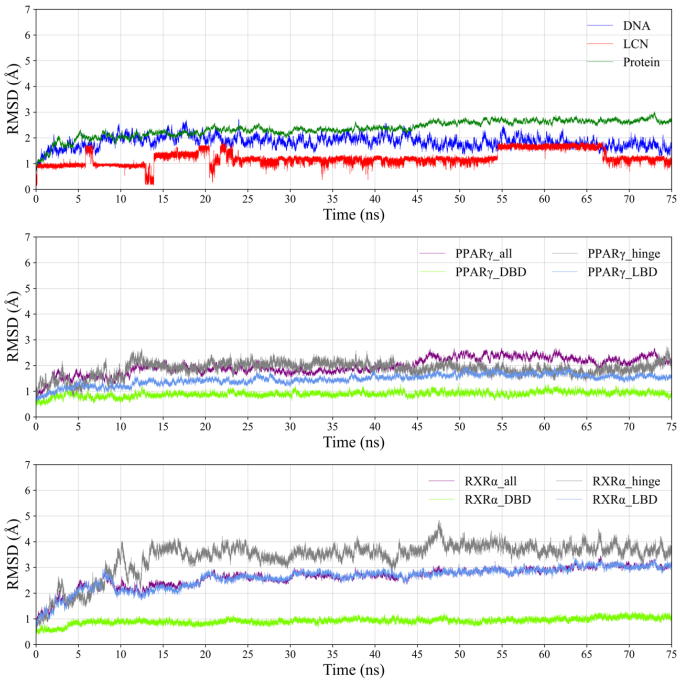


Figure S3.14 L-carnitine (LCN) RMSD plots for Pocket 1.

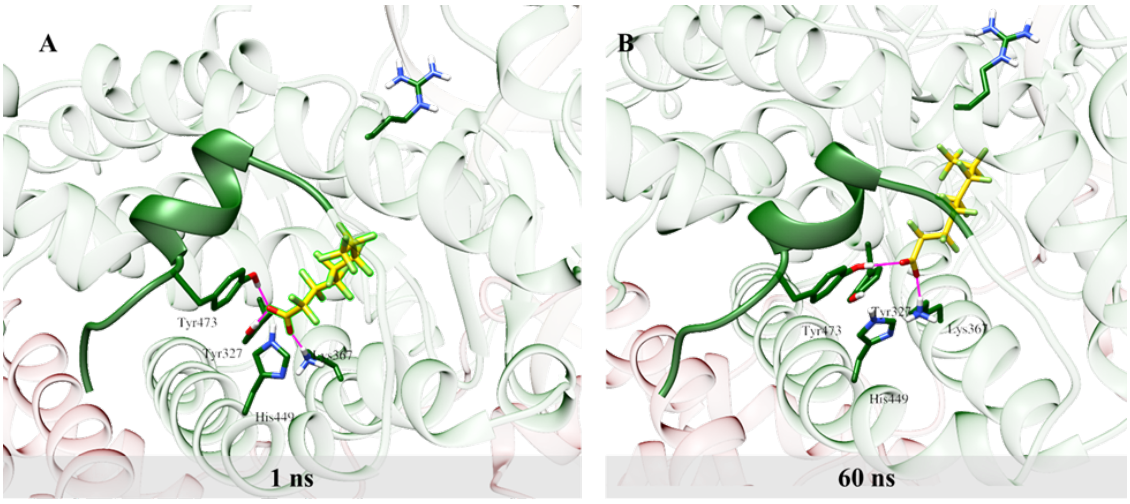


Figure S3.15 The orientations of PFOA at 1 ns (A) and 60 ns (B) of the simulation. The PPARγ is shown in green and RXRα is shown in red. The H12 helix from PPARγ is highlighted, along with important residues around PFOA (shown in yellow). The hydrogen bonding is indicated with a pink line.

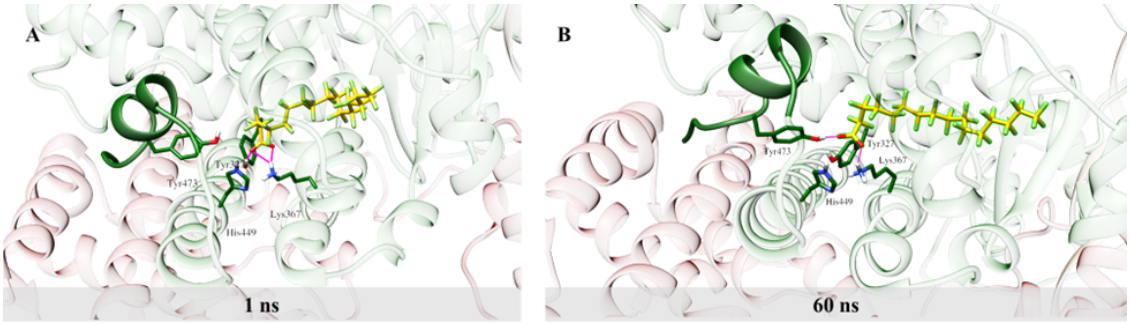


Figure S3.16 The orientations of PFHxDA at 1 ns (A) and 60 ns (B) of the simulation. The PPARγ is shown in green and RXRα is shown in red. The H12 helix from PPARγ is highlighted, along with important residues around PFHxDA (shown in yellow). The hydrogen bonding is indicated with a pink line.

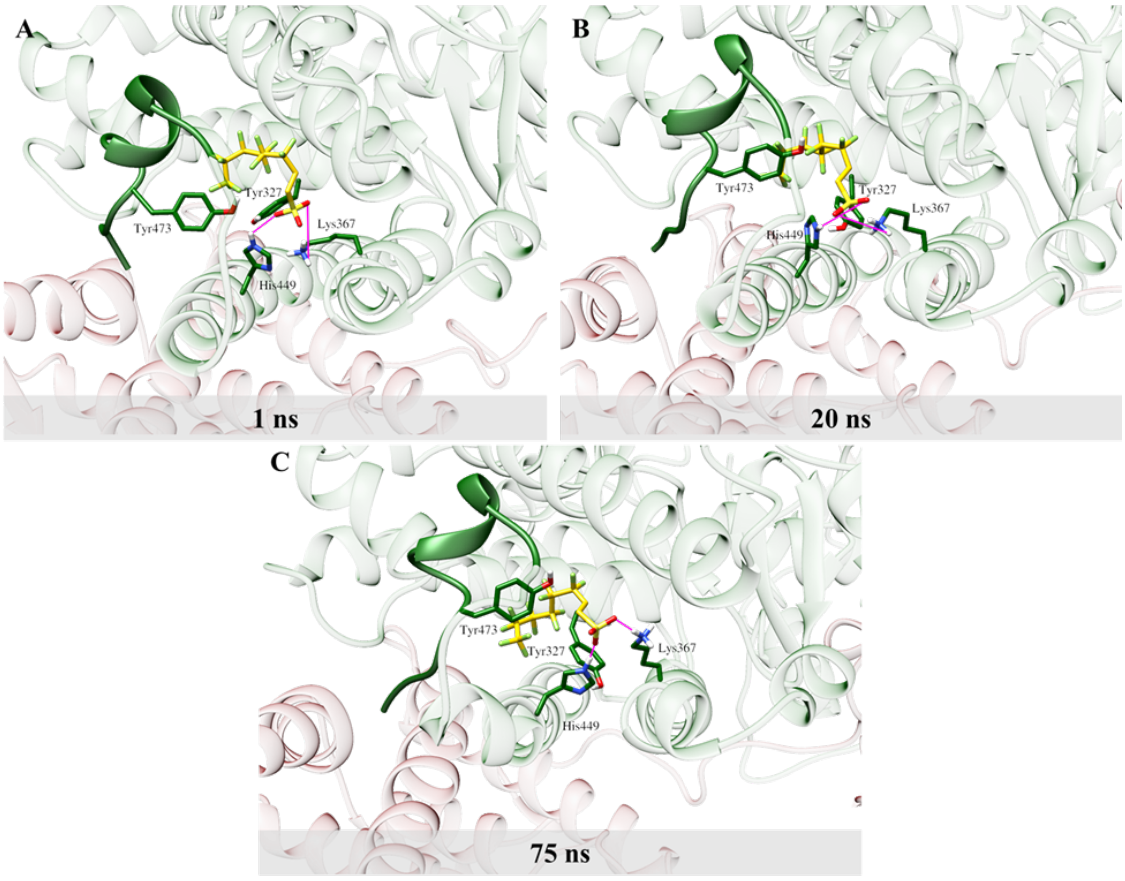


Figure S3.17 The orientations of 6:2 FTSA at 1 ns (A), 20 ns (B), and 75 ns (C) of the simulation. The PPARγ is shown in green and RXRα is shown in red. The H12 helix from PPARγ is highlighted, along with important residues around 6:2 FTSA (shown in yellow). The hydrogen bonding is indicated with a pink line.

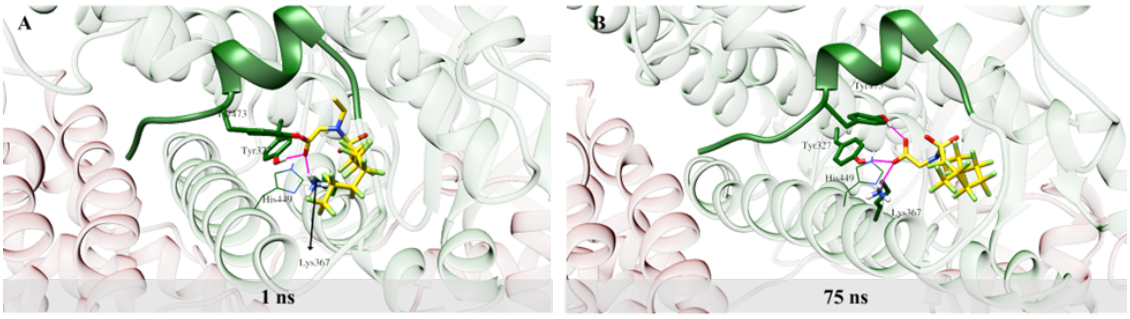


Figure S3.18 The orientations of Et-PFOSAAcOH at 1 ns (A), and 75 ns (B) of the simulation. The PPARγ is shown in green and RXRα is shown in red. The H12 helix from PPARγ is highlighted, along with important residues around Et-PFOSAAcOH (shown in yellow). The hydrogen bonding is indicated with a pink line.

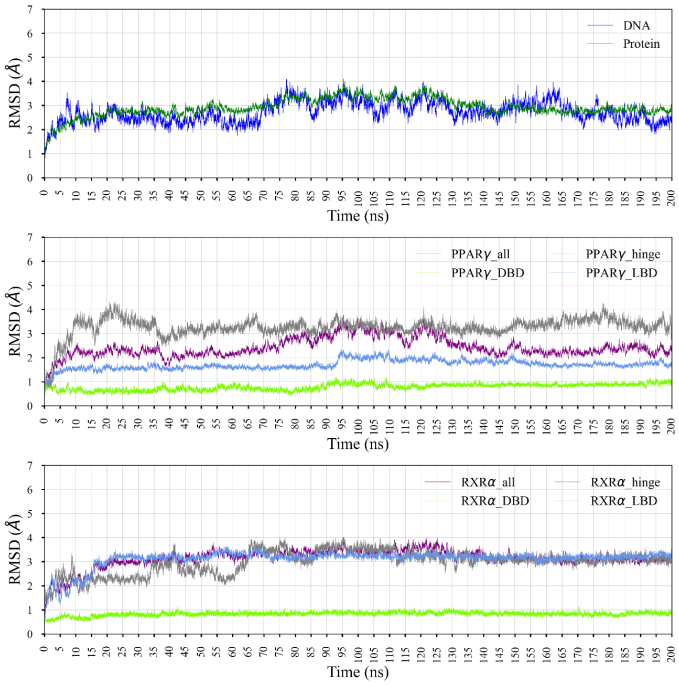


Figure S3.19 RMSD plots for the apo PPARγ-RXRα/DNA complex.

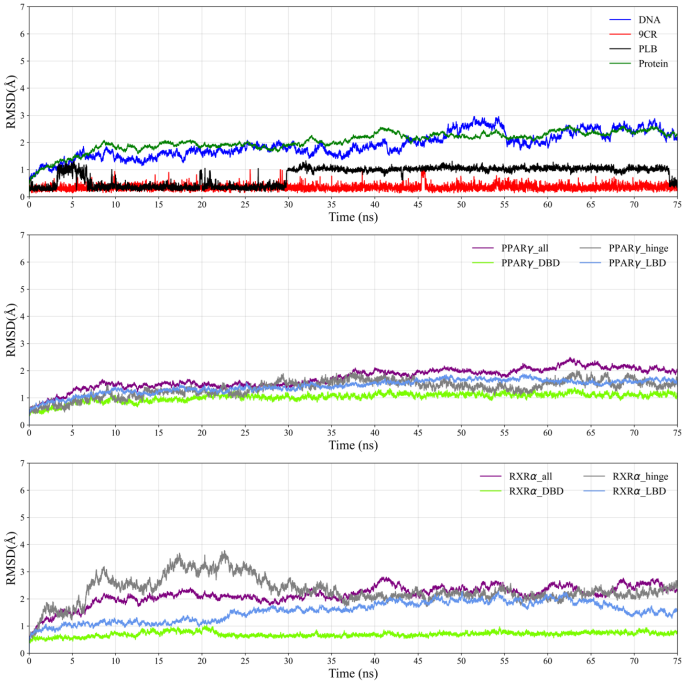


Figure S3.20 RMSD plots for the PPARγ-RXRα/DNA complex with its corresponding respective natural ligands.

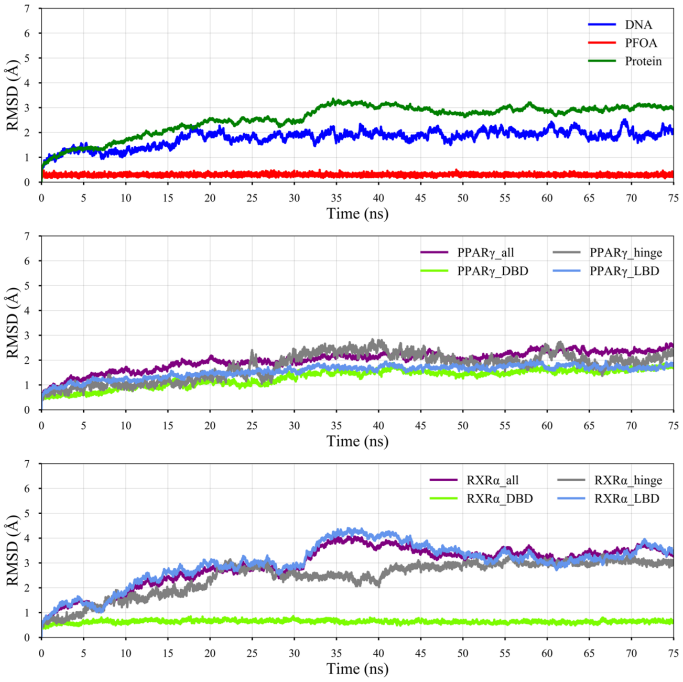


Figure S3.21 PFOA RMSD plots for the investigated DBD pocket (Pocket 3).

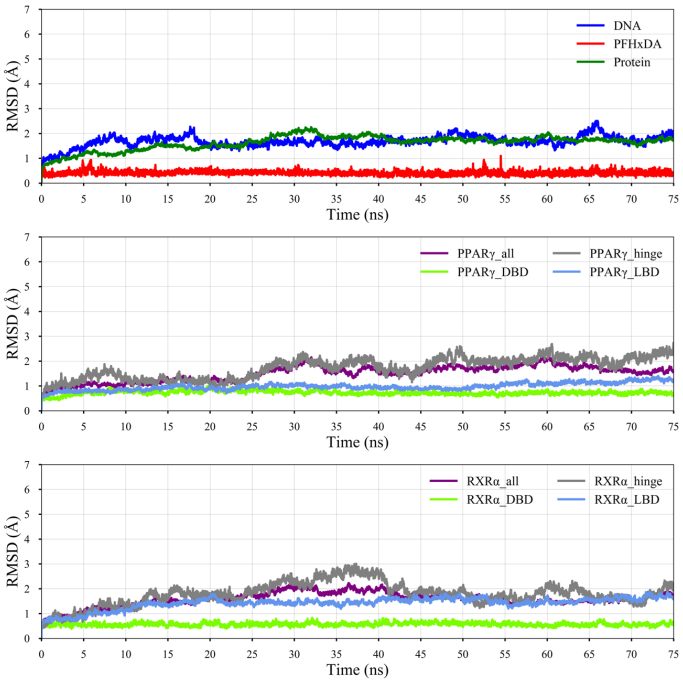


Figure S3.22 PFHxDA RMSD plots for the investigated DBD pocket (Pocket 3).

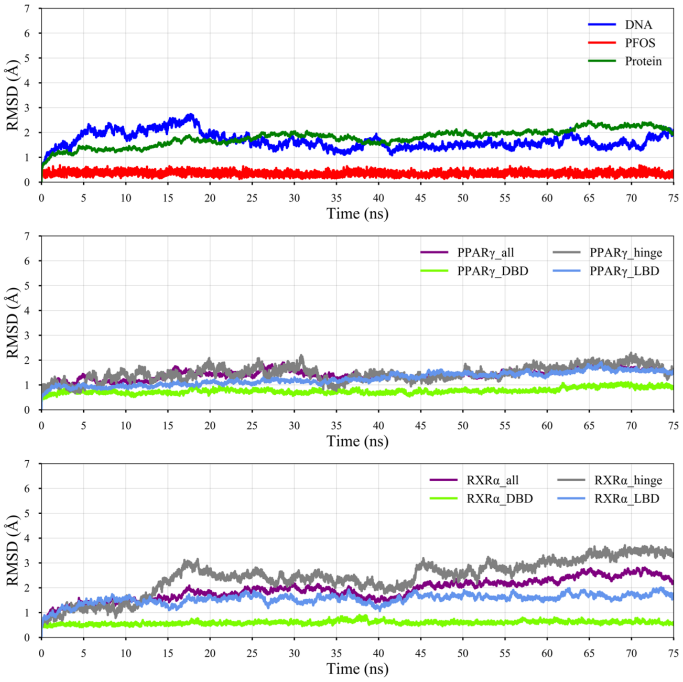


Figure S3.23 PFOS RMSD plots for the investigated DBD pocket (Pocket 3).

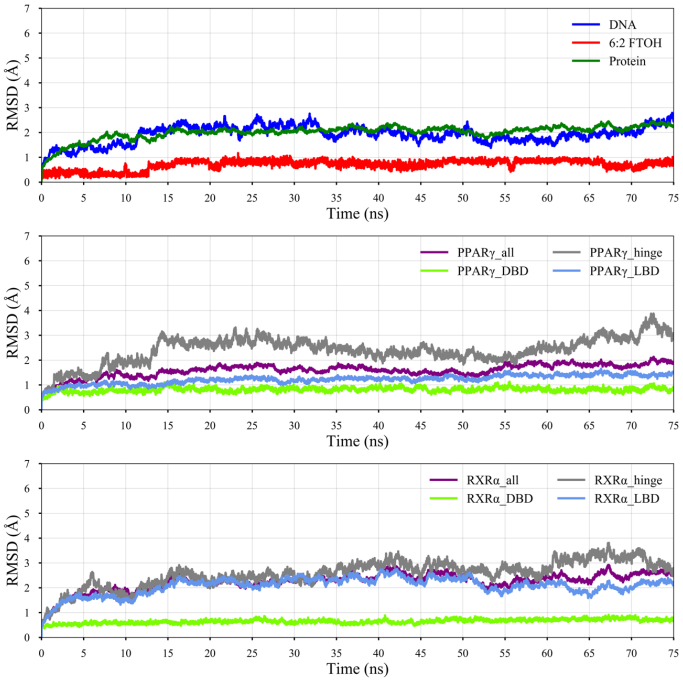


Figure S3.24 6:2 FTOH RMSD plots for the investigated DBD pocket (Pocket 3).

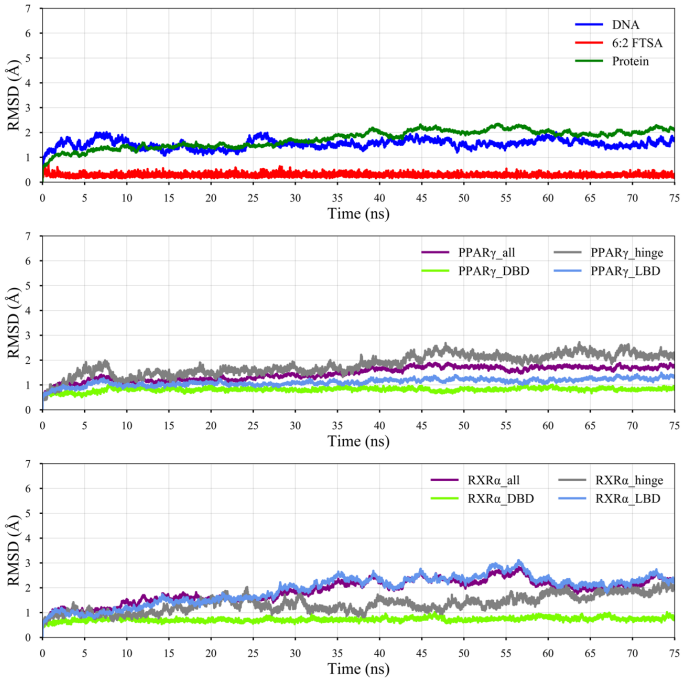


Figure S3.25 6:2 FTSA RMSD plots for the investigated DBD pocket (Pocket 3).

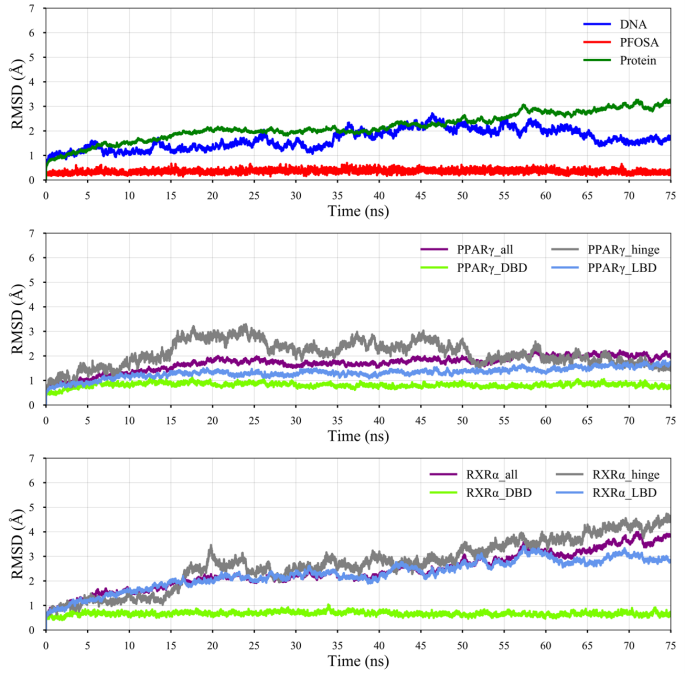


Figure S3.26 PFOSA RMSD plots for the investigated DBD pocket (Pocket 3).

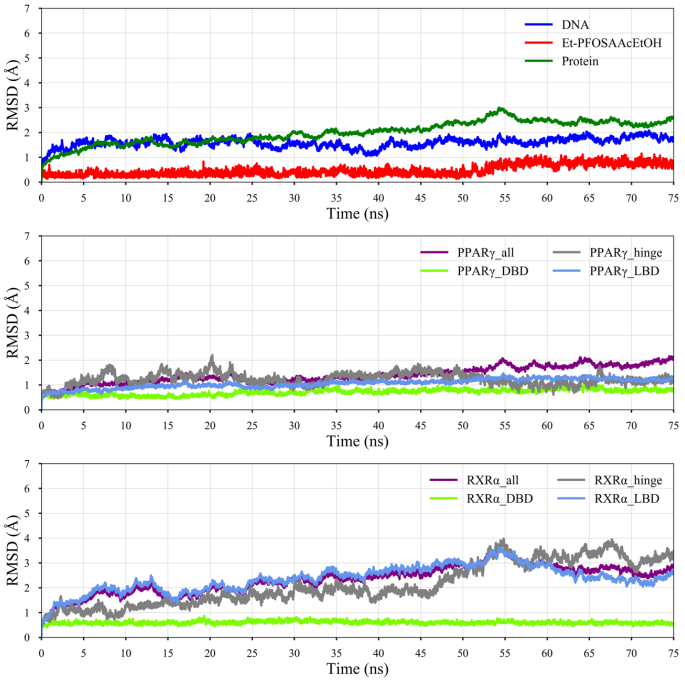


Figure S3.27 Et-PFOSAAcEtOH RMSD plots for the investigated DBD pocket (Pocket 3).

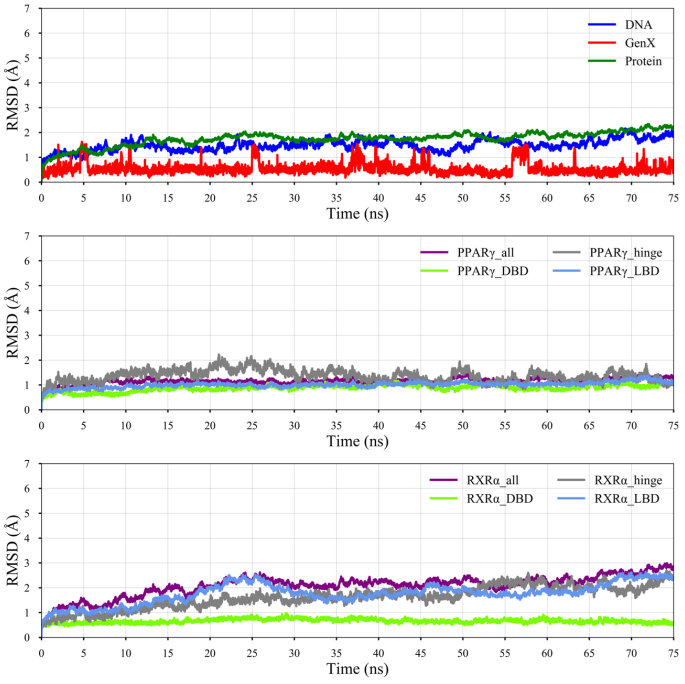


Figure S3.28 GenX RMSD plots for the investigated DBD pocket (Pocket 3).

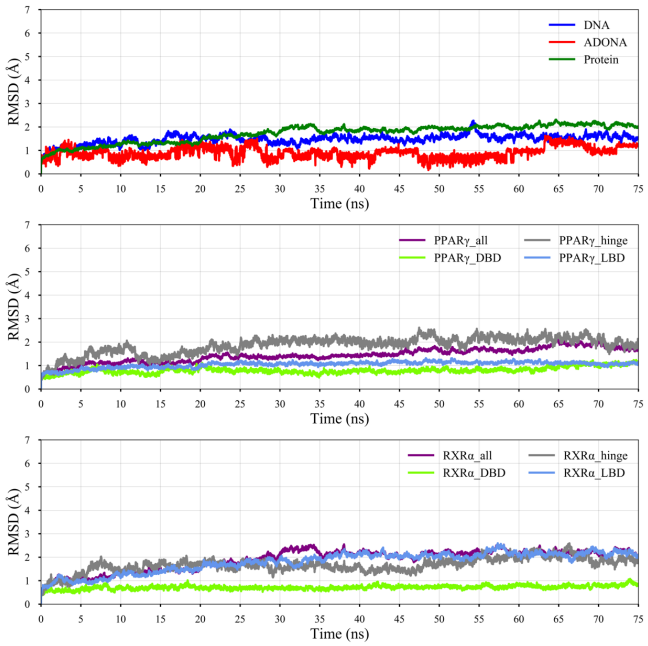


Figure S3.29 ADONA RMSD plots for the investigated DBD pocket (Pocket 3).

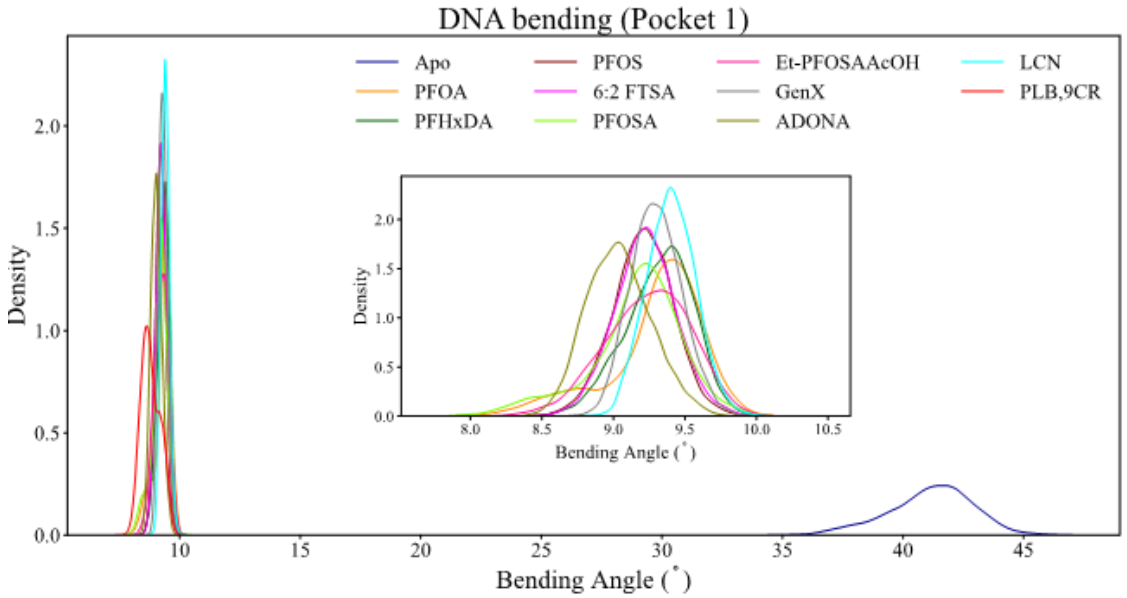


Figure S3.30 Distribution of the DNA bending angle in the Pocket 1 simulations with PFAS and L-carnitine. LCN: L-carnitine, PLB, 9CR: Natural ligands.

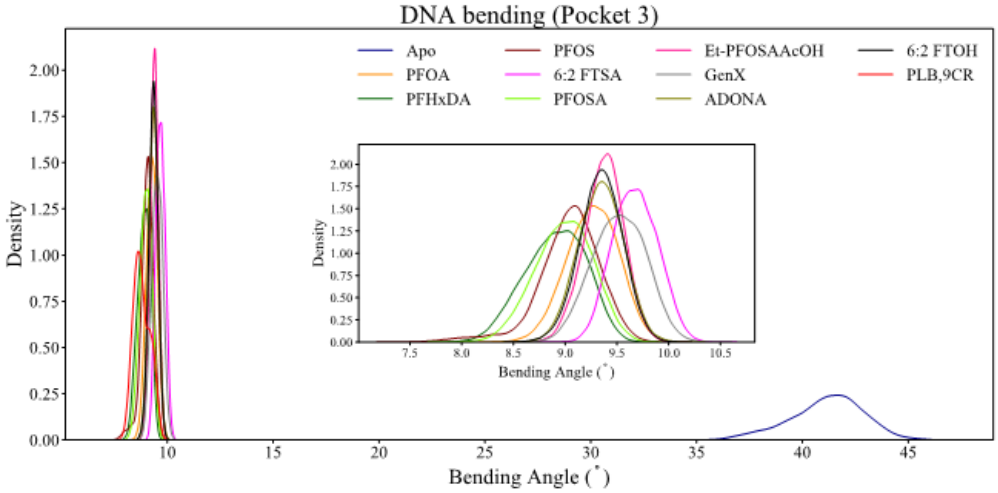


Figure S3.31 Distribution of the DNA bending angle in the Pocket 3 simulations with PFAS. PLB, 9CR: Natural ligands.

CHAPTER 4

INFLUENCE OF PFAS ON HUMAN THYROGLOBULIN PROTEIN: IMPACT ON THYROID HORMONE SYNTHESIS

4.1 Introduction

Environmental pollutants can significantly impact the health of living organisms in the ecosystem and human populations. Some of the most recent health concerns related to environmental pollutants are attributed to per- and polyfluoroalkyl substances (PFAS), a group of man-made chemicals with wide industrial applications due to their unmatched water- and oil-repellent properties as well as heat-resistance.[1–3] There are more than 14,000 compounds listed in the EPA PFASTRUCT database as of June 2023, however, remarkably, only approximately one percent of them have been tested for their toxicities.[EPA,5] PFAS can be found in many products with non-stick and water-repellent surfaces, including food packaging, water-resistant clothing and shoes, and firefighting foams, to provide only a small number of examples, and are often referred to as "forever chemicals" or "zombie chemicals" due to their resistance to degradation. The resistance to degradation, consequently, has resulted in bioaccumulation of PFAS compounds in humans and animals, which has been linked to disruptions of glucose and bile acid metabolisms, immune, reproductive, and thyroid systems, and lipid homeostasis.[6–13] To provide a backdrop for the impact of PFAS on thyroid systems, a functional thyroid gland is crucial for neurodevelopment, cognitive and behavioral growth, as well as for regulating metabolic rate.[11] The synthesis of thyroid hormones, thyroxine (T4) and triiodothyronine (T3) is performed by thyroglobulin, which is a highly conserved protein in vertebrates, and thyroglobulin is located in the lumen of the thyroid follicles.[14] In humans, the thyroglobulin protein – called the human thyroglobulin (hTG) protein - is a homodimer and has four hormonogenic sites (sites A to D as shown in Figure 1) - the four sites where the T4 hormone is produced.[15–17] These sites on hTG are the locations where the thyroid hormones are synthesized. Although the exact mechanism is still not fully understood, the available cryo-EM structures indicate that the orientation of ITY residues as well as neighboring lysine and phenylalanine residues are crucial for the mechanism to take place.[15,17] Current research on the impact of PFAS on thyroid function is mainly based on epidemiological studies and clinical data, with mixed conclusions as to whether PFAS leads to an increase or decrease of the thyroid hormone levels. One study investigating the associations between PFAS exposure during pregnancy and the

neurodevelopment in infants indicated a relationship with PFHxS and PFBS exposure, linking to thyroid hormone-mediated neurodevelopment problems.[12] Prior studies have observed that during pregnancy, there is an association between the maternal levels of thyroid stimulating hormone and the PFHxS, PFNA, and PFOA concentrations.[18–23] Animal studies in rats indicate a decrease in T3 and T4 levels upon PFOA and PFOS exposure,[24?,25] while long-term exposure to PFNA was linked to an increase in T3 levels in zebrafish.[25,26] While there is no single mechanism in which PFAS could disrupt the thyroid system, there are in silico and in vitro studies addressing various potential targets.[27] One study investigated the sodium-iodide symporters for rat and human thyroid cell lines and found that PFOS and PFHxS inhibited this protein.[27?,28] In a number of prior studies, PFAS exposure has been proposed to alter the expression of proteins important for iodide removal and thyroid hormone signaling.[24,29–31] A study of PFAS' effects on the thyroid was performed on common carp fish,[32?] and Manera et al. suggested that the PFOA concentration can cause significant effects on the thyroid follicles of carp by disrupting production as well as reabsorption of thyroglobulin. As the PFAS toxicity on thyroid chemistry is a complicated and mainly uncharted process, the source of the thyroid hormone production, namely hTG, and the influence of PFAS on the thyroid hormone synthesis has been investigated. Understanding how PFAS can impact homeostasis in humans will provide insight towards the development of potential mitigation strategies, such as targeted treatments and interventions for thyroid-related health issues.

4.2 Computational Protocols

The dimeric human hTG protein atomistic structure (PDB ID: 6SCJ, 3.6Å resolution) was obtained from the RSCB Protein Data Bank.[15] The missing loops of the structure were modeled using the I-TASSER server separately by including ten amino acids from each end of the missing loops.[33] The prepared structure was solvated and then minimized using Amber20 as described in the Simulation Details section.

4.2.1 Simulation details

The initial step of this investigation involved selecting a list of carboxylic PFAS (PFCA) and sulphonic PFAS (PFSA) with carbon chains varying from four to twelve, and their structures are

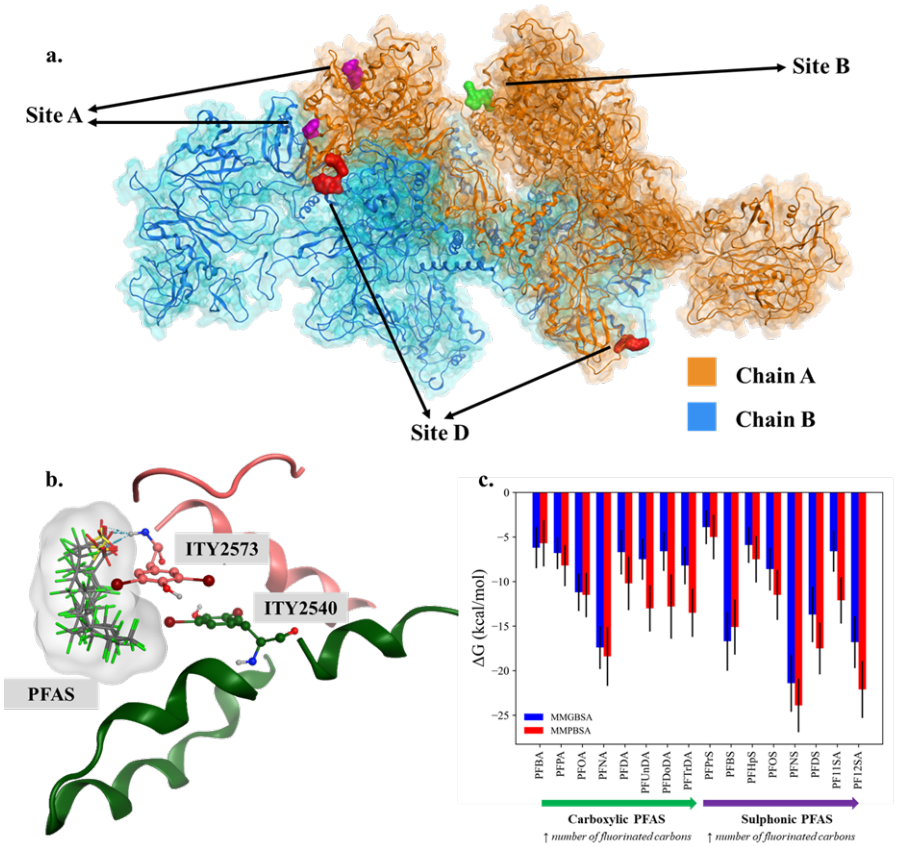


Figure S3.1 (a) The dimeric structure of human thyroglobulin (hTG) and the three hormonogenic sites on Chain A are shown. Among the identified hormogenic sites, Site A has two potential donor residues, and Site D has hormonogenic tyrosines from both chains. (b) The docking poses for PFAS in Site B along with ITY residues. (c) The binding energies for investigated PFAS, calculated with MM-GBSA and MM-PBSA methods are shown, along with the standard deviations. Carboxylic and sulphonic PFAS are listed.

provided in Table S1. Molecular Operating Environment (MOE) software was used for the docking procedures and protonation state determination.[34] The minimized hTG dimer structure was used for the docking procedures. The binding pocket was defined by using a pharmacophore docking strategy to place the PFAS head groups near hormonogenic Tyr residues. The pharmacophore consisted of two features to place the head group. For the short-chain PFAS, a docking procedure with no pharmacophore was also performed. The pharmacophore was used for the initial placement process with a London dG scoring function to obtain 100 poses, which were further refined to five poses with an induced fit method and Generalized Born Volume integral/Weighted Surface Area (GBVI/WSA) scoring function[34] The highest scoring poses for each PFAS were selected

for Molecular Dynamics (MD) simulations. For the docking procedure without pharmacophore placement, Triangle Matcher method was used for the initial PFAS placement. Both monomeric and dimeric structures were considered in the modeling of the binding of PFAS to Site B in hTG protein to understand the effects of the dimer structure. The binding energies of the PFPA and PFBA compounds to the dimer hTG structure is reported in Table S4.10.

The dimer hTG apo, monomeric apo, and PFAS-bound hTG monomeric systems were prepared using the tleap module of Amber20/AmberTools22 software.[35] The partial charges of PFAS compounds and iodinated Tyr residues (ITY) were calculated using the AM1-BCC method as implemented in the antechamber module of AmberTools22 with gaff2.[36,37] The protein, PFAS, and waters were modeled using ff14SB, gaff2, and TIP4P-EW force fields, respectively.[35? –39] NaCl ions were added to each a 0.15M of salt concentration to mimic the natural environment. On average, a monomeric system consisted of 680,000 atoms while a dimeric system had 1,020,000 atoms, including the solvent molecules. The minimization and heating steps were performed in a stepwise fashion as follows: (i) The minimization was done in four steps with the following restraints (100, 50, 10 ,0 kcal mol-1 Å-2), with each step having 20,000 cycles. (ii) The systems were heated up from 0 K to 20 K in 160 ps with a 3 kcal mol-1 Å-2 restraint applied on all atoms. Then, the systems were heated to 200 K for 250 ps with restraints applied to the backbone atoms only, followed by a short equilibrium simulation at 200 K for 200 ps with no restraints. Finally, heating to 300 K was done for 900 ps with no restraints applied. (iii) Before the production step, a 500 ps long equilibrium simulation was performed at 300 K. The minimized and equilibrated structures were simulated for 20 ns at 300 K and 1 atm using 1 fs timesteps with the SHAKE algorithm.[40] A duplicate set of simulations was performed by reinitializing the velocities after the heating step. The Langevin thermostat and isotropic position scaling were selected for the temperature and pressure controls, respectively.[41,42] All simulations were performed using the pmemd.cuda module as implemented in Amber20 suite[35,43]

4.2.1.1 Analysis

The binding energies were calculated by selecting every tenth frame for the last 5 ns of simulations, resulting in a total of 500 frames for a single simulation. The Molecular Mechanics-Poisson Boltzmann Surface Area/Generalized Born Surface Area (MM-PBSA/GBSA) method was used to estimate the binding strengths of the PFAS, as implemented in Amber20/AmberTools22.[44] As the focus of this work is not the exact estimation of the binding energies, but rather to provide a ranking of the binding strengths, MM-PBSA/GBSA methodology is useful in providing insight about binding pockets with partial solvent exposure.[45–50] The root-mean-square distances (RMSD), per-residue root-mean-square fluctuations (RMSF), and hydrogen bond analysis were calculated using the cpptraj module from AmberTools22 with the default parameters.[50] The per-residue decomposition energies were calculated by taking the non-bonded interactions into account for the residues within 10 Å of the PFASs. Clustering was performed to obtain the most dominant orientations of the ITY residues and PFAS using a hierarchical agglomerative algorithm with epsilon value of 3.0. The total energy convergence, as well as the structural convergence of the simulated systems were considered, and the last 5 ns of the trajectories were utilized for all analysis. Clustering of the trajectories was performed using dbscan (Density-Based Spatial Clustering of Applications with Noise) method as implemented in the cpptraj module, using six minimum samples and an epsilon value of 2.[50] The number of minimum samples were determined by a k-distribution plot. Only the last 5 ns of the trajectories were considered for this analysis.

4.2.1.2 Tyr orientation

The positions of ITY residues were clustered and the angle and distance between them were measured for the last 5 ns of each simulation, as shown in Figure S4.5. The distance between the center-of-mass of the side chain atoms of ITY residues and the distance between the reactive atoms were measured. To measure the angle between the ITY residues, a plane for each ITY residue was described by two vectors: each extending from the CG atom to the iodine atoms (Figure S4.5). Then, the angle between the two planes was calculated to estimate the relative orientations of ITY residues.

4.3 Results and Discussion

4.3.1 PFAS Binding

The location of Sites A, B, and D, and the docking poses of PFAS are shown; all of the functional groups that point towards the selected PFAS ITY2573 residue are shown in Figure 1(a). Site B of the hTG protein was selected for the suitability of the initial positioning of tyrosine residues, as Site A and Site D have either two donor ITY residues or have tyrosine residues from different chains. As the hTG monomer is a large protein with 2,700 residues, the RMSDs and total energies were calculated for the whole simulation length to assess if the simulated systems converged structurally and energetically (Figure S4.9-4.10 for RMSD, Figure S4.11-4.12 for total energies). For the majority of the hTG simulations, the total energy reached a plateau within the last 5 ns of the simulations as well as the RMSD time-series reported in the SI; hence, the last 5 ns of the PFAS systems were considered for analysis. Both PFAS and binding site showed no significant conformation change during this simulation period. The binding energies for each hTG/PFAS complex were calculated using end-point methods (MM-GBSA/PBSA) to estimate the relative binding strength of carboxylic and sulphonic PFAS with various fluorinated carbon chain lengths, as per Figure 1(c). Current literature indicates that the thyroid hormone synthesis in hTG can be affected negatively by the exposure to PFAS with eight to nine fluorinated carbons.[18–23] In our simulations, carboxylic PFAS showed an increase in binding strengths as the fluorinated carbons increased from PFBA to PFNA. However, this observation was different for PFCA with more than nine carbons. For PFDA, PFUnDA, PFDoDA, and PFTrDA, the binding energies were -10 and -13 kcal/mol (MM-PBSA). The binding energy analyses of PFSA are quite different than for PFCA. Among the investigated sulphonic PFAS, the strongest binding energy was observed for PFNS. Interestingly, PFBS was also among the strong binders, which was previously noted by Yao et al.[12] The binding strengths of longer-chain PFSA were also higher than their PFCA counterparts with the same fluorinated carbon chain length. These differences in binding strengths indicate that PFCA and PFSA compounds have different impacts on the binding site, and consequently, to the thyroid hormone synthesis.

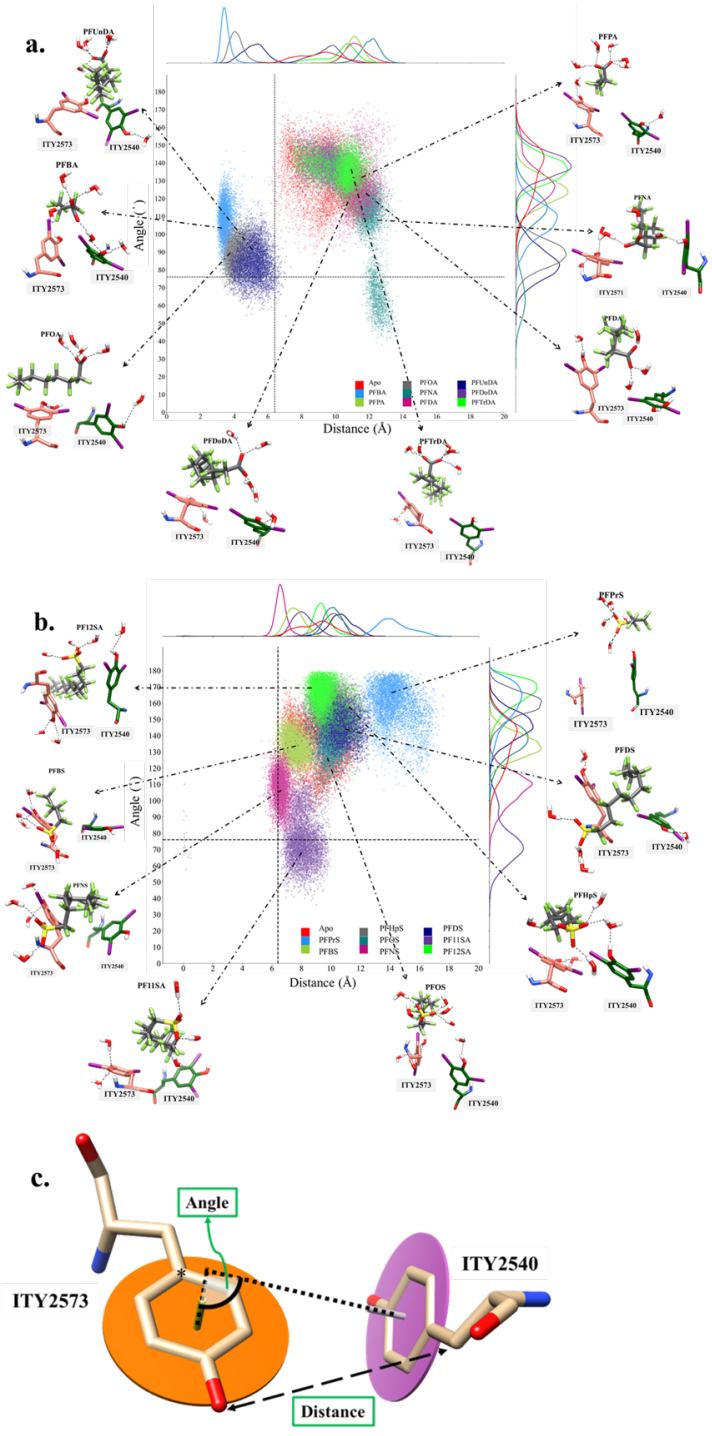


Figure S3.2 The angle/distance distribution of ITY side chains when PFCA (a) or PFSA (b) is present in the pocket. The angles are calculated between the normal vectors of the planes, as described in (c) and Figure S4.5. The most dominant orientations for each PFAS are also shown. The horizontal and vertical line intersection indicates the angle/distance calculated from the cryo-EM orientation (6.4 Å and 76°).

Residue decomposition can help understand some of the energetic differences observed in Figure 1a), so they were calculated for each simulation of PFCA and PFSA compounds with the binding pocket residues, as shown in Table 1, and Table S2-S5. The pocket residues are separated into four groups based on their polarity and acidity: polar, non-polar, basic, and acidic residues. The strengths of the electrostatic interactions and van der Waals interactions made by PFAS and residues within 10 Å radius suggest that the basic residues showed the strongest interaction among all, specifically the K2536 had the highest interaction energy with all of the PFAS. The acidic residues mainly had weak and non-stabilizing interactions, with values larger than zero. In general, electrostatic interactions with charged residues had stronger interactions with PFCA, while PFSA molecules interacted with polar and non-polar residues in the binding site through the C-F tail. Considering the fact that PFSA showed slightly higher MM-PBSA energies, the tail group of PFAS provides stronger anchoring to surrounding residues than the head groups of PFAS. The contributions from the ITY residues were also identified, as they play a pivotal role in thyroid hormone synthesis. PFAS primarily formed stabilizing interactions with ITY residues, although these interactions were weaker than those with charged residues. Among the PFAS with highest ITY interaction energies, PFNA, PFNS, PF12SA also exhibited a high MM-PBSA energies, suggesting that ITY interactions could be the determining factor in predicting PFAS binding to Site B. The interactions between the PFAS and ITY were established between the diiodotyrosine side chains and the C-F tails (Figure S4.7), and the total contribution from ITY residues increased as the fluorinated carbon chain length increased in PFCA molecules, with the exception of PFNA and PFTrDA. This finding provides further evidence supporting the crucial role of the C-F tail in stabilizing the PFAS binding. The hydrogen bond interactions that were formed by PFAS during the simulations were also investigated and are reported in Table S6. PFAS with 8 to 10 carbons predominantly formed direct hydrogen bonds, and as the chain length increased or decreased, the number of hydrogen bonds formed by PFAS with the protein decreased. During the simulations, PFDS exhibited the highest number of interactions with pocket residues, followed by PFNS and PFBS. All three compounds formed hydrogen bonds with S430, Q431, and ITY2573 residues, which also had high interaction energies with PFDS,

PFNS, and PFBS. Among the PFCA compounds, the highest number of interactions were observed for PFNA. These observations further support the notion that for PFAS with certain carbon chain length, the local interactions made with the head group and, more importantly, through the C-F tail are important determinants of being the strong binders, compared to other PFAS. Furthermore, the strong interactions between C-F tail of PFSA compounds and polar/nonpolar residues allowed sulphonic PFAS to have higher binding energies than PFCA.

4.3.2 Changes in local interaction patterns upon PFAS binding

Understanding how the presence of PFAS in the selected positions of the thyroglobulin protein change structural interactions, the residues located nearby PFAS were divided into three regions: Regions 1, 2, and 3 (Figure S4.4). Region 1 has a loop secondary structure, and Regions 2 and 3 have alpha helix structures, and the calculated hydrogen bond percentages are reported Table S4.7-S4.9. In Region 1, the interactions observed in the presence of PFAS were not significantly different: the apo system has interactions that were not observed in PFAS bound cases, or vice versa. For this region, the interactions did not show a distinct pattern either for head group type or the carbon chain length. On the other hand, the interactions within Region 2, clearly showed a noteworthy pattern: as the fluorinated carbon chain length increased, the number of interactions observed within the binding site increased (Table S4.8). The highest interaction percentage in apo simulations was for the S2534/A2538 residue pair, which is part of the alpha helix in Region 2, with the interaction occurring through their backbone atoms. S2534/A2538 interaction persisted in the majority of PFAS simulations, with the exception of the PFOA, PFNA, PFOS, and PFDA simulations. The rest of the dominant interactions in the apo system persisted for 25 to 35% of the simulation and were observed in the majority of PFAS simulations as well. The ITY2540/K2536 interaction persisted in 25% of the simulations in the apo system, and the interaction percentage increased as the carbon chain length of PFAS increased. A higher number of interactions among the residues in Region 2 results in a more stabilized helix-loop-helix structure in the presence of longer PFAS only. The importance of Region 2 for the thyroid hormone synthesis was observed in a recent study where the crystal structure of bovine TG was obtained after the formation of

T4 hormone.[17] Upon comparison of hTG and bovine TG with T4, one significant difference was observed for Region 2: to allow for T4 formation, Region 2 was shifted and three residues from the helix were unfolded and become part of the loop: S2534, S2535, and K2536. These are the residues that formed new hydrogen bonds in the presence of longer chain PFAS; in essence, the binding of longer chain PFAS triggers the formation of more interactions within Region 2, making it more rigid. Hence, by preventing the required flexibility, PFAS would be able to interfere with thyroid hormone synthesis. The interaction pattern observed for Region 3 is similar for Region 1. While the interactions observed in the apo system were protected in most PFAS-bound simulations, the percentages were generally higher in the presence of PFAS (Table S4.9). In general, however, the hydrogen bond percentages did not show significant interaction differences between the Apo system and PFAS simulations within this region. One interesting observation for Region 3 is that the orientations of PFSA compounds were usually towards the residues within this region (Figure S4.7), however, PFCA compounds showed preferences towards the Lysine residues in Region 2. This orientation preference, as will be explained in the following section, results in a characteristic distribution of distance and angles between ITY residues in the presence of PFCA and PFSA (Figure S4.5).

4.3.3 Impact of PFAS Binding on ITY orientations

As the disturbance of thyroid hormone levels has been identified as one of the health consequences of PFAS exposure, understanding how the presence of PFAS could affect the thyroid hormone synthesis in the investigated hormonogenic site is fundamental. The proposed mechanism for T4 synthesis indicates that the acceptor and donor ITY residues should be within 6 Å distance and nearly be parallel to one another, based on the available cryo-EM structures of hTG 15. While the angle between the ITY planes provides insight about the respective positioning of the side chains of these hormonogenic residues, the distances between the donor and acceptor atoms are also an important feature in assessing the thyroid hormone formation. Therefore, the distance between the oxygen from the donor ITY2540 and the carbon from the acceptor ITY2573 were tracked for all simulations. The angle between the ITY side chains was also calculated, and their distributions as

well as the dominant orientations of residues are shown in Figure 2 The angle/distance distribution plots show that the PFCA and PFSA compounds impact the ITY orientation. The ideal positioning of ITY residues in Site B which would allow for the formation of T3 and T4 hormones have a 76° angle and 6 Å distance, based on the available cryo-EM structure of human Thyroglobulin (Figure 2). The presence of PFAS generally limited the conformational space of ITY residues, in terms of the distance and angle tracked here. The apo system has a single peak at 145° along with a shoulder at 120° with a wide distribution. The distance range of the apo simulation was observed to be between 6-14 Å. PFCAs had a broader distance distribution (3-14 Å), while PFSA compounds displayed a narrower one, around 6 to 12 Å, with the exception of PFPrS. The correlation between the fluorinated carbon chain length and angle, however, shows different preferences between PFCA and PFSA compounds. The smallest angle in the distribution observed for PFCA was 60° (small peak of PFNA), and it was 70° for PFSA (PF11SA). On the other hand, the largest angles observed were for PFDoDA and PFTrDA (140°), and PFPrS, PFHpS, and PF12SA (170°) among the PFSA compounds. Overall, the smallest angle/distance distribution among PFCA was observed for PFBA, PFOA, and PFUnDA, while among PFSA, it was PFBS, PFNS, and PF11SA. The two clusters formed by PFCA compounds (Fig. 3) can be distinguished by the distance threshold of 6 Å. Only three PFCA compounds had distances smaller than 6 Å: PFUnDA, PFBA, and PFOA. However, only in PFBA bound simulations, which is a weak binder, ITY residues show distance/angle distribution that would allow for the formation of thyroid hormones. On the other hand, PFOA has an average binding strength, as per MM-PBSA energies, and it has strong interactions with the ITY2573 residue. Similarly, PFUnDA has strong binding energy and strong interactions with ITY2540. The strong interactions with ITY residues could prevent them from forming T3 and T4 thyroid hormones. The other cluster seen in Figure 2(a) has large distance (8-14 Å) and angle (100-140°). PFNA, among those compounds, showed a strong peak around 120° with a smaller peak around 70°. The interesting fact about PFNA interactions that played a role in the bimodal distribution is the stronger interactions with ITY2540, instead of ITY2573, as mentioned before (Table S4.2). The interaction preference also contributes to the ‘sandwiching’

Table S3.1 The sum of per-residue decomposition energies for charged residues and polar & non-polar residues (in kcal mol $^{-1}$).

	PFBA	PFPA	PFOA	PFNA	PFDA	PFUnDA	PFDoDA	PFTrDA
Charged Res.	-49.45	-26.79	-20.23	-46.05	-33.59	-27.55	-60.72	-2.16
Polar and Non-polar Res.	-14.05	-27.67	-32.90	-44.80	-33.93	-33.88	-42.54	-34.51
Sum	-63.50	-54.46	-53.13	-90.84	-67.52	-61.43	-103.26	-36.68
	PFPrS	**PFBS**	**PFHpS**	**PFOS**	**PFNS**	**PFDS**	**PF11SA**	**PF12SA**
Charged Res.	-29.38	-7.53	-29.40	-11.94	14.92	-19.83	-30.31	-32.76
Polar and Non-polar Res.	-13.13	-57.99	-33.03	-44.97	-75.64	-50.17	-32.79	-49.90
Sum	-42.51	-65.52	-62.43	-56.90	-60.72	-70.00	-63.10	-82.65

behavior that was seen in PFNA simulations, where PFAS places itself in-between two ITY residues (Fig. S4.7). Furthermore, the stronger MM-PBSA binding energy of PFNA can also be attributed to sandwiching interaction. Other PFCA has mainly stronger interaction with ITY2573 through their tail groups, and do not show 'sandwiching' behavior. Among the PFSA species, PFNS and PF12SA had a similar interaction type where the intercalation between ITY residues happened (Fig. S4.7). In this case, however, PFNS exhibited strong interaction with ITY2573, while PF12SA had interactions with both ITY, with comparable strengths (Table S4.77). Many of the PFAS bound systems did not show any distances and/or angles closer to those observed in cryo-EM structure, except for PFUnDA, which had a 80° angle and 5 Å distance. For the PFSA, no system had values close to those of the experimental structure, indicating that the presence of various PFAS near hormonogenic site B can prevent the conformational space that would allow the formation of thyroid hormones. Furthermore, based on our analysis of the investigated PFAS-bound systems, the degree in which the PFAS can impact this conformational space depends on (i) the interaction mode of PFAS with the surrounding residues, including ITYs, (ii) the length of the tail group of PFAS, and (iii) the hydrogen bond interactions of head group of PFAS.

The binding energies indicate that PFSA molecules have stronger interactions with the investigated site than with PFCA compounds, as shown in Figure 1(c). Furthermore, there is a chain-length dependent impact on the binding strength, although this dependence is not completely linear. As the chain length increased from three to eight or nine fluorinated carbons (PFNA and PFNS, respec-

tively), the binding energies showed a linear increase. And as the chain gets longer than eight or nine carbons, however, there is a drop in the binding strength, indicating that PFAS with eight and nine carbons can impact the hTG Site B by binding more strongly than shorter chain PFAS and forming key interactions with surrounding residues.[7,12,18,19,31] A 2023 study by Vollmar et al. suggests that PFOS and PFOA have the potential to disrupt the T4 levels.[51] Our study for the first time shows that the disruption by PFAS occurs through binding to the hTG protein and, thus, interfering with the thyroid hormone synthesis. The presence of PFAS, overall, causes the conformational space of the distance and angle between two ITY residues to narrow, as compared to the distribution observed for the apo system. While ITY residues do require the Thyroid Peroxidase (TPO) enzyme to form the thyroid hormones through a mechanism that is still unknown, the proximity and relative orientation of ITY residues are still important for successfully producing T3/T4 hormones.[15,16,52] The majority of PFAS-bound systems did not show distance and angle distribution was close to cryo-EM structure. Moreover, the influence of PFCA compounds on the conformational space of ITY residues suggests a wider range of distances compared to of PFSA molecules, pointing that these two series formed interactions with the different residues. While PFCA head groups prefer to orient towards ITY2540, PFSA compounds pointed towards the loop structure near binding site. The interactions of PFAS also impacted the distance and angle distributions significantly. PFNA and PFNS, for instance, showed a particular 'sandwiching' behavior between two ITY residues, that was not observed for any other PFAS investigated. These two PFAS also had strong hydrogen bonds with the surrounding residues. Together, these different types of interactions lead them to have strong binding energies, and consequently, have more pronounced adverse effects on thyroid hormone synthesis on Site B. The local interaction changes within the binding area indicate that the longer chain PFAS could lead to more rigid helix structure in Region 2. A comparison with a recent cryo-EM structure of the bovine TG with T4 formed in Site B shows that there is a shift in Region 2 associated with the formation of the thyroid hormone.[18] Therefore, for the first time, we suggest that the changes to the hydrogen bond network within Region 2 upon long-chain PFAS binding could inhibit the required motion for the formation of thyroid hormones. As the linkages

between the PFAS exposure and health problems are increasing and the governments in both the United States and the European Union are proposing restrictions on PFAS production due to these adverse health effects, a detailed molecular understanding of PFAS toxicity through computational methods is necessary to establish effective mitigation strategies. To the best of our knowledge, this work is the first of its kind to investigate the influence of PFAS binding to Site B of hTG and the potential impact of PFAS on thyroid hormone synthesis by causing rigidity in binding region. We observed that PFAS with eight to nine carbons with a distinct binding mode showed higher binding energies. The longer chain PFAS, on the other hand, resulted in a change in the rigidity of Region 2, which is important for thyroid hormone synthesis. Understanding these governing factors of PFAS toxicity on thyroid hormone synthesis would help enable the development of effective mitigation strategies and understand harmful influences of PFAS in humans better.

BIBLIOGRAPHY

[1] Sajid, M. and Ilyas, M. (2017). Ptfe-coated non-stick cookware and toxicity concerns: a perspective. *Environmental Science and Pollution Research*, 24:23436–23440.

[2] Schaider, L. A., Balan, S. A., Blum, A., Andrews, D. Q., Strynar, M. J., Dickinson, M. E., Lunderberg, D. M., Lang, J. R., and Peaslee, G. F. (2017). Fluorinated compounds in u.s. fast food packaging. *Environmental Science & Technology Letters*, 4:105–111.

[3] Rao, N. S. and Baker, B. E. (1994). *Textile finishes and fluorosurfactants*, pages 321–338. Springer US.

[EPA] Us environmental protection agency epa's per- and polyfluoroalkyl substances (pfas) action plan 2019 no. february.

[5] Houck, K. A., Patlewicz, G., Richard, A. M., Williams, A. J., Shobair, M. A., Smeltz, M., Clifton, M. S., Wetmore, B., Medvedev, A., and Makarov, S. (2021). Bioactivity profiling of per- and polyfluoroalkyl substances (pfas) identifies potential toxicity pathways related to molecular structure. *Toxicology*, 457:152789.

[6] Sunderland, E. M., Hu, X. C., Dassuncao, C., Tokranov, A. K., Wagner, C. C., and Allen, J. G. (2019). A review of the pathways of human exposure to poly- and perfluoroalkyl substances (pfass) and present understanding of health effects. *Journal of Exposure Science & Environmental Epidemiology*, 29:131–147.

[7] Rappazzo, K., Coffman, E., and Hines, E. (2017). Exposure to perfluorinated alkyl substances and health outcomes in children: a systematic review of the epidemiologic literature. *International Journal of Environmental Research and Public Health*, 14:691.

[8] Duan, X., Sun, W., Sun, H., and Zhang, L. (2021). Perfluorooctane sulfonate continual exposure impairs glucose-stimulated insulin secretion via sirt1-induced upregulation of ucp2 expression. *Environmental Pollution*, 278:116840.

[9] Anderko, L. and Pennea, E. (2020). Exposures to per-and polyfluoroalkyl substances (pfas): potential risks to reproductive and children's health. *Current Problems in Pediatric and Adolescent Health Care*, 50:100760.

[10] Guo, H., Chen, J., Zhang, H., Yao, J., Sheng, N., Li, Q., Guo, Y., Wu, C., Xie, W., and Dai, J. (2022). Exposure to genx and its novel analogs disrupts hepatic bile acid metabolism in male mice. *Environmental Science & Technology*, 56:6133–6143.

[11] Coperchini, F., Croce, L., Ricci, G., Magri, F., Rotondi, M., Imbriani, M., and Chiovato, L. (2021). Thyroid disrupting effects of old and new generation pfas. *Frontiers in Endocrinology*, 11:1077.

[12] Yao, Q., Vinturache, A., Lei, X., Wang, Z., Pan, C., Shi, R., Yuan, T., Gao, Y., and Tian, Y. (2022). Prenatal exposure to per- and polyfluoroalkyl substances, fetal thyroid hormones, and infant neurodevelopment. *Environmental Research*, 206:112561.

[13] Byrne, S. C., Miller, P., Seguinot-Medina, S., Waghiyi, V., Buck, C. L., von Hippel, F. A., and Carpenter, D. O. (2018). Exposure to perfluoroalkyl substances and associations with serum thyroid hormones in a remote population of alaska natives. *Environmental Research*, 166:537–543.

[14] Luo, Y., Ishido, Y., Hiroi, N., Ishii, N., and Suzuki, K. (2014). The emerging roles of thyroglobulin. *Advances in Endocrinology*, 2014:1–7.

[15] Coscia, F., Taler-Verčič, A., Chang, V. T., Sinn, L., O'Reilly, F. J., Izoré, T., Renko, M., Berger, I., Rappsilber, J., Turk, D., and Löwe, J. (2020). The structure of human thyroglobulin. *Nature*, 578:627–630.

[16] ul Kim, H., Jeong, H., Chung, J. M., Jeoung, D., Hyun, J., and Jung, H. S. (2022). Comparative analysis of human and bovine thyroglobulin structures. *Journal of Analytical Science and Technology*, 13:1–8.

[17] Marechal, N., Serrano, B. P., Zhang, X., and Weitz, C. J. (2022). Formation of thyroid hormone revealed by a cryo-em structure of native bovine thyroglobulin. *Nature Communications*, 13:1–7.

[18] Wang, Y., Rogan, W. J., Chen, P. C., Lien, G. W., Chen, H. Y., Tseng, Y. C., Longnecker, M. P., and Wang, S. L. (2014). Association between maternal serum perfluoroalkyl substances during pregnancy and maternal and cord thyroid hormones: Taiwan maternal and infant cohort study. *Environmental Health Perspectives*, 122:529–534.

[19] Webster, G. M., Venners, S. A., Mattman, A., and Martin, J. W. (2014). Associations between perfluoroalkyl acids (pfass) and maternal thyroid hormones in early pregnancy: a population-based cohort study. *Environmental Research*, 133:338–347.

[20] Lewis, R. C., Johns, L. E., and Meeker, J. D. (2015). Serum biomarkers of exposure to perfluoroalkyl substances in relation to serum testosterone and measures of thyroid function among adults and adolescents from nhanes 2011–2012. *International Journal of Environmental Research and Public Health 2015, Vol. 12, Pages 6098-6114*, 12:6098–6114.

[21] Lopez-Espinosa, M. J., Fitz-Simon, N., Bloom, M. S., Calafat, A. M., and Fletcher, T. (2012). Comparison between free serum thyroxine levels, measured by analog and dialysis methods, in the presence of perfluorooctane sulfonate and perfluorooctanoate. *Reproductive Toxicology*, 33:552–555.

[22] Wang, Y., Starling, A. P., Haug, L. S., Eggesbo, M., Becher, G., Thomsen, C., Travlos, G., King, D., Hoppin, J. A., Rogan, W. J., and Longnecker, M. P. (2013). Association between perfluoroalkyl substances and thyroid stimulating hormone among pregnant women: a cross-

sectional study. *Environmental Health: A Global Access Science Source*, 12:1–7.

[23] Kim, S., Choi, K., Ji, K., Seo, J., Kho, Y., Park, J., Kim, S., Park, S., Hwang, I., Jeon, J., Yang, H., and Giesy, J. P. (2011). Trans-placental transfer of thirteen perfluorinated compounds and relations with fetal thyroid hormones. *Environmental Science and Technology*, 45:7465–7472.

[24] Yu, W. G., Liu, W., and Jin, Y. H. (2009). Effects of perfluorooctane sulfonate on rat thyroid hormone biosynthesis and metabolism. *Environmental Toxicology and Chemistry*, 28:990–996.

[25] Boas, M., Feldt-Rasmussen, U., and Main, K. M. (2012). Thyroid effects of endocrine disrupting chemicals. *Molecular and Cellular Endocrinology*, 355:240–248.

[26] Liu, Y., Wang, J., Fang, X., Zhang, H., and Dai, J. (2011). The thyroid-disrupting effects of long-term perfluorononanoate exposure on zebrafish (danio rerio). *Ecotoxicology*, 20:47–55.

[27] Buckalew, A. R., Wang, J., Murr, A. S., Deisenroth, C., Stewart, W. M., Stoker, T. E., and Laws, S. C. (2020). Evaluation of potential sodium-iodide symporter (nis) inhibitors using a secondary fischer rat thyroid follicular cell (frtl-5) radioactive iodide uptake (raiu) assay. *Archives of Toxicology*, 94:873–885.

[28] Conti, A., Strazzeri, C., and Rhoden, K. J. (2020). Perfluorooctane sulfonic acid, a persistent organic pollutant, inhibits iodide accumulation by thyroid follicular cells in vitro. *Molecular and Cellular Endocrinology*, 515:110922.

[29] Du, G., Hu, J., Huang, H., Qin, Y., Han, X., Wu, D., Song, L., Xia, Y., and Wang, X. (2013). Perfluorooctane sulfonate (pfos) affects hormone receptor activity, steroidogenesis, and expression of endocrine-related genes in vitro and in vivo. *Environmental Toxicology and Chemistry*, 32:353–360.

[30] Spachmo, B. and Arukwe, A. (2012). Endocrine and developmental effects in atlantic salmon (salmo salar) exposed to perfluorooctane sulfonic or perfluorooctane carboxylic acids. *Aquatic Toxicology*, 108:112–124.

[31] Ballesteros, V., Costa, O., Iñiguez, C., Fletcher, T., Ballester, F., and Lopez-Espinosa, M. J. (2017). Exposure to perfluoroalkyl substances and thyroid function in pregnant women and children: a systematic review of epidemiologic studies. *Environment International*, 99:15–28.

[32] Manera, M., Castaldelli, G., and Giari, L. (2022). Perfluorooctanoic acid affects thyroid follicles in common carp (cyprinus carpio). *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 9049*, 19:9049.

[33] Yang, J. and Zhang, Y. (2015). I-tasser server: new development for protein structure and function predictions. *Nucleic acids research*, 43:W174–W181.

[34] (2022). Molecular operating environment (moe), 2022.02 chemical computing group ulc,

1010 sherbooke st. west, suite 910, montreal, qc, canada, h3a 2r7.

[35] (2020). Amber 2020.

[36] He, X., Man, V. H., Yang, W., Lee, T.-S., and Wang, J. (2020). A fast and high-quality charge model for the next generation general amber force field. *The Journal of Chemical Physics*, 153:114502.

[37] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25:1157–1174.

[38] Dopke, M. F., Moultos, O. A., and Hartkamp, R. (2020). On the transferability of ion parameters to the tip4p/2005 water model using molecular dynamics simulations. *The Journal of Chemical Physics*, 152:024501.

[39] Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *The Journal of Chemical Physics*, 120:9665–9678.

[40] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23:327–341.

[41] Wu, X., Brooks, B. R., and Vanden-Eijnden, E. (2016). Self-guided langevin dynamics via generalized langevin equation. *Journal of Computational Chemistry*, 37:595–601.

[42] Berendsen, H. J., Postma, J. P., Gunsteren, W. F. V., Dinola, A., and Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81:3684–3690.

[43] Mermelstein, D. J., Lin, C., Nelson, G., Kretsch, R., McCammon, J. A., and Walker, R. C. (2018). Fast and flexible gpu accelerated binding free energy calculations within the amber molecular dynamics package. *Journal of Computational Chemistry*, 39:1354–1358.

[44] Onufriev, A., Bashford, D., and Case, D. A. (2004). Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins: Structure, Function and Genetics*, 55:383–394.

[45] Almeida, N. M. S., Eken, Y., and Wilson, A. K. (2021). Binding of per- and polyfluoro-alkyl substances to peroxisome proliferator-activated receptor gamma. *ACS Omega*, 6:15103–15114.

[46] Lai, T. T., Eken, Y., and Wilson, A. K. (2020). Binding of per- and polyfluoroalkyl substances to the human pregnane x receptor. *Environmental Science & Technology*, 54:15986–15995.

[47] Bali, S. K., Marion, A., Ugur, I., Dikmenli, A. K., Catak, S., and Aviyente, V. (2018).

Activity of topotecan toward the dna/topoisomerase i complex: a theoretical rationalization. *Biochemistry*, 57:1542–1551.

[48] Findik, B. K., Cilesiz, U., Bali, S. K., Atilgan, C., Aviyente, V., and Dedeoglu, B. (2022). Investigation of iron release from the n- and c-lobes of human serum transferrin by quantum chemical calculations. *Organic & Biomolecular Chemistry*, 20:8766–8774.

[49] Bali, S. K., Haslak, Z. P., Cifci, G., and Aviyente, V. (2023). Dna preference of indenoisoquinolines: a computational approach. *Organic & Biomolecular Chemistry*, 21:4518–4528.

[50] Roe, D. R. and Cheatham, T. E. (2013). Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9:3084–3095.

[51] Vollmar, A. K. R., Lin, E. Z., Nason, S. L., Santiago, K., Johnson, C. H., Ma, X., Pollitt, K. J. G., and Deziel, N. C. (2023). Per- and polyfluoroalkyl substances (pfas) and thyroid hormone measurements in dried blood spots and neonatal characteristics: a pilot study. *Journal of Exposure Science & Environmental Epidemiology 2023 33:5*, 33:737–747.

[52] Kim, K., Kopylov, M., Bobe, D., Kelley, K., Eng, E. T., Arvan, P., and Clarke, O. B. (2021). The structure of natively iodinated bovine thyroglobulin. *Acta Crystallographica Section D: Structural Biology*, 77:1451–1459.

APPENDIX A

SUPPORTING TABLES

Table S4.1 The list of PFAS used in this work, the total number of fluorinated carbons, and their 2D structures are shown.

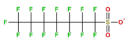
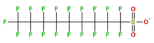
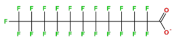
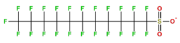
Carboxylic PFAS Name	# of Fluorinated Carbons	Structure	Sulphonic PFAS Name	# of Fluorinated Carbons	Structure
PFBA	3		**PFPrS**	3	
PFPA	4		**PFBS**	4	
PFOA	7		**PFHpS**	7	
PFNA	8		**PFOS**	8	
PFDA	9		**PFNS**	9	
PFUnDA	10		**PFDS**	10	
PFDoDA	11		**PF11SA/PFUnS**	11	
PFTrDA	12		**PF12SA/PFDoS**	12	

Table S4.2 Average per-residue decomposition energies for each PFAS simulation of non-polar pocket residues along with ITY residues.

PDB	PFBA	PFPA	PFOA	PFNA	PFDA	PFUnDA	PFDoDA	PFTrDA	PFPrS	PFBS	PFHpS	PFOS	PFNS	PFDS	PF11SA	PF12SA
P677	-0.99	-3.53	-2.35	-6.45	-3.80	-3.84	-5.89	-2.89	-2.01	-4.38	-5.37	-3.45	-4.04	-4.49	-3.90	-6.12
A678	0.98	1.97	1.59	3.86	2.06	1.15	2.80	1.42	1.81	1.63	1.98	0.48	1.03	1.89	1.53	3.89
I2517	0.26	-0.32	-0.45	-0.44	-0.18	0.00	-0.17	0.03	0.14	-0.18	-0.01	-0.03	-0.31	-0.22	-0.20	-0.47
A2538	-0.60	-0.77	-0.18	-0.24	-0.72	-0.51	-0.56	-0.41	-0.22	-0.38	-0.61	-0.67	-0.17	-0.54	-0.55	-0.52
F2539	-0.89	-1.18	-0.66	-1.18	-1.24	-1.57	-1.35	-0.66	-0.62	-0.89	-1.30	-0.93	-0.77	-0.96	-1.00	-1.41
ITY2540	-1.94	-1.00	-4.33	-11.25	-4.44	-5.26	-12.78	-2.70	-1.84	-2.39	-4.23	-3.29	-5.91	-2.99	-4.78	-11.40
A2542	-0.57	-0.64	-0.41	-0.76	-0.64	-0.85	-0.71	-0.40	-0.57	-0.53	-0.65	-0.55	-0.42	-0.51	-0.52	-0.82
L2543	-0.65	-0.82	-1.17	-1.60	-0.89	-1.29	-1.03	-0.56	-0.76	-0.80	-0.93	-0.74	-0.82	-0.80	-0.80	-1.43
ITY2573	-1.68	-8.67	-5.09	-8.93	-8.46	-10.96	-9.30	-8.82	-3.19	-11.16	-12.17	-5.95	-16.59	-12.68	-8.26	-10.26
A2574	-1.45	-1.45	0.10	-0.16	-1.16	-1.34	-1.51	-3.45	-1.02	-7.97	-2.24	-3.37	-9.42	-5.27	-1.41	0.49
F2576	-0.66	-1.57	-2.24	-4.16	-1.47	-1.47	-2.37	-1.03	-0.72	-2.37	-2.12	-1.32	-2.52	-2.03	-1.87	-3.71

Table S4.3 Average per-residue decomposition energies for each PFAS simulation of polar pocket residues.

PDB	PFBA	PFPA	PFOA	PFNA	PFDA	PFUnDA	PFDoDA	PFTrDA	PFPrS	PFBS	PFHpS	PFOS	PFNS	PFDS	PF11SA	PF12SA
Q429	0.23	-0.06	0.23	0.64	0.11	-1.91	1.63	1.55	0.24	-0.57	0.21	0.89	2.36	1.28	1.47	1.49
S430	-0.75	-0.81	-0.95	-0.69	-2.94	-1.27	-1.57	-3.72	-0.18	-11.48	-0.87	-1.61	-11.88	-2.10	-1.02	-1.48
Q431	-0.44	-1.84	-0.37	-0.75	-3.72	-0.39	-4.10	-1.43	-0.53	-4.63	-0.68	-6.29	-12.75	-6.59	-2.16	-0.51
Q432	0.07	-3.62	-1.39	-0.91	-1.43	-1.03	-0.34	-3.38	-0.43	-7.32	-1.44	-7.26	-6.11	-6.04	0.07	-0.26
S675	0.84	2.19	0.50	1.00	1.24	-0.29	2.36	1.54	0.99	3.61	2.45	2.08	2.47	2.68	2.93	1.02
Q676	-0.78	-1.15	-0.40	-0.51	-1.30	-0.39	-1.21	-4.47	-1.07	-0.18	-0.16	-5.36	-2.06	-1.84	-4.36	-1.42
S2529	-1.37	-0.24	-0.26	-0.58	-0.63	-0.31	-0.54	0.20	-0.62	-0.25	-0.19	-0.32	-0.14	-0.34	-0.38	-0.65
T2537	-1.37	-2.77	-2.42	-1.29	-2.24	0.02	-0.92	-2.04	-1.59	-1.07	-1.77	-2.28	-1.17	-2.08	-2.91	-2.19
Q2541	-1.08	-0.46	-3.37	-2.59	-0.28	-0.51	-0.25	-0.65	-0.17	0.04	-0.46	-0.89	-0.45	-0.24	-0.46	-2.53
Q2544	-0.05	-0.10	-2.15	-0.50	-0.11	-0.19	-0.17	0.11	0.04	-0.54	0.03	0.15	-0.16	0.29	0.17	-1.72
N2545	-0.73	-0.57	-1.38	-0.88	-0.43	-0.79	-0.40	-0.28	-0.54	-0.45	-0.55	-0.55	-0.29	-0.37	-0.39	-0.66
S2569	0.65	1.84	-3.92	-0.23	1.23	1.29	0.73	-0.24	0.54	0.04	1.83	-0.85	0.73	-1.12	0.25	0.17
T2570	0.59	0.46	-0.22	-0.91	0.18	0.35	0.15	-0.08	0.44	-0.23	0.17	-0.26	-0.03	-0.63	-0.26	-0.17
S2575	-0.67	-1.00	-0.63	-0.46	-0.87	-0.78	-1.06	-0.89	-0.49	-2.45	-1.31	-0.97	-3.47	-1.74	-0.38	-0.25
S2577	-1.03	-1.58	-0.97	-4.83	-1.80	-1.71	-3.96	-1.25	-0.75	-3.10	-2.64	-1.65	-2.76	-2.74	-3.59	-8.96

Table S4.4 Average per-residue decomposition energies for each PFAS simulation of basic pocket residues.

PDB	PFBA	PFPA	PFOA	PFNA	PFDA	PFUnDA	PFDoDA	PFTrDA	PFPrS	PFBS	PFHpS	PFOS	PFNS	PFDS	PF11SA	PF12SA
R2530	-46.14	-18.12	-17.90	-14.36	-15.07	-17.41	-16.65	-15.94	-17.38	-14.43	-16.78	-15.12	-18.06	-15.72	-16.93	-16.82
K2534	-26.36	-25.99	-30.52	-48.58	-35.14	-30.31	-49.27	-20.12	-41.65	-31.29	-45.98	-26.50	-25.59	-31.41	-32.77	-36.70
K2536	-44.72	-64.53	-45.77	-57.27	-64.72	-58.71	-75.28	-33.87	-38.89	-43.09	-58.03	-44.36	-33.62	-54.14	-49.82	-59.24
R2578	-17.71	-25.04	-22.09	-30.16	-26.20	-22.38	-30.77	-33.34	-23.77	-36.52	-31.41	-29.83	-30.06	-32.78	-31.91	-29.29

Table S4.5 Average per-residue decomposition energies for each PFAS simulation of acidic pocket residues.

PDB	PFBA	PFPA	PFOA	PFNA	PFDA	PFUnDA	PFDoDA	PFTrDA	PFPrS	PFBS	PFHpS	PFOS	PFNS	PFDS	PF11SA	PF12SA
E2528	30.67	35.79	27.23	32.37	38.26	35.05	41.96	31.23	29.90	27.40	41.10	30.01	26.09	30.73	30.84	34.27
D2571	19.41	25.05	21.93	22.71	22.41	22.61	20.36	24.58	16.61	33.53	24.00	26.26	32.29	31.01	20.79	21.96
D2572	19.71	27.24	24.24	27.77	30.33	24.78	28.56	27.81	24.89	37.32	38.18	29.80	41.30	32.12	29.30	29.67
E2581	15.70	18.60	19.82	23.78	18.86	18.19	21.78	18.47	21.66	19.48	20.15	18.62	18.93	19.76	18.98	23.40

Table S4.6 Hydrogen bond percentages of PFAS head group oxygen atoms. Res. ID: Residue ID of the amino acids.

	Res. ID	PFBA	PFPrS	PFPA	PFBS	PFOA	PFHpS	PFNA	PFOS	PFDA	PFNS	PFUnDA	PFDS	PFDoDA	PF11SA	PFTrDA	PF12SA
	S430	---	---	---	0.46	---	---	---	---	---	0.20	---	0.13	---	---	---	---
	Q431	---	---	---	0.31	---	---	0.10	---	---	0.55	---	0.23	---	---	---	---
	T681	---	0.14	---	---	---	---	---	---	---	---	---	---	---	---	---	---
	K2524	---	---	---	---	---	0.05	---	---	---	---	---	---	---	---	---	---
	R2530	0.5	---	---	---	---	0.05	---	---	---	---	---	---	---	---	---	---
PFAS@O	**K2536**	---	0.07	---	---	0.13	---	0.28	---	0.10	---	---	0.21	---	---	---	---
	ITY2540	---	---	---	---	---	---	---	---	---	---	---	---	0.43	---	---	---
	S2569	---	---	---	---	0.14	---	---	---	---	---	---	---	---	---	---	---
	ITY2573	---	---	---	0.74	---	---	---	0.12	---	0.24	---	0.37	---	---	---	---
	A2574	---	---	---	0.24	---	---	---	---	---	---	---	---	---	---	---	---
	A2575	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.40
PFAS@O1	**K2536**	---	---	0.12	---	---	---	0.08	---	0.19	---	---	---	0.40	---	---	---
	A2574	---	---	---	---	---	---	---	---	---	---	---	0.15	---	---	---	---
PFAS@O2	**A2572**	---	---	---	---	---	---	---	---	---	0.44	---	---	---	---	---	---

Table S4.7 Average hydrogen bond %100 fractions of Region 1. From left to right, the fluorinated carbon chain length increases.

	Residue Pair	APO	PFBA	PFPrS	PFPA	PFBS	PFOA	PFHpS	PFNA	PFOS	PFDA	PFNS	PFUnDA	PFDS	PFDoDA	PF11SA	PFTrDA	PF12SA
Region 1	Q676/R2578	0.75	0.66	0.84	0.21	0.69	0.42	0.40	0.71	0.82	0.63	0.67	0.38	0.63	0.65	0.66	0.70	0.13
	A678/S2577	0.42	0.00	0.17	0.59	0.56	0.55	0.00	0.51	0.54	0.55	0.44	0.40	0.56	0.54	0.61	0.45	0.37
	Q676/S680	0.00	0.00	0.11	0.06	0.32	0.16	0.21	0.00	0.34	0.16	0.28	0.07	0.29	0.57	0.39	0.24	0.00
	P677/S680	0.00	0.41	0.16	0.08	0.00	0.10	0.50	0.33	0.77	0.00	0.34	0.12	0.29	0.00	0.07	0.00	0.11
	G679/K2524	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.38	0.14	0.00	0.00	0.24	0.30
	G674/R2578	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.31	0.42	0.00	0.00
	A678/K2524	0.26	0.00	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.24	0.00	0.00	0.00	0.24	0.00	0.00	0.00
	S680/K2524	0.28	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.11	0.34	0.00	0.00	0.00	0.00	0.00	0.23
	Q676/T681	0.00	0.10	0.06	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	S680/T681	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	T681/L682	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	P677/T681	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table S4.8 Average hydrogen bond %100 fractions of Region 2. From left to right, the fluorinated carbon chain length increases.

	Residue Pair	APO	PFBA	PFPrS	PFPA	PFBS	PFOA	PFHpS	PFNA	PFOS	PFDA	PFNS	PFUnDA	PFDS	PFDoDA	PF11SA	PFTrDA	PF12SA
Region 2	T2537/Q2541	0.35	0.27	0.24	0.19	0.37	0.17	0.34	0.37	0.42	0.15	0.36	0.43	0.38	0.13	0.33	0.34	0.45
	S2535/P2539	0.33	0.10	0.31	0.24	0.24	0.33	0.12	0.50	0.40	0.64	0.36	0.41	0.43	0.43	0.32	0.29	0.33
	ITY2540/K2536	0.25	0.19	0.29	0.29	0.00	0.06	0.03	0.36	0.41	0.57	0.51	0.54	0.36	0.46	0.39	0.52	0.33
	S2534/A2538	0.48	0.37	0.37	0.29	0.35	0.00	0.44	0.00	0.06	0.00	0.28	0.71	0.25	0.10	0.31	0.32	0.18
	G2524/T2533	0.00	0.00	0.11	0.00	0.12	0.22	0.00	0.00	0.06	0.20	0.00	0.29	0.30	0.18	0.33	0.12	0.05
	T2533/T2537	0.27	0.56	1.39	0.12	0.00	0.48	0.00	0.00	0.00	0.00	0.86	0.45	0.00	0.21	0.70	0.75	0.33
	ITY2540/A678	0.00	0.00	0.00	0.62	0.71	0.61	0.00	0.00	0.05	0.54	0.00	0.00	0.00	0.19	0.31	0.37	0.46
	S2534/T2537	0.00	0.00	0.00	0.22	0.33	0.00	0.00	0.00	0.05	0.27	0.00	0.17	0.48	0.00	0.00	0.00	0.23
	R2532/K2536	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.22	0.00	0.00
	R2532/S2535	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.63	0.12	0.57	0.51	0.10	0.26

Table S4.9 Average hydrogen bond %100 fractions of Region 3. From left to right, the fluorinated carbon chain length increases.

	Residue Pair	APO	PFBA	PFPrS	PFPA	PFBS	PFOA	PFHpS	PFNA	PFOS	PFDA	PFNS	PFUnDA	PFDS	PFDoDA	PF11SA	PFTrDA	PF12SA
Region 3	ITY2573/S2577	0.45	0.40	0.60	0.70	0.84	0.70	1.29	0.68	0.34	0.17	0.56	0.77	0.57	0.45	0.42	0.46	0.79
	H2568/D2571	0.44	0.53	0.31	0.36	0.14	0.75	0.81	0.21	0.00	0.53	0.28	0.28	0.00	0.24	0.95	0.00	0.43
	H2568/D2572	0.25	0.68	0.32	0.38	0.33	0.23	0.40	0.19	0.00	0.24	0.73	0.19	0.26	0.29	0.55	0.00	0.18
	A2574/S2577	0.44	0.70	0.00	0.64	0.56	0.81	0.00	0.44	0.85	0.00	0.34	0.82	0.40	0.35	0.44	0.52	0.29
	T2570/D2571	0.33	0.00	0.03	0.12	0.53	0.30	0.00	0.00	0.36	0.28	0.00	0.09	0.39	0.22	0.38	0.67	0.05
	A2574/R2578	0.21	0.14	0.58	0.10	0.00	0.30	0.22	0.08	0.00	0.28	0.00	0.08	0.22	0.06	0.47	0.25	0.00
	D2572/S2575	0.33	0.67	0.99	0.28	0.00	0.77	1.14	0.89	0.00	0.95	0.00	0.46	0.00	1.60	0.23	0.53	0.90
	D2572/P2576	0.39	0.00	0.07	0.47	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.41	0.24	0.00	0.19	0.40	0.00
	D2572/A2574	0.03	0.11	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.49	0.00	0.00	0.00	0.00	0.00	0.00	0.16
	S2569/D2572	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00
	ITY2573/Q2541	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table S4.10 The MM-GBSA/PBSA binding energies of PFBA and PFPA with dimer hTG protein. The compounds did not form strong interactions with the binding site residues and did not reside within the region. These systems were simulated for 10 ns in two different poses, and their average is reported here.

Simulation	MM-PBSA (kcal mol $^{-1}$)		MM-GBSA (kcal mol $^{-1}$)	
	dG	std	dG	std
PFBA (monomer hTG)	-5.93	2.73	-5.30	1.86
PFBA (monomer hTG)	-3.69	2.95	-7.60	2.33
PFBA (dimer hTG)	10.80	5.67	-1.57	3.07
PFPA (dimer hTG)	-7.77	3.19	-9.02	2.31

APPENDIX B

SUPPORTING FIGURES

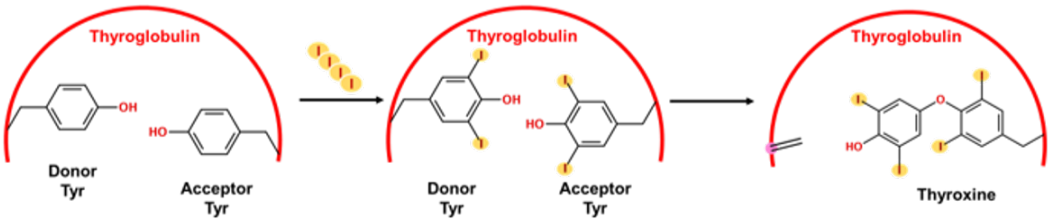


Figure S4.1 Formation of T4 by hormonogenic Tyrosine residues. Iodine is shown with yellow spheres.

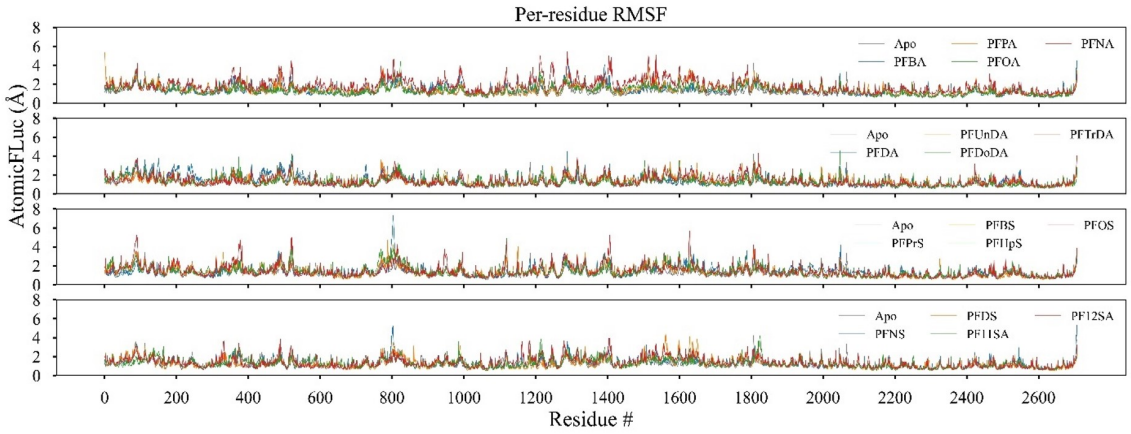


Figure S4.2 Per-residue RMSF plot of first simulation set.

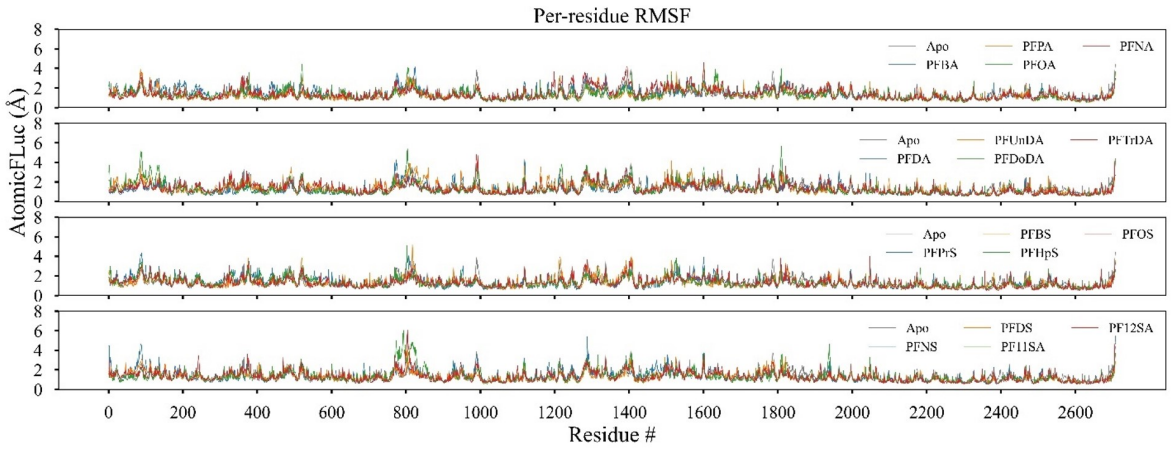


Figure S4.3 Per-residue RMSF plot of first simulation set.

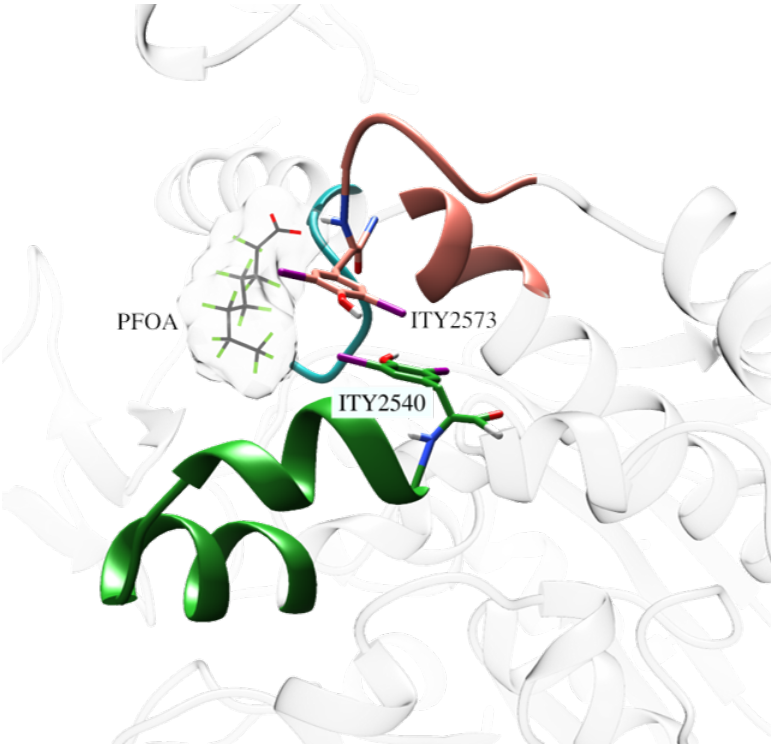


Figure S4.4 The regions for which the hydrogen bond patterns were investigated. Region 1 is shown in blue and includes S675, Q676, P677, A678, G679, and S680 residues. Region 2 is shown in green and includes V2523, K2524, Q2525, F2526, E2527, E2528, S2529, R2530, G2531, R2532, T2533, S2534, S2535, K2536, T2537, A2538, F2539, and ITY2540. Region 3 is depicted in pink loop representation and has the following residues: H2568, S2569, T2570, D2571, D2572, ITY2573, A2574, S2575, F2576, S2577, and R2578. The rest of the Thyroglobulin protein is shown as cartoon in grey color. PFOA is shown in wire representation, and ITY residues are shown in stick representation.

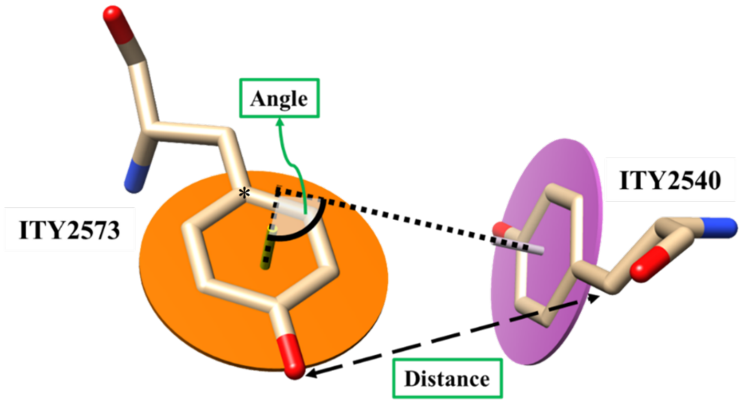


Figure S4.5 Representation of angle and distance measurements between ITY2540 and ITY2573. The residues are shown in stick representation, and the planes, shown as disks, are created by considering the side chain ring atoms. The normal of planes are shown as sticks. The distance between the OH atom of ITY2540 and CB atom of ITY2573 is shown with a dashed black arrow. CG atom is indicated by asterisk (*) on ITY2573, for reference. The structure of ITY residues is taken from the cryo-EM structure (PDB ID: 6SCJ). The distance calculated for the structure is 6.4 and the angle is 76 degrees.

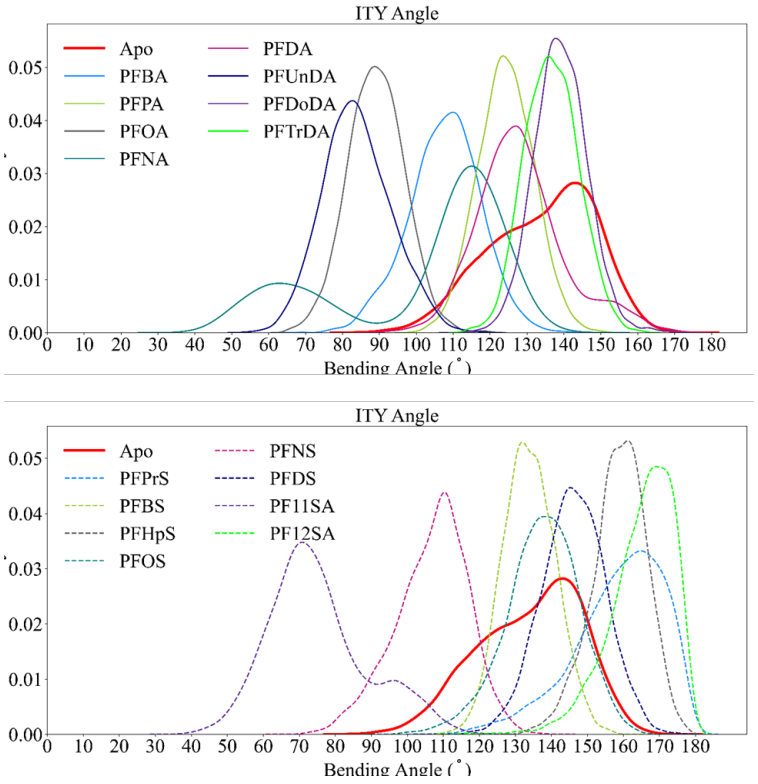


Figure S4.6 The kernel density plot of the distribution of calculated angles between the ITY residues. Above: PFCA compounds, below: PFSA compounds. Apo system is shown in solid red line in both plots.

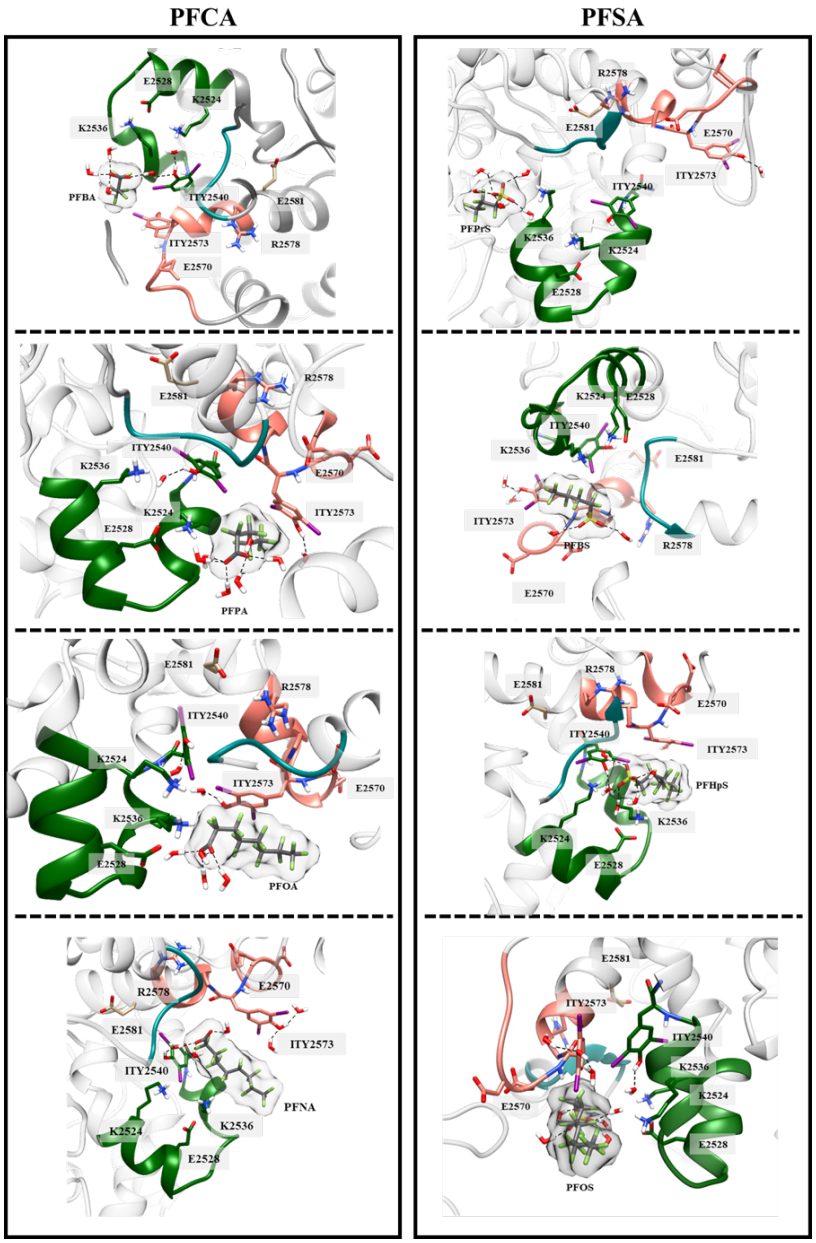
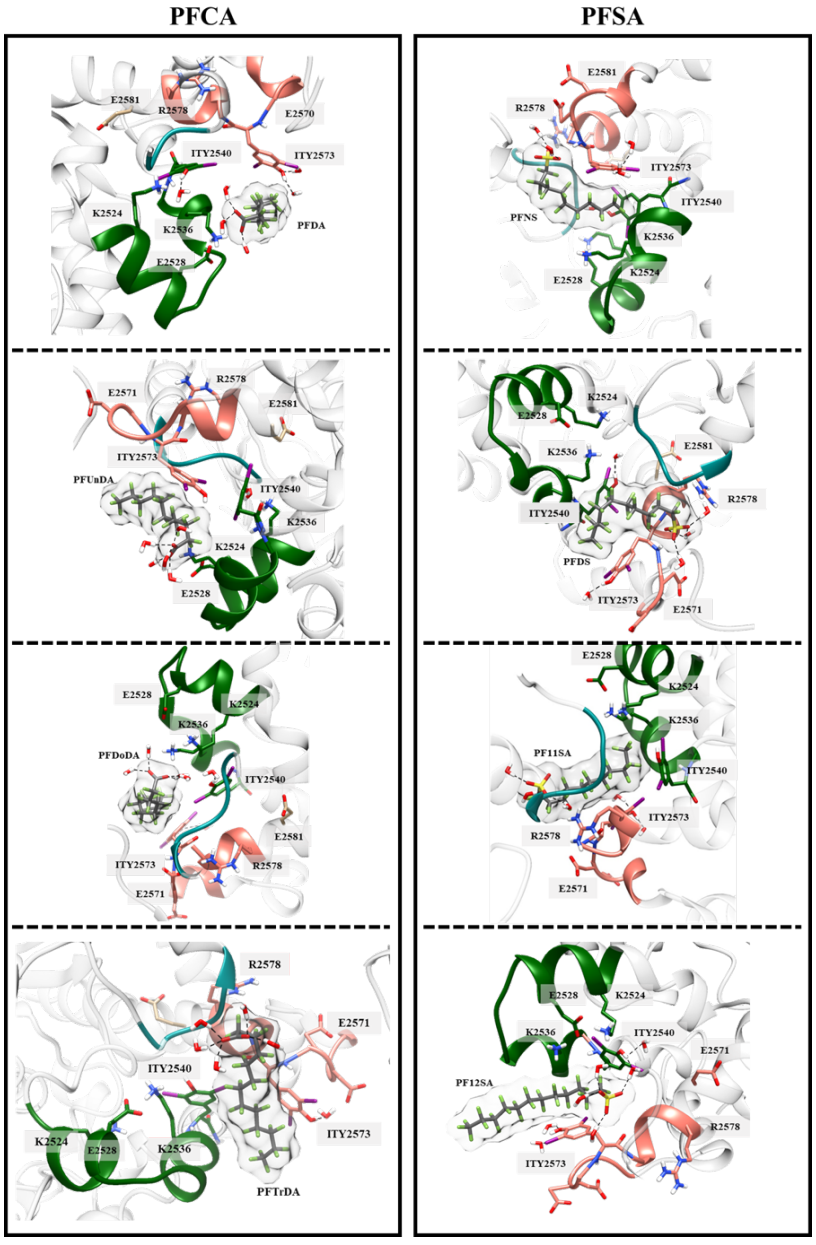


Figure S4.7 The dominant orientations of ITY residues and PFAS compounds, extracted by clustering the last 5ns of the simulations. The key residues that have either the highest/lowest interaction with PFAS or that make hydrogen bonds with PFAS were shown in stick representation. The coloring of the secondary structures was based on the scheme shown in Figure S4.4. PFCA: per-fluoroalkyl carboxylic acid, PFSA: per-fluoroalkyl sulphonic acid. The figure is obtained using Chimera.

Figure S4.7 (cont'd)



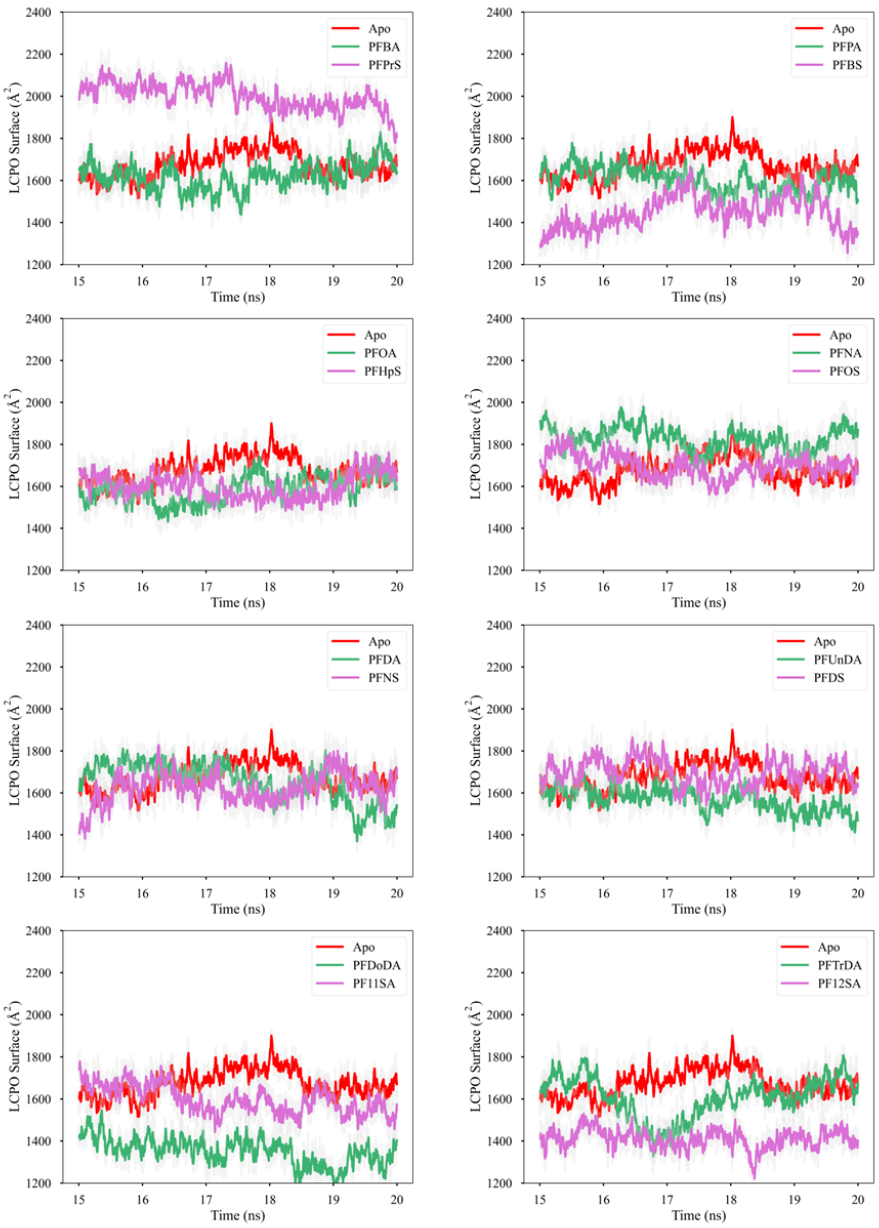


Figure S4.8 Surface Area calculations of ITY residues in the presence of PFAS.

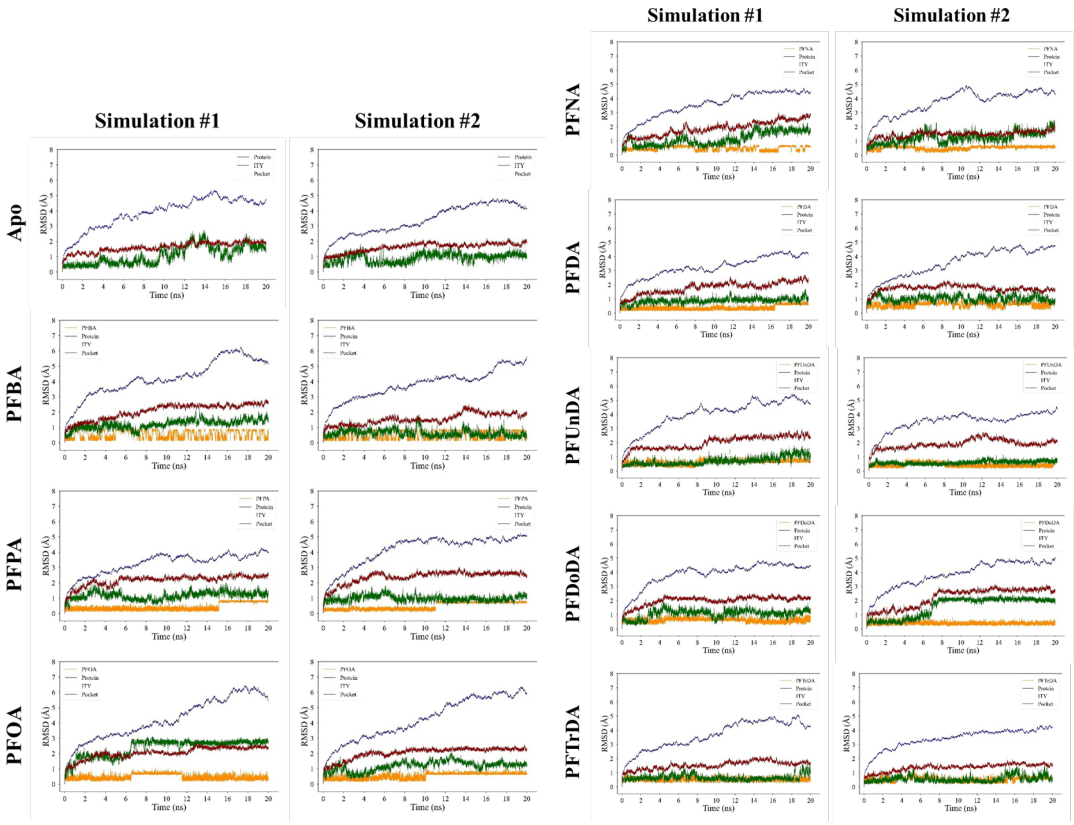


Figure S4.9 The RMSD time-series of apo system and carboxylic acid PFAS simulations.

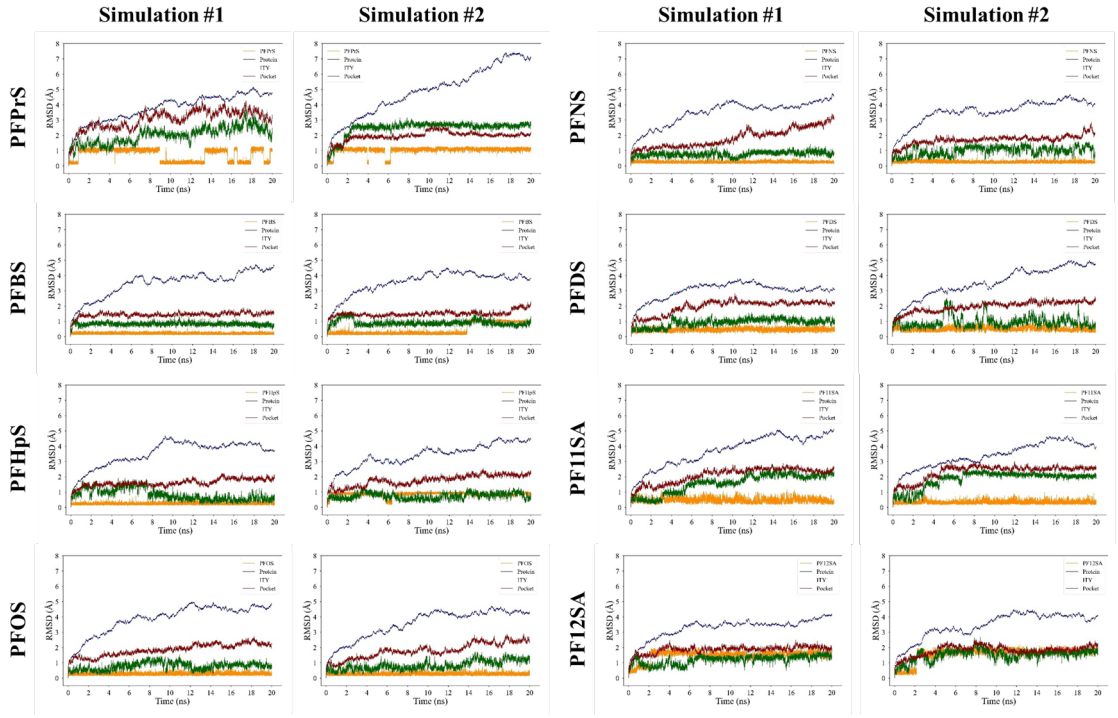


Figure S4.10 The RMSD time-series of apo system and sulphonic acid PFAS simulations.

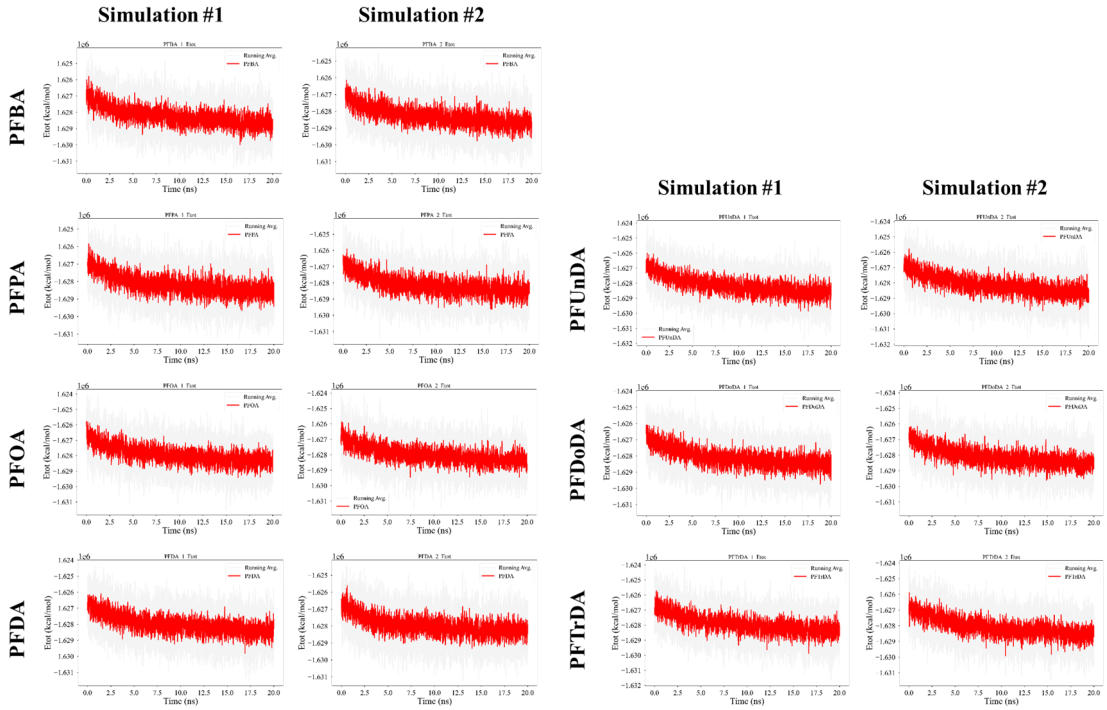


Figure S4.11 The total energy of carboxylic acid PFAS simulations.

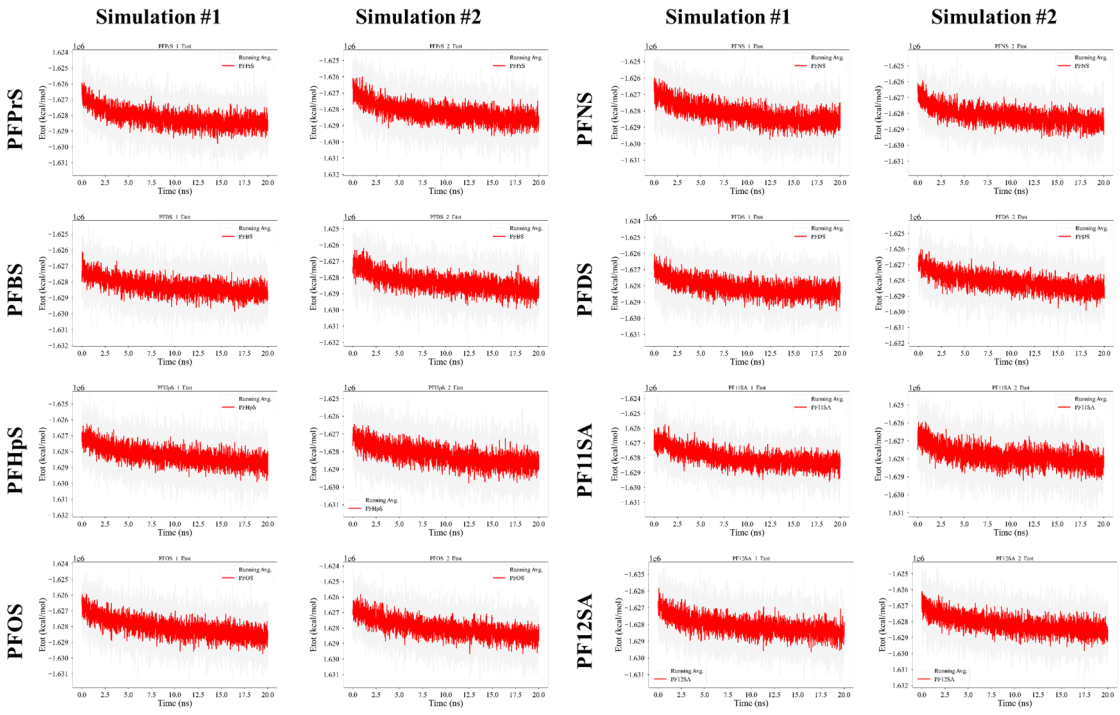


Figure S4.12 The total energy of sulphonic acid PFAS simulations.

CHAPTER 5

FISHING FOR ANSWERS: DIFFERENT BINDING MODES OF PFAS TARGETING RAINBOW TROUT ESTROGEN RECEPTORS

5.1 Introduction

Per- and polyfluoroalkyl substances (PFAS) are a class of synthetic organic fluorinated chemicals that were first created in the 1940s and quickly gained popularity in consumer and industrial products, such as food packaging, nonstick cookware, and water- and stain-proof textiles, due to their desirable water and oil repellent properties[1]. The high stability of PFAS in a variety of environments and their resistance to heat and degradation has resulted in their use in applications including firefighting foams.[2–4] In fact, PFAS are widely referred to as 'forever chemicals' or 'zombie chemicals' due to their high stability and resistance to degradation in the environment. As a direct consequence of this feature, PFAS show high levels of accumulation in water, soil, and living organisms, including humans.[5] The most well-known PFAS, perfluorooctanoic acid (PFOA) and perfluorooctane sulfonic acid (PFOS), were phased out of production by U.S. manufacturers in the mid-2000's and in 2022, were proposed to be considered as hazardous by the U.S. Environmental Protection Agency (EPA) under the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) as they present substantial danger to human health.[EPA] However, though banned for some time, PFOA and PFOS still persist in living organisms and in the environment, including the Great Lakes.[7,8] Despite the widespread use of PFAS over the past 70 years, only in the past decade have the health and environmental impacts of PFAS been widely studied. PFAS exposure in humans was shown to be linked to health problems, including high cholesterol levels, thyroid problems, certain types of cancers, and disruptions of the endocrine system.[9–15] According to the Public Health and Safety Organization, drinking PFAS-contaminated water may result in developmental problems in embryos of pregnant women.[16–19] Prior studies have shown that a major biological implication of the presence of PFAS in blood serum is the activation of certain nuclear receptors, such as Pregnane X Receptor.[9,10,14,20] A recent in vitro study testing for PFAS activation of human peroxisome proliferator-activated receptor α (PPAR α), peroxisome proliferator-activated receptor- γ (PPAR γ), and estrogen receptors (ER) indicated that multiple PFAS, both legacy and new, can result in activation of PPAR α, PPAR γ, and ER at certain concentrations.[21] Due to their roles in regulation of growth and lipid metabolism, the premature activation of these proteins can

have adverse effects on hormonal regulation and lipid metabolism. Though the direct mechanisms of PFAS toxicity have not been fully elucidated, the existing literature mentions several adverse effects that they have on human health through various nuclear receptors.[22,23] As PFAS contamination has become a growing public health concern, attention has turned to ecological areas of great significance, including the Great Lakes. The levels of PFAS contamination among the Great Lakes does vary – e.g., the northern lakes, such as Lake Superior, have the lowest concentrations of PFAS, while the highest concentrations were found in Lake Erie and Ontario, which lie in close proximity to areas of high industrial activity.[7,24–36] Furthermore, PFOS was the most common contaminant found to bioaccumulate in Great Lakes fish, such as lake trout, due to years of prior widespread use of PFOS. However, the bioaccumulation potential of PFAS was observed to vary based on the functional group and the carbon chain length.[7?] In terms of PFAS exposure to estrogen receptors in fish species, there are a limited number of studies.[37–44] A recent in vivo study highlights that several PFAS, including FC8-diol and HFPO-DA, exhibited varying levels of estrogenic activities in Fathead minnows.[45] In another study on the effect of PFAS on zebra fish, it was found that PFDA or PFTrDA can modulate the sex hormone balance by altering the steroidogenesis in a sex-dependent way; in male zebra fish, estradiol concentrations significantly increased upon PFAS exposure, but no such increase was observed for females.[46] Furthermore, a mixture of PFOS, PFNA, PFBA and PFOA was shown to cause an increase in endocrine-disruption biomarker levels, which was hypothesized to occur through either estrogen receptor binding and/or induction of estrogen expression.[47] In Tilapia, the exposure to PFOS, PFOA, and FTOHs resulted in anti-estrogenic activities in the presence of Estradiol.[41] However, despite the accumulating evidence on PFAS toxicity, the molecular-level details of the PFAS exposure in Estrogen receptors is not fully understood. The two subtypes of ERs, Estrogen receptor alpha and Estrogen receptor beta, have distinct roles in mammals and other vertebrates, including fish. In mammals, it is known that Estrogen receptor alpha is dominantly present in reproductive, bone, liver, and breast tissues, and involved in the development of secondary sexual characteristics; Estrogen receptor beta is found in the central nervous system, the immune system, and the cardiovascular system, playing an important

role in cardiac function.[48,49] In rainbow trout, on the other hand, ER alpha is found dominantly in the testes, liver, and spleen, and the ER beta was expressed more prominently in the kidney and liver.[50] Moreover, it has been shown that these two ERs also have different affinities towards the natural ligand, estradiol, bringing the question of whether the impact of PFAS exposure would affect the ER alpha and ER beta differently.[50–52] Given the pivotal involvement of ERs in essential processes, exploring how PFAS may interfere with the functions of ERs in fish is crucial for elucidating the endocrine-disrupting impact of these substances on aquatic organisms and the ecosystem. In this work, due to significant presence of PFAS contaminants in critical environmental areas including the Great Lakes, detailed insight into PFAS binding and toxicity towards the two Estrogen receptor subtypes, rainbow trout Estrogen receptor alpha (rER α) and Estrogen receptor beta (rER β), will be obtained about rainbow trout using molecular dynamics (MD) simulations and structural analysis. As a predatory fish, rainbow trout consume other organisms which makes them increasingly susceptible to higher doses of PFAS exposure and potential harmful effects. Understanding the impact of PFAS exposure on Estrogen receptors specifically, which are responsible for not only reproductive systems of fish but also many other important physiological functions including the immune system, enables a more comprehensive view of the impact of PFAS on the endocrine system, and may lead to the development of in vivo mitigation strategies. The observations from this work can also provide more insight on how endocrine disruption through ERs can occur in humans as well.

5.2 Computational Protocols

5.2.1 System preparation and docking protocols

The rER α and rER β sequences of rainbow trout (sp. Oncorhynchus mykiss) were obtained from the UniProt database with the accession numbers P16058 and P57782, respectively.[53] I-TASSER server was used for homology modeling of the protein structures.[54–56] The resulting structures were overlapped with the human ER α (PDB ID: 1G50) and ER β (PDB ID: 2J7X) structures co-crystallized with 17-β-Estradiol (E2) structures to determine the binding pocket residues in the ligand binding pockets.[57,2j7] Molecular Operating Environment (MOE) was used

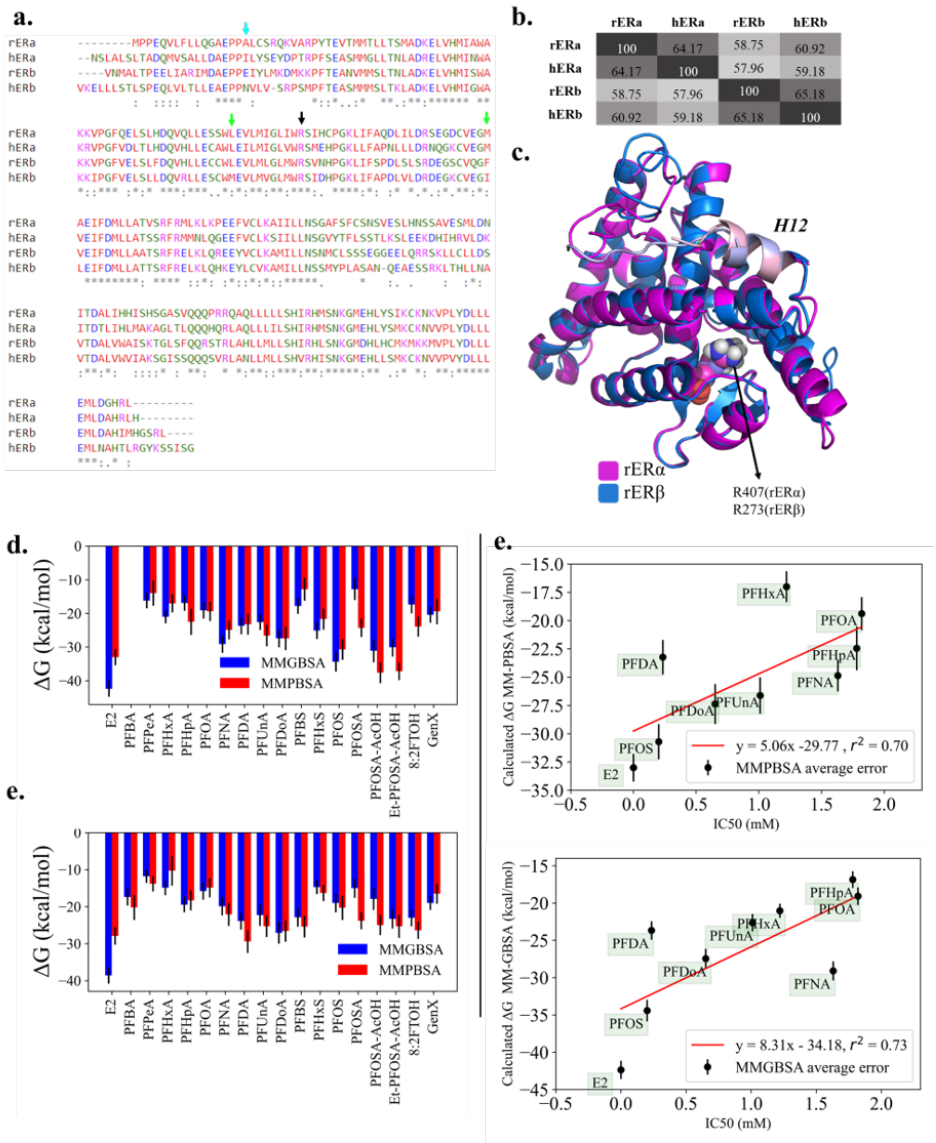


Figure S3.1 a. Sequence alignment of ligand binding domains of fish (rER α, rER β) and human (hER α, hER β) Estrogen receptors. The R407(rER α)/R273(rER β) residue used for the pharmacophore modeling is indicated with a black arrow. The blue arrow shows the mutated residue that causes the conformation change for R407(rER α)/R273(rER β) in rainbow trout estrogen receptors: A339/E205. Green arrows indicate the pocket residues that are not conserved between hER α and hER β four proteins: L384/M291 and M421/I328, for hER α and hER β, respectively. b. The ClustalW percent identity matrix for the multiple sequence alignment of rER α, rER β, hER α, and hER β ligand binding domains (LBDs).76,77 c. Superimposed structure of ligand binding domains of rER α (pink) and rER β (blue). R408(rER α)/R274(rER β) residue is shownwith van der Waals surface representation. d. The MM-PBSA/GBSA binding energies for rER α LBD are reported. e. The MM-PBSA/GBSA binding energies for rER β LBD are reported. f. The correlation between the experimental IC_{50} values and the calculated binding energies with MM-PBSA and MM-GBSA methods for PFAS bound to rER α LBD.[37] Kendall's tau is 0.66 for both correlation plots. The values of binding energies and the standard deviations are provided in Table S1.

for the minimization of homology models, determination of protonation states, and for docking procedures.59,60[59,60] rER α and rER β models were minimized with the AMBER10:Extended Huckel Theory (EHT) forcefield where Amber ff10 was used for the protein structure.[61–63] Once the binding pocket residues were identified by overlapping the homology models with human proteins, the selected PFAS compounds (Table S1) were docked using a pharmacophore approach that places the negatively charged head group of PFASs near R407 (rER α)/R273 (rER β) residues. R407 (rER α)/R273 (rER β) residues (R394 in hER α, R301 in hER β) is known to orient towards the OH group of the E2 ligand, as seen in the crystal structures, therefore it was selected for the orientation of the PFASs within the pocket.[57,2j7]The pharmacophore approach was used during the initial placement process with a London dG scoring function to obtain 100 poses.64 The further refinement was performed with an induced fit method and Generalized Born Volume integral/Weighted Surface Area (GBVI/WSA) scoring function, and the top 10 poses were reported.[59,64]The highest scoring pose for each PFAS were selected for Molecular Dynamics (MD) simulations.

5.2.2 Simulation details

AM1-BCC partial charges were calculated using antechamber module of Amber18/AmberTools20 using Generalized Amber Force Field (gaff2) to obtain the partial charges for the PFAS molecules.[65,62,66] The simulation boxes for each system were generated using the tleap module.[65] ff14SB, gaff2, and TIP4PEW force fields were selected for protein, PFAS, and water molecules, respectively.[62,67–69] Each system was neutralized using 0.1 M NaCl salt.[70] The minimization and heating steps were performed in a stepwise fashion as follows: (i) The systems were minimized with restraints (500, 200, 20, 10, 5, 0 kcal mol-1 Å-2). (ii) Heating from 100 K to 283.15 K was performed in 30 ps. (iii)The systems were equilibrated for 100 ps at 283.15 K. The production step consisted of a 30 ns long equilibrium MD simulation using 2 fs timestep at 300 K and 1 atm. A set of duplicate simulations were also performed by randomizing the initial velocities after the heating step. The temperature and pressure during the simulations were controlled by Langevin thermostat and isotropic position scaling. The bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[71] The MD simulations were performed using AMBER 2018 with pmemd.cuda module.[65]

5.2.3 Analysis of the trajectories

Molecular Mechanics - Generalized Born Surface Area and Molecular Mechanics - Poisson Boltzmann Surface Area (MM-GBSA/PBSA) methods were used to estimate the binding strength of PFASs in rER α and rER β ligand binding domains (LBDs), as implemented in Amber18/AmberTools20.[65] The binding energies were calculated for the last 1 ns of the simulations, including the duplicate simulations, and averaged for each PFAS. Root mean square distances (RMSD) and per-residue root mean square fluctuations (RMSF) were calculated with default settings as implemented in cpptraj module of AmberTools20.[72] Hydrogen bonds between the PFASs and the binding pocket residues were analyzed using cpptraj as well. The last 5 ns of the simulations were clustered using k-means clustering algorithm to obtain a representative frame. All time-series data were plotted using Python's matplotlib library, and the figures were obtained using UCSF Chimera 1.13.1 and MOE 2022.02.[59,73]

5.3 Results and Discussion

5.3.1 Stability of Investigated Complexes

The common Arginine residue used in pharmacophore docking has been identified based on the existing co-crystal structures of human ER α and ER β LBDs. The docking of E2 ligand yielded a pose similar to what was observed in human Estrogen receptors, and the charged head group of PFAS positioned near the side chain of arginine during the docking.[55,56] These poses yielded a comparable starting point for the PFAS simulations. In order to draw meaningful conclusions from our simulations, assessing both structural and energetic stability of the simulated systems is important. The structural stabilities of the systems were measured by RMSD analysis, as per Figure S2-S5, and this analysis indicated that the RMSD of simulations reached a plateau during the last 5 ns of the simulations. Similarly, the time-series data of the total energies that were tracked and reported in Figure S6-S7 indicated the systems reach an energetically equilibrated state during the first 10 ns and continued to stay stable until the end of the simulations. Therefore, only the last 5 ns of each simulation were considered for further analysis.

5.3.2 Binding strength and modes of PFAS

The sequence comparison of rER α and rER β from rainbow trout indicated that the two sequences have 50% identity for the whole sequence, and 58% identity for LBDs only.[50–52] In Figure 1, the sequence overlap and the structural superposition of both LBDs is shown. The binding pocket and the surrounding regions of both proteins are highly similar to each other, except for the residues shown and mutations listed in Figure S1. While some mutations do not change the chemical nature of the amino acid, others such as Glu to Gly or Lys to Met can cause changes in either the charge or polarity of the sidechains, which, in turn, can impact the binding strength as well as the binding mode of the investigated PFAS. While these mutations do not seem to alter the volume of the binding pockets significantly, 85 Å3 and 92 Å3 for rER α and rER β, respectively, the orientations of pocket residues including R407/R273 is highly impacted. The obtained MM-PBSA/GBSA binding energies for the E2 and PFAS are reported in Figure 1(d) and 1(e) for rER α and rER β, respectively. The prediction powers of MM-PBSA/GBSA approaches for PFAS binding to the rER α protein was analyzed by comparing the calculated binding affinities to experimentally available half-maximal inhibitory concentration (IC_{50}) values from studies by Benninghoff et al., and reported in Figure 1(e), for both MM-PBSA and MM-GBSA values for PFHxA, PFHpA, PFOA, PFNA, PFDA, PFUnA, PFDoA, and PFOS.[37] A strong correlation between the experimental IC_{50} values and calculated binding affinities of MM-PBSA and MM-GBSA methods were observed with coefficients of determination (R^2) of 0.70 and 0.73, respectively. While the MM-GBSA method has a slightly higher R^2, PFDA, PFNA, and E2 were outliers based on the correlation plot. Meanwhile, the MM-PBSA approach with 0.70 R^2 only had two outliers, PFDA and PFHxA, and the binding strength of E2 was consistent with the IC_{50} value, prompting the consideration of MM-PBSA binding energy values for further analysis. Prior studies indicate that the binding affinity of the natural agonist E2 differs between the two subtypes.[51,52,74] Our calculations show that the binding energy differences between rER α and rER β slightly favor the E2 binding for rER α protein, as shown in Figure 1. For both subtypes, E2 is among the strongest binders, indicating that the preference for the natural ligand is higher than for PFAS. Still, especially for rER α, the

predicted binding strength of certain PFAS, such as PFOS, PFOSA-AcOH, and Et-PFOSA-AcOH, were observed to be strong. PFDA and PFDoA for rER β protein were among the strongest binders after the E2 ligand. Weaker binding PFAS, however, were common between the two subtypes. PFBS, PFPeA, and PFHxA were the top three weak binders in rER α, and PFHxA, PFPeA, and PFOA were predicted to have the weakest binding energies in rER β (Figure 1). Interestingly, for carboxylic PFAS, having a higher number of fluorinated carbons resulted in a stronger binding energy when binding to both isoformsn.Still, due to the limited size of binding pockets, there was no increase in binding affinities after PFNA/PFDA when binding to rER α and rER β, respectively. For the PFAS with sulphonic acid group, however, only the binding energies with rER α pointed a relationship between the binding strength and chain length. The binding energies with rER β, on the other hand, showed no correlation with the length of the fluorinated carbon chains for sulphonic PFAS. Another interesting observation was related to the PFAS head group type and its relation to the binding strength. The type of head group of the PFAS impacted the binding strength, as shown in the comparison of the PFOS, PFOSA, PFOSA-AcOH, and Et-PFOSA-AcOH molecules. All of these compounds have eight fluorinated carbons, with different head groups (Table S1). However, the predicted binding strengths for rER α protein are different, pointing to the importance of the head groups and the interactions they are forming. The aforementioned four PFAS have stronger binding affinities than the majority of carboxylic PFAS for rER α LBD, but their affinities towards rER β protein were on par with those of long-chain carboxylic PFAS, such as PFDA, due to the mutations in the binding site, hence forming different interactions. These differences in binding affinities highlight the fact that binding free energies of PFAS compounds depend on (i) the carbon chain length, which is limited by the pocket size, and (ii) the type of the functional group. For the specific case of binding to rER α and rER β LBDs, overall binding affinity of carboxylic PFAS is quite comparable between two subtypes while the sulphonic and sulphonamide PFAS showed strong affinity towards rER α.[37]

Table S3.1 List of residues with largest energy contribution to the binding of E2 and PFAS.

Compound Name	**rERα**
E2	E365, H537, L400
PFPeA	R407, K365, K414, K542, K544, K546, R561
PFHxA	R407, K365, K414, K542, K544, K546, R561
PFHpA	R407, K365, K414, K542, K544, K546, R561
PFOA	R407, K365, K414, K542, K544, K546, R561
PFNA	R407, K365, K414, K542, K544, K546, R561
PFDA	R407, K365, K414, K542, K544, K546, R561
PFUnA	R407, K365, K414, K542, K544, K546, R561
PFDoA	R407, K365, K414, K542, K544, K546, R561
PFBS	R407, K365, K414, K542, K544, K546, R561
PFHxS	R407, K365, K414, K542, K544, K546, R561
PFOS	R407, K365, K414, K542, K544, K546, R561
PFOSA	E366, P417, I437, L538
PFOSA-AcOH	R407, K365, K414, K542, K544, K546, R561
Et-PFOSA-AcOH	R407, K365, K414, K542, K544, K546, R561
8:2 FTOH	R407, K365, K414, K542, K544, K546, R561
GenX	R407, K365, K414, K542, K544, K546, R561
Compound Name	**rERβ**
E2	L225, L266, L270, H403
PFPeA	K231, R273, K280, H403, K408, K410, K411
PFHxA	K231, R273, K280, H403, K408, K410, K411
PFHpA	K231, R274, K280, K408, K410, K411
PFOA	K231, R273, K280, H403, K408, K410, K411
PFNA	K231, R273, K280, H403, K408, K410, K411
PFDA	K231, R273, K280, H403, K408, K410, K411, G294, S295
PFUnA	K231, R273, K280, H403, K408, K410, K411
PFDoA	K231, R273, K280, K408, K410, K411
PFBS	K231, R273, K280, H403, K408, K410, K411
PFHxS	K231, R273, K280, K408, K410, K411
PFOS	K231, R273, K280, K408, K410, K411
PFOSA	L225, T226, H403
PFOSA-AcOH	K231, R273, K280, K408, K410, K411
Et-PFOSA-AcOH	K231, R273, K280, K408, K410, K411
8:2 FTOH 1	E232, L266, H403, L404
GenX	K231, R273, K280, H403, K408, K410, K411

5.3.3 Changes in interaction patterns upon PFAS binding

The binding energy predictions provided insight about the overall ranking of the affinities of investigated PFAS, and investigating the local interactions made by these fluorinated compounds is key for deciphering the molecular recognition by the rER α and rER β proteins. The results of the residue decomposition analysis that was employed to obtain the energetic contributions from the surrounding residues are reported in Table S8-S13, and the direct interactions between the PFAS and binding pocket amino acids were identified using hydrogen bond analysis, as shown in Figure 2. In the rER α and rER β LBD pockets, the binding of E2 is mainly driven by hydrogen bonds with E366/E232 and H537/H403, as shown in Figure 2(c), indicating that the orientation of the E2 molecule within the binding pockets as well as the recognition may be similar for both proteins. The histidine residue in two subtypes, H537/H403, provided an anchor point for the hydroxyl group of E2, while the other end is oriented towards the E366/E232 (Figure 2(c)). The largest energy contributions among the pocket residues for E2 binding were with E365, H537, L400 for rER α, and L225, L266, L270, H403 for rER β, as shown in Table 1, pointing out that although the anchor residues are the same, the strongest interactions with surrounding residues were different for rER α and rER β, potentially due to the mutations within the binding pocket. In general, however, the interaction energies of E2 ligand with the pocket residues were between 0 kcal mol-1 and -10 kcal mol-1, for both LBDs. The majority of PFAS had stabilizing interaction energies (smaller than zero) with the surrounding basic residues and unfavorable interaction energies (larger than zero) with acidic residues. This observation was valid for both rER proteins (Table S8, S11). One notable exception to this is PFOSA in rER α, and 8:2FTOH and PFOSA in rER β. PFOSA In rER β LBD, however, PFOSA did not form any notable interactions with either basic or acidic residues, and 8:2FTOH had strong interaction with E232 residue (Table S11). As these two PFAS lack a charged group, it was expected not to have any prominent interactions with acidic and basic residues.The binding of PFAS in rER α LBD was mainly facilitated by a direct hydrogen bonding between R407 side chain and the negatively charged head group of PFAS, while majority of PFAS formed stabilizing interactions with the surrounding basic residues and unfavorable interactions

with the acidic ones (Table S8-S11).

For sulfonamide (PFOSA) and fluorotelomer (8:2 FTOH) compounds, the anchor residue was E366 (Figure 2(a)) as those two PFAS have sulphonamide and alcohol head groups, respectively. PFOSA also formed interactions that are very similar to E2 ligand, i.e. strong stabilizing interaction with E366 in rER α (Table S8).To the contrary, the recognition of PFAS in rER β mainly involved the hydrogen bonding with the side chain of H403, located on Helix-11 (Figure 2(d)), and there was no direct interaction between R273 and PFAS head groups, except for the PFDoA compound. Moreover, the strongest interaction was also with H403 (Table S9, S12), and the interaction strength with this histidine was generally stronger for rER β, due to the direct hydrogen bonding between PFAS head group and the H403 residue. Interestingly, the majority of sulphonic PFAS, namely PFDA, PFUnA, PFHxS, PFOS, PFOSA, PFOSA-AcOH, and Et-PFOSA-AcOH, did not form any direct hydrogen bonds with the pocket residues of rER β. As shown in Figure S13, these sulphonic PFAS went through a slight rotation within the rER β binding pocket and their head groups were oriented towards the Helix-7 and Helix-11 (Figure S6). The other PFAS, specifically short chain sulphonic compound PFBS and carboxylic PFAS with up to eight fluorinated carbons, were able to fully rotate their head groups to interact with H403. This 'tumbling' motion of PFAS that was observed only within rER β binding pocket and not within rER α is a direct consequence of the amino acid differences in the pocket residues. In apo rER α binding pocket, R407 is in proximity of two glutamate residues, E336 and E366, forming a direct hydrogen bond with the latter, as depicted in Figure 2(e). On the other hand, there are three glutamate residues, E202, E205 and E232, in the proximity of R273 of apo rER β binding pocket. The arginine interacts directly with E205, and consequently shifting the orientation of R273 side chain. This E205 residue of rER β is modified to an alanine (A339) in rER α protein, explaining why (i) there is no direct interaction with R273 when natural ligand and PFAS are present in the pocket, and (ii) the charged PFAS undergo the 'tumbling' motion in rER β pocket. The reorientation within the pocket of rER β is not just limited to the R273 residue. Upon PFAS binding, the phenylalanine residue (F283) located near the arginine also goes through a conformational change, as shown in Figure 2(f).

In rER α, the orientation of the PFAS as well as E2 allowed an interaction between F417 side chain and the PFAS tail group to from stabilizing interactions (Table S10), and the orientation of F417 side chain further away from R407. Meanwhile, the strength of the interactions with F283 was weaker in rER β pocket (Table S13) and the F283 side chain had a similar orientation to what was observed in apo rER β simulations (Figure 2(f)). The role of this phenylalanine residue is not well-defined in the literature.; However, our simulations indicate that it may have a role in stabilizing the ligand within the binding pocket for both subtypes. To the best of our knowledge, this is the first study assessing molecular-level details of PFAS binding and toxicity in rainbow trout Estrogen receptors alpha and beta. The two subtypes have slightly different affinities against the E2 ligand and various PFAS due to the modifications in amino acid sequences within the binding pocket. The most commonly known mutations identified in human Estrogen receptor studies, L397/L263 and M434/F300 for rER α to rER β, respectively, were found to impact the binding strengths and the orientations of PFAS in rainbow trout. The most striking observation, however, was the amino acid modification of A339 to E205 from rER α to rER β, resulting in the complete reorientation of R273/R407 and F417/F283 residues and causing PFAS to 'tumble' within the binding pocket of rER β. This orientation change may explain the affinity differences between two subtypes, and further, it may indicate different downstream impacts of PFAS exposure. It is the first time in the literature that this amino acid modification was identified to impact the PFAS binding for Estrogen receptors. Human ER α and ER β proteins also do not have a conserved residue at this location with Ile to Asn modification, respectively (Figure 1(a), blue arrow). The impact of this residue on the mobility and orientation of important and conserved arginine amino acid may have a role in developing subtype-selective binders for both human and rainbow trout estrogen receptors.

5.3.4 Environmental Impact

Understanding how PFAS exert toxic effects on living organisms and ecosystems is crucial for developing effective mitigation strategies. The Great Lakes, with its central role for local biodiversity, faces a significant threat due to persistent accumulation of PFAS. This accumulation not only poses a risk to the ecosystem and biodiversity, but also to human health through the

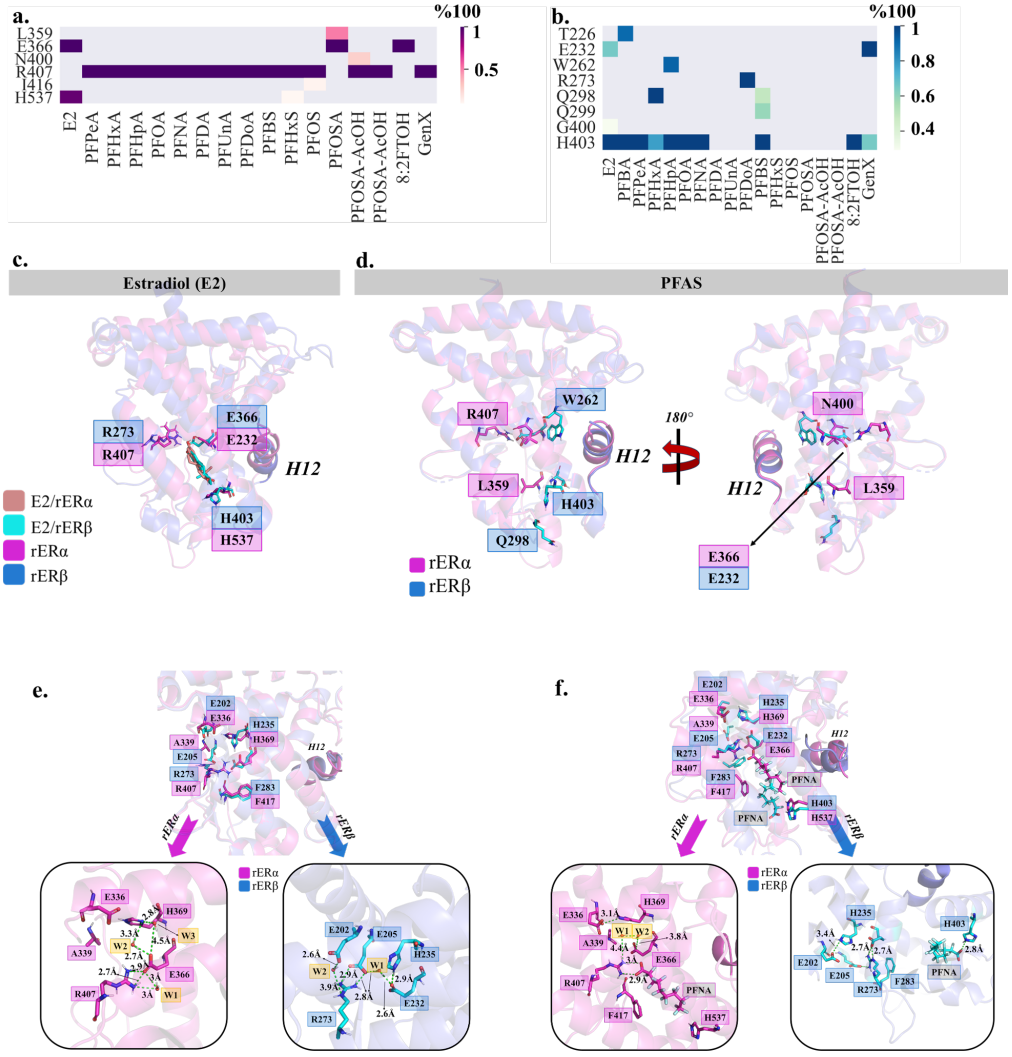


Figure S3.2 The hydrogen bond percentage heatmap of the PFAS in (a) rER α LBD and (b) rER β LBD pockets. The y-axis shows the pocket residue names and the x-axis shows the PFAS bound to the pocket. c. The locations of the residues that form direct hydrogen bond with E2 in rER α and rER β, overlapped. d. The locations of the residues that form direct hydrogen bond with PFAS in rER α and rER β. Helix 12 (H12) is also shown in light pink and light blue colors for rER α and rER β, respectively. e. The detailed depiction of the orientations of binding pocket residues in apo rER α and rER β LBDs is pictured. The distances between hydrogen bonding heavy atoms were shown in green dash lines. f. The detailed depiction of the orientations of binding pocket residues in PFNA-bound rER α and rER β LBDs is shown. The distances between hydrogen bonding heavy atoms were shown in green dash lines. PFNA was selected as a representative of the majority of PFAS simulations.

consumption of contaminated fish. Therefore, addressing the impact of PFAS on fish health and ecosystems is vital for protecting the environment and human health. As a first step in understanding PFAS toxicity, molecular details of how PFAS binds to target proteins in fish needs to be addressed. Here, we focused on Estrogen receptors: as one of the nuclear receptors, they play a fundamental role not only in reproductive system but also cytoplasmic signal transduction as well as in regulating the immune system. The current study sheds light to the different binding modes of PFAS within rER α to rER β LBDs and molecular details of PFAS interactions within the binding pocket. Significantly, the mutations of binding pocket residues not only caused PFAS to bind differently in Estrogen receptor subtypes, but also in different orientations, emphasizing that PFAS exert its impact through different mechanisms. This understanding is central for devising targeted interventions for PFAS toxicity and creating regulatory mechanisms that can effectively mitigate PFAS-associated risks.

BIBLIOGRAPHY

[1] Brennan, N. M., Evans, A. T., Fritz, M. K., Peak, S. A., and von Holst, H. E. (2021). Trends in the regulation of per-and polyfluoroalkyl substances (pfas): A scoping review. *International Journal of Environmental Research and Public Health*, 18:10900.

[2] Coggan, T. L., Moodie, D., Kolobaric, A., Szabo, D., Shimeta, J., Crosbie, N. D., Lee, E., Fernandes, M., and Clarke, B. O. (2019). An investigation into per- and polyfluoroalkyl substances (pfas) in nineteen australian wastewater treatment plants (wwtps). *Heliyon*, 5:e02316.

[3] Pelch, K. E., Reade, A., Wolffe, T. A., and Kwiatkowski, C. F. (2019). Pfas health effects database: Protocol for a systematic evidence map. *Environment International*, 130:104851.

[4] Gaines, L. G. T. and Gaines, C. G. L. T. (2023). Historical and current usage of per- and polyfluoroalkyl substances (pfas): A literature review. *American Journal of Industrial Medicine*, 66:353–378.

[5] Calafat, A. M., Kuklenyik, Z., Reidy, J. A., Caudill, S. P., Tully, J. S., and Needham, L. L. (2007). Serum concentrations of 11 polyfluoroalkyl compounds in the u.s. population: Data from the national health and nutrition examination survey (nhanes) 1999-2000. *Environmental Science and Technology*, 41:2237–2242.

[EPA] Epa proposes designating certain pfas chemicals as hazardous substances under superfund to protect people's health | us epa.

[7] Remucal, C. K. (2019). Spatial and temporal variability of perfluoroalkyl substances in the laurentian great lakes. *Environmental Science: Processes & Impacts*, 21:1816–1834.

[8] Point, A. D., Holsen, T. M., Fernando, S., Hopke, P. K., and Crimmins, B. S. (2021). Trends (2005–2016) of perfluoroalkyl acids in top predator fish of the laurentian great lakes. *Science of The Total Environment*, 778:146151.

[9] Rappazzo, K., Coffman, E., and Hines, E. (2017). Exposure to perfluorinated alkyl substances and health outcomes in children: A systematic review of the epidemiologic literature. *International Journal of Environmental Research and Public Health*, 14:691.

[10] Duan, X., Sun, W., Sun, H., and Zhang, L. (2021). Perfluorooctane sulfonate continual exposure impairs glucose-stimulated insulin secretion via sirt1-induced upregulation of ucp2 expression. *Environmental Pollution*, 278:116840.

[11] Sunderland, E. M., Hu, X. C., Dassuncao, C., Tokranov, A. K., Wagner, C. C., and Allen, J. G. (2019). A review of the pathways of human exposure to poly- and perfluoroalkyl substances (pfass) and present understanding of health effects. *Journal of Exposure Science & Environmental Epidemiology*, 29:131–147.

[12] Anderko, L. and Pennea, E. (2020). Exposures to per-and polyfluoroalkyl substances (pfas): Potential risks to reproductive and children's health. *Current Problems in Pediatric and Adolescent Health Care*, 50:100760.

[13] Guo, H., Chen, J., Zhang, H., Yao, J., Sheng, N., Li, Q., Guo, Y., Wu, C., Xie, W., and Dai, J. (2022). Exposure to genx and its novel analogs disrupts hepatic bile acid metabolism in male mice. *Environmental Science & Technology*, 56:6133–6143.

[14] Almeida, N. M. S., Eken, Y., and Wilson, A. K. (2021). Binding of per- and polyfluoro-alkyl substances to peroxisome proliferator-activated receptor gamma. *ACS Omega*, 6:15103–15114.

[15] Munoz, G., Liu, J., Duy, S. V., and Sauvé, S. (2019). Analysis of f-53b, gen-x, adona, and emerging fluoroalkylether substances in environmental and biomonitoring samples: A review. *Trends in Environmental Analytical Chemistry*, 23:e00066.

[16] Chen, M. H., Ha, E. H., Wen, T. W., Su, Y. N., Lien, G. W., Chen, C. Y., Chen, P. C., and Hsieh, W. S. (2012). Perfluorinated compounds in umbilical cord blood and adverse birth outcomes. *PLOS ONE*, 7:e42474.

[17] Sagiv, S. K., Rifas-Shiman, S. L., Fleisch, A. F., Webster, T. F., Calafat, A. M., Ye, X., Gillman, M. W., and Oken, E. (2018). Early-pregnancy plasma concentrations of perfluoroalkyl substances and birth outcomes in project viva: Confounded by pregnancy hemodynamics? *American Journal of Epidemiology*, 187:793–802.

[18] (2018). Prenatal exposure to perfluoroalkyl substances and birth outcomes; an updated analysis from the danish national birth cohort. *International Journal of Environmental Research and Public Health 2018, Vol. 15, Page 1832*, 15:1832.

[19] Johnson, P. I., Sutton, P., Atchley, D. S., Koustas, E., Lam, J., Sen, S., Robinson, K. A., Axelrad, D. A., and Woodruff, T. J. (2014). The navigation guide—evidence-based medicine meets environmental health: Systematic review of human evidence for pfoa effects on fetal growth. *Environmental Health Perspectives*, 122:1028–1039.

[20] Wen, L.-L., Lin, C.-Y., Chou, H.-C., Chang, C.-C., Lo, H.-Y., and Juan, S.-H. (2016). Perfluorooctanesulfonate mediates renal tubular cell apoptosis through ppargamma inactivation. *PLOS ONE*, 11:e0155190.

[21] Evans, N., Conley, J. M., Cardon, M., Hartig, P., Medlock-Kakaley, E., and Gray, L. E. (2022). In vitro activity of a panel of per- and polyfluoroalkyl substances (pfas), fatty acids, and pharmaceuticals in peroxisome proliferator-activated receptor (ppar) alpha, ppar gamma, and estrogen receptor assays. *Toxicology and Applied Pharmacology*, 449:116136.

[22] Amenyogbe, E., Chen, G., Wang, Z., Lu, X., Lin, M., and Lin, A. Y. (2020). A review on sex steroid hormone estrogen receptors in mammals and fish.

[23] Davidsen, N., Ramhøj, L., Lykkebo, C. A., Kugathas, I., Poulsen, R., Rosenmai, A. K., Evrard, B., Darde, T. A., Axelstad, M., Bahl, M. I., Hansen, M., Chalmel, F., Licht, T. R., and Svingen, T. (2022). Pfos-induced thyroid hormone system disrupted rats display organ-specific changes in their transcriptomes. *Environmental Pollution*, 305:119340.

[24] Furdui, V. I., Stock, N. L., Ellis, D. A., Butt, C. M., Whittle, D. M., Crozier, P. W., Reiner, E. J., Muir, D. C., and Mabury, S. A. (2007). Spatial distribution of perfluoroalkyl contaminants in lake trout from the great lakes. *Environmental Science and Technology*, 41:1554–1559.

[25] Houde, M., Czub, G., Small, J. M., Backus, S., Wang, X., Alaee, M., and Muir, D. C. (2008). Fractionation and bioaccumulation of perfluorooctane sulfonate (pfos) isomers in a lake ontario food web. *Environmental Science and Technology*, 42:9397–9403.

[26] Silva, A. O. D., Spencer, C., Scott, B. F., Backus, S., and Muir, D. C. (2011). Detection of a cyclic perfluorinated acid, perfluoroethylcyclohexane sulfonate, in the great lakes of north america. *Environmental Science and Technology*, 45:8060–8066.

[27] Myers, A. L., Crozier, P. W., Helm, P. A., Brimacombe, C., Furdui, V. I., Reiner, E. J., Burniston, D., and Marvin, C. H. (2012). Fate, distribution, and contrasting temporal trends of perfluoroalkyl substances (pfass) in lake ontario, canada. *Environment International*, 44:92–99.

[28] Martin, J. W., Whittle, D. M., Muir, D. C., and Mabury, S. A. (2004). Perfluoroalkyl contaminants in a food web from lake ontario. *Environmental Science and Technology*, 38:5379–5385.

[29] Silva, A. O. D., Muir, D. C., and Mabury, S. A. (2009). Distribution of perfluorocarboxylate isomers in select samples from the north american environment. *Environmental Toxicology and Chemistry*, 28:1801–1814.

[30] Codling, G., Vogt, A., Jones, P. D., Wang, T., Wang, P., Lu, Y. L., Corcoran, M., Bonina, S., Li, A., Sturchio, N. C., Rockne, K. J., Ji, K., Khim, J. S., Naile, J. E., and Giesy, J. P. (2014). Historical trends of inorganic and organic fluorine in sediments of lake michigan. *Chemosphere*, 114:203–209.

[31] Codling, G., Sturchio, N. C., Rockne, K. J., Li, A., Peng, H., Tse, T. J., Jones, P. D., and Giesy, J. P. (2018). Spatial and temporal trends in poly- and per-fluorinated compounds in the laurentian great lakes erie, ontario and st. clair. *Environmental Pollution*, 237:396–405.

[32] Yeung, L. W., Silva, A. O. D., Loi, E. I., Marvin, C. H., Taniyasu, S., Yamashita, N., Mabury, S. A., Muir, D. C., and Lam, P. K. (2013). Perfluoroalkyl substances and extractable organic fluorine in surface sediments and cores from lake ontario. *Environment International*, 59:389–397.

[33] Guo, R., Megson, D., Myers, A. L., Helm, P. A., Marvin, C., Crozier, P., Mabury, S., Bhavsar, S. P., Tomy, G., Simcik, M., McCarry, B., and Reiner, E. J. (2016). Application of a

comprehensive extraction technique for the determination of poly- and perfluoroalkyl substances (pfass) in great lakes region sediments. *Chemosphere*, 164:535–546.

[34] McGoldrick, D. J. and Murphy, E. W. (2016). Concentration and distribution of contaminants in lake trout and walleye from the laurentian great lakes (2008–2012). *Environmental Pollution*, 217:85–96.

[35] Asher, B. J., Wang, Y., Silva, A. O. D., Backus, S., Muir, D. C., Wong, C. S., and Martin, J. W. (2012). Enantiospecific perfluorooctane sulfonate (pfos) analysis reveals evidence for the source contribution of pfos-precursors to the lake ontario foodweb. *Environmental Science & Technology*, 46:7653–7660.

[36] Gewurtz, S. B., Silva, A. O. D., Backus, S. M., McGoldrick, D. J., Keir, M. J., Small, J., Melymuk, L., and Muir, D. C. (2012). Perfluoroalkyl contaminants in lake ontario lake trout: Detailed examination of current status and long-term trends. *Environmental Science and Technology*, 46:5842–5850.

[37] Benninghoff, A. D., Bisson, W. H., Koch, D. C., Ehresman, D. J., Kolluri, S. K., and Williams, D. E. (2011). Estrogen-like activity of perfluoroalkyl acids in vivo and interaction with human and rainbow trout estrogen receptors in vitro. *Toxicological Sciences*, 120:42–58.

[38] Wei, Y., Dai, J., Liu, M., Wang, J., Xu, M., Zha, J., and Wang, Z. (2007). Estrogen-like properties of perfluorooctanoic acid as revealed by expressing hepatic estrogen-responsive genes in rare minnows (gobiocypris rarus). *Environmental Toxicology and Chemistry*, 26:2440–2447.

[39] Xin, Y., Ren, X. M., Wan, B., and Guo, L. H. (2019). Comparative in vitro and in vivo evaluation of the estrogenic effect of hexafluoropropylene oxide homologues. *Environmental Science and Technology*, 53:8371–8380.

[40] Han, J. and Fang, Z. (2010). Estrogenic effects, reproductive impairment and developmental toxicity in ovoviparous swordtail fish (xiphophorus helleri) exposed to perfluorooctane sulfonate (pfos). *Aquatic Toxicology*, 99:281–290.

[41] Liu, C., Du, Y., and Zhou, B. (2007). Evaluation of estrogenic activities and mechanism of action of perfluorinated chemicals determined by vitellogenin induction in primary cultured tilapia hepatocytes. *Aquatic Toxicology*, 85:267–277.

[42] Qiu, Z., Qu, K., Luan, F., Liu, Y., Zhu, Y., Yuan, Y., Li, H., Zhang, H., Hai, Y., and Zhao, C. (2020). Binding specificities of estrogen receptor with perfluorinated compounds: A cross species comparison. *Environment International*, 134:105284.

[43] Qu, K., Song, J., Zhu, Y., Liu, Y., and Zhao, C. (2019). Perfluorinated compounds binding to estrogen receptor of different species: a molecular dynamic modeling. *Journal of Molecular Modeling*, 25:1–10.

[44] Cocci, P., Mosconi, G., and Palermo, F. A. (2021). An in silico and in vitro study for investigating estrogenic endocrine effects of emerging persistent pollutants using primary hepatocytes from grey mullet (mugil cephalus). *Environments - MDPI*, 8:58.

[45] Villeneuve, D. L., Blackwell, B. R., Cavallin, J. E., Collins, J., Hoang, J. X., Hofer, R. N., Houck, K. A., Jensen, K. M., Kahl, M. D., Kutsi, R. N., Opseth, A. S., Rodriguez, K. J. S., Schaupp, C., Stacy, E. H., and Ankley, G. T. (2023). Verification of in vivo estrogenic activity for four per- and polyfluoroalkyl substances (pfas) identified as estrogen receptor agonists via new approach methodologies. *Environmental Science and Technology*, 57:3794–3803.

[46] Jo, A., Ji, K., and Choi, K. (2014). Endocrine disruption effects of long-term exposure to perfluorodecanoic acid (pfda) and perfluorotridecanoic acid (pftrda) in zebrafish (danio rerio) and related mechanisms. *Chemosphere*, 108:360–366.

[47] Lee, J. W., Lee, J.-W., Shin, Y.-J., Kim, J.-E., Ryu, T.-K., Ryu, J., Lee, J., Kim, P., Choi, K., and Park, K. (2017). Multi-generational xenoestrogenic effects of perfluoroalkyl acids (pfaas) mixture on oryzias latipes using a flow-through exposure system. *Chemosphere*, 169:212–223.

[48] Jia, M., Dahlman-Wright, K., and Åke Gustafsson, J. (2015). Estrogen receptor alpha and beta in health and disease. *Best Practice & Research Clinical Endocrinology & Metabolism*, 29:557–568.

[49] Chen, P., Li, B., and Ou-Yang, L. (2022). Role of estrogen receptors in health and disease. *Frontiers in Endocrinology*, 13:839005.

[50] Nagler, J. J., Cavileer, T., Sullivan, J., Cyr, D. G., and Rexroad, C. (2007). The complete nuclear estrogen receptor family in the rainbow trout: Discovery of the novel er α2 and both er β isoforms. *Gene*, 392:164–173.

[51] Shyu, C., Cavileer, T. D., Nagler, J. J., and Ytreberg, F. M. (2011). Computational estimation of rainbow trout estrogen receptor binding affinities for environmental estrogens. *Toxicology and Applied Pharmacology*, 250:322–326.

[52] Shyu, C., Brown, C. J., and Ytreberg, F. M. (2010). Computational study of evolutionary selection pressure on rainbow trout estrogen receptors. *PLOS ONE*, 5:e9392.

[53] Consortium, T. U. (2023). Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51:D523–D531.

[54] Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E. W., and Zhang, Y. (2021). Folding non-homologous proteins by coupling deep-learning contact maps with i-tasser assembly simulations. *Cell Reports Methods*, 1:100014.

[55] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). The i-tasser suite:

protein structure and function prediction. *Nature Methods 2015 12:1*, 12:7–8.

[56] Yang, J. and Zhang, Y. (2015). I-tasser server: new development for protein structure and function predictions. *Nucleic Acids Research*, 43:W174–W181.

[57] Eiler, S., Gangloff, M., Duclaud, S., Moras, D., and Ruff, M. (2001). Overexpression, purification, and crystal structure of native er α lbd. *Protein Expression and Purification*, 22:165–173.

[2j7] Rcsb pdb - 2j7x: Structure of estradiol-bound estrogen receptor beta lbd in complex with lxxll motif from ncoa5.

[59] (2022). Molecular operating environment (moe).

[60] Labute, P. (2009). Protonate3d: Assignment of ionization states and hydrogen coordinates to macromolecular structures. *Proteins: Structure, Function, and Bioinformatics*, 75:187–205.

[61] Hoffmann, R. (2004). An extended hückel theory. i. hydrocarbons. *The Journal of Chemical Physics*, 39:1397.

[62] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25:1157–1174.

[63] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65:712–725.

[64] Corbeil, C. R., Williams, C. I., and Labute, P. (2012). Variability in docking success rates due to dataset preparation. *Journal of Computer-Aided Molecular Design*, 26:775–786.

[65] York, D. and P.A. Kollman, D.A. Case, e. a. (2020). Amber 2018.

[66] He, X., Man, V. H., Yang, W., Lee, T.-S., and Wang, J. (2020). A fast and high-quality charge model for the next generation general amber force field. *The Journal of Chemical Physics*, 153:114502.

[67] Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11:3696–3713.

[68] Döpke, M. F., Moultos, O. A., and Hartkamp, R. (2020). On the transferability of ion parameters to the tip4p/2005 water model using molecular dynamics simulations. *The Journal of Chemical Physics*, 152:024501.

[69] Horn, H. W., Swope, W. C., Pitera, J. W., Madura, J. D., Dick, T. J., Hura, G. L., and

Head-Gordon, T. (2004). Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. *The Journal of Chemical Physics*, 120:9665–9678.

[70] Joung, I. S. and Cheatham, T. E. (2008). Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *The Journal of Physical Chemistry B*, 112:9020–9041.

[71] Ryckaert, J.-P., Ciccotti, G., and Berendsen, H. J. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23:327–341.

[72] Roe, D. R. and Cheatham, T. E. (2013). Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9:3084–3095.

[73] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004). Ucsf chimera: A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25:1605–1612.

[74] Petit, F., Valotaire, Y., and Pakdel, F. (1995). Differential functional activities of rainbow trout and human estrogen receptors expressed in the yeast saccharomyces cerevisiae. *European Journal of Biochemistry*, 233:584–592.

APPENDIX A

SUPPORTING TABLES

Table S5.1 The list of PFAS used in this study. The average calculated binding energies and standard deviations in kcal/mol, and experimental IC_{50} obtained from Ref. 16 are provided as well.

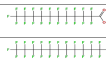
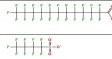
Carboxylic Acid PFAS Name	# of Fluorinated	Structure	rERα		rERβ		IC_{50} (mM)[Ref. 16]
			MM-PBSA (kcal mol^{-1})	MM-GBSA (kcal mol^{-1})	MM-PBSA (kcal mol^{-1})	MM-GBSA (kcal mol^{-1})	
PFBA	3		---	---	-21.18±3.31	-17.34±2.20	---
PFPeA	4		-14.04±3.74	-16.24±2.35	-13.80±2.10	-11.84±1.76	---
PFHxA	5		-17.00±2.73	-21.03±1.88	-10.27±4.05	-14.89±1.97	1.220
PFHpA	6		-22.45±3.88	-16.85±2.25	-18.31±2.70	-19.53±2.01	1.780
PFOA	7		-26.82±2.74	-27.32±2.51	-16.42±2.53	-16.48±2.39	1.820
PFNA	8		-24.44±2.76	-25.13±2.71	-23.61±2.62	-18.05±2.40	1.630
PFDA	9		-22.38±3.44	-25.74±2.47	-30.58±3.19	-24.01±2.31	0.234
PFUnA	10		-30.87±3.92	-25.63±3.63	-26.14±3.21	-24.42±2.94	1.010
PFDoA	11		-30.51±2.87	-30.91±2.62	-24.74±3.08	-24.52±2.74	0.651
PFBS	4		-12.82±3.44	-17.76±2.28	-25.43±2.91	-22.95±2.45	---
PFHxS	6		-14.91±4.02	-20.51±2.29	-17.07±2.38	-16.73±2.04	---
PFOS	8		-10.84±3.53	-32.00±3.02	-21.61±3.49	-18.53±2.47	0.201
PFOSA	8		-22.03±3.05	-13.14±3.11	-23.79±2.34	-15.05±2.63	---
PFOSA-AcOH	8		-37.57±3.12	-31.14±3.36	-24.09±2.94	-18.65±3.56	---
Et-PFOSA-AcOH	8		-39.01±3.32	-31.25±3.18	-26.27±2.86	-24.40±2.59	---
GenX	5		-19.40±3.76	-20.43±2.42	-16.47±2.75	-18.99±1.89	---
8:2 FTOH	8		-26.21±4.15	-19.57±3.35	-26.38±2.29	-22.96±2.29	N/A

Table S5.2 Average residue decomposition energies of charged rERα pocket residues. The color gradient goes from blue to red as the values change from negative to positive.

Res. #	E2	PFPeA	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUnA	PFDoA	PFBS	PFHxS	PFOS	PFOSA	PFOSAAcOH	Et-PFOSAAcOH	82FTOH	GenX
Asp364	0.05	27.84	29.23	28.12	25.23	26.70	27.00	25.06	24.83	27.01	26.93	23.24	-0.78	25.11	24.23	27.56	24.78
Glu366	-10.19	39.59	40.93	39.86	44.61	49.08	42.08	48.99	49.49	38.79	39.10	44.76	-10.91	48.11	41.71	39.32	36.64
Glu398	0.20	24.47	24.60	24.49	24.95	24.34	26.02	24.48	22.83	24.39	24.04	22.66	0.44	25.55	25.27	24.44	24.19
Glu427	-0.19	19.42	18.17	19.17	20.52	18.34	18.87	17.62	20.64	20.17	18.57	22.57	0.14	16.78	19.76	19.67	22.17
Asp429	-0.51	20.31	19.61	20.17	18.66	19.16	18.24	16.68	18.57	20.74	19.47	18.33	0.36	18.77	18.38	20.45	21.87
Glu432	-0.60	21.40	20.92	21.30	20.50	18.79	20.84	18.53	20.54	21.68	22.68	18.65	0.25	18.35	20.43	21.49	22.44
Glu436	-0.68	18.05	17.94	18.03	19.22	16.87	17.59	16.66	17.31	18.12	19.03	16.65	0.54	17.68	17.86	18.07	18.31
Asp439	-0.89	19.48	18.55	19.29	19.54	18.56	18.66	18.75	18.56	20.03	18.84	19.15	0.47	18.33	19.23	19.66	21.52
Glu536	-0.29	18.54	18.13	18.46	18.09	16.66	19.09	16.31	17.86	18.78	19.76	16.63	0.47	17.25	17.47	18.62	19.43
Lys365	0.20	-27.13	-27.56	-27.21	-29.17	-30.37	-27.98	-29.28	-29.32	-26.87	-29.84	-25.49	1.36	-27.52	-28.05	-27.04	-26.17
Arg407	-3.52	-81.80	-84.62	-82.36	-83.10	-76.49	-75.31	-77.67	-76.90	-80.11	-76.34	-75.67	2.37	-88.30	-90.27	-81.24	-75.59
Lys414	0.01	-19.10	-18.72	-19.02	-20.22	-18.84	-19.01	-19.28	-20.85	-19.33	-18.99	-23.82	-0.17	-18.75	-18.14	-19.18	-19.93
Lys542	-0.38	-17.75	-17.28	-17.65	-17.77	-16.40	-18.74	-16.12	-17.29	-18.03	-18.74	-15.86	-0.27	-17.18	-19.10	-17.84	-18.77
Lys544	0.30	-17.59	-16.15	-17.31	-16.03	-16.80	-18.03	-16.33	-16.36	-18.46	-19.45	-15.56	-0.47	-15.35	-17.11	-17.88	-20.77
Lys546	0.08	-16.24	-15.67	-16.13	-17.21	-15.77	-17.06	-14.85	-17.48	-16.59	-17.39	-15.21	-0.23	-16.02	-18.07	-16.36	-17.50
Arg561	-0.19	-15.36	-15.51	-15.39	-14.52	-13.74	-15.41	-14.37	-14.93	-15.28	-15.46	-13.75	-0.18	-13.99	-14.57	-15.34	-15.05

Table S5.3 Average residue decomposition energies of polar rERα pocket residues. The color gradient goes from blue to red as the values change from negative to positive.

Res. #	E2	PFPeA	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUnA	PFDoA	PFBS	PFHxS	PFOS	PFOSA	PFOSAAcOH	Et-PFOSAAcOH	82FTOH	GenX
Thr357	-0.02	0.90	1.07	0.94	0.52	0.69	0.72	0.55	0.68	0.80	0.66	0.36	-0.16	0.55	0.55	0.87	0.53
Thr360	-0.87	-2.40	-3.06	-2.53	-0.55	-0.71	-0.49	-0.41	-0.32	-2.00	-1.19	-0.58	-1.23	-1.08	-2.22	-2.27	-0.94
Ser361	-0.14	-1.16	-1.43	-1.21	-0.43	-0.39	-0.48	0.30	-0.20	-0.99	-0.47	-0.35	-0.13	-0.28	-0.29	-1.10	-0.56
Ser394	0.00	0.22	0.18	0.21	0.05	0.25	0.34	0.50	0.79	0.25	0.46	0.41	0.02	0.26	0.31	0.23	0.32
Ser395	0.01	0.80	0.68	0.77	0.76	0.53	1.17	0.69	0.82	0.87	0.75	0.70	0.01	0.42	0.30	0.82	1.06
Ser408	-0.02	-1.23	-0.99	-1.18	-0.82	-1.65	-1.06	-1.76	-0.83	-1.37	-0.89	-2.84	-0.08	-1.38	-1.02	-1.28	-1.75
His410	0.00	-0.19	-0.47	-0.25	0.80	0.64	0.67	0.42	0.47	-0.03	0.43	0.92	0.02	0.26	0.60	-0.14	0.42
Cys411	-0.02	-0.23	-0.29	-0.24	-0.10	-0.27	-0.16	-0.30	-0.30	-0.19	-0.12	-0.34	0.01	-0.20	-0.17	-0.22	-0.09
Gln419	0.00	-0.16	-0.30	-0.19	-0.49	-0.35	-0.10	-0.05	0.10	-0.09	-0.05	0.27	0.03	0.26	0.21	-0.14	0.12
Ser426	-0.01	-0.04	-0.17	-0.06	-0.27	-0.24	-0.17	-0.17	-0.31	0.04	-0.20	-0.32	-0.01	-0.39	-0.10	-0.01	0.25
Cys430	0.00	-0.33	-0.21	-0.31	-0.31	-0.41	-0.34	-0.39	-0.51	-0.40	-0.23	-0.62	-0.04	0.00	-0.40	-0.35	-0.59
Thr444	0.08	-1.10	-0.97	-1.08	-1.18	-1.01	-0.89	-1.20	-1.01	-1.18	-0.98	-1.07	-0.06	-0.99	-0.94	-1.13	-1.40
His537	-6.24	-3.56	-3.11	-3.47	-2.20	-2.74	-3.22	-2.13	-4.25	-3.83	-2.90	-1.17	-1.47	-2.18	-4.18	-3.65	-4.54
Tyr539	-0.02	-0.42	-0.41	-0.42	-0.31	-0.31	-0.34	-0.30	-0.36	-0.42	-0.42	-0.34	-0.10	-0.40	-0.24	-0.42	-0.44
Ser540	0.01	-1.00	-1.12	-1.03	-0.78	-0.76	-1.01	-0.69	-1.11	-0.93	-1.29	-0.79	-0.12	-0.63	-1.02	-0.98	-0.74
Cys543	0.00	-0.59	-0.53	-0.58	-0.54	-0.48	-0.62	-0.48	-0.53	-0.63	-0.69	-0.50	-0.04	-0.51	-0.58	-0.61	-0.74
Asn545	0.00	-0.25	-0.23	-0.24	-0.20	-0.09	-0.34	-0.10	-0.19	-0.26	-0.20	-0.18	-0.03	-0.21	-0.30	-0.25	-0.28
His560	-0.06	-0.85	-0.65	-0.81	-0.44	-0.43	-0.39	-0.49	-0.45	-0.96	0.18	0.04	-0.03	-0.17	-0.65	-0.89	-1.27
Gln563	0.04	-0.53	-0.49	-0.52	-0.06	-0.34	-0.16	-0.29	-0.16	-0.55	-0.36	-0.22	-0.02	-0.23	-0.07	-0.54	-0.62

Table S5.4 Average residue decomposition energies of non-polar rERα pocket residues. The color gradient goes from blue to red as the values change from negative to positive.

Res. #	E2	PFPeA	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUnA	PFDoA	PFBS	PFHxS	PFOS	PFOSA	PFOSA AcOH	Et-PFOSA AcOH	KETO H	GenX
Met356	0.04	2.27	2.45	2.30	1.15	1.65	0.74	1.02	1.17	2.15	1.57	0.68	-0.86	0.71	0.21	2.23	1.86
Leu358	-0.13	1.19	1.42	1.24	0.79	1.36	1.13	1.07	1.28	1.06	1.03	0.62	-0.53	0.96	0.55	1.15	0.69
Leu359	-2.81	1.64	2.69	1.85	-1.10	0.12	0.07	-0.16	-0.48	1.02	0.05	-0.73	-6.46	-0.69	-3.32	1.43	-0.65
Met362	-2.21	-5.53	-5.58	-5.54	-4.17	-4.35	-3.99	-2.73	-5.69	-5.50	-4.43	-3.54	-1.22	-2.09	-4.65	-5.52	-5.43
Ala363	-1.87	-2.60	-2.60	-2.60	-2.60	-2.10	-2.69	-1.08	-1.91	-2.60	-3.08	-1.29	-1.40	-2.07	-2.07	-2.60	-2.61
Trp396	-0.19	1.73	1.55	1.69	1.93	1.96	2.31	1.99	1.88	1.84	2.11	1.96	-0.72	1.56	1.14	1.77	2.13
Leu397	-0.89	1.32	1.16	1.29	0.98	0.78	1.55	0.98	1.29	1.42	1.61	1.38	-1.25	1.17	1.01	1.35	1.68
Val399	-0.09	0.84	0.84	0.84	0.76	1.14	0.79	1.43	0.79	0.85	0.65	1.25	-0.13	1.37	0.88	0.85	0.86
Leu400	-5.88	-0.15	-0.75	-0.27	-0.24	0.06	-0.41	0.81	0.31	0.21	-0.59	0.54	-2.64	-3.17	-0.95	-0.03	1.18
Met401	-2.07	0.50	0.43	0.48	-0.47	-1.33	-1.16	-1.72	-0.42	0.54	-1.28	-0.82	-1.51	-2.60	-2.13	0.51	0.64
Ile402	-0.33	0.06	-0.04	0.04	-0.10	-0.17	-0.16	-0.31	-0.14	0.12	0.01	0.11	-0.07	-0.51	-0.23	0.08	0.27
Gly403	-0.87	-2.76	-2.93	-2.80	-2.99	-3.80	-3.13	-4.44	-3.07	-2.66	-2.39	-3.31	-0.14	-6.61	-4.12	-2.73	-2.38
Leu404	-2.32	-4.61	-4.45	-4.58	-4.50	-5.05	-5.13	-3.72	-4.63	-4.70	-4.96	-3.69	-2.12	-5.89	-5.79	-4.64	-4.96
Ile405	-0.16	-0.32	-0.37	-0.33	-0.35	-0.69	-0.47	-0.68	-0.30	-0.28	-0.42	-0.22	-0.10	-0.98	-0.61	-0.31	-0.20
Trp406	-0.34	-1.75	-1.66	-1.73	-1.74	-2.59	-1.82	-2.78	-2.26	-1.80	-1.68	-2.41	0.02	-2.18	-1.77	-1.76	-1.94
Ile409	-0.02	0.04	-0.10	0.01	0.26	0.34	0.22	0.32	0.19	0.13	0.24	0.38	0.02	0.17	0.39	0.07	0.36
Pro412	-0.03	-0.16	-0.19	-0.16	-0.10	-0.37	-0.16	-0.42	-0.29	-0.14	-0.07	-0.31	0.02	-0.17	-0.21	-0.15	-0.09
Gly413	-0.04	0.11	0.01	0.09	0.29	0.21	0.20	0.16	0.01	0.16	0.43	0.15	0.03	0.12	0.40	0.13	0.32
Leu415	-0.78	-1.96	-1.71	-1.91	-1.83	-0.88	-1.60	-2.52	-3.23	-2.11	-2.05	-7.60	-0.66	-1.69	-1.84	-2.01	-2.51
Ile416	-0.02	2.41	2.76	2.48	3.81	1.83	1.43	1.92	3.79	2.21	1.55	-1.49	-0.18	2.83	2.82	2.34	1.65
Phe417	-3.50	-0.12	0.60	0.02	-0.47	-0.39	-2.51	-0.56	-2.43	-0.55	-1.64	-4.04	-3.34	0.12	1.95	-0.26	-1.69
Ala418	0.02	0.98	1.11	1.01	0.79	0.27	0.67	0.61	1.14	0.91	0.83	1.41	-0.06	0.56	-0.01	0.96	0.70
Gly428	-0.08	0.59	0.94	0.66	0.10	0.02	-0.06	-0.06	-0.29	0.38	0.44	-0.47	-0.16	0.15	0.31	0.52	-0.18
Val431	-0.18	-1.72	-1.57	-1.69	-1.07	-0.86	-1.62	-0.97	-1.49	-1.82	-1.82	-1.51	-0.16	-0.86	-1.65	-1.76	-2.06
Gly433	0.11	-0.69	-0.47	-0.65	-0.52	-0.81	-0.87	-0.55	-1.01	-0.83	-0.19	-0.12	-0.62	-0.65	-1.61	-0.74	-1.18
Met434	-0.22	0.93	0.97	0.94	1.05	1.08	0.71	0.72	0.15	0.91	0.48	0.93	-1.02	0.40	0.76	0.93	0.86
Ala435	-0.12	0.65	0.63	0.65	0.23	0.59	0.18	0.46	0.39	0.66	0.40	0.58	-0.03	0.25	0.51	0.65	0.68
Ile437	-0.23	-0.50	-0.61	-0.52	-1.89	-0.77	-1.47	-1.20	-1.82	-0.43	-1.30	-0.54	-1.52	-1.15	-0.77	-0.47	-0.24
Phe438	-2.14	-0.59	-0.25	-0.52	-0.53	-0.57	-0.75	-0.61	-1.89	-0.79	-0.61	-0.95	-0.77	-0.33	-0.79	-0.66	-1.32
Met440	0.05	-0.50	-0.57	-0.51	-0.35	-0.26	-0.87	-0.12	-0.58	-0.45	-0.51	-0.11	-0.16	-0.33	-0.31	-0.48	-0.33
Leu441	-0.90	-1.28	-0.98	-1.22	-1.66	-1.50	-1.01	-1.85	-1.77	-1.46	-1.22	-1.49	-0.44	-1.11	-1.22	-1.34	-1.93
Leu442	-0.12	-0.12	-0.03	-0.10	-0.44	-0.09	-0.04	0.08	-0.15	-0.17	-0.12	-0.15	-0.05	0.01	-0.11	-0.14	-0.31
Ala443	0.04	-0.35	-0.34	-0.35	-0.36	-0.31	-0.41	-0.28	-0.37	-0.35	-0.42	-0.13	-0.04	-0.37	-0.41	-0.35	-0.36
Val445	0.01	-0.68	-0.57	-0.66	-0.77	-0.58	-0.52	-0.79	-0.63	-0.75	-0.55	-0.88	-0.04	-0.50	-0.57	-0.70	-0.93
Leu538	-2.53	-1.64	-1.43	-1.60	-2.22	-2.08	-2.73	-2.21	-2.37	-1.77	-2.38	-2.26	-1.80	-2.68	-2.57	-1.69	-2.12
Ile541	-0.37	-1.22	-0.95	-1.17	-1.10	-1.14	-1.66	-1.38	-1.53	-1.38	-1.72	-1.07	-0.54	-1.35	-2.42	-1.28	-1.82
Val547	0.01	0.26	0.36	0.28	-0.38	-0.05	-0.30	-0.23	-0.03	0.20	-0.40	-0.09	-0.03	-0.15	-0.51	0.24	0.04
Pro548	-0.01	0.49	0.35	0.46	1.18	0.71	1.08	0.69	0.71	0.57	1.03	0.67	-0.04	0.84	0.85	0.52	0.79
Leu549	-0.11	-1.64	-1.88	-1.69	-0.91	-0.95	-1.11	-0.82	-0.88	-1.50	-1.13	-0.78	-0.07	-1.05	-1.40	-1.59	-1.12
Leu553	-0.20	-0.96	-1.16	-1.00	-0.71	-0.89	-0.66	-0.45	-0.37	-0.84	-0.69	-0.41	-0.27	-0.78	-0.68	-0.92	-0.52
Leu557	-0.13	-0.86	-1.04	-0.90	-0.45	-0.51	-0.79	-0.29	-0.32	-0.75	-0.45	-0.34	-0.07	-0.76	-1.02	-0.82	-0.47
Gly559	-0.04	-0.37	-0.32	-0.36	-0.61	-0.15	-0.56	-0.37	-0.31	-0.40	-0.57	-0.26	-0.01	-0.58	-0.59	-0.38	-0.48
Leu562	-0.01	-0.44	-0.39	-0.43	-0.35	-0.35	-0.68	-0.29	-0.62	-0.47	-0.67	-0.34	-0.03	-0.19	-0.17	-0.45	-0.55

Table S5.5 Average residue decomposition energies of charged rERβ pocket residues. The color gradient goes from blue to red as the values change from negative to positive.

Res. #	E2	PFBA	PFPeA	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUnA	PFDoA	PFBS	PFHxS	PFOS	PFOSA	PFOSAAcOH	Et-PFOSAAcOH	8:2FTOH	GenX
Asp230	0.11	24.02	21.53	18.68	26.48	16.99	23.22	20.11	19.98	23.44	21.19	18.35	20.86	-0.34	21.89	18.73	-0.27	27.35
Glu232	-1.76	21.82	20.96	19.79	23.43	19.49	20.66	22.38	20.30	34.85	21.16	21.79	22.01	0.12	22.53	21.79	-12.72	23.98
Glu264	-0.20	22.86	21.73	24.35	27.13	18.88	20.59	19.40	18.95	21.03	21.73	20.13	18.73	0.45	20.21	21.10	0.19	29.43
Asp292	0.00	19.1	23.40	21.94	18.78	30.75	23.90	34.51	27.65	22.46	22.43	31.39	26.51	0.15	26.35	31.02	0.28	18.41
Glu293	0.02	24.78	26.54	24.98	22.46	29.57	28.67	30.64	25.66	27.02	27.43	28.38	29.11	0.11	25.79	24.36	0.38	20.52
Glu302	-0.13	22.21	23.95	27.37	21.36	27.15	25.21	21.26	25.38	20.16	23.19	23.31	25.73	0.38	21.29	25.10	0.62	22.00
Asp305	0.20	21.94	26.61	27.83	20.21	34.97	24.30	28.39	31.12	25.60	25.16	28.36	28.17	0.27	27.86	36.97	0.52	20.39
Asp402	-0.17	24.5	24.13	25.91	24.58	20.18	23.17	18.13	20.48	16.60	24.91	18.78	21.57	-0.03	19.48	19.09	0.71	24.31
Asp417	-0.11	19.82	20.52	18.48	21.95	13.49	18.33	14.36	15.24	16.36	19.29	14.04	16.46	-0.30	15.97	14.25	0.09	21.66
Glu421	-0.08	17.03	16.89	17.69	21.34	13.04	15.89	12.85	13.84	14.00	16.74	13.17	14.37	-0.08	14.53	13.01	0.16	21.55
Lys231	-0.56	-24.01	-21.87	-18.44	-25.03	-17.60	-24.38	-22.52	-20.67	-26.77	-19.95	-20.34	-24.10	0.35	-22.97	-20.75	0.47	-22.61
Arg273	-2.00	-20.24	-20.40	-18.50	-23.17	-22.12	-22.82	-26.62	-23.28	-55.35	-19.74	-25.89	-26.20	-0.12	-25.58	-25.13	3.34	-19.24
Lys280	-0.28	-19.09	-21.50	-19.51	-15.66	-26.96	-22.79	-28.54	-20.52	-31.39	-23.07	-28.65	-22.96	-0.12	-25.58	-26.55	0.23	-16.69
Lys408	0.21	-21.95	-20.55	-22.29	-34.04	-16.70	-20.10	-16.58	-17.37	-14.91	-21.12	-16.42	-19.38	0.05	-17.47	-15.81	-0.56	-25.37
Lys410	0.41	-28.7	-26.37	-26.91	-22.77	-23.99	-32.59	-19.86	-25.45	-15.86	-29.90	-19.12	-22.90	-0.05	-20.67	-20.05	-0.60	-22.90
Lys411	0.22	-17.88	-16.73	-16.95	-18.88	-14.63	-16.85	-13.94	-14.55	-12.42	-17.56	-13.05	-14.74	0.05	-14.13	-13.25	-0.36	-17.11

Table S5.6 Average residue decomposition energies of polar rERβ pocket residues. The color gradient goes from blue to red as the values change from negative to positive.

Res. #	E2	PFBA	PFPeA	PFHxA	PFHpA	PFOA	PFNA	PFDA	PFUnA	PFDoA	PFBS	PFHxS	PFOS	PFOSA	PFOSAcOH	Et-PFOSAcOH	8:2FTOH	GenX
Ser224	-0.02	-0.26	-0.41	-0.84	-0.34	-0.75	-1.71	-0.97	-1.43	-0.03	-0.79	-0.75	0.21	-0.37	-0.27	-0.73	-0.29	0.02
Thr226	-1.29	-8.38	-3.99	-1.81	-4.28	-0.49	-5.18	-1.41	-2.71	-1.09	-3.27	-0.98	-2.46	-6.28	-3.60	-1.85	-1.12	-3.99
Asn227	-0.05	-1.82	-1.32	-0.72	-1.35	-0.52	-1.51	-0.42	-1.19	0.30	-0.97	-0.09	-1.35	-0.08	-1.27	-0.50	-0.29	-0.93
Cys260	0.01	0.24	0.32	0.16	0.04	0.33	0.34	0.22	0.16	0.06	0.27	0.07	-0.05	0.00	-0.07	0.09	0.01	0.00
Cys261	-0.01	-0.41	0.29	0.13	-0.20	0.24	-0.07	0.23	0.46	0.54	-0.29	0.11	0.21	0.00	0.38	0.18	-0.05	-0.45
Ser274	-0.07	-0.92	-1.26	-0.90	-0.23	-2.42	-0.99	-1.25	-2.24	-2.29	-1.15	-2.28	-1.72	-0.06	-1.24	-2.60	-0.01	-0.60
Asn276	-0.01	0.41	0.09	0.34	0.21	0.83	1.08	1.05	1.58	0.82	0.57	1.21	0.64	-0.01	1.03	0.63	0.03	0.63
His277	-0.01	0.41	0.36	0.33	0.26	1.13	0.53	0.84	0.74	0.65	0.49	0.86	0.68	0.00	0.56	1.26	0.01	0.31
Ser284	-0.02	0.04	-0.15	0.30	0.11	0.20	0.79	0.34	0.63	0.02	1.21	0.79	0.16	-0.03	0.85	0.24	-0.13	0.00
Ser295	-0.06	0.5	0.23	1.07	0.10	-1.13	-0.69	-11.75	-1.97	-0.69	0.70	-1.87	-0.88	-0.06	-2.11	-1.27	-0.04	0.09
Cys296	-0.03	1.08	-0.34	-0.41	-0.45	-1.88	-1.62	-2.75	-1.44	-1.40	1.65	-1.56	-1.95	-0.10	-0.88	-0.71	-0.07	-0.17
Gln298	-0.50	-8.89	1.06	-5.67	-0.39	0.53	2.80	0.74	3.65	0.81	-13.35	0.49	1.08	-0.06	-0.53	0.16	0.06	-0.52
Thr311	-0.05	-0.53	-0.63	-0.85	-0.32	-1.42	-0.91	-0.76	-1.11	-0.88	-0.65	-0.70	-0.72	-0.04	-0.78	-1.77	-0.10	-0.36
His403	-6.54	-14.81	-17.93	-14.81	-11.00	-9.97	-14.23	-3.98	-10.49	-3.47	-14.64	-4.64	-8.13	-3.58	-5.22	-4.31	-2.00	-12.68
His405	-0.09	-0.08	0.14	-0.21	0.77	-0.09	0.12	0.13	-0.05	0.07	-0.20	-0.08	0.03	-0.08	-0.22	-0.06	-0.05	-0.21
Cys406	0.03	-1.34	-1.53	-1.86	-1.37	-0.91	-1.12	-0.64	-1.17	-0.49	-1.70	-0.75	-1.10	0.04	-0.89	-0.80	-0.12	-1.41
Tyr416	-0.05	0.35	0.27	0.61	0.68	0.38	0.16	-0.05	0.33	-0.05	0.22	-0.17	-0.11	-0.02	-0.42	-0.01	-0.03	0.21

Table S5.7 Average residue decomposition energies of non-polar rERβ pocket residues. The color gradient goes from blue to red as the values change from negative to positive.

Res. #	E2	PFBA	PFPrA	PFBtA	PFHpA	PFOA	PFNA	PFDA	PFUnA	PFDoA	PFBS	PFHxS	PFOS	PFOSA	PFOSAAcOH	N-PFOSAAcOH	8:2FTOH	GenX
Met222	-1.81	-1.53	0.13	-1.22	-0.81	0.40	-3.13	0.64	-2.84	0.49	-0.98	0.45	1.12	-2.17	1.30	0.34	-1.19	0.13
Met223	-0.09	0.38	0.66	0.10	1.07	-0.19	0.85	0.02	-0.08	0.98	0.70	0.24	0.74	-0.43	0.75	0.12	-0.07	1.63
Leu225	-2.48	-6.03	-3.51	-2.81	-3.61	-2.57	-4.71	-2.91	-4.40	-2.40	-4.05	-3.08	-3.55	-5.27	-5.39	-4.38	-3.86	-2.22
Leu228	-2.06	-2.34	-2.12	-1.13	-2.18	-1.12	-3.66	-2.02	-1.97	-1.59	-1.50	-1.60	-3.95	-0.07	-3.23	-2.19	-2.39	-1.60
Ala229	-2.11	-2.96	-2.21	-1.36	-3.29	-0.74	-3.98	-1.57	-2.00	-1.53	-2.19	-1.01	-2.65	-1.05	-2.67	-1.69	-1.73	-3.07
Trp262	-0.44	-0.3	0.73	0.40	-9.14	0.66	0.17	0.53	0.13	1.48	0.45	0.78	0.21	-0.62	0.17	0.79	-0.41	-14.39
Leu263	-2.03	-0.62	0.16	-0.62	-2.81	0.29	-0.41	0.01	-0.51	0.76	0.29	0.52	0.05	-1.23	-0.60	-0.20	-1.50	-2.30
Val265	-0.28	-0.24	0.01	-0.12	-0.85	0.36	0.05	0.39	0.29	0.96	0.03	0.42	0.22	-0.04	0.24	0.37	-0.25	-1.04
Leu266	-3.55	-1.3	-1.29	-1.56	-3.75	-0.26	-1.43	-0.75	-1.15	0.31	-1.15	-0.08	-1.45	-1.36	-1.31	-1.30	-3.34	-4.38
Met267	-2.40	-1.32	-1.55	-3.11	-2.95	0.52	-0.75	0.15	0.01	0.63	-1.96	0.47	0.12	-1.27	-0.52	-0.53	-2.32	-2.32
Leu268	0.02	-0.13	0.05	-0.05	-0.44	0.51	0.16	0.54	0.41	0.21	0.06	0.61	0.23	-0.04	0.32	0.36	-0.05	-0.59
Gly269	0.47	-0.75	-0.80	-0.76	-1.32	-0.14	-0.66	-0.31	-0.44	-2.24	-0.81	-0.30	-0.59	-0.07	-0.59	-0.47	0.15	-1.46
Leu270	-4.05	-1.1	-1.76	-1.59	-2.52	-1.01	-2.16	-2.34	-2.02	-2.41	-1.74	-1.99	-2.41	-1.31	-2.35	-2.71	-2.17	-1.95
Met271	-0.04	-0.82	-0.55	-1.11	-0.67	0.25	-0.51	-0.16	-0.09	-0.63	-0.93	0.02	-0.32	-0.09	-0.30	-0.67	-0.10	-0.61
Tpr272	0.00	-0.45	-0.41	-0.41	-0.65	-0.09	-0.43	-0.28	-0.29	-1.70	-0.43	-0.27	-0.46	-0.02	-0.44	-0.19	0.15	-0.55
Val275	0.00	0.37	0.38	0.45	-0.04	1.13	0.63	1.09	0.76	0.59	0.50	1.00	0.58	0.00	0.66	1.24	0.01	0.35
Pro278	-0.02	0.27	0.27	0.36	0.06	1.72	0.41	0.99	0.35	0.80	0.33	0.89	0.77	0.01	0.36	-0.61	0.03	0.14
Gly279	-0.02	0.63	0.68	0.67	0.33	1.84	0.94	1.95	0.67	0.96	0.68	2.16	1.29	0.01	1.25	2.12	0.03	0.48
Leu281	-0.63	-0.34	-0.76	-0.84	-0.47	-1.58	-0.86	-2.57	-1.79	-3.32	-0.79	-2.53	-1.71	-0.13	-1.72	-4.23	-0.26	-0.72
Ile282	-0.02	0.35	0.15	0.26	0.71	0.54	0.69	1.05	0.66	1.74	0.47	0.96	1.00	-0.01	0.92	0.74	-0.31	-0.08
Phe283	-2.53	-0.84	-1.13	-1.34	-0.51	-0.14	-0.63	-0.56	-0.53	-2.30	-2.05	-0.30	-0.94	-0.11	-1.17	-0.32	-1.40	-1.30
Pro285	-0.02	-0.24	-0.34	0.18	0.06	-0.17	0.39	-0.18	0.18	-0.58	0.14	0.13	-0.13	0.00	0.43	0.35	-0.01	0.09
Gly294	-0.12	1.38	0.50	2.39	1.28	-2.53	-0.30	-10.31	-0.80	-1.64	1.91	-3.48	-1.46	-0.24	0.56	-2.74	0.02	0.99
Val297	-0.07	-6.9	-3.48	-1.74	-3.90	-1.33	-2.58	-0.45	-3.29	-2.89	-9.59	-0.73	-1.86	-0.81	-1.90	0.73	-0.80	-3.18
Gly299	-0.10	-4.9	-2.45	-2.24	0.03	2.66	0.09	-0.51	-0.37	0.18	-7.81	-2.32	-2.08	-0.08	-2.93	-0.44	-0.05	-2.26
Phe300	-0.14	1.66	2.97	3.39	1.15	3.09	1.77	0.91	3.54	0.54	2.29	1.46	3.58	-0.60	0.97	1.40	-0.01	1.62
Val301	0.01	0.3	0.97	0.22	0.61	1.66	-2.20	1.23	1.21	-0.24	0.82	1.49	-0.59	-0.07	0.76	1.99	-0.03	0.46
Ile303	-1.15	-2.7	-3.85	-6.86	-2.03	-3.28	-4.69	-0.43	-4.34	-2.66	-1.67	-1.76	-6.34	-0.91	-1.09	-2.31	-1.68	-2.52
Phe304	-0.95	-1.98	-2.86	-4.86	-1.63	-3.88	-3.40	-1.44	-3.39	-2.87	-2.98	-2.82	-5.49	-2.46	0.12	-3.43	-1.88	-1.72
Met306	-0.10	-0.9	-1.18	-2.20	-0.82	-1.23	-0.97	-0.77	-1.26	-0.67	-1.17	-1.28	-1.32	-0.09	-1.20	-1.09	-0.15	-0.92
Leu307	-1.10	-1.76	-2.68	-3.30	-2.15	-4.22	-2.19	-4.05	-3.96	-2.40	-2.75	-4.52	-2.85	-0.93	-4.22	-5.85	-1.00	-1.31
Leu308	-0.04	-0.71	-0.96	-1.17	-0.50	-2.09	-0.76	-0.86	-1.57	-0.68	-1.05	-0.99	-1.03	-0.06	-1.25	-2.09	-0.08	-0.34
Ala309	-0.03	-0.61	-0.77	-0.95	-0.56	-0.80	-0.71	-0.53	-0.76	-0.42	-0.69	-0.66	-0.80	-0.03	-0.75	-0.86	-0.06	-0.60
Ala310	-0.07	-0.82	-1.05	-1.18	-0.80	-1.29	-0.96	-1.20	-1.33	-0.82	-1.02	-1.30	-1.07	-0.05	-1.20	-1.46	-0.08	-0.81
Leu404	-2.29	-1.84	-1.85	-2.92	-3.92	-0.93	-1.81	-1.59	-1.71	-1.13	-1.94	-1.26	-1.87	-1.98	-1.71	-1.92	-1.81	-2.36
Met407	-0.51	-2.77	-3.05	-2.31	-3.10	-1.71	-3.10	-0.98	-1.08	-0.59	-3.11	-0.93	-1.48	-0.76	-0.74	-0.71	-0.43	-3.64
Met409	0.02	-0.81	-0.76	-0.95	-0.96	-0.60	-0.88	-0.53	-0.76	-0.45	-1.05	-0.55	-0.71	0.01	-0.62	-0.56	-0.04	-0.89
Met412	-0.05	-1.47	-0.08	-0.09	0.06	-1.03	-0.75	-0.85	-0.90	-0.25	-0.52	-0.48	-0.63	-0.39	-0.64	-0.40	-0.07	-0.36
Val413	-0.10	1.99	1.68	1.46	2.45	0.82	1.39	0.71	0.80	0.56	0.56	0.60	0.86	-0.14	0.77	0.32	-0.03	1.44
Pro414	-0.03	0.63	0.65	-0.21	0.12	0.18	0.15	0.33	0.33	0.26	1.08	0.23	0.34	-0.03	0.16	0.45	0.01	-0.27
Leu415	-0.07	-2.71	-2.35	-2.16	-3.68	-0.73	-1.85	-1.02	-1.55	-1.63	-1.92	-0.88	-1.51	-0.42	-1.76	-0.94	-0.09	-2.89

APPENDIX B

SUPPORTING FIGURES

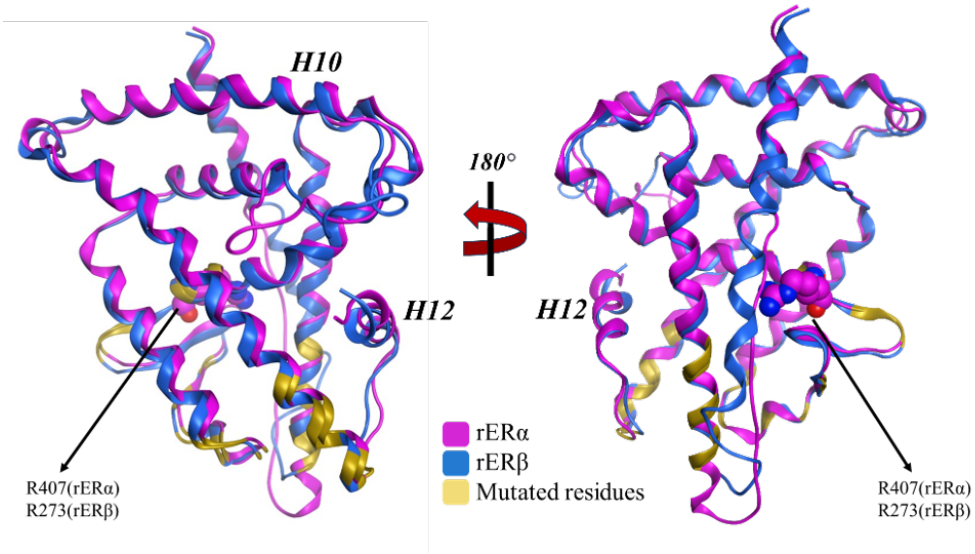


Figure S5.1 The overlap of rERα and rERβ LBDs is shown. Van der Walls ball representation was used for the arginine residues used in pharmacophore docking. The locations of mutated residues are shown in yellow. The volume of the binding pockets is 85 Å^3 and 92 Å^3 for rERα and rERβ, respectively. The mutated residues between two isoforms with numbering of rERα /rERβ are: V353/A219, T354/N220, M355/V221, T357/M223, L358/S224, S361/N227, M362/L228, S394/C260, S395/C261, I402/L268, I405/M271, I409/V275, H410/N276, C411/H277, A418/S284, Q419/P285, I422/S288, D424/S290, S426/D292, D429/S295, E432/Q298, M434/F300, A435/V301, T444/A310, V445/T311, E536/D402, Y539/H405, S540/C406, I541/M407, C553/M409, N545/K411, K546/M412, G559/A418, R561/I420,L562/E421, Q563/M422.

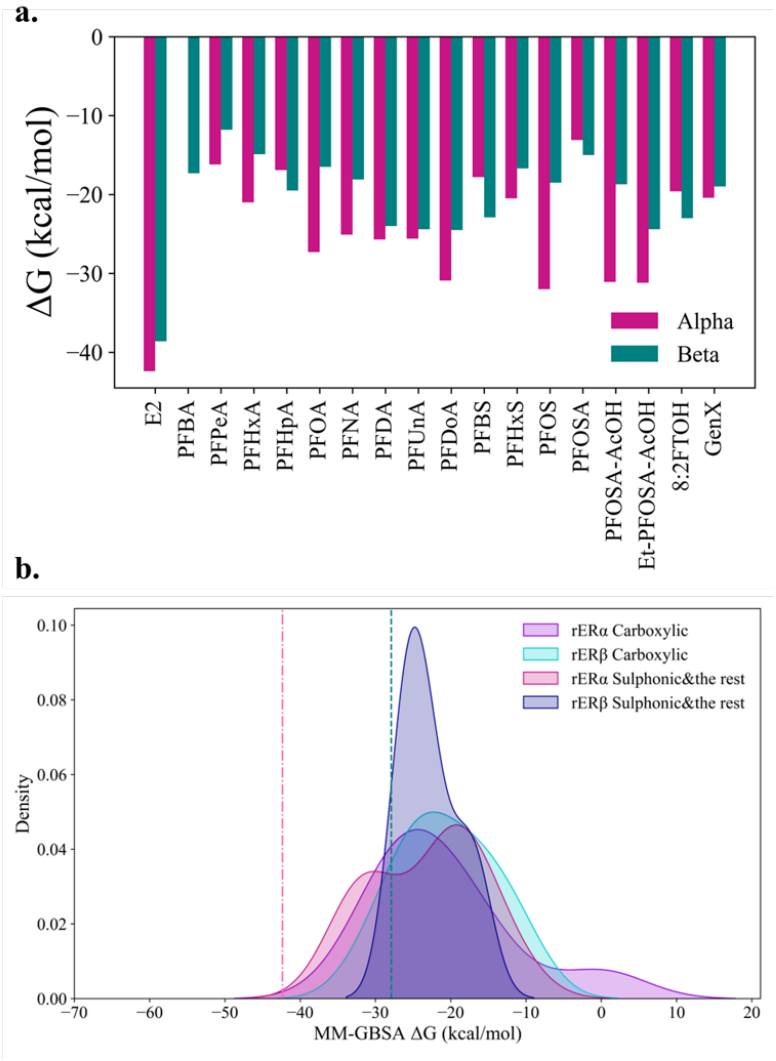


Figure S5.2 (a) MM-GBSA binding energies of rERα and rERβ proteins. (b) The kde distribution of MM-GBSA energies with respect to the PFAS type: carboxylic, and sulphonic along with the rest of the PFAS. The pink dashed line corresponds to E2 binding energy to rERα and blue dotted line indicates the binding energy of E2 to rERβ.

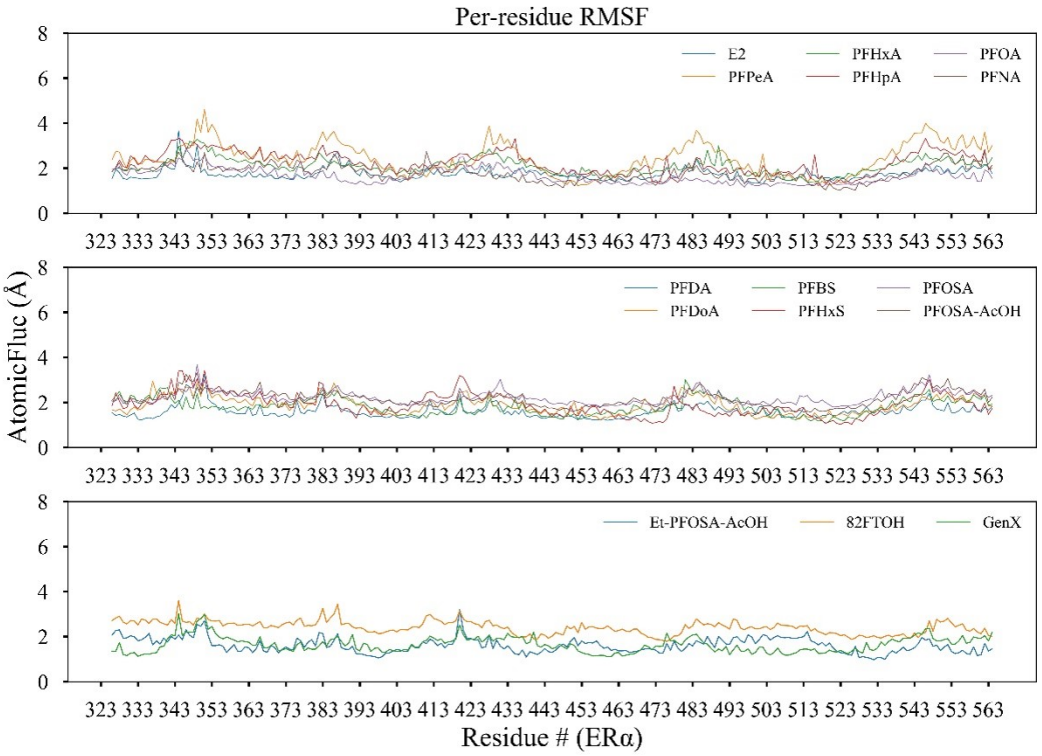


Figure S5.3 Per-residue root-mean square fluctuation (RMSF) of rERα residues of the first simulation sets.

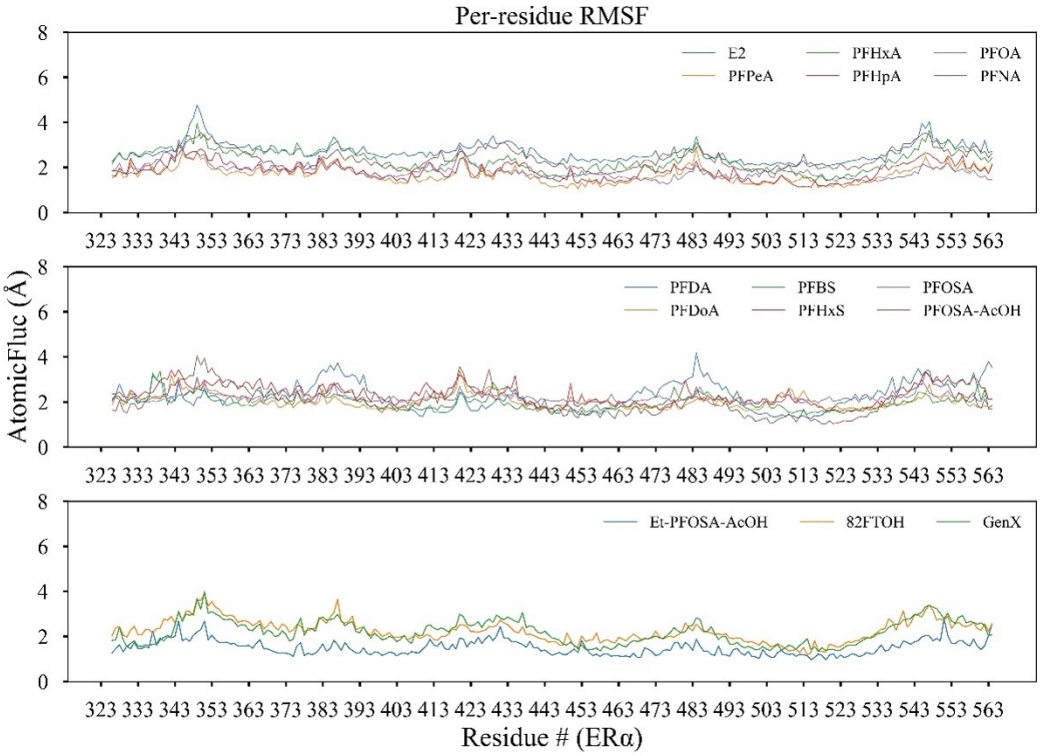


Figure S5.4 Per-residue root-mean square fluctuation (RMSF) of rERα residues of the second simulation sets.

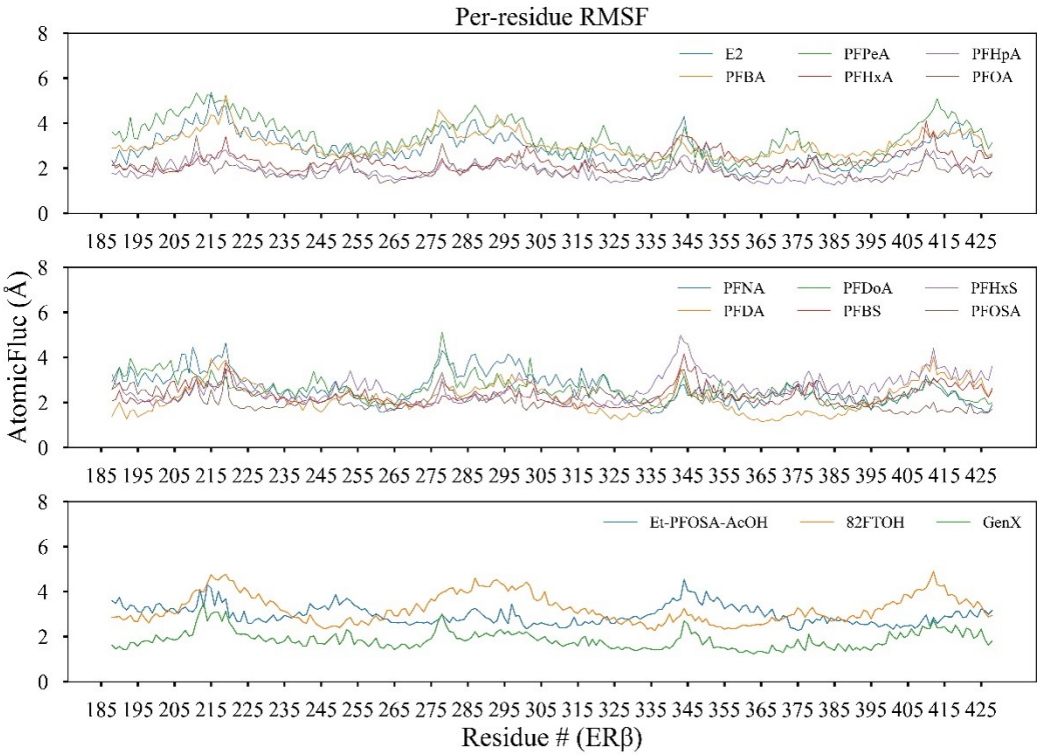


Figure S5.5 Per-residue root-mean square fluctuation (RMSF) of rERβ residues of the first simulation sets.

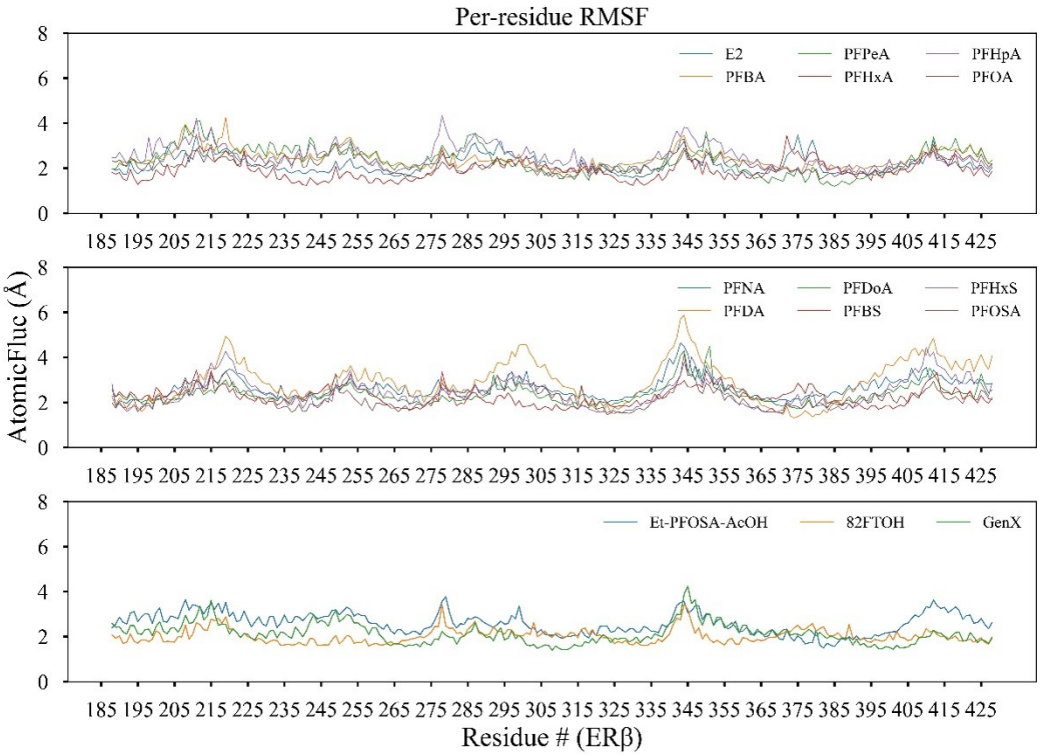


Figure S5.6 Per-residue root-mean square fluctuation (RMSF) of rERβ residues of the second simulation sets.

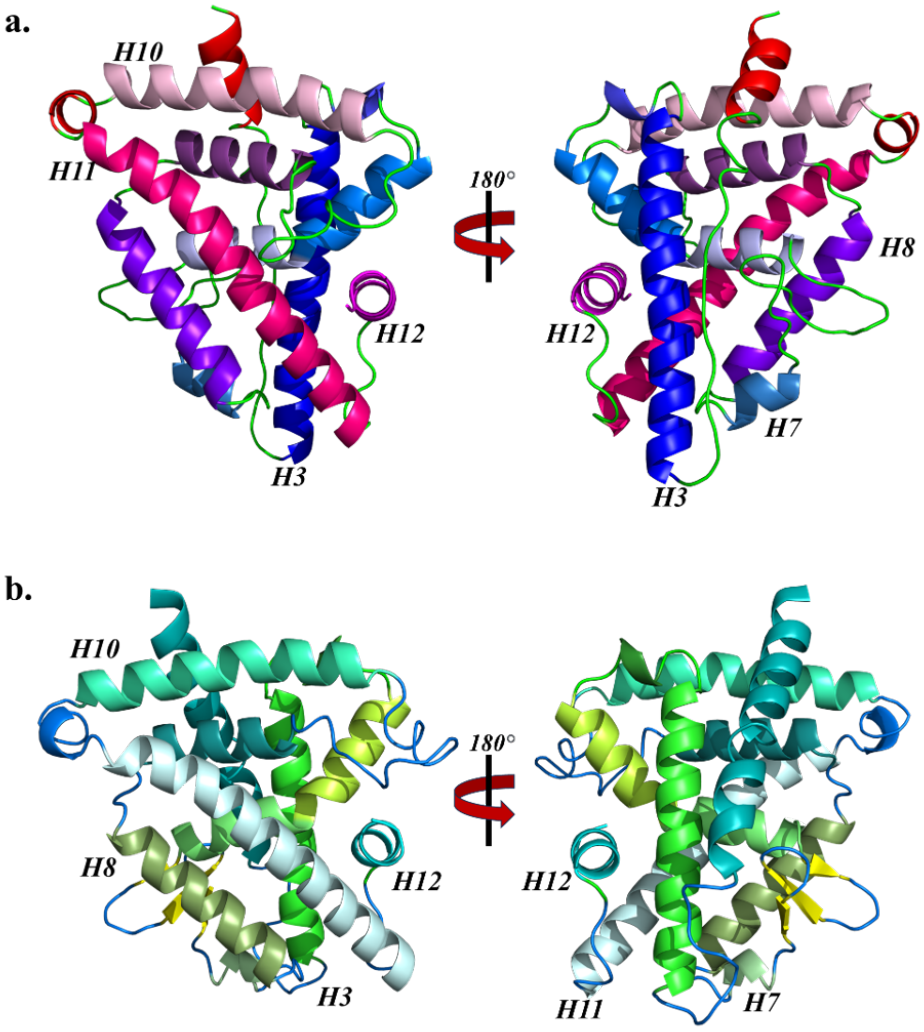


Figure S5.7 The helix numbering of (a) rERα and (b) rERβ LBDs is used for hydrogen bond analysis.

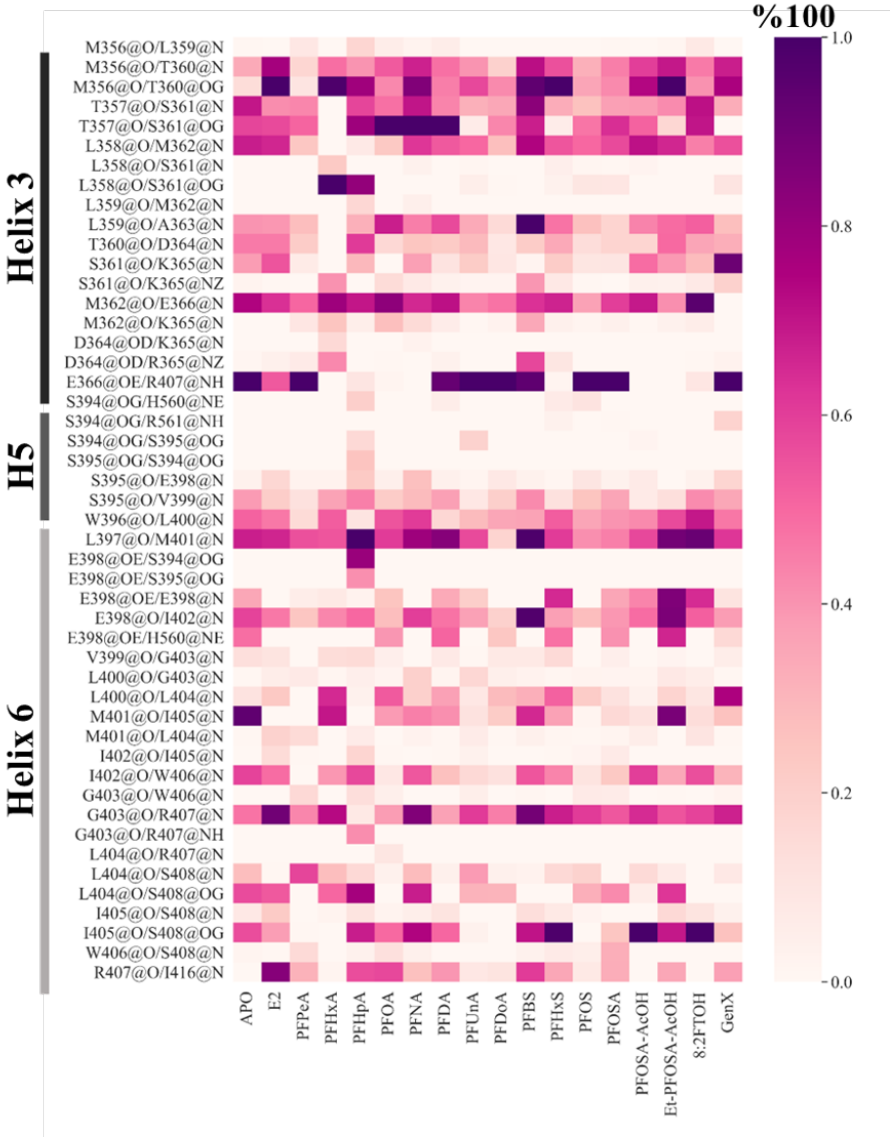


Figure S5.8 Hydrogen bond heatmap for rERα Helix 3,5, and 6. The residue and atom pairs that form hydrogen bonding are shown with the following nomenclature: Res1@Atom1/Res2@Atom2.

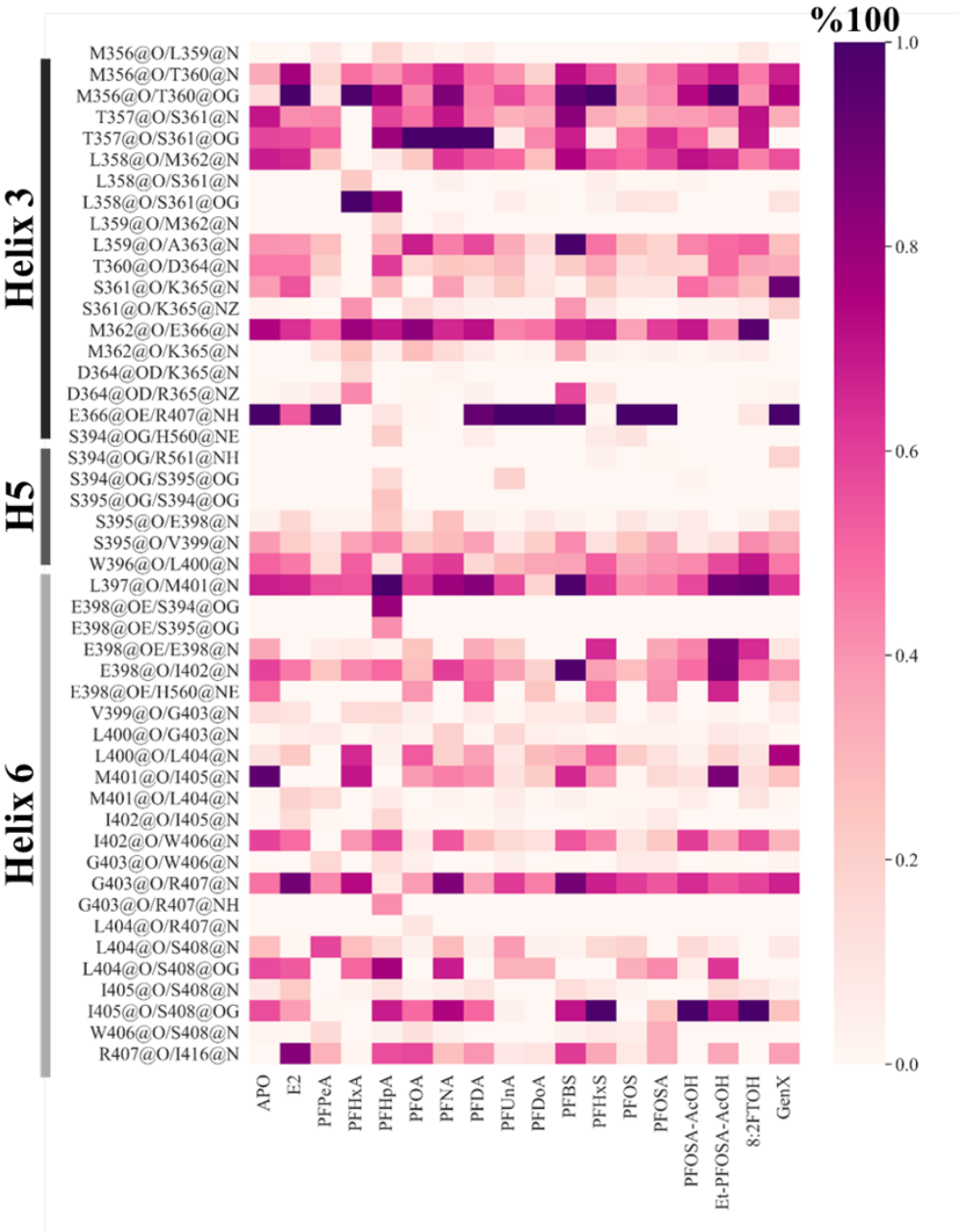


Figure S5.9 Hydrogen bond heatmap for rERα Helix 3,5, and 6. The residue and atom pairs that form hydrogen bonding are shown with the following nomenclature: Res1@Atom1/Res2@Atom2.

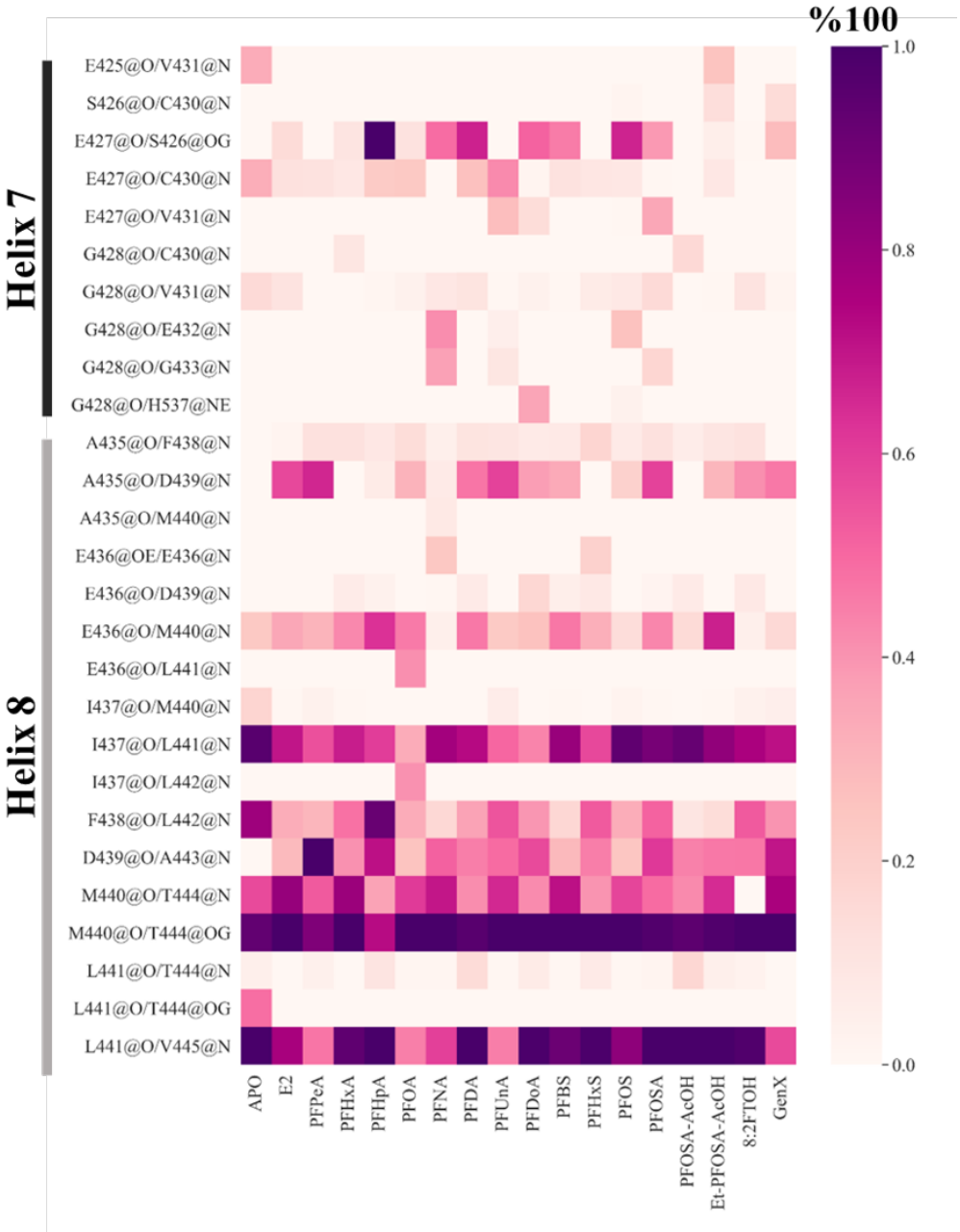


Figure S5.10 Hydrogen bond heatmap for rERα Helix 7 and 8. The residue and atom pairs that form hydrogen bonding are shown with the following nomenclature: Res1@Atom1/Res2@Atom2.

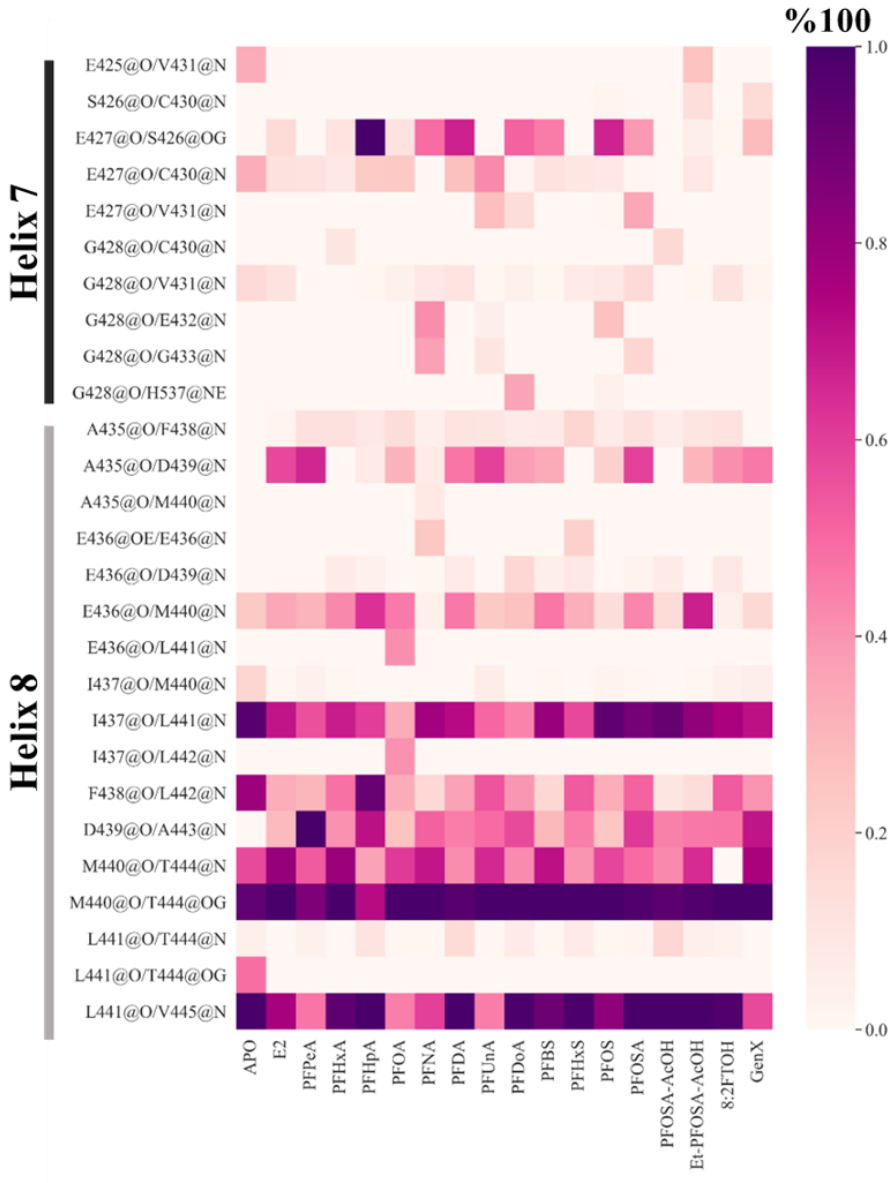


Figure S5.11 Hydrogen bond heatmap for rERα Helix 11 and 12. The residue and atom pairs that form hydrogen bonding are shown with the following nomenclature: Res1@Atom1/Res2@Atom2.

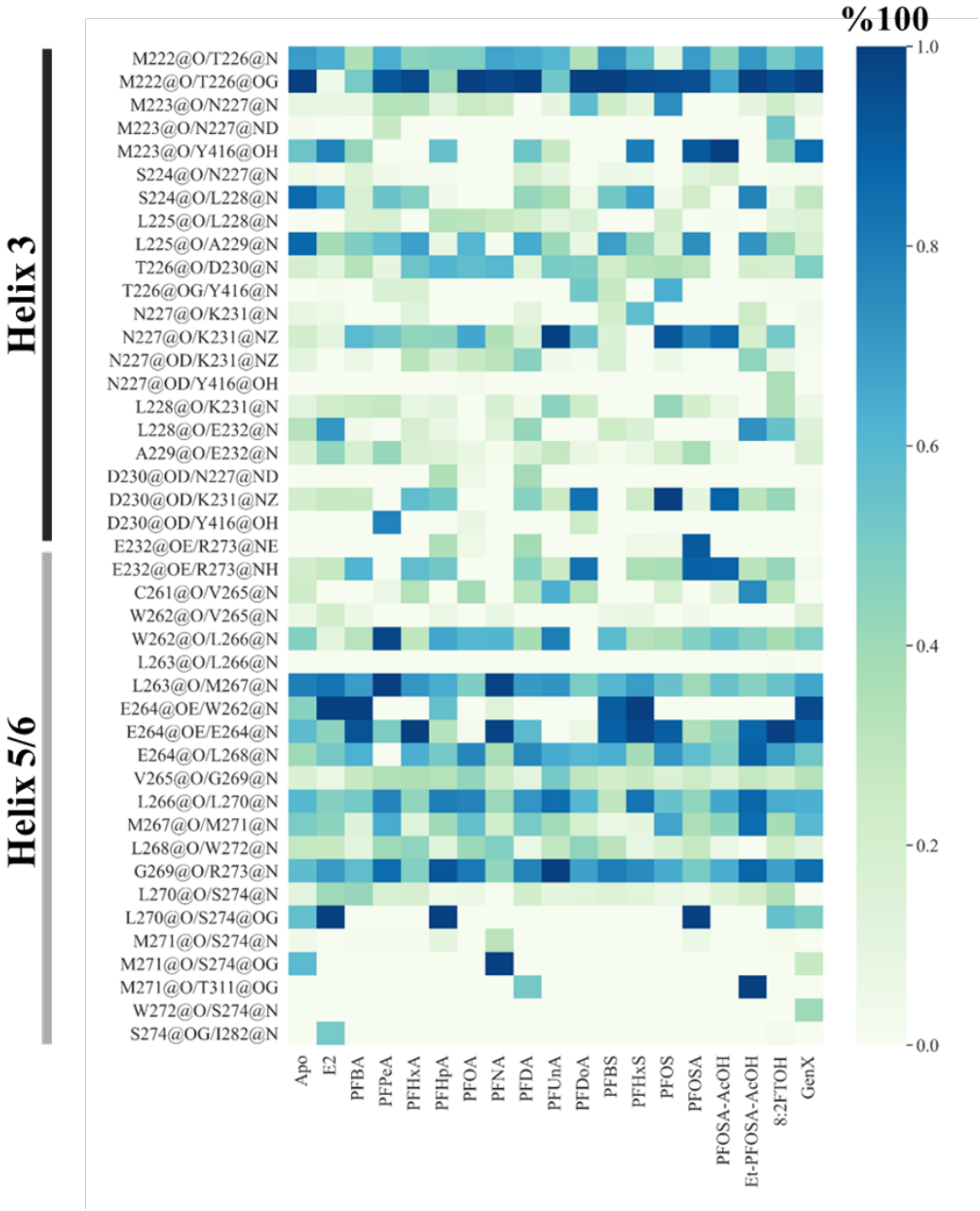


Figure S5.12 Hydrogen bond heatmap for rERβ Helix 3,5, and 6. The residue and atom pairs that form hydrogen bonding are shown with the following nomenclature: Res1@Atom1/Res2@Atom2.

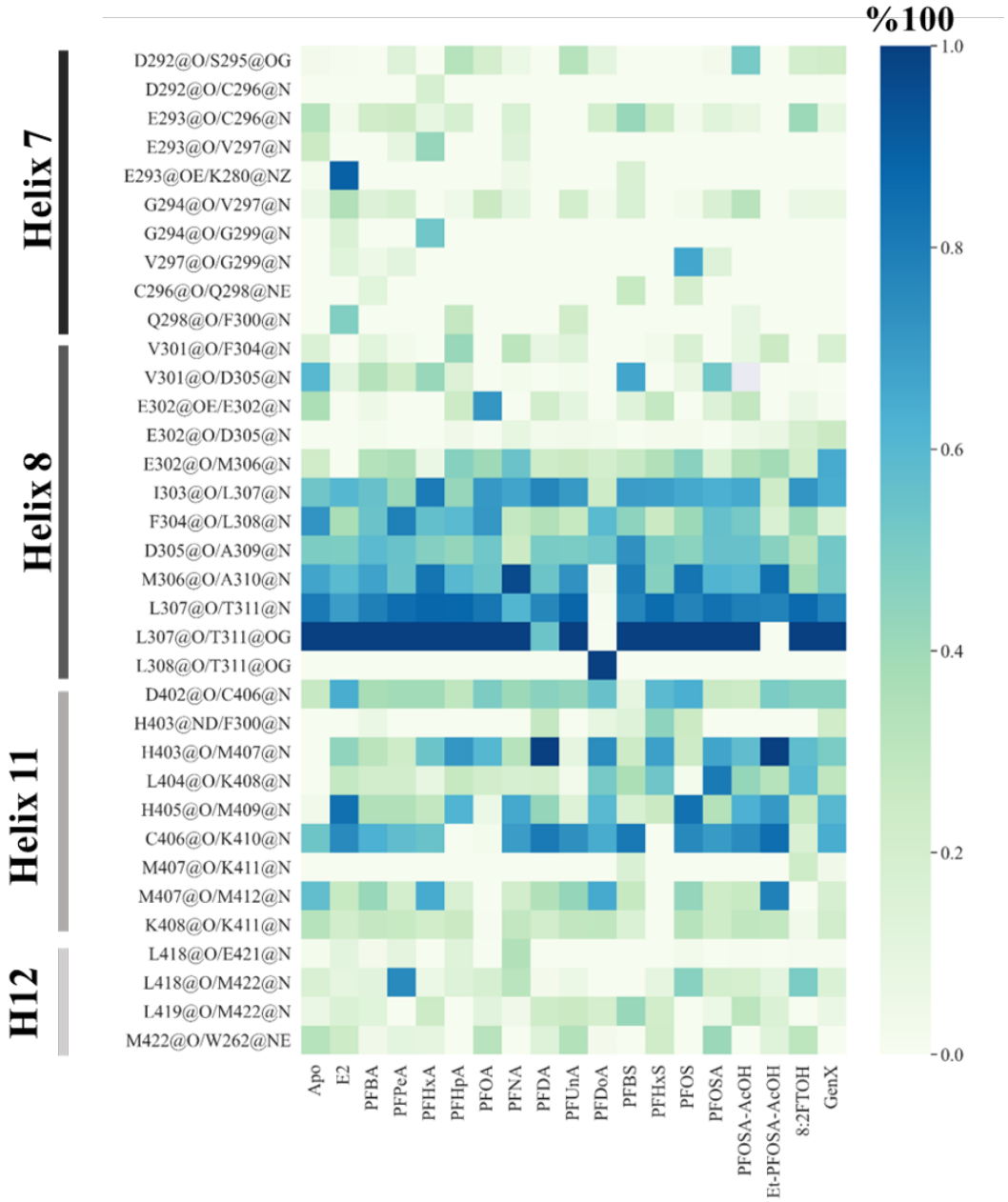


Figure S5.13 Hydrogen bond heatmap for rERβ Helix 7, 8, 11, and 12. The residue and atom pairs that form hydrogen bonding are shown with the following nomenclature: Res1@Atom1/Res2@Atom2.

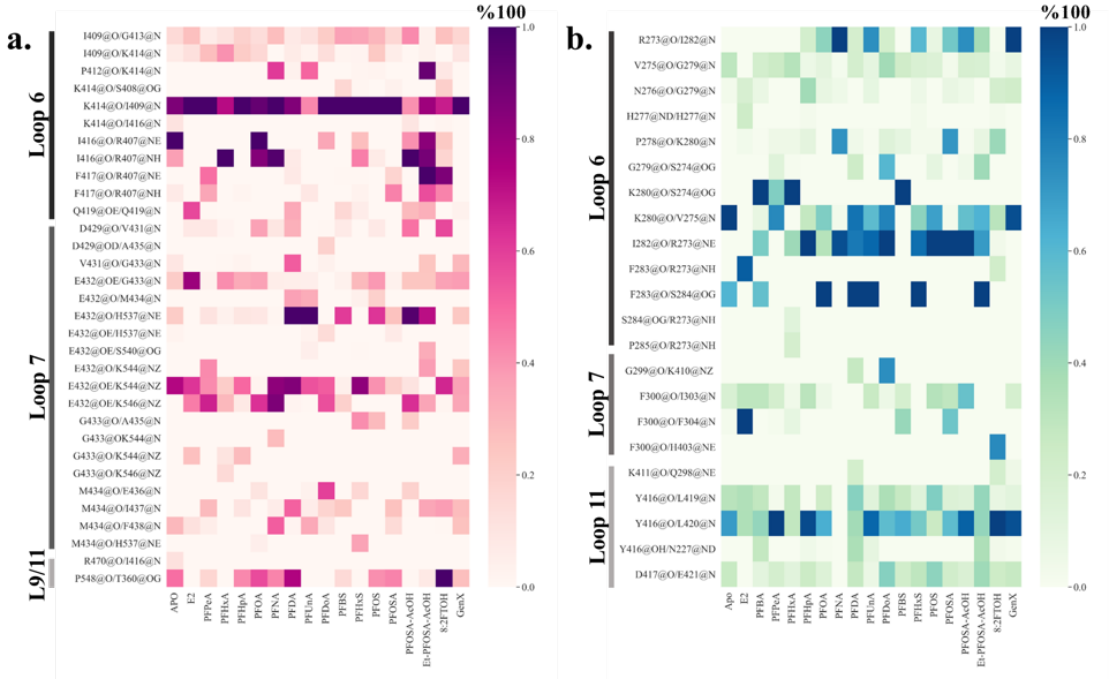


Figure S5.14 Hydrogen bond heatmap of loop regions of (a) rERα and (b) rERβ. The residue and atom pairs that form hydrogen bonding are shown with the following nomenclature: Res1@Atom1/Res2@Atom2.

Figure S5.15 Comparison of the orientation of investigated PFAS in rERα and rERβ binding pockets. The poses were obtained by clustering the last 5 ns of the simulations, and the most populated cluster was selected. The beta ones are not looking towards the Arg.

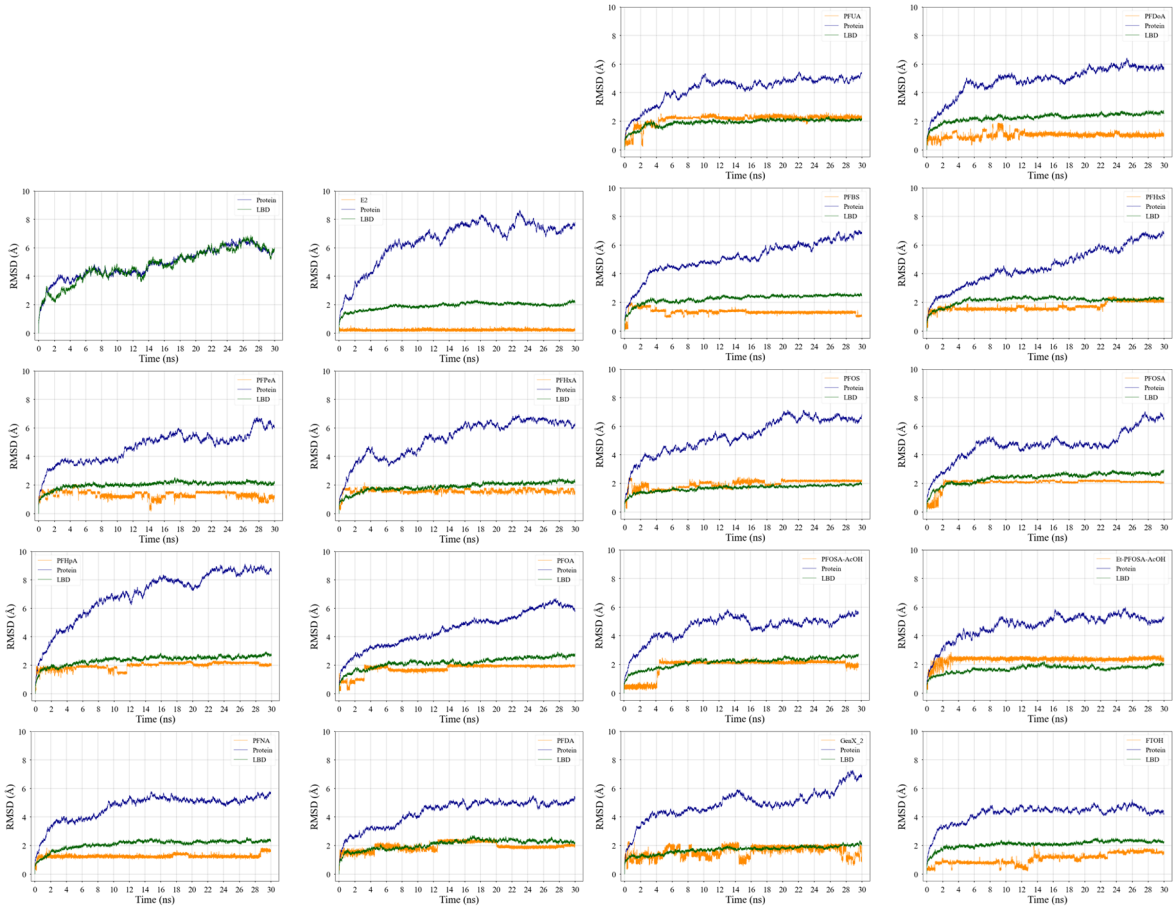


Figure S5.16 RMSD of PFAS bound to ERα-LBD, primary simulation set.

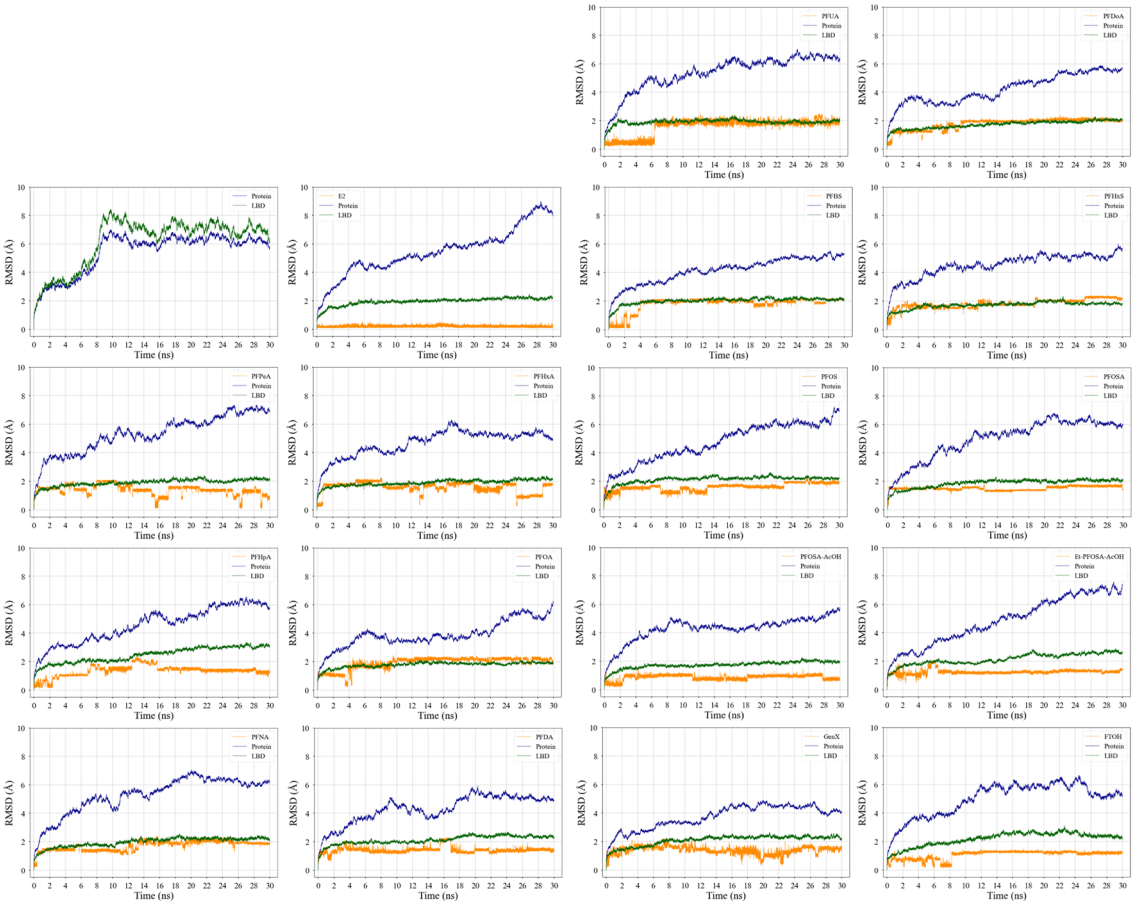


Figure S5.17 RMSD of PFAS bound to ERα-LBD, duplicate simulation set.

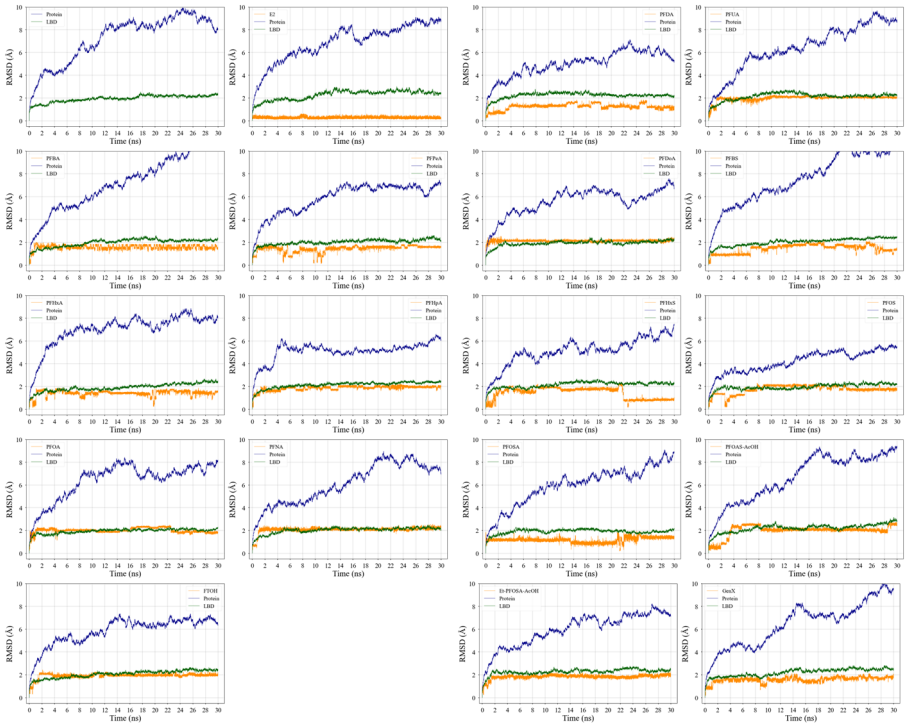


Figure S5.18 RMSD of PFAS bound to ERβ-LBD, primary simulation set.

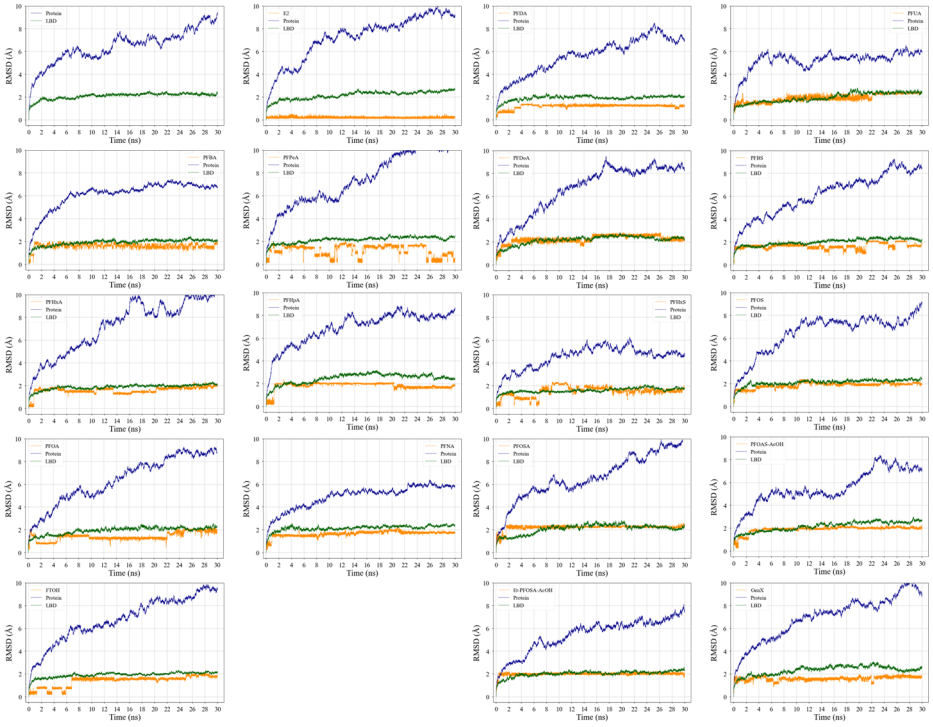


Figure S5.19 RMSD of PFAS bound to ERβ-LBD, duplicate simulation set.

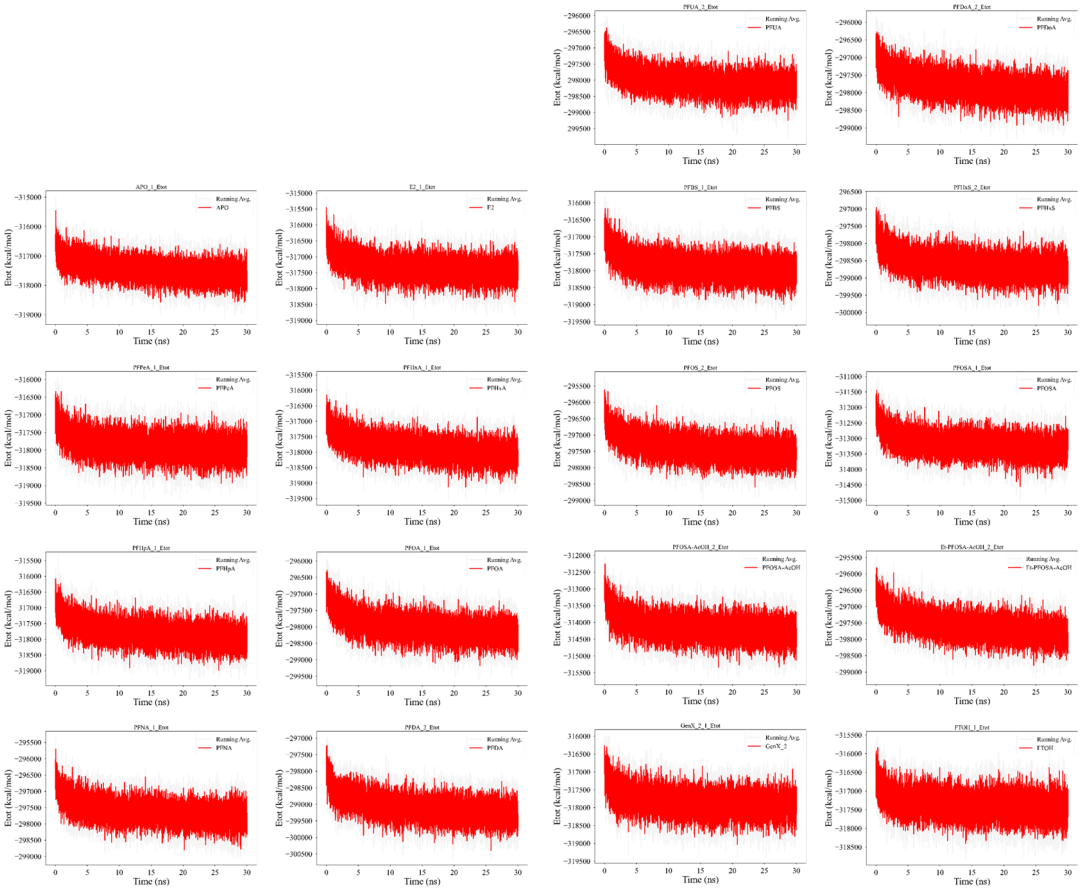


Figure S5.20 Total energies of PFAS bound simulations of ERα-LBD.

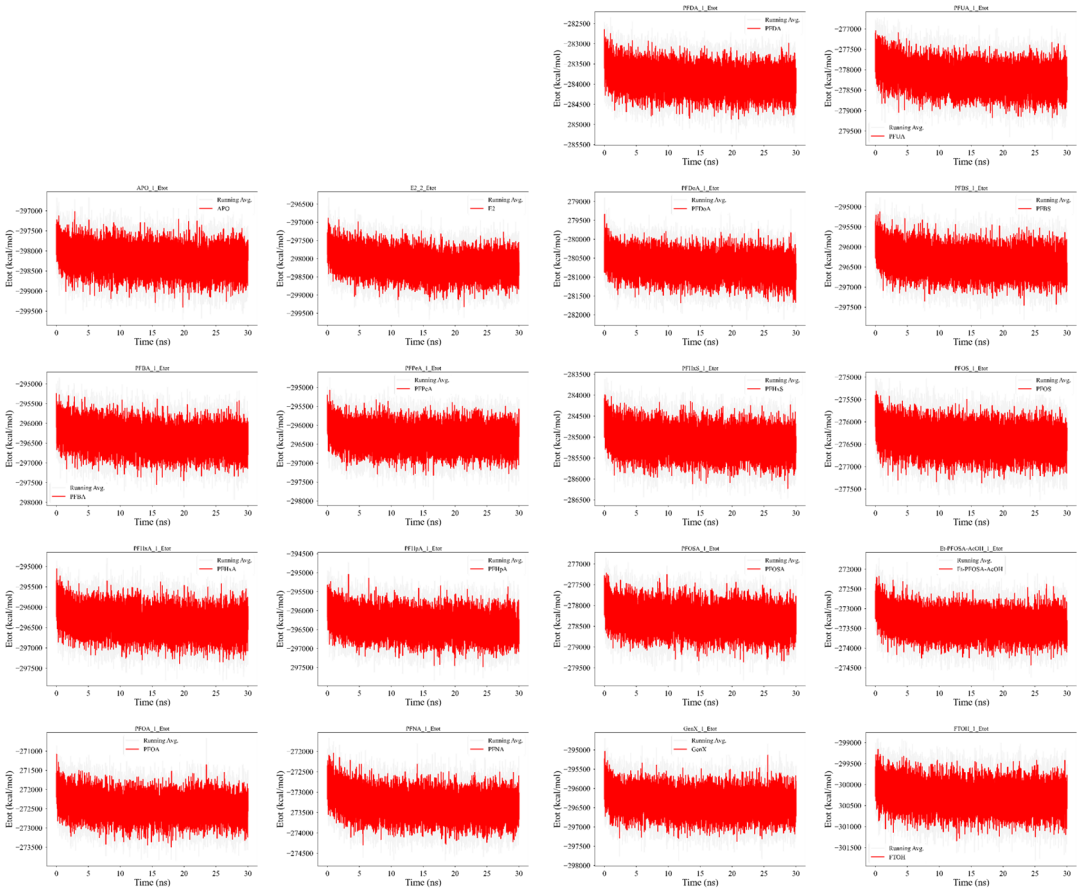


Figure S5.21 Total energies of PFAS bound simulations of ERβ-LBD.

CHAPTER 6

COMPUTATIONAL PATHWAYS TOWARDS NEW THERAPEUTIC COMPOUNDS: ADDRESSING TUBERCULOSIS VIA MMPL3 INHIBITION

6.1 Introduction

Tuberculosis (TB) has been one of the most widespread infections around the world[1]. According to the World Health Organization (WHO) Report, TB is still one of the top 10 causes of death worldwide. It is an airborne disease and continues to infect around 10 million people each year around the world. The main cause of TB is *Mycobacterium Tuberculosis*, a bacterium which was first isolated by Robert Koch in 1882[2]. Once an individual has the bacteria in their body, the disease lasts for their lifetime, and the bacteria can also result in formation of tubercules[2]. The bacteria mainly infect the lungs causing pulmonary TB. However, TB is not a completely new disease. Fossil and skeletal records showed abnormalities in the skeletons that are characteristics of TB, indicating that the infection has existed for a very long time. However, TB is mostly known for turning into an epidemic in 1700s and 1800s.

The first vaccine was developed by French scientists following the isolation of *M. tuberculosis*, called BCG (Bacillus Calmette–Guérin) vaccine[2]. The vaccine is currently being used in countries with high TB prevalence[3]. Although the BCG vaccine is protective against meningitis and disseminated TB if it is administrated during infancy, its effectiveness in adults is variable[1,3]. In addition, the vaccine does not inhibit the primary infection and reactivation. WHO Report also shows that the highest incidences are occurring in rural areas such as sub-Saharan Africa and Asia. However, WHO also indicates that almost one third of the world population is infected but only 10% of those who are infected show active symptoms[1].

Currently, TB treatment includes the combination of various drugs to be taken for 6-9 months[4]. The first line of treatment includes the use of Isoniazid, Rifampicin, Pyrazinamide and Ethambutol with various doses, and their structures can be seen in Figure 1. Rifampicin acts by inhibiting the DNA-dependent RNA synthesis by binding to the RNA polymerase[5]. The rest of the drugs are known to disrupt the cell wall, although their mechanisms are still not completely known[6–8]. This treatment is suitable for patients that have drug-susceptible pulmonary TB. The treatment should be followed rigorously, and if not, the bacteria could gain resistance to the first-line treatment drugs, resulting in drug-resistant bacteria. The drug-resistant TB, described as an infection caused by

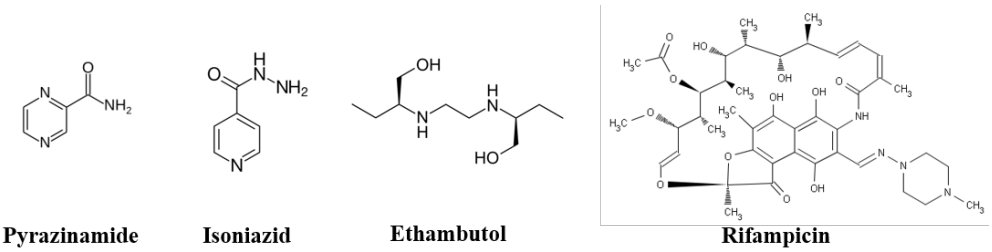


Figure S5.1 Compounds that are being used as first-line treatment against TB.

bacteria resistant to at least Isoniazid and Rifampicin, or one other TB treatment drug, requires a more extensive treatment regimen[4]. In addition, the cost associated with TB treatment is high, CDC reports that the treatment of a patient with drug-susceptible TB is approximately $ 20.000[9]. Apart from the cost of treatment, the side effects are a common concern, as are the interactions with other drugs. For instance, liver, ocular, skin and peripheral nerve toxicities can be seen in patients who take Ethambutol and Isoniazid.

Given the prominence of the disease and existence of only a few options for treatment, WHO started an initiative in 2006 to reduce TB worldwide and to cure almost 85% of TB positive cases[1]. With this initiative, the efforts to find a better treatment for TB gained momentum. Currently, there are many different treatment strategies that include targeting new proteins with new mechanism of actions (MOA) and creating new treatment regiments with known drugs. The treatment strategies that are currently under clinical trial are shown in Figure 2. Among those, linezolid, lavofloxacin and ofloxacin are repurposed drugs, and TMC-207 (Bedaquilin) as well as SQ109 have new MOA. While TMC-207 targets ATP synthase, SQ109 inhibits the activity of Mycobacterium Membrane Protein Large 3 (MmpL3)[1]. However, the bottleneck when developing new TB inhibitory drugs is to find a target with whole-cell active compounds. MmpL3, at the time being, fills this gap due to its role in TMM relocation. In addition, the drug candidates should also aim to reduce the treatment duration and occurrence of resistance.

6.1.1 Treating TB by Targeting MmpL3

MmpL proteins belong to the RND (resistance, nodulation and cell division) superfamily which exists in bacteria, archaea, and eukaryotes. The RND family consists of multidrug resistance

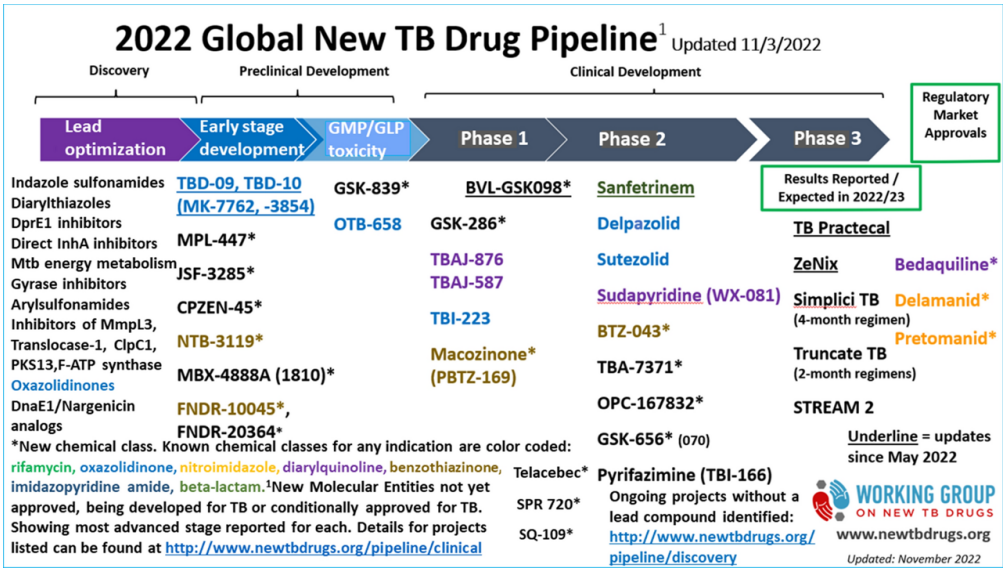


Figure S5.2 The global clinical development pipeline for new anti-TB drugs and drug regimens to treat TB disease. Reproduced with permission from [10].

pumps which are used to transport drugs, heavy metals, fatty acids and detergents, and in general, they use the electrochemical proton gradient across the mycobacterium inner membrane. In *M. tuberculosis*, 13 genes are found to code MmpL proteins. In a prior study, in order to understand the roles of those MmpL proteins, they were knocked out one by one, and cells were grown without the gene [11]. However, when the MmpL3 gene was removed, the cells did not grow. This led to the conclusion that MmpL3 protein is important for cell growth and viability in *M. tuberculosis*, and it is conserved within all mycobacterial genome. Later, it was determined that the main function of MmpL3 protein is to act as a flipase for lipids called trehalose monomycolates (TMM) [12].

Mycobacteria has two membranes, an inner membrane (IM) and an outer membrane (OM) (Fig. 3). The IM mainly consists of Ac_2PIM_2 , and other major phospholipids, with Ac_2PIM_2 being the most abundant [14]. The OM, on the other hand, does not have any common lipids, but has mainly glycopeptidolipids and mycolic acid containing lipids [14]. The current treatment strategy with Isoniazid and Ethambutol focuses on the inhibition of mycolic acid synthesis to disrupt the cell membrane composition [13]. However, the MmpL3 protein is further in the mycolic acid pathway (Fig. 4). The MmpL3 is located in the IM and thought to be responsible for trehalose monomycolate translocation to the periplasmic domain in between the IM and OM. Once trehalose monomycolates

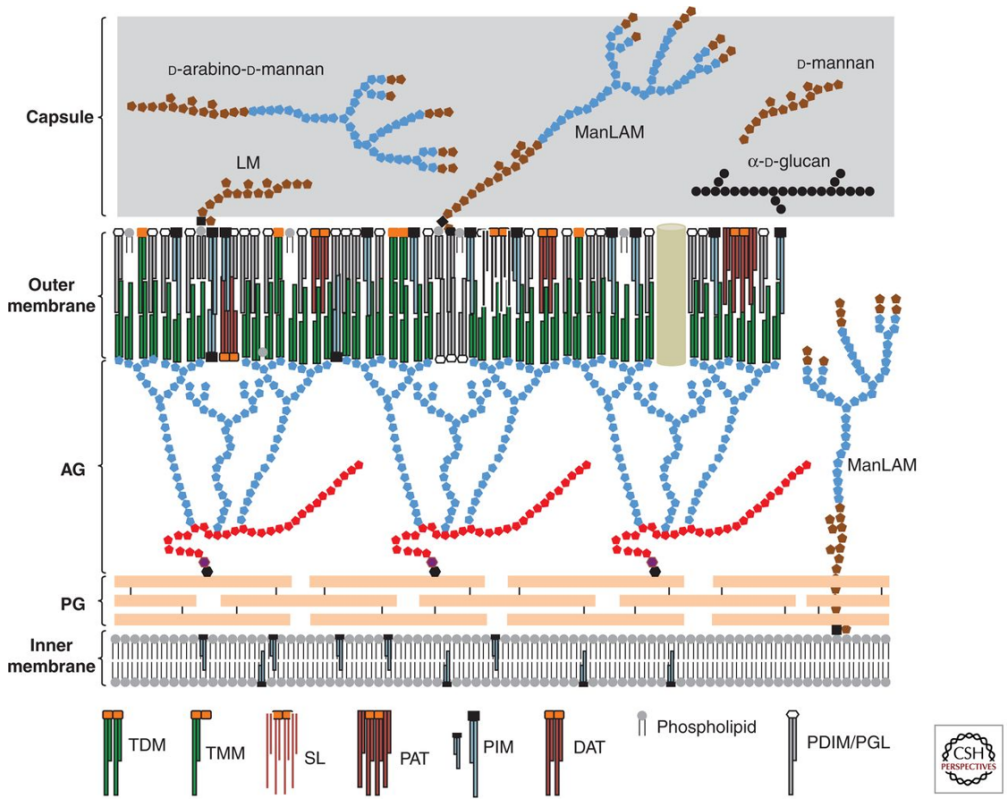


Figure S5.3 Schematic representation of the Mtb cell envelope. Reproduced with permission from[13].

reaches the OM, it dimerizes and forms trehalose dimycolates (TDM)[13]. trehalose dimycolates and other mycolic acids are important for the permeability of the OM (by making the membrane extremely hydrophobic) as well as the formation of biofilms[13,15,16]. Therefore, the transportation of trehalose monomycolates has great importance.

As stated, MmpL3 protein exports trehalose monomycolates, and it does so with the help of proton motive force (PMF). The crystal structure obtained from *Mycobacterium smegmatis* revealed that there are two Asp-Tyr pairs located in helices IV and X, and they are hypothesized to play an important role in PMF, showed by mutation studies[17,18]. This structure can be seen in Fig.3.5. The C-terminal domain, which is located on the cytoplasmic side, however is not shown. High-throughput screening (HTS) resulted in the identification of an adamantyl urea (AU1235) against both drug-susceptible and drug-resistant *M. tuberclosis* by targeting MmpL3[19]. Later, two other compounds, BM212 and SQ109, are found to inhibit MmpL3 protein (Fig. 5)[20,21]. The

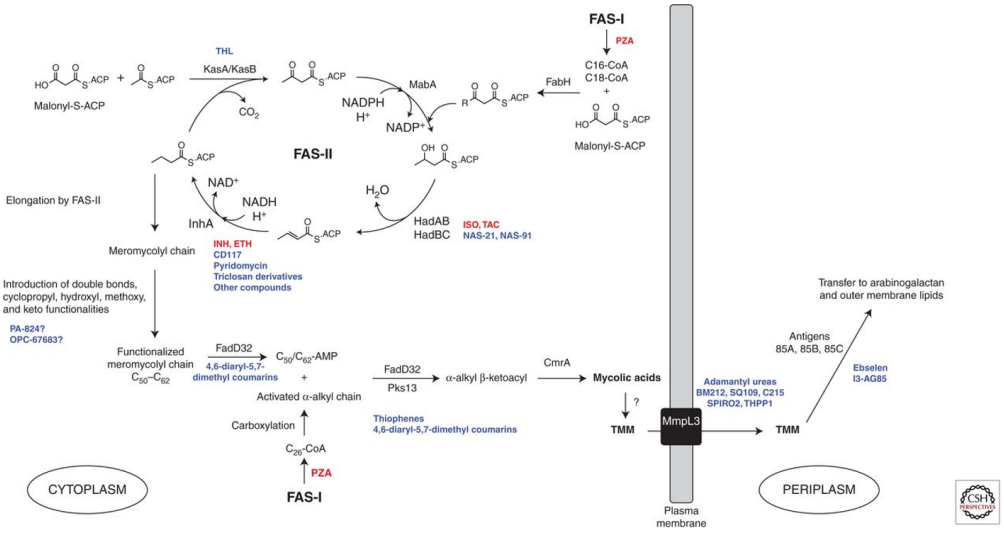


Figure S5.4 Biosynthetic pathway of mycolic acids in Mtb and site of action of anti-TB drugs[13]. The drugs that are used for TB treatment currently are indicated with red text. 'Reproduced with permission from[13]

MmpL3 *M. smegmatis* structure co-crystallized with these compounds showed that despite different chemical scaffold, they all bind to the same pocket, and inhibit PMF by disrupting hydrogen bonds between Asp-Tyr pairs[17]. Since then, many groups have been working on creating compounds that target MmpL3 protein. A recent study showed that HTS helped to identify new types of MmpL3 inhibitors, and the study was also successful in identifying the known hit compounds[22]. In order to understand the mechanism of action of those hit compounds, currently, computational investigation of the effect of the identified compounds was done.

6.2 Computational Details

6.2.1 Homolog Modeling

Homology modeling essentially targets building a three-dimensional structure for proteins by using the available structures of closely related proteins. Since not all proteins have their 3-D structures experimentally determined, being able to predict them successfully with in silico methods is extremely useful in drug discovery studies. There are currently many available tools for homology modeling, and each of them uses a different approach. One of the most successful one is I-TASSER by Zhang Lab[23–25]. The amino acid sequence is first matched with the sequence

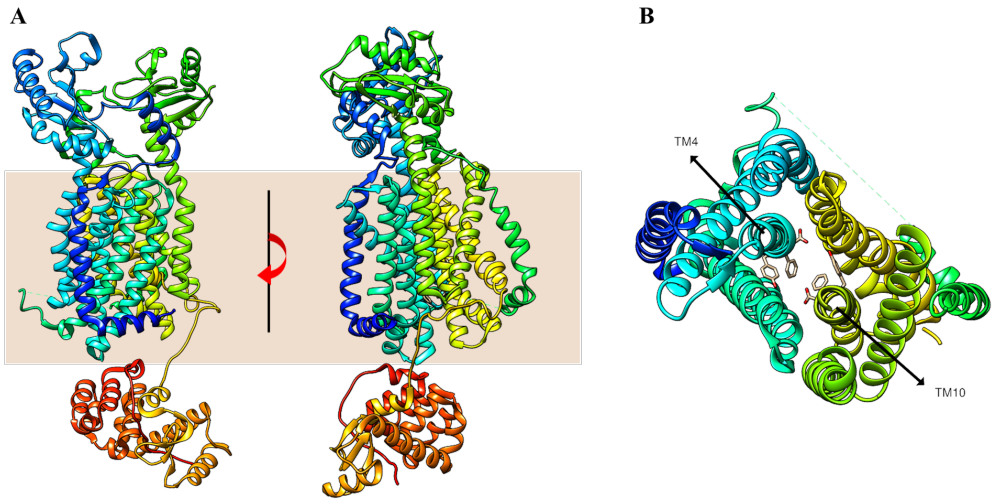


Figure S5.5 Left: Overall Structure of MmpL3 from M. Smegmatis (PDB ID: 6AJG)[17]. A: The side view of the crystal structure. Both periplasmic domain as well as transmembrane domain are divided into two regions, N and C. B: The top view of the transmembrane domain. Two Asp-Tyr pairs are shown in stick representation. The star indicates the pocket inhibitor molecules bind to. Right: Structures of SQ109, BMB212, and AU1235.

of available crystal structures in Protein Data Bank, producing fragments. Then, these fragments obtained from PDB templates are combined to form full-length structure models with Monte Carlo simulations, and the clustering is used to obtain a model. In the final step, this model is used to re-assemble the structures to obtain the final model with the lowest energy. Given the success of this approach in CASP (Critical Assessment of Techniques for Protein Structure Prediction) competitions, it was used to model *M. tuberculosis* the MmpL3 protein structure.

6.2.2 Molecular Docking

MOE (Molecular Operating Environment) is a commercial software that is designed for in silico research. The docking suite of MOE is capable of performing induced fit docking with a user-friendly GUI. The algorithm used for ligand placement is Triangle Matcher which uses alpha-spheres to define the binding site[26]. The ligand is positioned so that the triplets of ligand atoms are superposed on alpha spheres, and if there is a steric clash with protein, that pose is removed. The scoring function for placement step is called London dG that includes the terms for ligand flexibility, hydrogen bonds and desolvation[27,28]. After the placement step, a specified number of

poses (usually 10% of the initial placements) are refined for final ranking. For the refinement step, force-field based GBVI/WSA dG (Generalized-Born Volume Integral/Weighted Surface area) scoring function is used[28].

The docking procedure is applied as follows unless stated otherwise:

- Compounds are drawn in MOE software and the structures are minimized at Amber10:ETH level as implemented in MOE.
- The binding pocket is selected using "SiteFinder" package that utilizes alpha-spheres.
- For placement, Triangle Matcher method is used with London dG scoring function. 100 poses are generated.
- For the refinement step, induced fit method is used with GBVI/WSA dG scoring function. Top 10 poses are reported.

6.2.3 Molecualr Dynamics Simulations and Binding Free Energy Calculations

The MD simulations systems are prepared as following unless stated otherwise:

- Protein crystal structures are prepared in MOE.
- The partial charges of the ligand molecules are calculated using AM1-BCC with *antechamber* module.
- The protein and ligand are combined, ff14SB, gaff2 and TIP4P-EW force-fields are used for protein, ligand and water molecules, respectively.
- Minimization is performed in 4 steps.
    1. All heavy atoms are restrained (100 kcal/mol/A^2) and the system is minimized for 20000 steps.
    2. All heavy atoms are restrained (50 kcal/mol/A^2) and the system is minimized for 20000 steps.

3. Only ligand is restrained (10 kcal/mol/A^2) and the system is minimized for 20000 steps.
4. The system is minimized with no restraint.

- Heating is performed in a step-wise fashion from 0K to 300K. Langevin thermostat is used for temperature control. The time step is 1fs.

1. All atoms are restrained (3 kcal/mol/A^2) and the system is simulated at 0K for 10ps.
2. All atoms are restrained (3 kcal/mol/A^2) and the system is heated up to 5K for 50ps.
3. All atoms are restrained (3 kcal/mol/A^2) and the system is heated up to 10K for 50ps.
4. All atoms are restrained (3 kcal/mol/A^2) and the system is heated up to 20K for 50ps.
5. All heavy atoms except for the solvent atoms are restrained and the system is heated up to 50K for 50ps.
6. All heavy atoms except for the solvent atoms are restrained and the system is heated up to 100K for 100ps.
7. All heavy atoms except for the solvent atoms are restrained and the system is heated up to 200K for 100ps.
8. No restraint is applied, and the system is equilibrated at 200K for 200ps.
9. No restraint is applied, and the system is heated up to 300K for 400ps.
10. No restraint is applied, and the system is equilibrated at 300K for 500ps.
11. A short 500ps long simulation is performed before the production run at 300K.

- The production run is performed with 1fs time step under NPT conditions. SHAKE is applied, and the Langevin thermostat is used.

The MM-GBSA/PBSA methods were used to estimate the binding energies and rank the affinities of the investigated compounds by selecting every tenth frame from the simulation. The root-mean-square-distances (RMSD), root-mean-square-fluctuations (RMSF), hydrogen bonds, and residue decompositions were calculated using the cpptraj module.

6.2.4 Fragment Search

To grow the ligands within the binding pocket, a selected molecule was selected to for fragment addition within MAB mmpL3 protein. Two available fragment libraries within MOE were selected: ChEMBL fragment library (778760 fragments) and MOE fragment library (40626 fragments). MM/GBVI values were calculated for each generated compound and used for ranking them. The compounds with lowest MM/GBVI values were selected for docking.

6.3 Results and Discussion

6.3.1 Preliminary investigation of MmpL3 inhibition with computational modeling

The only available MmpL3 crystal structure is for *Mycobacterium smegmatis*[17], as of 2021. *M. smegmatis* (Mtb) and *M. tuberculosis* MmpL3 proteins share 60% identity and 76% similarity. Therefore, I-TASSER server is used to obtain the model structure for M. tuberculosis. The model has a 1.3Åoverall similarity (Fig. 6). As stated in the Introduction, the binding pocket for SQ109, for example, lies in the middle of the transmembrane domain[17]. The model structure is simulated for 20ns to sample a better orientation for the two Phe residues that rest at the bottom of the binding pocket that would allow docking of the compounds. Performing a short simulation also allows for the model system to equilibrate better. As can be observed in Fig.6, a better orientation for Phe side chains is successfully obtained, and it was also similar to the orientations observed in structure co-crystallized with an inhibitor, SQ109. The orientation of Phe residues is important since when they have positioned upwards, they do block the binding pocket. However, a downwards orientation allows for the docking of the molecules into the pocket. For the following studies, 20ns simulated model structure is used. The binding pocket overall is hydrophobic at the top part, polar in the middle due to Asp-Tyr pairs, and slightly hydrophobic at the bottom part again.

After obtaining an appropriate structure, co-crystallized ligands introduced in Fig. 3 are docked to make sure that the docking procedure is working successfully. For SQ109, the orientation as well as the ligand interactions of the compound in the pocket is the same when compared to co-crystallized SQ109 (Fig. 7). Both docked pose and the co-crystallized SQ109 are seen to directly interact with Asp640 residue in the binding pocket. The 50ns long simulation of the docked system

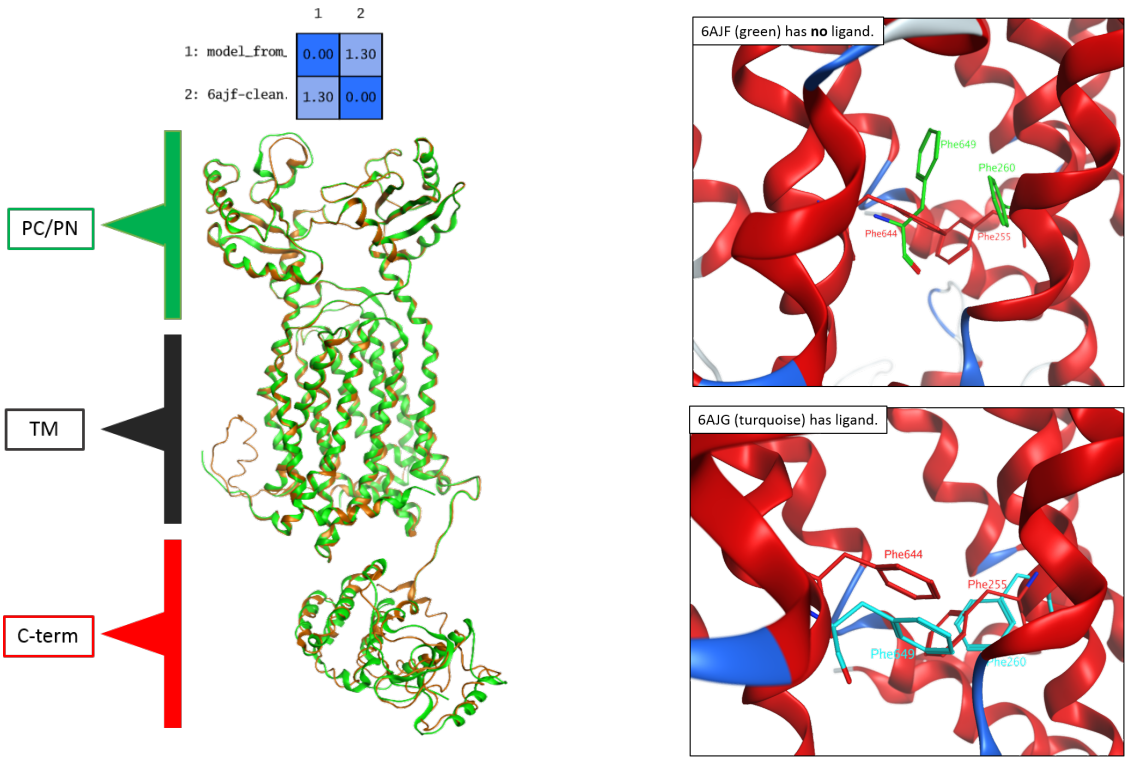


Figure S5.6 Left: The model structure (orange) and 6AJF crystal (green) structure are overlapped[17]. The total RMSD can be seen at the top of the figure. PC/PN is the periplasmic region, TM is the transmembrane and C-terminal is the cytoplasmic region of MmpL3. Right: Overlap of the binding pocket Phe side groups with 6AJF (top right, green) and 6AJG (bottom right, cyan) crystal structures. 20ns simulated model structure is shown in red.

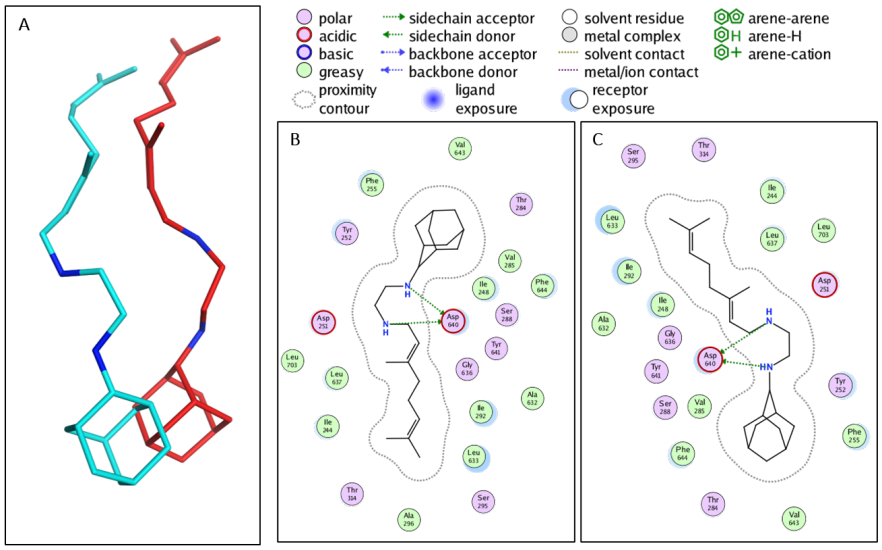


Figure S5.7 A. Comparison of SQ019 orientation between model (red) and 6AJG (cyan) structures. B,C. Ligand interaction map of docked and co-crystallized SQ109, respectively.

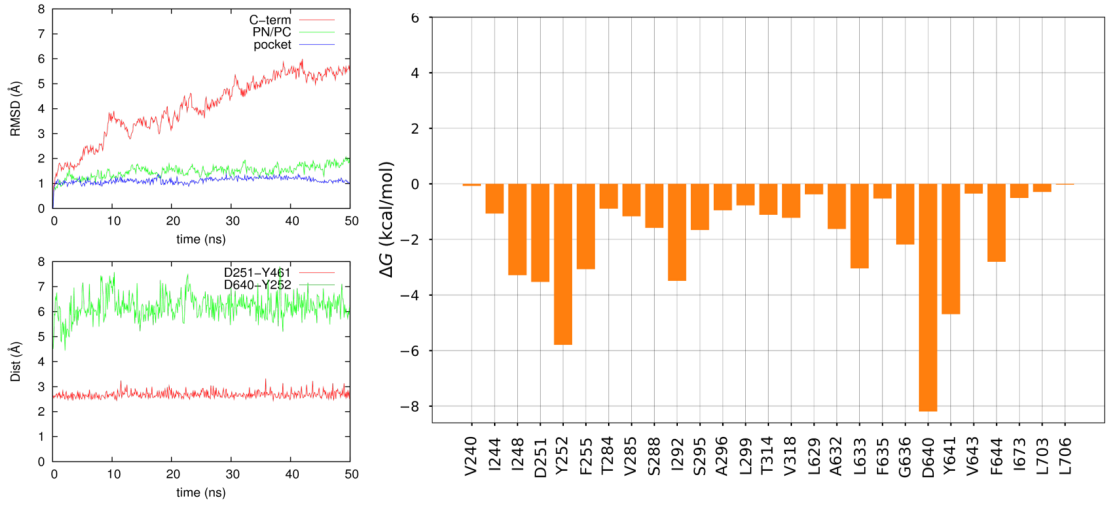


Figure S5.8 Left: RMSD (top) and distance (bottom) plots. Right: Residue interaction plot for the pocket residues.

shows a stable N-terminal and binding pocket, however, the C-terminal RMSD keeps increasing. The distances between Asp-Tyr pairs are also tracked throughout the simulation. While one pair stays very close, the interaction between the other pair is completely broken due to the interference of SQ109 (Fig. 8). The residue interactions suggest that Asp640, Tyr252 and Tyr641 provide the highest stabilizing contributions for SQ109, while there is no positive contribution from any surrounding residues. This indicates that the compound is very well positioned and stable in the pocket, and it is successful in disrupting one Asp-Tyr pair. The binding free energy results (MM-PBSA) are reported in Table 1, and SQ109 results align with the experimental energies.

Following the success of the docking procedure, the ligands presented in Williams *et al.* are investigated. The names of the compounds are given in Table 1, and their structures along with EC_{50} values are reported in the cited paper[22]. The docking of HC2060 ligand resulted in two different poses. In order to understand which pose is the preferred one, both are simulated and analyzed. Based on free energy calculations, pose 2 is more stable in the pocket than pose 1. During the simulations, it is also observed that the pose 1 is folding within the pocket, trying to maximize the interaction between the polar region and the carbonyl azepane & piperidine rings. In the pose 2, they are located closer to Asp-Tyr pairs and can form interactions easily (Fig. 9). Therefore, pose 2 is considered for further analysis. This observation also gives support regarding

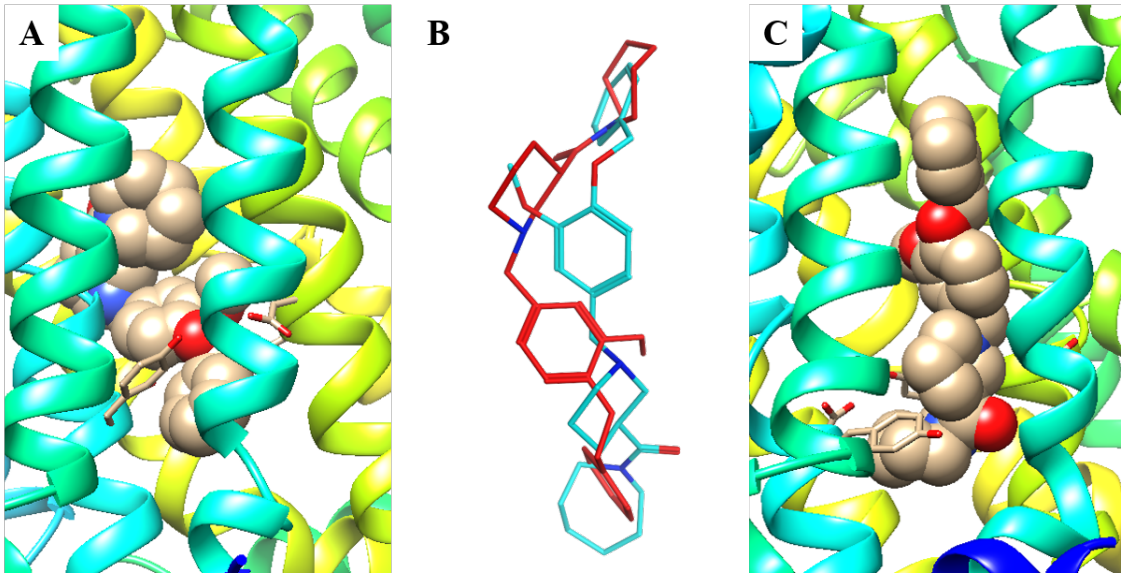


Figure S5.9 Orientation of HC2060 pose 1 and pose 2. A. The snapshot from the simulation of HC2060 pose 1. B. The comparison of docking orientations of pose 1(red) and pose 2 (cyan). C. The snapshot from the simulation of HC2060 pose 2. HC2060 compound is shown in vdW representation, Asp-Tyr pairs are shown in stick representation.

the pose selection. Further analysis shows that the interactions with the surrounding residues are favorable, and Asp-Tyr interactions are disrupted throughout the simulation.

Similar analysis is performed for HC2183 compound as well. It is selected due to the low differences between EC_{50}s of the mutant pool and wild-type[22]. HC2183 docking provided two different poses with very similar scores. Binding free energies, however, indicates that pose 2 is more stable in the pocket (Table 1). Furthermore, the disruption of Asp-Tyr interactions is lot more stable in pose 2 simulations. The pose 1 orientation forces the acetamide group to face the hydrophobic residues; however, in pose 2, acetamide is positioned near to polar region in the pocket (Fig. 10). Comparison of binding pocket surfaces of SQ109, HC2060 and HC2183 show that the pocket can be extended upwards easily, but the bottom part is blocked by two Phe residues. At this point, it was assumed that the pocket has more or less a cylindrical shape with a small protrusion to the middle region between Asp-Tyr pair. This results in a rotation of the compounds especially during docking procedure. However, docking results of the derivatives of compounds published by Zheng *et al.* revealed that a small extension in-between the Asp-Tyr pair can accommodate if functional group (oxane or cyclopentane) is bulky for the pocket[29]. In Figure 11, the comparison of

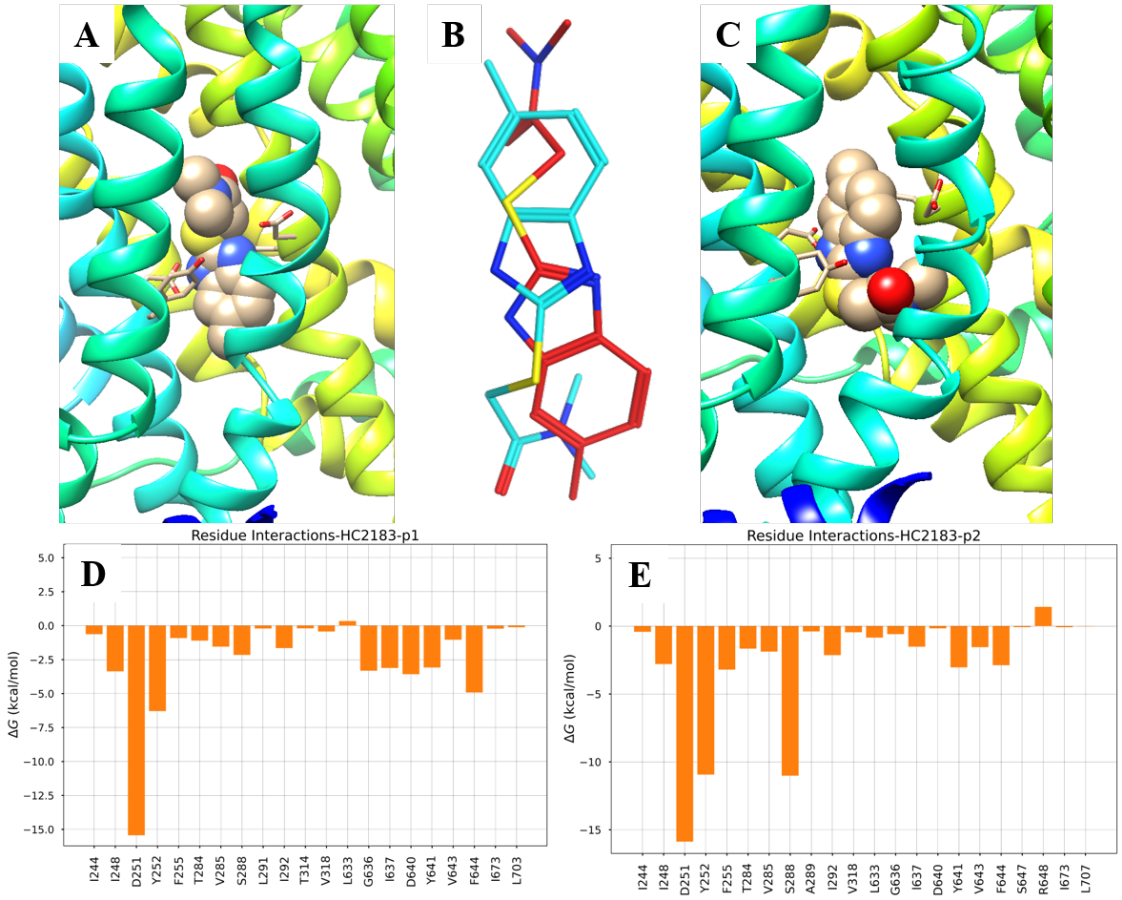


Figure S5.10 Orientation of HC2183 pose 1 and pose 2. A. The snapshot from the simulation of HC2183 pose 1. B. The comparison of docking orientations of pose 1(red) and pose 2 (cyan). C. The snapshot from the simulation of HC2060 pose 2. HC2183 compound is shown in vdW representation, Asp-Tyr pairs are shown in stick representation. D,E: The residue interaction plot for pose1 and pose 2, respectively.

pockets for SQ109 and a derivative indicates that the pocket has a capacity to expand upwards. We also see that the middle section of the pocket is negatively charged while the upper side is mostly neutral.

In conclusion, the orientation of the compounds can be determined based on the binding free energies, which is further supported by the electrostatic surface of the pocket. The key interactions are determined for the pocket residues based on the MD simulations. The pocket has a flexibility to expand in one direction, and this can be utilized when designing new compounds. Although a direct comparison of EC_{50} values with MM-GBSA/PBSA values is not feasible, a similar trend was observed (Table 1).

Table S5.1 MM-GBSA energy values for simulated systems.

Compound	AVG MM-GBSA (kcal/mol)	std (kcal/mol)
SQ109	-62.05	3.10
HC2134	-64.29	2.70
HC2060	-64.73	2.95
HC2183	-34.39	2.79
MSU43085	-54.89	3.38
HC2138	-66.57	3.05
HC2091	-49.40	2.66
HC2099	-48.64	3.05
MSU43557	-53.23	2.91
MSU44147	-49.43	2.70
OCT_01	-48.14	2.89
OCT_02	-53.47	2.69
OCT_03	-51.59	3.06
OCT_04	-51.29	3.29
OCT_05	-49.22	2.67
OCT_06	-50.20	3.05
OCT_07	-53.74	2.68
OCT_08	-46.83	2.74
MSU43107	-51.91	2.87
MSU43165	-48.57	2.87
MSU43557	-50.55	2.97
MSU43644	-49.36	2.95
HC2149	-45.95	3.22

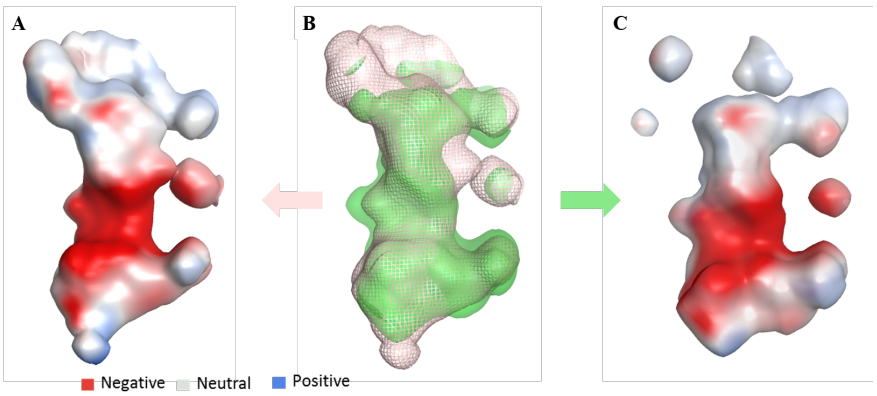


Figure S5.11 The pocket surface is shown for SQ109 and a derivative compound. A,C. Electrostatic surface of the pocket when SQ109 and a derivate are present, respectively. B. The overlap of the pockets. SQ109 is shown with light pink mesh surface, and other compound is shown in green solid surface.

6.3.2 Inhibitory differences among mycobacterium species: Mtb vs MAB

While the *mmpl3* gene is conserved among the mycobacterium species, there are certain differences of the protein sequences that lead to different reposes to the inhibitor compounds that are being tested. The sequence alignment of *Mycobacterium tuberculosis*(Mtb) and *Mycobacterium abscessus* (MAB) mmpL3 proteins share 62% sequence identity, as can be seen in Figure 12. Similarly, *Mycobacterium tuberculosis* (Mtb) and *Mycobacterium avium complex* (MAC) mmpL3 proteins have 73% sequence identity. When the TM helices around the binding pockets were compared, Mtb and MAB Helix 4 sequences were the most similar and Helix 5 was the least similar. Similarly, we do observe that there is a high sequence similarity for Helix 4 and 12, the Helix 5 had the lowest similarity between Mtb and MAC (Figure 12). When only the binding pocket residues were compared, all three species show high identity with some differences including Ser295 (Mtb) $\rightarrow$ Ala (MAB); Ala296 (Mtb) $\rightarrow$ Ser (MAC) and Leu (MAB); Leu299 (Mtb) $\rightarrow$ Met (MAB); Thr314 (Mtb) $\rightarrow$ Gly (MAB) and Ile (MAC); Ser317 (Mtb) $\rightarrow$ Ala (MAC); Ala632 (Mtb) $\rightarrow$ Val (MAB); Leu633 (Mtb) $\rightarrow$ Val (MAB, MAC); Ile673 (Mtb) $\rightarrow$ Leu (MAC). Mutations such as Ile to Leu may not affect the binding of the inhibitor compounds significantly, however, Ser to Ala, or Thr to Gly mutations could affect the binding strength or the ability of the inhibitor compounds. Despite the differences of binding site TM helices, the RMSD of the pocket residues highlights the fact that the residue orientations were similar (Fig. 12(C)). In addition, the three MmpL3 structures

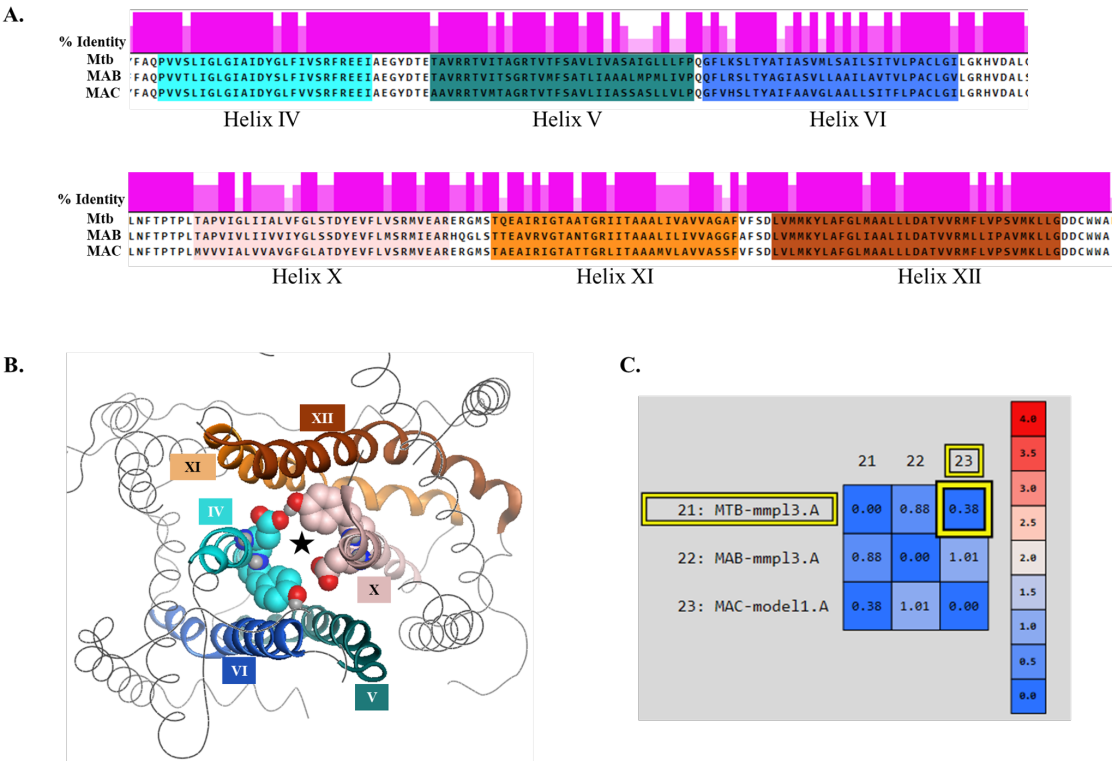


Figure S5.12 The sequence comparison of Mtb, MAB, and MAC mmpL3 proteins. A. Sequence overlap of transmembrane helices around the binding pocket of mmpL3 protein. B. The top-down view of the TM region of Mtb mmpL3 protein. The Asp-Tyr pairs and Phe residues are shown in vdW representation, and the inhibitor binding site was shown with star. The colors of helices correspond to the colors used in A. C. The RMSD of superimposing the binding pocket residues of Mtb, MAB, and MAC.

obtained with homology modeling suggest that the initial pocket opening between the TM4-5-6 and TM10-11-12 for inhibitor binding is larger in MAB than of Mtb and MAC (Fig. S1(B,C)). This might play a role in the accessibility of the pocket to the inhibitors.

The apo mmpL3 proteins from both MAB and Mtb were also simulated analyzed. The apo MAB mmpL3 protein simulations show that the C-terminal is very flexible, similar to the Mtb mmpL3 C-terminal. On the other hand, the distances between the atoms interacting on the Asp-Tyr pairs have larger distances in MAB mmpL3 protein with more stability throughout the simulations, compared to the Mtb mmpL3 case (Figure 13). In addition, the most dominant orientations of Asp-Tyr pairs during the apo simulations can be seen, and they indicate that the Asp-Tyr positionings are not very different from each other. As shown in Figure 14, the RMSD of the transmembrane helices 4-6, and 10-12 are also investigated and found to be slightly different between Mtb and

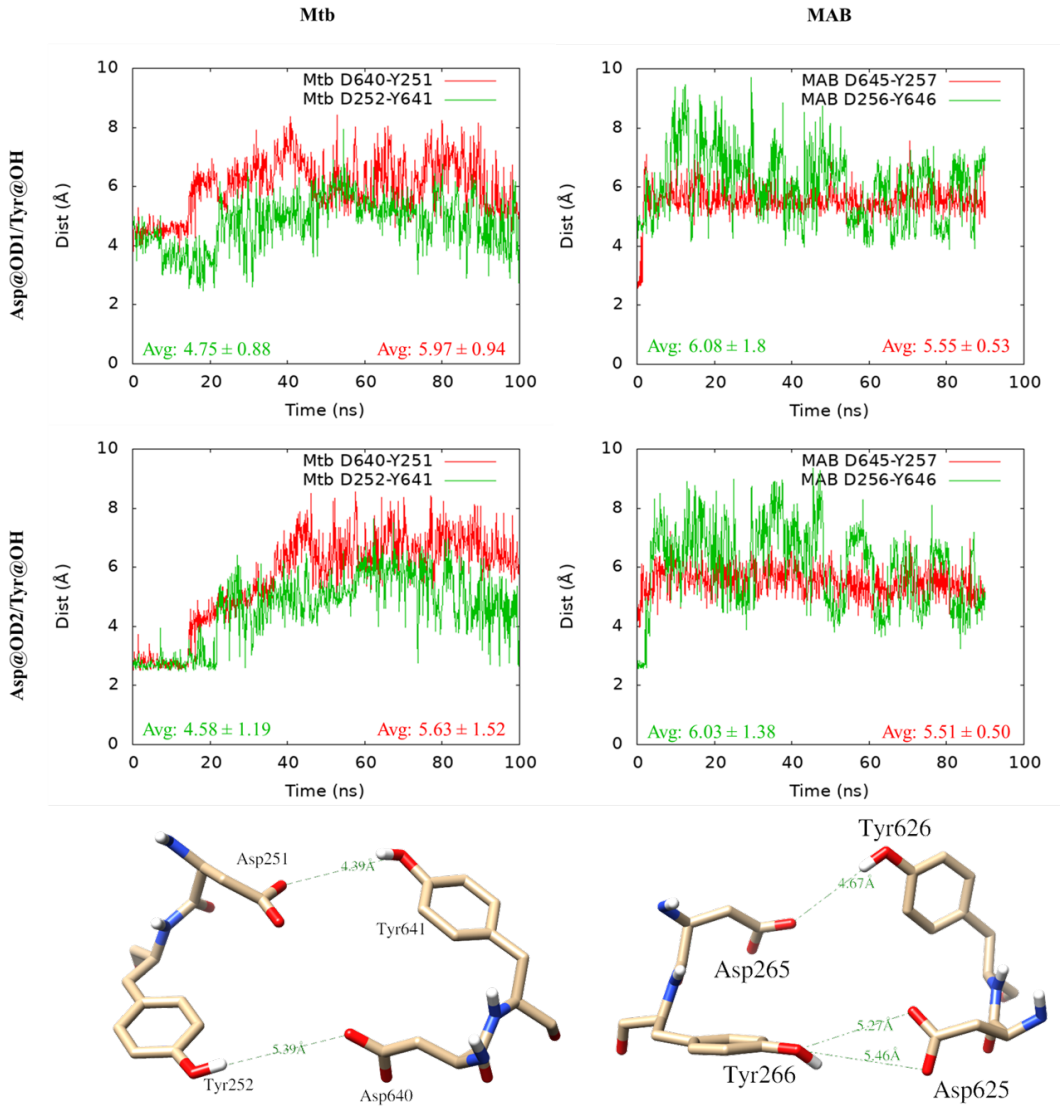


Figure S5.13 The distance time series plots for Asp-Tyr pairs from Mtb and MAB mmpL3 apo simulations. The average values for each distance is provided with the corresponding color within the plots.

MAB simulations, potentially due to the residue differences between the two proteins.

Next, to understand the different inhibitory effects of the selected compounds on these two species, all were docked to both Mtb and MAB mmpL3 model proteins, two poses for each compound were selected and simulated for 20 ns. While the molecular docking gives an idea about the possible orientations of a compound in the binding pocket and surrounding residues, MD simulations are useful to observe the interactions between the inhibitor and the protein as well as the overall behavior of the protein over time. For the selected four compounds, MSU43107,

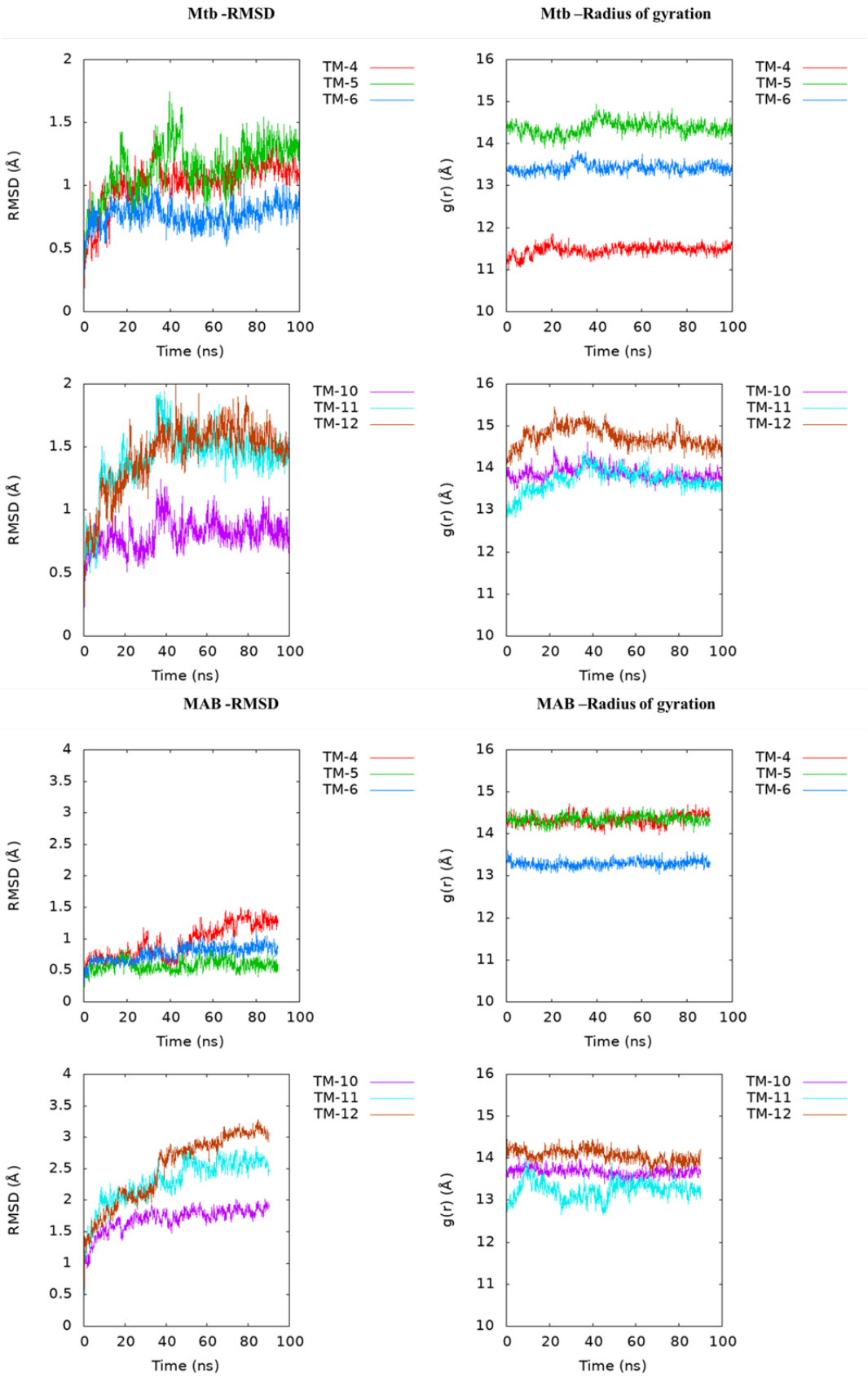


Figure S5.14 The RMSD and radius of gyration time series plots for TM helices 4,5,6,10,11, and 12 from Mtb and MAB mmpL3 apo simulations.

MSU43165, MSU43557, and MSU43644, molecular docking was used to obtain binding poses using MOE. On average, the MM-PBSA binding energies of the investigated compounds are lower for MAB MmpL3 protein that is accompanied by lower residue interaction energies with Asp-Tyr pairs in MAC system (Fig. 15). This could be a result of the amino acid differences around the binding pocket which could affect the inhibitor binding. The MM-PBSA energies suggested that certain binding poses are energetically more stable than others, but the other orientations cannot be completely discarded given the shape of the pocket. This preference, however, is most likely due to the placement of -NH on the imidazole ring, which prefers to be closer to Asp residues in the pocket to form hydrogen bonds. The hydrogen bond analysis also showed that the Asp251 is one of the residues that forms strong interactions with the inhibitors through aforementioned -NH group, which consequently implicate that the loss of -NH in this position would hinder the compound's ability to form strong interactions with the pocket residues. In the absence of a such hydrogen-donating group, the hydrogen bond percentages with Asp residues dropped almost 50% when compared to MSU43557. In addition, in the presence of an inhibitor, while Asp641-Tyr251 interaction persisted at different percentages throughout the simulations, the hydrogen bond percentage for Asp252-Tyr640 interaction was dropped from 80% in apo-MmpL3 to 0%, except for MSU43557. This could be attributed to the orientation of the amine group towards the Asp252 and the positioning of the cyclohexane towards the Asp251 causing it to change orientation to interact with -NH on the imidazole as well as the backbone of Ile248. However, in all cases, the interaction between Ser288 and Asp640 was not disturbed in all simulations, except for HC2099 where the interaction percentage is lower than the other cases. Furthermore, the interaction strengths with each pocket residue were also calculated and the results can be seen in Figure 15. The dominant interaction seen was with Asp residues from the Asp-Tyr pairs, supporting the observations of the hydrogen bond analysis.

The simulations of HC2091 and HC2099 provided an interesting situation. While the interaction pattern on HC2099 was similar to MSU-43644 except for the interaction with Ser288, HC2091 showed no hydrogen bonding with the surrounding. A close inspection of the simulation shows that

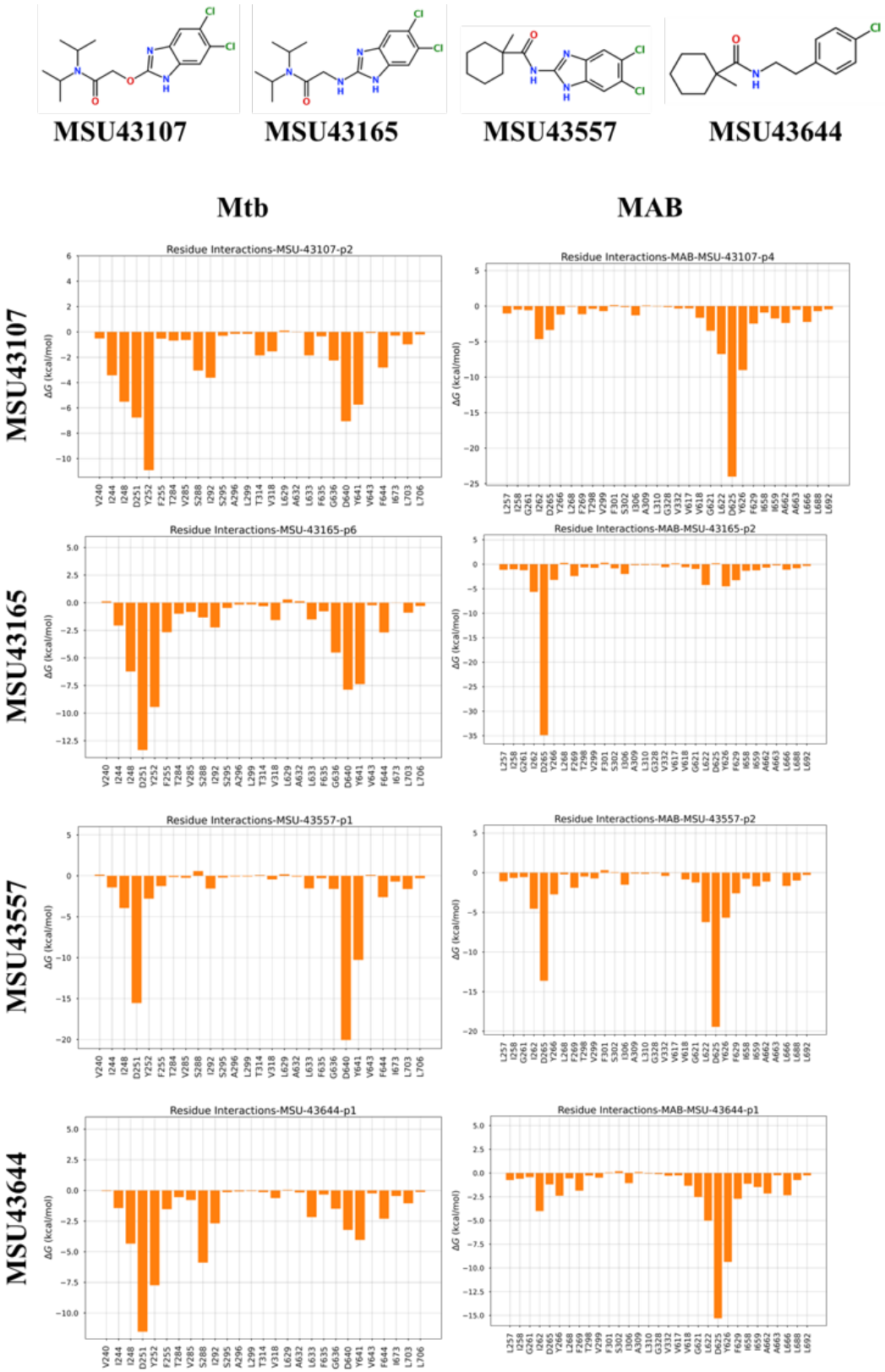


Figure S5.15 Residue decomposition energies of compounds MSU43107, MSU43165, MSU43557, and MSU43644 for Mtb and MAB mmpL3 protein.

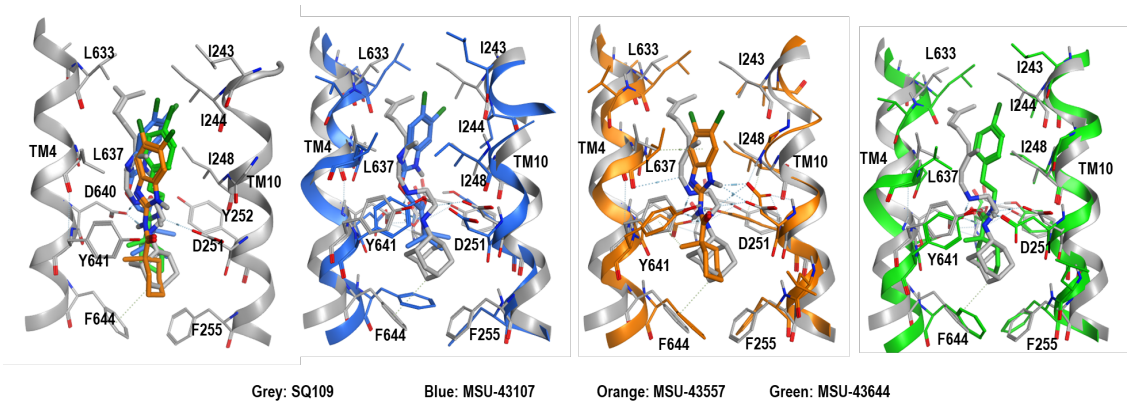


Figure S5.16 Comparison of binding poses of SQ109 (grey) with MSU43107 (blue), MSU43557 (orange), and MSU42644 (green).

HC2091 shifts in the binding pocket towards Tyr252 residue and pushing it away from the original position. This shift also causes Phe644 side group to move towards the pocket area that is usually occupied by cyclohexane in the MSU-43557 case. The presence of the tetrahydropyran ring as well as thiophene might cause the compound to have a polar region near the Phe residues, hence causing a shift in the pocket and creating space for Phe644 to occupy. In addition, this orientation of Phe644 is also observed in the MmpL3 proteins with no inhibitor compound.

The calculated MM-GBSA energies for Mtb and MAB mmpL3 protein are shown in Table 2. Overall, the MAB binding affinities measured with MM-GBSA method provided a reasonable ranking of the compounds, when compared to the EC_{50} values. To validate, we compared binding free energies to our EC_{50} (Mtb and MAB) values for a small set of analogs (Table X). Despite EC_{50} being subject to features such as cell wall permeability and protein-binding, the data roughly rank similar within series and from compound-to-compound for both Mtb and Mab with a few exceptions. Analysis of the core of the binding domain showed that combinations of Asp251 and Tyr252 and their partner residues Asp640 and Tyr641 make similar interactions with each inhibitor with either the benzimidazole of MSU-43107 and -43557 or the amide of MSU-43644. Relative to SQ109, each inhibitor does not fully fill the lipophilic binding space available. A fragment search to investigate different moieties that can fill the periplasmic side of the pocket, as will be shown in the next section.

Another interesting observation from the apo versus inhibitor-bound simulations is the water

access to the channel of mmpL3 proteins from Mtb and MAB species as well as the periplasmic region of the protein. One of the most potent inhibitors (MSU43085) was selected against Mtb mmpL3 for comparison with apo Mtb mmpL3 simulations. The 20 ns long simulations were clustered and the highest populated cluster was selected for the analysis of the water access to the channel where the inhibitors are bound, and the results are shown in Figure S3 . The periplasmic region of both the apo Mtb and MAB mmpL3 proteins clearly indicate the pockets where the TMM lipid can bind and be moved from the inner membrane to the periplasmic region. The water density around these regions was not impacted by the presence of the inhibitor in the channel. On the other hand, the water access to the channel in Mtb mmpL3 is significantly blocked by the presence of the inhibitor molecule (Fig. S3 A,B). One interesting difference between Mtb and MAB channels is that the water occupancy in MAB case covers a slightly larger volume than the Mtb case, supporting the results that the pocket volume is larger in MAB mmpL3 protein. When MSU43085 compound is bound to the protein, however, while both proteins have limited water access to the channel due to the bulky presence of the inhibitor in the channel, MAB mmpL3 protein seems to have slightly more access for the water molecule around the pocket helices, as shown in Figure S3(D). These observations support the fact that the binding pockets have different sizes between Mtb and MAB proteins, and also provide insight about the mechanism of inhibition: the physical presence of the inhibitor molecule can act as a "bottle stopper" to prevent the water access between the periplasmic and cytoplasmic sides and hence, inhibiting the H^+ transport.

6.3.3 Understanding the residue mutations in Mtb mmpL3 and its influence on inhibitor binding

Using our identified Mtb mmpL3 mutant library (Fig. 2) of resistant organisms, 15 of the mutant positions of both Clade 1 and 2 were modeled into the binding domain of Mtb MmpL3 to provide a 3D model. Most mutations are concentrated on TM5 and TM10, the remaining were on TM helices (TM4, 6, 11, 12). The Clade 1 mutants grouped on the cytosolic side of the binding domain and most Clade 2 mutants grouped on the periplasmic end or on the periphery (R373W). Based on the docking experiments for HC2099 and HC2091 (Fig. 7), it is anticipated that many Clade 1 mutants

(Fig 2) would be and are resistant to treatment with HC2099 and HC2091. An alanine scan of both Clade 1 and 2 mutants was also performed. Overall, Clade 1 mutations have a greater effect on ligand binding energies than those of Clade 2, supporting the data reported in Fig. 2. The biggest impact is observed for Tyr252Ala and Phe644Ala mutations. For Tyr252Ala system located on TM4, HC2099 showed loss of binding strength around -6 kcal/mol, while this loss was -4 kcal/mol, -3.5 kcal/mol, and -2 kcal/mol for HC2091, MSU-43664, and MSU-43557, respectively. Fo the Phe644Ala located on TM10, both HC2099 and MSU-43557 have -2.5 kcal mol^{-1} less interaction strength, and it was -2 kcal/mol and -1.8 $kcal^{-1}$ for MSU-43664 and HC-2091, respectively. Both Tyr252 and Phe644 are residues located in the binding pocket, and in the close proximity of the compound. The simulation results show that there is a trend in which the investigated compounds are forming strong hydrogen bonds with Tyr residues in the pocket, and mutation to Ala would cause the loss of this prominent interaction. Similarly, Phe644 is one of the two Phe residues that is located at the cytosolic side of the pocket, and they determine the border of the pocket. During the simulations, interactions between the compound and the Phe644 residue are observed, although transient. Nevertheless, Phe644 and Phe252 stay in the close proximity of the compound forming long range interactions with the phenyl group or isopropyl groups, and the loss of the functional group of Phe residue would disrupt the interactions.

Another interesting mutation from Clade I is Val285Ala located on TM5. The residue decomposition analysis showed that the energy contribution of Val285 to the binding of MSU-43557 and MSU-43644 is ≈0.2 kcal/mol and -0.8 kcal/mol, respectively. However, the interaction loss from Val285Ala mutation is -0.6 kcal/mol and -1kcal/mol for MSU-43557 and MSU-43644, respectively. This would indicate that although there is not much of a direct contact between Val285 and the compounds, however, as it can be seen from Figure 2, the positioning of this residue is near Tyr252, and it can form long range interactions with the compounds in the pocket.

In Clade II, the highest change of interaction energies was observed for the Ile244Ala mutations. Ile244 is located on TM4, and the side chain of the residue is oriented towards the binding pocket where it is interacting with the inhibitor compound. In addition, the energy contribution of Ile244 for

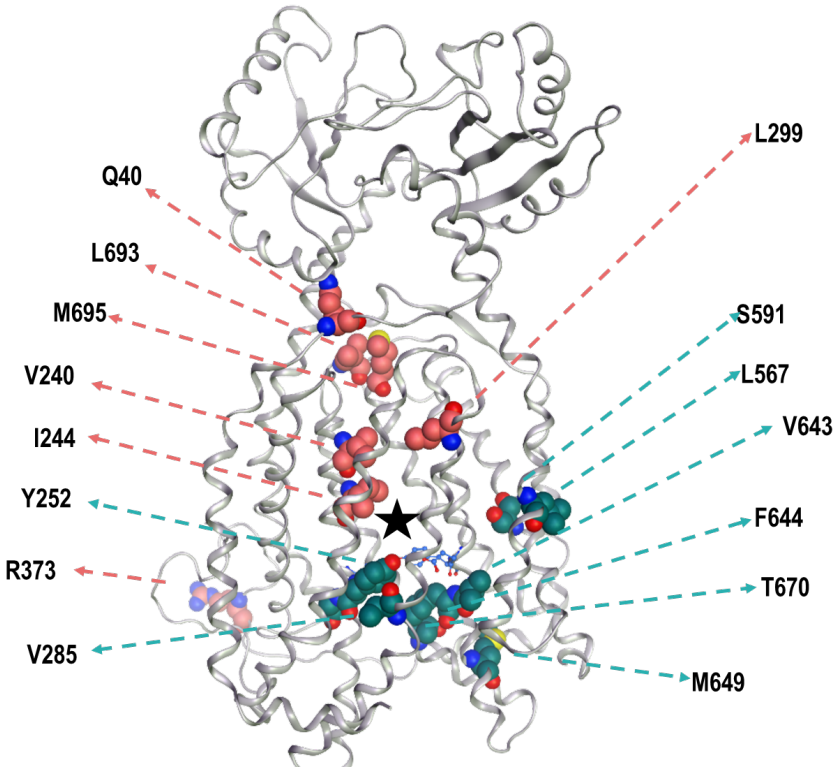


Figure S5.17 The residues tested from the cross-resistance Z-matrix data from CITE. Blue corresponds to the residues selected from Clade I and the red ones correspond to the Clade II resdiues. The Mtb mmpL3 protein is shown with grey ribbon representation,and the binding pocket is indicated with a star.

both MSU-43557 and MSU-43644 is 1.4 kcal/mol, indicating that the residue is forming interactions with both compounds. The mutation of Ile244 to Ala showed that the biggest loss of interaction energy was for HC2091 with -1.7 kcal/mol, followed by -1 kcal/mol for MSU-43644, -0.7 kcal/mol for MSU-43557, and -0.3 kcal/mol for HC2099. This observation indicates that the compounds from the HC-2091 series with the chlorobenzene group tends to have stronger interaction energies than the molecules with benzimidazole. Finally, for MSU-43557 and MSU-43644 compounds, the same mutations on MAB MmpL3 were testes. While all of the investigated residues have a similar interaction loss upon mutated to Ala, only Ile244 (Mtb)/Ile255(MAB) showed a difference in MSU-43644. The simulations of this compound for MAB and Mtb MmpL3 proteins revealed that Ile244(Mtb) moves towards the compound during the simulations, however, Ile255 (MAB) moves towards the opposite direction. The interaction energies with this residue also observed to be lower (-0.4 kcal/mol) for MAB MmpL3 simulations, while the Mtb protein had -1.4 kcal/mol.

Table S5.2 MM-GBSA energy values for simulated systems in Mtb and MAB mmpL3 binding pockets.

	Mtb		MAB	
Compound	**EC50 (µM)**	**AVG MM-GBSA (kcal/mol)**	**EC50 (µM)**	**AVG MM-GBSA (kcal/mol)**
MSU44271	1.2	-29.48	30	-21.90
MSU44124	0.7	-28.35	10	--
MSU44233	0.48	-34.09	0.8	-26.34
MSU44582	0.36	-26.34	1.30	-18.39
MSU43644	0.28	-24.98	--	---
MSU44605	4.1	-31.10	25	-22.59
MSU44607	0.36	-26.18	9.6	-29.79
MSU43557	0.088	-39.82	--	--
MSU43107	0.116	-27.44	--	--
MSU43165	0.38	-23.81	--	--

While the results show that the mutations of the pocket residues can have a detrimental effect on the interaction energies of the investigated compounds, more sophisticated methods, including but not limited to Molecular Dynamics and enhanced sampling, would be needed to further understand the effect caused by the residues located away from the binding site. Furthermore, with the publication of a crystal structure of mmpL3 protein with TMM lipid (PDB ID: 7N6B), the importance of some resistant mutations that are positioned further away from the binding site can be explained (Figure S2).

6.3.4 Modeling new compounds targeting Mtb/MAB MmpL3 with fragment search

To expand the chemical space of the compounds, a fragment search was performed by taking the selected compounds of interest as basis. For the fragment search, the focus was the periplasmic side of the pocket on the MAB mmpL3 protein as it was observed that the pocket can accommodate larger compounds than the pocket of Mtb mmpL3 (Fig. S1). Two different fragment databases available on MOE were used: ChEMBL fragment database and MOE fragment database. The obtained fragments were rescored using the GBVI/WSA dG scoring function, and the top 25 compounds were selected for docking with Mtb and MAB mmpL3 proteins. One common theme that was observed for the top 25 compounds was the presence of a hydrogen donor attached to the phenyl ring. Docking of these compounds revealed that this -OH group prefers to orient towards

Val618 and interact with the backbone of the amino acid in MAB mmpL3 in the majority of the poses, however, this was not observed as commonly in Mtb mmpL3, indicating that this residue may not form the hydrogen bond (Fig. 18).

While the docking scores were favorable with both protein pockets, the orientations of the compounds within the pocket were observed to be different. most likely due to the shape as well as the pocket volume differences caused by the mutations in the protein (Fig. S1(B)). The poses obtained for MAB mmpL3 pocket indicated a "narrower" and more extended conformations for the compounds while Mtb mmpL3 pocket had a wider conformations around the middle region of the pocket (Figure 18). The added fragments extend towards the same region within the pocket in MAB and coordinate to Val618 backbone carbonyl. When the compounds generated with MOE fragment database were docked to MAB mmpL3 protein, a similar trend in which there is a hydrogen donor (not an alcohol necessarily this time) that interacts with Val618 backbone upon docking (Fig. S5). These results support the fact that the pocket of MAB mmpL3 protein is able to fit larger compounds due to mutations and side chain orientations, and more polar groups can be used to grow the inhibitors towards that region.

6.4 Conclusions

Computational modeling approaches are instrumental in understanding the affinities and motions of compounds of interest in their binding sites. Here, docking and molecular dynamics simulations were used to explain and investigate the selected mmpL3 inhibitor molecules and provided a basis for their inhibitory mechanisms. Furthermore, the affinity differences observed for Mtb and MAB species towards certain candidate compounds were also analyzed.

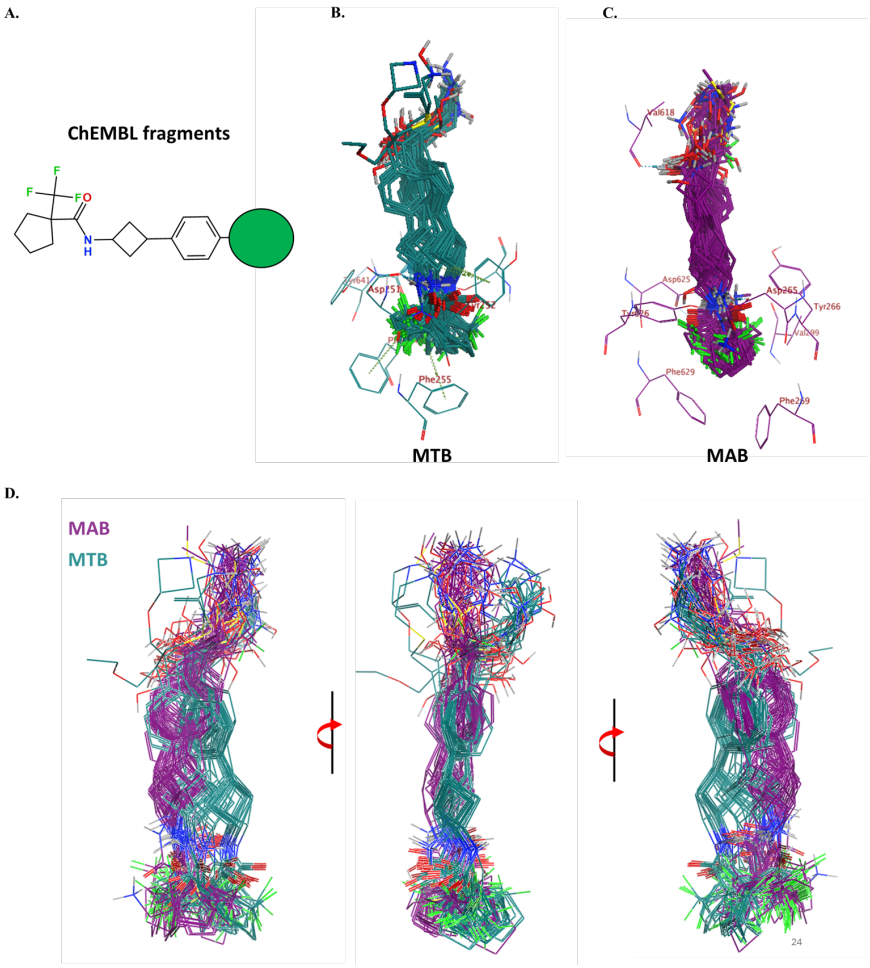


Figure S5.18 A. The scaffold used for the fragment search using ChEMBL fragment database on MOE. B. Superimposed docked poses of selected compounds obtained from fragment search, docked on Mtb mmpL3 pocket. The Phe residues and Asp-Tyr pairs are shown in line representation and the docked compounds shown in stick representation. C. Superimposed docked poses of selected compounds obtained from fragment search, docked on MAB mmpL3 pocket. The Phe residues and Asp-Tyr pairs are shown in line representation and the docked compounds shown in stick representation. D. Overlay of Mtb (dark cyan) and MAB (purple) docked poses. Compounds can be seen in Figure S4.

BIBLIOGRAPHY

[1] Suarez, L. Y. T. (2020). Global tuberculosis report 2020.

[2] Barberis, I., Bragazzi, N. L., Galluzzo, L., and Martini, M. (2017). The history of tuberculosis: From the first historical records to the isolation of koch's bacillus. *Journal of Preventive Medicine and Hygiene*, 58:E9–E12.

[3] CDC (2018). Infection control & prevention, fact sheet - bcg vaccine.

[4] (WHO), W. H. O. (2016). Treatment of tuberculosis: guidelines, 4th edition.

[5] Wehrli, W. (1983). Rifampin: Mechanisms of action and resistance. *Reviews of Infectious Diseases*, 5:S407–S411.

[6] Schubert, K., Sieger, B., Meyer, F., Giacomelli, G., Böhm, K., Rieblinger, A., Lindenthal, L., Sachs, N., Wanner, G., and Bramkamp, M. (2017). The antituberculosis drug ethambutol selectively blocks apical growth in CMN group bacteria. *mBio*, 8(1).

[7] Zhang, Y., Shi, W., Zhang, W., and Mitchison, D. (2014). Mechanisms of Pyrazinamide Action and Resistance. *Microbiology Spectrum*, 2(4):1.

[8] Vilchèze, C. and Jacobs, W. R. (2019). The Isoniazid Paradigm of Killing, Resistance, and Persistence in Mycobacterium tuberculosis.

[9] Marks, S. M., Flood, J., Seaworth, B., Hirsch-Moverman, Y., Armstrong, L., Mase, S., Salcedo, K., Oh, P., Graviss, E. A., Colson, P. W., Armitige, L., Revuelta, M., and Sheeran, K. (2014). Treatment practices, outcomes, and costs of multidrug-resistant and extensively drug-resistant tuberculosis, United States, 2005-2007. *Emerging Infectious Diseases*, 20(5):812–821.

[10] Edwards, B. D. and Field, S. K. (2022). The struggle to end a millennia-long pandemic: Novel candidate and repurposed drugs for the treatment of tuberculosis. *Drugs 2022 82:18*, 82:1695–1715.

[11] Domenech, P., Reed, M. B., and Barry, C. E. (2005). Contribution of the Mycobacterium tuberculosis MmpL protein family to virulence and drug resistance. *Infection and Immunity*, 73(6):3492–3501.

[12] Su, C. C., Klenotic, P. A., Bolla, J. R., Purdy, G. E., Robinson, C. V., and Yu, E. W. (2019). MmpL3 is a lipid transporter that binds trehalose monomycolate and phosphatidylethanolamine. *Proceedings of the National Academy of Sciences of the United States of America*, 166(23):11241–11246.

[13] Jackson, M. (2014). The mycobacterial cell envelope-lipids. *Cold Spring Harbor Perspectives in Medicine*, 4(10).

[14] Bansal-Mutalik, R. and Nikaido, H. (2014). Mycobacterial outer membrane is a lipid bilayer and the inner membrane is unusually rich in diacyl phosphatidylinositol dimannosides. *Proceedings of the National Academy of Sciences of the United States of America*, 111(13):4958–4963.

[15] Daffé, M., Crick, D. C., and Jackson, M. (2014). Genetics of Capsular Polysaccharides and Cell Envelope (Glyco)lipids. *Microbiology Spectrum*, 2(4).

[16] Yang, X., Hu, T., Yang, X., Xu, W., Yang, H., Guddat, L. W., Zhang, B., and Rao, Z. (2020). Structural Basis for the Inhibition of Mycobacterial MmpL3 by NITD-349 and SPIRO. *Journal of Molecular Biology*, 432(16):4426–4434.

[17] Zhang, B., Li, J., Yang, X., Wu, L., Zhang, J., Yang, Y., Zhao, Y., Zhang, L., Yang, X., Yang, X., Cheng, X., Liu, Z., Jiang, B., Jiang, H., Guddat, L. W., Yang, H., and Rao, Z. (2019). Crystal Structures of Membrane Transporter MmpL3, an Anti-TB Drug Target. *Cell*, 176(3):636–648.e13.

[18] Xu, Z., Meshcheryakov, V. A., Poce, G., and Chng, S. S. (2017). MmpL3 is the flippase for mycolic acids in mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 114(30):7993–7998.

[19] Grzegorzewicz, A. E., Pham, H., Gundi, V. A., Scherman, M. S., North, E. J., Hess, T., Jones, V., Gruppo, V., Born, S. E., Korduláková, J., Chavadi, S. S., Morisseau, C., Lenaerts, A. J., Lee, R. E., McNeil, M. R., and Jackson, M. (2012). Inhibition of mycolic acid transport across the Mycobacterium tuberculosis plasma membrane. *Nature Chemical Biology*, 8(4):334–341.

[20] Rosa, V. L., Poce, G., Canseco, J. O., Buroni, S., Pasca, M. R., Biava, M., Raju, R. M., Porretta, G. C., Alfonso, S., Battilocchio, C., Javid, B., Sorrentino, F., Ioerger, T. R., Sacchettini, J. C., Manetti, F., Botta, M., Logu, A. D., Rubin, E. J., and Rossi, E. D. (2012). Mmpl3 is the cellular target of the antitubercular pyrrole derivative bm212. *Antimicrobial Agents and Chemotherapy*, 56(1):324–331.

[21] Tahlan, K., Wilson, R., Kastrinsky, D. B., Arora, K., Nair, V., Fischer, E., Barnes, S. W., Walker, J. R., Alland, D., Barry, C. E., and Boshoff, H. I. (2012). SQ109 Targets MmpL3, a Membrane Transporter of Trehalose Monomycolate Involved in Mycolic Acid Donation to the Cell Wall Core of Mycobacterium tuberculosis. *Antimicrobial Agents and Chemotherapy*, 56(4):1797–1809.

[22] Williams, J. T., Haiderer, E. R., Coulson, G. B., Conner, K. N., Ellsworth, E., Chen, C., Alvarez-Cabrera, N., Li, W., Jackson, M., Dick, T., and Abramovitch, R. B. (2019). Identification of new MMPL3 inhibitors by untargeted and targeted mutant screens defines MMPL3 domains with differential resistance. *Antimicrobial Agents and Chemotherapy*, 63(10).

[23] Zhang, Y. (2008). I-tasser server for protein 3d structure prediction. *BMC Bioinformatics*, 9:1–8.

[24] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, 12(1):7–8.

[25] Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: A unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738.

[26] Edelsbrunner, H. (1992). Weighted alpha shapes. Technical report, Technical paper of theDepartment of Computer Science of the University of Illinois atUrbana-Champaign, Urbana, Illinois.

[27] Corbeil, C. R., Williams, C. I., and Labute, P. (2012). Variability in docking success rates due to dataset preparation.

[28] Labute, P. (2008). The generalized born/volume integral implicit solvent model: Estimation of the free energy of hydration using London dispersion instead of atomic surface area. *Journal of Computational Chemistry*, 29(10):1693–1698.

[29] Zheng, H., Williams, J. T., Coulson, G. B., Haiderer, E. R., and Abramovitch, R. B. (2018). HC2091 kills mycobacterium tuberculosis by targeting the MmpL3 mycolic acid transporter. *Antimicrobial Agents and Chemotherapy*, 62(7).

APPENDIX A

SUPPORTING TABLES

Table S6.1 MM-GBSA energy values for addtionoal simulated systems in Mtb and MAB mmpL3 binding pockets.

	Mtb		MAB	
Compound	**AVG MM-GBSA (kcal/mol)**	**std (kcal/mol)**	**AVG MM-GBSA (kcal/mol)**	**std (kcal/mol)**
MSU45675	-42.21	2.25	-39.34	2.62
MSU45677	-45.05	2.75	-35.59	2.68
MSU45518	-47.10	3.16		
MSU45518_7	-44.41	3.58	-36.91	2.52
MSU45518_4	-48.22	3.42		
MSU45518_3	-45.79	3.07	-45.09	2.88
MSU45518_2	-45.82	5.19		
MSU45518_1	-41.23	3.67	-37.91	2.92
MSU45435	-45.79	2.41	-43.42	2.87
MSU45434	-42.91	2.97	-37.17	2.37
MSU45416	-49.46	2.80	-39.81	2.85
MSU45416_1	-46.47	2.65	-40.20	2.85
MSU45370	-49.97	2.87	-39.65	2.63
MSU45287	-45.47	2.92	-44.38	2.41
MSU43085	-54.89	3.38	-47.35	2.91

APPENDIX B

SUPPORTING FIGURES

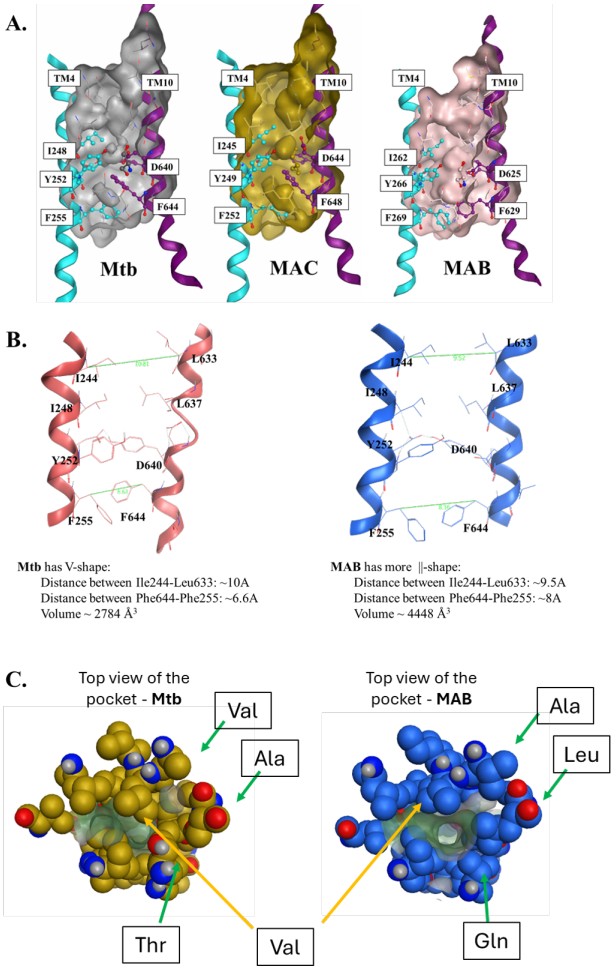


Figure S6.1 A. The pocket surfaces for Mtb, MAC, and MAB mmpL3 proteins along with TM4 and TM10 are shown. The residues are shown in stick and line representation. B. The distances between the TM4 and TM10 helices in Mtb and MAB mmpL3 proteins. The cytoplasmic side of the pocket was defined by the two Phe residues, and the periplasmic side was defined by the Ile-leu residues. C. The top view of the pockets of Mtb and MAB mmpL3 proteins. The binding pocket was shown in green mesh, and the pocket residues are shown in vdW representation. Due to the mutations showed above, MAB mmpL3 has more accessible space on the periplasmic side of the pocket.

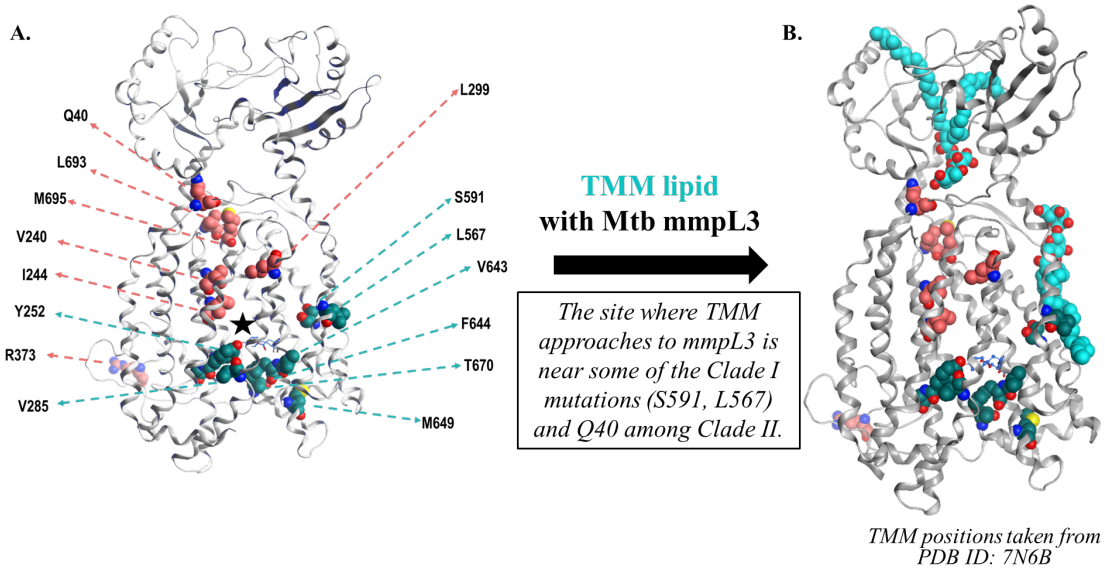


Figure S6.2 A. The residues tested from the cross-resistance Z-matrix data from CITE. Blue corresponds to the residues selected from Clade I and the red ones correspond to the Clade II resdiues. The Mtb mmpL3 protein is shown with grey ribbon representation,and the binding pocket is indicated with a star. B. The superimposed structure of the Mtb mmpL3 protein from (A) with mmpL3 protein co-srystallized with TMM lipids. The positions of some identified resistant mutations and TMM lipid binding sites do overlap.

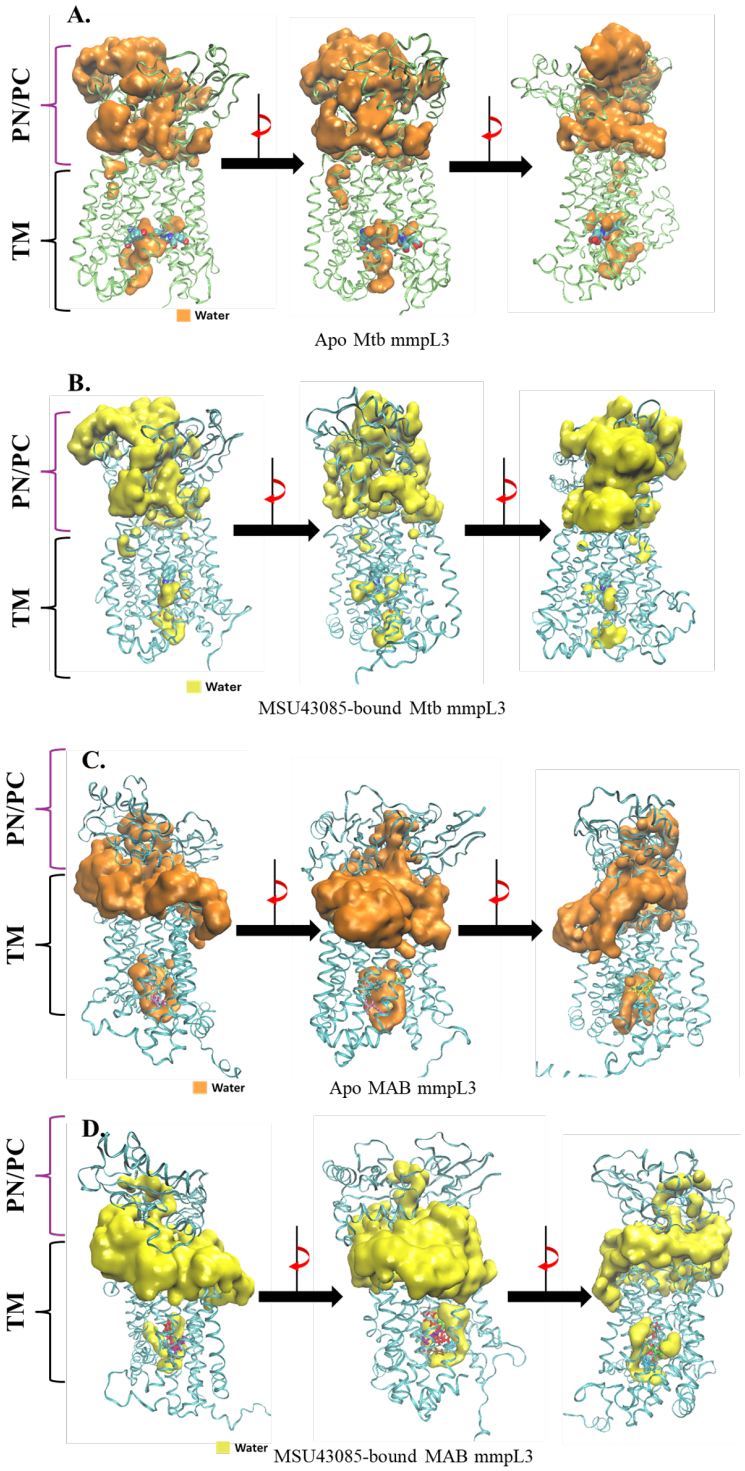


Figure S6.3 Water exposure surfaces for the periplasmic domains and the TM channel of Mtb and MAB proteins. A. Apo Mtb mmpL3 protein, B. MSU43085-bound Mtb mmpL3 protein, C. Apo MAB mmpL3 protein, and D. MSU43085-bound MAB mmpL3 protein. PC/PN:perilasmic C/periplasmic N terminal. TM: transmembrane region.

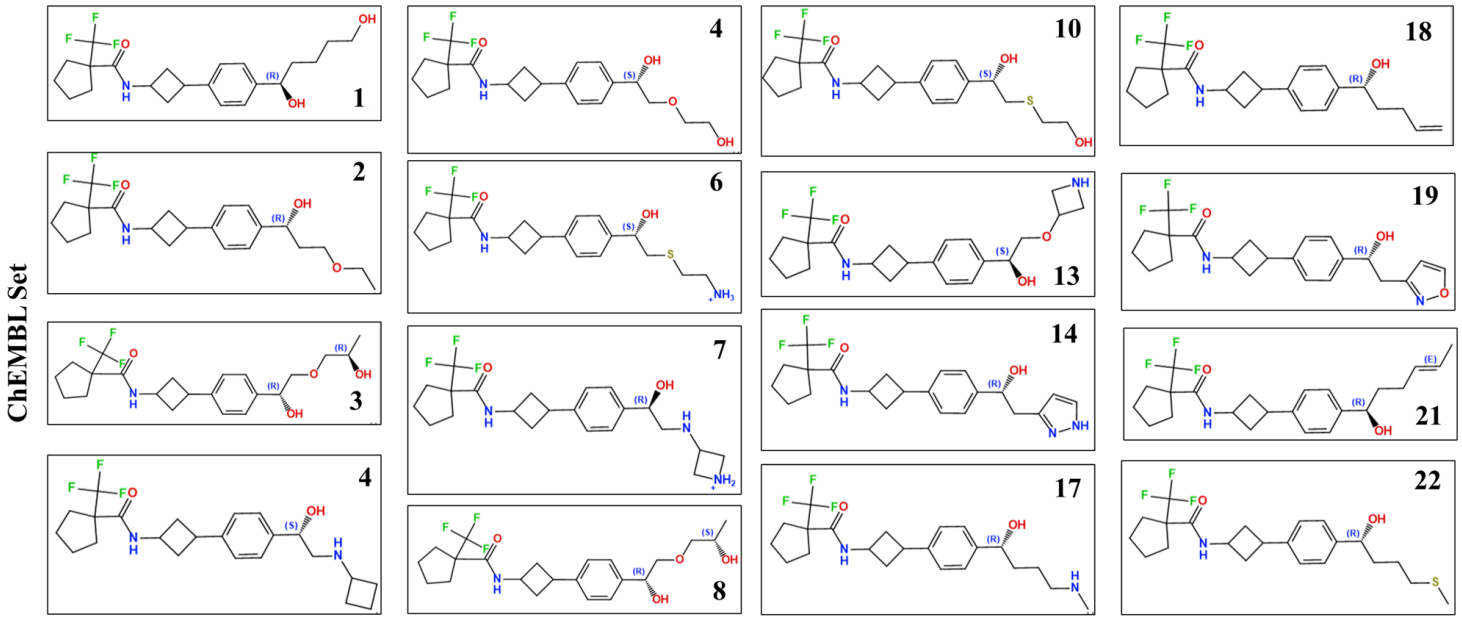


Figure S6.4 Examples of compounds generated using fragment search with ChEMBL database in MOE.

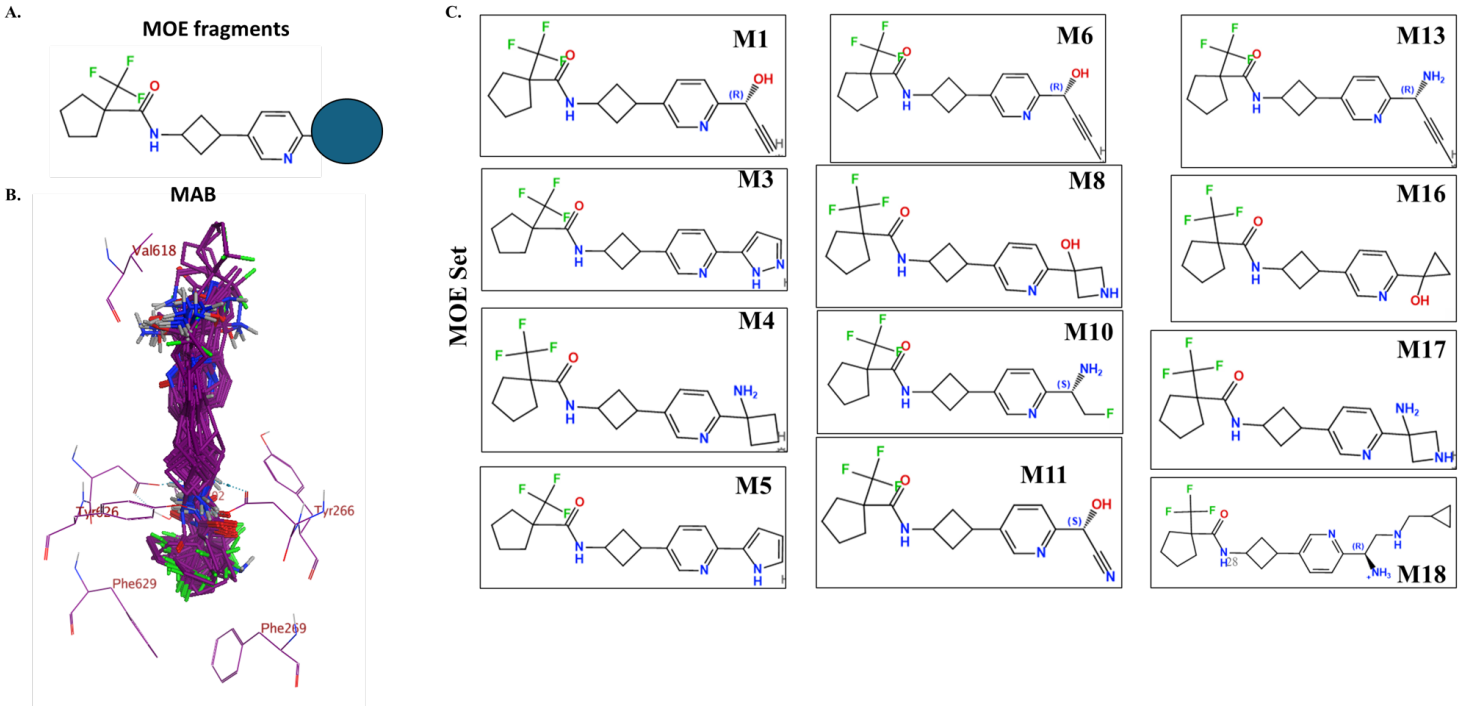


Figure S6.5 A. The scaffold used for the fragment search using MOE fragment database in MOE. B. Superimposed docked poses of selected compounds obtained from fragment search, docked on MAB mmpL3 pocket. The Phe residues and Asp-Tyr pairs are shown in line representation and the docked compounds shown in stick representation. C.Examples of compounds generated using fragment search with MOE database in MOE.

CHAPTER 7

MODELING OF DOSS INTERACTIONS WITH SMALL MOLECULE INHIBITORS AS A SUPPLEMENTARY TREATMENT STRATEGY AGAINST TB

7.1 Introduction

7.1.1 Sensing the Environment: DosRST Proteins

While many of the relevant references have already been provided in Chapter 7, for this chapter, only new references relevant to DosRST have been included.

Another critical target for TB treatment that has been identified belongs to the DosRST two-component system. Two-component systems (TCS) generally consist of one sensing protein, histidine kinase, and one response regulatory element. A significant amount of homology is shared among the TCS proteins in different bacteria, indicating that TCS is an important regulatory system. In Mycobacterium species, TCS is involved in the regulation of intracellular multiplication during the early infection period, the regulation of genes that are involved in pathogenesis, and the adaptation to pH and hypoxia hence controlling the non-replicating persistence (NRP)[1]. TCS changes the expression of the NRP genes that would allow the bacteria to survive non-optimal conditions, and this plays an important role in TB pathogenicity and treatment length. In *M. tuberculosis*, there are 11 known TCS, among them two are essential (MtrAB and DosRST)[1]. Disrupting the environmental sensing for TB treatment has been the focus for some time, and in this report, we will cover DosRST systems as well.

DosRST is a member of TCS, however, it differs by having two sensing histidine kinases instead of one. These kinases, DosS and DosT, react to the change in hypoxia (lack of oxygen) and redox change in the environment by binding to O_2, CO and NO through the heme group. Upon binding, the proteins switch to the "active" mode and autophosphorylate themselves. Then, the phosphate is transferred to the regulatory response element, DosR. Phosphorylated DosR dimerizes and binds to a specific conserved region on DNA to regulate the expression of almost 50 genes[1]. DosS and DosT share about 60% sequence identity, and they also have structural homology. Both DosS and DosT have GAF domain containing heme group in N-terminal, histidine kinase domain, and ATP-binding domain in the C-terminal. Although the specifics of the sensing through heme and how the autophosphorylation is triggered are not known currently, the hypothesis is that ligand-heme interactions induce a conformational change that would cause changes in the overall structure to

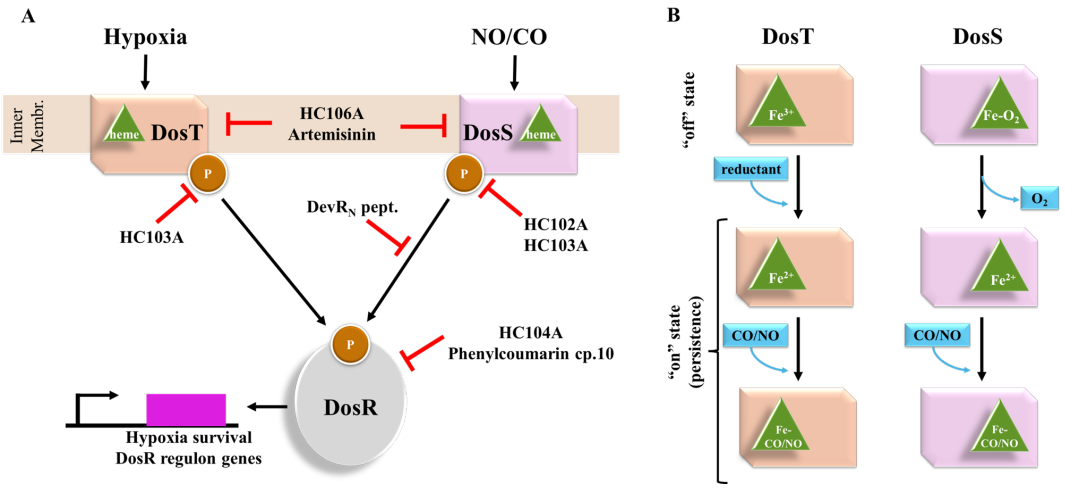


Figure S6.1 A: Schematic for the DosRST signaling pathway, with examples of where small molecules and peptides interfere with DosRST signaling[1]. B: Proposed mechanism for the role of *M. tuberculosis* DosS and DosT in the shift down of tubercle bacilli to the persistent state[2,3]. Normoxia: Normal oxygen conditions, hypoxia: low oxygen condition.

trigger the autophosphorylation. The specific ligands of DosS and DosR are not exactly known, but a recent study shows that DosS functions in reduced Fe^{2+} state, indicating that it could be a redox sensor[2,3]. The proposed mechanisms of activation for DosS and DosR is shown in Fig.3.6. The current proposed mechanism for DosS is as follows: in the off state, Fe is oxidized, and with the help of a reducing agent (such as flavin nucleotides), Fe is reduced to Fe^{2+} switching the protein "on". During an "on" state, either NO or CO can bind to the reduced iron causing a conformational change that will trigger the autophosphorylation[2]. For DosT, the off-state corresponds to O_2-bound Fe. Under hypoxia, O_2 is released, and iron is reduced, triggering the autophosphorylation[4]. After the transfer of phosphate from either DosS or DosT to DosR, DosR dimerizes and binds to the regulatory sequence on DNA. The mutation studies with DosRST proteins showed *dosR* mutation did not inhibit the virulence. However, *dosRS* mutants caused growth defects[1].

There are a number of molecules that have been linked to the DosRST system using HTS. "Compound 10" has been found to inhibit DosR-regulated gene expression under normoxia - normal oxygen conditions. In another HTS study that scanned more than 540,000 compounds, six distinct molecules were discovered to inhibit DosRST[5]. Among them, Artemisinin acted by oxidizing the heme group in both DosS and DosT, causing loss of sensing ability[5]. HC102A and

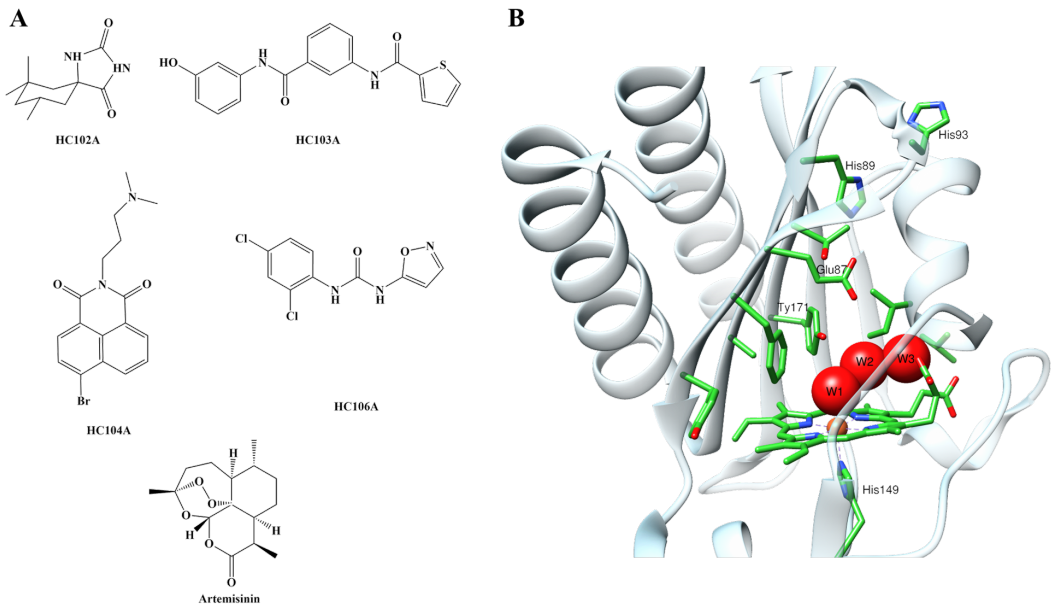


Figure S6.2 A: The structures of the compounds that target DosRST system. B: Crystal structure of the ferric DosS GAF domain (PDB ID: 2W3E).

HC103A that are shown in Figure 2 did not act on the GAF domains, but they mainly inhibited the autophosphorylation activity in DosS and DosS/T, respectively. HC104 is found to interfere with the DNA-binding mechanism of DosR, but only for a specific operon hence does not impact the overall bacterial survival. In a UV-visible spectroscopy assay for DosS protein, the presence of HC106A caused a shift in Fe^{2+} Soret peak, which is also observed when NO or CO binds to heme. However, it is believed that the mechanism of action of HC106 is different than of Artemisinin. When a residue located in the heme channel, G117, is mutated to Leucine, a resistance to both compounds has been observed. The aforementioned study also showed that both compounds access to heme through the same path[5].

7.2 Computational Details

7.2.1 Protein preparation and Docking Procedures

The existing crystal structures of DosS protein (PDB ID: 2W3D, 2W3F, 2W3E, 4YNR) were prepared using the Molecular Operating Environment (MOE) protein preparation suite at pH 7 at their appropriate iron oxidation states.[2,6,7] The iron metal center was selected as an anchor for the docking using a pharmacophore approach, which was utilized to place the investigated molecules.

Then, the top 100 poses were refined using induced fit approach and the final docking scores were calculated with Generalized-Born volume integral/weighted surface area score (GBVI/WSA dG).[7] The poses were visually analyzed and selected for further investigation. For the mutation docking studies, the residues of interest were mutated and the docking procedure was repeated.

7.2.2 Parametrization of heme group

To model the non-bonded iron interactions, 12-6 LJ parameters were used for iron-isoxazole interactions. For the bonded heme model, MCPB.py was used for streamlined parametrization of the heme and iron at different oxidation states and different number of coordinations.[8] The small model containing iron, heme, and axial coordinating histidine was optimized using B3LYP/6-31G* to calculate the force constant. The large model with the second axial coordinating moiety was used to do the RESP charge calculations[9]. Then, the Seminario method was used to generate the force field parameters.[10] As a final step, the RESP charge fitting was performed.

7.2.3 Constant pH Molecular Dynamics Simulations

To determine the pKa values hence the correct protonation states of pocket residues, CpHMD[11] simulations were performed for the following systems: Fe^{+2}, Fe^{+3}, CO-bound Fe, and Fe^{+2}-isoxazole ring. The protein was modeled with constph ff based on the ff10 force field with PBradii mbondi2 and parameters for titratable residues (histidine, aspartic acid, and glutamic acid). The CpHMD method uses Monte Carlo sampling of discrete protonation states along with a molecular dynamics simulation in an implicit solvent defined by igb = 2 ("OBC" model).[12] The leaprc.constph force field was used for protein and gaff2 force field was used for small molecules.[13] Heme and iron were parametrized based on the oxidation state of the iron, as described in the previous section. Each prepared system was simulated for the pKa calculations from pH 1 to 14 with an increment of one. For each pH, the simulations were performed for 5 ns with 2 fs timestep with an attempt to change the protonation steps at every 5 steps.

The minimization is performed in four steps with decreasing positional restraints on the heavy atoms (100, 50, 10, 0 kcal mol^{-1} $Å^{-2}$) and for each step, minimization was performed for 200000 steps with steepest descent algorithm. The systems were then heated up to 300 K using the same

ten-step heating procedure used in the mmpL3 simulations, with the addition of Gibbs implicit solvent. No protonation state change was attempted during the minimization and the heating processes.

7.2.4 Classical Molecular Dynamics Simulations

The MD simulations were performed for 100 ns with an explicit solvent described by the TIP3P water model.[14] The ff14SB force field was used for protein, and gaff2 force field was used for small molecules.[15] AM1-BCC was used to calculate the partial charges on the inhibitor compounds.[16] The minimization is performed in four steps with decreasing positional restraints on the heavy atoms (100, 50, 10, 0 kcal mol $^{-1}$ Å^{-2}) and for each step, minimization was performed for 200000 steps with steepest descent algorithm. The systems were then heated up to 300 K using the same ten-step heating procedure used in the mmpL3 simulations. In the production run, 1 fs timestep was used along with Langevin thermostat and isobaric barostat. The 100 trajectories were saved for each nanosecond of the simulation.

7.2.5 Analysis of Trajectories

The cpptraj suite of AmberTools20 was used for analysis of the trajectories. RMSD, RMSF, radial distribution plots, and residue decomposition energies were calculated with cpptraj module and plotted using gnuplot.[17,18]

7.3 Results and Discussion

7.3.1 New Therapeutics for DosRST Inhibition with Docking Studies

In this section, the docking results for HC106A compound and its derivatives to test the docking parameters used are shown. The DosS crystal structure GAF domain (PDB ID: 2W3E) is selected and prepared using MOE. The potential binding site is found using MOE Site Finder tool by taking the mechanism of action of known compounds into account. This resulted in a pocket where a compound can coordinate with the iron in the heme group. The compounds are docked in the pocket using a pharmacophore approach. The docking approach is exactly the same as described in the Methodology section, with the exception of the placement step in which a pharmacophore is used. The docking of HC106A shows that the compound coordinates with iron through the

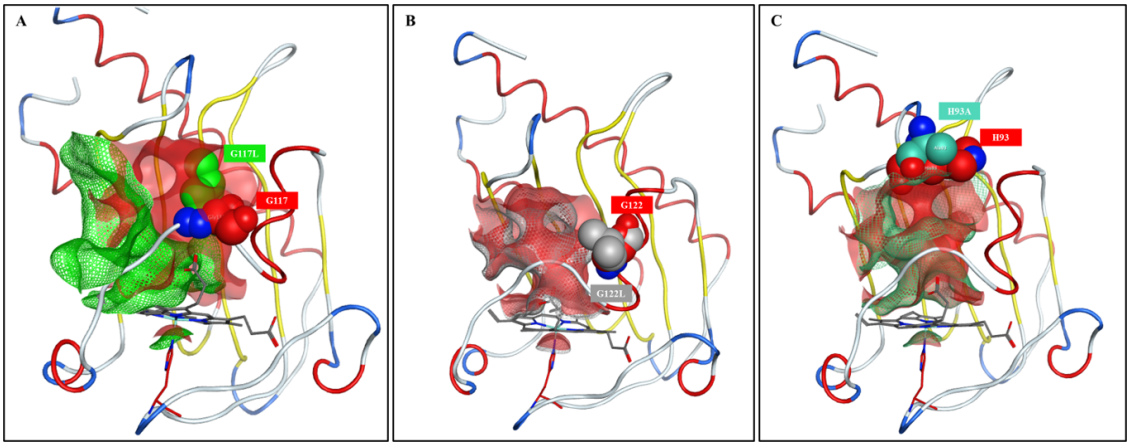


Figure S6.3 Overlapping of binding pockets with respect to WT structure. Mutated residues are shown with vdW sphere representation. WT is shown in red color in A-C. A. Gly117Leu mutation is shown in green. B. Gly122Leu mutation is shown in gray. C. His93Ala mutation is shown in cyan.

isoxazole ring, and the urea group coordinates with the carboxyl groups in the heme (Fig.4). The pocket consists of a relatively hydrophobic region around the isoxazole ring, and a more solvent exposed area located near aromatic ring. It is known in the literature that Gly117 plays an important role for DosS resistance against HC106A. Therefore, Gly117Leu mutation is used for docking. In addition, Gly122Leu and His93Ala mutations are tested for their effect on docking of HC106A compound. The docking results show that the Gly177Leu mutation is indeed causing an unstable binding of HC106A compound (Fig. 4). The docking score is above zero and compared to WT and other mutations, it is very high. The reason for this is that Leu is has a bulkier side chain than Gly, therefore, it occupies the binding pocket (Fig. 3), making it harder for a compound to dock. Furthermore, the interaction with iron is completely lost due to the rotation of the isoxazole ring. On the other hand, when compared to Gly117Leu mutation, Gly122 to Leu mutation did not cause a significant difference for docking of HC106A in terms of binding orientation. Similarly, the His93 residue located at the top of the pocket is mutated into Ala, and it did not cause any change in the binding of HC106A compound. One problem with H93A docking is that the hydrogens in the urea group in the WT interacts with carboxyl of heme, however, in His93Ala, one of them rotates and interacts with the Glu87 residue.

The modeling of the synthesized compounds (Table 1) was performed using molecular docking.

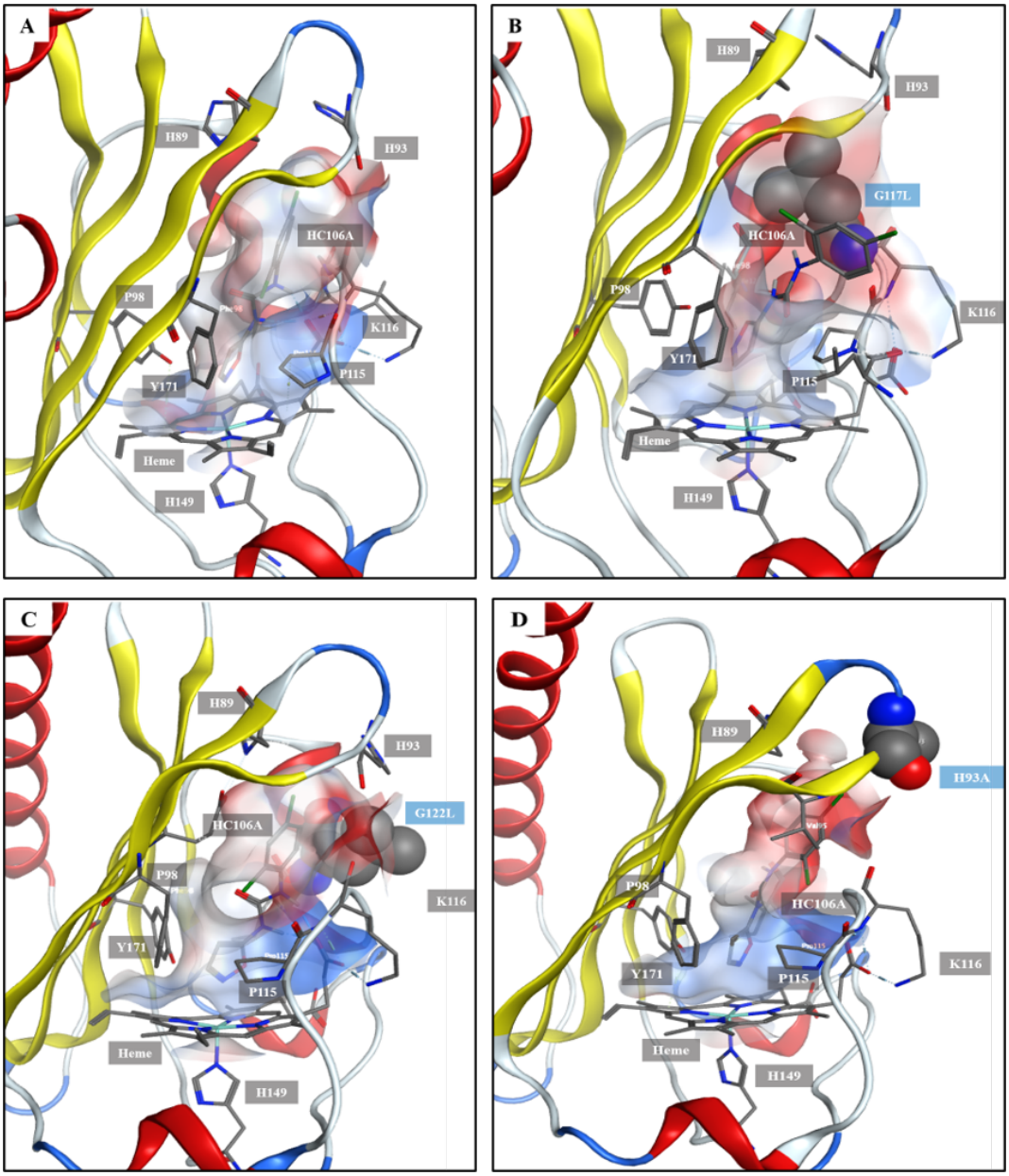


Figure S6.4 The pocket surface is shown for HC106A in WT and mutated structures. The coloring follows the same scheme used in Figure 13. A. Wild-type DosS (docking score: -0.87). B. G117L mutation (docking score: 4.55). C. G122L mutation (docking score: -1.31). D. H93A mutation (docking score: -1.33).

The binding pocket consists of mainly non-polar residues (Phe98, Val95, Pro115, Gly117, Ile121, Ile125, and Tyr171) and three polar residues (Glu87 and His 89 and His93). For all of the compounds that are shown to be active, coordination with a heme iron through the isoxazole ring was observed. The urea linker region provided another contact with the heme carboxylate groups by forming hydrogen bonds, and this particular orientation was not observed for most of the inactive compounds (Fig. 5). The lipophilic-binding domain for each compound was located at the interface between the protein and the solvent, therefore, as observed from the SAR studies, the addition of a hydrogen bond donor or acceptor group provides better interactions with the polar backbone of the protein (Fig. 6). The compounds MSU-43572, MSU-43419, MSU-43424, for example, include a phenol or benzyl alcohol that is found to form interactions with the polar backbone atoms from His93 or Gly117/Lys116, depending on the orientation of the ring. The presence of the benzyl alcohol group increased the hydrophobic contact area with Ile125. Furthermore, the meta positioning of the –OH group on the phenyl ring was found to provide less polar contact area with the Val95 residue, as compared to para positioning. Replacing the –OH group with an amine group, MSU-43423, hydrogen bonding with the surrounding residues increases, however, the polar contact area with Ile125 also increases, which is not favorable.

Overall, although the docking scores are not reliable in ranking the affinities, the molecular docking studies provided a stable initial conformation for the further modeling of the compounds of interest. The orientation of key moieties and interactions with the surrounding residues, how the compounds would "sit" in the pocket, and preliminary information regarding the key residues and their mutations.

7.3.2 Determining the protonation states of the pocket residues

The next step in understanding the mechanism of inhibition of these compounds is to run a molecular dynamics simulation to obtain the time-dependent behaviors of the protein and the inhibitor molecules. However, as the oxidation state of the iron center in the DosS protein is known to change to trigger the "on" switch in the protein, the titratable residues within or near the iron center in heme may be changing their protonation states. The comparison of the crystal

Table S6.1 The docking scores and EC50s for the compounds of interest.

Compound	Docking Score	EC_{50}
MSU43423	-2.27	6.16
MSU43424	-2.71	0.21
MSU43427	-2.42	1.79
MSU43432	-3.13	0.42
MSU43452	-1.83	0.74
MSU43573	-0.20	0.13
MSU43574	-1.62	0.13
MSU43577	-2.77	4.37
MSU43578	-3.01	4.02
MSU43579	-1.87	0.13
MSU43580	-1.00	0.13
MSU43582	-0.86	0.12
MSU43612	-2.11	0.73
MSU43613	-2.75	2.03
MSU43614	-2.15	0.48
MSU43616	-3.00	0.58
MSU43617	-2.35	0.093
MSU43618	-2.43	0.093
MSU43619	-3.03	0.09
MSU43620	-3.05	0.09
MSU43672	-2.75	0.25
MSU39451	-1.95	4.10
MSU39452	-0.68	2.47
MSU39453	0.08	16.60
MSU39449	-0.61	5.20
MSU33189	-3.02	0.63
MSU39447	-2.40	0.61
MSU39446	-2.85	0.54
MSU39445	-2.35	0.75
MSU41443	-0.75	11.20
MSU41442	1.56	2.08
MSU41462	-1.67	1.12
MSU42004	-1.96	2.36
MSU41545	-0.70	4.75
MSU41542	-3.06	1.34
MSU41546	-1.24	2.15
MSU42002	0.12	1.21
MSU39444	-3.25	1.70

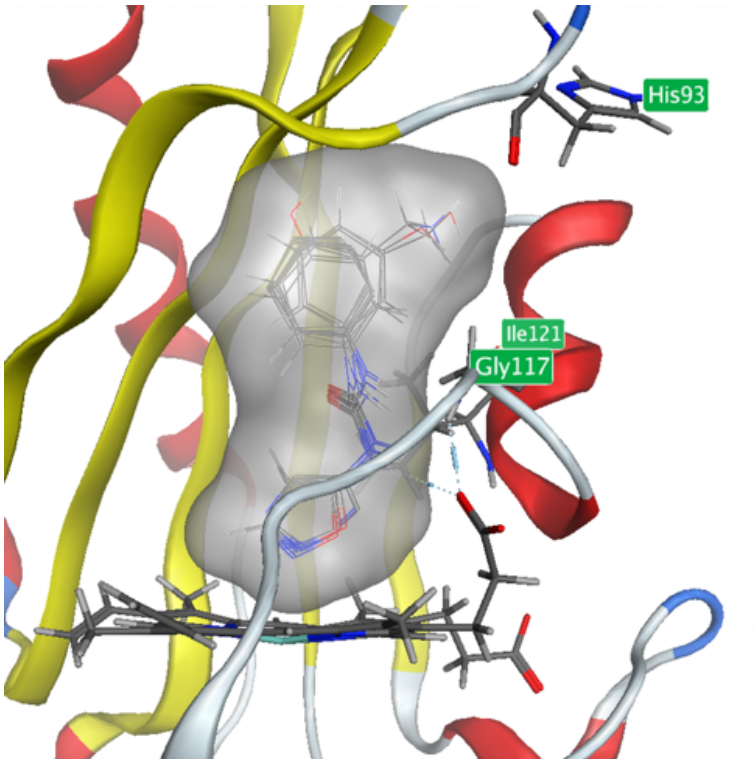


Figure S6.5 The overlapped docking orientations of selected compounds. The heme group and the residues are shown with stick representation, and the compounds are shown with line representation.

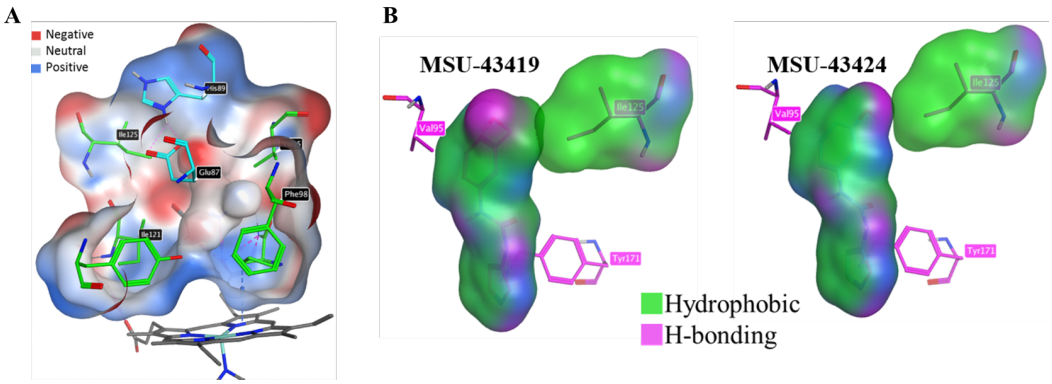


Figure S6.6 (A) The electrostatic surface of the binding pocket. (B) The lipophilic surfaces of MSU-43419 and MSU-43424 with Ile125. The residues, the heme group and the compounds are shown with stick representation.

structures of DosS at "on" and "off" states indicate that some pocket residues, for example Glu87 and His90, go through a conformation change that accompanies to the oxidation state change of iron. Constant-pH Molecular Dynamics (CpHMD) simulations are useful in determining the pKa values of titratable residues within the protein. As the iron center changes the oxidation states, accompanied by the potential conformation changes of pocket residues, during the on-off shift of the DosS protein, it is reasonable to hypothesize that the protonation states of certain titratable pocket residues, i.e. histidines, can change. Four systems were investigated to understand the pKa change in the His,Glu, and Asp residues in the protein: Fe^{+2} (5-coordinated on-state; PDB ID 2W3F), water-coordinated Fe^{+3} (6-coordinated off-state; PDB ID 2W3D), CO-Fe^{+2} (CO-bound on-state; PDB ID 4YNR), and isoxazole-coordinated Fe^{+2} (6-coordinated inhibited protein).

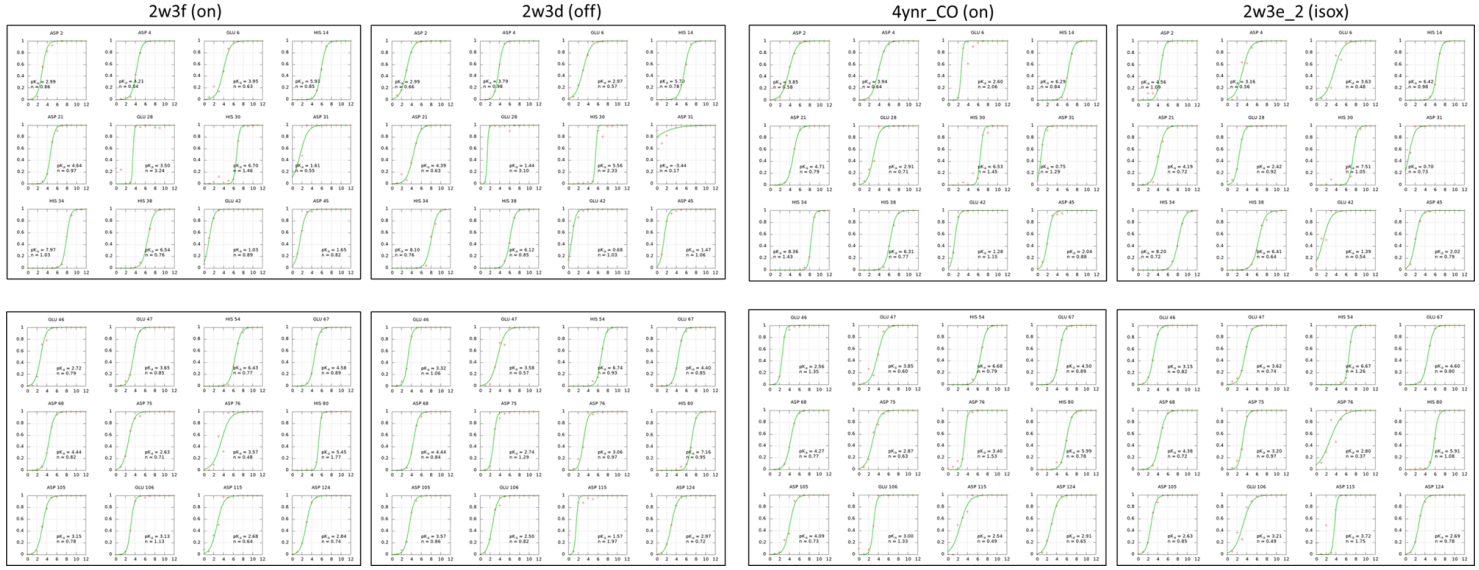


Figure S6.7 The titration plots obtained with CpHMD simulations for four systems: Fe^{+2} (5-coordinated on-state; PDB ID 2W3F), water-coordinated Fe^{+3} (6-coordinated off-state; PDB ID 2W3D), CO-Fe^{+2} (CO-bound on-state; PDB ID 4YNR), and isoxazole-coordinated Fe^{+2} (6-coordinated inhibited protein).

Comparison of 2w3f ("on" state) with 2w3d ("off" state) of the pKa values in Figure 7 reveals that the protonation state changes occur for His117 and His139. At neutral pH, His117 had very close populations for the protonated state at the Nϵ and both Nδ and Nϵ, whereas His139 was mainly protonated at Nϵ. In the "on" state, however, His117 was predominantly protonated at Nδ while His139 was double-protonated. The presence of the CO molecule coordinated to iron further changes the protonation states of these histidines. Going from the "on" state to CO-bound state, His117 has close populations for Nϵ protonated and double-protonated states, and His139 lost Nδ proton. If we compare the "on" state with isoxazole-bound state, His117 has similar populations for Nδ and doubly protonated states, while His139 loses the Nδ proton.

The addition of the isoxazole-based compounds when the DosS protein is in the "on" state was shown to shift the oxidation state of iron towards +3 from +2 state.[?] Although isoxazole itself may not be considered an inhibitor, its coordination to iron may trigger pKa shifts and conformation changes in the pocket residues. If we compare the protonation states of "off" state, which is Fe^{+3} with isoxazole-bound state, His92 is doubly protonated and in "off" state, it only has proton in Nδ.

The simulation of apo systems were performed using the corresponding protonation states of titratable residues. The "off" state and the CO-bound state of the DosS protein were analyzed to gain a better understanding of the differences between the two states (Fig. S1). While the per-residue fluctuations are quite similar in both cases, in "off" state, the RMSD of the protein is more stable than the "on" state. Furthermore, the distance between Glu87 and His89 is quite stable in both while the His87-His93 distance showed more instability for the "on" state (Fig. S1(D,H)). The water density around the aforementioned histidines is also very similar in both systems. The water presence around the iron center on the other hand is different. While the water entry to the pocket in "off" state seem to occur through where His89 and His93 are located as well as near the loop of Gly117 residue, in CO-bound systems, there was no water bridge towards the pocket (Fig. S1).

One drawback of this CpHMD approach is that the simulations were performed in implicit water, therefore, the impact of the water molecules in the pocket residues cannot be well understood. The crystal structures of both "on" and "off" states of the DosS protein has crystal water molecules,

which leads to hypothesize that to understand the conformation changes associated with the on-off switch, the CpHMD simulations with explicit waters need be performed.

7.3.3 Modeling of Heme group for Inhibitor Binding

After determining the protonation states of specific residues that may be involved in the sequence of events to change the oxidation state of iron, the methods of modeling the iron-inhibitor interactions for calculating the binding affinities of the compounds of interest were studied. The choice of modeling approach for iron metal-inhibitor coordination determines the approach that can be used for binding energy approximations. Here, two different approaches were investigated: unbound metal coordination and bound metal coordination approaches. The different oxidation states of iron also were tested with each approach. First, the observations from the unbound modeling approach will be addressed.

With the unbound modeling of the inhibitor compounds, methods such as absolute binding free energies or MM-GBSA can be used to estimate the binding affinities. However, during the simulations of unbound model, most compounds did not stay coordinated to iron during the 100 ns and either left the pocket or moved in different areas within the pocket, as can be seen in Figure 8 and Figure 9. These systems were useful in understanding the conformations that the inhibitor compounds may take before forming the interactions with iron. (+3) oxidation state of iron led to more stable complexes with the inhibitor molecule and iron, for some of the tested compounds. For instance, MSU-43686 compound with Fe^{+3} state (Fig. S2) was more stable and sustained a shorter distance between iron and isoxazole for a longer time than the Fe^{+2} counterpart. However, overall, the majority of the compounds were either leaving the pocket during the 100 ns simulation time or losing the interactions and the coordination with heme group and trying to leave the pocket. Therefore, this type of unbound modeling with iron, even with the correct oxidation state, was deemed not to be an appropriate approach to investigate the inhibitor molecules.

On the other hand, it provided some evidence in terms of which paths can be used to enter/exit the pocket. Two unique paths were identified: one near the space between the Gly117 and His93 is located, and the other one is where the water channel was formed in the apo simulations. This is

shown in the Fe^{+2} unbound modeling of MSU-43683 (Fig.8) molecule where it is first losing the coordination and interaction with iron and heme, respectively, and then, exit the pocket through a path between the His89 and His93 residues. This path was identified as a water entry path to the pocket in the simulations of the "off" state. Another potential exit path was seen in HC106A simulations (Fig. 9). The molecule is quite stable during the frst 30 ns of the simulation time, however, after that, HC106A starts to lose its interaction with the heme and the iron and assumes an almost parallel placement with the heme plane by orienting the upper end of the molecule towards the Pro115.

Following the non-bonded models, next, the bonded model using the MCPB.py package to model iron-isoxazole interaction was teste; it was then used it to model the rest of the compounds. With this methodology, to estimate the binding affinities, the ABFE approach is not suitable due to the bond existing between the nitrogen of isoxazole and the iron atom, therefore, either MM-GBSA or Relative binding free energy (RBFE) methods can be applied. The rationale behind selecting the isoxazole for parametrization instead of the whole compound is (i) to save time and computational cost as these parametrization procedures involve QM-based optimization and charge calculations, (ii) any other ring replacement in place of isoxazole does not provide any activity against DosS. Therefore, the iron-isoxazole group for the bonded model was parametrized. To understand the dynamics of the inhibitor compounds within the pocket, 100 ns-long simulations were performed, and the interaction energies were calculated for the pocket residues, including the heme group. For all compounds listed in Figure 10, the coordination with Fe was stable throughout the simulations. Overall, the interactions of the compounds with the surrounding pocket residues (shown in Figure S3) suggest that the –NH on the indazole ring provides a strong H-bond with His93 backbone. When hydrogen on indazole –NH is removed, the interaction is lost. The presence of the inhibitor molecules did not prevent waters from entering the pocket, and those water molecules act as water bridges between compounds and the pocket residues, including providing a coordination between heme with the urea group occurs through a water bridge. The nitrogen closer to the isoxazole group provides the majority of the interaction with water (it shows that the second nitrogen may not be as

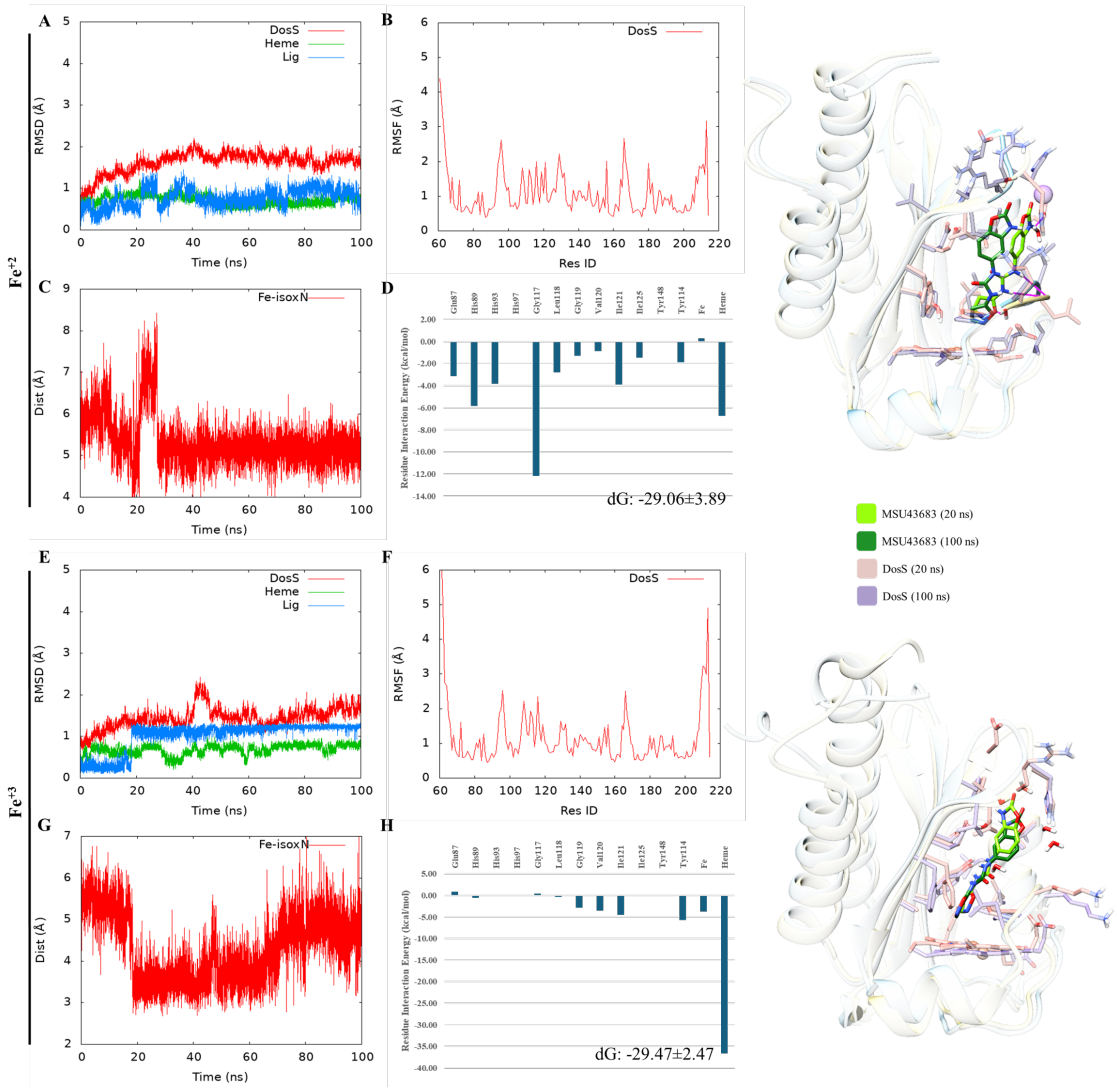


Figure S6.8 Fe^{+2} and Fe^{+3} unbound modeling for MSU-43683 molecule. A, E: RMSD time-series of the DosS protein, heme and iron, and inhibitor molecule. B, F: Per-residue RMSF plot. B, G: The distance between the iron and nitrogen from isoxazole. D, H: Per-residue interaction energies. The MM-GBSA binding energies are also given as dG values.

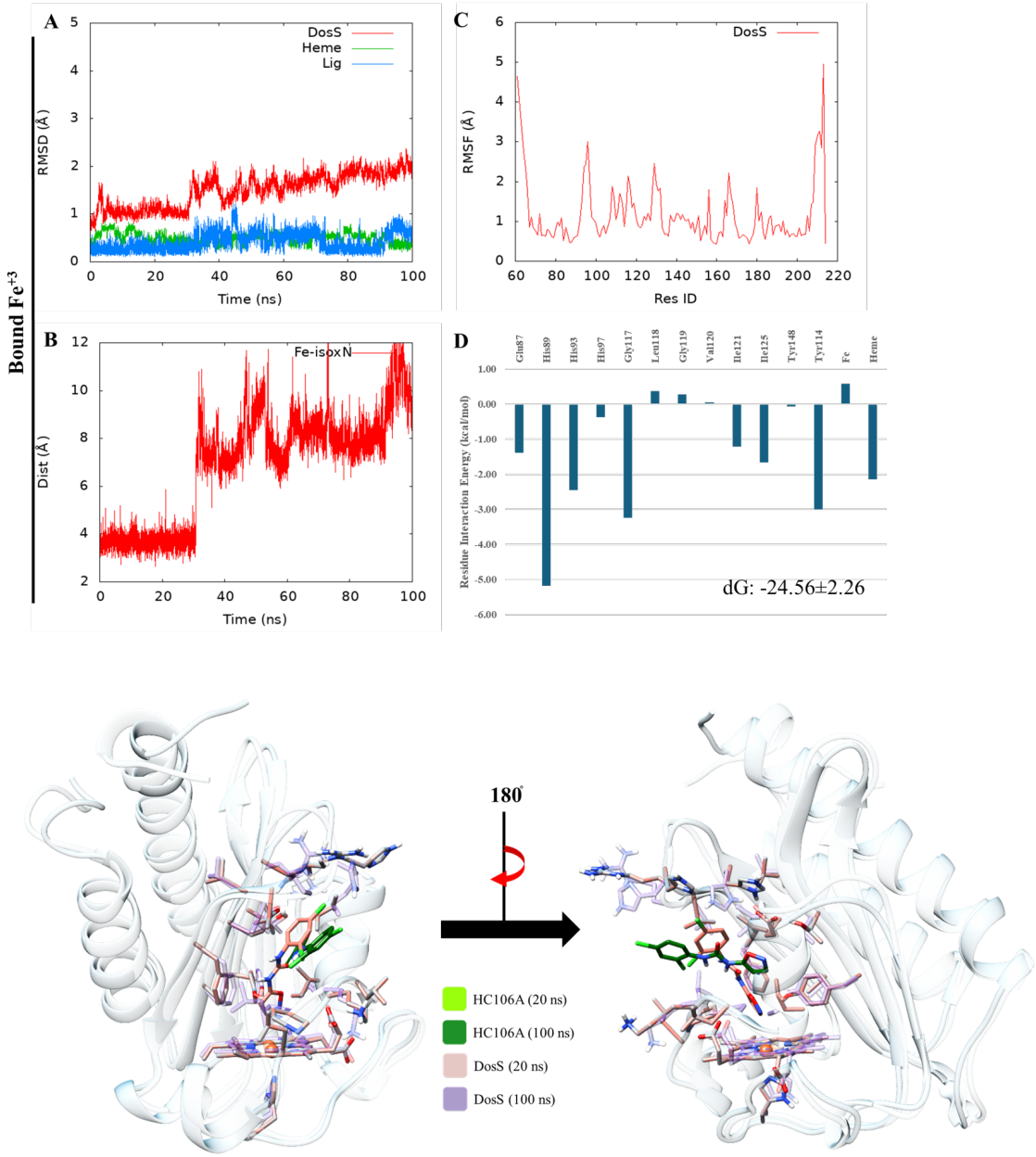
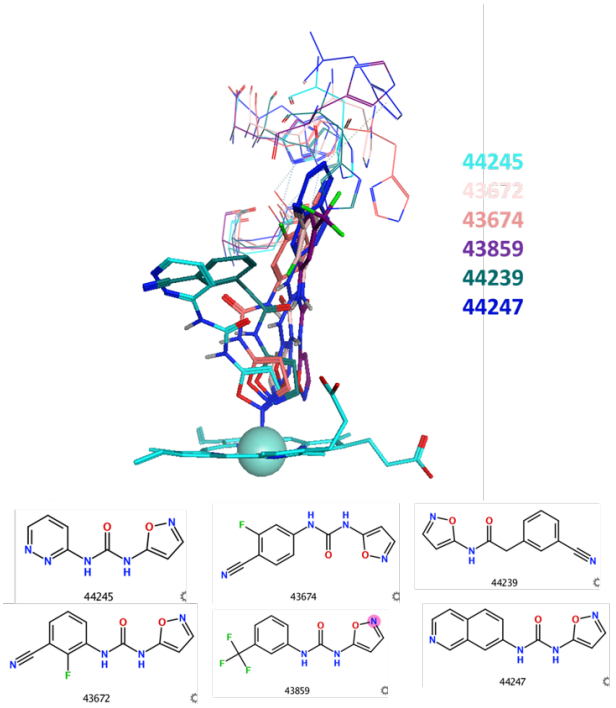


Figure S6.9 Fe^{+3} unbound modeling for HC106A molecule. A: RMSD time-series of the DosS protein, heme and iron, and inhibitor molecule. B: Per-residue RMSF plot. C: The distance between the iron and nitrogen from isoxazole. D: Per-residue interaction energies. The MM-GBSA binding energy are also given as the dG value. The binding orientation of HC106A at two different timepoints during the simulation are shown at the bottom of the figure.



Compound	MM-GBSA (kcal/mol)	MM-PBSA (kcal/mol)	ABFE (kcal/mol)	EC50 (uM)
MSU43419	-38.04±2.81	-23.76±2.98	--	0.01
MSU43672	-33.42±2.34	-19.34±2.69	-71.17±0.25	1.17
MSU43674	-38.82±2.37	-26.46±2.83	-50.17±0.19	0.37
MSU43859	-36.79±2.80	-28.39±2.50	--	2.64
MSU44239	-36.49±2.12	-26.62±2.35	--	--
MSU44245	-28.97±2.10	-20.95±2.13	--	--
MSU44247	-37.06±3.34	-24.55±3.56	--	--

Figure S6.10 Fe^{+3} simulations with isoxazole parameters and corresponding binding affinities estimated with different methods.

important). Furthermore, no conformation change was observed for Glu87 residue in the presence of the investigated compounds. The majority of the compounds formed a stable interaction with His89.

The RMSD time-series data from HC106A shown in Figure 11 and MSU-43672 shown in Figure 12 indicate that both the protein and the ligands have reached an equilibrium within the investigated simulation time. Furthermore, their per-residue RMSF values are quite similar as well. The most interesting difference between these two simulations is the interaction energies with the pocket residues. While HC106A formed a very strong interaction with the heme group due to the direct hydrogen bond formation with its urea group, MSU-43672 has multiple strong interactions with other pocket residues. For both compounds, however, interaction with the iron has a destabilizing contribution (larger than zero).

7.4 Conclusions

Modeling of metal-containing protein active sites using classical methods is indeed a challenging problem. Here, a protein sensor with a heme group that changes the oxidation state of its iron center

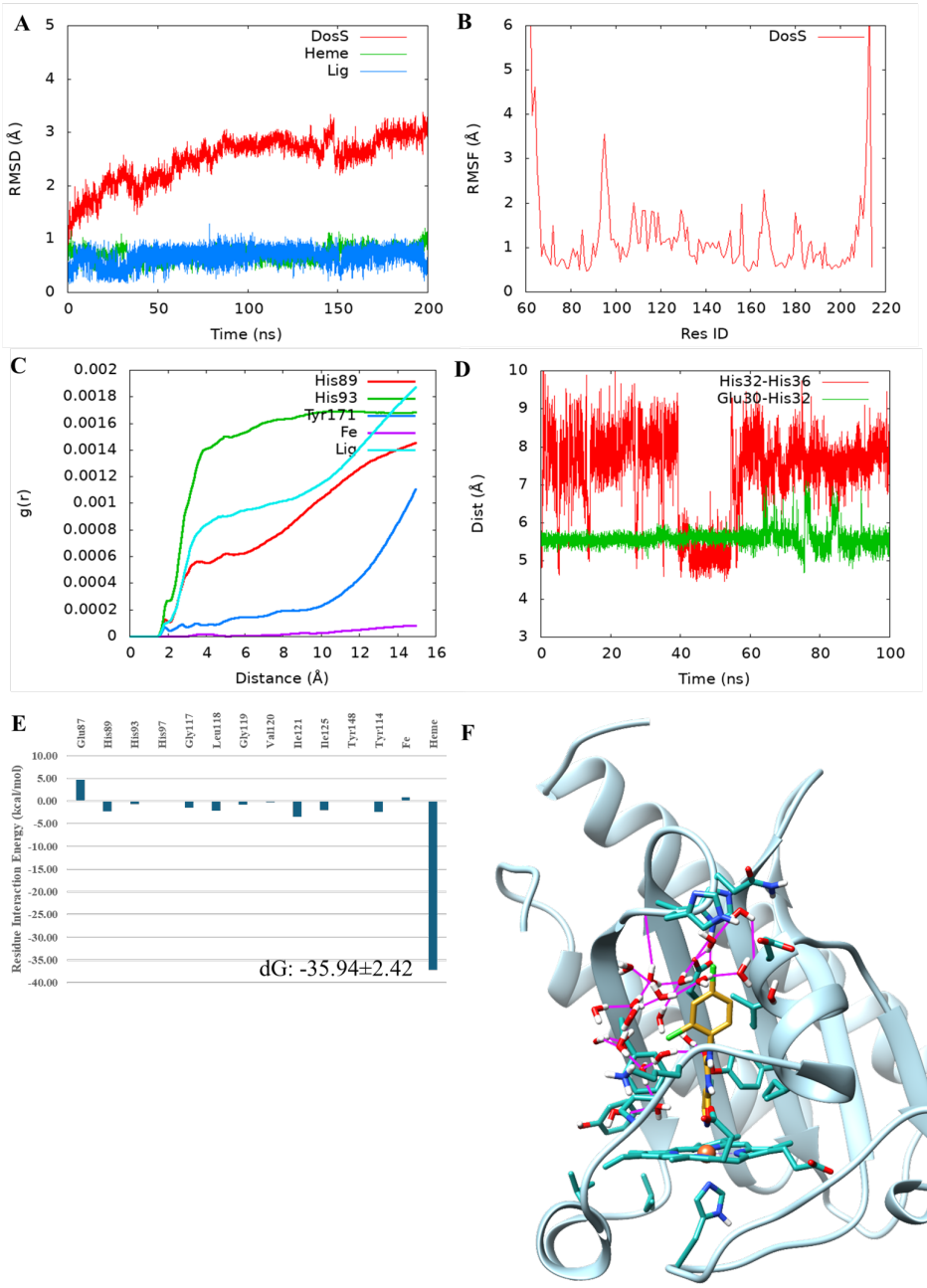


Figure S6.11 Results for HC106A bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and HC106A. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the HC106A simulations. The interactions between the water around the pocket residues are shown with pink lines.

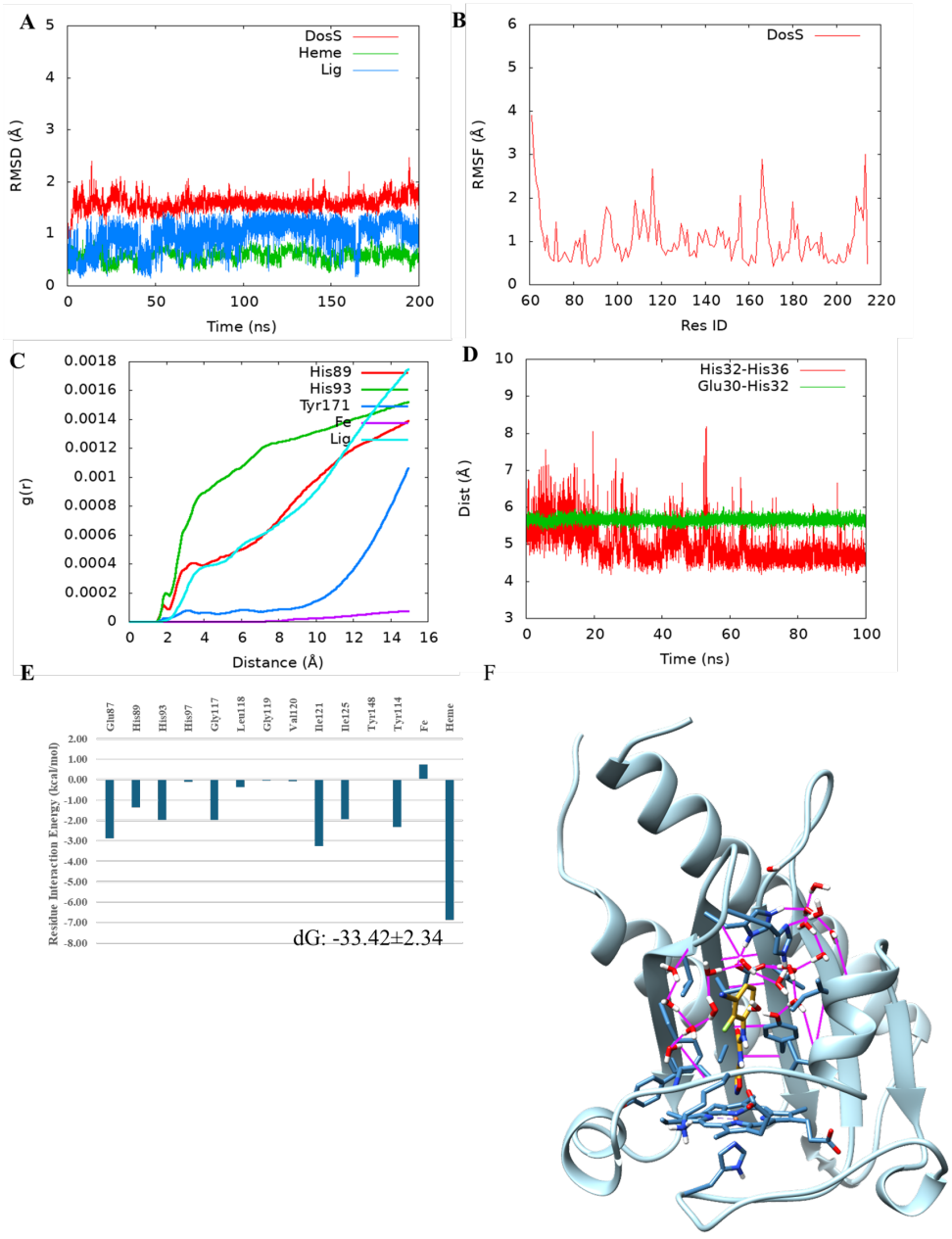


Figure S6.12 Results for MSU-43672 bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and MSU-43672. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the MSU-43672 simulations. The interactions between the water around the pocket residues are shown with pink lines.

when the protein switches to "on" state was studied. The reduction of iron is accompanied by a proposed confirmation change of the side chains of the pocket residues and a water network rearragement of the active site. The water rearrangement within the active may indicate that the certain titratable residues may change their protonation states during the switch from "off" to "on" states of DosS protein. In addition, the protonation states of histidines do change when the protein is in the "on" state.

The binding of the selected inhibitor compounds with various models was also investigated. While the docking approach gave initial structures for use in further simulations, the scores were not reliable enough to rank the compounds based on their affinities. Next, different methods for modeling the interaction with the iron center were tested. The bonded model was the most reliable method, and while it does not allow for ABFE to estimate binding affinities, RBFE methods can be used in this setting.

BIBLIOGRAPHY

[1] Zheng, H. and Abramovitch, R. B. (2020). Inhibiting DosRST as a new approach to tuberculosis therapy.

[2] Cho, H. Y., Cho, H. J., Kim, Y. M., Oh, J. I., and Kang, B. S. (2009). Structural insight into the Heme-based redox sensing by DosS from Mycobacterium tuberculosis. *Journal of Biological Chemistry*, 284(19):13057–13067.

[3] Kumar, A., Toledo, J. C., Patel, R. P., Lancaster, J. R., and Steyn, A. J. (2007). Mycobacterium tuberculosis DosS is a redox sensor and DosT is a hypoxia sensor. *Proceedings of the National Academy of Sciences of the United States of America*, 104(28):11568–11573.

[4] Podust, L. M., Ioanoviciu, A., and Ortiz De Montellano, P. R. (2008). 2.3 ÅX-ray structure of the heme-bound GAF domain of sensory histidine kinase DosT of Mycobacterium tuberculosis. *Biochemistry*, 47(47):12523–12531.

[5] Zheng, H., Colvin, C. J., Johnson, B. K., Kirchhoff, P. D., Wilson, M., Jorgensen-Muga, K., Larsen, S. D., and Abramovitch, R. B. (2017). Inhibitors of Mycobacterium tuberculosis DosRST signaling and persistence. *Nature Chemical Biology*, 13(2):218–225.

[6] Basudhar, D., Madrona, Y., Kandel, S., Lampe, J. N., Nishida, C. R., and Montellano, P. R. O. D. (2015). Analysis of cytochrome p450 cyp119 ligand-dependent conformational dynamics by two-dimensional nmr and x-ray crystallography. *Journal of Biological Chemistry*, 290:10000–10017.

[7] ULC, C. C. G. (2022). Molecular operating environment (moe).

[8] Li, P. and Merz, K. M. (2016). Mcpb.py: A python based metal center parameter builder. *Journal of Chemical Information and Modeling*, 56:599–604.

[9] Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P., and Dupradeau, F.-Y. (2011). 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 Research*, 39:W511–W517.

[10] Seminario, J. M. (1996). Calculation of intramolecular force fields from second-derivative tensors. *International Journal of Quantum Chemistry*, 60.

[11] Mongan, J., Case, D. A., and McCammon, J. A. (2004). Constant ph molecular dynamics in generalized born implicit solvent. *Journal of computational chemistry*, 25:2038–2048.

[12] Nguyen, H., Roe, D. R., and Simmerling, C. (2013). Improved generalized born solvent model parameters for protein simulations. *Journal of chemical theory and computation*, 9:2020.

[13] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25:1157–1174.

[14] Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., and Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79:926–935.

[15] Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., and Simmerling, C. (2015). ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of Chemical Theory and Computation*, 11:3696–3713.

[16] Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. am1-bcc model: Ii. parameterization and validation. *Journal of Computational Chemistry*, 23:1623–1641.

[17] York, D. and Case, P. K. . D. (2020). Amber 2020.

[18] Roe, D. R. and Cheatham, T. E. (2013). Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9:3084–3095.

APPENDIX A

SUPPORTING TABLES

Table S8.1 The docking scores and EC50s for the compounds of interest.

Compound	Docking Score	EC_{50}
MSU43423	-2.27	6.16
MSU43424	-2.71	0.21
MSU43427	-2.42	1.79
MSU43432	-3.13	0.42
MSU43452	-1.83	0.74
MSU43573	-0.20	0.13
MSU43574	-1.62	0.13
MSU43577	-2.77	4.37
MSU43578	-3.01	4.02
MSU43579	-1.87	0.13
MSU43580	-1.00	0.13
MSU43582	-0.86	0.12
MSU43612	-2.11	0.73
MSU43613	-2.75	2.03
MSU43614	-2.15	0.48
MSU43616	-3.00	0.58
MSU43617	-2.35	0.093
MSU43618	-2.43	0.093
MSU43619	-3.03	0.09
MSU43620	-3.05	0.09
MSU43672	-2.75	0.25
MSU39451	-1.95	4.10
MSU39452	-0.68	2.47
MSU39453	0.08	16.60
MSU39449	-0.61	5.20
MSU33189	-3.02	0.63
MSU39447	-2.40	0.61
MSU39446	-2.85	0.54
MSU39445	-2.35	0.75
MSU41443	-0.75	11.20
MSU41442	1.56	2.08
MSU41462	-1.67	1.12
MSU42004	-1.96	2.36
MSU41545	-0.70	4.75
MSU41542	-3.06	1.34
MSU41546	-1.24	2.15
MSU42002	0.12	1.21
MSU39444	-3.25	1.70

APPENDIX B

SUPPORTING FIGURES

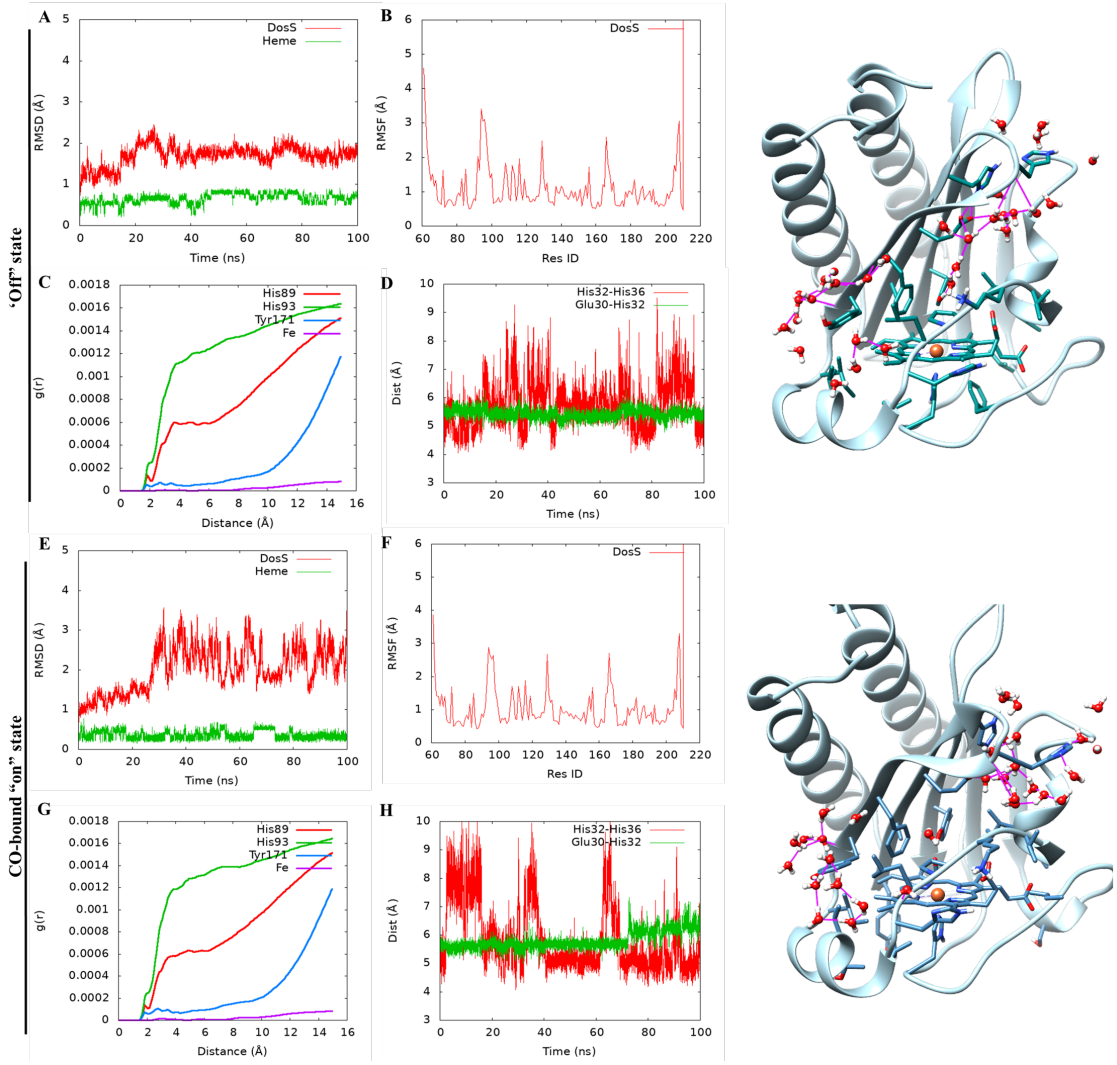


Figure S8.1 "Off" state and CO-bound "on" state modeling for DosS protein. A, E: RMSD time-series of the DosS protein, and heme and iron. B, F: Per-residue RMSF plot. C, G: Calculated water density around the given residues. D, H: Distances between Glu87-His89 and His89-His93 residues.

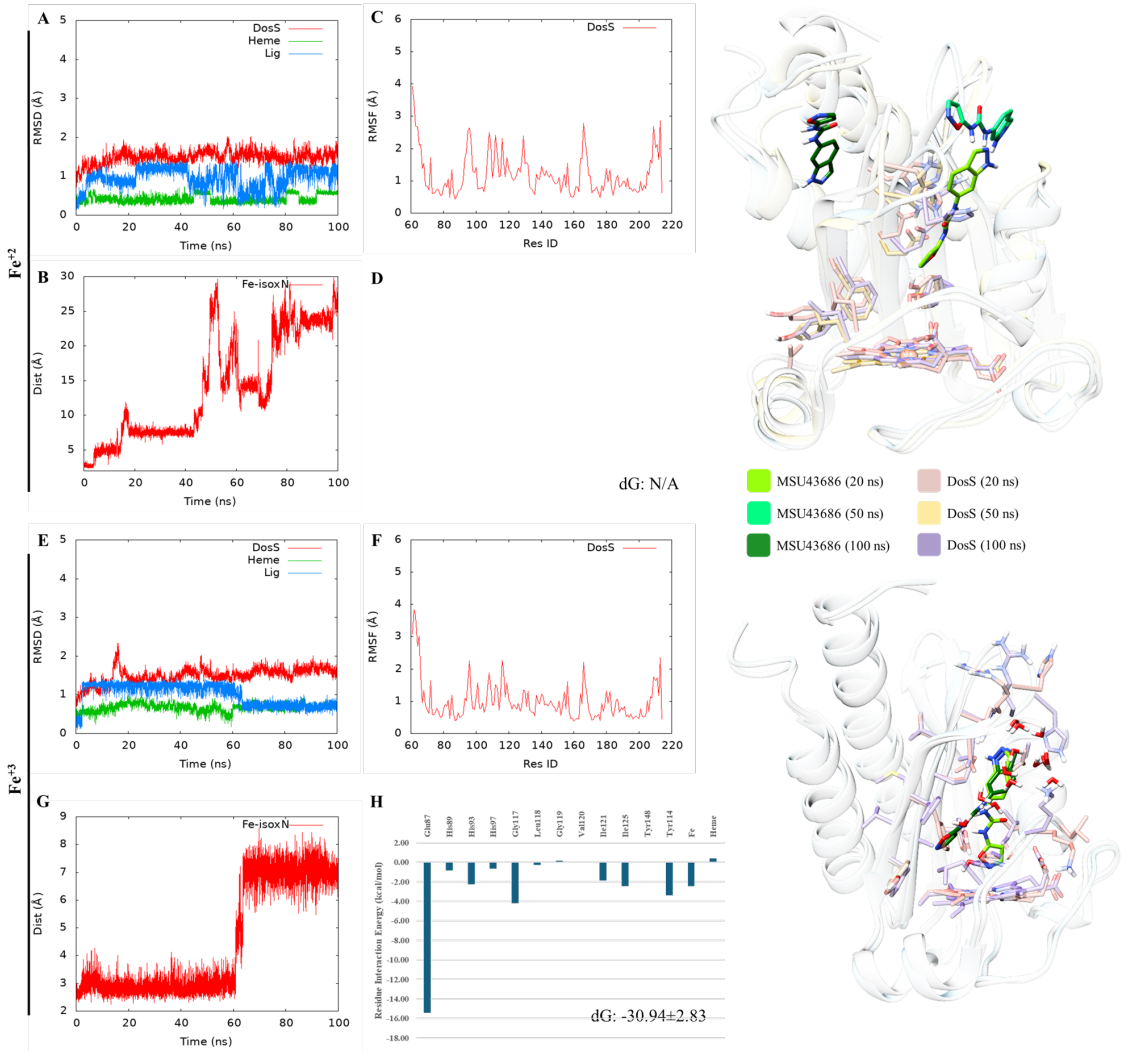


Figure S8.2 Fe^{+2} and Fe^{+3} unbound modeling for MSU-43686 molecule. A, E: RMSD time-series of the DosS protein, heme and iron, and inhibitor molecule. B, F: Per-residue RMSF plot. C, G: The distance between the iron and nitrogen from isoxazole. D, H: Per-residue interaction energies, this was not provided for Fe^{+2} simulations as the compound left the pocket. The MM-GBSA binding energies are also given as dG values for Fe^{+3}.

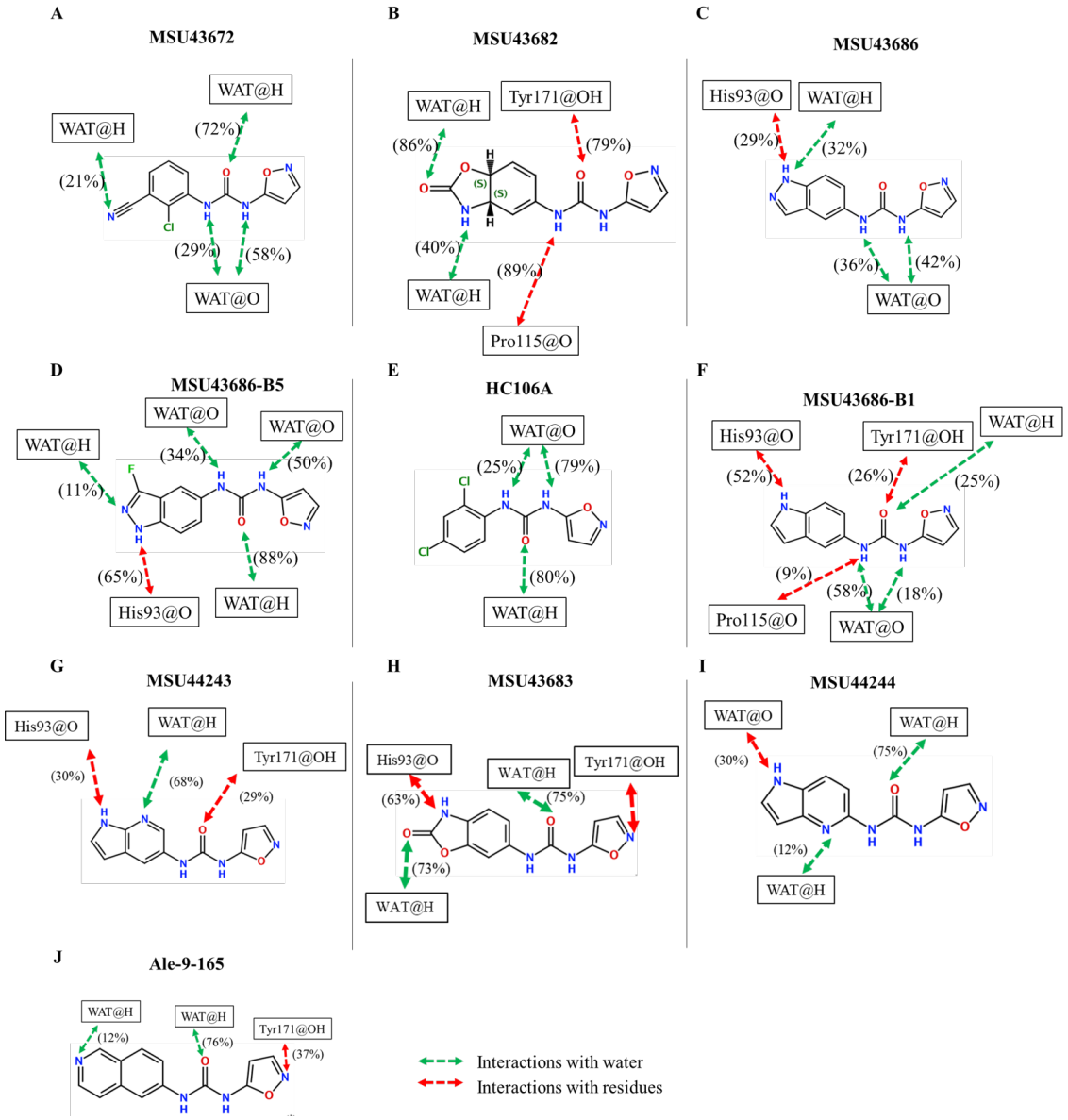


Figure S8.3 Direct interactions observed during the simulations of highlighted compounds. (A) MSU-43672, (B) MSU-43682, (C) MSU-43686, (D) MSU-43686-B5 derivative, (E) HC106A, (F) MSU-43686-B1 derivative, (G) MSU-43243, (H) MSU-43683, (I) MSU-44244, (J) Ale-9-165.

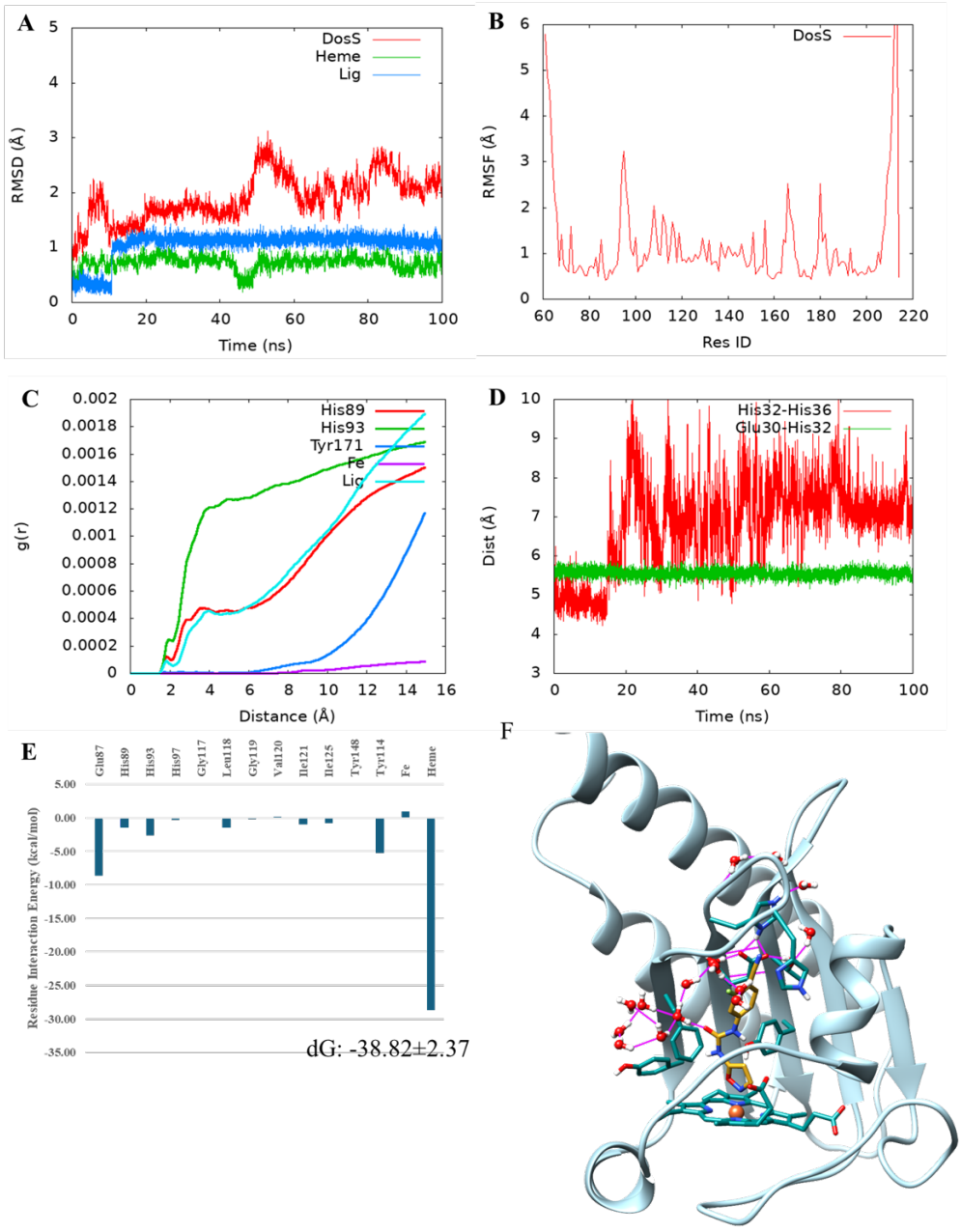


Figure S8.4 Results for MSU-43674 bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and MSU-43674. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the MSU-43674 simulations. The interactions between the water around the pocket residues are shown with pink lines.

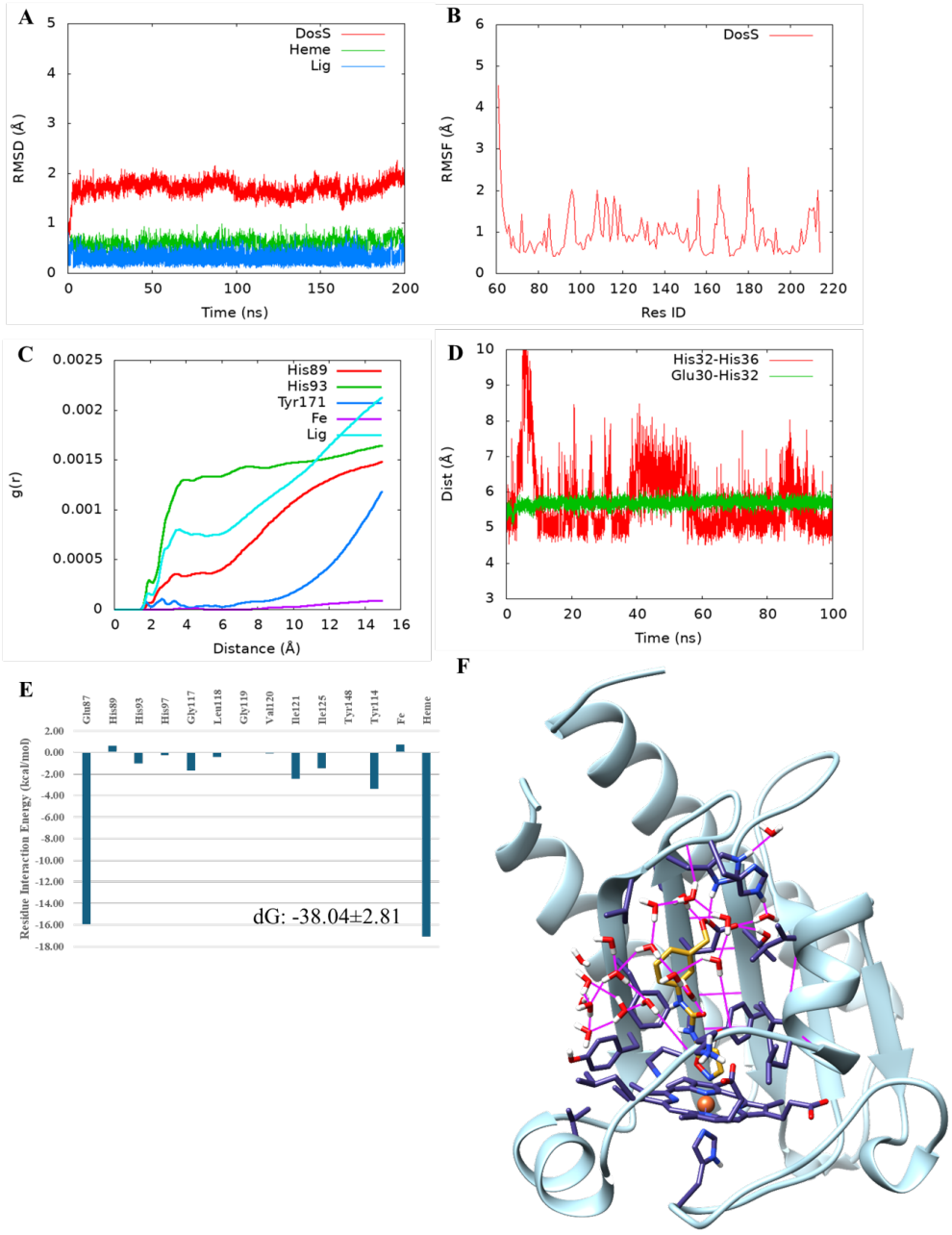


Figure S8.5 Results for MSU-43419 bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and MSU-43419. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the MSU-43419 simulations. The interactions between the water around the pocket residues are shown with pink lines.

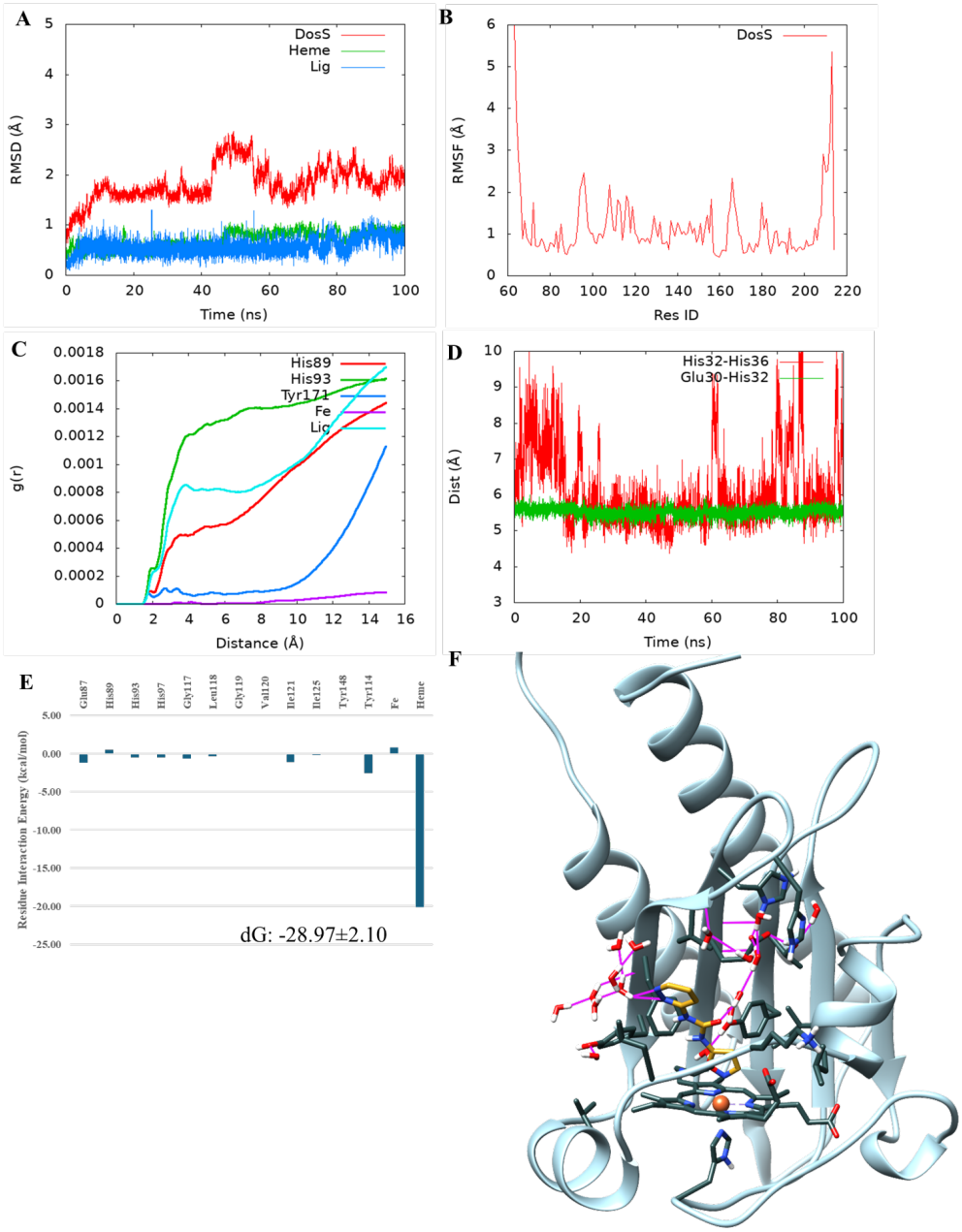


Figure S8.6 Results for MSU-44245 bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and MSU-44245. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the MSU-44245 simulations. The interactions between the water around the pocket residues are shown with pink lines.

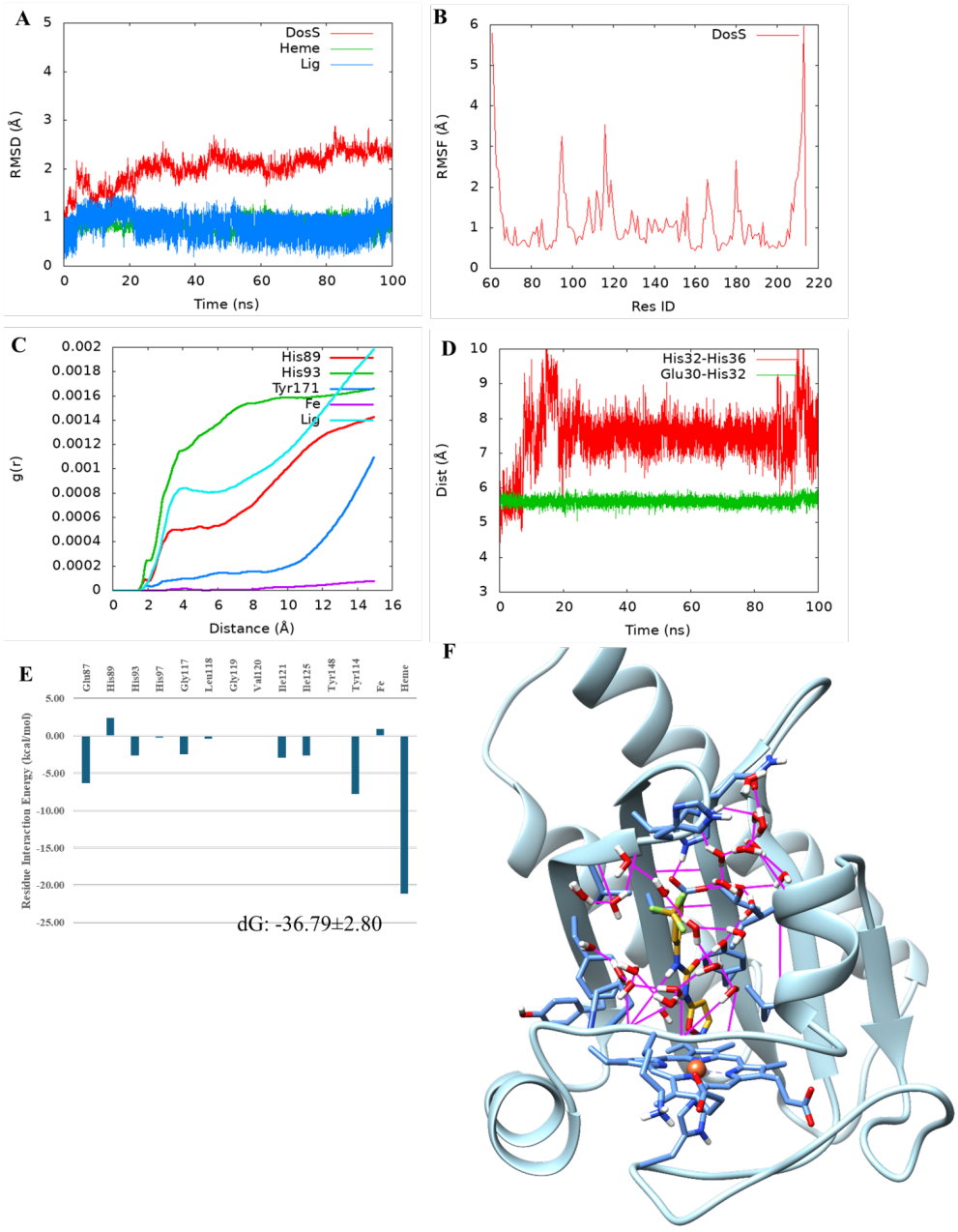


Figure S8.7 Results for MSU-43859 bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and MSU-43859. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the MSU-43859 simulations. The interactions between the water around the pocket residues are shown with pink lines.

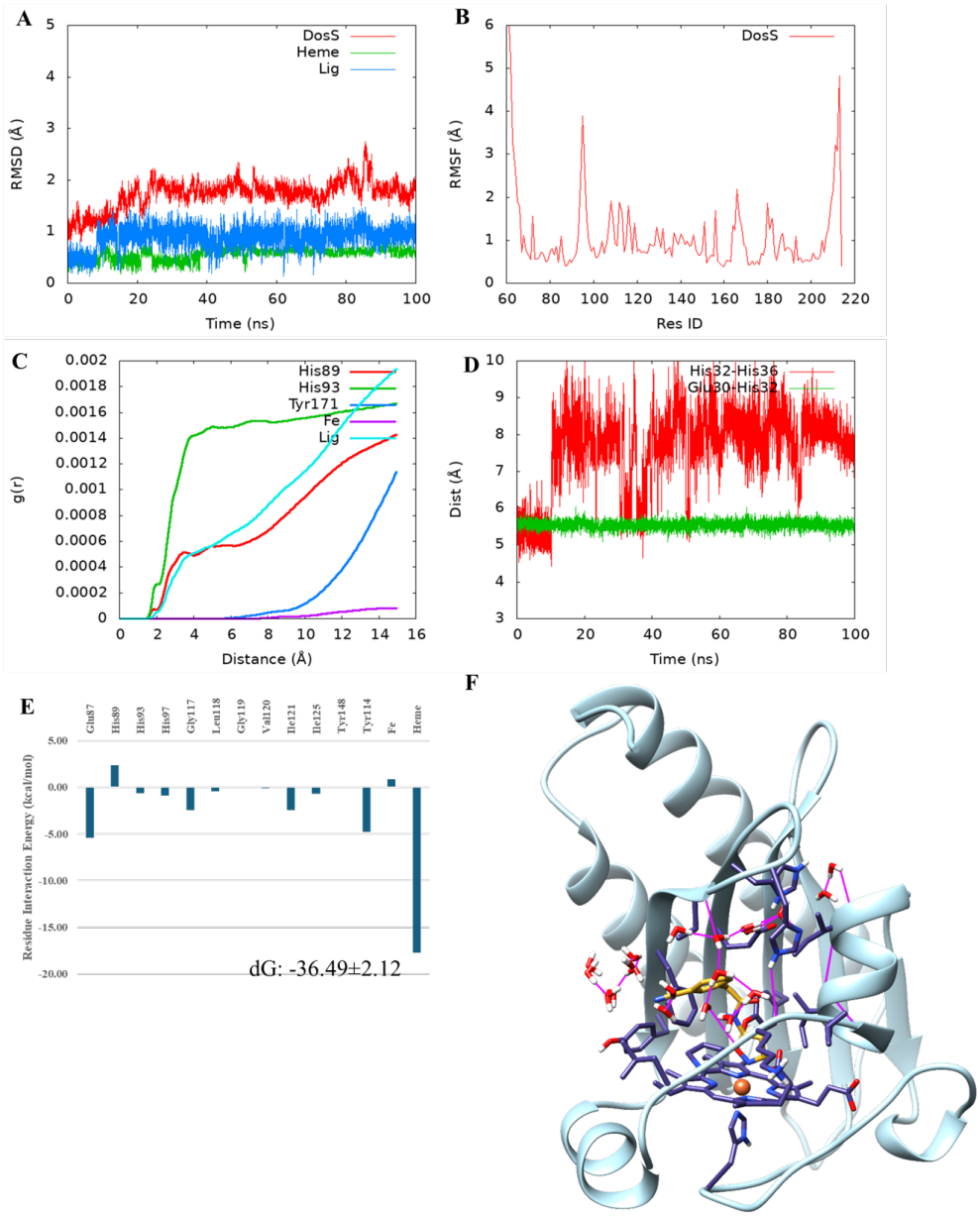


Figure S8.8 Results for MSU-44239 bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and MSU-44239. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the MSU-44239 simulations. The interactions between the water around the pocket residues are shown with pink lines.

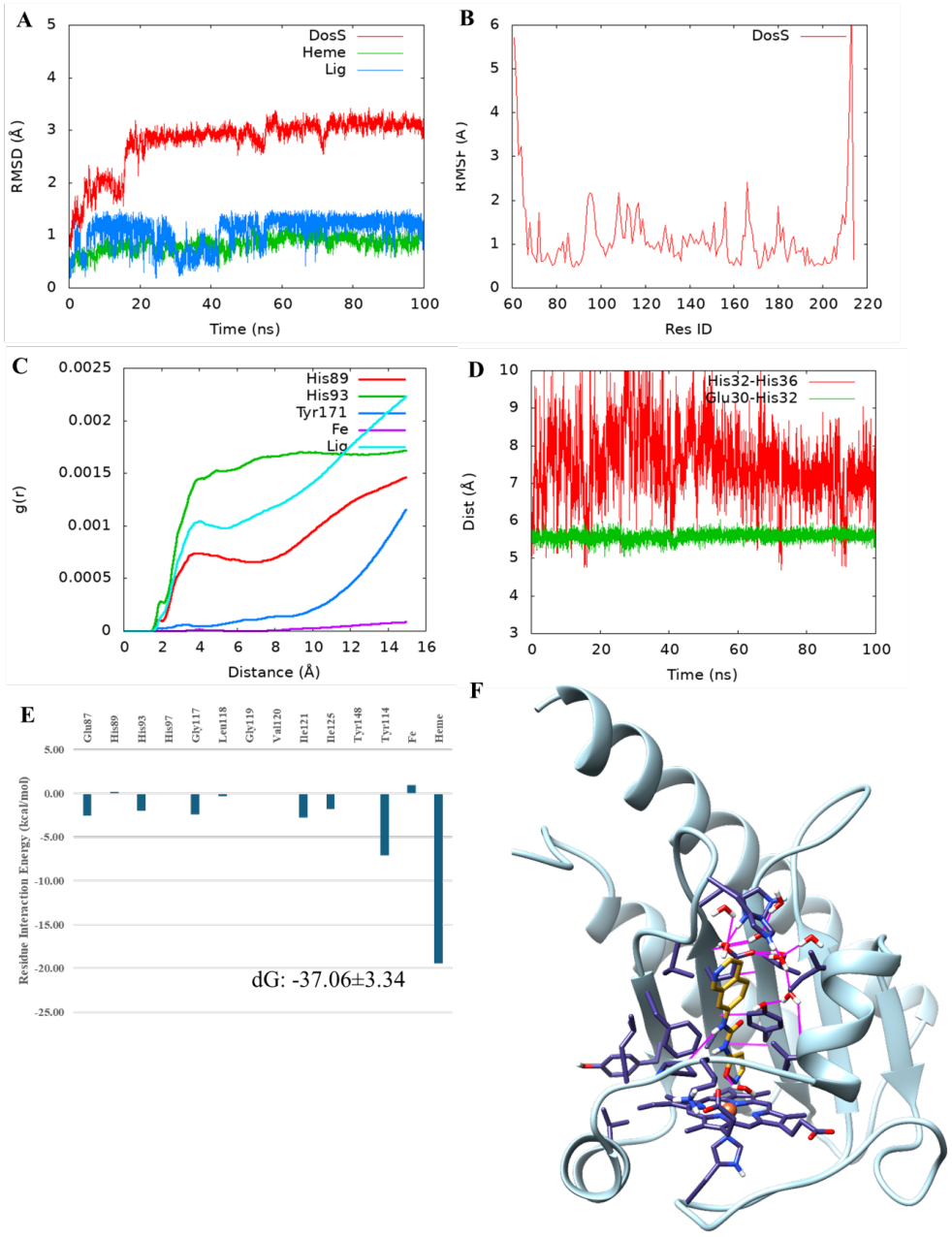


Figure S8.9 Results for MSU-44247 bonded-Fe^{+3} state modeling for DosS protein. A : RMSD time-series of the DosS protein, heme and iron, and MSU-44247. B: Per-residue RMSF plot. C: Calculated water density around the given residues. D: Distances between Glu87-His89 and His89-His93 residues. E. Per-residue interaction energies. F. The most populated cluster from the MSU-44247 simulations. The interactions between the water around the pocket residues are shown with pink lines.

CHAPTER 8

INVESTIGATION OF HOST-GUEST BINDING AFFINITIES WITH GEOMETRIC AND END-POINT BINDING FREE ENERGY CALCULATIONS

8.1 Introduction

Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) is a blind challenge that provides computational researchers the opportunity to gauge and improve computational methods that are can be applied to drug discovery.[1–5] The existing and newly developed strategies to investigate properties such as solvation free energies, binding affinities, pKa values, and other physicochemical properties can be assessed for a series of given compounds by participants, then the predicted values are compared with the experimentally determined values to assess the performance of each approach. One challenge that is assessed by the SAMPL9 competition is the estimation of the binding affinities between a small molecule (guest) and its target molecule (host).[6,7,SAM] Binding affinity prediction is one of the cornerstones of computer-aided drug discovery (CADD)[9,10]. Being able to predict the binding energies accurately for a given target helps to reduce the number of compounds to be investigated in a wet-lab setting and hence can help to reduce the cost and the timeframe of a drug discovery program. Therefore, it is critical to assess the reliability of various approaches and schemes used for binding affinities.

The SAMPL9 challenge includes two host and five guest molecules, as are shown in Figure 1[SAM]. Host molecules are β-cyclodextrin (bCD) and Hexakis-2,6-dimethyl-β-cyclodextrin (HbCD) compounds that are heptasaccharides made of glucose connected by a 1-4 glycosidic bond[6]. Cyclodextrins are used as a reactant involved in inclusion complexes, assessing the diffusion and single-molecule interactions with certain proteins, and improving the solubility and stability of drugs. The five guest molecules that are given in this challenge share the same phenothiazine core with substitutions and a cationic arm (Fig. 1(B)). Among the guest molecules, Promethazine hydrochloride (PMT) is a drug used to treat nausea and vomiting, Thioridazine hydrochloride (TDZ), Chlorpromazine hydrochloride (CPZ), and Trifluoperazine dihydrochloride (TFP) are used for certain mood disorders, Promazine hydrochloride (PMZ) is used as a tranquillizer used in veterinary medicine.[11–15] The goal of the SAMPL9 host-guest challenge is to predict the binding affinities of these host-guest molecules. There are numerous computational methods with varying levels of complexity and cost that can be used to estimate binding affinities: docking methods,

alchemical methods such as free-energy perturbation and thermodynamic integration, geometric methods including steered molecular dynamics (SMD) and umbrella sampling, and end-state methods. The majority of these methods rely on atomistic simulations such as equilibrium or non-equilibrium MD.

In this work, a number of schemes involving end-state methods are investigated: Molecular Mechanics – Gibbs Born Surface Area/Poisson Boltzmann Surface Area (MM-GBSA/PBSA). Furthermore, an SMD approach was also considered with two different pulling speeds: 5 Å/ns and 10 Å/ns, with a positional restraint only on the host molecules along with a cylindrical restraint to keep the guest molecules on the axis of the collective variable. Approaches tested here aimed to answer the following questions using the host-guest dataset:

- How does the frame sampling impact the performance of end-state methods, namely MM-GBSA and MM-PBSA?
- How does the orientation of the guest molecules affect the performance of MM-GBSA/PBSA methods?
- Can a simple alternative SMD protocol be created to capture the binding affinities?

The end-state methods used here have been the method of choice of many drug discovery programs to perform a quick compound screen, as they have been shown to perform well in ranking the compounds based on their affinities.[16] However, as the success of MM-GBSA/PBSA methods are heavily dependent on the frame selection from the MD trajectories, it is important to test and assess the accuracies of the frame sampling schemes. Furthermore, as an extension of that, the selection of the correct binding pose is another significant criterion that changes the performance of the end-state methods for affinity predictions.

Geometric methods used in free energy estimations are based on creating a potential-of-mean force (PMF) to calculate the strength of interaction between two molecules.[17,18]These methods require a selection of a collective variable (can be distance, angle, dihedral angle) that is used to perturb the system, and there are well-established schemes such as pAPRika that can be used.[17,18]

Here, the aim was to create an SMD scheme with the lowest number of restraints possible on the host molecules to estimate the binding affinities.

8.2 Computational Details

8.2.1 Preparation of systems

The host and guest molecules were obtained from the SAMPL9 GitHub page (SAMPL9 CD dataset), and their 2D structures are shown in Figure 1(A,B).[SAM] The protonation states of host and guest molecules were determined by Protonate3D module as implemented in Molecular Operating Environment version 2019.01 (MOE).[19] The partial charges of the molecules were calculated using the RESP method with REDSERVER.[20]**?**

8.2.2 Docking Scheme

The host-guest complexes were obtained with the docking using the MOE software.[19] The docking procedure involves two steps: placement and refinement. Triangle matcher was used to place the guest molecules in the host cavities and scores were calculated using the London dG method. Then, the top 100 poses were selected and refined with induced fit approach and the final docking scores were calculated with Generalized-Born volume integral/weighted surface area score (GBVI/WSA dG).[21] The selected poses were minimized with molecular mechanics using AMBER10: Extended Hückel Theory (EHT) force field implemented in MOE. For each host-guest complex, along with the highest scoring poses, a pose that orients differently was also selected to investigate the impact of the guests' orientation with the within the host molecules.

8.2.3 Classical Molecular Dynamics Procedure

Each selected pose was prepared for MD simulations using tleap module of Amber18.[22] The host and guest molecules were modeled using the gaff2 force field, and TIP4P-EW was used as water model to solvate the complexes in a 14Å cubic box.[23,24] Required counter ions, Na+ and Cl-, were added to neutralize the systems. 1fs timestep was applied during the heating and the production steps. Langevin thermostat and isotropic position scaling was selected for the thermostat and barostat, respectively.[25–27] Each system was minimized in four steps with decreasing positional restraints on the solute molecules (100, 50, 10, 0 (kcal mol^{-1} $Å^{-2}$). Then, the systems were heated

up to 300 K in a stepwise fashion in 1.51 ns, as described in previous papers.[28–31] The production runs were 20 ns long at 300K, 1atm pressure in triplicates. Nonbonded interactions were truncated with a 10.0 Å cutoff value. Particle-mesh Ewald was used for long-range electrostatic interactions. The pmemd.cuda module was used to perform the simulations, as implemented in Amber 18.[22]

8.2.4 Binding Energy Estimations

The binding energies for the host-guest complexes were calculated using end-state methods, i.e. MM-GBSA/PBSA.[32] Three different frame sampling schemes were tested: (i) sampling 500 frames from first 5 ns, (ii) sampling 500 frames from last 5 ns, and (iii) sampling 1000 frames from the 20 ns trajectory. These calculations were performed for the duplicate simulations and the values are averaged.

8.2.5 Steered Molecular Dynamics Procedure

Amber18 along with Plumed was used for the Steered Molecular Dynamics (SMD) calculations.[33–35] The final frames for each set of simulations were used as a starting point for SMD, and the host-guest molecules were re-solvated in 30 Å TIP4P-EW water box. The pulling direction was set to be on the z-axis and the guest molecules were oriented accordingly. The pulling direction is determined by the orientation of the cationic side chains of guest molecules. During the SMD, a small positional restraint (5 kcal mol^{-1} $Å^{-2}$) was applied to the host molecules to keep them at the center of the box.

The collective variable is defined as the distance between the center-of-mass of sulfur and nitrogen atoms of phenothiazine core (COM-guest) for each guest molecule and a stationary point within the host molecule that is positioned initially 2 Å away from the COM-guest. A wall constraint was added to the mean distance between the COM-guest and the center-of-mass of the oxygen atoms from glycosidic bond (COM-host) (Figure S2). A time-dependent harmonic restraint with force constant of 10 was applied to the collective variable on the z axis to move the guest molecule 20 Å away from the host with the speed of 10 Å/ns for 2 ns followed by equilibrium at 20 Å for 600 ps and 5 Å/ns for 3.6 ns followed by 1.2 ns equilibrium. The calculations were performed independently four times for each replica of a pose, and the final energies were obtained by averaging based on

Table S8.1 The estimated binding energies and statistics for primary selected poses for bCD and listed guest molecules are shown. RMSE: root mean square errors, MAE: mean absolute errors, ME: mean errors, r^2: correlation coefficient, m: slope of the correlation plots, τ : Kendall's Tau rank correlation coefficient.

	Pose	dG (exp, kcal/mol)	MM-GBSA (avg, kcal/mol)	MM-PBSA (avg, kcal/mol)	MM-GBSA (first 5ns, kcal/mol)	MM-PBSA (first 5ns, kcal/mol)	MM-GBSA (last 5ns, kcal/mol)	MM-PBSA (last 5ns, kcal/mol)	SMD (10Å/ns, avg, kcal/mol)
bCD	Tdz-p2	-5.70	-23.79±2.37	-20.79±2.47	-24.07±2.12	-20.98±2.24	-23.60±2.60	-20.73±2.58	-7.99±1.52
	Tfp-p1	-5.06	-23.42±3.98	-18.07±3.70	-23.12±3.62	-18.07±2.80	-25.70±2.88	-20.15±2.70	-15.25±1.52
	Pmz-p2	-4.97	-20.76±2.68	-18.40±2.71	-20.97±2.80	-18.19±2.90	-19.95±2.59	-17.95±2.46	-14.79±0.88
	Pmt-p1	-4.48	-19.93±2.55	-16.80±2.49	-19.07±1.95	-16.60±2.00	-19.77±2.45	-16.18±2.36	-11.46±1.03
	Cpz-p1	-5.42	-23.22±2.84	-20.52±3.10	-23.25±2.84	-20.39±3.18	-23.94±2.27	-21.45±2.37	-16.79±1.00
Statistics	RMSE	--	17.18	13.84	17.04	13.78	17.60	14.26	8.76
	MAE	--	17.13	13.79	16.97	13.72	17.47	14.16	8.13
	ME	--	17.13	13.79	16.97	13.72	17.47	14.16	8.13
	r^2	--	0.77	0.93	0.87	0.96	0.38	0.80	0.03
	m	--	3.36	3.52	4.09	3.79	3.48	4.18	-1.39
	τ	--	0.79	0.79	0.99	0.79	0.39	0.79	0.19

the Jarzynski equality.

8.2.6 Analysis

The estimated binding affinities were compared with the experimental values and the root mean square errors (RMSE), mean absolute errors (MAE), mean errors (ME), correlation coefficient (r^2), slope of the correlation plots (m), and Kendall's Tau rank correlation coefficient (τ) were calculated by bootstrapping with replacement. The statistical analysis was performed for bCD and HbCD systems separately and together as well to assess the performance of methods with respect to each host. The root mean square deviations (RMSD), root mean square fluctuations (RMSF), hydrogen bonds (HBOND), and contact maps were obtained using cpptraj module of Amber18.[22,36]

Table S8.2 The estimated binding energies and statistics for primary selected poses for HbCD and listed guest molecules are shown. RMSE: root mean square errors, MAE: mean absolute errors, ME: mean errors, r^2: correlation coefficient, m: slope of the correlation plots, τ : Kendall's Tau rank correlation coefficient.

	Pose	dG (exp, kcal/mol)	MM-GBSA (avg, kcal/mol)	MM-PBSA (avg, kcal/mol)	MM-GBSA (first 5ns, kcal/mol)	MM-PBSA (first 5ns, kcal/mol)	MM-GBSA (last 5ns, kcal/mol)	MM-PBSA (last 5ns, kcal/mol)	SMD (10Å/ns, avg, kcal/mol)
HbCD	Tdz-p2	-6.46	-26.32±2.38	-24.89±2.56	-26.44±2.32	-25.15±2.52	-26.75±2.32	-25.03±2.47	-17.72±1.52
	Tfp-p1	-5.54	-26.25±4.06	-21.14±3.75	-23.95±2.30	-19.18±2.51	-24.09±3.18	-19.09±2.82	-18.85±1.52
	Pmz-p1	-5.05	-23.54±2.86	-21.43±2.80	-22.71±2.76	-21.00±2.60	-23.38±2.82	-21.39±2.72	-13.77±1.22
	Pmt-p1	-5.36	-24.70±2.52	-21.81±2.48	-24.48±2.47	-21.38±2.60	-24.59±2.29	-21.90±2.48	-19.98±1.51
	Cpz-p1	-5.40	-25.30±3.54	-21.81±3.26	-25.84±2.74	-22.49±2.86	-24.74±3.42	-21.05±3.26	-13.84±1.52
Statistics	RMSE	--	19.68	16.68	19.15	16.37	19.16	16.21	11.54
	MAE	--	19.66	16.66	19.13	16.28	19.15	16.13	11.27
	ME	--	19.66	16.66	19.13	16.28	19.15	16.13	11.27
	r^2	--	0.60	0.85	0.60	0.55	0.90	0.52	0.12
	m	--	1.68	2.62	2.16	3.05	2.24	2.89	1.86
	τ	--	1.00	0.32	0.60	0.40	0.60	0.00	0.20

8.3 Results and Discussion

8.3.1 Host-guest docking poses

The structures of the investigated host and guest molecules from SAMPL8 CD challenge are provided in Figure 1 (A) and (B), respectively. The alcohol groups in bCD are modified to methoxy in HbCD host, and an example of these modifications is highlighted with red and blue circles in Fig. 1(A). These modifications change the charge distribution and hydrogen bonding ability of the host molecules, as shown in Fig. 1 (C) and (D) and impact the orientation as well as interactions with the guest molecules. The evaluated docking poses highlight that there are preferred orientations for each host molecule. The primary poses (dominant orientation obtained from docking) in both bCD and HbCD docking, the bulky cationic side chains of guest molecules reside at the secondary face of host molecules, as can be seen in Figure 2. On the other hand, the phenothiazine core substituent did not have a specific preference. In bCD: TDZ and TFP have phenothiazine core substituent that orients towards primary face, and CPZ has the substituent pointing towards the secondary face (Fig. 2(A)). In HbCD, however, the core substituents in TDZ and TFP are oriented towards the secondary face and the CPZ substituent occupies the primary face (Fig. 2(B)). The secondary poses observed

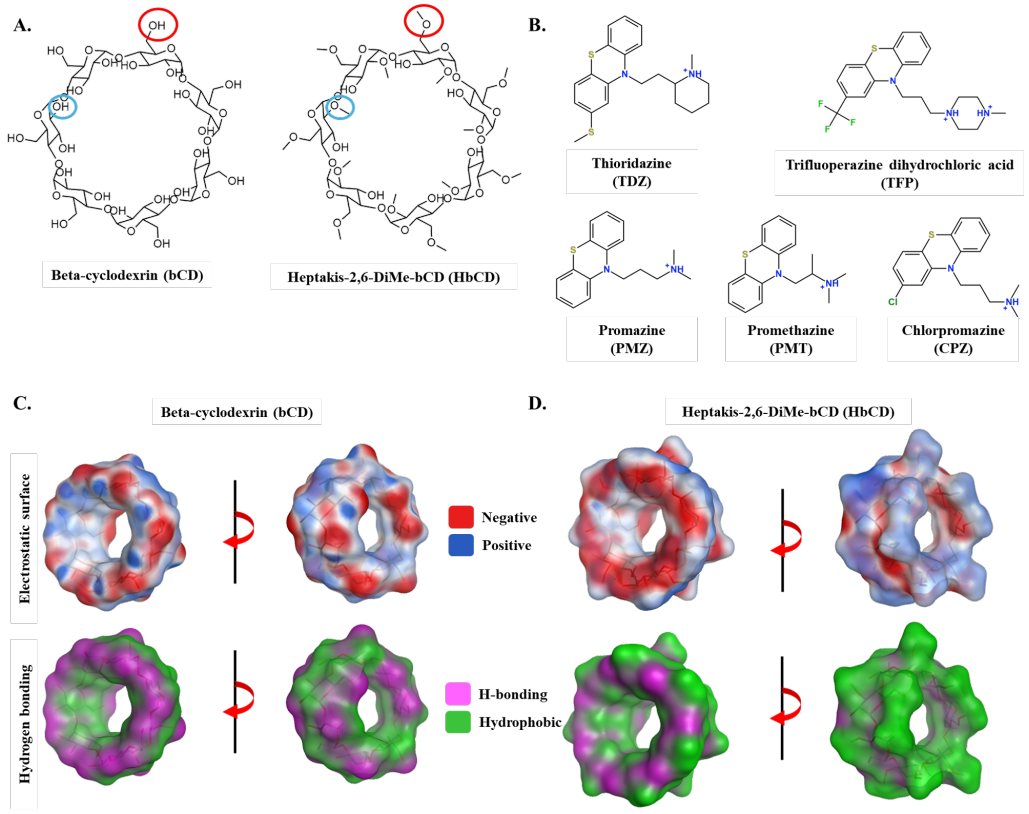


Figure S8.1 2D structures of (A) host and (B) guest molecules for SAMPL9 challenge. The electrostatic and hydrophobic surfaces of host molecules are shown in (C) and (D) for bCD and HbCD, respectively.

for each system as shown in Figure S1 shows that the cationic side chains still prefer the secondary face of the host molecules. The charge distribution and the hydrophobicity of each face in the host molecules can be attributed for this preference: the dominant negative charge distribution of the secondary face can lead the cationic side chain of guest molecules to orient towards this face (Fig. 1(A, B)).

8.3.2 Orientations of guest molecules

The selected orientations of the guest molecules bound to bCD and HbCD are depicted in Figure S1. In general, within the primary poses, a preference is observed: for both bCD-bound and HbCD-bound molecules, their cationic side chains are oriented towards the secondary face. This orientation towards the side chains also occurs for the secondary poses as well. On the other hand, as mentioned in the previous section, the core substituents show different orientation preferences

between the bCD and HbCD hosts.

As PMZ and PMT do not have a substituent on the phenothiazine core, their orientations are mainly based on the cationic side chains – and in all cases, PMZ and PMT poses face the secondary face consistently. bCD-PMZ and bCD-PMT also have very similar binding affinities, except for the initial 5 ns of the PMZ simulations. HbCD-PMZ poses on average have very similar MM-GBSA/PBSA energies, however, at the beginning and end of the simulations, the second pose of PMZ (HbCD-PMZ-p2) has a binding energy 2 kcal mol^{-1} stronger. HbCD-PMT poses, on the other hand, have very similar MM-GBSA/PBSA energies regardless of the frame sampling. The primary poses obtained for TFP have a single binding conformation in both bCD and HbCD with a difference in orientation for their phenothiazine core substituent (Fig. S1). For SP vs SS orientations obtained for bCD-TFP-p1 and HbCD-TFP-p1, respectively, their average MM-GBSA/PBSA values align with the experimental binding affinities. Another guest molecule with a different orientation is for the phenothiazine core substituent is the primary poses of CPZ. The experimental binding affinities of CPZ with bCD and HbCD are very similar (-5.42 vs -5.40 kcal mol^{-1}. The MM-GBSA/PBSA results also follow a similar order regardless of the frame sampling. The observations from TFP and CPZ compounds and their respective binding orientations with bCD and HbCD host molecules leads to the conclusion that the preferred orientation of the phenothiazine core substituent is different in each host molecule, and also depends on the substituent type. The phenothiazine core substituent in bCD-CPZ allows more hydrogen bonds to be made with the secondary face of bCD, and in HbCD-CPZ system it allows -Cl to be positioned away from the highly charged HbCD secondary face. All in all, the orientation preferences for guest molecules depend on the substituent changes on the host as well.

8.3.3 Binding energy estimations with end-state methods

The selected poses, primary poses and an alternative pose for each host-guest system, were prepared and minimized according to the protocol outlined in Methods section, and 20 ns long classical MD simulations were run in duplicates for each pose. The binding affinities of each pose were calculated for both simulations and averaged. The final results are listed in Table 1 and Table

2 for bCD and HbCD systems, respectively. For both MM-GBSA and MM-PBSA binding energy calculations, a number of frame samplings were tested: frames sampled from first 5 ns and last 5 ns of the simulations, or frames sampled throughout the entirety of the simulations. In addition, the SMD approach was considered with two different pulling speeds: 5 Å/ns and 10 Å/ns.

The estimated binding affinities listed in Table 1 and their statistical analysis show that all MM-GBSA estimations have high RMSE ($\tilde{1}$7 kcal mol^{-1}), and MM-PBSA results RMSE values were calculated to be $\tilde{1}$3 kcal mol^{-1}. These observations do not change significantly upon changing the frame sampling (avg. vs first 5ns vs last 5ns in Table 1). However, the predicted binding energies are still able to capture the ranking and correlate with the experimental affinities well, as is demonstrated from the Kendall's tau and r^2 values, respectively. Specifically, the approach with the lowest RMSE, MM-PBSA (first 5ns), has a 0.96 r^2 and a Kendall's tau score of 0.79. This indicates that although the exact binding affinities were not predicted, the ranking of them can be calculated with this approach. Another observation from Table 1 is that the selection of frames does not seem to impact the statistics significantly, with the exception of MM-GBSA (last 5 ns). The results obtained for HbCD host are shown in Table 2. The RMSE values are approximately 2 kcal mol^{-1} higher than the bCD results, highlighting the fact that predictions for HbCD are slightly more difficult for the end-state approach used here. Considering the RMSE results, the lowest RMSE was for MM-PBSA sampled from the last 5 ns. On the other hand, the highest r^2 value was observed for MM-GBSA sampled from the last 5 ns, with Kendall's tau of 0.6. The best ranking of the binding affinities however was observed for MM-GBSA energies averaged from 20 ns of the simulations. The cumulative statistics of the results (Table 3) shows that:

- MM-PBSA energies have 2-3 kcal mol^{-1} lower RMSE values than the MM-GBSA energies.
- The best r^2 value is obtained with MM-PBSA method with frames sampled from the 20 ns-long trajectories. This is followed by the value obtained with the MM-PBSA method with frames sampled from the initial 5 ns of the simulations.
- Both of these schemes have comparable Kendall's tau values: 0.45 and 0.47, respectively

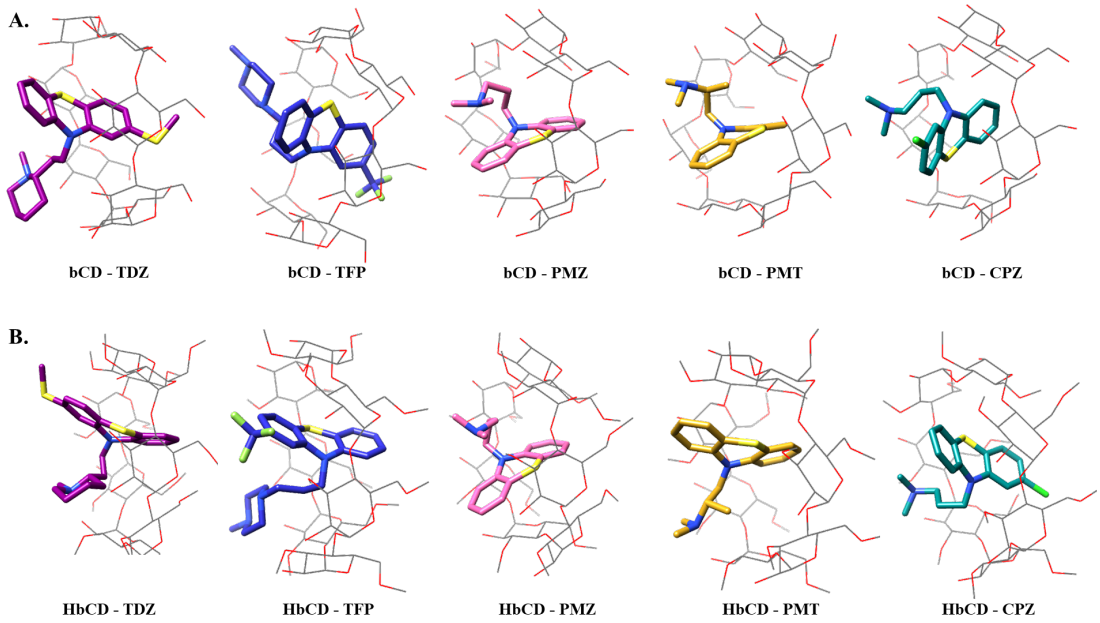


Figure S8.2 The primary poses considered for (A) bCD and (B) HbCD molecules. Oreintations (arm/core) are as follows: bCD: SP; SP; S-; S-; SS. HbCD: SS; SS; S-; S-; SP.

Comparing these values with the statistics of the secondary poses (Table S3), it is clear that the pose selection has great significance. While the RMSE values did not improve or deteriorate with pose selection, the highest values obtained for r^2 (0.24) and Kendall's tau (0.38) values were quite low. Even when the lowest energy values for each binding energy calculation scheme was selected for each complex, as reported in Table S4, the correlation coefficient and tau values are still not able to compare with the results from the primary poses.

8.4 Conclusions

Based on the challenge results of the submitted computational protocols in SAMPL9 competition, MM-GBSA and MM-PBSA methods provide a high Kendall's tau (0.45-0.65) along with good r^2 estimations (0.78-0.65) with MAE values $\tilde{1}7$ kcal mol^{-1}, highlighting the fact that these end-state methods are successful in obtaining trends in the binding affinities, but not necessarily the exact binding energies. Results obtained in this study also showed that the success of these end-state methods is dependent on the choice of the initial poses and the sampling of the frames from MD simulations.

Table S8.3 Cumulative statistics of both bCD and HbCD results with primary poses are listed. RMSE: root mean square errors, MAE: mean absolute errors, ME: mean errors, r^2: correlation coefficient, m: slope of the correlation plots, τ : Kendall's Tau rank correlation coefficient.

	Analysis	MM-GBSA (avg, kcal/mol)	MM-PBSA (avg, kcal/mol)	MM-GBSA (first 5ns, kcal/mol)	MM-PBSA (first 5ns, kcal/mol)	MM-GBSA (last 5ns, kcal/mol)	MM-PBSA (last 5ns, kcal/mol)	SMD (10Å/ns, avg, kcal/mol)
Cumulative Statistics	**RMSE**	18.47	15.33	18.13	15.13	18.40	15.27	10.24
	MAE	18.40	15.23	18.05	15.00	18.31	15.15	9.70
	ME	18.40	15.23	18.05	15.00	18.31	15.15	9.70
	r^2	0.65	0.78	0.74	0.74	0.74	0.71	0.08
	m	3.25	3.88	3.56	4.04	3.17	3.84	1.93
	τ	0.64	0.45	0.73	0.47	0.42	0.42	0.29

BIBLIOGRAPHY

[1] Nicholls, A., Wlodek, S., and Grant, J. A. (2009). The samp1 solvation challenge: Further lessons regarding the pitfalls of parametrization†. *Journal of Physical Chemistry B*, 113:4521–4532.

[2] Geballe, M. T., Skillman, A. G., Nicholls, A., Guthrie, J. P., and Taylor, P. J. (2010). The sampl2 blind prediction challenge: Introduction and overview. *Journal of Computer-Aided Molecular Design*, 24:259–279.

[3] Yin, J., Henriksen, N. M., Slochower, D. R., Shirts, M. R., Chiu, M. W., Mobley, D. L., and Gilson, M. K. (2017). Overview of the sampl5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design*, 31:1–19.

[4] Muddana, H. S., Fenley, A. T., Mobley, D. L., and Gilson, M. K. (2014). The sampl4 host-guest blind prediction challenge: An overview. *Journal of Computer-Aided Molecular Design*, 28:305–317.

[5] Muddana, H. S., Varnado, C. D., Bielawski, C. W., Urbach, A. R., Isaacs, L., Geballe, M. T., and Gilson, M. K. (2012). Blind prediction of host-guest binding affinities: A new sampl3 challenge. *Journal of Computer-Aided Molecular Design*, 26:475–487.

[6] Andrade, B., Chen, A., and Gilson, M. K. (2024). Host-guest systems for the sampl9 blinded prediction challenge: phenothiazine as a privileged scaffold for binding to cyclodextrins. *Physical chemistry chemical physics : PCCP*, 26:2035–2043.

[7] Amezcua, M., Setiadi, J., and Mobley, D. L. (2024). The sampl9 host–guest blind challenge: an overview of binding free energy predictive accuracy. *Physical Chemistry Chemical Physics*, 26:9207–9225.

[SAM] samplchallenges/sampl9: 0.8.

[9] Beveridge, D. L. and DiCapua, F. M. (1989). Free energy via molecular simulation: applications to chemical and biomolecular systems. *Annual review of biophysics and biophysical chemistry*, 18:431–492.

[10] DiMasi, J. A., Grabowski, H. G., and Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of r&d costs. *Journal of Health Economics*, 47:20–33.

[11] Kiningham, K. K. (2007). Promethazine. *xPharm: The Comprehensive Pharmacology Reference*, pages 1–6.

[12] Cheng, H. W., Liang, Y. H., Kuo, Y. L., Chuu, C. P., Lin, C. Y., Lee, M. H., Wu, A. T., Yeh, C. T., Chen, E. T., Whang-Peng, J., Su, C. L., and Huang, C. Y. (2015). Identification of thioridazine, an antipsychotic drug, as an antiglioblastoma and anticancer stem cell agent using

public gene expression data. *Cell Death & Disease 2015 6:5*, 6:e1753–e1753.

[13] Mann, S. K. and Marwaha, R. (2023). Chlorpromazine. *Encyclopedia of Toxicology: Third Edition*, pages 925–929.

[14] Koch, K., Mansi, K., Haynes, E., Adams, C. E., Sampson, S., and Furtado, V. A. (2014). Trifluoperazine versus placebo for schizophrenia. *The Cochrane Database of Systematic Reviews*, 2014.

[15] Sibilio, J. P., Andrew, G., Stehman, V. A., Dart, D., and Moore, K. B. (1957). Treatment of chronic schizophrenia with promazine hydrochloride. *A.M.A. Archives of Neurology & Psychiatry*, 78:419–424.

[16] Eken, Y., Almeida, N. M., Wang, C., and Wilson, A. K. (2021). Sampl7: Host–guest binding prediction by molecular dynamics and quantum mechanics. *Journal of Computer-Aided Molecular Design*, 35:63–77.

[17] Henriksen, N. M., Fenley, A. T., and Gilson, M. K. (2015). Computational calorimetry: High-precision calculation of host-guest binding thermodynamics. *Journal of Chemical Theory and Computation*, 11:4377–4394.

[18] Velez-Vega, C. and Gilson, M. K. (2013). Overcoming dissipation in the calculation of standard binding free energies by ligand extraction. *Journal of Computational Chemistry*, 34:2360–2371.

[19] (2022). Molecular operating environment (moe), 2022.02 chemical computing group ulc, 1010 sherbooke st. west, suite 910, montreal, qc, canada, h3a 2r7.

[20] Vanquelef, E., Simon, S., Marquant, G., Garcia, E., Klimerak, G., Delepine, J. C., Cieplak, P., and Dupradeau, F.-Y. (2011). 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 Research*, 39:W511–W517.

[21] Corbeil, C. R., Williams, C. I., and Labute, P. (2012). Variability in docking success rates due to dataset preparation. *Journal of Computer-Aided Molecular Design*, 26:775–786.

[22] York, D., Kollman, P., and et al, D. C. (20218). Amber 2018.

[23] He, X., Man, V. H., Yang, W., Lee, T.-S., and Wang, J. (2020). A fast and high-quality charge model for the next generation general amber force field. *The Journal of Chemical Physics*, 153:114502.

[24] Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25:1157–1174.

[25] Woodcock, L. V. (1971). Isothermal molecular dynamics calculations for liquid salts. *Chemical Physics Letters*, 10:257–261.

[26] Verlet, L. (1967). Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159:98–103.

[27] Horn, H. W., Swope, W. C., and Pitera, J. W. (2005). Characterization of the tip4p-ew water model: Vapor pressure and boiling point. *Journal of Chemical Physics*, 123:194504.

[28] Almeida, N. M., Bali, S. K., James, D., Wang, C., and Wilson, A. K. (2023). Binding of per- and polyfluoroalkyl substances (pfas) to the pparγ/rxrα-dna complex. *Journal of Chemical Information and Modeling*, 63:7423–7443.

[29] Bali, S. K., Marion, A., Ugur, I., Dikmenli, A. K., Catak, S., and Aviyente, V. (2018). Activity of topotecan toward the dna/topoisomerase i complex: A theoretical rationalization. *Biochemistry*, 57:1542–1551.

[30] Bali, S. K., Haslak, Z. P., Cifci, G., and Aviyente, V. (2023). Dna preference of indenoisoquinolines: a computational approach. *Organic & Biomolecular Chemistry*, 21:4518–4528.

[31] Findik, B. K., Cilesiz, U., Bali, S. K., Atilgan, C., Aviyente, V., and Dedeoglu, B. (2022). Investigation of iron release from the n- and c-lobes of human serum transferrin by quantum chemical calculations. *Organic & Biomolecular Chemistry*, 20:8766–8774.

[32] Miller, B. R., McGee, T. D., Swails, J. M., Homeyer, N., Gohlke, H., and Roitberg, A. E. (2012). Mmpbsa.py : An efficient program for end-state free energy calculations. *Journal of Chemical Theory and Computation*, 8:3314–3321.

[33] Bonomi, M., Bussi, G., Camilloni, C., Tribello, G. A., Banáš, P., Barducci, A., Bernetti, M., Bolhuis, P. G., Bottaro, S., Branduardi, D., Capelli, R., Carloni, P., Ceriotti, M., Cesari, A., Chen, H., Chen, W., Colizzi, F., De, S., Pierre, M. D. L., Donadio, D., Drobot, V., Ensing, B., Ferguson, A. L., Filizola, M., Fraser, J. S., Fu, H., Gasparotto, P., Gervasio, F. L., Giberti, F., Gil-Ley, A., Giorgino, T., Heller, G. T., Hocky, G. M., Iannuzzi, M., Invernizzi, M., Jelfs, K. E., Jussupow, A., Kirilin, E., Laio, A., Limongelli, V., Lindorff-Larsen, K., Löhr, T., Marinelli, F., Martin-Samos, L., Masetti, M., Meyer, R., Michaelides, A., Molteni, C., Morishita, T., Nava, M., Paissoni, C., Papaleo, E., Parrinello, M., Pfaendtner, J., Piaggi, P., Piccini, G. M., Pietropaolo, A., Pietrucci, F., Pipolo, S., Provasi, D., Quigley, D., Raiteri, P., Raniolo, S., Rydzewski, J., Salvalaglio, M., Sosso, G. C., Spiwok, V., Šponer, J., Swenson, D. W., Tiwary, P., Valsson, O., Vendruscolo, M., Voth, G. A., and White, A. (2019). Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods 2019 16:8*, 16:670–673.

[34] Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. (2013). Plumed 2: New feathers for an old bird. *Computer Physics Communications*, 185:604–613.

[35] Bonomi, M., Branduardi, D., Bussi, G., Camilloni, C., Provasi, D., Raiteri, P., Donadio, D.,

Marinelli, F., Pietrucci, F., Broglia, R. A., and Parrinello, M. (2009). Plumed: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180:1961–1972.

[36] Roe, D. R. and Cheatham, T. E. (2013). Ptraj and cpptraj: Software for processing and analysis of molecular dynamics trajectory data. *Journal of Chemical Theory and Computation*, 9:3084–3095.

APPENDIX A

SUPPORTING TABLES

Table S8.1 The estimated binding energies and statistics for secondary selected poses for bCD and listed guest molecules are shown. RMSE: root mean square errors, MAE: mean absolute errors, ME: mean errors, r^2: correlation coefficient, m: slope of the correlation plots, τ : Kendall's Tau rank correlation coefficient.

	Pose	dG (exp, kJ/mol)	dG (exp, kcal/mol)	MM-GBSA (avg, kcal/mol)	MM-PBSA (avg, kcal/mol)	MM-GBSA (first 5ns, kcal/mol)	MM-PBSA (first 5ns, kcal/mol)	MM-GBSA (last 5ns, kcal/mol)	MM-PBSA (last 5ns, kcal/mol)	SMD (10Å/ns, avg, kcal/mol)
bCD	Tdz-p3	-23.86	-5.70	-23.76±2.45	-20.84±2.59	-23.48±2.22	-20.28±2.26	-24.34±2.71	-21.42±2.85	-18.62±1.29
	Tfp-p2	-21.18	-5.06	-22.55±3.50	-17.49±3.19	-24.07±3.03	-18.25±2.82	-22.07±2.66	-17.51±2.45	-18.95±1.51
	Pmz-p3	-20.81	-4.97	-22.53±4.30	-19.56±4.56	-26.01±3.81	-23.51±4.04	-21.20±3.00	-17.86±2.97	-14.92±1.51
	Pmt-p2	-18.73	-4.48	-21.26±2.63	-18.07±2.80	-20.65±2.38	-17.00±2.42	-20.62±2.22	-17.50±2.25	-15.44±0.88
	Cpz-p2	-22.66	-5.42	-20.22±2.91	-17.52±2.67	-19.76±2.69	-17.20±2.43	-20.12±2.90	-17.19±2.53	-10.34±1.51
	RMSE	--	--	16.98	13.62	17.82	14.44	16.60	13.24	10.98
	MAE	--	--	16.94	13.57	17.62	14.24	16.55	13.17	10.53
	ME	--	--	16.94	13.57	17.62	14.24	16.55	13.17	10.53
	r^2	--	--	0.12	0.21	0.005	0.01	0.35	0.42	0.001
	m	--	--	1.004	1.44	0.405	0.57	2.10	2.44	0.27
	τ	--	--	0.4	0.2	-0.2	0	0.4	0.2	0

Table S8.2 The estimated binding energies and statistics for secondary selected poses for HbCD and listed guest molecules are shown. RMSE: root mean square errors, MAE: mean absolute errors, ME: mean errors, r^2: correlation coefficient, m: slope of the correlation plots, τ : Kendall's Tau rank correlation coefficient.

	Pose	dG (exp, kJ/mol)	dG (exp, kcal/mol)	MM-GBSA (avg, kcal/mol)	MM-PBSA (avg, kcal/mol)	MM-GBSA (first 5ns, kcal/mol)	MM-PBSA (first 5ns, kcal/mol)	MM-GBSA (last 5ns, kcal/mol)	MM-PBSA (last 5ns, kcal/mol)	SMD (10Å/ns, avg, kcal/mol)
HbCD	Tdz-p2	-27.03	-6.46	-26.32±2.38	-24.89±2.56	-26.44±2.32	-25.15±2.52	-26.75±2.32	-25.03±2.47	-17.72±1.52
	Tfp-p2	-23.17	-5.54	-22.27±5.00	-17.68±3.64	-24.60±4.17	-18.81±3.23	-24.24±2.78	-18.99±2.68	-14.86±1.26
	Pmz-p2	-21.15	-5.05	-24.58±3.96	-22.45±3.77	-25.42±2.27	-23.00±2.55	-27.72±2.51	-25.44±2.18	-20.47±1.42
	Pmt-p3	-22.42	-5.36	-22.76±2.42	-20.19±2.64	-22.57±2.35	-20.39±2.56	-22.64±2.33	-19.88±2.67	0.00±0.00
	Cpz-p4	-22.6	-5.40	-29.76±4.18	-26.33±3.94	-28.79±2.64	-25.37±2.69	-32.01±2.70	-28.41±1.78	-18.16±1.39
	RMSE	--	--	19.76	17.02	20.11	17.16	21.37	18.34	11.34
	MAE	--	--	19.58	16.75	20.01	16.99	21.11	17.99	10.82
	ME	--	--	19.58	16.75	20.01	16.99	21.11	17.99	8.68
	r^2	--	--	0.03	0.09	0.04	0.13	0.03	0.00	0.02
	m	--	--	1.05	1.92	0.88	1.95	1.05	0.44	2.02
	τ	--	--	0.00	0.00	0.20	0.00	0.00	-0.20	-0.20

Table S8.3 Cumulative statistics of both bCD and HbCD results with secondary poses. RMSE: root mean square errors, MAE: mean absolute errors, ME: mean errors, r^2: correlation coefficient, m: slope of the correlation plots, τ : Kendall's Tau rank correlation coefficient.

	Analysis	MM-GBSA (avg, kcal/mol)	MM-PBSA (avg, kcal/mol)	MM-GBSA (first 5ns, kcal/mol)	MM-PBSA (first 5ns, kcal/mol)	MM-GBSA (last 5ns, kcal/mol)	MM-PBSA (last 5ns, kcal/mol)	SMD (10Å/ns, avg, kcal/mol)
Cumulative Statistics	RMSE	18.42	15.42	19.00	15.86	19.13	16.00	11.17
	MAE	18.26	15.16	18.83	15.61	18.83	15.58	10.68
	ME	18.26	15.16	18.83	15.61	18.83	15.58	9.61
	R^2	0.17	0.24	0.12	0.18	0.15	0.19	0.001
	m	2.18	2.97	1.76	2.49	2.77	3.36	0.40
	τ	0.29	0.25	0.07	0.11	0.38	0.29	-0.07

Table S8.4 Cumulative statistics of both bCD and HbCD results with lowest energy poses for each calculation type. RMSE: root mean square errors, MAE: mean absolute errors, ME: mean errors, r^2: correlation coefficient, m: slope of the correlation plots, τ : Kendall's Tau rank correlation coefficient.

	Analysis	MM-GBSA (avg, kcal/mol)	MM-PBSA (avg, kcal/mol)	MM-GBSA (first 5ns, kcal/mol)	MM-PBSA (first 5ns, kcal/mol)	MM-GBSA (last 5ns, kcal/mol)	MM-PBSA (last 5ns, kcal/mol)	SMD (10Å/ns, avg, kcal/mol)
Cumulative Statistics	RMSE	19.35	16.19	19.53	16.31	19.99	16.79	12.76
	MAE	19.24	16.03	19.44	16.14	19.77	16.5	12.65
	ME	19.24	16.03	19.44	16.14	19.77	16.50	12.65
	r^2	0.30	0.41	0.23	0.27	0.16	0.23	0.08
	m	2.48	3.26	1.95	2.71	2.48	3.18	0.97
	τ	0.51	0.36	0.13	0.16	0.24	0.29	-0.02

APPENDIX B

SUPPORTING FIGURES

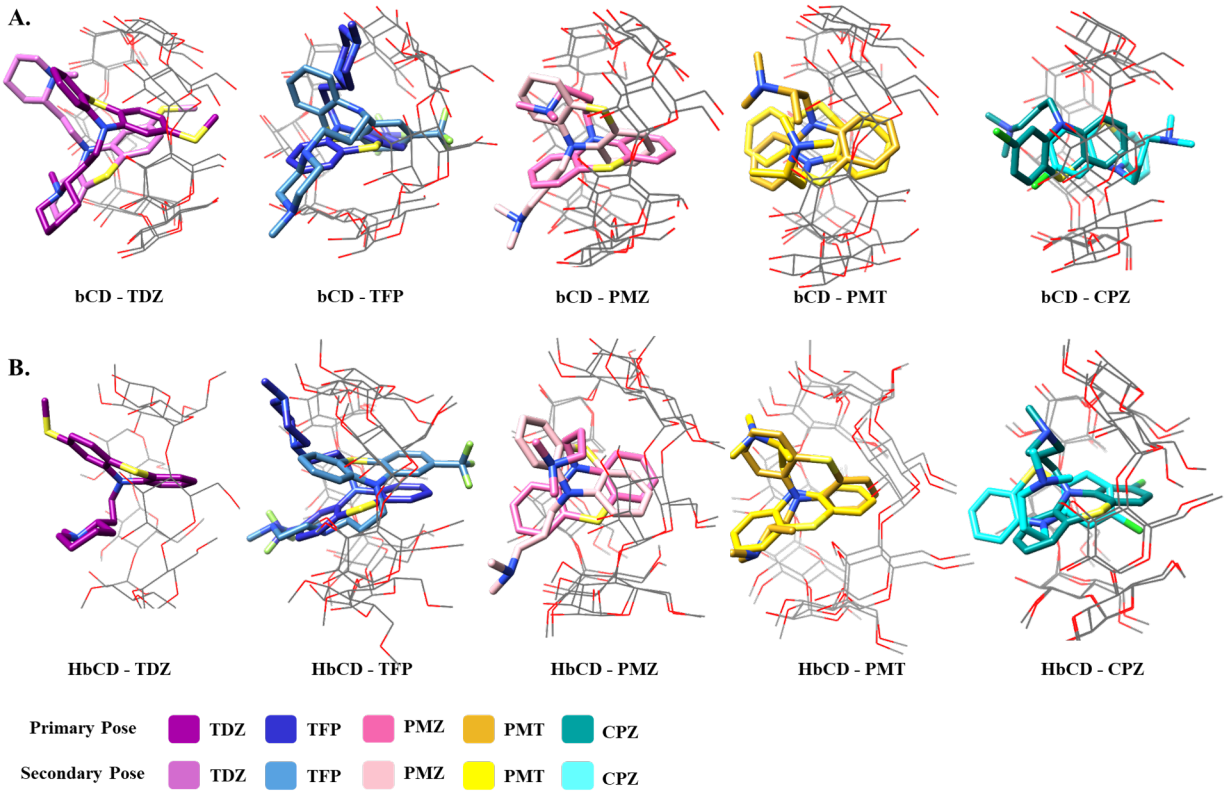


Figure S8.1 The primary and secondary poses considered for (A) bCD and (B) HbCD molecules. Oreintations (arm/core) are as follows: bCD: SP/SP; SP/SP; S-/S-; S-/S-; SS/PS. HbCD: SS; SS/SP; S-/S-; S-/S-; SP/SP

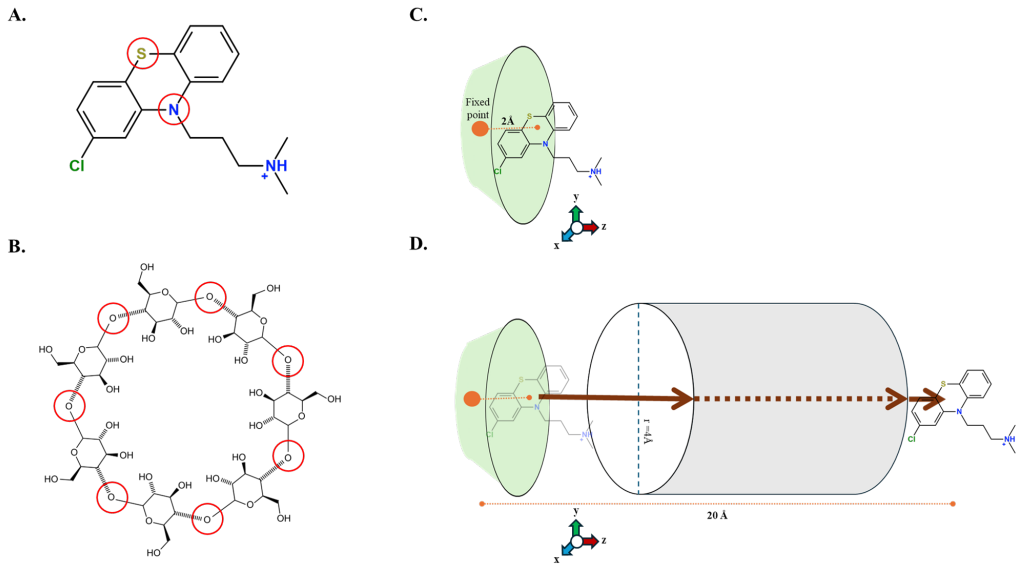


Figure S8.2 An overview of SMD protocol. A. Atoms selected for the COM-guest calculation. CPZ is given here as an example. B. Atoms selected for COM-host calculation. bCD is given here as an example.C. Details of the starting point for SMD calculations. (D) SMD path of the guest molecules. A cylindrical restraint was applied to the mean of the xy distance between center-of-mass of sulfur and nitrogen atoms of phenothiazine core and center-of-mass of oxygen atoms from glycosidic bond. CPZ is shown here as an example.

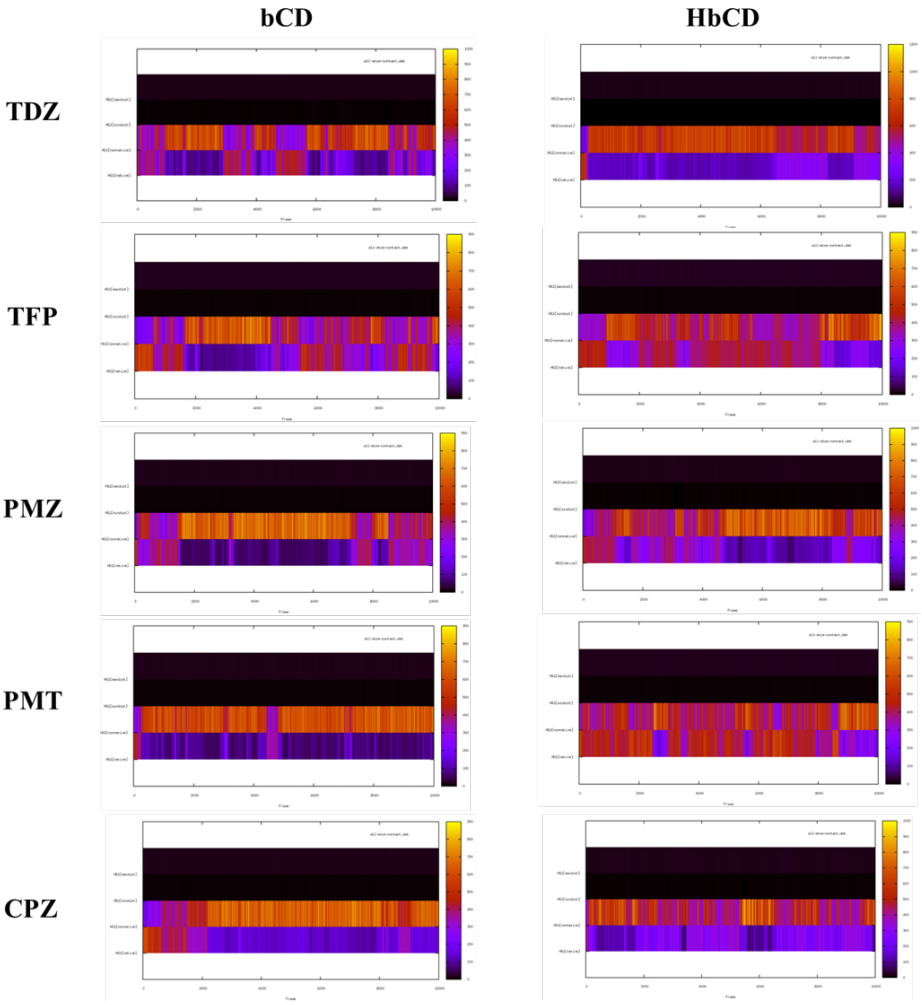


Figure S8.3 The maps for the total contact numbers (native and non-native) between host and guest molecules for primary poses.

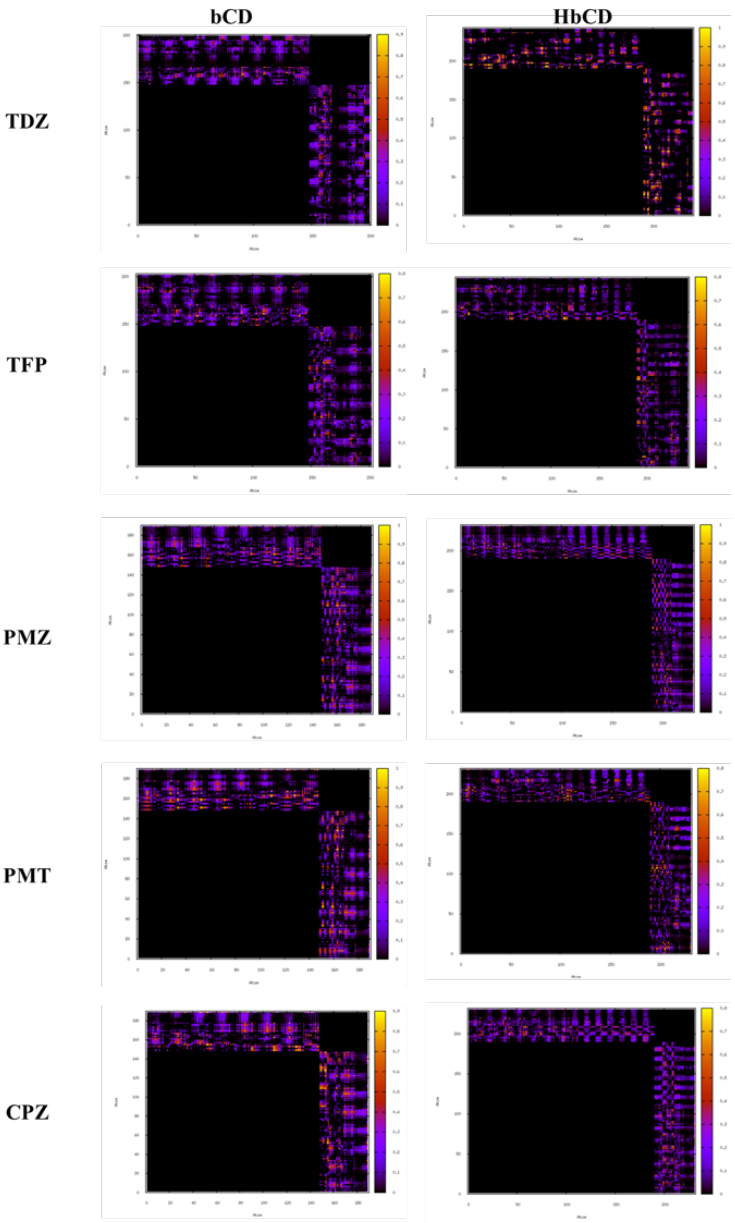


Figure S8.4 The atom contact maps (non-native) between host and guest molecules for primary poses.

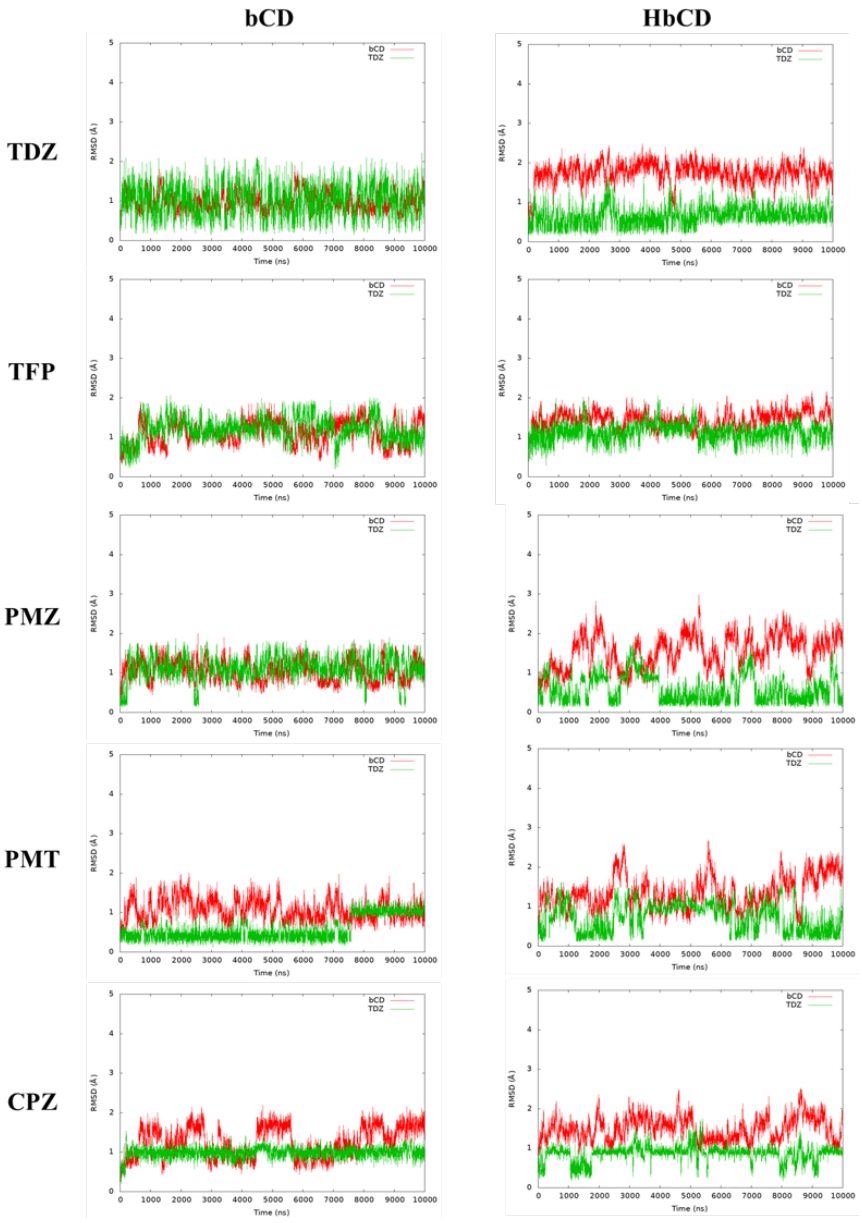


Figure S8.5 The RMSD time-series of host and guest molecules for primary poses.

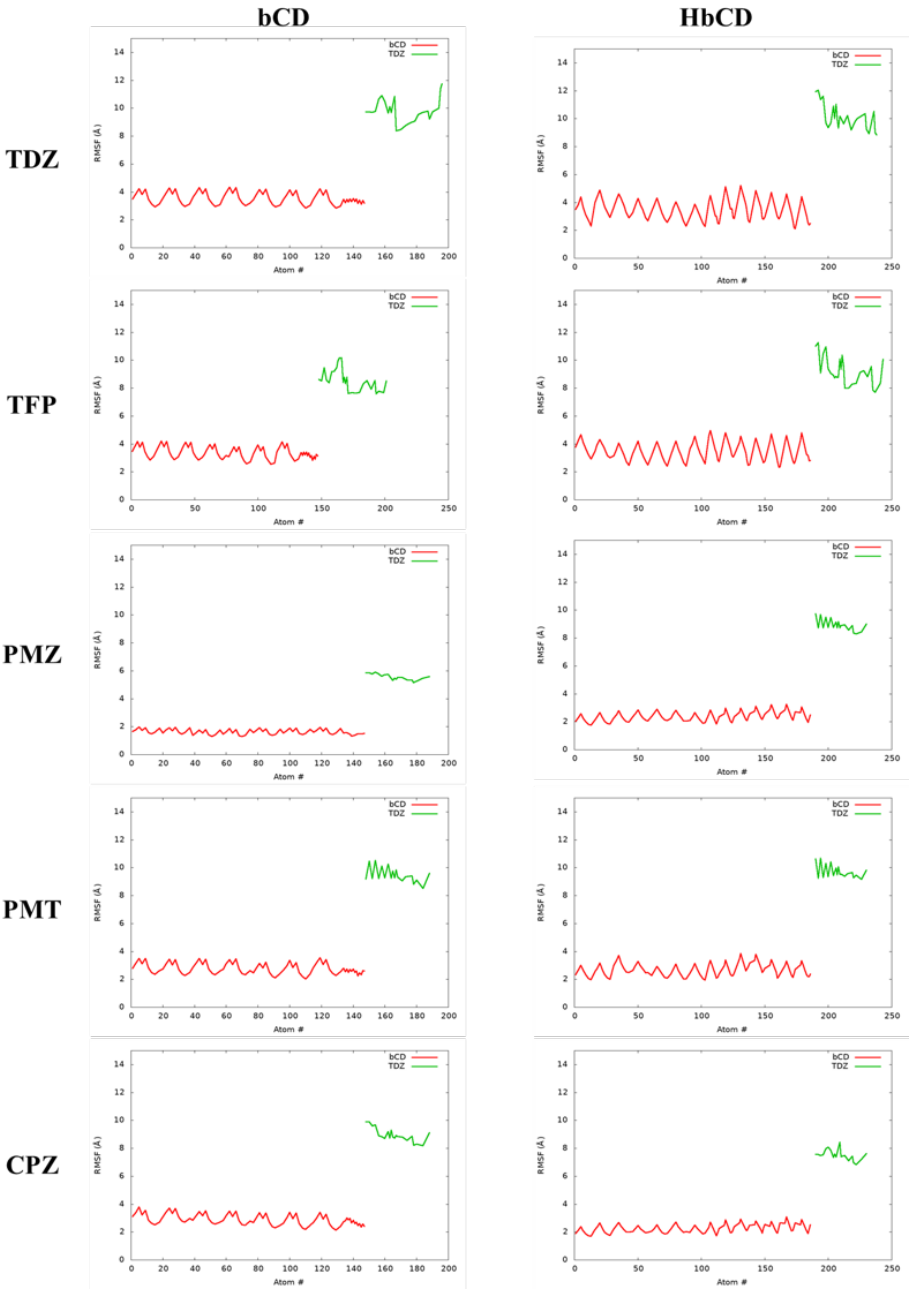


Figure S8.6 The per-atom RMSF plots of host and guest molecules for primary poses.

CHAPTER 9

CONCLUDING REMARKS AND FUTURE DIRECTIONS

Computational chemistry provides us a set of versatile tools study various phenomenons, including developing of new compounds to target certain diseases and understanding toxicological effects of specific chemicals on various organism. Describing the dynamics of the proteins and their interactions with ligands or inhibitor compounds allows us to have better understanding of the conformational and energetic spaces in which ligands can bind to proteins. With the appropriate use of these tools, we can gain insight for the ligand binding phenomena and the protein dynamics associated with it.

In Chapters 3, 4, and 5, the impact of a class of environmental pollutants (PFAS) on human and fish proteins were investigated. We have targeted three different proteins: PPARγ-RXRα-DNA complex, Estrogen receptors α and β, and thyroglobulin protein. The results highlight that the PFAS are binding strong enough to be able to exert toxic effects and impact the conformational flexibility of the target proteins. Our results outline the important residues in PFAS recognition and binding in these protein pockets.

In Chapter 6 and 7, in collaboration with Dr. Robert Abramovitch's and Dr. Edmund Ellsworth's groups, compounds aimed to kill *Mycobacterium tuberculosis* infections were developed and their interactions with the target proteins were investigated. One of the target proteins, DosS, is a heme protein hypothesized to sense the redox change in the environment. Our calculations suggested that the isoxazole moiety is favored for iron coordination. Furthermore, to elucidate the mechanism of inhibition of DosS, we employed constant pH simulations as well as classical MD simulations at different states of the protein, ad provided insight into the protonation state changes and conformational differences of the pocket residues. The other target, mmpL3, is a membrane protein responsible for relocating TMM lipid from inner membrane to periplasmic region of Mtb bacteria. The compounds developed by Dr. Ellsworth's group were modeled and key residues and interactions were identified. Furthermore, we also investigated the other tuberculous bacteria and compared their binding sites to allow for the more efficient targeting with the developed compounds.

In Chapter 8, we tested the performance of our methods as well as new SMD protocols using the dataset from Statistical Assessment of the Modeling of Proteins and Ligands 9 (SAMPL9) blind

challenge. Our results show good performance in terms of ranking the binding affinities of tested host-guest systems.

Besides the work that was presented here, we also worked collaboratively with Reata Pharmaceuticals until October 2023 to assist their drug discovery programs, and the contents of those work cannot be shared.

Going forward, PFAS still continues to be a threat to human well-being and ecological health. Understanding the detailed mechanisms in which PFAS exerts its toxicity is crucial to develop appropriate mitigation strategies at different exposure levels. Computational modeling is a very powerful approach in providing molecular level understanding of toxicities and can be used to uncover the impact of many PFAS towards important target proteins. The nuclear receptors have been the main focus for the PFAS toxicity in humans and other vertebrae, and comparison of all the data available on these nuclear receptors would highlight the key characteristics of the protein pocket features as well PFAS features that is associated with different degrees of toxicities. Besides the nuclear receptors that are known targets for PFAS, understanding how specific signaling mechanisms are disturbed by PFAS exposure is needed. For instance, as mentioned in Chapter 4, thyroid hormone levels in human body is impacted by PFAS exposure. However, there are many layers to understand the molecular details of the cascade of events that would lead to such health problems. Investigation of thyroid hormone signaling after their production by hTG protein is also a key consideration to obtain a better picture on PFAS toxicity on thyroid system.

While TB is not an immediate public health threat in developed countries, it is still a concerning problem in most parts of the rural areas. Considering the shortcomings of the current treatment strategies, novel and more effective drugs are required. mmpL3 and DosS are two crucial targets in combating TB, and understanding how the current compounds interact with these two targets are crucial in developing candidate compounds. Methods for which the computational challenges associated with the two aforementioned targets were explained in this dissertation. Moving this one step further, a bigger model system together with the relevant lipid molecules is needed to fully grasp the impact of inhibitor compounds on the mmpL3 protein. Furthermore, addition of TMM

lipid to the membrane will also be a key step to further uncover how the lipid binding to mmpL3 would be impacted by the inhibitory compounds. DosS protein is a challenging target to study using classical MD modeling due to the presence of an iron center. A workflow that can be built on top the work that has been presented in this thesis is required to accurately predict the binding affinities of investigated compounds.