Development of *In Silico* Tools to Predict the Behavior of Per- and Polyfluoroalkyl Substances (PFAS) in Biological Systems

by

Weixiao Cheng

Bachelor of Science, Shandong University, 2011

Master of Engineering, Zhejiang University, 2014

Submitted to the Graduate Faculty of the

Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Weixiao Cheng

It was defended on

June 29, 2021

and approved by

Dr. Carla Ng, Assistant Professor, Department of Civil and Environmental Engineering, Department of Environmental and Occupational Health

Dr. Radisav Vidic, William Kepler Whiteford Professor and Chair, Department of Civil and Environmental Engineering

Dr. Vikas Khanna, Associate Professor, Department of Civil and Environmental Engineering, Department of Chemical and Petroleum Engineering

Dr. Murat Akcakaya, Assistant Professor, Department of Electrical and Computer Engineering

Dissertation Director: Dr. Carla Ng, Assistant Professor, Department of Civil and Environmental Engineering, Department of Environmental and Occupational Health

Development of *In Silico* Tools to Predict the Behavior of Per- and Polyfluoroalkyl Substances (PFAS) in Biological Systems

Weixiao Cheng, PhD

University of Pittsburgh, 2021

Per and polyfluoroalkyl substances (PFAS) are a group of chemicals that have been widely used in industrial and consumer products for decades. Recent estimates suggest there are over 4000 PFAS on the global market. However, many of these have very little information available about their potential hazards. Given the vast number of PFAS, a three-level hierarchical framework that includes permeability-limited physiologically based toxicokinetic (PBTK) model, molecular dynamics (MD) based workflow and machine learning (ML) based quantitative structure-activity relationships (QSAR) was proposed to inform the toxicokinetics, bioaccumulation and toxicity of PFAS. The PBTK model was developed to estimate the toxicokinetic and tissue distribution of perfluorooctanoic acid (PFOA) in male rats; the hierarchical Bayesian analysis was used to reduce the uncertainty of parameters and improve the robustness of the PBTK model. By comparing with different experimental studies, most of the predicted plasma toxicokinetic (e.g., half-life) and tissue distribution fell well within a factor of 2.0 of the measured data.

Moreover, a modeling workflow that combines molecular docking and MD simulation techniques was developed to estimate the binding affinity of PFAS for liver-type fatty acid binding protein (LFABP). The results suggest that EEA and ADONA are at least as strongly bound to rat LFABP as perfluoroheptanoic acid (PFHpA), and to human LFABP as PFOA; both F-53 and F-53B have similar or stronger binding affinities than perfluorooctane sulfonate (PFOS). In addition, human, rat, chicken, and rainbow trout had similar binding affinities to one another for each tested

PFAS, whereas Japanese medaka and fathead minnow had significantly weaker LFABP binding affinity for some PFAS.

Finally, the ML-based QSAR model was developed to predict the bioactivity of around 4000 PFAS from the OECD report. Based on the collected PFAS dataset, a total of 5 different machine learning models were trained and validated that cover a variety of conventional models (i.e., logistic regression, random forest and multitask neural network) and advanced graph-based models (i.e., graph convolutional network and weave model). The model indicated that most of the biologically active PFAS have perfluoroalkyl chain lengths less than 12 and are categorized into fluorotelomer-related compounds and perfluoroalkyl acids.

Table of Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I am extremely grateful to my advisor Dr. Carla Ng for her invaluable guidance and continuous support through my PhD research journey. Dr. Ng is one of the nicest people I know in my life, I feel very lucky to be one member of the Ng Lab. In the past five years, I have learned many things from her, from research techniques to academic writing to communication skills. She was also very supportive of me when I wanted to pursue a second degree. I truly appreciate everything she has done during my PhD study.

I would like to express my sincere gratitude to committee members Dr. Radisav Vidic, Dr. Vikas Khanna, and Dr. Murat Akcakaya for their helpful discussion and suggestions on completing this dissertation.

I am also grateful to my lab members Manoochehr Khazaee, Trevor Sleight, Megha Bedi, Zhaokai Dong, Hajar Smaili, Yuexin Cao and Lacey W. Heinsberg for their support and encouragement. It has been such a wonderful experience working with them.

Finally, I want to thank my parents, my wife, and my sister for always being there for me. The PhD journey is not always smooth, but they always gave me unconditional support when needed. Thank you!

1.0 Dissertation Introduction

1.1 Motivation

1.1.1 PFAS as Emerging Contaminants

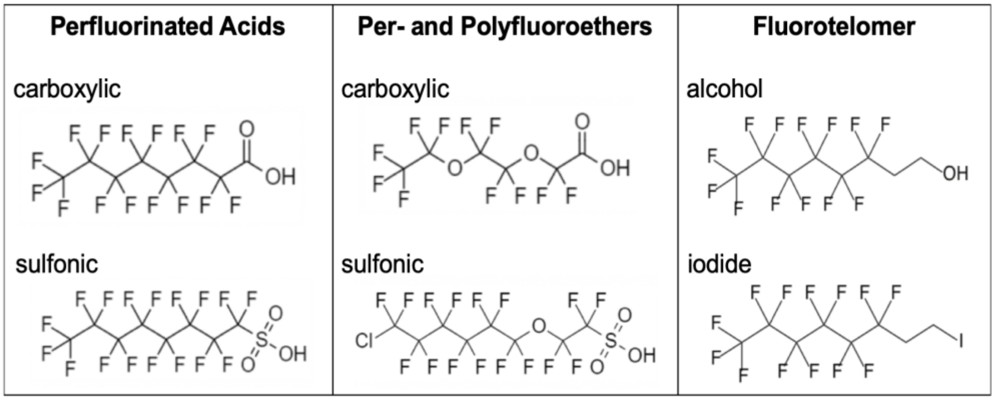


Figure 1. Example of PFAS structures.

Per- and polyfluoroalkyl substances (PFAS, C_nF_{2n+1}-R) are a diverse group of chemicals that have been widely used in a variety of industrial and consumer products, from fire-fighting foams to food contact materials to apparel.[1-6] Based on a recent Organization for Economic Cooperation and Development (OECD) report, there are 4730 PFAS that have been in some way registered and/or produced since the late 1940s.[7] Some of their chemical structures are indicated in Figure 1; broadly, those chemicals can be divided into perfluoroalkyl acids (PFAA, e.g., perfluoroalkyl carboxylic acids), PFAA precursors (e.g., fluorotelomer-based substances) and others (e.g., fluoropolymers).[8] Due to the strength of the carbon-fluorine bond, fully fluorinated PFAS, such as perfluoroalkyl carboxylic acids (PFCA) and perfluoroalkane sulfonic acids (PFSA), are extremely persistent in the environment and hard to remove, while other PFAS with partially fluorinated structures, such as fluorotelomer alcohols, usually break down to form more persistent, fully fluorinated PFAS.[9] As a result, those PFAS have been detected ubiquitously in the environment, wildlife and in humans.[10]

Moreover, unlike other persistent organic pollutants, many PFAS are acids that are almost fully ionized at environmentally relevant pH.[11] Instead of passively accumulating in fat tissue, those chemicals interact with different kinds of proteins such as serum albumin and membrane transporters, which results in tissue-specific accumulation patterns.[12] Experimental studies have found that many long-chain PFAA (defined as PFCAs with ≥7 perfluorinated carbons and PFSAs with ≥6 perfluorinated carbons) like perfluorooctanoic acid (PFOA) and perfluorooctanesulfonic acid (PFOS) mainly accumulate in blood, liver and kidneys,[13] and their biological half-lives were estimated to be several years for humans.[14] Finally, toxicological studies in animals have shown that exposure to long-chain PFAS and short-chain PFAS (in higher concentrations) cause toxic effects on reproduction and development, and on the nervous, endocrine, and immune systems, among others.[15-23]

The unique properties of PFAS and concerns about their persistence, bioaccumulation and toxicity have led to a phase-out of the production of long-chain PFAAs for the majority of uses.[24, 25] As a result, manufacturers have started using replacement substances that include shorter-chain homologues of PFOA and PFOS, as well as perfluoroether carboxylic acids (PFECAs) and perfluoroether sulfonic acids (PFESAs).[25, 26] However, to date little information is disclosed about the identity and frequency of use of those alternative PFAS, let alone their potential hazards such as bioaccumulation potential and toxicity.[26] Given the large number of PFAS, the scarce knowledge about their potential risks and the lack of effective control measures to eliminate them from the environment, PFAS are considered a potentially intractable, never-ending chemicals management issue that challenges society.[8]

1.1.2 Paradigm Shift in Toxicology

The field of toxicology is undergoing a paradigm shift from primarily *in vivo* animal studies to *in vitro* assays and sophisticated modeling approaches for toxicity assessments.[27, 28] The current risk assessment system, which relies on whole-animal-based toxicity-testing approaches for hazard identification and dose-response assessment, suffers from important limitations. Traditional toxicity testing methods are expensive, time-consuming, and use many laboratory animals, which limit their application for risk assessment of large numbers of chemicals like PFAS.[27] Moreover, they provide little insight into the modes of action that are important for interpreting interspecies differences in toxicity and little information for assessing variation among individuals in specific susceptible groups.[27] Alternative approaches to *in vivo* animal testing are needed for risk assessment of environmental contaminants.

1.2 Objective

Given the large number of PFAS (nearly 5000 [7]) and our limited resources (e.g., time and cost), it is not feasible to evaluate all PFAS individually through experiments. Therefore, *in silico* methods based on computational biology hold great promise for the hazard and risk assessment of those PFAS. As part of the efforts to tackle PFAS management issues and a response to the paradigm shift in toxicology, this dissertation focused on the development of reliable and efficient *in silico* tools to predict the toxicokinetics, bioaccumulation and toxicity of PFAS. Specifically, the objectives of this project include:

(1) **Development of mechanistic physiologically based toxicokinetic (PBTK) models that aim at providing a quantitative simulation of the toxicokinetics and tissue distributions of PFAS in mammals.** PBTK modeling is a promising tool to inform risk assessment of chemicals like PFAS.[29] With appropriate specification of species- and chemical-specific parameters, PBTK models simulate absorption, distribution, metabolism, and excretion (ADME) of compounds in animals and humans, providing a useful tool to understand and extrapolate toxicokinetics across different species and dosing scenarios.[29] Although PBTK models have been developed for PFAS in humans, rats and monkeys,[30-36] none of these models explicitly take multiple protein-PFAS interactions into account. In addition, existing models use *in vivo* test data to fit many of their parameters, and thus the predictive power of these models largely relies on the quality of available *in vivo* data. When the training set is poor, predictions of traditional PBTK models are not satisfactory.[34] To overcome these problems, we developed mechanistic PBTK models that explicitly consider protein-PFAS interactions and rely on no *in vivo* test data for parameterization (other than data on organism physiology compiled from the literature). PFAS-related parameters are instead obtained from *in vitro* studies or *in silico* predictions, and then used to predict *in vivo* toxicokinetics. These mechanistic PBTK models are intended to provide an effective framework to conduct *in vitro*-*in vivo* and *in silico*-*in vivo* extrapolation.

(2) **Development of a molecular dynamics (MD) based workflow that provides reliable and efficient estimation of protein-PFAS binding affinity.** While many protein-PFAS interaction parameters are required to build the PBTK model, those data are very limited, especially for many untested PFAS. One way to obtain these parameters is by conducting *in vitro*

experiments. For example, a number of techniques such as fluorescence spectroscopy[37], equilibrium dialysis[38], and NMR[39] have been widely used in recent years to measure protein binding affinity. The kinetics of PFAS uptake or efflux facilitated by transporters can also be obtained from cell-based *in vitro* assays, by incubation under a range of test compound concentrations[40-42]. However, chemical standards for the majority of those untested PFAS are still not widely available, limiting our ability to investigate protein-PFAS interactions experimentally. Here a computational approach that combines molecular dynamics (MD) and molecular docking is proposed for prediction of protein-PFAS interactions in a reliable and relatively efficient way, requiring only information on the three-dimensional structure of PFAS and target proteins.

(3) **Development of machine learning (ML)-based quantitative structure-activity relationships (QSAR) to predict the bioactivity of PFAS for screening purposes.** Although the MD approach is a solid method to estimate protein-PFAS binding affinity, it is still too computationally expensive to handle the large PFAS dataset containing over 4000 chemicals, especially when the size of the protein of interest is large (e.g., serum albumin with molecular weight of around 65000 Daltons) — in such cases, the time it takes to run simulations and calculate free energies of binding are much longer (on the order of days per protein-PFAS pair).[43, 44] To solve this problem, we applied much more cost-effective ML-based QSAR techniques that are able to screen a large number of chemicals based on their bioactivity prediction results. ML-based QSAR modeling is a data-driven approach that applies ML algorithms to model the relationship between physical and biological properties of compounds and their chemical structures.[45] It is very cost-effective and therefore optimal for dealing with large datasets, and it has been successfully applied for decades in many areas like drug discovery and chemical toxicity predictions.[46, 47] In this project, we build QSAR models to predict the bioactivity for a large number of PFAS.

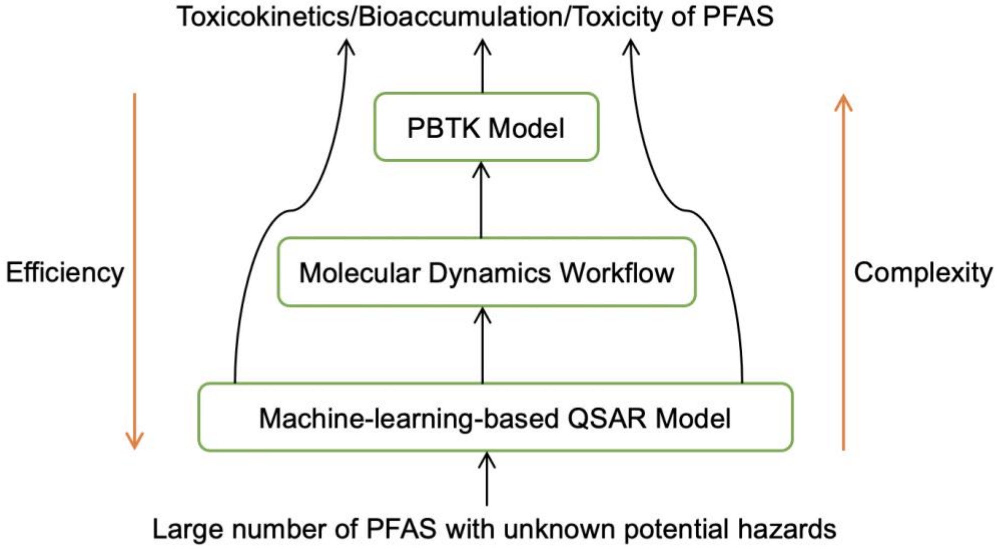

Toxicokinetics/Bioaccumulation/Toxicity of PFAS
PBTK Model
Molecular Dynamics Workflow
Machine-learning-based QSAR Model
Large number of PFAS with unknown potential hazards
Efficiency
Complexity

Figure 2. Three-level hierarchical framework for PFAS assessment.

As shown in Figure 2, the QSAR models, MD-based approach, and PBTK models form a three-level hierarchical framework to tackle the PFAS management issue. The bottom level is the ML-based QSAR model. As a statistical method, the QSAR cannot provide insights into the toxicokinetics for PFAS in biological system, Therefore, this level is mainly used for high-throughput screening and prioritization for the large PFAS collection. The intermediate level is the MD-based approach to calculate protein-PFAS binding affinity. The protein-PFAS interaction results generated at this level can be used as parameters for the PBTK models in the top level. On the top level, the PBTK model is used to predict the toxicokinetics and bioaccumulation of the prioritized PFAS in biological systems. Since our PBTK modeling takes protein-PFAS interactions into consideration, it can be used to identify the proteins that play critical roles in potential PFAS bioaccumulation and toxicity. Unlike ML-based QSAR, the PBTK model can provide insight into the behavior of PFAS in different tissues of the biological system and integrate this distribution into a whole-organism picture of the biological fate of PFAS.

1.3 Organization

The dissertation is organized as follows:

In Chapter 2.0, the permeability-limited PBTK model that was developed to estimate the toxicokinetics and tissue distribution of PFOA in male rats is presented. The model explicitly considered multiple protein-PFAS interactions including the cellular uptake and efflux of PFOA via transporters and the protein binding of PFOA in different tissues. In addition, rather than requiring *in vivo* data fitting, all PFOA-related parameters in our model were obtained from *in vitro* studies. Finally, the hierarchical Bayesian framework with Markov chain Monte Carlo (MCMC) was employed to reduce the uncertainty of parameters and improve the robustness of the PBTK model. The model performance was evaluated by comparing its predictions with experimental data from three different studies.

In Chapter 3.0, the molecular modeling workflow that combines molecular docking and molecular dynamics simulation techniques was developed to predict the protein binding affinity for PFAS in a reliable and relatively efficient way. To demonstrate the power of the workflow, the interactions between a total of 15 legacy and replacement PFAS (i.e., PFBA, PFPA, PFHxA, PFHpA, PFOA, PFNA, 2m-PFOA, PFBS, PFHxS, PFOS, ADONA, GenX, EEA, F-53 and F-53B) and 2 well-studied liver-type fatty acid binding proteins (LFABP) (i.e., rat and human LFABP) were tested to inform the bioaccumulation potential for novel PFAS. Moreover, the model was employed to probe the bioaccumulation potential across different species by examining LFABP interactions across 7 different species (i.e., human, rat, chicken, zebrafish, rainbow trout, fathead minnow, and Japanese medaka) for 9 PFAS (i.e., PFBA, PFPA, PFHxA, PFHpA, PFOA, PFNA, PFBS, PFHxS, and PFOS). The model prediction results were evaluated by comparing with experimental data extracted from three different studies. An approach that can predict PFAS impacts across different species can be used both to inform ecosystem protection (e.g. to identify most vulnerable wildlife species) and to identify appropriate model species for human toxicity studies.

In Chapter 4.0, a machine-learning-based quantitative structure-activity relationship (QSAR) model was developed to predict the bioactivity of around 4000 PFAS structures originally reported by the Organization for Economic Cooperation and Development (OECD) in 2018. By examining a number of available experimental data sets on chemical bioactivity, the first PFAS-specific database was constructed that contains bioactivity information for 1012 PFAS from 26 bioassays. Based on the collected PFAS dataset, a total of 5 different machine learning models were trained and validated that cover a variety of conventional models (i.e., logistic regression, random forest and multitask neural network) and advanced graph-based models (i.e., graph convolutional network and weave model).

Finally, in Chapter 5.0, the key results in the dissertation are summarized, the significance of the work and some recommendations for future research are discussed.

2.0 Development of a Permeability-Limited Physiologically Based Toxicokinetic (PBTK) Model for Perfluorooctanoic Acid (PFOA) in Male Rats

This chapter is reproduced in part with permission from:

Cheng, W. and Ng, C. A. A permeability-limited physiologically based pharmacokinetic (PBPK) model for perfluorooctanoic acid (PFOA) in male rats. *Environmental Science & Technology* 2017, 51 (17), 9930-9939. Copyright 2017 American Chemical Society. [https://pubs.acs.org/doi/abs/10.1021/acs.est.7b02602]

and

Cheng, W. and Ng, C. A. Bayesian refinement of a permeability-limited physiologically based pharmacokinetic (PBPK) model for perfluorooctanoic acid (PFOA) in male rats. *Chemical Research in Toxicology*, submitted for publication. Unpublished work copyright 2021 American Chemical Society.

Physiologically based toxicokinetic (PBTK) modeling is a powerful *in silico* tool that can be used to simulate the toxicokinetics and tissue distribution of xenobiotic substances, such as perfluorooctanoic acid (PFOA) in organisms. However, most existing PBTK models have been based on the flow-limited assumption. Given the large molecular weight and high extent of ionization of PFOA, we develop a permeability-limited PBTK model to estimate the toxicokinetics and tissue distribution of PFOA in male rats. Our model considers the uptake and efflux of PFOA via both passive diffusion and active transport facilitated by various transporters, association with serum albumin in circulatory and extracellular spaces, and association with intracellular proteins in liver and kidney. In addition, hierarchical Bayesian framework with Markov chain Monte

Carlo (MCMC) was employed to reduce the uncertainty of parameters and improve the robustness of the PBTK model. With the optimized posterior parameters, the PBTK model was evaluated by comparing its prediction with experimental data from three different studies. The results show that the uncertainties of the posterior model parameters were reduced substantially. Moreover, the PBTK model became more robust: with the posterior parameters, most of the predicted plasma toxicokinetic (e.g., half-life) and tissue distribution fell well within a factor of 2.0 of the experimental data. This study presents the first permeability-limited PBTK model that explicitly considers PFOA-protein interaction for rats. Given all parameters used were obtained from *in vitro* assays rather than *in vivo* data fitting, our model provides an effective framework to test *in vitro-in vivo* extrapolation and holds great promise for predicted toxicokinetics of per- and polyfluorinated alkyl substances in humans.

2.1 Introduction

Perfluorooctanoic acid ($C_7F_{15}COOH$, PFOA) is one of the most well-studied per- and polyfluoroalkyl substances (PFAS) that was widely used in industrial and consumer products.[48] The strong carbon-fluorine bonds in PFOA make it very resistant to metabolic and environmental degradation, which, coupled with its widespread use, results in its worldwide presence.[3, 10, 49] Although production of PFOA has been eliminated by many manufacturers,[24] worldwide human exposure to PFOA is still continuing.[50-53]

PFOA toxicokinetics have been studied extensively in mammals, and the results show that the substance is well absorbed orally and not metabolized.[54-56] It is primarily accumulated in plasma, liver, and kidney, with lowest levels distributed to adipose and muscle.[54, 57-61] In addition, PFOA can be eliminated through urine and feces, with urine being the major route. It has been reported that renal elimination rates are species- and sex-dependent.[62] For example, the half-life of PFOA in human blood was estimated to be about 3.5 years and no significant gender difference was observed.[63] However, the clearance of PFOA in rats is considerably sex-dependent, with reported half-lives of several days and hours for male and female rats, respectively.[64, 65]

Two principal underlying molecular mechanisms have been identified for PFOA toxicokinetics: protein binding and cell membrane transport. Studies revealed that PFOA is strongly bound to serum albumin as well as cytosolic fatty acid binding proteins (FABPs), which are pervasive in different tissues such as liver and kidney.[37, 39, 66, 67] Therefore, binding to different proteins could be an important determinant for high accumulation in blood, liver, and kidney. For membrane transport, both passive diffusion and active transport facilitated by transporter proteins play important roles in cellular uptake of PFOA.[40-42, 68] A number of transporters, such as organic anion transporters (Oats) and organic anion transporting polypeptides (Oatps) have been identified that are responsible for renal tubular excretion and reabsorption of PFOA in humans and rats.[41, 42, 68]

Due to its persistence and bioaccumulation, the potential human health risks of PFOA have received intense attention from environmental scientists and regulatory agencies.[24] Especially, a number of physiologically based toxicokinetic (PBTK) modeling tools have been developed to inform risk assessment for PFOA in different species such as humans, rats and monkeys.[30-36] All those models assume the chemical uptake rate to each tissue is mainly determined by blood flow rate rather than cell membrane permeability. By neglecting membrane permeability and its associated parameters, the flow-limited assumption simplifies the PBTK model process significantly. However, for chemicals with large molecular weights and/or ionic charges (e.g., PFOA which has a molecular weight of 414.09 Da and is negatively charged at environmentally and physiologically relevant pH[12]), cell membrane permeability becomes the rate-limiting process for uptake and needs to be included in the PBTK model.[29]

Based on the above considerations, we developed a permeability-limited PBTK model that explicitly considers cellular membrane permeability of PFOA through different tissues as well as the molecular mechanisms for PFOA toxicokinetics (i.e., protein binding and active transport process). We applied the permeability-limited PBTK model to estimate the plasma toxicokinetics and patterns of tissue distribution of PFOA in male rats, for which most protein-related parameters were available.

Finally, one potential limitation of this PBTK model is that some mechanism-related parameters are either based on a single study (e.g., the equilibrium association constant of albumin for the rat) or extrapolated from *in vitro* studies (e.g., active transport rates in the kidney). The limited knowledge about those key parameters leads to a substantial amount of uncertainty in the PBTK model. To address this issue, hierarchical Bayesian analysis with Markov chain Monte Carlo (MCMC) was applied to reduce the uncertainty of parameters and improve the robustness of the PBTK model.[69] With the optimized posterior parameters, the improved PBTK model was evaluated by comparing its prediction with both the prior model results and the experimental data from three separate studies,[54, 58, 59] where different PFOA dose levels and administration routes were used (providing a total of even data sets) for male rats.

2.2 Materials and Methods

2.2.1 Rat Model Structure

As indicated in Figure 3, the model includes seven tissues: blood, kidney, liver, gut, muscle, adipose and the rest of the body. Since this is a permeability-limited model, the consideration of tissue subcompartments is required. Except for blood, each tissue contains both a vascular space and tissue space, the latter of which can be further divided into two subcompartments: interstitial fluid and tissue. To characterize absorption and elimination processes of PFOA, gut lumen, filtrate and bile compartments were incorporated in our model.

The blood compartment functions as systemic circulation connecting each tissue compartment. In blood, PFOA binds to serum albumin based on the equilibrium association constant, K_a. Interstitial fluids of other compartments also contain albumin to which PFOA could bind.[70, 71]

Enterohepatic circulation may play a role in the distribution of PFOA in liver[72] and thus was considered in our model. Due to scarcity of data, we only included two transporters that could be associated with the cycling of PFOA in liver: Oatp1a1 and Ntcp, both of which are located at the basolateral membrane of hepatocytes.[73] Oatp1a1 has been demonstrated to be able to transport PFOA,[68] while for Ntcp, only the interactions with perfluorooctane sulfonate (PFOS) were reported.[74] Given the structural similarity between PFOA and PFOS, we assume that Ntcp could also transport PFOA. Once in the hepatocyte, PFOA can bind to liver-type fatty acid binding protein (LFABP) in hepatocytes while the free fraction is available for excretion into the bile duct via passive diffusion. Biliary PFOA is then circulated to gut lumen, where reabsorption of PFOA from the intestine back to systemic circulation can occur, as well as elimination of PFOA through defecation.

The kidney is another major elimination tissue, involving glomerular filtration, renal clearance, renal reabsorption, and renal efflux processes. The free fraction of PFOA can transport from blood into filtrate through both glomerular filtration and renal clearance. The latter process is mainly mediated via organic anion transporters (Oat1 and Oat3) located at the basolateral membrane of proximal tubular cells.[62] In filtrate, PFOA is actively reabsorbed by Oatp1a1 back to the tissue compartment (i.e., renal reabsorption),[62] where PFOA can bind to two different proteins, L-FABP and α2μ-globulin (traditionally but erroneously called kidney fatty acid binding protein[75]), both of which are present in rat kidney tissue.[76, 77] The free fraction of PFOA in kidney tissue might be excreted into blood through organic solute transporters (Ostα/β) (i.e., renal efflux). Based on the observation of lower kidney:blood PFOA concentration in male rats compared to female rats, it is hypothesized that male rats have more effective efflux transporters on the renal basolateral membrane excreting intracellular PFOA back to blood;[42] Ostα/β and Mrp6 are proposed to be promising candidates for PFOA efflux.[62] Given available kinetics data for Ostα/β, it was included in our model.

Finally, muscle and adipose were selected for comparison to other tissues, since they typically have the lowest levels of PFOA.[54, 59] All the remaining tissues were lumped into a single compartment, rest of body.

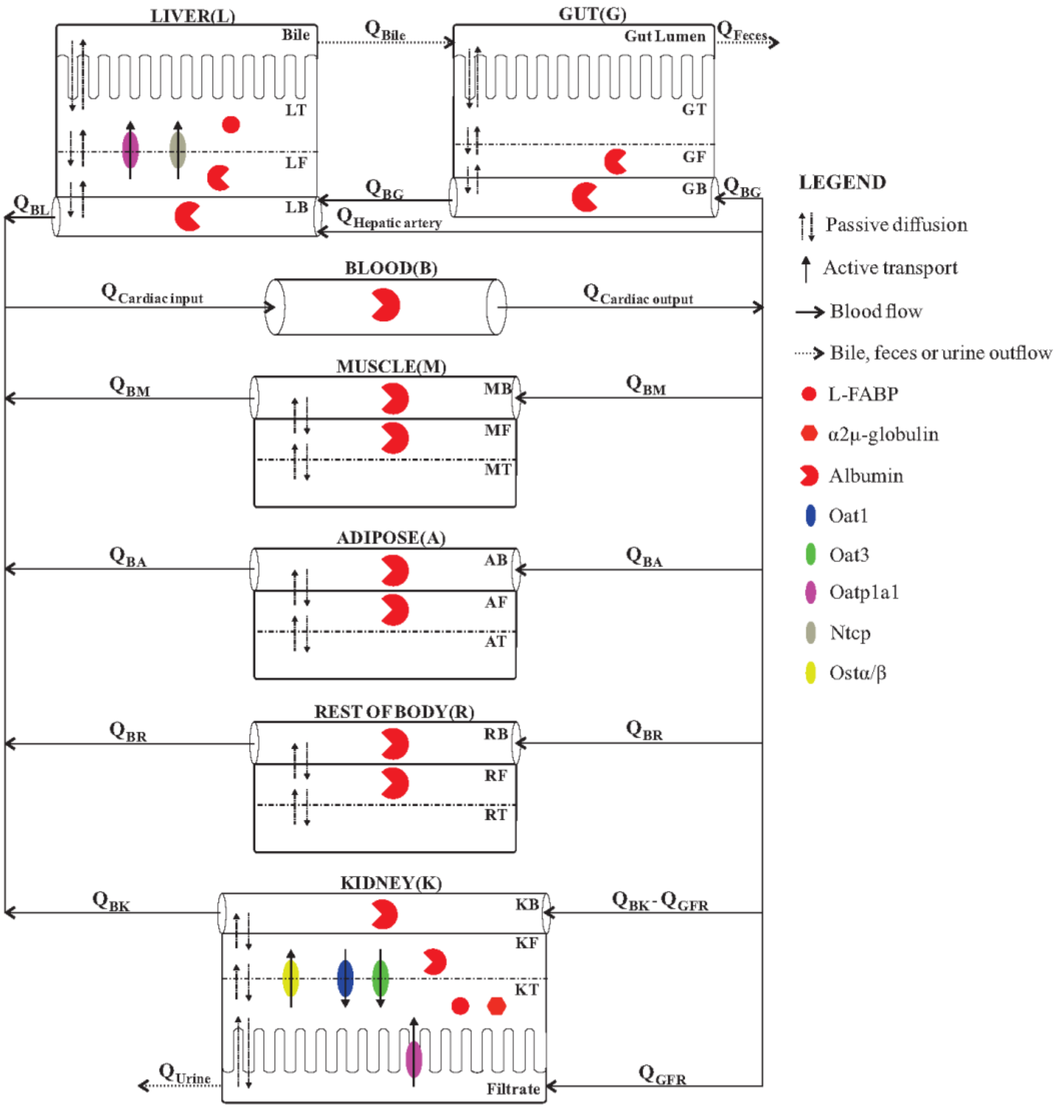


Figure 3. Rat model structure.

There are seven tissues including blood (B), liver (L), gut (G), kidney (K), muscle (M), adipose (A), and "rest of body" (R). All tissues except blood contain a vascular space (e.g. KB for kidney), interstitial fluid (e.g. KF for kidney) and tissue space (e.g. KT for kidney). Blood flow rate for each tissue is indicated (e.g. Q_{BK} is the blood flow rate to kidney).

2.2.2 Model Parameterization

All parameters used in our model were rat-specific, except capillary surface area, albumin concentrations in interstitial fluid compartments, and the transport kinetics of Ostα/β. The first two parameters were estimated from other mammalian studies,[70, 71, 78] and Ostα/β kinetics was based on an uptake study in human cells.[74] All these parameters are explained in detail below.

2.2.2.1 Rat Physiology

Physiological parameters were obtained from the literature. The average body weight of rats in each experimental dataset (Section 2.2.5) was used for model simulation. Fractional tissue volume, blood flow rate, interstitial fluid and blood volumes, and capillary surface areas for each compartment are summarized in Appendix Table 1. Other parameters including volume of bile, renal filtrate, and gut content as well as urinary, biliary, and fecal flow rates are also indicated. Detailed information on derivation of physiological parameters is provided in Appendix Table 1.

2.2.2.2 Protein Binding

In terms of the protein binding of PFOA, a total of three proteins were considered including albumin (in blood and interstitial compartments),[39] liver-type fatty acid binding protein, LFABP (in both liver and kidney tissue compartment),[37] and α2μ-globulin (only in the kidney tissue compartment);[75] those proteins have all been demonstrated as important determinants for PFAS accumulation in blood, liver, and kidney. The linear model parameterized by free fraction of PFOA was used to model the protein binding process.[79] As shown in equation 1, the free fraction of PFOA (*ff*), i.e., the ratio of free PFOA concentration (C_{free}) to the total PFOA concentration in tissue

(C_{total}), is determined by the equilibrium association constant (K_a) and the maximum binding capacity (B_m, which is considered as the protein concentration in tissue).

$$ff = \frac{C_{free}}{C_{total}} = \frac{1}{1+ K_a \times B_m} \tag{2-1}$$

Here, *ff* is considered to be independent of C_{free} and the protein binding for PFOA is therefore characterized by a constant parameter predefined by K_a and B_m. The K_a of albumin, LFABP and α2μ-globulin as well as the concentration of those proteins in different tissues (i.e., B_m) were obtained from the literature and shown in Appendix Table 2 and 3.

2.2.2.3 Membrane Transport

Both passive diffusion and active transport facilitated by proteins play essential roles in membrane transport of PFOA.[80-84] To derive the mass balance equations for those processes, passive diffusion rates for each tissue and active transport rates for each relevant membrane transporter are required. For passive diffusion, the effective permeability (P_{eff}) for each tissue was used to calculate the passive diffusion rate. As shown in equation 2, permeability is estimated based on Fick's Law:

$$P_{eff} = \frac{J}{A\Delta C} \tag{2-2}$$

Where J, the initial passive diffusion flux, was empirically determined by extracting *in vitro* data from Weaver et al.;[80] the average value of J is around 0.13 nmol/mg protein/min and is converted to mol/s by scaling to the protein content of each tissue-specific cell type (Appendix Table 2). A is the cellular surface area, which is assumed to be 4000 μm^2 for a single cell.[82] ΔC is the concentration of PFOA in the exposure medium (i.e., 10 μmol/L in the Weaver et al. study).[80]

Once P$_{\mathrm{eff}}$ is determined, the passive diffusion rate (k) between connected tissue compartments can be calculated as follows.

For diffusion between blood (B) and the interstitial fluid compartment in each tissue (iF):

$$k^{iF-B} = k^{B-iF} = \left(\frac{1}{Q_B^i} + \frac{1}{P_{eff}^B A^{B-iF}}\right)^{-1} \tag{2-3}$$

Where Q_B^i is the blood flow to each tissue, and A^{B-iF} is the surface area of exchange between blood and fluid compartment (Appendix Table 1).

For diffusion between the fluid (iF) and tissue (iT) subcompartment in each tissue, only permeability accounts for the overall mass transfer:

$$k^{iF-iT} = k^{iT-iF} = P_{eff}^i A^{iF-iT} \tag{2-4}$$

For tissues containing filtrate, bile, or gut lumen, the diffusion between tissues and those subcompartments (iS) are:

$$k^{iS-iT} = P_{eff}^i A^{iS-iT} \tag{2-5}$$

$$k^{iT-iS} = \frac{k^{iS-iT}}{CR_{SS}^{C-W}} \tag{2-6}$$

Where CR_{ss}^{C-W} is the steady-state cell-water concentration ratio, which can be extrapolated from *in vitro* data.[82] The values of CR_{ss}^{C-W} for liver to bile, kidney to filtrate, and enterocyte to gut lumen are shown in Appendix Table 4.

Regarding active transport, a total of five transporters are responsible for four active transport processes: (i) renal clearance, which involves Oat1 and Oat3 located at the basolateral membrane of proximal tubular cells; (ii) renal reabsorption, which involves Oatp1a1 located at the apical membrane of proximal tubular cells; (iii) renal efflux, which relates to Ostα/β located at the basolateral membrane of proximal tubular cells; and (iv) hepatocyte absorption, which relates to Oatp1a1 and Ntcp located at the basolateral membrane of hepatocytes.[62] Similar to the passive diffusion process, Fick's Law was used to derive the coefficients for those four active transport processes; the only difference is that the J parameter in equation 2 corresponds to the flux of transporter-expressing cell rather than the passive diffusion flux. The J value for each transporter and the calculated active transport coefficients are shown in Appendix Table 5. Finally, the active transport rate for each process can be derived by multiplying the transport coefficient by the surface area for exchange of the corresponding tissue.

2.2.3 Mass Balance Equations

Our PBTK model is based on permeability-limited equations that consider three or four subcompartments for each tissue. Each compartment is described by mass balance equations based on protein binding of PFOA and the transport between connected compartments.

For an individual tissue *i* (*i* = A, G, K, L, M, and R, for adipose, gut, kidney, liver, muscle, and rest of body, respectively), the mass of free PFOA in each subcompartment *j* (*j* = F and T, for fluid and tissue subcompartment, respectively) can be determined by: $\mathrm{M_{free}} = \mathrm{M} \times ff$, where *ff* is the free fraction parameter of PFOA. The value of *ff* for subcompartment ij (i.e., ff^{ij}) is calculated as follows:

$$ff^{Blood} = \frac{1.0}{1.0 + \mathrm{C}_{\mathrm{albumin}}^{\mathrm{Blood}} \times K_a^{albumin}} \tag{2-7}$$

$$ff^{iF} = \frac{1.0}{1.0 + \mathrm{C}_{\mathrm{albumin}}^{\mathrm{iF}} \times K_a^{albumin}} \tag{2-8}$$

$$ff^{KT} = \frac{1.0}{1.0 + \mathrm{C}_{\mathrm{LFABP}}^{\mathrm{KT}} \times K_a^{LFABP} + \mathrm{C}_{\mathrm{globulin}}^{\mathrm{KT}} \times K_a^{globulin}} \tag{2-9}$$

$$ff^{LT} = \frac{1.0}{1.0 + \mathrm{C}_{\mathrm{LFABP}}^{\mathrm{LT}} \times K_a^{LFABP}} \tag{2-10}$$

Where C_p^{ij} is the concentration of protein p (i.e., albumin, L-FABP or α2μ-globulin) in subcompartment ij. K_a^p represents the association constant of protein p.

Using the free fraction parameter to calculate $\mathrm{M}_{\mathrm{free}}$ in each compartment, the mass balances can then be expressed as detailed in equations 11-19 below.

In the blood compartment (B), which is in contact with and can be considered as a subcompartment of each tissue:

$$\frac{\mathrm{dM}_{\mathrm{free}}^{\mathrm{B}}}{\mathrm{dt}} = \sum\nolimits_{\mathrm{i}} \mathrm{b}^{\mathrm{iF-B}} \mathrm{M}_{\mathrm{free}}^{\mathrm{iF}} - \sum\nolimits_{\mathrm{i}} \mathrm{b}^{\mathrm{B-iF}} \mathrm{M}_{\mathrm{free}}^{\mathrm{B}} + \mathrm{b}^{\mathrm{Filtrate-B}} \mathrm{M}_{\mathrm{free}}^{\mathrm{Filtrate}} - \mathrm{b}^{\mathrm{B-Filtrate}} \mathrm{M}_{\mathrm{free}}^{\mathrm{B}} \tag{2-11}$$

In interstitial fluid subcompartment *i*F:

$$\frac{\mathrm{dM}_{\mathrm{free}}^{\mathrm{iF}}}{\mathrm{dt}} = \mathrm{b}^{\mathrm{B-iF}} \mathrm{M}_{\mathrm{free}}^{\mathrm{B}} - \mathrm{b}^{\mathrm{iF-B}} \mathrm{M}_{\mathrm{free}}^{\mathrm{iF}} + \mathrm{b}^{\mathrm{iT-iF}} \mathrm{M}_{\mathrm{free}}^{\mathrm{iT}} - \mathrm{b}^{\mathrm{iF-iT}} \mathrm{M}_{\mathrm{free}}^{\mathrm{iF}} + \mathrm{b}_{\mathrm{active}}^{\mathrm{iT-iF}} \mathrm{M}_{\mathrm{free}}^{\mathrm{iT}} - \mathrm{b}_{\mathrm{active}}^{\mathrm{iF-iT}} \mathrm{M}_{\mathrm{free}}^{\mathrm{iF}} \tag{2-12}$$

In tissue subcompartment *i*T (*i* = A, M, and R):

$$\frac{\mathrm{dM_{free}^{iT}}}{\mathrm{dt}} = \mathrm{b^{iF-iT}M_{free}^{iF}} - \mathrm{b^{iT-iF}M_{free}^{iT}} + \mathrm{b_{active}^{iF-iT}M_{free}^{iF}} - \mathrm{b_{active}^{iT-iF}M_{free}^{iT}} \tag{2-13}$$

For tissues (*i* = K, L, and G) that contain the additional compartments of filtrate, bile, or gut lumen (GL), the mass balance equations in tissue subcompartment *i*T and its corresponding additional compartments are:

$$\begin{aligned}\frac{\mathrm{dM_{free}^{KT}}}{\mathrm{dt}} = {} & \mathrm{b^{KF-KT}M_{free}^{KF}} - \mathrm{b^{KT-KF}M_{free}^{KT}} + \mathrm{b_{active}^{KF-KT}M_{free}^{KF}} - \mathrm{b_{active}^{KT-KF}M_{free}^{KT}} + \mathrm{b^{Filtrate-KT}M_{free}^{Filtrate}} \\ & -\mathrm{b^{KT-Filtrate}M_{free}^{KT}} + \mathrm{b_{active}^{Filtrate-KT}M_{free}^{Filtrate}}\end{aligned} \tag{2-14}$$

$$\begin{aligned}\frac{\mathrm{dM_{free}^{Filtrate}}}{\mathrm{dt}} = {} & \mathrm{b^{B-Filtrate}M_{free}^{B}} - \mathrm{b^{Filtrate-B}M_{free}^{Filtrate}} + \mathrm{b^{KT-Filtrate}M_{free}^{KT}} - \mathrm{b^{Filtrate-KT}M_{free}^{Filtrate}} \\ & - \mathrm{b_{active}^{Filtrate-KT}M_{free}^{Filtrate}} - \frac{\mathrm{Qurine}}{\mathrm{VFiltrate}}\mathrm{M_{free}^{Filtrate}}\end{aligned} \tag{2-15}$$

$$\begin{aligned}\frac{\mathrm{dM_{free}^{LT}}}{\mathrm{dt}} = {} & \mathrm{b^{LF-LT}M_{free}^{LF}} - \mathrm{b^{LT-LF}M_{free}^{LT}} + \mathrm{b_{active}^{LF-LT}M_{free}^{LF}} \\ & + \mathrm{b^{bile-LT}M_{free}^{bile}} - \mathrm{b^{LT-bile}M_{free}^{LT}}\end{aligned} \tag{2-16}$$

$$\frac{\mathrm{dM_{free}^{Bile}}}{\mathrm{dt}} = \mathrm{b^{LT-Bile}M_{free}^{LT}} - \mathrm{b^{Bile-LT}M_{free}^{Bile}} - \frac{\mathrm{Qbile}}{\mathrm{Vbile}}\mathrm{M_{free}^{Bile}} \tag{2-17}$$

$$\frac{\mathrm{dM_{free}^{GT}}}{\mathrm{dt}} = \mathrm{b^{GF-GT}M_{free}^{GF}} - \mathrm{b^{GT-GF}M_{free}^{GT}} + \mathrm{b^{GL-GT}M_{free}^{GL}} - \mathrm{b^{GT-GL}M_{free}^{GT}} \tag{2-18}$$

$$\frac{\mathrm{dM_{free}^{GL}}}{\mathrm{dt}} = \mathrm{b^{GT-GL}M_{free}^{GT}} - \mathrm{b^{GL-GT}M_{free}^{GL}} + \frac{\mathrm{Qbile}}{\mathrm{Vbile}}\mathrm{M_{free}^{Bile}} - \frac{\mathrm{Qfeces}}{\mathrm{VGL}}\mathrm{M_{free}^{GT}} \tag{2-19}$$

Where $\mathrm{M}_{\mathrm{free}}^{ij}$ represents free PFOA in *ij* subcompartment, which can freely transport between compartments. Parameters $\mathrm{b}^{ij\text{-}im}$ and $\mathrm{b}_{\mathrm{active}}^{ij\text{-}jm}$ are first-order rate constants for passive diffusion and active transport between subcompartments *ij* and *im* ($j \neq m$), respectively. Those rate constants are determined by the transport rate parameters divided by the volume of the corresponding compartment. For example, $b^{iF-iT} = \frac{k^{iF-iT}}{V^{iF}}$, where k is the passive diffusion rate from *i*F to *i*T subcompartment. Finally, Q and V are the flow rate and volume of the corresponding compartment, respectively.

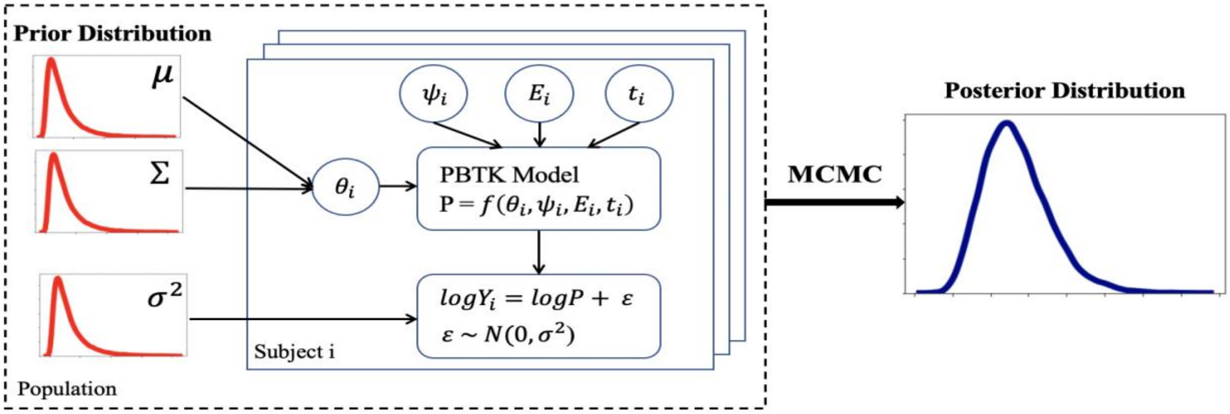


Figure 4. Workflow of hierarchical Bayesian framework for the permeability-limited PBTK model. The PBTK model is a function of chemical-related parameters (θ_i), physiological covariables (ψ_i), exposure scenarios (E_i) and sampling time points (t_i). The individual parameters for each subject i (θ_i) are drawn from the population distribution with mean (μ) and variance (Σ). The experimental error term (ε) has a normal distribution with mean 0 and variance (σ^2). Based on the prior information of μ, Σ and σ^2 and experimental data points (Y_i), the hierarchical Bayesian framework with MCMC simulation was used to generate the posterior distribution for those parameters.

2.2.4 Hierarchical Bayesian Framework

The hierarchical Bayesian approach[85] was employed to reduce the uncertainty and variability of the permeability-limited PBTK model for PFOA in the male rat. The Bayes Rule is shown in the following equation:[85]

$$P(\theta|Y) = \frac{P(\theta)P(Y|\theta)}{\int P(\theta)P(Y|\theta)d\theta} \qquad (2\text{-}20)$$

Where θ is the PBTK model parameter vector to be estimated. Y is the measured toxicokinetic data for PFOA. P(θ|Y) is the posterior distribution of the model parameters, and P(θ) is the prior distribution that describes the prior knowledge of parameters. P(Y|θ) is the likelihood of the experimental dataset. In order to perform Bayesian inference, first of all, prior distributions of model parameters need to be defined. Here, we mainly focus on the key parameters that are related to PFOA toxicokinetics mechanism due to the high uncertainty and sensitivity of those parameters. The prior distributions for those parameters were assumed based on the literature (e.g., most of biological parameters were assumed to be log-normal distributions). Next, the likelihood of the observed data set can be calculated based on the PBTK model. Given the likelihood and prior distribution of the parameter, the posterior distribution for those parameters can be inferred from the Bayes Rule. However, it is almost impossible to obtain an analytical expression for P(θ|Y). For this reason, the MCMC technique will be employed to estimate the posterior distribution for the parameters. MCMC is a powerful computational tool to provide samples of parameters without the analytical results of P(θ|Y).[85]

As indicated in Figure 4, the hierarchical structure consists of two major parts: the subject level and the population level. At the subject level, for each individual i, the PBTK model (i.e., function *f*) was used to predict the PFOA concentration-time profiles based on given parameters including chemical-related parameters (θ_i), physiological covariables (ψ_i), exposure scenarios (E_i) and sampling time points (t_i). The prediction results are related to the experimentally measured concentration data (Y_i) through the following error model:

$$logY_i = logP + \varepsilon \tag{2-21}$$

Where the error term ε is a normal variable with mean set equal to 0 and variance to σ^2. At the population level, to reflect the interindividual variability, chemical-related parameters (θ_i) were considered to be generated from a multivariate population distribution, with population mean (μ) and variance (Σ). The prior distribution of the μ and Σ for each parameter is discussed in Section 2.2.4.1. With the prior knowledge of the model parameters and the above hierarchical Bayesian framework, the Markov chain Monte Carlo (MCMC) technique was employed to numerically estimate the posterior distribution for the model parameters. Finally, the resulting change in the central estimate and the uncertainty and variability of those parameters were analyzed.

Table 1. Summary of the parameters of the PBTK model selected for MCMC analysis.

Parameters	Symbol	Values	Unit	Confidence Factor (Cf)
Effective permeability of blood	PeffB	0.179	mm/h	5
Effective permeability of kidney	PeffK	0.158	mm/h	5
Effective permeability of liver	PeffL	0.185	mm/h	5
Steady-state cell-water concentration ratio of kidney	CRssK	6.19	unitless	5
Steady-state cell-water concentration ratio of liver	CRssL	7.28	unitless	5
Renal clearance rate constant	Pbclear	0.994	mm/h	5
Renal reabsorption rate constant	Pbreab	0.425	mm/h	5
Renal efflux rate constant	Pbefflux	0.497	mm/h	5
Absorption rate constant of hepatocyte	Pbabs	0.641	mm/h	5
Association constant of albumin	K_a^{Alb}	2.418×10^4	M^{-1}	5
Association constant of LFABP	K_a^{LFABP}	1.35×10^5	M^{-1}	5

Note: for the association constants, the values represent measured association constants multiplied by the number of binding sites.

2.2.4.1 Prior Distributions

As described above, many parameters are involved in the PBTK model (68 in all). To reduce the computational cost for the MCMC simulation, only the chemical-related parameters to which the model was previously shown to be most sensitive were selected for uncertainty analysis. Based on the sensitivity analysis results (Appendix Table 6), these parameters include the equilibrium association constants, K_a, between PFOA and albumin and LFABP, P_{eff} of blood, liver and kidney, CR_{ss}^{C-W} of liver and kidney, and active transport rates of the four active transport processes discussed above (Table 1). Other parameters, such as physiological parameters, K_a of α2μ-globulin, and P_{eff} of gut, muscle and adipose, were considered as fixed values in the hierarchical Bayesian framework since those parameters were well-studied (low uncertainty) or much has much less influence on the model performance (Appendix Table 6).

Next, as described in Figure 4, the population mean (μ) of each selected parameter was assigned with a log-normal prior distribution with hyperparameter mean (M) and standard deviation (S). The M value for each parameter was derived from the literature, as shown in Table 1; the S value was calculated based on equation: $S = e^{\ln(\sqrt{Cf})}$, where Cf represents confidence factor, which is an intuitive measure of variance in log-normal distributions.[86] For example, a Cf of 2 indicates that 95% of the values lie between ½ and 2 times the median. Given the scarcity of the available data for those parameters, a value of 5 was assigned for them indicating the high uncertainty of their prior distributions (Table 1).

The prior distributions for the population variance of those parameters (Σ^2) were assumed to be inverse gamma distribution: $\Sigma^2 \sim InvGamma(\alpha, \beta)$, where the shape parameter α is set to 3, and the scale parameter β is set to 0.5 based on previous studies.[69, 87] The quantities M, S, α and

β are hyperparameters that embody prior knowledge of the uncertainty and variability of the model parameters.

Finally, considering the high uncertainty and variation of experimental data among different studies (e.g., 1 mg PFOA/kg BW IV and oral dose scenarios from the Kemper[54] and the Kim et al.[58] study), the prior distribution of the experimental error term (σ^2) was modeled as a noninformative uniform distribution with a lower bound of 0.01 and upper bound of 3.3 for all experimental measurements.[69, 87]

2.2.4.2 MCMC Simulation

With the prior information of the population mean (μ) and variance (Σ^2) and experimental error term (σ^2), the joint posterior distribution given the experimental data (Y) can be determined based on Bayes' theorem, as shown in equation 22.[85]

$$p(\theta, \mu, \Sigma^2, \sigma^2 \mid Y) \propto p(Y \mid \theta, \sigma^2) \times p(\theta \mid \mu, \Sigma^2) \times p(\mu) \times p(\Sigma^2) \times p(\sigma^2) \qquad (2\text{-}22)$$

$p(\mu)$, $p(\Sigma^2)$ and $p(\sigma^2)$ are the probability calculated from corresponding prior distributions. $p(\theta \mid \mu, \Sigma^2)$ is the probability of individual chemical-related parameter θ, which is assumed to be log-normally distributed as $log(\theta) \sim N(\log(\mu), \Sigma^2)$. Finally, $p(Y \mid \theta, \sigma^2)$ is the likelihood of the experimental data Y, which is determined based on $log(Y) \sim N(\log(P), \sigma^2)$, where P is the predicted concentration-time data from the PBTK model given a set of parameters (i.e., $P = f(\theta, \psi, E, t)$, as in Figure 4).

Due to the nonlinearity of the PBTK model, it is impossible to acquire an analytical expression for $p(\theta, \mu, \Sigma^2, \sigma^2 \mid Y)$. Instead, the Delayed Rejection Adaptive Metropolis (DRAM) algorithm,[88] a commonly used MCMC sampling technique, was employed to numerically approximate the joint posterior distribution. DRAM was selected because it is highly efficient and has been successfully applied in toxicokinetic models.[69] Here, a total of four Markov chains were constructed in the simulation. For each chain, the total number of iterations was set to 300000, with the first 150000 iterations as a "burn-in" period and the last 50000 iterations as the output samples for posterior distribution analysis.

2.2.4.3 Posterior Analysis and Evaluation

After an MCMC simulation, the convergence of the posterior distributions needs to be verified before further analysis. The Gelman-Rubin diagnostic was used to assess the samples generated from the MCMC method.[89] Specifically, the potential scale reduction factor ($\hat{R}$) was calculated for each parameter distribution. When the posterior distribution becomes stationary, $\hat{R}$ is close to 1. An $\hat{R}$ value of 1.2 or less is considered to be converged for the distribution, as recommended by Gelman et al.[90]

Based on the MCMC output, the posterior quantiles and density plots for the distribution of each selected model parameter were generated for analysis. The PBTK model was then run with the updated parameter distributions and its output was compared with the model results generated with prior parameter distributions. Finally, based on the new predicted concentration-time data, toxicokinetic parameters including half-life, clearance, the maximum PFOA concentration in plasma (C_{max}), and the time required to reach the peak concentration (T_{max}) were calculated and compared with experimental data from other studies.

2.2.5 Experimental Data

Several studies on toxicokinetics and tissue distribution of PFOA in male rats were reported for both oral and IV dosing.[54, 58, 59] In order to compare different administration routes, data on single oral and IV doses of 1 mg PFOA/kg body weight (BW) were chosen for our model evaluation.[54, 58] Moreover, data from a single oral dose of 0.1 mg PFOA/kg BW and IV dose of 0.041 mg PFOA/kg BW were incorporated to verify model performance at low doses.[54, 59] This is of particular toxicological relevance since people in Europe and North America are exposed to low levels of PFOA, with estimated daily intakes in the range of 1 to 130 ng PFOA/kg BW.[91] A total of seven experimental datasets collected from three different studies were used and are briefly described below.

Table 2. Summary of PFOA toxicokinetics studies for male rats.

Administration Routes	Dose Scenarios*	Sampling Time for Tissues	Rat Strain	References
oral	1 mg/kg	Sample from blood at 0.25, 0.5, 1, 2, 4, 8, 12, 16, 24, 36, 48, 72, 96, 120, 144, 168, 192, 240, 288, 336, 384, 432, 480, 528 hours	Sprague-Dawley	Kemper[54]
oral	0.1 mg/kg			
IV	1 mg/kg			
oral	1 mg/kg	Sample from liver, kidney, gut, muscle and adipose after 672 hours		
oral	1 mg/kg	Sample from blood at 6, 12, 24, 48, 96, 144, 192, 240, 288 hours; Sample from liver and kidney after 288 hours	Sprague-Dawley	Kim et al.[58]
IV	1 mg/kg			
IV	0.041 mg/kg	Sample from blood at 5, 15, 45, 90, 120, 150, 210, 270, 300 minutes; Sample from liver and kidney after 120 minutes	Wistar	Kudo et al.[59]

*All experiments used single doses.

As summarized in Table 2, we extracted four datasets from the Kemper study,[54] in which toxicokinetics and distribution of PFOA at different dose levels were investigated for both male and female rats. The first two datasets are the toxicokinetic data for male rats administered by two routes: oral and IV dose. Specifically, four male Sprague-Dawley (SD) rats were dosed 1 mg PFOA/kg BW through oral or IV administration, respectively, then blood samples were collected at different times and analyzed. In addition, a third experiment measuring terminal tissue distribution was chosen from the same study for comparison to predicted PFOA levels in each organ. In this experiment, 1 mg ^{14}C-FPOA/kg BW oral dose was administered to four male SD rats. After 28 days, tissue samples were collected for analysis. Finally, the fourth dataset is the toxicokinetic experiment with a dose level of 0.1 mg PFOA/kg BW, where four male SD rats were orally dosed, and the PFOA concentration in blood collected at different times were analyzed.

In the second study conducted by Kim et al.,[58] two datasets, corresponding to IV and oral administration routes, respectively were extracted. Briefly, 1 mg PFOA/kg BW oral and IV dose were administrated to 5 male and 5 female SD rats. Blood samples were collected and analyzed at different time points. At the end tissue samples including liver, kidney, heart, lung, and spleen were collected for analysis.

To further assess our model, we selected another dataset from a third study[59] where four male Wistar rats were intravenously dosed with the low dose of 0.041 mg [1-^{14}C]PFOA/kg BW. PFOA concentration in blood collected at different time points and tissues, including liver, kidney, intestine, testis, spleen, fat, heart, lung, brain, stomach, and carcass, all analyzed after 2 hours, were available.

For all three studies (seven datasets), data were directly taken from tables, where available, or extracted from plots by WebPlotDigitizer tool (https://automeris.io/WebPlotDigitizer/).

2.2.6 Software

The PBTK model and MCMC simulation were programmed in R (https://www.r-project.org/) using mrgsolve (https://mrgsolve.github.io/), a package designed for solving ordinary differential equations, for PBTK model development. The MCMC simulation was coded using the FME package, which provides convenient functions for the DRAM algorithm.[92]

2.3 Results

2.3.1 Convergence Diagnosis

The trace plots for Markov chains in MCMC are shown in Appendix Figure 1-3. As indicated, no visible trends or changes were observed in the trace plot for each model parameter, suggesting good convergence of the distribution for each parameter. In addition, Gelman-Rubin diagnostic results (Appendix Table 7) show that all parameters have potential scale reduction factor (PSRF) values between 1.001 and 1.02, with upper confidence limits between 1.002 and 1.056. The multivariate PSRF value, which forms the upper bound of PSRF for any linear combination of the parameters, is 1.09. All PSRF values are less than 1.2, indicating the posterior distributions in MCMC have reached equilibrium and can be used for further analysis.

2.3.2 Posterior Distribution Analysis

Table 3. Percentiles of the prior and posterior distribution for each parameter.

Parameters	Prior Distribution			Posterior Distribution		
	2.50%	**50%**	**97.50%**	**2.50%**	**50%**	**97.50%**
PeffB	0.037	0.179	0.868	0.560	0.791	0.865
PeffK	0.033	0.158	0.764	0.038	0.186	0.672
PeffL	0.038	0.185	0.898	0.042	0.192	0.827
CRssL	1.504	7.280	35.246	1.634	7.720	32.756
CRssK	1.279	6.190	29.969	1.521	6.776	26.044
Pbclear	0.205	0.994	4.810	0.304	1.705	4.563
Pbreab	0.088	0.425	2.057	0.094	0.228	1.423
Pbabs	0.132	0.641	3.102	0.138	0.337	2.273
Pbefflux	0.103	0.497	2.405	0.108	0.267	1.233
$\mathrm{K}_\mathrm{a}^\mathrm{Alb}$	4.994×10^3	2.418×10^4	1.171 ×10^5	2.28×10^4	3.582×10^4	5.686×10^4
$\mathrm{K}_\mathrm{a}^{LFABP}$	2.788×10^4	1.35×10^5	6.536×10^5	3.289×10^4	1.439×10^5	4.769×10^5

The percentiles (2.5%, 50% and 97.5%) and density plot comparisons between prior and posterior distribution for each model parameter are shown in Table 3 and Figure 5, respectively. After updating with experimental data, the posterior distributions of the population mean for all parameters were substantially narrower than their prior distributions, indicating that the uncertainties of those parameters were substantially reduced. In addition, an obvious shift was observed in the density plot for some parameters (e.g., PeffB and $\mathrm{K}_\mathrm{a}^\mathrm{Alb}$ in Figure 5). The percentiles of distributions also showed significant changes (defined by larger than ± 20% of prior values[93]) between prior and posterior for some parameters. Specifically, the posterior median of the effective permeability of blood (PeffB) was 4.4 times higher than its prior median. In addition, the posterior median of the $\mathrm{K}_\mathrm{a}^\mathrm{Alb}$ value (i.e., the association constant of albumin multiplied by the number of binding sites) is 3.582×10^4 M^{-1}, which is increased by 48% compared to its prior median. Finally, the posterior medians of all the active transport parameters were substantially different from their prior values: after updating with experimental data, the renal clearance rate constant (Pbclear) increased by 71.5%, the absorption rate constant of hepatocyte (Pbabs), renal reabsorption rate (Pbreab) and renal efflux rate constant (Pbefflux) decreased by 47.4%, 46.4% and 46.3%, respectively.

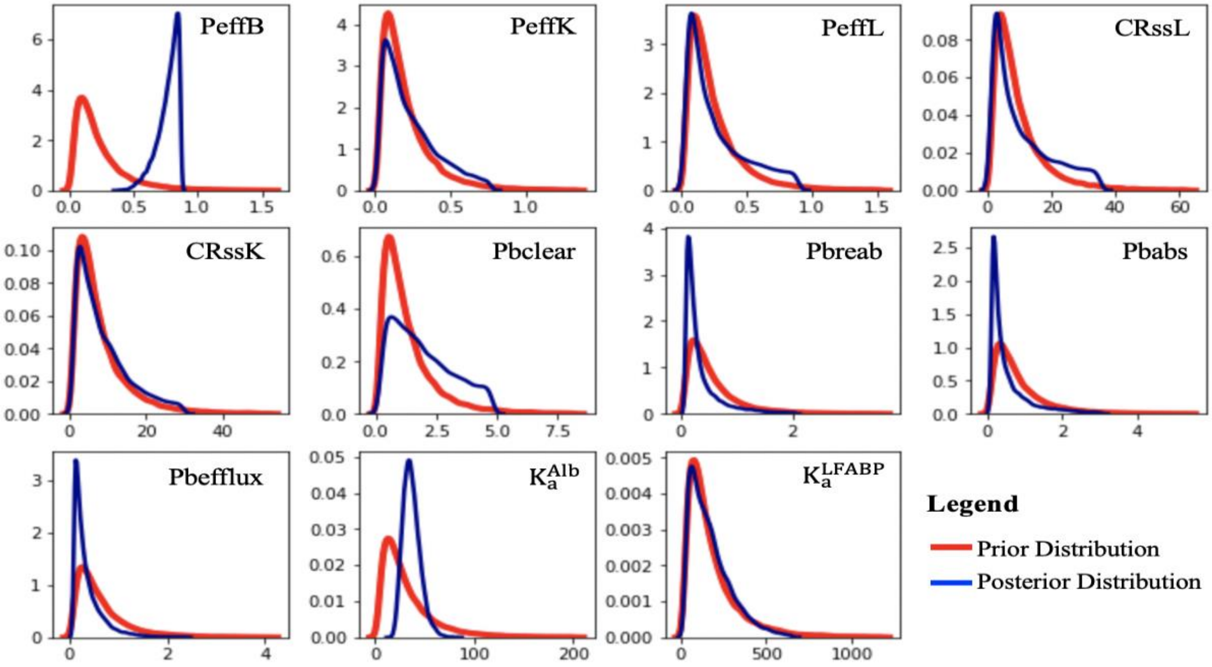


Figure 5. Density plots of the prior and posterior distribution for each parameter.

2.3.3 Model Evaluation

With the generated posterior parameter distributions, the PBTK model was used to simulate the plasma toxicokinetics and tissue distribution of PFOA in male rats and the model was evaluated by comparing with both experimental toxicokinetic data and model predictions based on prior parameter information.

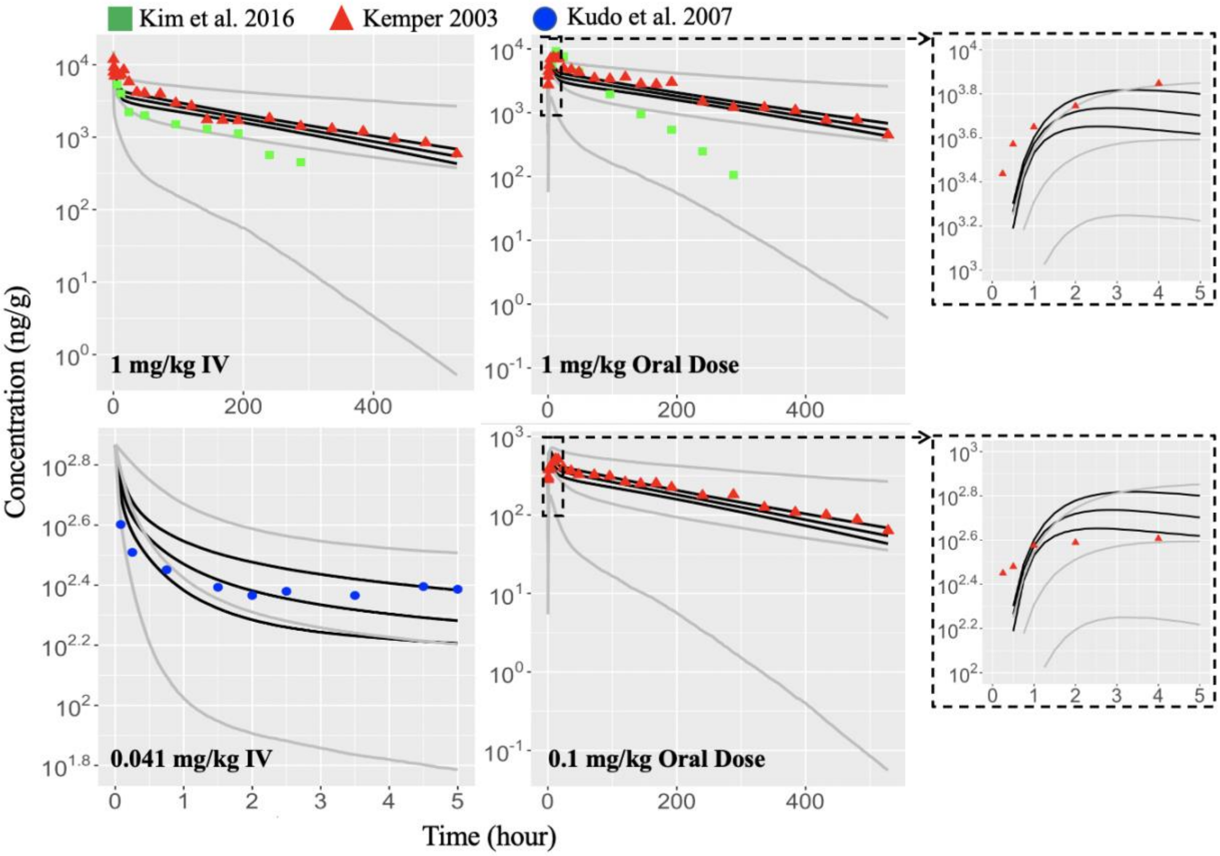


Figure 6. PFOA toxicokinetics in plasma under different dose scenarios.

The grey line represents model results using prior parameter distributions; the black line is with the posterior distributions. The upper, middle, and lower lines indicate the 97.5th, 50th, and 2.5th percentiles of the predicted results, respectively. Red triangles, green squares, and blue circles represent the data sets extracted from the works of Kemper 2003, Kim et al. 2016, and Kudo et al. 2007, respectively. The first 5 hours time-course behavior for oral dose was zoomed to show its up trend at the beginning phase.

2.3.3.1 Plasma toxicokinetics

As shown in Figure 6, all experimental data sets fall within the 95% range of the prior model predictions. Both prior and posterior model predictions indicated similar time-course behavior to the experimental data. However, the 95% range of the posterior prediction (black line) was substantially smaller than that of the prior prediction (grey line), demonstrating a significant decrease in the model uncertainty. In addition, most experimental data fall within the 95% range of the posterior prediction of PFOA concentration in plasma, except for data from the Kim et al study,[58] which show a higher elimination rate. It is worth noting that even under the same dose scenarios (e.g., 1 mg PFOA/ kg BW IV and oral dose), the PFOA concentration profiles are quite different between Kim et al[58] and the Kemper study[54], illustrating the significant variation that can be found among different experimental studies for PFOA toxicokinetics.

Finally, based on the predicted PFOA concentration profiles in plasma, different toxicokinetic parameters were estimated and compared with the experimental results from the Kemper study.[54] As shown in Table 4, in comparison with experimental data, the posterior model results demonstrate much improvement from the prior model for the half-life, clearance of PFOA, and maximum plasma concentration of PFOA (C_{max}); the posterior predicted values for those toxicokinetic parameters fall well within a factor of 1.5 of the experimental data for three different dose scenarios. The half-life seems to be independent of dose scenarios and is calculated as 7.90 days, which is in very good agreement with the half-life values from other experimental studies (range from 5.63 to 15 days).[56, 94-96] The posterior clearance of PFOA is also very similar under different dose cases and has an average value of 25.70 mL/day/kg, which falls within the range of other experimental data (from 21.5 to 50.5 mL/day/kg).[94, 95, 97] For the parameter of the time required to reach the maximum concentration (T_{max}), the prior and posterior model predictions are underestimated by a factor of around 2 and 3.5, respectively.

Table 4. Comparison of toxicokinetic parameters between model prediction and experimental data.

Dose Scenario		Half-life (day)	Clearance (mL/day/kg)	C_{max} (ng/g)	T_{max} (h)
0.1 mg/kg oral	Prior	14.739 ± 15.547	69.310 ± 71.964	424.651 ± 147.354	5.154 ± 3.319
	Posterior	7.900 ± 0.662	26.948 ± 1.715	546.219 ± 50.851	2.874 ± 0.261
	Experiment	8.41 ± 1.56	23.10 ± 5.76	598 ± 127	10.25 ± 6.45
1 mg/kg oral	Prior	15.223 ± 19.857	65.829 ± 71.657	4245.631 ± 1463.333	5.095 ± 3.111
	Posterior	7.912 ± 0.673	25.175 ± 1.545	5479.220 ± 531.690	2.885 ± 0.272
	Experiment	5.76 ± 1.33	20.9 ± 3.79	8431 ± 1161	9.00 ± 3.83
1 mg/kg IV	Prior	15.410 ± 17.516	68.775 ± 91.513	-	-
	Posterior	7.894 ± 0.668	24.990 ± 1.553	-	-
	Experiment	7.73 ± 0.82	21.51 ± 1.97	-	-

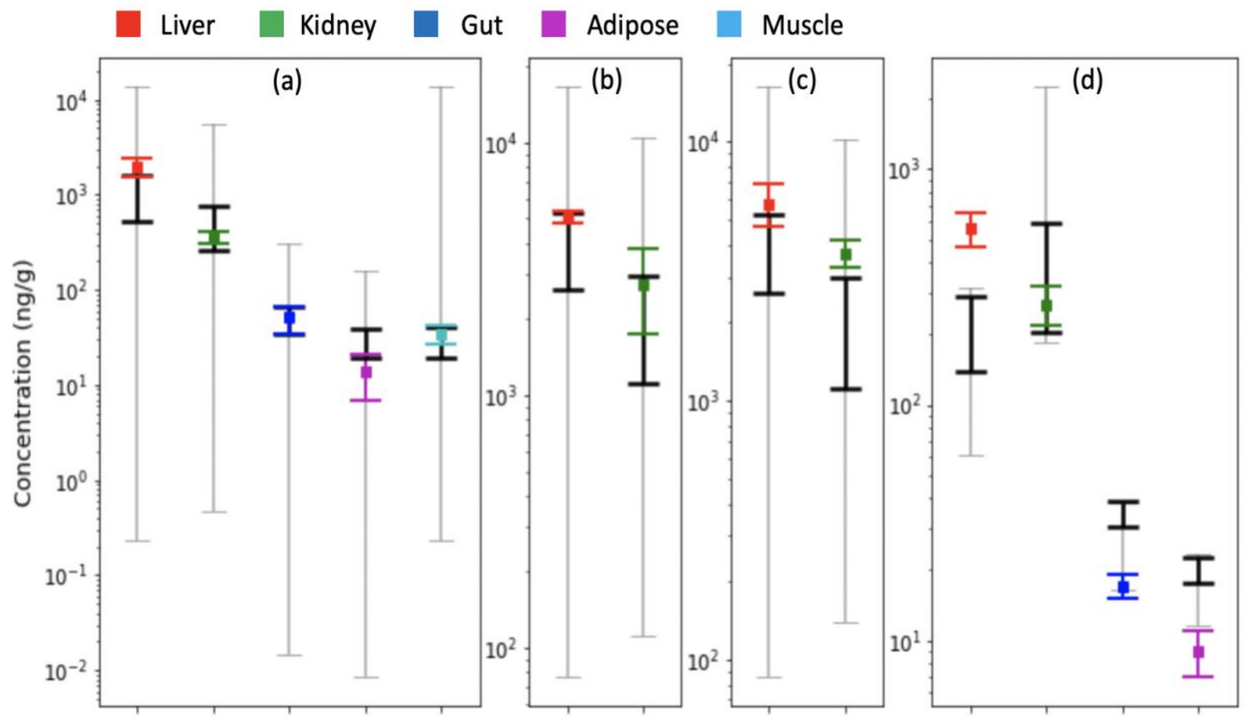


Figure 7. PFOA terminal tissue distribution under different dose scenarios.

(a) 28 days after 1 mg PFOA/kg BW oral dose; (b) 12 days after 1 mg PFOA/kg BW IV dose; (c) 12 days after 1 mg PFOA/kg BW oral dose; (d) 2 h after 0.041 mg PFOA/kg BW IV dose. The grey and black lines represent the 95% range of the model predictions using prior and posterior parameter distributions, respectively. Color bars are experimental data sets from different studies (a: Kemper 2003, b and c: Kim et al. 2016, and d: Kudo et al. 2007) All experimental data are shown as mean ± standard deviation.

2.3.3.2 Tissue Distribution

Figure 7 shows the comparison of PFOA tissue distribution between model predictions (with both prior and posterior parameter distributions) and experimental results under different dose scenarios. As indicated in Figure 7, our model was able to successfully predict the tissue distribution patterns for PFOA in long-term simulations (i.e., after 12 or 28 days): liver > kidney > gut > muscle ≈ adipose. For a short-term dosing scenario (e.g., 2 hours), the predicted PFOA concentration in liver was significantly lower than the measured data. Similar to the plasma toxicokinetics results, the uncertainty for posterior model predictions was reduced substantially compared with the prior model. Most measured PFOA concentrations in each tissue fall well within or overlap with the 95% prediction range, except for the data from the Kudo et al. study.[59]

A further comparison was performed between the mean of experimental data in different tissues and the mean of the model predictions. Both prior and posterior models are able to predict most of PFOA tissue distributions within a factor of 4 of the experimental data. However, the posterior model indicates better performance compared with prior model results. For long-term simulations, the posterior model predictions are well within a factor of 2 of measured concentrations for both oral and IV dose. For short-term dosing (i.e., the Kudo et al. study[59]), the hepatic PFOA concentration was underestimated by the PBTK model, but is within a factor of 2.6.

2.4 Discussion

In this study we developed a permeability-limited PBTK model that explicitly considers PFOA-protein interactions for toxicokinetics and distribution of PFOA in male rats. In addition, the hierarchical Bayesian framework with MCMC simulation was employed to reduce the uncertainty of the model and improve its robustness. With the help of the Bayesian framework, not only were the uncertainties of the posterior parameters substantially reduced, but the PBTK model predictions also became more reliable and meaningful. For example, the toxicokinetic parameters such as half-life and clearance of PFOA estimated with posterior parameters are well within a factor of 1.5 of the experimental data, while the prior calculated toxicokinetic parameters fall within a factor of 1.8 to 3.2 of the experimental data. The good agreement between the simulation results and experimental data illustrates our model's ability to predict the toxicokinetics and tissue distribution of PFOA in rats. Although the PBPK model with posterior parameters demonstrates better performance than the prior model, it is worth pointing out that this is because the posterior model used the experimental data to update the parameter distributions. In other words, the posterior model performance relies on the accuracy of the available experimental datasets. Given the substantial difference between the Kim et al[58] and Kemper studies,[54] if one of those studies turns out to be unreliable, it may be tuning the model in the wrong direction. On the other hand, the prior PBPK model, while it has high uncertainty, relies on no *in vivo* toxicokinetics data and thus could be less subject to bias.

2.4.1 Molecular Mechanisms Driving PFOA Toxicokinetics

With the Bayesian statistical framework, the PBTK model provides more insight into the molecular mechanisms that result in the observed PFOA toxicokinetics. As indicated in Table 3, the posterior median of the association constant for albumin (K_a^{Alb}) increased substantially compared to its prior value. It is worth pointing out that only one study was available for the prior knowledge of K_a^{Alb} in rats, and its association constant value (3.1×10^3 M^{-1} × 7.8 binding sites[39]) is much smaller compared with the Ka values in humans (e.g., 3.12×10^4 M^{-1} × 13 binding sites[98] and 1.26×10^{4} M^{-1} × 2.4 binding sites[66]) and bovines (e.g., 4.36×10^{4} M^{-1} × 1 binding sites[99]). Both our model prediction and the comparison with other experimental data seem to indicate the current $\mathrm{K_a^{Alb}}$ value for rat is a little low and more studies are needed to measure $\mathrm{K_a^{Alb}}$ for PFOA with rat serum albumin.

Another important insight is about the renal elimination of PFOA. From Table 4, compared with the prior half-life parameter (a mean of around 15 days), the posterior values (7.9 days) decreased substantially, indicating an increase in the renal elimination of PFOA. The major reason for this is due to the significant increase in the renal clearance rate constant (Pbclear) and the decrease in both the renal reabsorption rate (Pbreab) and efflux rate (Pbefflux), as shown in Table 3. All those active transport processes were facilitated by different transporters. Although a total of five transporters were considered for the renal elimination process, other transporters such as Oatp4c1 and multidrug resistance-associated proteins (Mrps) located at the proximal tubular cells were not included due to limited information on their transport kinetics.[62] However, our model results indicate that those transporters may have the potential to significantly affect the elimination of PFOA, and more *in vitro* data are needed to evaluate that possibility.

Finally, our model performed very well for long-term dosing simulations, however, in the short-term dosing case (i.e., the 2-hour experiment from Kudo et al.[59]), the PFOA concentration in the liver was substantially underestimated by the model. This could be attributed to cellular membrane binding of PFOA at the beginning phase of distribution to the liver. In fact, Kudo et al.[59] showed that 2 hours after dosing, around 97% of PFOA was found in the membrane fraction. Therefore, PFOA might bind to some membrane components (e.g., protein or phospholipids[100]), which slows down the distribution of PFOA to liver in a short period. In long-term simulation, it seems the membrane binding of PFOA has a negligible effect on the tissue distribution (Figure 7).

2.4.2 Model Limitations

The first limitation on the Bayesian framework is that the prior knowledge is very limited for some model parameters, especially those related to protein binding and active transport processes. For example, the active efflux transporter Mrps, which is located at both the basolateral and apical membranes of proximal tubular cells, are dominant in female rats and could be responsible for the substantial gender difference in PFOA elimination between male and female rats.[62] However, due to the lack of information on the transport kinetics of Mrps, a female rat model for PFOA was not considered in this study. In addition, the computational cost of MCMC simulations is very large, especially for a complex PBTK model. In this study, all the physiological parameters were fixed during MCMC simulation to reduce the computational burden, so the opportunity to refine all parameters in the model was missed.

2.4.3 Call for Data

More data are required to further improve the PBTK model and generalize it to other species and other PFAS. First, data are needed on more PFAS-protein interactions, such as the transporter Mrps, which has the potential to significantly affect PFAS elimination, but for which very limited information is currently available. Given the importance of the equilibrium association constant of albumin in the PBTK model, more accurate measurements are also necessary for model validation. PFAS-protein interaction data could be obtained through *in vitro* studies or estimated with molecular modeling tools (e.g., molecular docking and molecular dynamics). In addition, measurements of the membrane permeability for different tissues (e.g., blood vessels, liver, and kidney) are needed to validate the estimated posterior distributions for those parameters (Table 3). Those measurements could be obtained using cell-based *in vitro* experiments, as demonstrated in the Weaver et al. study.[80] Finally, more *in vivo* toxicokinetic data on PFAS are needed for Bayesian analysis of the PBTK model. As shown in Figure 3, even under the same dose scenarios, there is a huge difference between the toxicokinetics data from Kim et al.[58] and Kemper[54] (e.g., under the same 1 mg/kg oral dose scenario, after 12 days, the PFOA concentration in plasma of the Kemper study is 10 times higher than that of the Kim et al. study). More experimental data are needed to reduce the variability in observations, as well as to better understand actual inter-individual and intra-population variability.

In conclusion, this Chapter focused on the PBTK model, which serve as the top level of the three-level hierarchical framework for PFAS risk assessment (Figure 2). Specifically, we developed a permeability-limited PBTK model that can be successfully used to predict the toxicokinetics and tissue distribution of PFOA in male rats. With the help of the hierarchical Bayesian framework, not only were the uncertainties of the posterior parameters substantially reduced, but the PBTK model predictions also became more robust: with the posterior parameters, most of the predicted plasma toxicokinetic parameters (e.g., half-life) and tissue distributions fell well within a factor of 2.0 of the experimental data. In addition, the PBTK model could provide insights into the molecular mechanisms that result in the observed PFOA toxicokinetics: PFAS-protein binding, membrane permeability and active transport. As discussed above, the large difference in the optimized K_a^{Alb} shows that more especially species-specific data are needed for PFAS-protein binding, which is challenging given the large number of PFAS. Therefore, reliable methods are needed to predict these interactions, and Chapter 3.0 will try to address this persistent gap using molecular modeling.

3.0 Development of a Molecular Dynamics Workflow to Predict Relative Protein Affinity for Per- and Polyfluoroalkyl Substances (PFAS)

This chapter is reproduced in part with permission from:

Cheng, W. and Ng, C. A. Predicting relative protein affinity of novel per- and polyfluoroalkyl substances (PFASs) by an efficient molecular dynamics approach. *Environmental Science & Technology* 2018, 52 (14), 7972-7980. Copyright 2018 American Chemical Society. [https://pubs.acs.org/doi/abs/10.1021/acs.est.8b01268]

This chapter is also reproduced in part based on:

Cheng, W., Doering, J. A., LaLone, C. and Ng, C. A. Integrative Computational Approaches to Inform Relative Bioaccumulation Potential of Per-and Polyfluoroalkyl Substances Across Species. *Toxicological Sciences*, 2021, 180 (2), 212-223. By permission of Society of Toxicology.

With the phasing out of long-chain per- and polyfluoroalkyl substances (PFAS), a wide variety of alternative PFAS have increased production to fill market demand. However, very little is known about the bioaccumulation potential of these replacement compounds. Here, we developed a modeling workflow that combines molecular docking and molecular dynamics simulation techniques to estimate the relative binding affinity of PFAS for liver-type fatty acid binding protein (LFABP). To demonstrate the power of the workflow, we first tested 15 legacy and replacement PFAS and two well-studied LFABPs (i.e., hLFABP and rLFABP for human and rat LFABP) to inform the bioaccumulation potential for novel PFAS. Moreover, we probed the bioaccumulation potential across different species by examining LFABP interactions across 7 different species (i.e., human, rat, chicken, zebrafish, rainbow trout, fathead minnow, and Japanese medaka) for 9 PFAS. The predicted results were evaluated by comparing with experimental data extracted from three different studies. There was good correlation between predicted free energies of binding and measured binding affinities, with correlation coefficients of 0.97, 0.79, and 0.96, respectively. With respect to replacement PFAS, our results suggest that EEA and ADONA are at least as strongly bound to rLFABP as perfluoroheptanoic acid (PFHpA), and to hLFABP as perfluorooctanoic acid (PFOA). For F-53 and F-53B, both have similar or stronger binding affinities than perfluorooctane sulfonate (PFOS). Given that interactions of PFAS with proteins (e.g., LFABPs) are important determinants of bioaccumulation potential in organisms, these alternatives could be as bioaccumulative as legacy PFAS, and are therefore not necessarily safer alternatives to long-chain PFAS. For bioaccumulation potential across species, Human, rat, chicken and rainbow trout had similar binding affinities to one another for each PFAS, whereas Japanese medaka and fathead minnow had significantly weaker LFABP binding affinity for some PFAS. This result indicates that human, rat, chicken, zebrafish or rainbow trout seem to be better representative species of the higher range of vertebrate bioaccumulation potential of PFAS than Japanese medaka and fathead minnow.

3.1 Introduction

Per- and polyfluoroalkyl substances (PFAS) are a diverse group of compounds that have been used in a broad range of industrial and consumer products (e.g., fire-fighting foams, food contact materials, and apparel).[1-6] Due to their persistence, bioaccumulation and toxicity, the long-chain PFAS have been phased out for the majority of uses.[24, 25, 57] To take their place, manufacturers have started using alternatives that include shorter-chain homologues of PFOA and PFOS, as well as perfluoroether carboxylic acids (PFECAs) and perfluoroether sulfonic acids (PFESAs).[24, 25, 101, 102] A number of those fluorinated alternatives used in industrial and consumer products have been identified in a recent review paper by Wang et al.[25] And these alternatives have also been widely detected in the environment and organisms.[103-106] However, the identity and frequency of use of many other alternative PFAS remains largely unknown, leading scientists to employ extensive non-target analytical techniques to puzzle out the structures present in complex environmental mixtures of PFAS.[26, 107-109] In addition, limited information is known on the potential impacts of alternative PFAS on humans and the environment; especially, there is a lack of information on the bioaccumulation potential and toxicity of PFECAs and PFESAs.[25]

Given the large number of PFAS and the scarce knowledge about their potential hazards, a rapid and reliable method to predict the behavior of these chemicals in the environment and organisms would be of great benefit. In our previous work, we developed mechanistic physiologically based toxicokinetic (PBTK) models that explicitly consider binding with serum albumin, liver-type fatty acid binding protein (LFABP), and organic anion transporters to predict the bioaccumulation of PFCAs and PFSAs in different tissues of both fish and rat.[81, 82] The success of our models demonstrated that the interaction of PFAS with proteins plays an essential role in determining their bioaccumulation potential in organisms, and thus could be used as a proxy for bioaccumulation assessment. However, the protein binding parameters used to build the models were limited to a small subset of PFAS (e.g., PFOA and PFOS).[13]

To provide insights into the bioaccumulation potential of novel PFAS and generate more protein binding parameters for PBTK models, we proposed an *in silico* method based on molecular dynamics (MD) simulations to predict PFAS-protein interactions. Specifically, we employed the molecular mechanics combined with Poisson-Boltzmann surface area (MM-PBSA) continuum solvation method to calculate binding affinities between ligands (i.e., PFAS) and proteins. As a starting point, we focused on LFABPs in this study because of their available 3-dimensional crystal structures and experimental binding affinity data with different PFAS, which can be used for method evaluation. In the MM-PBSA method, the free energy of binding, ΔG_{bind}, for a chemical reaction: P + L = PL (P denotes the protein and L the ligand) is calculated from:

$$\Delta G_{bind} = G^{PL} - G^{P} - G^{L} \tag{3-1}$$

where the free energy of a state (i.e., G^P, G^L, and G^{PL}) is derived from post-processing an ensemble of representative protein-ligand snapshots generated from MD simulations.[44] This method is more computationally efficient than rigorous alchemical perturbation methods (e.g., free energy perturbation and thermodynamic integration methods), but more robust compared to molecular docking based on scoring functions.[43] It is worth noting that MM-PBSA is a continuum solvation method and involves several thermodynamics approximations, which makes the absolute binding energies unreliable.[43] However, many studies have demonstrated that MM-PBSA is able to successfully predict the relative binding affinities of ligands.[43, 44, 110-113] Therefore, the primary goal of this study is to rank the binding affinities of PFAS bound to LFABPs. Given the large number of PFAS, an efficient method like MM-PBSA would be of great benefits.

In this chapter, we first developed the following MD-based workflow to estimate ΔG_{bind} for LFABP-PFAS interactions: the initial structure of the LFABP-PFAS complex was generated from molecular docking, a powerful tool to predict the binding mode between a protein and a ligand;[114] based on the complex structure, the MD simulation was then carried out; finally, MM-PBSA was used to calculate the ΔG_{bind}.

Then to test the power of the MD workflow and apply it to inform bioaccumulation potential for alternative PFAS, we considered 15 PFAS with different functional head groups and fluorinated carbon chain lengths including 10 legacy PFAS (7 PFCAs: PFBA, PFPA, PFHxA, PFHpA, PFOA, PFNA, and 2m-PFOA; and 3 PFSAs: PFBS, PFHxS, and PFOS) and 5 alternatives (3 PFECAs: ADONA, GenX, and EEA; and 2 PFESAs: F-53 and F-53B). The 2-dimensional structures of these chemicals are shown in the Appendix Figure 4. The binding affinities of these chemicals were evaluated for 2 different LFABPs (hLFABP and rLFABP for human and rat LFABP, respectively) which have been previously experimentally determined.[37, 115, 116]

Finally, we explore the application of the MD workflow in estimating the bioaccumulation potential across different species. In this part, we selected LFABPs of 7 species (i.e., human, rat, chicken, zebrafish, rainbow trout, Japanese medaka, and fathead minnow) as protein proxies for bioaccumulation assessment. The MD workflow was employed to predict the interactions between those LFABPs and 9 PFAS with different functional head groups and fluorinated carbon chain lengths including (i.e., 6 PFCAs: PFBA, PFPA, PFHxA, PFHpA, PFOA, and PFNA; and 3 PFSAs: PFBS, PFHxS, and PFOS). This application could enhance understanding of LFABP-PFAS interactions across species and thus can inform research and decision-making.

3.2 Materials and Methods

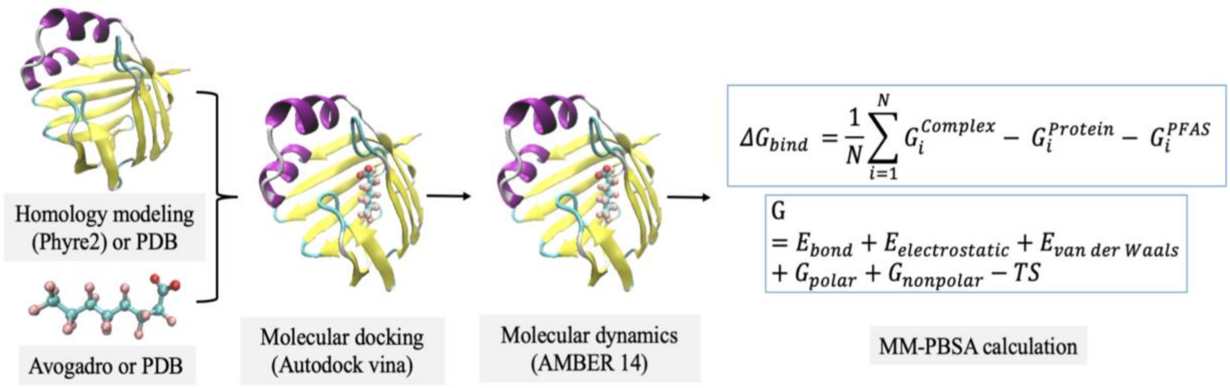

Homology modeling (Phyre2) or PDB
Avogadro or PDB
Molecular docking (Autodock vina)
Molecular dynamics (AMBER 14)
$\Delta G_{bind} = \frac{1}{N}\sum_{i=1}^{N} G_i^{Complex} - G_i^{Protein} - G_i^{PFAS}$
$G = E_{bond} + E_{electrostatic} + E_{van\ der\ Waals} + G_{polar} + G_{nonpolar} - TS$
MM-PBSA calculation

Figure 8. Molecular Dynamics Workflow.

The molecular dynamics workflow was used to estimate the LFABP binding affinity for different PFAS. As shown in Figure 8, the workflow consists of four major steps: curation of structures, molecular docking, molecular dynamics, and molecular mechanics combined with Poisson-Boltzmann surface area (MM-PBSA) calculation.

3.2.1 Curation of Structures

The 3-dimensional crystal structures were obtained from the Protein Data Bank (PDB, http://www.rcsb.org) for hLFABP (PDB code: 3STM[117]) and rLFABP (PDB code: 1LFO[118]). These structures were selected because their high resolution and completeness of key residues, as discussed in our previous docking study of LFABP interaction with PFAS.[119] For LFABP of other species, Phyre2[120] was used to construct 3-dimensional structures because it is one of the most popular protein structure prediction servers and very user-friendly[120]. The protein sequences used to build the 3-dimensional structures are shown in Table 5. We selected the structure with the highest confidence as the output of Phyre2. All protein structures constructed for this study have a confidence of 100%. For PFAS ligands, the 3-dimensional structures of PFOA and PFOS were extracted from 5JID (PDB code for crystal structure of human transthyretin in complex with PFOA[121]) and 4E99 (PDB code for crystal structure of human serum albumin in complex with PFOS[122]), respectively. The other 3-dimensional structures were constructed from scratch using Avogadro (v1.2.0)[123] and exported in pdb file format.

3.2.2 Molecular Docking

All PFAS ligands were docked to LFABPs with Autodock Vina (v1.1.2),[124] as described in our previous study.[119] Briefly, both protein and ligand structures produced above were first preprocessed using AutoDock Tools (v1.5.6),[125] the output pdbqt files were then used for docking. For each protein, the binding site boundaries were determined using the Grid menu in AutoDock Tools. According to the 3-dimensional structure of rLFABP, the binding cavity is a flattened rectangular box (roughly 13 × 9 × 4 Å).[118] Since there is no available dimensions for hLFABP and LFABP of other species, we assumed those LFABPs have similar binding pocket as rLFABP. For rLFABP, although there are two ligands in the binding cavity, they are not independent with each other, that is, the secondary binding site (for the ligand in the solvent-accessible location) would not exist until the primary binding site (for the ligand buried inside the pocket) is filled; while for hLFABP, only one ligand lies in the binding cavity (i.e., the one buried inside the pocket). For simplification, we only considered the primary binding site for those LFABPs. The ligands docked to the protein were assumed to be in their deprotonated forms, given their low pK_a.[8, 126, 127] The docking experiments output binding free energies and docking poses for each ligand-protein complex. The 3 binding modes with lowest energies (strongest associations) and distinct conformations were chosen as initial structures for MD simulations.

To assess the success of the docking experiment, we redocked the ligands from the crystalized complex (i.e., PDB code 3STM and 1LFO) back into their corresponding receptors and measured the root-mean-square deviation (RMSD) between original crystal structure and the docked ligands using PyMol;[128] the results showed that Autodock Vina can successfully predict the bound conformations of the ligands and LFABP with reasonable accuracy (RMSD < 2.5 Å). In addition, our previous study also demonstrated Autodock Vina can redock PFOS to human serum albumin with RMSD smaller than 2 Å.[119]

3.2.3 Molecular Dynamics Simulation

The system setup and simulation of the 90 ligand-protein complexes were performed with the Amber 14 suite.[129] For system setup, the *ff*14SB force field[130] was used for proteins and the general AMBER force field (GAFF)[131] was used for ligands. The atomic partial charges of ligands were derived by AM1-BCC (AM1-bond charge correction), which is an efficient method to reproduce HF/6-31G* RESP charges.[132] The whole complex system was explicitly solvated in a cubic box of TIP3P water molecules with a minimal distance of 12 Å from solute atoms to box edges. Na^+ counterions were added to neutralize the systems. Periodic boundary conditions were employed for all simulations. Long-range electrostatic interactions were handled by the particle mesh Ewald (PME) method.[133] The cutoff for nonbonded interactions was set to 8 Å.

The simulation was carried out by the GPU accelerated *pmemd* module.[134] First, the solvent molecules were subjected to 2000 steps of energy minimization, while the solute was constrained with the harmonic force constant of 500 kcal mol^{-1} $Å^{-2}$ to eliminate nonphysical contact between solute and solvent. Next, the whole system was minimized without restraint for 1500 cycles. Then, the system was heated from 0 to 300 K in 20 ps at constant volume; a weak harmonic force constant of 10 kcal mol^{-1} Å^{-2} was added on the complex. After the heating phase, the density of the system was adjusted to 1 g/cm^{3} at constant pressure (1 bar) for 100 ps with restraint (10 kcal mol^{-1} Å^{-2}) on the complex. Finally, the system was equilibrated at constant temperature (300 K) and pressure (1 bar) for 2 ns, the temperature was controlled by Langevin dynamics with a collision frequency of 2 ps^{-1},[129] and the pressure was controlled by the isotropic position scaling protocol.[129] Under the same condition as in the equilibration phase, the production run for the complex system was performed for 24 ns. The SHAKE bond length constraints were used to allow a larger timestep of 2 fs.[135] The trajectories were sampled at a time interval of 16 ps to ensure each snapshot is statistically independent.[43, 136] All simulations were run on an AMBER GPU Certified MD workstation (Exxact Corporation, CA, USA).

3.2.4 MM-PBSA Calculations

Free energy calculations were conducted using the *MMPBSA.py* program in Amber 14.[44] Specifically, ΔG_{bind} between PFAS ligands and LFABPs were calculated as follows:

$$\Delta G_{bind} = G^{Complex} - G^{LFABP} - G^{PFAS} \qquad (3\text{-}2)$$

where $G^{Complex}$, G^{LFABP}, and G^{PFAS} are the free energies of complex, LFABPs, and PFAS ligands, respectively. The free energy (G) of each state was estimated from the following sum:[137-139]

$$G = < E_{bond} + E_{el} + E_{vdw} + G_{polar} + G_{nonp} - TS > \qquad (3\text{-}3)$$

where the brackets indicate an average over MD trajectories. Inside the brackets, the first three terms are molecular mechanical energy terms for bonded, electrostatic and van der Waals (vdw) interactions, respectively. G_{polar} is the polar solvation free energy, which was calculated using the Poisson-Boltzmann (PB) implicit solvent method (which is a differential equation based on the Poisson continuum dielectric model and Boltzmann distribution for the ions in the solvent).[140] G_{nonp} is the nonpolar contribution, which was determined from a linear relation to solvent-accessible surface area. The last term is the absolute temperature (T) multiplied by the entropy (S), which was estimated by normal-mode analysis using the *nmode* program in Amber 14;[129] Specifically, the entropy change is calculated based on the change in three types of molecular motions (i.e., translational, rotational, and vibrational motion) of the system when ligand binds to the receptor.[141] For the thermodynamic variables to control the calculation, the recommended values for Amber 14 were employed (e.g., the external and internal dielectric constants are 80.0 and 1.0, respectively).[129] Finally, a single trajectory protocol (STP) approach was used for the free energy calculation.[44] That is, the calculation of $G^{complex}$, G^{LFABP}, and G^{PFAA} were based on a single MD trajectory of the complex system, rather than on 3 trajectories generated from 3 separate MD simulations. We employed STP method based on the following considerations. First, since no study is available indicating that the binding of PFAS to LFABPs causes a significant conformation change, we assumed the proteins and ligands are comparable in the bound and unbound states. Second, all ligands in our study have similar chemical structure (i.e., fluorinated carbon chain and functional head group). Finally, The STP method is less computationally expensive, and it also leads to a cancellation of the E_{bond} term in equation 3, which improves the precision of the results tremendously.[43, 44]

As described in the MD simulation section, a total of 1500 independent snapshots (24 ns production divided by the timestep of 16 ps) were generated for an individual complex system. These MD snapshots were evenly divided into 3 groups (i.e., the first 4 ns or 250 snapshots is group one, the second 4 ns or 250 snapshots is group two and so forth, each group could be considered as an independent simulation phase). In each group, the binding free energy was calculated using equation 2 and 3 and then averaged over all snapshots; for the entropy calculation, because it is very computationally expensive,[44] only 10 snapshots in each group were considered for normal-mode analysis. Those 10 snapshots were collected as a subset of the total of 250 snapshots in each group based on an interval of 25 snapshots. All MM-PBSA calculations were carried out on Bridges, part of the Pittsburgh Supercomputing Center (www.psc.edu).

3.2.5 Free Energy Decomposition

To gain insights into the contributions to the binding free energy, energy decomposition on a per-residue basis was performed using the *decomp* program in Amber 14.[44] The per-residue decomposition scheme can decompose calculated free energies into specific residue contributions based on the Poisson-Boltzmann implicit solvent model.[142, 143] The contribution of each residue in LFABP to the total free energy of the complex system was estimated.

3.2.6 Data Analysis

The final binding free energies estimated by MM-PBSA were averaged over 3 MD simulation phases and 3 binding modes for each LFABP-PFAS pair (i.e., a total of 9 ΔG_{bind} values for each LFABP-PFAS pair). The standard error of the mean was then calculated. Moreover, the correlation analysis for predicted free energies versus carbon chain lengths and predicted free energies versus experimental binding affinities were conducted. For comparison and correlation analysis, the experimental data represented as equilibrium dissociation constant (K_d) with units of µM were translated into free energy of binding (ΔG_{bind}, in kcal/mol) as follows:[144, 145]

$$\Delta G_{bind} = RTln\frac{K_d}{c_0} \qquad (3\text{-}4)$$

where R is gas constant (1.987 cal K^{-1} mol^{-1}), T is temperature (which is assumed to be 300 K), and c_0 is the standard state concentration (1 M).

Finally, one-way ANOVA was conducted to test for significant differences among the 7 different species of LFABP (i.e., human, rat, chicken, zebrafish, rainbow trout, Japanese medaka, and fathead minnow) for 9 PFAS ligands (i.e., 6 PFCAs: PFBA, PFPA, PFHxA, PFHpA, PFOA, and PFNA; and 3 PFSAs: PFBS, PFHxS, and PFOS). In addition, multiple comparisons with Tukey's test were performed to identify which groups are significantly different from each other for cross-species effects. The Python package SciPy (https://scipy.org/index.html) and statsmodels (https://github.com/statsmodels/statsmodels) were used for ANOVA and Tukey's test, respectively; and both tests were conducted based on the 9 different ΔG_{bind} values for each LFABP-PFAS pair.

Table 5. Amino acid sequences for different species.

Species	Amino Acid Sequences
Human	MAGSSFSGKYQLQSQENFEAFMKAIGLPEELIQKGKDIKGVSEIVQNGKHFKFTITAGSKVIQNEFTVGEECELETMTGEKVKTVVQLEGDNKLVTTFKNIKSVTELNGDIITNTMTLGDIVFKRISKRIGT
Rat	XMNFSGKYQVQSQENFEPFMKAMGLPEDLIQKGKDIKGVSEIVHEGKKVKLTITYGSKVIHNEFTLGEECELETMTGEKVKAVVKMEGDNKMVTTFKGIKSVTEFNGDTITNTMTLGDIVYKRVSKRI
Chicken	MSFTGKYELQSHENFEPFMKALGLPDDQIQKGKDIKSISEIVQNGNKFKITVTTGSKVMTNEFTIGEECEMELLTGEKAKCIVNMEGNNKLVANLKGLKSVTELNGDTITHTMTKGDLTYKRISKRI
Zebrafish	MAFTGKYQLESHENFEAFMKAVGVPDDEVEKGKDIKSISEIHQDGKDFKVTVTAGTKVILYSFTVGEECELETFTGDRAKTVVQMDGNKLTAFVKGIESVTELDGDTISNTLSFNGIVYKRISRRIS
Rainbow trout	MAFTGKYQLESQENFEPFMKAIGLPDDLIQKGKDIKSVSEIEQNGDHFKVTVTTGTKVMVNSFTVGQEAELETLTGEKIKSTVNLVGNKLMVSLKGIESVTEFNGDTIIATMMLGPIVYKRISKRI
Japanese medaka	MDFNGTWQVYSQENYEEFLRALELSEDIIKLAKDVKPVTEIKQTGNDFVITSKTPGRTVTNSFTIGKEAEISTMDGKKLKCVVNMEGGKLVCKTGKFCHVQEIKGGEMVETMTVGSTTLIRKSKKM
Fathead minnow	VYLQENYEEFLPAIPLPEDIIKLAKDVKPVTEIQQKGNDFTITSKTPGKTVTNSFTVGKEAEITTMDGKKLKCIVKLEGGKLVCNTERFSHIQEIKGGEMVETLTVAGTTMVRKSKKI

Notes: The sequences can be downloaded from PDB website (https://www.rcsb.org/) for human and rat and from NCBI website (https://www.ncbi.nlm.nih.gov/) for other species.

3.3 Results

3.3.1 Method Evaluation

To evaluate the effectiveness of the MD workflow, we conducted correlation analysis between the predicted ΔG_{bind} to the experimental ΔG_{bind} derived from three different studies (Figure 9).[37, 115, 116] For hLFABP, MM-PBSA performance varied between different experimental studies. In comparison with the Sheng et al. study (Figure 10A),[116] the correlation coefficient was excellent ($r = 0.97$), while the correlation analysis between computational results and the Zhang et al.[115] study indicated a coefficient of 0.79 (Figure 10B), which is also acceptable. The binding free energies between rLFABP and PFAS ligands from the simulation correlate very well with experimental data (Figure 10C, $r = 0.96$). The results also show that the predicted absolute binding energies of PFAS are generally lower than corresponding experimental values. However, it should be emphasized that our study is primarily focused on the relative binding affinities of PFAS rather than their absolute binding strengths.

3.3.2 Protein Binding Affinity for Legacy and Novel PFAS

The MD workflow was used to estimate the protein binding affinity for novel PFAS. In this part, the results of the interactions between 15 PFAS (i.e., PFBA, PFPA, PFHxA, PFHpA, PFOA, PFNA, 2m-PFOA, PFBS, PFHxS, PFOS, ADONA, GenX, EEA, F-53 and F-53B) and 2 LFABPs (rLFABP and hLFABP for rat LFABP and human LFABP, respectively) are described.

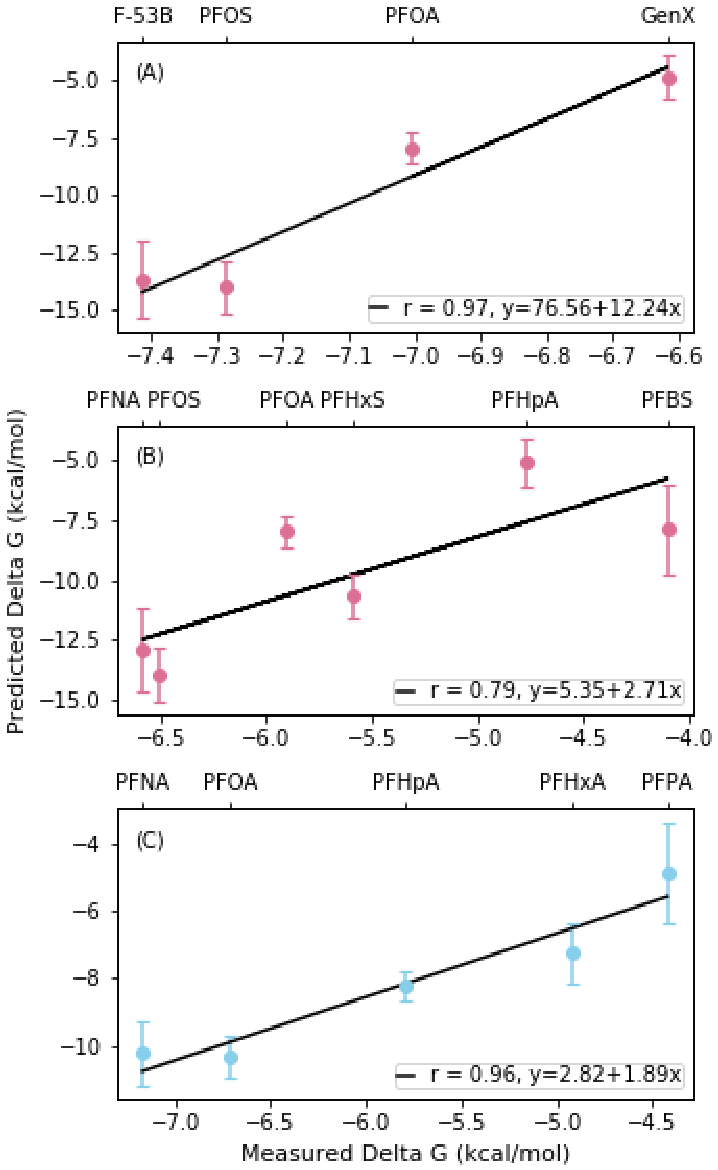


Figure 9. Correlations between the average ΔG_{bind} calculated by MM-PBSA and the experimental values. (A) hLFABP from Sheng et al.,[116] (B) hLFABP system from Zhang et al.,[115] and (C) rLFABP from Woodcroft et al.,[37] Error bars indicate the standard error for predicted ΔG_{bind}.

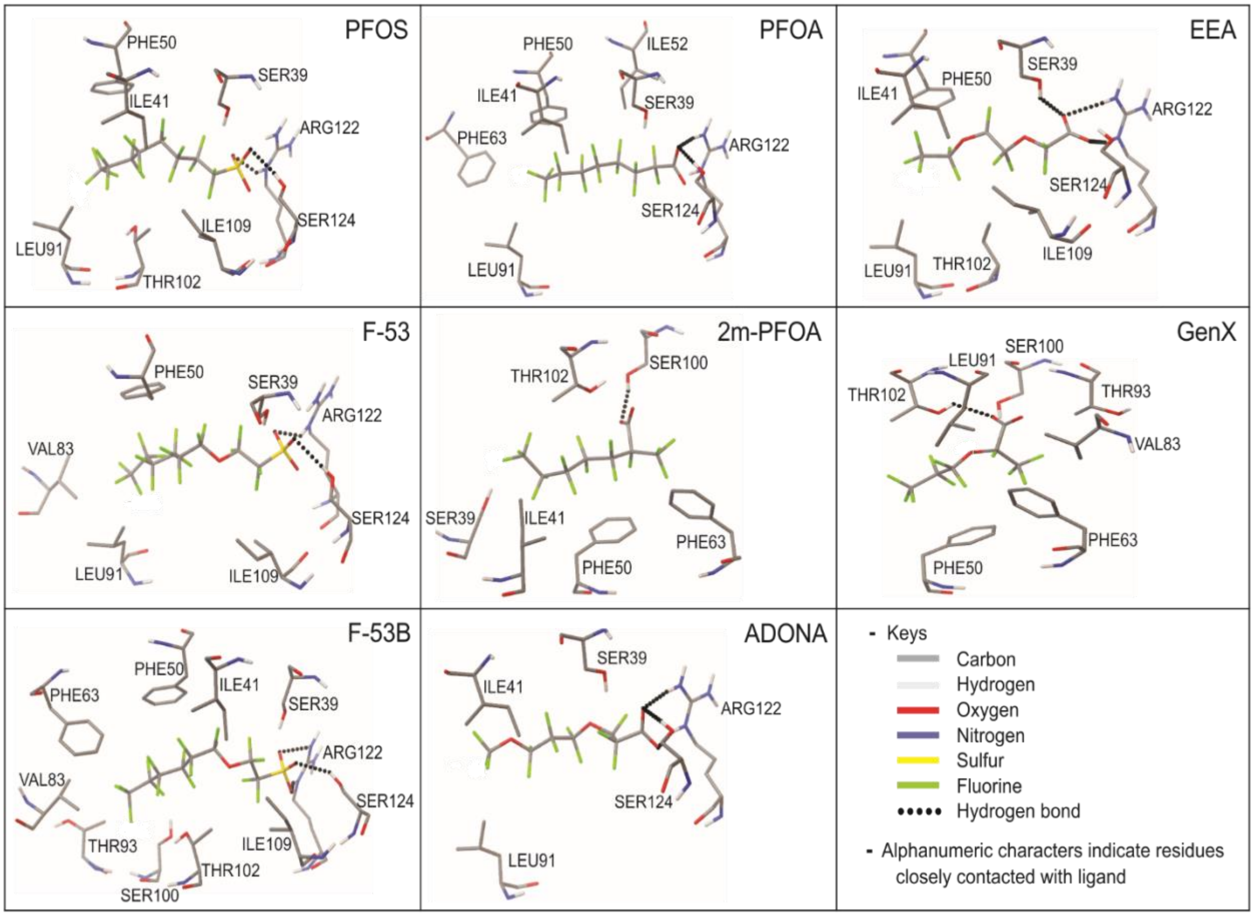


Figure 10. The interactions between human LFABP and PFAS ligands.

3.3.2.1 Molecular Docking

The docking experiment was conducted mainly to predict the interaction between LFABP and PFAS to find potential binding modes.[114] The interaction structures for the best binding mode of each LFABP-PFAS complex for rat and human LFABP are indicated in Figure 10, and Appendix Figures 5 and 6. As shown, the interactions between PFAS and the two LFABPs (rLFABP and hLFABP) are substantially different. For hLFABP, the residues closely interacting with PFAS include ARG 122, SER 39, SER 124, PHE 50, ILE 109, ILE 41, and LEU 91; while for rLFABP, the close contact residues consist of ARG 122, TYR 120, MET 74, LEU 28, and TYR 54 (except for the rLFABP-PFBS interaction, where the close contact residues are ARG 122, SER 39, and SER 124). With respect to hydrogen (H) bonding interactions (Table 6), all PFAS ligands formed H bonding with hLFABP, and most interactions occurred between ligands and residues of ARG 122, SER 124, or SER 39. On the other hand, only a few instances of H bonding occurred between PFAS ligands and rLFABP, and the major residue participating in H bonding interactions is TYR 120.

In addition, for both LFABPs, the predicted binding modes were similar among ligands with different functional head groups (i.e., carboxyl and sulfonate groups). The binding of alternative PFAS (which contain ether groups in their structures) and legacy PFAS (which contain no ether group) to LFABPs show little difference in terms of conformations. The only notable differences were observed for 2m-PFOA and GenX, both of which have branched structures. As indicated in Figure 10, the carboxyl group in 2m-PFOA and GenX mainly interacted with THR 102 and SER 100, not ARG 122, SER 124 and SER 39, which were the major residues interacting with the head group of other PFAS.

Table 6. The protein residues interacting with the PFAS ligands through H-bond and those having dominant energy contributions to ΔG_{bind} of each LFABP-PFAS complex.

Ligands	Human LFABP		Rat LFABP	
	H-bond interaction	Largest energy contribution	H-bond interaction	Largest energy contribution
PFBA	ARG 122, SER 124	ARG 122, SER 39, ILE 52	-	SER 57, LYS 58, LYS 32
PFPA	ARG 122, SER 124	ARG 122, VAL 83, PHE 50	-	ARG 122, TYR 55, ILE 53
PFHxA	ARG 122, SER 124	ARG 122, SER 39, SER 124	TYR 120	ARG 122, ILE 53, LYS 58
PFHpA	ARG 122, SER 124	ARG 122, SER 39, ILE 52	-	ARG 122, ILE 60, MET 74
PFOA	ARG 122, SER 124	ARG 122, SER 39, ILE 52	-	ARG 122, TYR 55, ILE 60
PFNA	ARG 122, SER 124	ARG 122, SER 39, ILE 52	-	ARG 122, ILE 60, ILE 53
PFBS	ARG 122, SER 39	ARG 122, SER 124, LEU 9	ARG 122, SER 39	ARG 122, SER 100, LEU 71
PFHxS	ARG 122	ARG 122, SER 124, SER 39	-	ARG 122, ASN 111, LEU 51
PFOS	ARG 122, SER 124	ARG 122, SER 124, ILE 52	TYR 120	ARG 122, ILE 60, ILE 53
EEA	ARG 122, SER 39	ARG 122, SER 39, ASN 111	-	ARG 122, MET 74, ILE 60
GenX	THR 102	ARG 122, ASN 111, THR 73	-	ARG 122, MET 74, ILE 53
ADONA	ARG 122, SER 124	ARG 122, SER 39, SER 124	-	ARG 122, MET 74, TYR 55
2m-PFOA	SER 100	ARG 122, SER 100, ASN 111	-	ARG 122, TYR 120, ILE 60
F-53	ARG 122, SER 124	ARG 122, PHE 50, SER 39	TYR 120	ARG 122, SER 124, ILE 53
F-53B	ARG 122, SER 124	ARG 122, SER 124, SER 39	TYR 120	ARG 122, TYR 55, ILE 60

3.3.2.2 MM-PBSA

The average ΔG_{bind} calculated based on MM-PBSA and five energy components (i.e., vdw, electrostatic, polar and nonpolar solvation energy, and entropy) of each LFABP-PFAS pair for rat and human LFABP are present in Appendix Tables 8 and 9. As indicated, the predicted free energies of ligands interacting with hLFABP and rLFABP range from -15.76 to -2.21 kcal/mol and from -11.26 to -3.74 kcal/mol, respectively. For each ligand, the predicted binding affinities with hLFABP are generally higher than that of rLFABP. For both LFABPs, the most significant contribution to the binding affinity is the electrostatic interaction, but this large change of electrostatic interaction upon binding is compensated by the polar solvation energy. The nonpolar solvation energies are very small and show a minor variation among ligands, thus having insignificant contribution to the ΔG_{bind}.

Figure 11 shows the distribution of vdw, the sum of electrostatic and polar solvation energy, and entropy changes for each ligand-protein system. The sum of electrostatic and polar solvation energy is shown instead of the separate contributions because both energy terms are strongly anti-correlated (r = -0.96). An obvious pattern was observed between vdw and carbon chain length: as carbon chain length increases, the vdw interaction energy decreases. The entropy term also indicated a similar trend, but the relationship is not as strong as vdw. With respect to electrostatic and polar solvation energy, wild fluctuations were observed in both LFABP systems. In particular, the sum of electrostatic and polar solvation energy for PFHxA bound to hLFABP is much lower compared with other ligands.

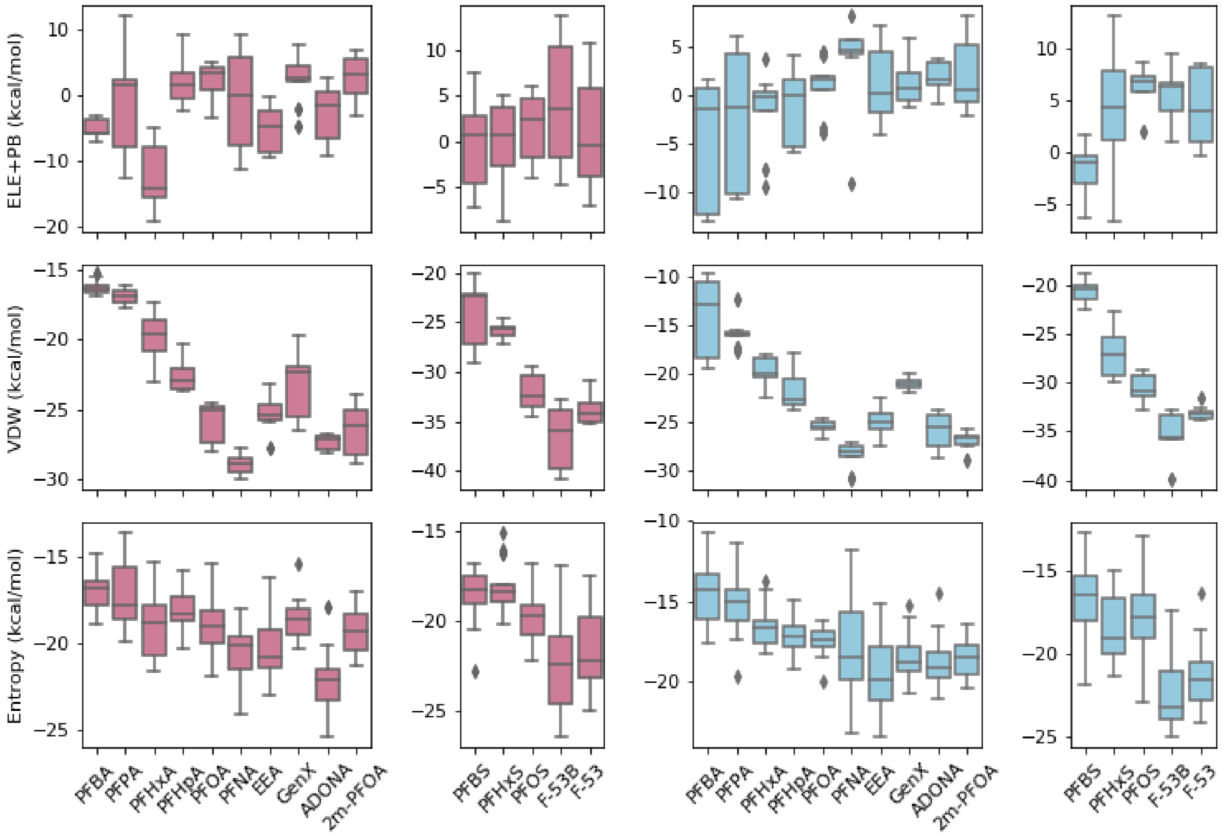


Figure 11. The distributions of the energy components of ΔG_{bind}.

Those energy components include the sum of electrostatic interactions and polar solvation free energy (ELE + PB), van der Waals energy (vdw), and entropy changes upon binding. Pink represents the hLFABP system, while blue is the rLFABP system.

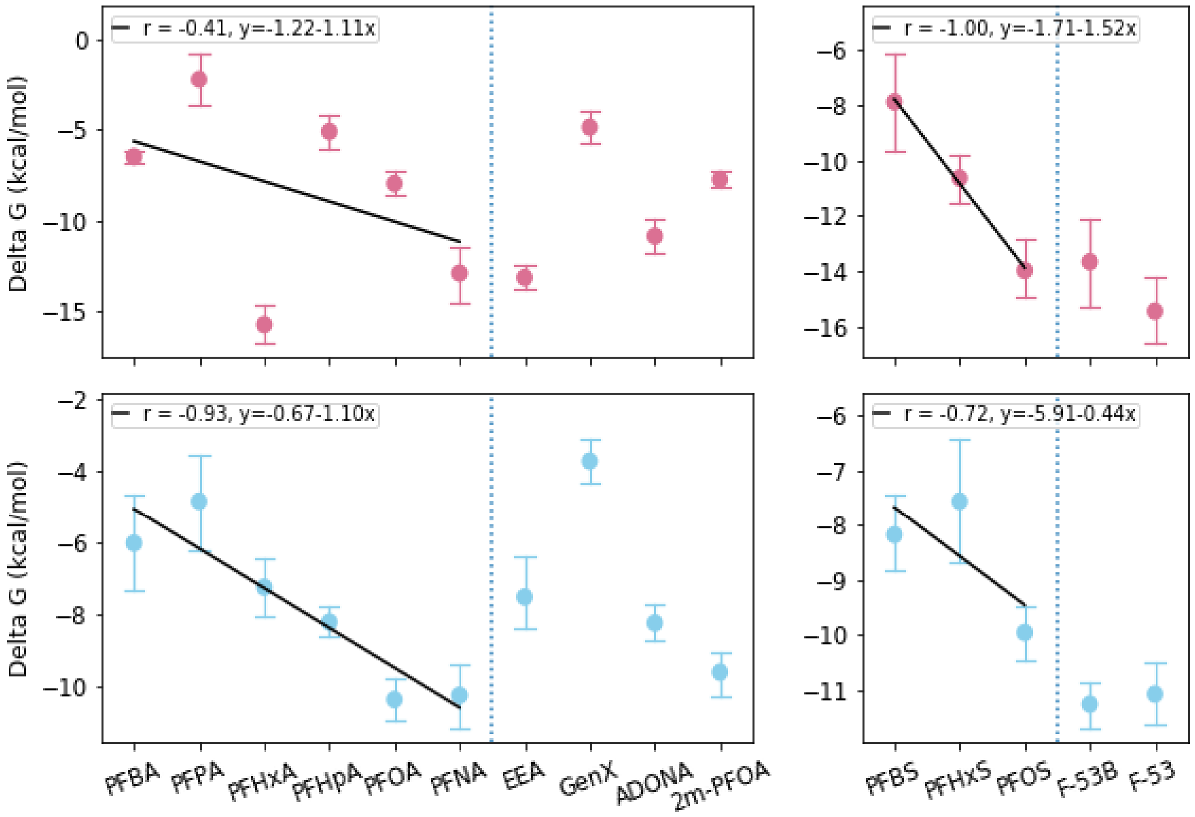


Figure 12. The distribution of ΔG_{bind} (mean ± standard error) for different LFABP-PFAS complexes and the correlation analysis between ΔG_{bind} and carbon chain length.

Pink represents hLFABP, and blue is rLFABP.

A correlation analysis was conducted between predicted ΔG_{bind} and carbon chain length. As shown in Figure 12, for both LFABP systems, the ΔG_{bind} indicates negative relationships with carbon chain length. A strong correlation is observed for PFCAs versus rLFABP, and PFSAs versus both hLFABP and rLFABP, with correlation coefficients of -0.93, -1.0, and -0.72, respectively. The correlation for PFCAs versus hLFABP is relatively weak (r = -0.41), and the major reason for this is due to the much lower predicted ΔG_{bind} for PFHxA, which can be further attributed to the much more favorable electrostatic interaction and polar solvation energy between PFHxA and hLFABP (Figure 11). In terms of predicted ΔG_{bind} for novel PFAS, 2 PFESAs exhibit comparable or stronger binding affinities than PFOS for both LFABPs. The ΔG_{bind} of EEA and ADONA are similar with PFNA when bound to hLFABP, and similar with PFHpA when bound to rLFABP, while the ΔG_{bind} of GenX is weaker compared to that of PFHxA, which has the same carbon number as GenX. Finally, 2m-PFOA has a comparable binding affinity with PFOA for both LFABPs.

3.3.2.3 Free Energy Decomposition

The contribution of each residue in the rat and human LFABPs to the binding free energy was determined based on a per-residue decomposition scheme. For each residue, all free energy components in Equation 3 except entropy and nonpolar solvation energy (the calculation of both terms were not available in the *decomp* program in Amber 14) were calculated. In each LFABP-PFAS complex system, the residues contributing most to the total free energy were determined (they account for 44 % to 85 % contribution among all protein residues). As shown in Table 6, for hLFABP the residues such as ARG 122, SER 39, SER 124, and ILE 52 contribute significantly to ΔG_{bind} among all ligands, while for the rLFABP system, the residues showing strong contributions include ARG 122, ILE 60, ILE 53, TYR 55, and MET 74.

3.3.3 Bioaccumulation Potential Across Species

The MD workflow was employed to probe the bioaccumulation potential across different species by examining LFABP interactions across 7 different species (i.e., human, rat, chicken, zebrafish, rainbow trout, Japanese medaka, and fathead minnow) for 9 PFAS with different chain length and functional groups (i.e., PFBA, PFPA, PFHxA, PFHpA, PFOA, PFNA, PFBS, PFHxS, and PFOS). The ΔG_{bind} value were calculated for each LFABP-PFAS complex and analyzed by different statistical tools. As shown in Table 7, the average ΔG_{bind} values over the 9 tested PFAS ligands for human, rat, chicken, and rainbow trout are smaller than -8.0 kcal/mol. This is a substantially lower value (i.e., stronger binding affinity) than that predicted for Japanese medaka and fathead minnow (average ΔG_{bind} values larger than or equal to -5.25 kcal/mol, Table 7). The binding affinity for zebrafish is between these two groups, with an average ΔG_{bind} value of -6.44 kcal/mol. A one-way ANOVA shows there is a significant difference across 7 species for all 9 tested PFAS in terms of their LFABP binding affinity, with P values ranging from 1.02E-10 to 0.021.

Table 7. The average, max, and min of ΔG_{bind} over all tested PFAS ligands for 7 different species of LFABP.

LFABPs	Max ΔG_{bind}	Min ΔG_{bind}	Mean ΔG_{bind}
human	-4.39333	-13.9894	-8.89
rat	-4.85333	-10.3439	-8.06698
chicken	-4.89333	-12.9956	-9.2
zebrafish	-3.12778	-10.8956	-6.44444
rainbow trout	-2.01111	-16.2389	-8.45975
Japanese medaka	2.956667	-12.9867	-3.86617
fathead minnow	1.024444	-10.9344	-5.25457

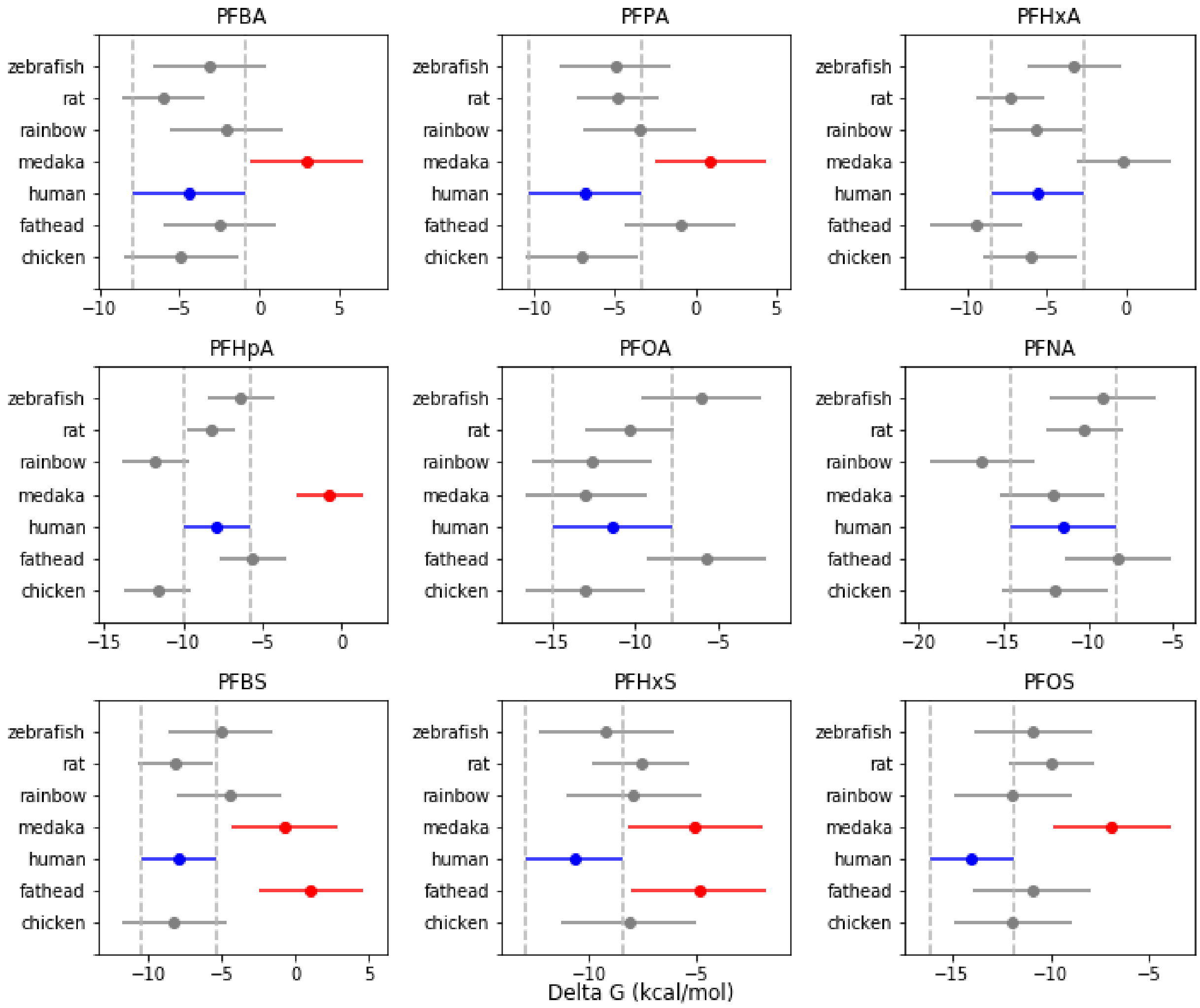


Figure 13. Multiple comparison (Tukey test) between human LFABP and other LFABPs for different PFAS. Blue is human LFABP; red indicates significant difference ($p < 0.05$); gray means no difference from human wild type ($p > 0.05$).

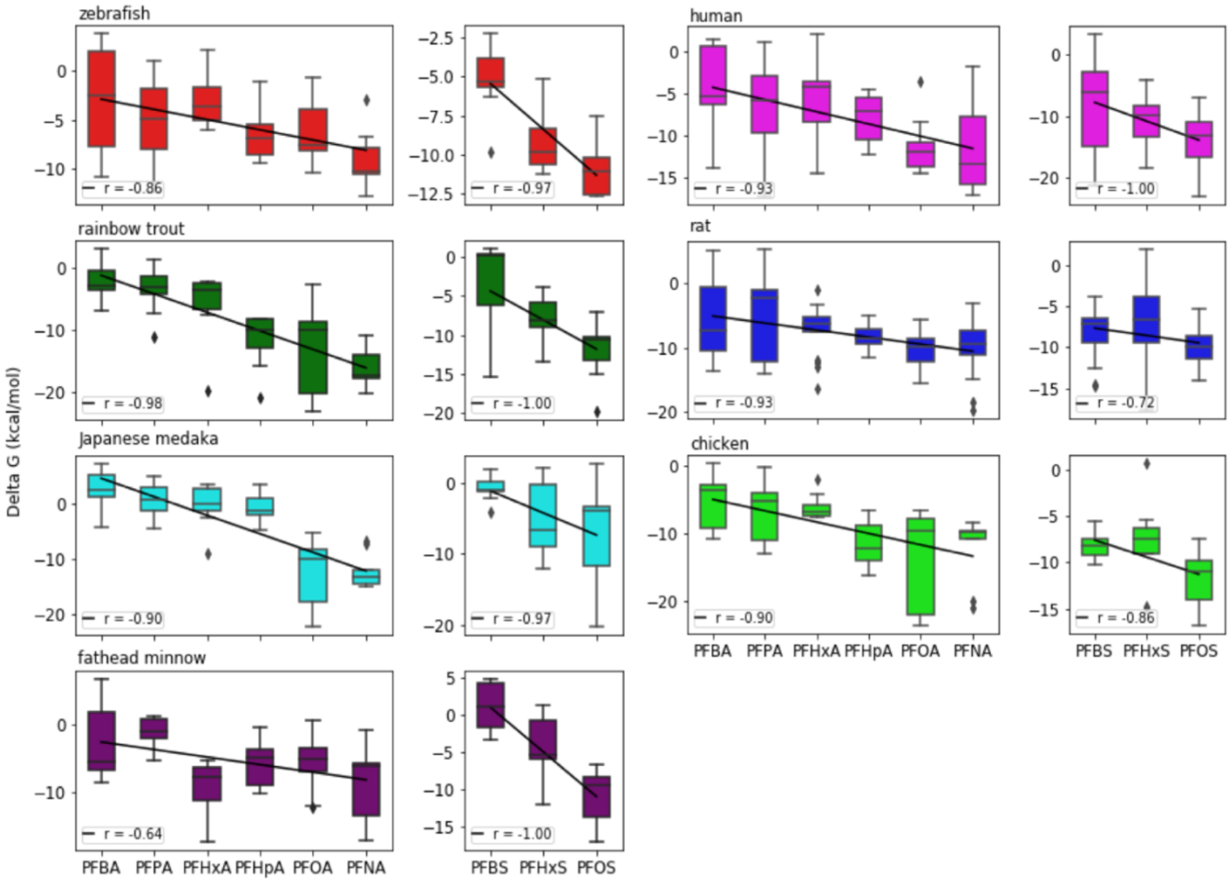


Figure 14. Distribution of ΔG_{bind} for different PFAS- LFABP complexes across species.

The multiple comparison Tukey test between human LFABP and the LFABPs for the other 6 species shows that Japanese medaka has significantly weaker LFABP binding affinity compared to human for all PFAS ligands ($P < 0.05$) except PFHxA, PFOA and PFNA (Figure 13). Fathead minnow also shows significantly weaker LFABP binding affinity than human for PFBS and PFHxS ($P < 0.05$), while LFABP of other species all indicate comparable binding affinity to human LFABP ($P > 0.05$) for all PFAS.

Finally, a correlation analysis was performed between ΔG_{bind} and carbon chain length. As indicated in Figure 14, in all LFABP systems, a quite strong negative relationship was observed for both LFABP versus PFCAs and LFABP versus PFSAs, with the correlation coefficient ranging from -0.64 to -1.0.

3.4 Discussion

In this study, we developed a workflow combining molecular docking and MM-PBSA based on MD simulation techniques to predict the binding affinity of legacy and replacement PFAS for LFABPs. Experimental data from three different studies[37, 115, 116] were used to evaluate this approach, and the performance is excellent for predicting PFAS binding affinity to rLFABP ($r = 0.96$). For hLFABP, predictions are different between Zhang et al.[115] ($r = 0.79$) and Sheng et al.[116] ($r = 0.97$). Both studies used fluorescence displacement assays to measure the dissociation constant. However, the binding affinity results (expressed as K_d values in unit of µM) of Sheng et al. were 3 to 8 times higher than those of Zhang et al., which reveals the variation among different experimental studies. Given that available experimental datasets for LFABP-PFAS complex are very small, we call for a broadening work on protein-PFAS interactions which will make validation of the predicted results more reasonable. The available data in those three studies cover most traditional PFAS and two novel PFAS (i.e., GenX and F-53B). The satisfactory performance of the MM-PBSA method we used demonstrates its ability to rank the binding affinity of both legacy and alternative PFAS.

3.4.1 Protein Binding for Legacy and Novel PFAS

This approach provided mechanistic understanding of how the molecular structures of PFAS influence their protein binding behavior. For legacy PFAS (i.e., PFCAs and PFSAs), as carbon chain length increases, the binding affinity also increases. Further analysis for each energy component of ΔG_{bind} showed that the strong relationship between carbon chain length and binding affinity was mainly caused by the vdw interaction energy and entropy change upon binding, both of which indicate a close correlation with carbon chain length (Figure 11). The sum of electrostatic interaction and polar solvation energy terms, on the other hand, seem to fluctuate around a certain value. The extreme low value for PFHxA bound to hLFABP may be because the simulation time is not long enough for hLFABP-PFHxA complex system. To validate this hypothesis, a twice longer molecular dynamics simulation was performed for the hLFABP-PFHxA complex. The results show that the new calculated average ΔG_{bind} of hLFABP-PFHxA is -5.55 kcal/mol, and with the updated ΔG_{bind}, the correlation coefficient for PFCAs versus hLFABP becomes -0.93, which is much better than previous results.

For most alternative PFAS, the addition of ether groups actually increased their binding affinities to LFABPs in comparison with their perfluoroalkyl counterparts with same carbon numbers. Binding free energy component analysis indicated that introducing oxygen in their backbone increases the chain length of PFAS; longer chain length indicates greater vdw interactions with the proteins and more favorable entropy changes (larger molecule has greater molecular motions and thus higher entropy),[141] therefore, the introduction of ether groups in PFAS could lead to a stronger binding free energy (Figure 12). It is interesting to note that distinct from other novel PFAS, GenX has a branched structure similar to 2m-PFOA, which imparts some special behaviors. Due to its structure, GenX (and 2m-PFOA) showed a significantly different binding mode from linear PFAS; instead of interacting with ARG or SER residues, the head group of GenX mainly interacted with THR residue through H-bonding (Figure 10). Furthermore, although inserting an oxygen atom, the binding affinity of GenX was comparable with or even a little weaker than PFHxA (which has the same carbon numbers as GenX). This implies that the binding affinity is closely related to the backbone chain length, not the total carbon number including branches.

Our results suggest that EEA and ADONA indicate at least as strong binding strength as PFHpA when bound to rLFABP, and as PFOA when bound to hLFABP. For F-53 and F-53B, both have similar or stronger binding affinities than PFOS. Based on our toxicokinetics model, these alternatives could be as bioaccumulative as legacy PFAS. In addition, toxicological studies of F-53B have shown a similar or stronger toxicity compared with PFOS,[102, 116] and the toxic effect (e.g., hepatotoxicity, genotoxicity, and developmental toxicity) of other alternatives were also reported and summarized by Wang et al.[24] The above bioaccumulation and toxicity results indicate that those substances are not necessarily safer alternatives to legacy PFAS, particularly when the backbone chain length is greater than 6.

Given the vast number of PFAS (more than 4000) potentially on the market and our limited resources (e.g., time and cost), it is not feasible to evaluate all PFAS individually through experimental study.[8] Therefore, *in silico* methods based on computational biology hold great promise for the hazard and risk assessment of non-tested PFAS. The MM-PBSA approach based on MD simulation provides reliable protein binding affinity prediction for legacy and alternative PFAS, and thus can be used for large-scale screening of protein-PFAS interactions. In addition, the binding affinities generated from this approach can be further used as parameters for our previous PBTK models, which were developed by considering the interactions between PFAS and proteins including serum albumin, LFABPs, and membrane transporters.[81, 82] The combination of MD simulation and PBTK modeling will provide a flexible framework that can be used to evaluate the bioaccumulation behaviors of non-tested PFAS and support their risk assessment.

3.4.2 Bioaccumulation Potential Across Species

By estimating PFAS binding affinity to LFABP across different species (i.e., human, rat, chicken, zebrafish, rainbow trout, Japanese medaka and fathead minnow), our workflow revealed that human LFABP has comparable PFAS binding affinity to all other vertebrate species evaluated, except Japanese medaka and fathead minnow. The LFABP of those two fish species indicated significantly weaker binding affinities than human for some PFAS ligands (Table 7 and Figure 13). A closer look at the binding mode of PFAS bound to human, Japanese medaka, and fathead minnow LFABP shows that the close contact residues are very similar across those species for different PFAS, but the positions of these residues are quite different between human and the two fish species (e.g., SER124 versus SER52, Table 8 and Figure 15). It seems that the position of key residues, which seem to drive the position of ligand binding, can cause significantly different binding affinities between humans and the two fish species. Because the identity, not the position, of close-contact residues is conserved (i.e., the residue is a serine in both cases in the example above), the specific amino acids are implicated in facilitating certain key interactions (e.g., hydrogen bonding). At the same time, when the position of ligands is closer to the bottom of the LFABP binding pocket, the binding affinity also tends to be stronger (Figure 15). Thus, we conclude that when the position of key residues facilitate binding in a region of the protein that is more energetically favorable (e.g., increases hydrophobic contacts), stronger binding affinities result. However, these observations should be tempered with an acknowledgment that molecular simulations have a degree of uncertainty and variations in the predictions of exact binding conformations can and do occur from simulation to simulation.

Based on the MD workflow results, human, rat, chicken, zebrafish, or rainbow trout seem to be better representative species of the higher range of vertebrate bioaccumulation potential of PFAS than Japanese medaka and fathead minnow. It is worth pointing out that this conclusion is based on the interaction between PFAS and LFABP. Other proteins such as serum albumin and membrane transporters also play important roles in determining the bioaccumulation behavior of PFAS [81] and should be included in future work.

In conclusion, in this Chapter we developed a modeling workflow that combines molecular docking and molecular dynamics simulation techniques to estimate the protein binding affinity of PFAS. By comparing the model prediction with different experimental data, the MD-based workflow has been demonstrated to be an efficient and reliable way to predict the protein binding affinity for PFAS. In addition, the MD-based workflow can be successfully used to inform about the bioaccumulation potential of replacement PFAS and the bioaccumulation potential of PFAS across species. As a critical component of the three-level hierarchical framework for PFAS risk assessment (Figure 2), the MD-based workflow provides a robust way to generate large numbers of protein-PFAS interaction data, which can be used as parameters for the PBTK model. For example, based on the results of Chapter 2.0, the protein-PFAS binding affinity for serum albumin and LFABPs as well as the active transport of PFAS facilitated by membrane transporters are critical processes for the tissue distribution of PFAS in biological systems; the parameters related to those processes can be derived from the output of the MD-based workflow (i.e., ΔG_{bind} for protein-PFAS complex). The combination of MD-based workflow and PBTK modeling will provide a flexible framework that can be used to evaluate the bioaccumulation behaviors of non-tested PFAS and support their risk assessment. A second important component of risk assessment is to understand the toxicity of the chemical being investigated. While the first two parts of this dissertation focused on the toxicokinetics, the final component considers their toxicodynamics through the lens of predicted bioactivity. In next Chapter, we will focus on predicting the bioactivity for large number of PFAS using statistical modeling tools.

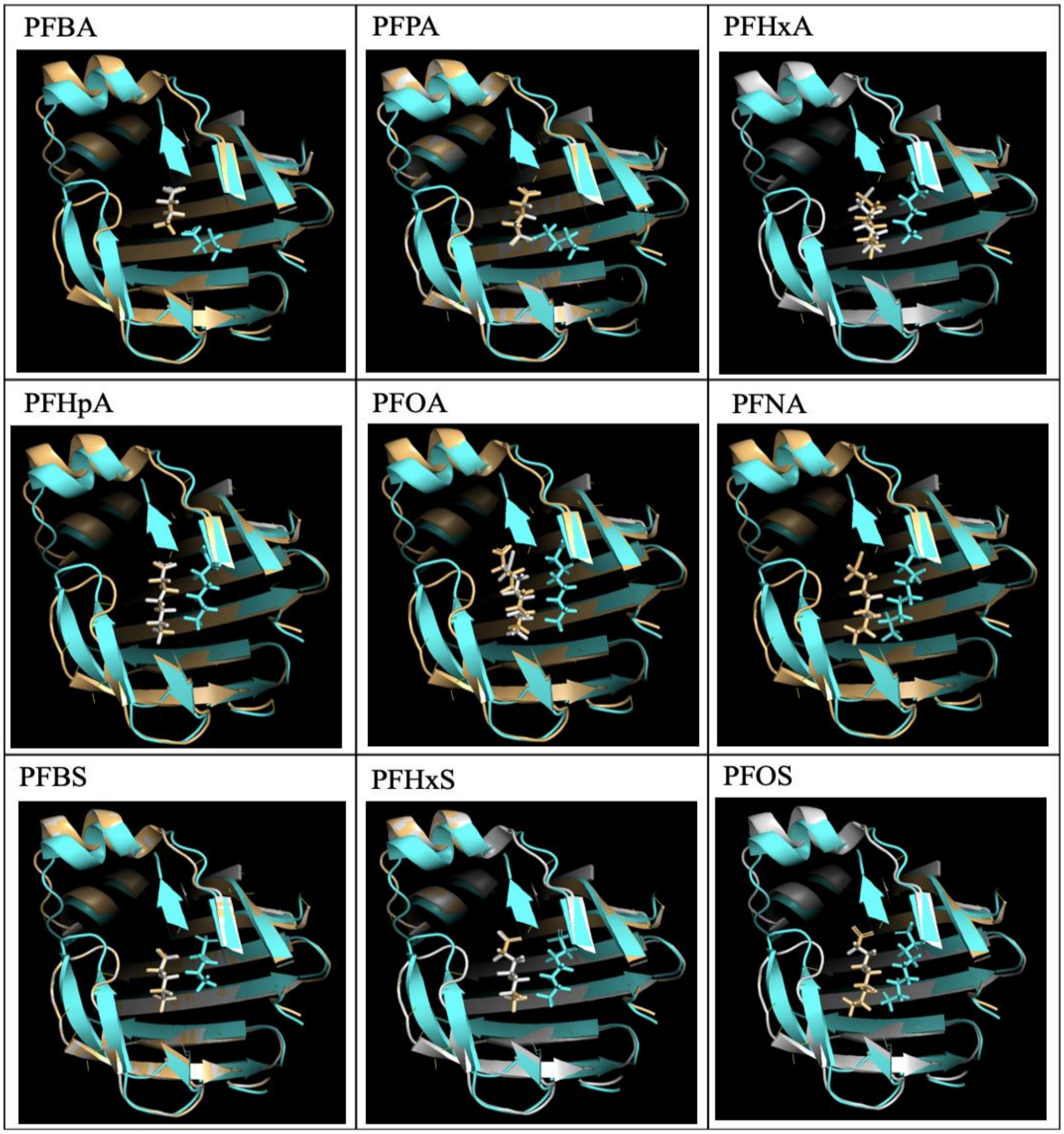


Figure 15. The PFAS binding poses for human (cyan color), Japanese medaka (orange color) and fathead minnow (grey color) LFABP after sequence alignment.

Table 8. The hydrogen bond and close contact residues for PFAS ligands bound to different LFABPs.

Ligands	Japanese medaka		Fathead minnow		Human	
	H-bond	Close Contact	H-bond	Close Contact	H-bond	Close Contact
PFBA	ARG121	ILE50, SER52, ARG121	ARG113	ILE42, SER44, ARG113	ARG122, SER124	SER39, ILE62, ILE109, ARG122, SER124
PFPA	ARG121	ILE50, SER52, ASN61, ARG121	ARG113	ILE42, SER44, ARG113	ARG122, SER124	SER39, ILE109, ARG122, SER124
PFHxA	ARG121	ASN61, PHE63, THR73, HIS99, ARG121	ARG113	ILE42, THR65, ARG113	ARG122, SER124	SER39, PHE50, ILE109, ARG122, SER124
PFHpA	ARG121	ILE50, SER52, ASN61, PHE63, THR73, HIS99, ARG121	ARG113	ILE42, SER44, ASN53, PHE55, THR65, HIS91, ARG113	ARG122, SER124	SER39, PHE50, ILE109, ARG122, SER124
PFOA	SER52, ARG121	ILE50, SER52, PHE63, THR73, HIS99, ARG121	ARG113	ILE42, SER44, ASN53, PHE55, ILE63, CYS84, HIS91, ARG113	ARG122, SER124	SER39, PHE50, PHE63, LEU91, THR102, ILE109, ARG122, SER124
PFNA	SER52, ARG121	ILE50, SER52, ASN61, PHE63, ILE71, THR73, CYS92, HIS99, ARG121	ARG113	ILE42, SER44, ASN53, PHE55, ILE63, THR65, CYS84, HIS91, ARG113	ARG122, SER124	SER39, ILE41, PHE50, ILE52, PHE63, LEU91, ARG122, SER124
PFBS	THR73	ILE50, ASN61, PHE63, THR73, HIS99, ARG121	-	ILE42, ASN53, PHE55, HIS91, ARG113	ARG122, SER124	SER39, ASN111, ARG122
PFHxS	ARG121	SER52, ASN61, PHE63, THR73, HIS99, ARG121	ARG113	SER44, ASN53, PHE55, THR65, HIS91, ARG113	ARG122, SER124	SER39, PHE50, ILE109, ARG122, SER124
PFOS	ARG121	ILE50, SER52, ASN61, PHE63, ILE71, THR73, CYS92, HIS99, ARG121	ARG113	ILE42, SER44, ASN53, PHE55, ILE63, THR65, CYS84, HIS91, ARG113	ARG122, SER124	SER39, ILE41, PHE50, LEU91, THR102, ILE109, ARG122, SER124

4.0 Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances (PFAS) from the OECD List

This chapter is reproduced in part with permission from:

Cheng, W. and Ng, C. A. Using Machine Learning to Classify Bioactivity for 3486 Per- and Polyfluoroalkyl Substances (PFASs) from the OECD List. *Environmental Science & Technology* 2019, 53 (23), 13970-13980. Copyright 2019 American Chemical Society. [https://pubs.acs.org/doi/abs/10.1021/acs.est.9b04833]

A recent OECD report estimated that more than 4000 per- and polyfluorinated alkyl substances (PFAS) have been produced and used in a broad range of industrial and consumer applications. However, little is known about the potential hazards (e.g., bioactivity, bioaccumulation, and toxicity) of most PFAS. Here, we built machine-learning-based quantitative structure-activity relationship (QSAR) models to predict the bioactivity of those PFAS. By examining a number of available molecular datasets, we constructed the first PFAS-specific database that contains the bioactivity information of 1012 PFAS for 26 bioassays. Based on the collected PFAS dataset, we trained 5 different machine learning models that cover a variety of conventional models (e.g., random forest and multitask neural network (MNN)) and advanced graph-based models (e.g., graph convolutional network). Those models were evaluated based on the validation dataset. Both MNN and graph-based models demonstrated the best performance and the average of the best area-under-curve score for each bioassay is 0.916. For predictions on the OECD list, most of the biologically active PFAS have perfluoroalkyl chain lengths less than 12 and are categorized into fluorotelomer-related compounds and perfluoroalkyl acids.

4.1 Introduction

According to a recent Organization for Economic Cooperation and Development (OECD) report, there are nearly 5000 PFAS that have been in some way registered and/or produced;[7] those chemicals can be broadly divided into perfluoroalkyl acids (PFAAs, e.g., perfluoroalkyl carboxylic acids), PFAA precursors (e.g., fluorotelomer-based substances) and others (e.g., fluoropolymers).[8] Although many long-chain PFAAs have been phased out due to their public health concerns,[24, 25, 57] the potential hazards (e.g., bioactivity, bioaccumulation, and toxicity) of those alternative PFAS remain largely unknown.

Due to the large number of PFAS involved and our limited resources, it is very difficult to evaluate all of them individually through experiments. Therefore, reliable quantitative structure-activity relationships (QSAR) or other predictive modeling approaches hold great promise to address the wide variety of PFAS structures in the environment. Computational methods are able to provide valuable insight into the behavior of those chemicals, and thus facilitate high-throughput screening and prioritization. In our previous work, we developed a molecular dynamics simulation based approach that can successfully predict the protein binding affinity for 15 legacy and alternative PFAS;[146] PFAS-protein interactions have been demonstrated to be critical factors in determining their bioaccumulation potential and could potentially be used as a proxy for bioaccumulation assessment.[81] Although the molecular-dynamics-based approach is a solid method to rank the relative protein binding affinity for PFAS, it is still too computationally expensive to handle a large dataset containing over 4000 chemicals, especially when the size of the protein of interest is large (e.g., serum albumin) — in such cases, the time it takes to run simulations and calculate free energies of binding are much longer.[43, 44]

A promising alternative to molecular-dynamics-guided protein interaction approach is machine learning (ML) based QSAR, which employs ML algorithms to model the relationship between physical and biological properties of compounds and their chemical structures.[45] It is much more cost-effective for dealing with large datasets and has been successfully applied for decades in many areas like drug discovery and chemical toxicity predictions.[46, 47] In particular, the advent of deep learning (DL) has revolutionized many research areas including computer vision, speech recognition and computational toxicity.[147] For example, in a Kaggle competition on molecular property predictions, the winning team employed the DL algorithm and achieved a relative accuracy improvement of approximately 15% over a random forest baseline.[26] In addition, the winner of the Tox21 Data Challenge 2014 also used the DL approach.[148] Those successes have brought tremendous attention from the academic community to ML.

In general, building a QSAR model involves three major steps: (i) collecting a training dataset (i.e., compounds with measured physical or biological properties), (ii) encoding those compounds into chemical descriptors (i.e., the feature of each molecule), and (iii) training the model to predict chemical properties based on their descriptors and assessing the model performance using a validation dataset.[149] Advances in high throughput screening (HTS) technologies (which can test hundreds to thousands of chemicals simultaneously) have produced a tremendous amount of assay data. For example, PubChem's BioAssay database contains 1 million bioassays for 96 million compounds. Each bioassay contains a collection of biological activity data (e.g., active or inactive) that are determined by testing against biological targets (e.g., genes, proteins, or cell lines) in *in vitro* experiments (https://pubchem.ncbi.nlm.nih.gov/). Such extensive databases are ideal for use as training datasets for the first step of QSAR development.

For the second step (encoding compounds), molecular featurization is also a well-studied problem and various approaches of encoding molecules into fixed-length vectors or mathematical graphs have been developed, such as extended-connectivity fingerprints (ECFP)[150] and molecular graph convolutions.[151] Once chemical descriptors have been computed, ML models can be trained to make predictions on chemical properties.

There are a number of QSAR models that have been developed to predict different biological or physical properties of PFAS such as interfacial adsorption coefficients,[152] interactions with transthyretin (which is a thyroid hormone transport protein),[153, 154] oral[155] and inhalation toxicity[156] and PFAS mixture toxicity.[157] However, all those studies only focused on a single bioassay and their PFAS dataset are also very small (ranging from 24 to 58). In addition, those QSAR models were built using simple multiple linear regression method and did not consider the state-of-the-art ML algorithms such as neural network,[158] let alone the more advanced graph-based models. To handle the 4730 PFAS in the OECD report, more powerful QSAR models are needed. The goal of this study is to develop powerful ML-based QSAR models to predict the bioactivity for those PFAS in the OECD list. Specifically, our work makes three major contributions:

(1) We construct the first PFAS-specific database for QSAR modeling. The PFAS dataset contains the bioactivity information of 1012 PFAS for 26 bioassays and was gathered from the curated datasets in MoleculeNet,[151] a benchmark collection for molecular machine learning that includes chemical property information for over 700 000 compounds.

(2) Based on the collected PFAS dataset, we trained and evaluated the performance of 5 different ML models that covers a variety of conventional models, including logistic regression,[159] random forests,[160] and multitask neural networks[158] (one of the deep learning algorithms), and also advanced graph-based models, including graph convolutional models[161] and weave models.[162]

(3) Using the models with the best performance for each bioassay, we predicted the bioactivity of the PFAS in the OECD report and analyzed the patterns in activity and PFAS structural features that emerged from the prediction results.

4.2 Methods

4.2.1 Original Data Sets

Table 9. Summary of original datasets obtained from the MoleculeNet benchmark.

Datasets	# Molecules	# Bioassays	# Bioactivities	Active rate (%)*
PCBA	437929	128	34017170	1.389 (0.009 - 33.124)
Tox21	7831	12	77946	7.521 (2.884 - 16.152)
MUV	93087	17	249886	0.196 (0.163 - 0.205)
BACE	1513	1	1513	45.671
BBBP	2050	1	2050	76.439

* Active rate means the percent of bioactive chemicals for an assay. The average of the active rate across all assays in each dataset is shown. Inside the parenthesis, the minimum and maximum active rates for each dataset are indicated.

The general workflow of ML-based QSAR construction and application to PFAS is shown in Figure 16. To the best of our knowledge, there is no PFAS-specific database currently available for QSAR modeling. In order to build such a dataset, we focused on a total of 6 different datasets including PubChem's BioAssay (PCBA),[163] the Maximum Unbiased Validation (MUV),[164] the human β-secretase 1 (BACE),[165] Blood-brain barrier penetration (BBBP),[166] and Toxicology in the 21st Century (Tox21) datasets.[167] All these datasets have been curated and incorporated into the MoleculeNet benchmark (See Table 9 for summary information) and thus can be accessed through MoleculeNet.[151] Briefly, PCBA is a small subset of the PubChem BioAssay database and includes 128 bioassays for over 400 thousand compounds;[163] MUV is also selected from PubChem BioAssay database and is designed for validation of virtual screening tasks using a refined nearest neighbor analysis method, the MUV minimizes the effect of the dataset bias (e.g., analogue bias and artificial enrichment) on validation results;[164] BACE is a membrane-bound aspartyl protease that plays a key role for brain amyloid β accumulation and thus is considered as potential drug target for Alzheimer's disease,[168] the BACE dataset provides BACE binding results for 1513 compounds;[165] BBBP includes the blood-brain barrier permeability properties for over 2000 chemicals;[166] and Tox21 contains toxicity measurements for 7831 compounds for 12 receptor targets including stress response pathways and nuclear receptors.[167] In addition, the bioassays for all datasets are qualitative: bioactivity results are binary, with a label of 1 meaning active and 0 meaning inactive. In other words, each bioassay can be considered as a binary classification task. Finally, all compounds in those datasets are uniquely identified by their Simplified Molecular Input Line Entry System (SMILES) strings, which can be encoded into more sophisticated electronic or topological features of molecules through the molecular featurization methods described below.

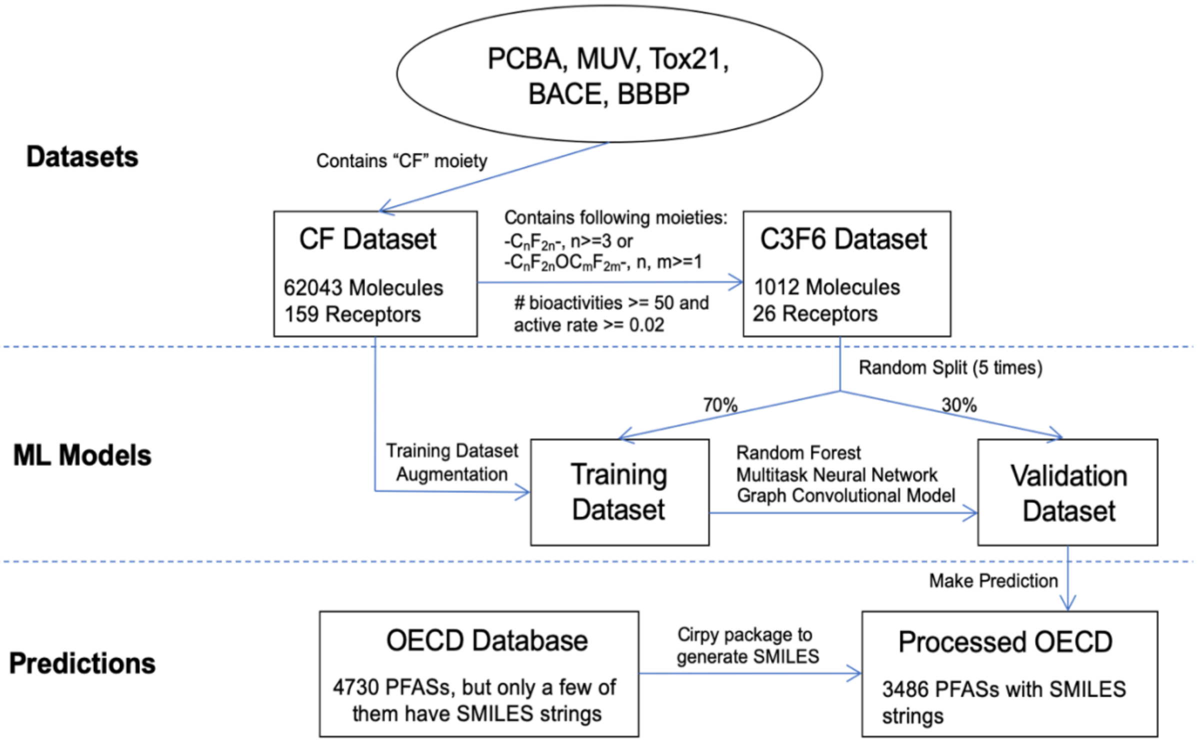


Figure 16. Workflow of machine-learning-based QSAR construction and application to PFAS.

4.2.2 Constructing CF and C3F6 Data Sets

Table 10. Summary of all datasets after -CF- screening process.

Datasets	**# Molecules**	**# Bioassays**	**# Bioactivities**	**Active rate (%)***
PCBA	60559	128	4465740	1.237 (0.002 - 51.415)
Tox21	524	12	4759	9.666 (3.363 - 23.438)
MUV	7756	17	16594	0.211 (0.077 - 0.463)
BACE	761	1	761	56.636
BBBP	285	1	285	91.228
CF	62043	159	4488139	1.257 (0.002 - 91.228)
C3F6	1012	26	14335	7.276 (2.038 - 43.000)

* Active rate means the percent of bioactive chemicals for an assay. The average of the active rate across all assays in each dataset is shown. Inside the parenthesis, the minimum and maximum active rates for each dataset are indicated.

Based on the SMILES strings, we first searched each dataset for compounds containing at least one -CF- moiety. The summary information for each screened dataset is shown in Table 10. We then merged all the reduced datasets into a single CF dataset, which includes 159 bioassays for 62043 molecules. This is the broadest set of fluorine-containing compounds we consider, which includes many compounds that would not be considered PFAS based on the current definition (i.e., -C_nF_{2n+1}-). The major goal of constructing the CF dataset is to use it for data augmentation[169] during the ML model training process. To build a PFAS dataset, we then followed the screening criteria in the OECD report and searched the CF dataset for the substances that contains a perfluoroalkyl moiety with three or more carbons (i.e., -C_nF_{2n}-, n >= 3) or a perfluoroalkyl ether moiety with two or more carbons (i.e., -$C_nF_{2n}OC_mF_{2m}$-, n and m >= 1).[7] After this filtering process, the generated dataset, which is referred as the C3F6 dataset, contains 159 bioassay results for 1012 molecules. For many bioassays in the C3F6 dataset, the bioactivity information was either not available or the active rate (i.e., the number of active results divided by the total number of bioactivities) was very low. To make our models more robust to random splitting, we added another filtering criterion to screen for the bioassays that contain at least 50 bioactivity results and have an active rate not smaller than 0.02. After this step, the final C3F6 dataset includes 26 bioassay results for 1012 compounds.

4.2.3 Machine Learning Models

We selected 5 different ML models including both conventional methods (i.e., logistic regression (LogReg), random forest (RF), and multitask neural network (MNN)) and graph-based methods (i.e., graph convolutional model (GC) and weave model) to conduct the QSAR task for the C3F6 dataset. Those models were selected because they had the best performance on the original datasets in the MoleculeNet (those models were evaluated on the test dataset generated from splitting methods).[151] For this study, we did not include the Kernel-based support vector machine (KernelSVM), which is one of the most widely used ML methods.[170] This is because in a preliminary test, the KernelSVM did not perform well on the small PFAS dataset (Appendix Table 10) and was too computationally expensive to train on the large CF dataset.

A brief description of the ML models follows:

(1) LogReg is a simple classification technique that applies the logistic sigmoid function to weighted linear combinations of the input feature vectors and output binary prediction results.[159] The L2 regularization was applied to overcome overfitting problems and increase the generalizability of LogReg.[171]

(2) RF is a popular ensemble predictive model that consists of a large number of individual decision trees. Each tree is trained using a sample with replacement from the training dataset and

outputs its prediction results. The final prediction of the full forest is determined by the result of each individual tree through majority voting algorithms.[160]

(3) Artificial neural network (ANN) is a powerful non-linear model that maps input feature vector to output vectors through repeated linear and non-linear transformation operations.[158] ANN architecture consists of an input layer, an output layer, and a number of hidden layers. In MNN, the final hidden layer of the network is attached with M softmax classifiers (which is a classification model that employs the softmax function, also known as normalized exponential function, to predict the probability for each class label), one for each bioassay task (M is the total number of bioassays). This architecture makes MNN able to make predictions for multiple bioassays simultaneously. For this study, we evaluated the performance of two different MNN architectures: one is the Pyramidal MNN (P-MNN), which contains two hidden layers, the first layer has 2000 neurons and the second one has 100 neurons;[158] the other one contains only one hidden layer with 1500 neurons (1-MNN). P-MNN demonstrated the best performance in a recent virtual screening task, while the 1-MNN is the default MNN model in the MoleculeNet benchmark.[151]

(4) GC is a graph-based model where the molecules can be modeled as mathematical graphs (i.e., atoms represent nodes, and bonds are edges).[161] In GC, the molecules are first transformed into molecular graph representations, as described in the Featurization section below. Then a number of graph convolution modules (which include graph convolution, batch normalization and graph pooling operations) are sequentially operated on the initial molecular graph. Finally, a graph feature vector is generated by summing up the feature vector of all atoms and then fed to a classification or regression layer.[161]

(5) Similar to GC, weave architecture is another graph-based model.[162] The main difference is the convolutional operation. In weave module, the atom feature vectors are updated by combining information from not only other atoms but also their corresponding pairs in the molecular graph. Adding the pair features makes the weave model more efficient at forwarding information between atoms; however, it also increases the complexity of weave models. In weave models, a number of weave modules (which take atom and pair feature vectors as inputs and output a new set of atom and pair features) are stacked in series and followed by a graph gather layer that combines atom features into molecule-level features. Finally, the molecular features are fed to a classification or regression layer.[162]

4.2.4 Featurization

A total of 3 molecular featurization methods were employed: extended-connectivity fingerprints (ECFP), graph convolutions and weave featurization. ECFP are very popular molecular characterizations in cheminformatics and were used in LogReg, RF, and MNN models to encode the SMILES strings into fixed-length vectors.[150] ECFP are circular topological fingerprints that work by hashing the fragments of the molecule into a fixed-length binary fingerprint. Those fragments represent the circular neighborhood of each atom up to a determined diameter.[150] ECFP are most commonly used with a diameter of 4 (i.e., ECFP4), and ECFP4 were employed in our study. Graph convolutions and weave featurization can encode the molecule into a molecular graph and were used here for the GC and weave models, respectively. In both methods, an initial feature vector is computed for each atom (node) based on its local environment, such as the atom-types, hybridization and valence structures.[151] However, these methods are different in terms of the connectivity information: graph convolution calculates a neighbor list to represent the connectivity of the whole molecule, while weave featurization utilizes more detailed pair feature vectors to represent connectivity information about any pair of atoms in the molecule.[151]

4.2.5 Training and Evaluation

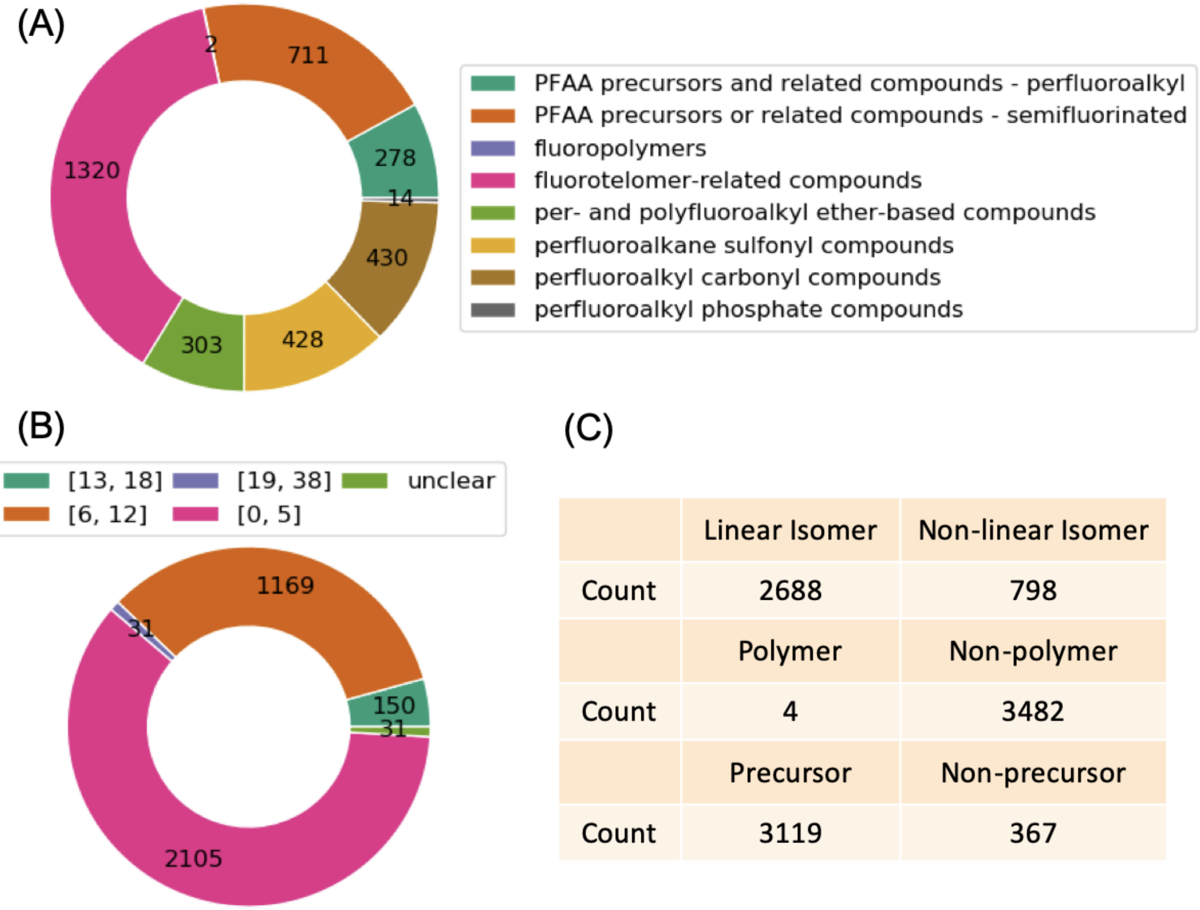
In this study, the C3F6 dataset was randomly split into 70% for the training set and 30% for the validation set. The training set was used to train ML models, and the validation set was used to tune hyperparameters and evaluate the performance of each model. During the training process, in addition to using just the training set, we also applied the data augmentation technique by including the whole CF dataset for training.[169] The model performance for those two training scenarios were compared. Given all bioassays in the datasets are classification tasks, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve was used as the performance metric and the mean AUC-ROC score over all tasks was reported.[172] Following the procedure in the MoleculeNet benchmark paper,[151] the random split was performed 5 times with different random seeds, and the final performance results for different ML models are the average of the 5 independent runs.

4.2.6 Hyperparameter Optimization

The hyperparameters for each ML model are indicated in Appendix Table 11. Both grid search and Bayesian optimization with Gaussian process techniques were employed to tune the hyperparameters for both CF and C3F6 datasets. For grid search, a range of values were considered, and the best hyperparameters were determined based on the mean AUC-ROC score on the validation set. For Bayesian optimization, the Matérn kernel and the expected improvement algorithm were selected as the covariance function and the acquisition function,[173] respectively, to maximize the mean AUC-ROC score for the validation set. After 20 iterations, the optimized hyperparameters were determined, as indicated in Appendix Table 11. For some models such as the RF and GC, the training process was much more computationally expensive for the large CF dataset than the C3F6 dataset. Given our limited resources, we only tuned the hyperparameters using the small C3F6 dataset. For the large CF dataset, we directly applied those optimized hyperparameters from C3F6 dataset.

4.2.7 Prediction on OECD Data Set

After hyperparameter optimization, we selected the ML model that demonstrated the best performance on each individual bioassay to make a prediction on the OECD dataset.[7] Although there are 4730 PFAS listed in the OECD database, only 1213 of them have SMILES string information. Therefore, we utilized the CIRpy package (https://github.com/mcs07/CIRpy), a Python interface for the Chemical Identifier Resolver, to convert the chemical name of the remaining PFAS to SMILES strings. Unfortunately, only a total of 3486 PFAS names (including those 1213 PFAS with available SMILES strings) could be successfully converted to SMILES strings in this way; for the other PFAS, their SMILES representations were not available in CIRpy (the resolver returned None for those chemicals). Our model predictions are based on this final processed OECD dataset containing 3486 individual substances. The overview of those PFAS, including their structure category, perfluoroalkyl chain length, and other features, are shown in Figure 17.



	Linear Isomer	Non-linear Isomer
Count	2688	798
	Polymer	Non-polymer
Count	4	3482
	Precursor	Non-precursor
Count	3119	367

Figure 17. Overview of 3486 PFAS in processed OECD dataset.

The dataset is grouped based on (A) 8 different structural categories (as defined in the OECD list); (B) different perfluoroalkyl chain length (e.g., [6, 12] means the chain length is larger than or equal to 6 and smaller than or equal to 12); and (C) categorization as linear vs. nonlinear, polymer vs. non-polymer, or PFAA precursor vs. non-precursor, shown as orange vs. green in each case.

4.2.8 Applicability Domain (AD)

The distance-based method was employed to determine the AD of the QSAR models.[174] Specifically, the Euclidean distances between the test compounds (i.e., the PFAS in the OECD list) and a defined point (i.e., the centroid) within the descriptor space of the training dataset (i.e., the CF dataset, which is the dataset used to train the QSAR models) were first calculated. Then the distance from each test compound to the defined point was compared with a pre-defined threshold. If the distance of a test compound is smaller or equal to the threshold, the compound is considered to be within the AD. A total of 5 different strategies were used to define the threshold value, as described in the Sahigara study.[174] Briefly, a distance vector was first generated to stores the distance between each compound in the training dataset and the centroid of the training dataset. The first threshold was based on the maximum distance of the vector (*maxdist*), The second and third threshold values were 2 and 3 times of the average distance of the vector (*2*avg* and *3*avg*), respectively. The fourth strategy considered the 95 percentiles (*p95*) of the vector (sorted in ascending order) as the threshold. The last method took the average (*avg*) and standard deviation (*std*) of the vector into consideration and defined the $avg + std * z$ (z is a parameter and set to 0.5 as default[175]) as the threshold.

4.2.9 Computational Resources and Packages

All models were run on the Google Cloud Platform (https://cloud.google.com/) or using resources from the University of Pittsburgh Center for Research Computing (https://crc.pitt.edu/). For conventional ML models (i.e., LogReg, RF, MNN), a virtual instance was created with 8 CPUs, 30 GB memory and 50 GB disk size; while for graph-based models, a virtual instance with

1 NVIDIA Tesla P100 GPU, 4 CPUs, 15 GB memory, and 100 GB disk was established. In addition, all ML models were constructed with the DeepChem package (https://github.com/deepchem/deepchem), an open-source library that provides high-quality implementation of various ML models as well as many useful functions for molecular featurization, dataset splitting, and metric calculation. For hyperparameter optimization, we chose the pyGPGO package,[176] which contains all the functions we need and is easy to implement.

4.3 Results and Discussion

The overview of the 3486 PFAS in the processed OECD list, including their structure category, perfluoroalkyl chain length, and other features, is shown in Figure 17. Fluorotelomer-related compounds are the most dominant structures, followed by PFAA precursors and related compounds (perfluoroalkyl ones) and perfluoroalkyl carbonyl compounds. The fluoropolymer structure category has the smallest numbers of structures (only 2). In terms of perfluoroalkyl chain length, most PFAS have a chain length less than 12, while the PFAS with chain length larger than 18 are very rare (only 31 in total). Most PFAS present belong to the groups of linear isomers, non-polymers, and potential precursors to PFAAs in the environment/biota. Finally, it is interesting to note that among those 3486 PFAS, only 11 of them were also found in the curated C3F6 dataset (which contains 1012 chemicals). Given that the C3F6 dataset was generated based on the same screening criteria as that in OECD report, it implies that some PFAS could be overlooked in the OECD list.

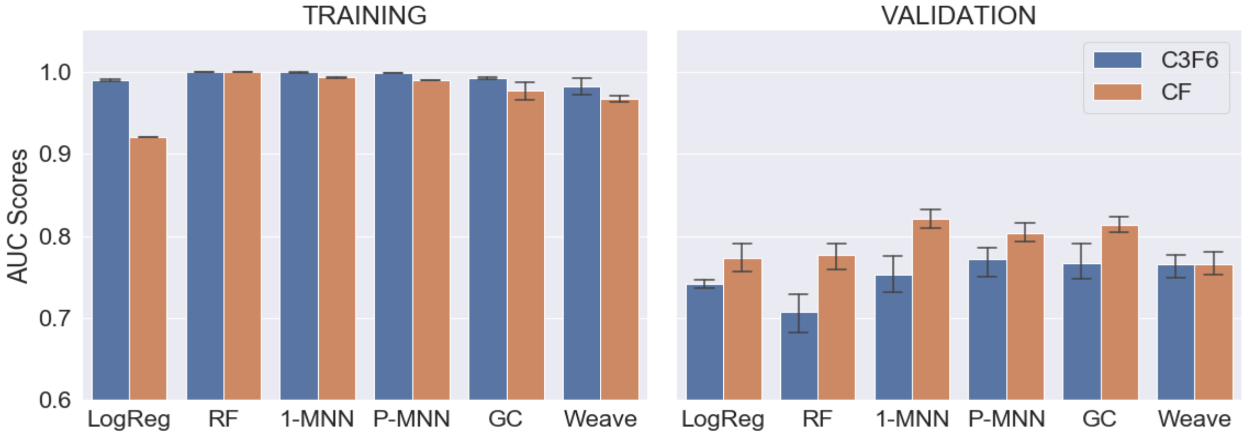


Figure 18. Training and validation AUC-ROC scores of different machine learning models for C3F6 and CF dataset.

Those models include LogReg (logistics regression), RF (random forest), 1-MNN (multitask neural network with 1 hidden layer), P-MNN (pyramidal multitask neural network with 2 hidden layers), and GC (graph convolutional model). Error bar indicates 95% confidence interval of the mean.

The performance of different ML models on both training and validation sets is shown in Figure 18 and Appendix Table 10. As a reminder, the CF dataset is the large dataset that contains at least one -CF- moiety, while the C3F6 dataset is a small subset of the CF dataset and contains either a perfluoroalkyl moiety with three or more carbons or a perfluoroalkyl ether moiety with two or more carbons. For all models, the average AUC scores over all tasks on training sets were substantially larger than that on validation sets, indicating that overfitting is a general problem. The reason for the overfitting issue could be that the training dataset is not large enough. To address this issue, more PFAS-related bioassay data is needed. For the C3F6 dataset, the GC weave and P-MNN models demonstrated better performance than other models. When trained on the CF dataset, all models showed a substantial performance boost compared to those trained on the C3F6 dataset. This illustrates that obtaining more data is of critical importance in improvement of the predictive power of ML models. The best model for the large CF dataset is 1-MNN, which achieved the average AUC score of 0.821 over all tasks.

It is not surprising that MNN demonstrated the best performance for PFAS QSAR tasks, especially when a large amount of training data was provided. MNN has been reported to outperform traditional models on many Big Data-based QSAR tasks.[148, 149, 158] In addition to the strength of the deep neural network architecture, the multi-task learning paradigm plays a critical role in the performance enhancement of MNN models. By incorporating multiple tasks into the learning process, multi-task network could facilitate feature information sharing between different tasks and thus benefit some tasks with limited or imbalanced training data.[148, 149, 158] In our application, the GC model indicated comparable performance to MNN. As discussed in the Methods section, GC is based on graph convolutional architecture, which processes molecules as mathematical graphs and constructs features with convolution layers. In contrast to fingerprint-based featurization methods (e.g., ECFP), graph convolutions take greater advantage of the information in molecular graphs and provide a flexible way to learn new molecular features.[162] A number of studies have shown that the learnable feature extracting architectures either outperform or perform competitively with traditional ML models.[162, 177, 178] As a more sophisticated graph convolution-based model, the weave model, however, did not perform well on the CF dataset. The main reason is that the weave model is very computationally expensive to train (more than 10 hours on a GPU platform). Given our limited resources, we only conducted optimization for a few hyperparameters within a small range (Appendix Table 11) for the weave model. We expect a significant boost in the performance of the weave model could be achieved with a more thorough optimization process.

Table 11. ROC-AUC scores for different machine learning models.

Target Class	Target	Best Models	Best AUC-ROC Scores
GPCR	NPSR	1-MNN	0.888
GPCR	GLP-1	P-MNN	0.923
ion channel	CNG	P-MNN	0.858
miscellaneous	DNA re-replication	1-MNN	0.902
other enzymes	ALDH1A1	1-MNN	0.986
other enzymes	G9a	1-MNN	0.890
other enzymes	CYP2C9	GC	0.988
other enzymes	CYP3A4	GC	0.910
other enzymes	CYP2D6	1-MNN	0.764
promoter	ELG1	P-MNN	0.919
promoter	ATXN	1-MNN	0.969
protein kinase	Plk1 PBD	GC	0.753
protein-protein interaction	K18	1-MNN	0.985
protein-protein interaction	HTTQ103	P-MNN	0.950
protein-protein interaction	JMJD2A	GC	0.902
signaling pathway	Gsgsp	GC	0.992
transcription factor	VP16	1-MNN	0.950
transcription factor	ROR gamma	P-MNN	0.950
transcription factor	Nrf2	1-MNN	0.959
transcription factor	Smad3	1-MNN	0.928
viability	HT-1080-NT	GC	0.913
viability	DT40-hTDP1	1-MNN	0.888
viability	DT40-hTDP1	1-MNN	0.898

The outstanding performances of MNN and GC models demonstrated their capability to estimate the bioactivity for PFAS. We therefore applied the well-trained (with CF dataset) 1-MNN, P-MNN, and GC model to make predictions on the bioactivity of the PFAS in the processed OECD substance list. Here, instead of applying directly the model with the highest average AUC score (i.e., 1-MNN) for all bioassays, we compared the performance of the models on each individual bioassay and selected the best one to make the prediction for each bioassay. The best model and corresponding AUC score for each task (i.e., biological target) are shown in Table 11. The average of the best AUC score across all bioassays was 0.916, which is much higher than the average AUC score of the single 1-MNN model. Finally, the prediction results for PFAS from the OECD dataset were analyzed.

Figures 19-20 and Appendix Figures 7-9 summarize the bioactivity of the 3486 PFAS in the processed OECD dataset. It is important to point out that the 26 targets in our originally compiled data set were chosen because they had available PFAS-related bioactivity data. Many of the targets were developed specifically for drug discovery (e.g., Marburg virus, orthopox virus, and cancer cell viability, see Appendix Table 12), not to identify toxic contaminants. Because 3 of the original 26 assay targets pertain to viruses, we focus here on the results for the 23 assays of more direct relevance to human or animal cells (detailed descriptions for all 26 targets are provided in Appendix Table 12). Each figure represents an individual categorization element (e.g., the structure and the chain length) that the PFAS were grouped into. As shown, the most significantly affected target class is cell viability, the active rate of PFAS for this target class ranges from 10.8% to 60.4%. A fairly large number of PFAS could inhibit the cellular growth of HT-1080-NT fibrosarcoma and chicken DT40 cell lines. However, it should be emphasized that those bioassays are related to cancer cell, which is different from normal human cells. Other targets significantly affected by PFAS (active rate >≈ 10%) include HTTQ103 (Huntington protein with 103 polyglutamines expansion) aggregates,[179] ROR (retinoic acid-related orphan receptor) gamma,[180] G9a (histone methyltransferase) enzyme,[181] and GLP-1 (glucagon-like peptide-1) receptor.[182] Briefly, HTTQ103 aggregation is closely related to neurodegenerative disorders like Huntington's disease;[179] ROR gamma is a transcription factor that plays a critical role in the differentiation of T helper 17 (Th17) cells, which are considered as the major inflammatory cells in autoimmune diseases;[180] G9a is histone methyltransferase that catalyzes the mono- and di-methylation of histone H3 lysine 9 (H3K9) and has been detected to be upregulated in cancer cells;[181] GLP-1 receptor is a Family B G protein-coupled receptor (GPCR) and is a potential therapeutic target for type 2 diabetes.[182] Those active PFAS could reduce the HTTQ103 aggregate formation, inhibit ROR gamma and G9a activity, and bind to GLP-1 receptor. However, the similar bioactivity of those PFAS to drugs could still pose potential threats to human health. For example, those PFAS might compete with the real drugs for binding to therapeutic targets and cause undesirable effects. This hypothesis is evidenced by a recent study that PFOA/PFOS could mimic the function of their structurally similar fatty acids and activate the uncoupling protein 1 in brown adipose tissue, which may lead to a series negative effects such as decreased metabolic efficiency and decreased fitness.[183]

Most of the PFAS predicted to be biologically active are fluorotelomer-related compounds, followed by PFAAs and PFAA precursors. The distribution of different active PFAS structures for each target is roughly proportional to their distribution in the processed OECD dataset (Figure 17A). In terms of perfluoroalkyl chain length, short-chain PFAS (length < 6) and medium-chain PFAS (6 <= length <= 12) are dominant classes in both the biologically active PFAS and the total PFAS in the OECD substance list. It is interesting to note that for cancer cell viability, the number of active PFAS are very similar between short-chain and medium-chain PFAS, however, the total number of short-chain PFAS is almost twice as much as that of medium-chain PFAS (Figure 17B). This indicates that medium-chain PFAS are more likely to impair cancer cell viability. Finally, we analyzed the chemical structures for those PFAS that indicate bioactivity for at least half of the 26 targets (Figure 21 and Appendix Figure 10). The screened 26 PFAS all have similar chemical structures: at least one benzene ring and the perfluoroalkyl structure with at least three carbons. A large portion of those PFAS belong to perfluoroalkyl carbonyl compounds, with the remaining PFAS being fluorotelomer-related compounds and PFAA precursors. In addition, 6 of those PFAS (CAS Numbers: 77758-84-0, 77758-89-5, 106376-38-9, 106376-37-8, 200862-70-0, 77758-79-3) were included in the KEMI (Swedish Chemicals Agency) 2015 report on occurrence and use of highly fluorinated substances and alternatives (https://www.kemi.se/global/rapporter/2015/report-7-15-occurrence-and-use-of-highly-fluorinated-substances-and-alternatives.pdf). It should also be pointed out that some of the highly bioactive PFAS have a relatively long carbon chain length (6 to 8). Given existing concerns about the bioaccumulation and toxicity of long-chain PFAS, additional scrutiny to determine whether the production and use of these chemicals should be limited is warranted. Finally, many short-chain PFAS (perfluoroalkyl chain lengths of 3, see Appendix Figure 10) were classified as bioactive for every target. This emphasizes that short-chain PFAS are not biologically inert, further supported by the recently released studies by the National Toxicology Program (https://ntp.niehs.nih.gov/results/areas/pfas/index.html).

Table 12. The applicability domain of the QSAR model based on distance-based approach with different pre-defined thresholds.

Strategies	Threshold values	Number of compounds outside the AD
maxdist	8.93	0
*2*avg*	11.84	0
*3*avg*	17.75	0
*avg+std*z*	6.20	162
p95	6.82	33

Notes: *maxdist*: maximum distance; *2*avg*: 2 times the average distance; *3*avg*: 3 times the average distance; *avg+std*z*: average distance plus standard deviation multiplying parameter z; *p95*: 95 percentile

Finally, the applicability domain (AD) of the QSAR models was determined by distance-based method. Different strategies were used to define the distance threshold. As shown in Table 12, using the maximum distance and 2 and 3 times the average distance as the threshold, all 3486 PFAS in the OECD list lie inside the AD. The threshold strategy of *avg + std * z* indicated the narrowest domain, with 162 PFAS outside the AD. Even so, there are still 95.4% PFAS falling inside the AD, indicating our models are able to provide reliable predictions for most PFAS in the OECD list.

In this study, we built a PFAS dataset (i.e., the C3F6 dataset) based on available chemical bioactivity datasets. With the PFAS dataset, we trained a number of state-of-the-art ML models including both conventional models and graph-based models, which were then used to make bioactivity predictions for untested PFAS in the OECD database. Given the limited data on the bioactivity and toxicity of PFAS, we believe the benchmarking PFAS dataset we created will facilitate the development of novel ML-based QSAR methods to predict PFAS properties in a more efficient and reliable way. It is worth noting that our current QSAR models are used for bioactivity classification only (yes/no for bioactivity) and cannot provide information about intensity of effect or dose-response. That will be the key in further understanding the impacts of a particular compound. In future work, we hope to expand the PFAS dataset to cover a broader range of molecular properties. For example, protein binding affinity is a critical factor in determining the bioaccumulation of PFAS,[81] and thus should be included in the future. In addition, we will conduct further hyperparameter optimization for the graph-based models to achieve better performance on the PFAS dataset.

In conclusion, in this Chapter we developed ML-based QSAR models to predict the bioactivity of 3486 PFAS in the OECD list. Based on the evaluation on validation dataset, both multi-task neural network and advanced graph-based models demonstrated the best performance and the average of the best area-under-curve score for each bioassay is 0.916. The ML-based QSAR is an important part of the three-level hierarchical framework for PFAS risk assessment (Figure 2); although it cannot provide insights into the toxicokinetics and tissue distribution of PFAS in biological systems, the ML-based QSAR is very cost-effective and can be used to screen large number of PFAS. After the screening process, the molecular dynamics workflow and physiologically based toxicokinetic model can be used to predict the toxicokinetics and bioaccumulation of the prioritized PFAS in biological systems. In addition, the ML-based QSAR can provide quantitative toxicodynamic information for PFAS such as whether the chemical is bioactive/toxic or not.

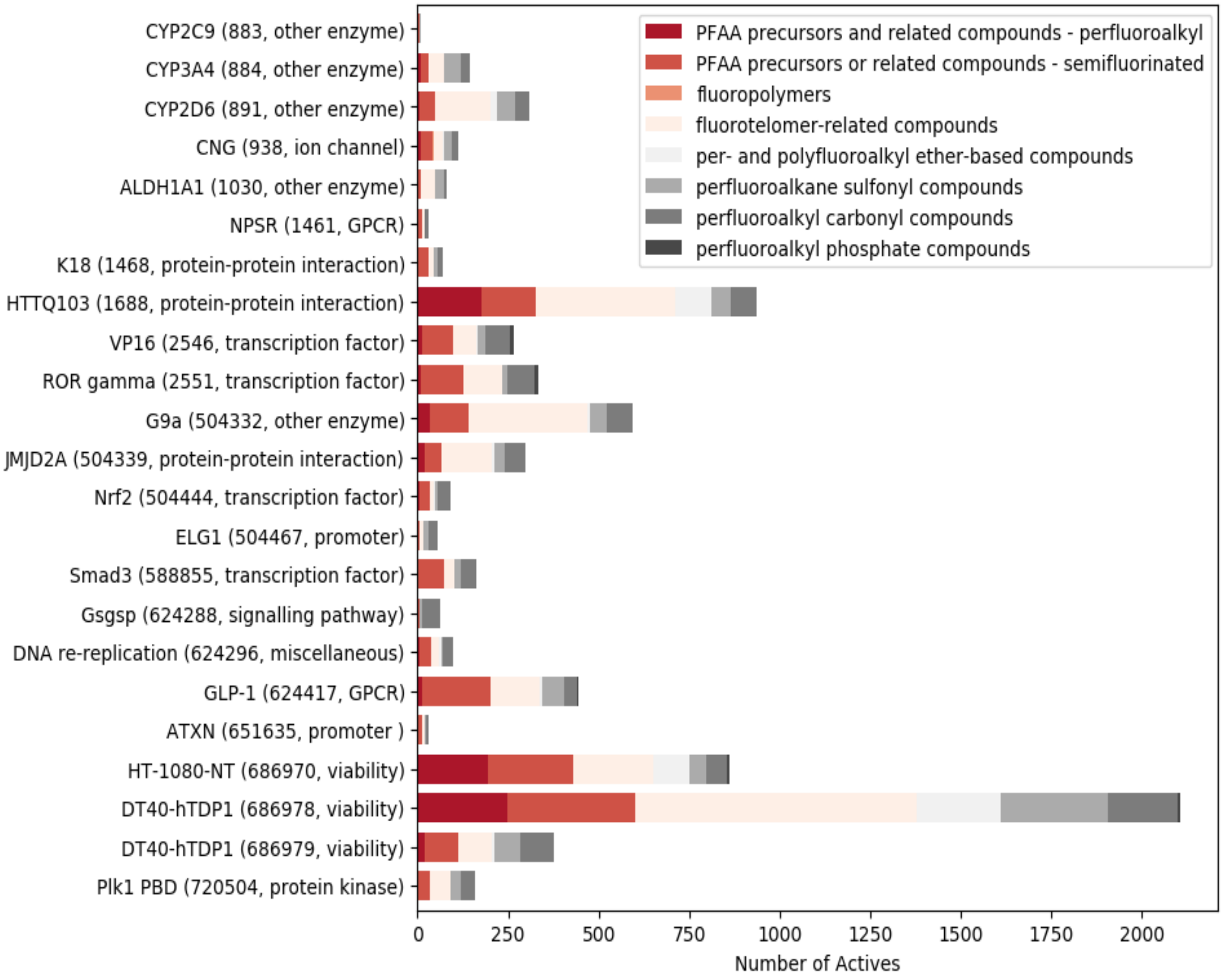


Figure 19. Predicted biological activity of all PFAS (grouped by structure categories) in processed OECD dataset for the 23 targets.

Inside the parentheses, the PubChem AID and the target class are indicated.

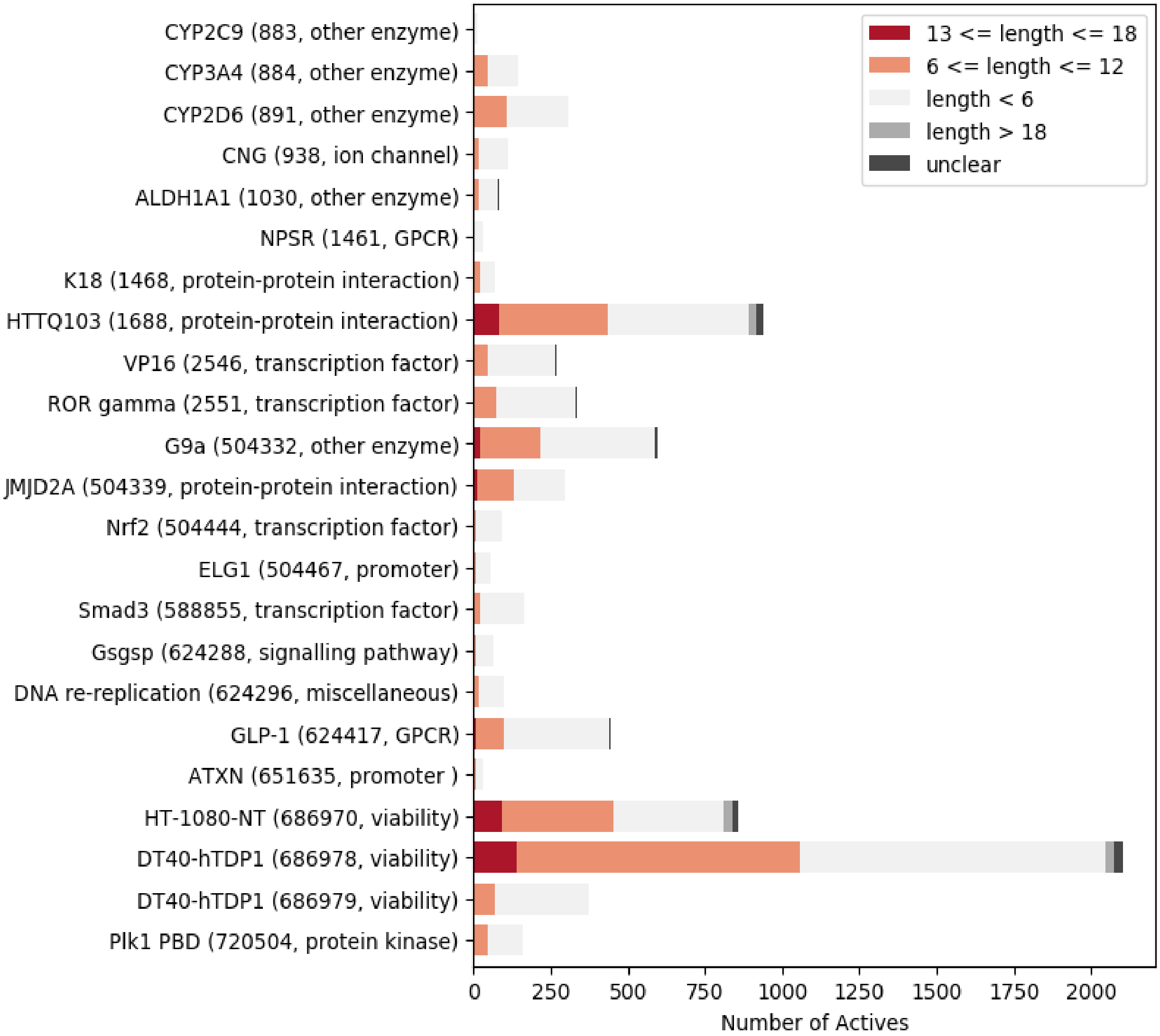


Figure 20. Predicted biological activity of all PFAS (grouped by perfluoroalkyl chain lengths) in processed OECD dataset for the 23 targets.

Inside the parentheses, the PubChem AID and the target class are indicated.

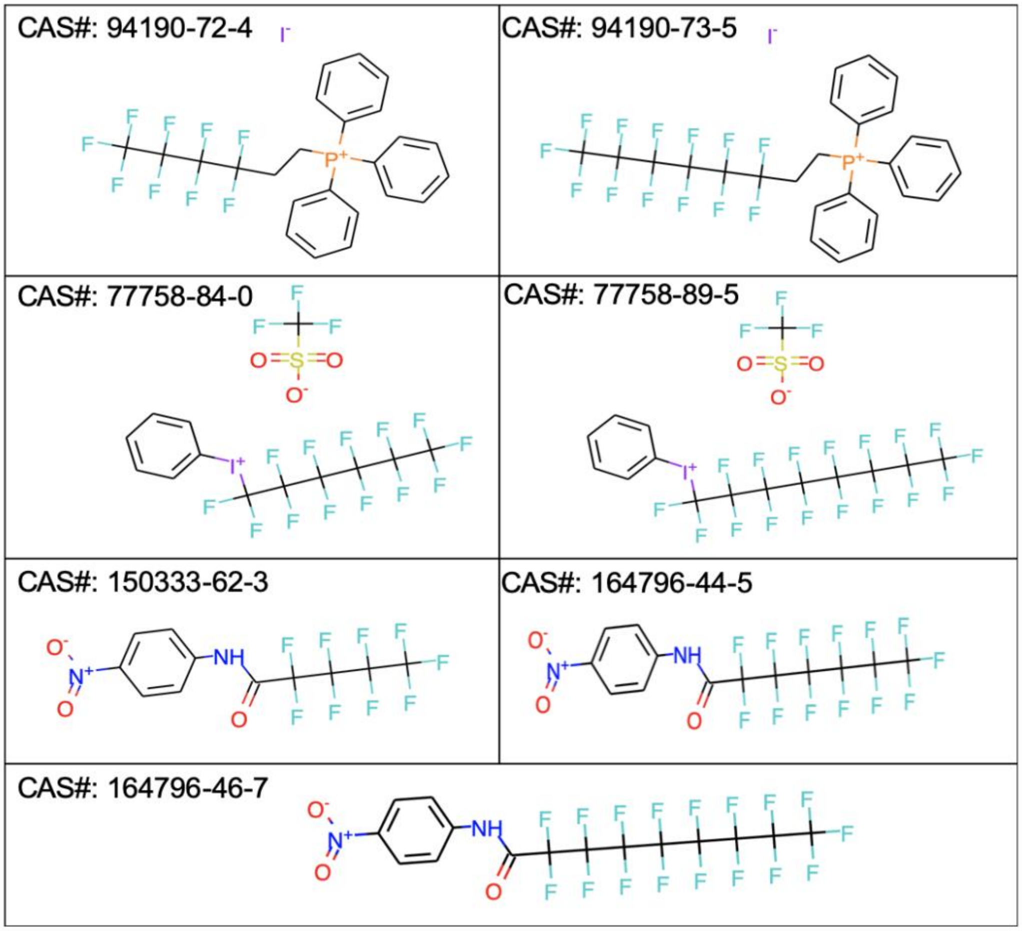


Figure 21. The chemical structures for PFAS predicted to have biological activity for at least half of the targets investigated in this study.

The chain length of each individual perfluoroalkyl substructure in the PFAS presented here is longer than 3; for an additional 19 structures with shorter perfluoroalkyl chain lengths see Appendix Figure 10.

5.0 Summary and Future Work

5.1 Summary

To evaluate the potential hazards of PFAS, different *in silico* techniques were developed in this dissertation including permeability-limited physiologically based toxicokinetic (PBTK) modeling to predict PFAS distribution and accumulation in the rat, a molecular dynamics (MD) based workflow to predict PFAS-protein interactions, and machine learning (ML) based quantitative structure-activity relationship (QSAR) models to predict PFAS bioactivity. The performance of these models and their application to PFAS risk assessment are summarized as follows:

The permeability-limited PBTK model was successfully used to predict the toxicokinetics and tissue distribution of PFOA in male rats. With the help of the hierarchical Bayesian framework, not only were the uncertainties of the posterior parameters substantially reduced, but the PBTK model predictions also became more robust: with the posterior parameters, most of the predicted plasma toxicokinetic (e.g., half-life) and tissue distribution fell well within a factor of 2.0 of the experimental data. In addition, the PBTK model could provide insights into the molecular mechanisms that result in the observed PFOA toxicokinetics: PFAS-protein binding, membrane permeability and active transport.

The MD-based workflow provides an efficient and reliable way to predict the protein binding affinity for PFAS. By comparing the prediction with experimental data, a good correlation between predicted free energies of binding and measured binding affinities was observed, with correlation coefficients ranging from 0.79 to 0.97. In addition, the MD-based workflow provides mechanistic understanding of how the molecular structures of PFAS influence their protein binding behavior. For example, it was observed that as carbon chain length increases, the binding affinity of PFAS also increases. Further analysis for each energy component of the free energy of binding showed that the strong relationship between carbon chain length and binding affinity was mainly caused by the van der Waals interaction energy and entropy change upon binding, both of which indicate a close correlation with carbon chain length. Finally, it was demonstrated that the MD-workflow can be used to inform about the bioaccumulation potential of replacement PFAS and the bioaccumulation potential of legacy and replacement PFAS across species. With respect to replacement PFAS, our results suggest that EEA and ADONA are at least as strongly bound to rLFABP as perfluoroheptanoic acid (PFHpA), and to hLFABP as perfluorooctanoic acid (PFOA). For F-53 and F-53B, both have similar or stronger binding affinities than perfluorooctane sulfonate (PFOS). Given that interactions of PFAS with proteins (e.g., LFABPs) are important determinants of bioaccumulation potential in organisms, these alternatives could be as bioaccumulative as legacy PFAS, and are therefore not necessarily safer alternatives to long-chain PFAS. For bioaccumulation potential across species, humans, rats, chickens and rainbow trout had similar binding affinities to one another for each PFAS, whereas Japanese medaka and fathead minnow had significantly weaker LFABP binding affinity for some PFAS. This result indicates that humans, rats, chicken, zebrafish and rainbow trout seem to be better representative species of the higher range of vertebrate bioaccumulation potential of PFAS than Japanese medaka and fathead minnow.

Finally, the ML-based QASR models, built on the first PFAS-specific bioassay database constructed in this work, showed good performance. Both multitask neural network and graph-based models demonstrated the best performance and the average of the best area-under-curve score for each bioassay is 0.916. For predictions on the OECD list, most of the PFAS predicted to be biologically active are fluorotelomer-related compounds, followed by PFAAs and PFAA precursors. In terms of perfluoroalkyl chain length, short-chain PFAS (length < 6) and medium-chain PFAS (6 <= length <= 12) are dominant classes in both the biologically active PFAS and the total PFAS in the OECD substance list.

As indicated above, the PBTK, MD-based workflow and ML-based QSAR models form a three-level hierarchical framework that has the potential to revolutionize risk assessment for PFAS by enabling fast *in silico* prediction of their bioaccumulation and toxicity and providing insights into the toxicokinetics and tissue distribution of those chemicals in biological systems. Specifically:

(1) Our models will relate external measures of exposure (e.g., concentration of PFAS in air, food, or water) to internal measures of biologically effective dose (e.g., concentration of PFAS in the tissue showing the toxic effects), and thus allow for more realistic internal dose-based exposure assessment.

(2) The parameterization strategies for these models substantially reduce reliance on *in vivo* animal data by making use of *in vitro* and *in silico* data sources in innovative ways. Moreover, our work will provide a more efficient protocol for risk assessment of large numbers of PFAS in a short time.

(3) Our models will allow for extrapolation of dosimetry among different exposure scenarios and animal species, as well as between healthy and susceptible groups, if the relevant physiological properties of the target population are available; this is very useful for estimating PFAS toxicokinetics in some species where *in vivo* data is very scarce (e.g., humans). It can also

be used to account for interindividual variability by defining the distributions for physiological and physiochemical parameters.

Therefore, our models will play essential roles in enabling new toxicity-testing paradigms for assessing risks posed by PFAS.

5.2 Future Work

Due to limited data availability for model parameters, the PBTK model was only tested for the toxicokinetics of PFOA in male rats. In the future, to generalize the framework to other PFAS and other species including humans, more protein-PFAS interaction data are needed. In addition to *in vitro* techniques such as fluorescence spectroscopy[37], equilibrium dialysis[38], and NMR[39] to measure protein binding affinity, the MD-based workflow provides an efficient way to estimate the protein binding affinity for a large number of PFAS. However, it should be pointed out that the values generated from the MD-based workflow are relative protein binding affinities for PFAS. This is because the molecular mechanics combined with Poisson-Boltzmann surface area (MM-PBSA) energy calculation method in the workflow involves several thermodynamic approximations for the purpose of computational efficiency[43] and thus the estimated free energy of binding values cannot be used directly as parameters. Two ways are proposed to deal with this issue. First, a relationship between MD-predicted parameter values and the actual values can be established using linear regression techniques based on the available *in vitro* experimental data and their MD-predicted values. Then the regression model can be used to estimate the actual values for those protein-PFAS interaction parameters that are not experimentally measured. Another way is to employ more rigorous method such as thermodynamic integration (TI) technique to estimate protein binding affinity for PFAS. TI is a theoretically rigorous method used to calculate free energy differences between two given states (e.g., bound and unbound state) based on a non-physical thermodynamic cycle;[184] although TI is more computational expensive compared to MM-PBSA, it has shown excellent accuracy in binding free energy prediction.[184] Both of those methods are worth exploring in the future work given the essential roles of the protein-PFAS interactions in the toxicokinetics model development for PFAS.

In addition, alternative methods to the Markov chain Monte Carlo (MCMC) simulation can be used for the Bayesian analysis of the PBTK model. MCMC is a powerful method to generate the posterior distribution of model parameters; however, it is not suitable for handling sequentially generated experimental datasets due to computational cost.[185] If new experimental data for PFAS toxicokinetics are available, to include those datasets for Bayesian inference, the previously estimated posterior parameter distribution cannot be reused and a new MCMC simulation must be conducted.[185] To make the Bayesian inference more cost-effective, we can employ the sequential Monte Carlo or particle filtering (PF) method, which is a simulation-based method that can be used to perform on-line estimation for posterior distributions of interest.[186] In PF, the importance sampling technique is employed to calculate the importance weights for each sample (i.e., particle) and then the samples with the least weights are filtered out.[186] When new data arrive, PF would update the estimation of the posterior distribution based on only the new data (no need to reconsider the old data), which makes PF very efficient to handle the sequential dataset.[185, 186] In the future, the Bayesian analysis with PF simulation can be tested for the PBTK model to improve the model performance.

Finally, for ML-based QSAR models, a more rigorous data splitting technique for model training and evaluation could improve the rigor of the approach. In the current ML model development, the dataset was split into 70% for training and 30% for validation and the validation set was used for both hyperparameter tuning and model evaluation. It is suggested that the dataset used for model evaluation need to be completely separated from other datasets to avoid any bias error caused by the test dataset.[187] In the future, to provide unbiased evaluation for the ML models, a more rigorous method should be applied to split the dataset into 70% for training, 15% for validation and 15% for testing. The validation is then used for hyperparameter optimization, and the test set for model evaluation.

The National Research Council's report, *Toxicity Testing in the 21st Century*, proposes a shift from *in vivo* animal tests to *in vitro* assays and *in silico* approaches.[27] Under this paradigm shift, more and more high throughput *in vitro* assays and sophisticated computational tools are used for chemical risk assessment by regulatory agencies. For example, the Tox21 program, a collaborative work between United States Environmental Protection Agency (EPA), National Center for Advancing Translational Sciences (NCATS), and National Toxicology Program (NTP), has generated over 100 million data samples for around 8500 chemicals with high throughput screening technology.[188] Based on those large toxicological data sets, advanced computational tools including mechanistic modeling (e.g., toxicokinetics model) and statistical modeling (e.g., machine learning-based QSAR) have been developed to inform chemical risk assessment.[189, 190] These *in vitro* assays and *in silico* methods hold great promise to tackle the PFAS management issue more efficiently. *In vitro* assays can be used to generate a large number of toxicity data with high throughput technology as well data for the parameterization of toxicokinetic models for PFAS (e.g., protein binding affinities). In addition, advancements in computing hardware and algorithms could enhance the predictive power of those *in silico* tools and reduce the computational costs of applying those models. With enough data and more advanced models, the toxicokinetics, bioaccumulation and toxicity of PFAS can be predicted in a more efficient and robust way, which could substantially improve the risk assessment of PFAS.

Appendix A Supporting Information for Chapter 2.0

Appendix A.1 PBTK Model Parameters

For rat physiology, most model parameters are representative mean values of the mean value reported in different studies.[191] In addition, our PBTK model was evaluated on experimental data from both Sprague-Dawley and Wistar rat species, so it represents some generic rat model in this study.

Appendix Table 1. Physiological parameters a generic for male rat.

	Blood	Liver	Kidney	Gut	Muscle	Adipose	Rest of Body
Fractional Volume (%BW)	5.4[a] [192]	3.66 [191]	0.73 [191]	2.69 [191]	40.43 [191]	7 [191]	40.09[b]
Blood Flow Rate [c] [191] (%)	-	2.4	14.1	15.1	27.8	7	33.6[b]
Interstitial Fluid (mL/g tissue)	-	0.049 [193]	0.13 [194]	0.28 [195]	0.054 [82]	0.174 [82]	0.18[d]
Blood Volume (mL/g tissue)	-	0.21 [191]	0.16 [191]	0.034 [196]	0.04 [191]	0.02 [191]	0.036[e]
Capillary Surface Area [f] [78] (cm^2/g)	-	250	350	100	70	70	100
Bile Duct Volume [193]	0.4% of liver tissue volume						
Renal Filtrate Volume [g]	0.25 mL						
Gut Lumen Volume [197]	4.5% BW						
Glomerular Filtration Rate [198]	10.74 mL/min/kg BW						
Urine Flow Rate [192]	200 mL/d/kg BW						
Bile Flow Rate [192]	90 mL/d/kg BW						
Feces Flow Rate [h]	5.63 mL water per day						

Volume calculations were based on density of 1 g/mL.[191]

[a] Plasma volume is 3.12 % of body weight (BW).[192]

[b] Fractional volume and blood flow rate were calculated by subtracting the fraction of other tissues from 1.

[c] Expressed as the percent of cardiac output (Q_c); $Q_c = 0.235 \times BW^{0.75}$ L/min, where the unit of BW is kg.[191]

[d] Based on data availability, it was assumed to be the weighted average of brain, heart and spleen fluids.[199]

[e] Calculated on the weighted average of blood volume of the "rest of body".[191]

[f] Only the data for liver, kidney and muscle were available; the capillary surface area of other tissues was assumed. In kidney, the surface area of glomerular capillary, through which blood filters into filtrate compartment, is 6890 mm^2/g kidney.[200] The area for exchange between each subcompartment was assumed to be the same as the capillary surface area of each tissue, except for the apical membrane of enterocytes and proximal tubules. The microvilli located on these two apical membranes could increase the corresponding surface area significantly. Taking this into consideration, the surface area of gut lumen would be 4.14 m^2/kg BW[201] and the area of the apical membrane of proximal tubule would be increased by a factor of 5; this was assumed to be the same as the enlargement factor used in describing area increasing due to the numerous projections formed on the intestinal wall.[201]

[g] To calculate the volume of the filtrate compartment, we considered the tubular lumen as a cylinder with length of 5.16 mm[202] and diameter of 45 µm[203], and there are about 30000 nephrons in adult rats,[204] therefore, the volume of the filtrate compartment for an adult rat was estimated to be 0.25 mL.

[h] We assumed the PFOA was associated with the water content of feces, which was estimated to be 45% of total fecal weight.[205] Additionally, we used an estimate of fecal production for male Sprague–Dawley rats of 6.88 g dry weight per day.[206]

Appendix Table 2. Protein concentrations in different tissues.

	C_{total} (mg/mL)[a] [207]	$C_{albumin}$ (μmol/L)	C_{LFABP} (μmol/L)	$C_{a2\mu}$ (μmol/L) [e]
Plasma	67	486[b] [192]		
Liver Fluid		243 [208]		
Liver Tissue	40		133[d] [209]	
Kidney Fluid		243		
Kidney Tissue	34		2.65 [76]	110 [77]
Gut Fluid		146[c]		
Gut Tissue	20.6			
Muscle Fluid		146 [71]		
Muscle Tissue	20.6			
Adipose Fluid		73 [71]		
Adipose Tissue	20.6			
Rest of Body Fluid		73[c]		
Rest of Body Tissue	20.6			

[a] Total protein content of liver, kidney and gut were estimated based on a study that investigated the distribution of heart FABP in different organs; in that study, heart FABP concentrations were normalized to both protein content and organ weight; based on that information the protein content of each tissue could be determined. For other tissues the same protein concentration as that in gut was assumed.

[b] Calculated assuming molecular weight of albumin is 65 kg/mol.[210]

[c] Kidney and gut were assumed to have similar albumin levels as liver and the "rest of body" was the same as muscle.

[d] Calculated assuming molecular weight of L-FABP is 14 kg/mol.[77]

[e] α2u-globulin is a male-specific protein, and its molecular weight is 15.5 kg/mol.[77]

Appendix Table 3. Association constant (K_a) and binding sites (n) of the proteins for PFOA.

Proteins	K_a (M^{-1})	n	References
Albumin	3.10E+03	7.8	Han et al. (2003)[39]
LFABP	1.20E+05	3	Woodcroft et al. (2010)[37]
	4.00E+04		
	1.90E+04		
$\alpha 2\mu$-globulin	5.00E+02	1	Han et al. (2004)[75]

Appendix Table 4. Effective permeability (P_{eff}) for different tissues and steady-state cell-water concentration ratios.

Tissue	Blood	Liver	Kidney	Gut	Muscle	Adipose	Rest of Body
P_{eff} (m/s)	4.98E-08	5.15E-08	4.38E-08	2.65E-08	2.65E-08	2.65E-08	2.65E-08
steady-state cell-water concentration ratios							
Hepatocyte to bile	7.28						
Kidney to filtrate	6.19						
Enterocyte to gut lumen	3.75						

Appendix Table 5. The *in vitro* net flux for different transporters.

Transporters	Net flux (nmol/mg protein/min)	References
Oat1	0.34	Weaver et al. (2010)[80]
Oat3	0.48	
Oatp1a1	0.35	
Ntcp	0.10	Zhao et al. (2015)[83]
Ostα/β	0.41	

Appendix A.2 PBTK Model Results

Appendix Table 6. Correlation analysis between parameters and estimated PFOA levels in blood, liver, and kidney.

Parameters[a]	Blood[b]	Liver[b]	Kidney[b]
Dose	-	0.13	0.11
Blood volume	0.18	-	-
Plasma volume	-0.16	-	-
Glomerular filtration rate	-0.26	-0.25	-
Urine flow rate	-0.28	-0.25	-0.36
Capillary surface area of kidney	0.27	0.26	-
Enlargement factor of apical membrane of proximal tubule	-	-	0.16
Effective permeability for blood	0.24	0.23	-0.17
Effective permeability for liver	0.10	-0.34	0.11
Effective permeability for kidney	-	-	-0.17
Steady-state cell-water concentration ratio for kidney	0.25	0.24	0.38
Albumin concentration in blood	0.29	-	-
Albumin concentration in kidney fluid	-	-	0.13
LFABP concentration in liver tissue	-	0.33	-0.10
LFABP concentration in kidney tissue	-	-	0.12
Association constant of LFABP	-	0.22	-
Association constant of albumin	0.35	-	0.15
Renal clearance rate constant	-0.16	-0.16	-
Renal reabsorption rate constant	0.26	0.26	0.39
Renal efflux rate constant	0.17	0.16	-
Absorption rate constant of hepatocyte	-0.11	0.33	-0.12

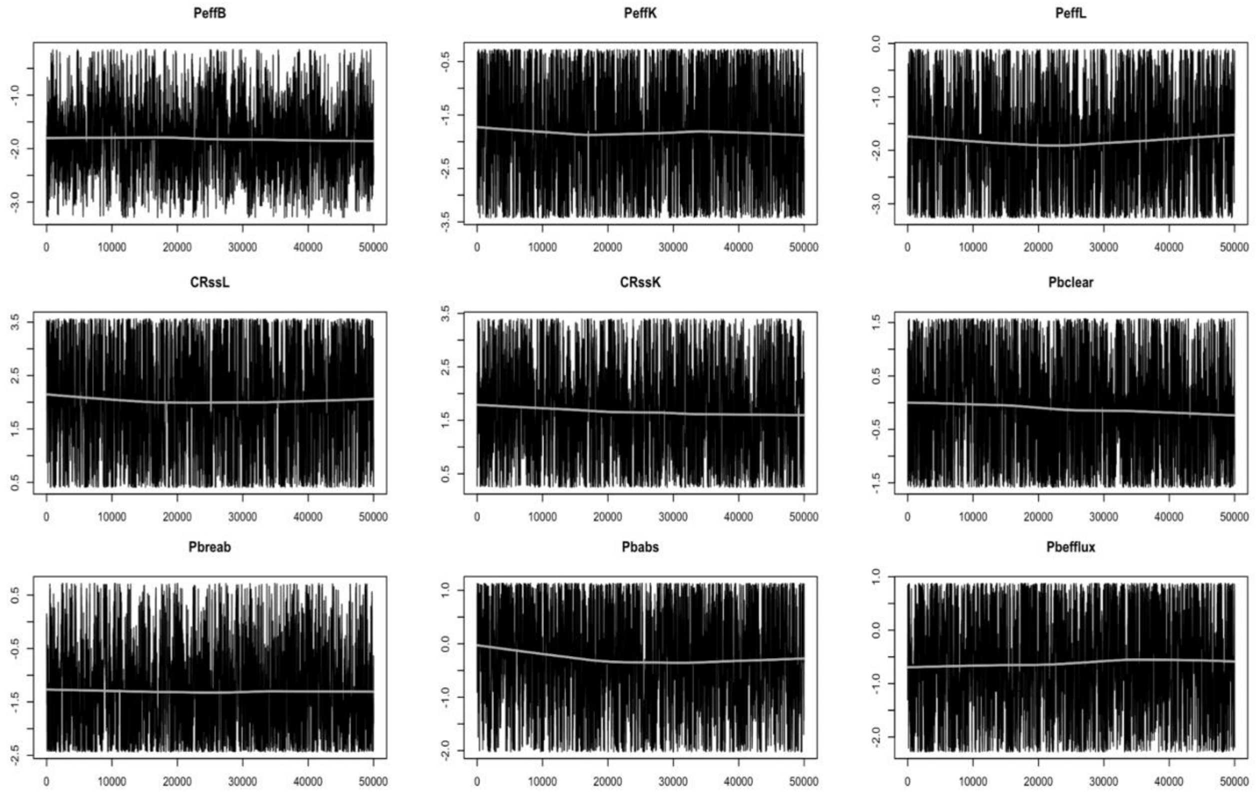
[a] Only parameters with correlation coefficient ≥ 0.1 and P value < 0.05 are indicated here.

[b] the coefficient values are the average of three simulation results at different dose and administration routes (1 mg/kg oral dose, 1 mg/kg IV dose and 0.1 mg/kg oral dose), given that the sensitivities were similar for the three scenarios.

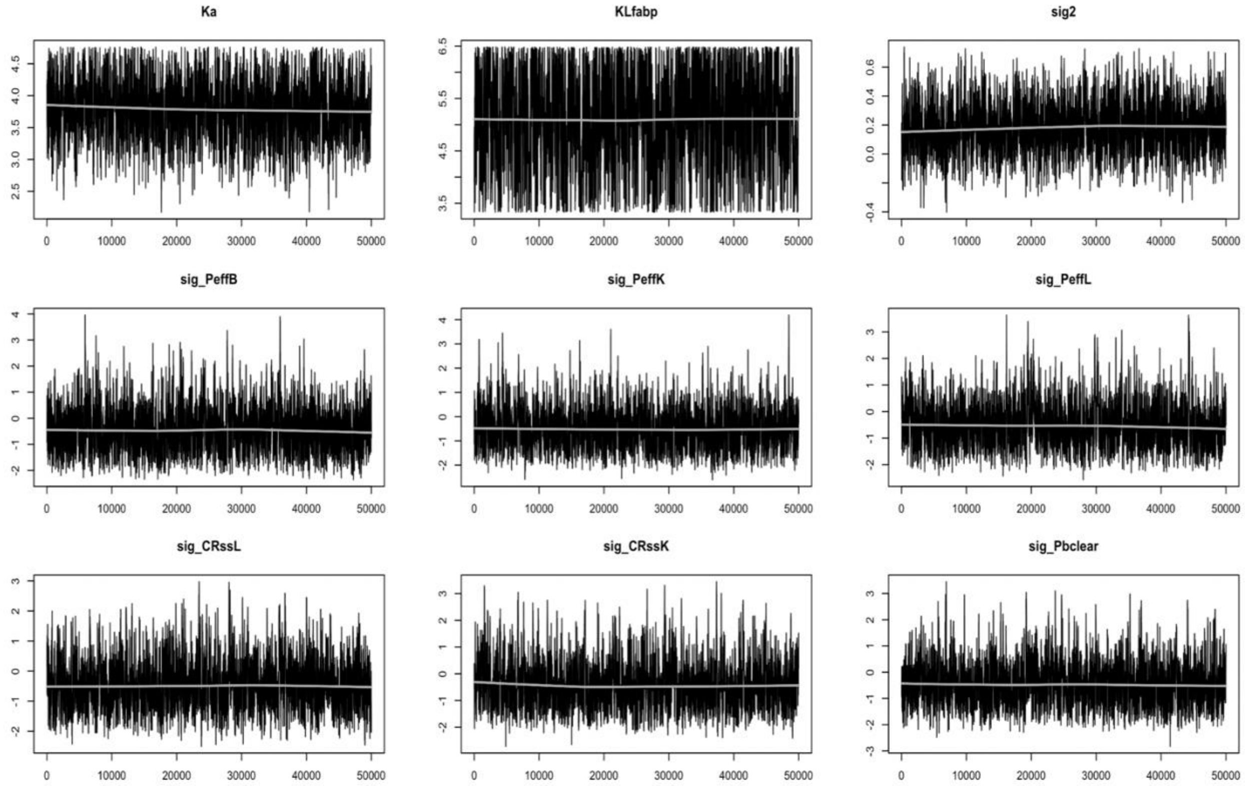
Appendix Table 7. The potential scale reduction factor (PSRF) and multivariate PSRF values of population mean and population variance for each selected model parameter.

Parameters	Population Mean		Population Variance	
	point estimation of PSRF	upper confidence interval	point estimation of PSRF	upper confidence interval
PeffB	1.002	1.003	1.038	1.086
PeffK	1.008	1.025	1.005	1.011
PeffL	1.010	1.031	1.007	1.017
CRssL	1.002	1.008	1.003	1.009
CRssK	1.013	1.039	1.011	1.028
Pbclear	1.011	1.032	1.003	1.009
Pbreab	1.020	1.056	1.013	1.035
Pbabs	1.008	1.023	1.008	1.023
Pbefflux	1.002	1.005	1.002	1.007
$\mathrm{K_a^{Alb}}$	1.005	1.015	1.001	1.003
$\mathrm{K_a^{LFABP}}$	1.009	1.026	1.009	1.029
Model Error	1.001	1.002		

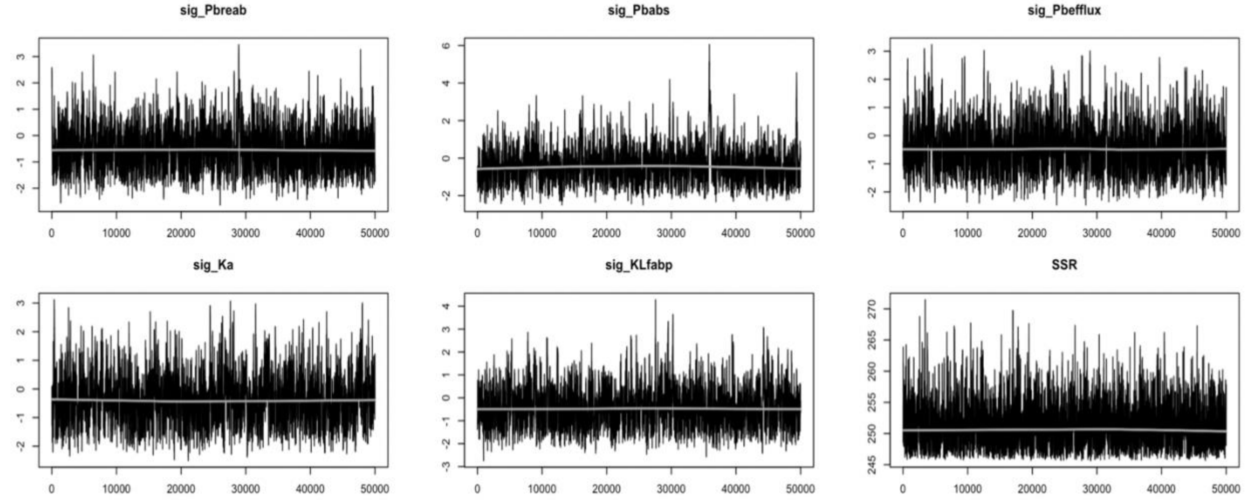
Note: Multivariate PSRF = 1.09



Appendix Figure 1. Trace plot of the last 50000 iterations from the MCMC simulation for selected model parameters.

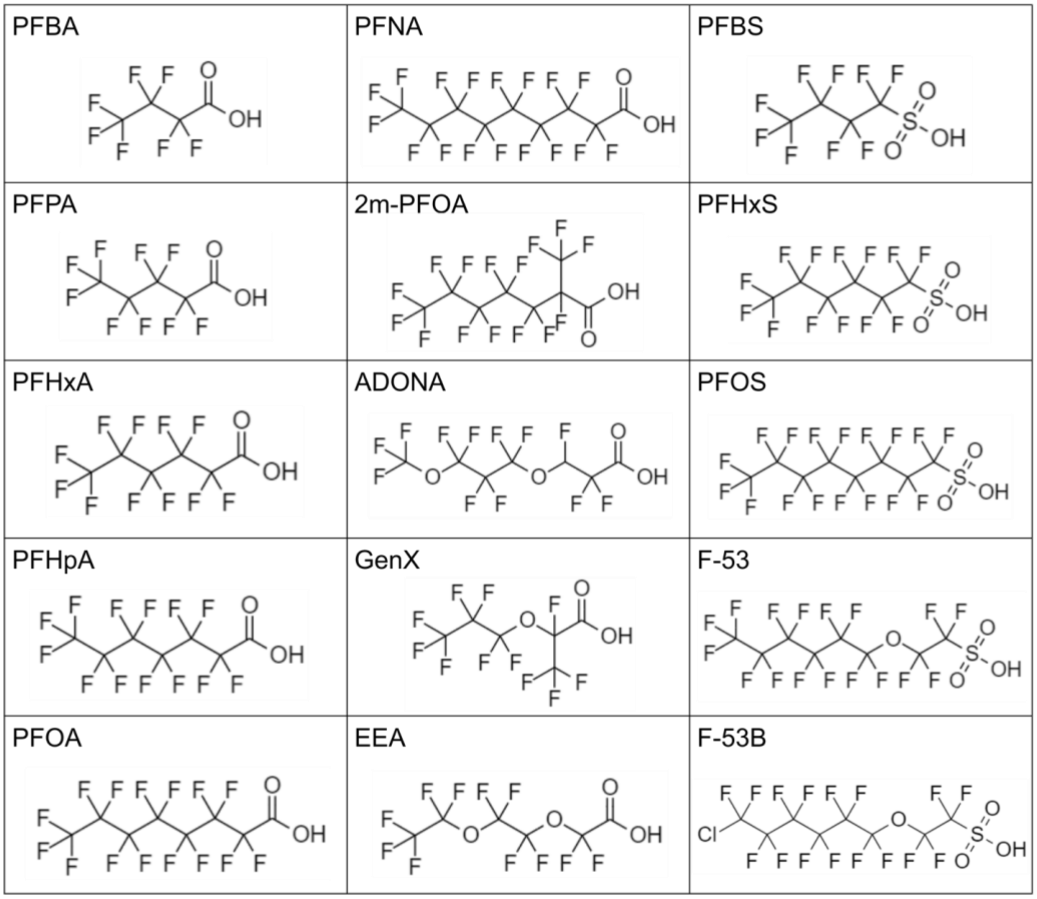


Appendix Figure 2. Trace plot of the last 50000 iterations from the MCMC simulation for selected model parameters.



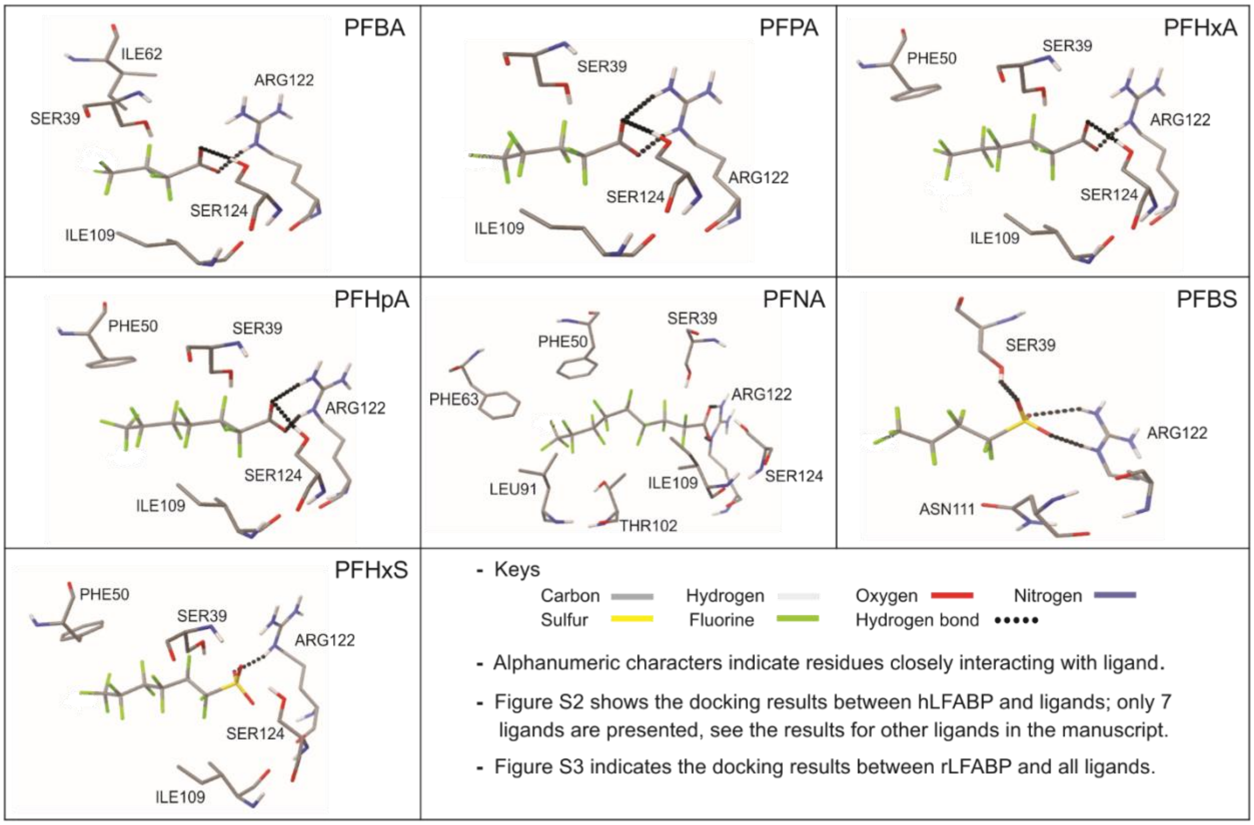
Appendix Figure 3. Trace plot of the last 50000 iterations from the MCMC simulation for selected model parameters.

Appendix B Supporting Information for Chapter 3.0



Appendix Figure 4. The chemical structures of PFAS ligands.

PFBA: perfluorobutanoic acid; PFPA: perfluoropentanoic acid; PFHxA: perfluorohexanoic acid; PFHpA: perfluoroheptanoic acid; PFOA: linear perfluorooctanoic acid; PFNA: perfluorononanoic acid; 2m-PFOA: perfluoro-2-methylheptanoic acid; ADONA: 3H-perfluoro-3-[(3-methoxy-propoxy)propanoic acid]; GenX: hexafluoropropylene oxide dimer acid; EEA: perfluoro-3,6-dioxaoctanoic acid; PFBS: perfluorobutane sulfonate; PFHxS: perfluorohexane sulfonate; PFOS: perfluorooctane sulfonate; F-53: 6:2 polyfluorinated ether sulfonate; F-53B: 6:2 chlorinated polyfluorinated ether sulfonate. All structures were generated from the online ChemDraw tool (https://chemdrawdirect.perkinelmer.cloud/js/sample/index.html).

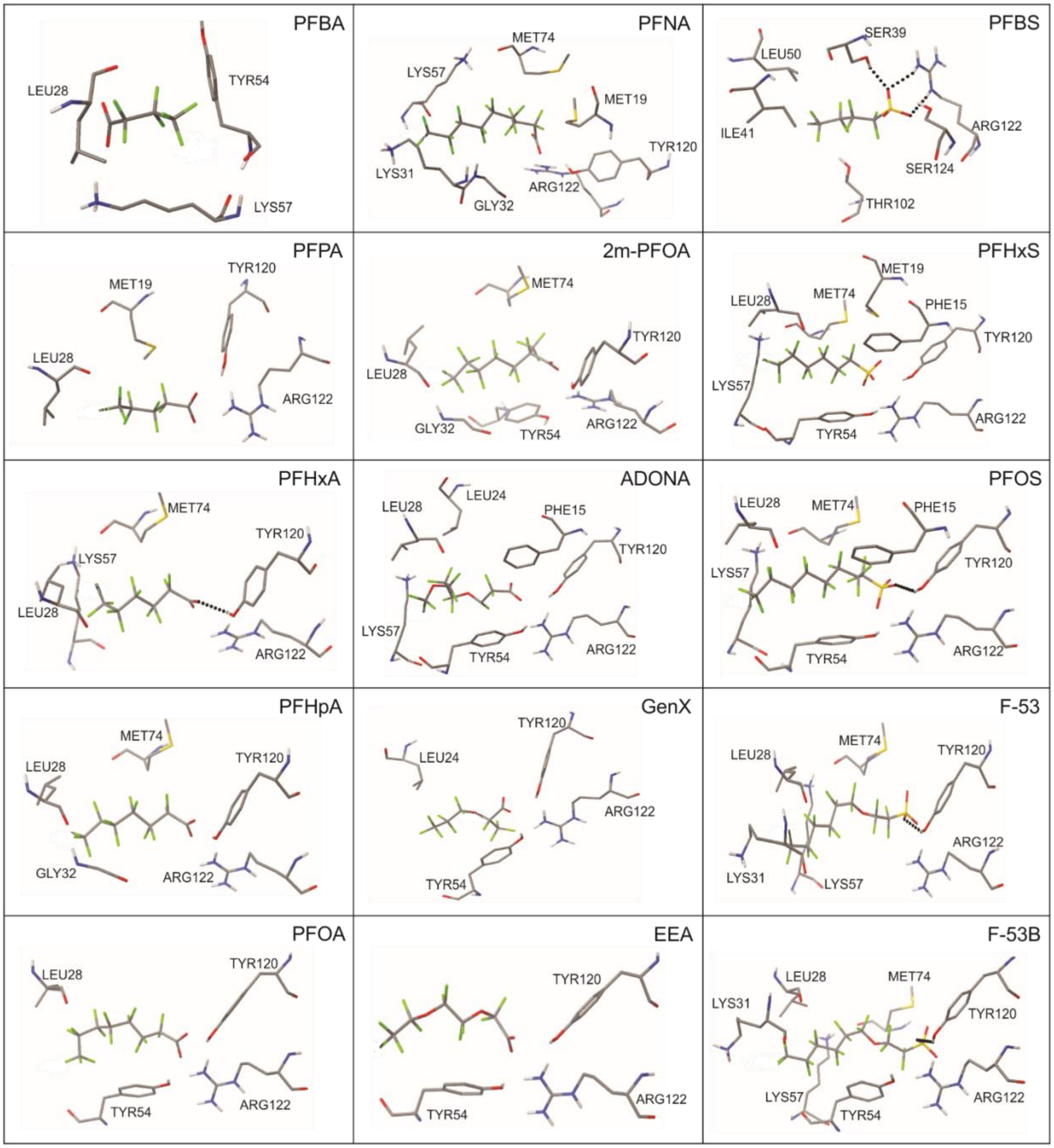


- Keys

  Carbon ▬ Hydrogen ▬ Oxygen ▬ Nitrogen ▬
  Sulfur ▬ Fluorine ▬ Hydrogen bond •••••

- Alphanumeric characters indicate residues closely interacting with ligand.
- Figure S2 shows the docking results between hLFABP and ligands; only 7 ligands are presented, see the results for other ligands in the manuscript.
- Figure S3 indicates the docking results between rLFABP and all ligands.

Appendix Figure 5. The interactions between human LFABP and PFAS ligands.



Appendix Figure 6. The interactions between rat LFABP and PFAS ligands.

Appendix Table 8. The mean of ΔG_{bind} and its five energy components calculated based on MM-PBSA for human LFABP.

Ligand	van der Waals	Electrostatic	Polar Solvation Energy	Nonpolar Solvation Energy	Entropy	Delta G
PFBA	-16.24	-78.09	72.93	-2.03	-16.93	-6.50
PFPA	-16.95	-69.17	68.98	-2.26	-17.20	-2.21
PFHxA	-19.74	-113.98	101.52	-2.55	-18.98	-15.76
PFHpA	-22.53	-76.66	78.90	-2.95	-18.13	-5.11
PFOA	-25.85	-89.69	91.67	-3.24	-19.13	-7.98
PFNA	-28.87	-102.23	101.09	-3.46	-20.55	-12.92
EEA	-25.21	-103.13	97.77	-3.01	-20.41	-13.16
GenX	-23.10	-65.18	67.61	-2.81	-18.61	-4.87
ADONA	-27.34	-108.41	105.77	-3.10	-22.21	-10.88
2m-PFOA	-26.33	-72.53	75.07	-3.20	-19.24	-7.75
PFBS	-23.96	-90.75	90.73	-2.44	-18.51	-7.90
PFHxS	-25.76	-88.32	88.32	-3.02	-18.12	-10.65
PFOS	-31.97	-88.90	90.63	-3.58	-19.82	-13.99
F-53B	-36.59	-84.63	88.79	-3.83	-22.58	-13.68
F-53	-33.59	-93.55	93.90	-3.80	-21.57	-15.47

Appendix Table 9. The mean of ΔG_{bind} and its five energy components calculated based on MM-PBSA for rat LFABP.

Ligand	van der Waals	Electrostatic	Polar Solvation Energy	Nonpolar Solvation Energy	Entropy	Delta G
PFBA	-13.87	-98.03	93.35	-1.86	-14.39	-6.02
PFPA	-15.72	-95.69	93.73	-2.23	-15.07	-4.85
PFHxA	-19.64	-100.33	98.76	-2.61	-16.58	-7.24
PFHpA	-21.70	-102.70	101.94	-2.94	-17.19	-8.21
PFOA	-25.48	-105.33	106.19	-3.21	-17.48	-10.34
PFNA	-28.43	-104.58	108.25	-3.50	-18.03	-10.23
EEA	-24.92	-106.84	107.72	-3.06	-19.59	-7.51
GenX	-20.95	-103.30	104.88	-2.83	-18.47	-3.74
ADONA	-25.80	-108.59	110.51	-3.20	-18.85	-8.23
2m-PFOA	-26.86	-104.63	106.51	-3.17	-18.55	-9.59
PFBS	-20.63	-101.67	99.92	-2.46	-16.66	-8.18
PFHxS	-26.84	-101.73	105.69	-3.07	-18.39	-7.57
PFOS	-30.56	-92.36	98.84	-3.64	-17.76	-9.96
F-53B	-35.21	-107.02	112.68	-3.86	-22.15	-11.26
F-53	-33.13	-105.26	109.71	-3.78	-21.39	-11.07

Appendix C Supporting Information for Chapter 4.0

In this section, additional information is provided for Chapter 4.0 on ROC-AUC score for different machine learning models, hyperparameters for each model, description of the receptors (bioassays), structural information of the PFAS in the OECD list and chemical structures for highly bioactive PFAS. In addition, the data files including the mean and standard deviation of the AUC score of each bioassay task for each machine learning model, the CF data set, the C3F6 data set, and the target names and targe classes for each bioassay target ID are available in the following link: https://pubs.acs.org/doi/abs/10.1021/acs.est.9b04833.

Appendix Table 10. ROC-AUC scores for different machine learning models.

Models	**Mean of ROC-AUC scores**				**Machine**
	C3F6 dataset		**CF dataset**		
	training	**validation**	**training**	**validation**	
Logistic Regression	0.989 ± 0.002	0.742 ± 0.006	0.920 ± 0.000	0.773 ± 0.023	CPU-8 (GCP)
Support Vector Machine	0.932 ± 0.037	0.719 ± 0.024	-	-	CPU-8 (GCP)
Random Forests	1.000 ± 0.000	0.707 ± 0.028	1.000 ± 0.000	0.777 ± 0.022	CPU-8 (GCP)
1-hidden-layer Multitask Network	0.999 ± 0.001	0.753 ± 0.027	0.993 ± 0.000	**0.821 ± 0.014**	CPU-8 (GCP)
Pyramidal Multitask Network	0.999 ± 0.000	**0.772 ± 0.021**	0.990 ± 0.001	0.803 ± 0.015	CPU-8 (GCP)
Graph Convolutional Model	0.992 ± 0.002	0.767 ± 0.024	0.977 ± 0.014	0.813 ± 0.013	GPU-1 (GCP)
Weave Model	0.982 ± 0.012	0.765 ± 0.016	0.967 ± 0.005	0.765 ± 0.018	GPU-1 (CRC)

Notes: Validation scores for best-performing models for each dataset in bold. GCP: Google Cloud Platform; CRC: University of Pittsburgh Center for Research Computing. The ROC-AUC score (Receiver Operating Characteristic-Area Under the Curve) is used to assess the performance of classification models; and ROC-AUC of 1 is perfect classification, whereas an ROC-AUC of 0.5 means a model cannot separate between the two classes at all.

Appendix Table 11. Summary of the hyperparameters for each machine learning model.

Model	Hyperparameters	
	C3F6 dataset	CF dataset
Logistic Regression	**penalty**: 30.26, penalty_type: 'l2'	**penalty**: 86.7, penalty_type: 'l2'
Support Vector Machine	**C**: 67.7, *gamma*: 0.01	-
Random Forests	**n_estimators**: 250	n_estimators: 250
1-hidden-layer Multitask Network	layer_sizes: [1500], weight_init_stddevs: [0.02], bias_init_consts: [1.], **dropouts**: [0.26], penalty: 0.1, penalty_type: 'l2', batch_size: 50, *nb_epoch*: 100, **learning_rate**: 0.007	layer_sizes: [1500], weight_init_stddevs: [0.02], bias_init_consts: [1.], *dropouts*: [0.25], penalty: 0.1, penalty_type: 'l2', batch_size: 50, *nb_epoch*: 100, *learning_rate*: 0.0005
Pyramidal Multitask Network (2 hidden layer)	layer_sizes: [2000, 100], weight_init_stddevs: [0.02, 0.02], bias_init_consts: [1., 1.], *dropouts*: [0.25, 0.1], penalty: 0.1, penalty_type: 'l2', batch_size: 50, *nb_epoch*: 100, **learning_rate**: 0.0005	layer_sizes: [2000, 100], weight_init_stddevs: [0.02, 0.02], bias_init_consts: [1., 1.], dropouts: [0.25, 0.1], penalty: 0.1, penalty_type: 'l2', batch_size: 50, *nb_epoch*: 100, *learning_rate*: 0.0005
Graph Convolutional Model	**batch_size**: 124, *nb_epoch*: 50, **learning_rate**: 0.001, **n_filters**: 227, **n_fully_connected_nodes**: 312	batch_size: 124, *nb_epoch*: 250, learning_rate: 0.001, n_filters: 227, n_fully_connected_nodes: 312
Weave Model	batch_size: 64, *nb_epoch*: 200, **learning_rate**: 0.00028, n_graph_feat: 128, n_pair_feat: 14	batch_size: 64, *nb_epoch*: 500, *learning_rate*: 0.00005, n_graph_feat: 128, n_pair_feat: 14

Notes:

(1) Bold parameters were tuned with Bayesian Optimization. *Italic parameters* were tuned with Grid Search. For other parameters, the default values of each model in DeepChem were used. Finally, underlined parameters for the CF dataset were taken from the C3F6 dataset without further optimization, given that the training process is much more computationally expensive for the large CF dataset.

(2) the meaning of each hyperparameter is available in the documentation of the DeepChem package (https://www.deepchem.io/docs/deepchem.html).

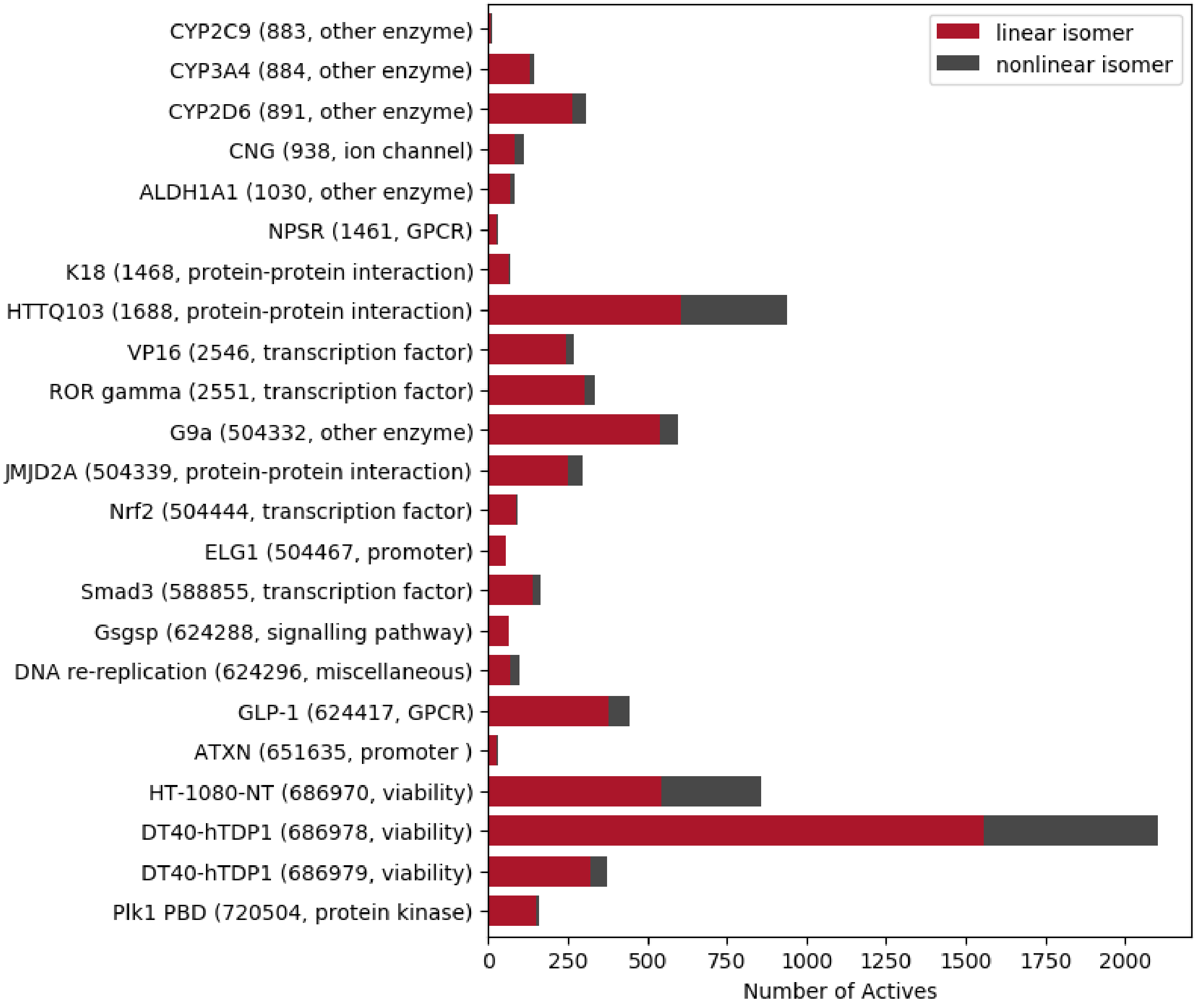
Appendix Table 12. Descriptions for the 26 receptors selected for showing some activity for the chemicals in the C3F6 dataset.

Target Class	Target	Description
G protein-coupled receptors (GPCRs)	NPSR	The neuropeptide S receptor (NPSR), which is highly expressed in brain areas involving modulation of arousal, stress and anxiety, could be a novel drug target for the treatment of sleep and anxiety disorders. This assay is conducted to identify NPSR antagonists.
	GLP-1	The overall aim of this assay is to discover ligands for class B1 GPCRs. Specifically, this assay focused on class B1 receptor for glucagon-like peptide-1 (GLP-1), which is a potential therapeutic target for diabetes and neurodegenerative disease.
ion channel	CNG	The cyclic nucleotide gated (CNG) ion channel was used as a biosensor for cAMP induction in this assay. The rationale is that cAMP stimulation will cause the CNG ion channel to open and subsequent membrane depolarization to occur.
miscellaneous	Marburg virus	Marburg virus could cause Marburg hemorrhagic fever in humans, this assay was developed to identify inhibitors that block the virus binding or entry into cells.
	DNA re-replication	This assay is used to screen small molecules that induce DNA re-replication, which can cause the DNA damage response, arrest cell proliferation, and trigger apoptosis.
	orthopoxvirus*	This assay (mCherry Reporter Primary qHTS) is used to screen molecules that inhibit Orthopox viruses, which are a genus of viruses including monkeypox and variola (the causative agent of smallpox).
	orthopoxvirus*	Another assay (Venus Reporter Primary qHTS) used to screen molecules that inhibit Orthopox viruses.
other enzymes	CYP2C9	Cytochromes P450 (CYP) are a group of heme-thiolate monooxygenases that oxidize a variety of substances including steroids, fatty acids, and xenobiotics. In these assays, three different CYPs (CYP2C9, CYP3A4, and CYP2D6) were used to screen inhibitors and substrates for those CYP enzymes.
	CYP3A4	
	CYP2D6	
	ALDH1A1	Aldehyde dehydrogenase 1 (ALDH1A1) is an enzyme that oxidizes a variety of endogenous and exogenous aldehydes to the corresponding carboxylic acids and is the critical step for retinoic acid metabolism. In this assay, inhibitors of ALDH1A1 were identified.
	G9a	G9a is a histone methyltransferase that is responsible for histone H3 lysine 9 (H3K9) mono- and di-methylation. It has been recognized as a potential drug target for several human diseases, including cancer. The goal of this assay is to identify inhibitors of G9a.
promoter	ELG1	As the major subunit of a Replication Factor C-like complex, ELG1 is critical to ensure genomic stability during DNA replication. This assay identifies small molecules that block ELG1 function.
	ATXN	Ataxin-2 protein (ATXN2) is encoded by the ATXN2 gene. The mutation in ATXN2 could cause Spinocerebellar ataxia type 2 (SCA2) disease. The objective of this assay is to identify compounds that inhibit the expression of ATXN2.
protein kinase	Plk1 PBD	Polo-like kinase 1 (Plk1) is a member of a conserved subfamily of serine / threonine protein kinases and plays a central role in cell proliferation. Plk1 is a potential target for anti-cancer therapy. This assay aimed to identify inhibitors that target the Plk1 polo-box domain (PBD).

Appendix Table 12 (continued)

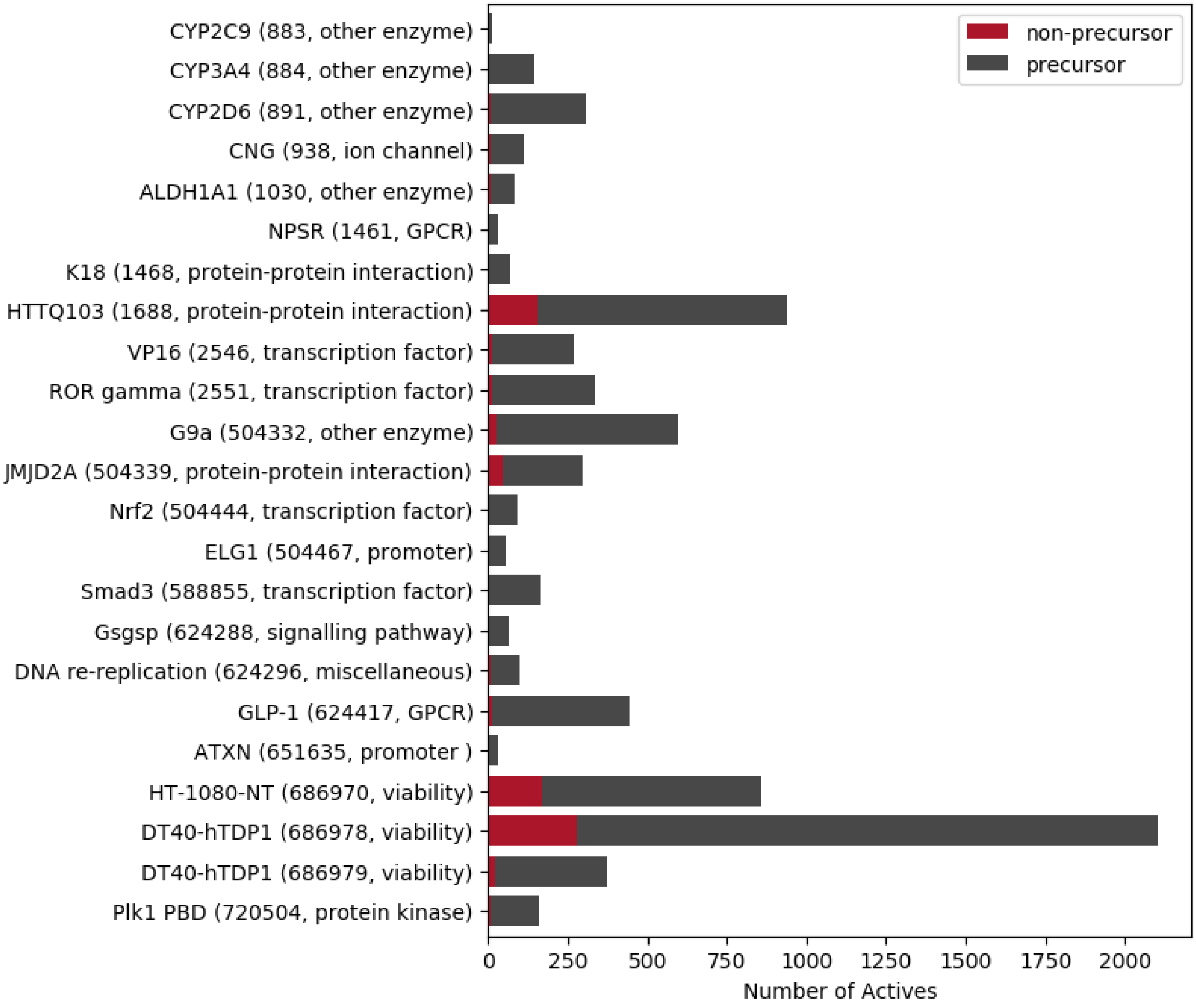
protein-protein interaction	K18	In this assay, a recombinantly expressed fragment of tau, K18 was used to identify inhibitors of tau (which is an abundant protein in the axons of neurons that stabilizes microtubules) aggregation.
	HTTQ103	When exon 1 of HTTQ103 (Huntingtin protein containing 103 polyglutamines expansion) is expressed, it causes cell death and GFP aggregates. This assay screens for small molecules that reduce aggregate formation.
	JMJD2A	JMJD2A is a jumonji-domain-containing lysine demethylase. In this assay, the inhibitors of JMJD2A-tudor domain interactions (which is helpful in probing the regulatory pathways of selective demethylation of a given methyllysine locus) were identified.
signaling pathway	Gsgsp	The objective of this assay is to identify molecules with inhibitory activity at gsp mutations, which are responsible for McCune-Albright syndrome.
transcription factor	ROR gamma	The goal of this assay is to identify small molecules that inhibit ROR (retinoic acid-related orphan receptor) gamma activity.
	VP16	The goal of this assay is to identify small molecules that inhibit components common to both ROR gamma and VP16 transcription factor.
	Nrf2	Nrf2 is a transcription factor that maintains cellular redox homeostasis and protects cells from xenobiotics. This assay is used to screen inhibitors of Nrf2 function, which could be potential therapeutic targets for improvement in cancer treatment.
	Smad3	TGF-b signaling pathway plays important roles in cellular and development pathways. Smad3 is the primary transducer of TGF-b's signals and regulates many functions related to TGF-b signaling. The goal of this assay is to identify Smad3-small molecule antagonists.
viability	HT-1080-NT	In this assay, a synthetic lethal screen was conducted for chemical probes specific for 2HG-producing tumor cells using HT-1080-NT fibrosarcoma cell line.
	DT40-hTDP1*	Human tyrosyl-DNA phosphodiesterase 1 (HTDP1) is a novel repair gene and can be used as a new target for anti-cancer drug development. In this assay, after exposure to small molecules in the absence of camptothecin, the growth kinetics of DT40-hTDP1 cells were evaluated to determine whether the molecules can inhibit the TDP1-mediated repair pathway.
	DT40-hTDP1*	In this assay, after exposure to small molecules in the presence of camptothecin, the growth kinetics of DT40-hTDP1 cells were evaluated to determine whether the molecules can inhibit the TDP1-mediated repair pathway.

Notes: The description of each receptor is obtained from National Center for Biotechnology Information, PubChem Database (https://pubchem.ncbi.nlm.nih.gov/). Those rows in grey indicate receptors that are of less relevance to PFAS-related human toxicity (e.g. viral assays) that are therefore not included in the discussion of results.



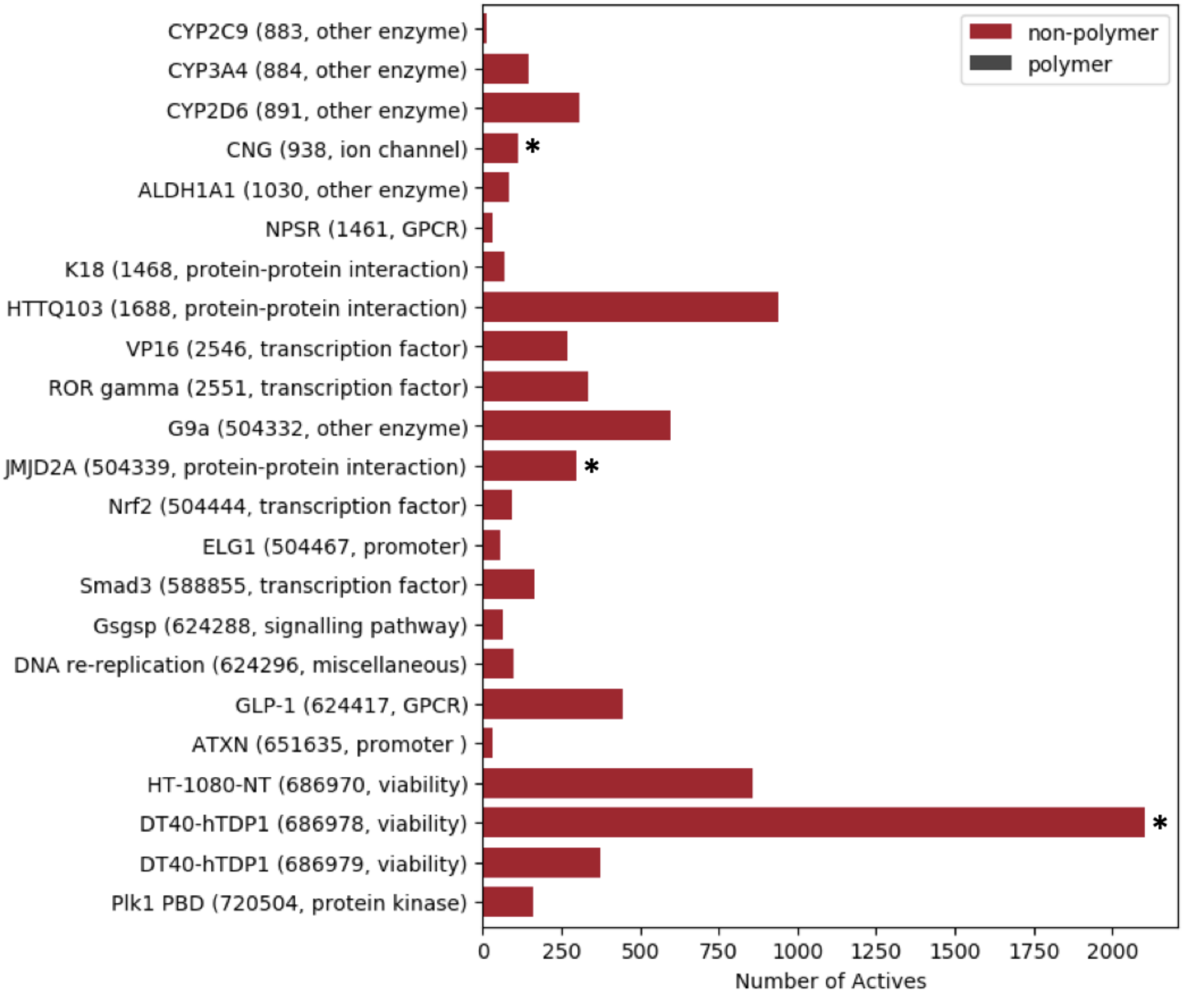
Appendix Figure 7. The biological activity information of all 3486 PFAS (2688 linear vs 798 nonlinear isomer) in the processed OECD dataset for the 23 targets.

Inside the parentheses, the PubChem AID and the target class are indicated.



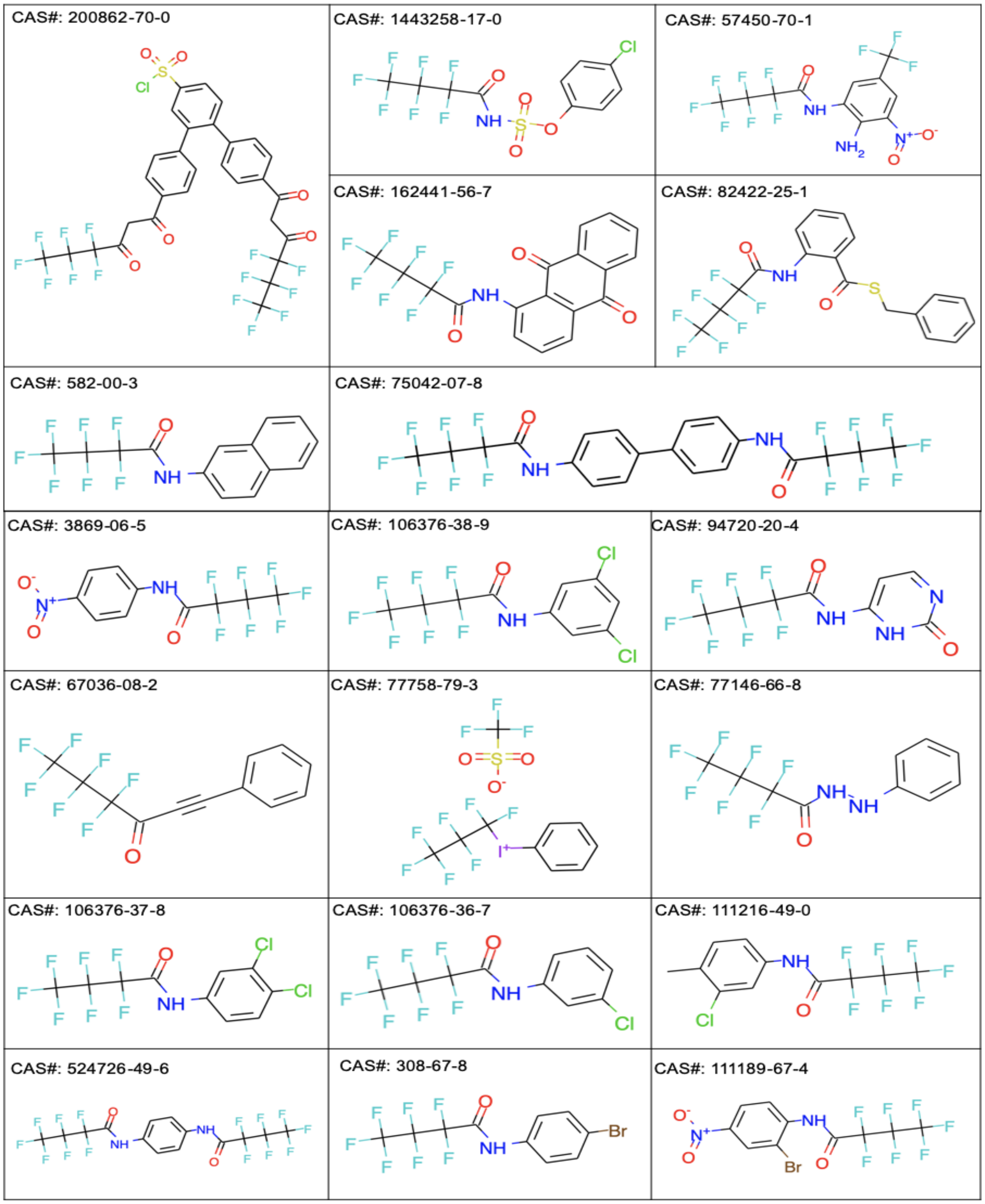
Appendix Figure 8. The biological activity information of all 3486 PFAS (3119 precursor vs 367 non-precursor) in the processed OECD dataset for the 23 targets.

Inside the parentheses, the PubChem AID and the target class are indicated.



Appendix Figure 9. The biological activity information of all 3486 PFAS (4 polymer vs 3482 non-polymer) in the processed OECD dataset for the 23 targets.

Inside the parentheses, the PubChem AID and the target class are indicated); Asterisk (*) indicates that there is 1 polymer corresponding to that target.



Appendix Figure 10. The chemical structures for PFAS that indicate biological activity for at least half of the targets investigated in this study.

The chain length of each individual perfluoroalkyl substructure in the PFAS presented in this figure is no longer than 3; for longer-chain PFAS see Figure 21 in Chapter 4.0.

Bibliography

(1) Place, B. J.; Field, J. A. Identification of novel fluorochemicals in aqueous film-forming foams used by the US military. *Environmental science & technology* **2012,** *46* (13), 7120-7127.

(2) Buck, R. C.; Franklin, J.; Berger, U.; Conder, J. M.; Cousins, I. T.; De Voogt, P.; Jensen, A. A.; Kannan, K.; Mabury, S. A.; van Leeuwen, S. P. Perfluoroalkyl and polyfluoroalkyl substances in the environment: terminology, classification, and origins. *Integrated environmental assessment and management* **2011,** *7* (4), 513-541.

(3) Kannan, K. Perfluoroalkyl and polyfluoroalkyl substances: current and future perspectives. *Environmental chemistry* **2011,** *8* (4), 333-338.

(4) D'eon, J. C.; Mabury, S. A. Is indirect exposure a significant contributor to the burden of perfluorinated acids observed in humans? *Environmental science & technology* **2011,** *45* (19), 7974-7984.

(5) Liu, X.; Guo, Z.; Folk IV, E. E.; Roache, N. F. Determination of fluorotelomer alcohols in selected consumer products and preliminary investigation of their fate in the indoor environment. *Chemosphere* **2015,** *129*, 81-86.

(6) Kotthoff, M.; Müller, J.; Jürling, H.; Schlummer, M.; Fiedler, D. Perfluoroalkyl and polyfluoroalkyl substances in consumer products. *Environmental Science and Pollution Research* **2015,** *22* (19), 14546-14559.

(7) OECD Summary report on the new comprehensive global database of Per- and Polyfluoroalkyl Substances (PFASs). Publications Series on Risk Management No. 39. **2018**.

(8) Wang, Z.; DeWitt, J. C.; Higgins, C. P.; Cousins, I. T., A never-ending story of per-and polyfluoroalkyl substances (PFASs)? In ACS Publications: 2017.

(9) Houtz, E. F.; Sutton, R.; Park, J.-S.; Sedlak, M. Poly-and perfluoroalkyl substances in wastewater: Significance of unknown precursors, manufacturing shifts, and likely AFFF impacts. *Water research* **2016,** *95*, 142-149.

(10) Krafft, M. P.; Riess, J. G. Per-and polyfluorinated substances (PFASs): Environmental challenges. *Current Opinion in Colloid & Interface Science* **2015,** *20* (3), 192-212.

(11) Goss, K.-U. The p K a values of PFOA and other highly fluorinated carboxylic acids. *Environmental science & technology* **2007,** *42* (2), 456-458.

(12) Armitage, J. M.; Erickson, R. J.; Luckenbach, T.; Ng, C. A.; Prosser, R. S.; Arnot, J. A.; Schirmer, K.; Nichols, J. W. Assessing the bioaccumulation potential of ionizable organic

compounds: Current knowledge and research priorities. *Environmental toxicology and chemistry* **2017,** *36* (4), 882-897.

(13) Ng, C. A.; Hungerbühler, K. Bioaccumulation of perfluorinated alkyl acids: observations and models. *Environmental science & technology* **2014,** *48* (9), 4637-4648.

(14) Olsen, G. W.; Burris, J. M.; Ehresman, D. J.; Froehlich, J. W.; Seacat, A. M.; Butenhoff, J. L.; Zobel, L. R. Half-life of serum elimination of perfluorooctanesulfonate, perfluorohexanesulfonate, and perfluorooctanoate in retired fluorochemical production workers. *Environmental health perspectives* **2007,** *115* (9), 1298.

(15) Abbott, B. D.; Wolf, C. J.; Das, K. P.; Zehr, R. D.; Schmid, J. E.; Lindstrom, A. B.; Strynar, M. J.; Lau, C. Developmental toxicity of perfluorooctane sulfonate (PFOS) is not dependent on expression of peroxisome proliferator activated receptor-alpha (PPARα) in the mouse. *Reprod. Toxicol.* **2009,** *27* (3-4), 258-265.

(16) Ankley, G. T.; Kuehl, D. W.; Kahl, M. D.; Jensen, K. M.; Linnum, A.; Leino, R. L.; Villeneuve, D. A. Reproductive and developmental toxicity and bioconcentration of perfluorooctanesulfonate in a partial life ‐ cycle test with the fathead minnow (Pimephales promelas). *Environmental toxicology and chemistry* **2005,** *24* (9), 2316-2324.

(17) Austin, M. E.; Kasturi, B. S.; Barber, M.; Kannan, K.; MohanKumar, P. S.; MohanKumar, S. M. Neuroendocrine effects of perfluorooctane sulfonate in rats. *Environmental Health Perspectives* **2003,** *111* (12), 1485.

(18) Fair, P. A.; Driscoll, E.; Mollenhauer, M. A.; Bradshaw, S. G.; Yun, S. H.; Kannan, K.; Bossart, G. D.; Keil, D. E.; Peden-Adams, M. M. Effects of environmentally-relevant levels of perfluorooctane sulfonate on clinical parameters and immunological functions in B6C3F1 mice. *J. Immunotoxicol.* **2011,** *8* (1), 17-29.

(19) Macon, M. B.; Villanueva, L. R.; Tatum-Gibbs, K.; Zehr, R. D.; Strynar, M. J.; Stanko, J. P.; White, S. S.; Helfant, L.; Fenton, S. E. Prenatal perfluorooctanoic acid exposure in CD-1 mice: low-dose developmental effects and internal dosimetry. *Toxicol. Sci.* **2011,** *122* (1), 134-145.

(20) Shi, G.; Cui, Q.; Pan, Y.; Sheng, N.; Sun, S.; Guo, Y.; Dai, J. 6: 2 Chlorinated polyfluorinated ether sulfonate, a PFOS alternative, induces embryotoxicity and disrupts cardiac development in zebrafish embryos. *Aquat. Toxicol.* **2017,** *185*, 67-75.

(21) Wang, M.; Chen, J.; Lin, K.; Chen, Y.; Hu, W.; Tanguay, R. L.; Huang, C.; Dong, Q. Chronic zebrafish PFOS exposure alters sex ratio and maternal related effects in F1 offspring. *Environmental toxicology and chemistry* **2011,** *30* (9), 2073-2080.

(22) (NTP), N. T. P. NTP technical report on the toxicity studies of perfluoroalkyl sulfonates (perfluorobutane sulfonic acid, perfluorohexane sulfonate potassium salt, and perfluorooctane sulfonic acid) administered by gavage to Sprague Dawley (Hsd:Sprague Dawley SD) rats. **2019**.

(23) (NTP), N. T. P. NTP technical report on the toxicity studies of perfluoroalkyl carboxylates (perfluorohexanoic acid, perfluorooctanoic acid, perfluorononanoic acid, and perfluorodecanoic acid) administered by gavage to Sprague Dawley (Hsd:Sprague Dawley SD) rats. **2019**.

(24) Wang, Z.; Cousins, I. T.; Scheringer, M.; Hungerbuehler, K. Hazard assessment of fluorinated alternatives to long-chain perfluoroalkyl acids (PFAAs) and their precursors: status quo, ongoing challenges and possible solutions. *Environment international* **2015,** *75*, 172-179.

(25) Wang, Z.; Cousins, I. T.; Scheringer, M.; Hungerbühler, K. Fluorinated alternatives to long-chain perfluoroalkyl carboxylic acids (PFCAs), perfluoroalkane sulfonic acids (PFSAs) and their potential precursors. *Environment international* **2013,** *60*, 242-248.

(26) Strynar, M.; Dagnino, S.; McMahen, R.; Liang, S.; Lindstrom, A.; Andersen, E.; McMillan, L.; Thurman, M.; Ferrer, I.; Ball, C. Identification of novel perfluoroalkyl ether carboxylic acids (PFECAs) and sulfonic acids (PFESAs) in natural waters using accurate mass time-of-flight mass spectrometry (TOFMS). *Environmental science & technology* **2015,** *49* (19), 11622-11630.

(27) Krewski, D.; Acosta Jr, D.; Andersen, M.; Anderson, H.; Bailar III, J. C.; Boekelheide, K.; Brent, R.; Charnley, G.; Cheung, V. G.; Green Jr, S. Toxicity testing in the 21st century: a vision and a strategy. *Journal of Toxicology and Environmental Health, Part B* **2010,** *13* (2-4), 51-138.

(28) Yoon, M.; Campbell, J. L.; Andersen, M. E.; Clewell, H. J. Quantitative in vitro to in vivo extrapolation of cell-based toxicity assay results. *Crit. Rev. Toxicol.* **2012,** *42* (8), 633-652.

(29) Lin, Z.; Gehring, R.; Mochel, J.; Lavé, T.; Riviere, J. Mathematical modeling and simulation in animal health–Part II: principles, methods, applications, and value of physiologically based pharmacokinetic modeling in veterinary medicine and food safety assessment. *J. Vet. Pharmacol. Ther.* **2016,** *39* (5), 421-438.

(30) Loccisano, A. E.; Campbell, J. L.; Butenhoff, J. L.; Andersen, M. E.; Clewell, H. J. Comparison and evaluation of pharmacokinetics of PFOA and PFOS in the adult rat using a physiologically based pharmacokinetic model. *Reprod. Toxicol.* **2012,** *33* (4), 452-467.

(31) Tan, Y.-M.; Clewell, H. J.; Andersen, M. E. Time dependencies in perfluorooctylacids disposition in rat and monkeys: a kinetic analysis. *Toxicol. Lett.* **2008,** *177* (1), 38-47.

(32) Worley, R. R.; Fisher, J. Application of physiologically-based pharmacokinetic modeling to explore the role of kidney transporters in renal reabsorption of perfluorooctanoic acid in the rat. *Toxicology and applied pharmacology* **2015,** *289* (3), 428-441.

(33) Andersen, M. E.; Clewell, H. J.; Tan, Y.-M.; Butenhoff, J. L.; Olsen, G. W. Pharmacokinetic modeling of saturable, renal resorption of perfluoroalkylacids in monkeys—probing the determinants of long plasma half-lives. *Toxicology* **2006,** *227* (1), 156-164.

(34) Fàbrega, F.; Kumar, V.; Benfenati, E.; Schuhmacher, M.; Domingo, J. L.; Nadal, M. Physiologically based pharmacokinetic modeling of perfluoroalkyl substances in the human body. *Toxicol. Environ. Chem.* **2015,** *97* (6), 814-827.

(35) Fàbrega, F.; Kumar, V.; Schuhmacher, M.; Domingo, J. L.; Nadal, M. PBPK modeling for PFOS and PFOA: Validation with human experimental data. *Toxicol. Lett.* **2014,** *230* (2), 244-251.

(36) Loccisano, A. E.; Campbell, J. L.; Andersen, M. E.; Clewell, H. J. Evaluation and prediction of pharmacokinetics of PFOA and PFOS in the monkey and human using a PBPK model. *Regulatory toxicology and pharmacology* **2011,** *59* (1), 157-175.

(37) Woodcroft, M. W.; Ellis, D. A.; Rafferty, S. P.; Burns, D. C.; March, R. E.; Stock, N. L.; Trumpour, K. S.; Yee, J.; Munro, K. Experimental characterization of the mechanism of perfluorocarboxylic acids' liver protein bioaccumulation: The key role of the neutral species. *Environmental Toxicology and Chemistry* **2010,** *29* (8), 1669-1677.

(38) Bischel, H. N.; MacManus-Spencer, L. A.; Luthy, R. G. Noncovalent interactions of long-chain perfluoroalkyl acids with serum albumin. *Environmental science & technology* **2010,** *44* (13), 5263-5269.

(39) Han, X.; Snow, T. A.; Kemper, R. A.; Jepson, G. W. Binding of perfluorooctanoic acid to rat and human plasma proteins. *Chem. Res. Toxicol.* **2003,** *16* (6), 775-781.

(40) Han, X.; Yang, C.-H.; Snajdr, S. I.; Nabb, D. L.; Mingoia, R. T. Uptake of perfluorooctanoate in freshly isolated hepatocytes from male and female rats. *Toxicol. Lett.* **2008,** *181* (2), 81-86.

(41) Yang, C.-H.; Glover, K. P.; Han, X. Organic anion transporting polypeptide (Oatp) 1a1-mediated perfluorooctanoate transport and evidence for a renal reabsorption mechanism of Oatp1a1 in renal elimination of perfluorocarboxylates in rats. *Toxicol. Lett.* **2009,** *190* (2), 163-171.

(42) Yang, C.-H.; Glover, K. P.; Han, X. Characterization of cellular uptake of perfluorooctanoate via organic anion-transporting polypeptide 1A2, organic anion transporter 4, and urate transporter 1 for their potential roles in mediating human renal reabsorption of perfluorocarboxylates. *Toxicol. Sci.* **2010,** *117* (2), 294-302.

(43) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert opinion on drug discovery* **2015,** *10* (5), 449-461.

(44) Miller III, B. R.; McGee Jr, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. MMPBSA. py: an efficient program for end-state free energy calculations. *J. Chem. Theory Comput.* **2012,** *8* (9), 3314-3321.

(45) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R. QSAR modeling: where have you been? Where are you going to? *Journal of medicinal chemistry* **2014,** *57* (12), 4977-5010.

(46) Mitchell, J. B. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014,** *4* (5), 468-481.

(47) Varnek, A.; Baskin, I. Machine learning methods for property prediction in chemoinformatics: quo vadis? *J. Chem. Inf. Model.* **2012,** *52* (6), 1413-1437.

(48) Wang, Z.; Cousins, I. T.; Scheringer, M.; Buck, R. C.; Hungerbühler, K. Global emission inventories for C 4–C 14 perfluoroalkyl carboxylic acid (PFCA) homologues from 1951 to 2030, Part I: production and emissions from quantifiable sources. *Environment international* **2014,** *70*, 62-75.

(49) Lau, C.; Anitole, K.; Hodes, C.; Lai, D.; Pfahles-Hutchens, A.; Seed, J. Perfluoroalkyl acids: a review of monitoring and toxicological findings. *Toxicol. Sci.* **2007,** *99* (2), 366-394.

(50) Olsen, G. W.; Lange, C. C.; Ellefson, M. E.; Mair, D. C.; Church, T. R.; Goldberg, C. L.; Herron, R. M.; Medhdizadehkashi, Z.; Nobiletti, J. B.; Rios, J. A. Temporal trends of perfluoroalkyl concentrations in American Red Cross adult blood donors, 2000–2010. *Environmental science & technology* **2012,** *46* (11), 6330-6338.

(51) Schröter-Kermani, C.; Müller, J.; Jürling, H.; Conrad, A.; Schulte, C. Retrospective monitoring of perfluorocarboxylates and perfluorosulfonates in human plasma archived by the German Environmental Specimen Bank. *Int. J. Hyg. Environ. Health* **2013,** *216* (6), 633-640.

(52) Wu, M.; Sun, R.; Wang, M.; Liang, H.; Ma, S.; Han, T.; Xia, X.; Ma, J.; Tang, L.; Sun, Y. Analysis of perfluorinated compounds in human serum from the general population in Shanghai by liquid chromatography-tandem mass spectrometry (LC-MS/MS). *Chemosphere* **2017,** *168*, 100-105.

(53) Toms, L.-M.; Thompson, J.; Rotander, A.; Hobson, P.; Calafat, A. M.; Kato, K.; Ye, X.; Broomhall, S.; Harden, F.; Mueller, J. F. Decline in perfluorooctane sulfonate and perfluorooctanoate serum concentrations in an Australian population from 2002 to 2011. *Environ. Int.* **2014,** *71*, 74-80.

(54) Kemper, R.; Jepson, G. Perfluorooctanoic acid: toxicokinetics in the rat. *Project ID: DuPont* **2003,** *7473*.

(55) Kuslikis, B. I.; Vanden Heuvel, J. P.; Peterson, R. E. Lack of evidence for perfluorodecanoyl ‐ or perfluorooctanoyl ‐ coenzyme a formation in male and female rats. *J. Biochem. Mol. Toxicol.* **1992,** *7* (1), 25-29.

(56) Heuvel, J. P. V.; Kuslikis, B. I.; Van Rafelghem, M. J.; Peterson, R. E. Tissue distribution, metabolism, and elimination of perfluorooctanoic acid in male and female rats. *Journal of biochemical toxicology* **1991,** *6* (2), 83-92.

(57) Vestergren, R.; Cousins, I. T. Tracking the pathways of human exposure to perfluorocarboxylates. *Environmental science & technology* **2009,** *43* (15), 5565-5575.

(58) Kim, S.-J.; Heo, S.-H.; Lee, D.-S.; Hwang, I. G.; Lee, Y.-B.; Cho, H.-Y. Gender differences in pharmacokinetics and tissue distribution of 3 perfluoroalkyl and polyfluoroalkyl substances in rats. *Food and Chemical Toxicology* **2016,** *97*, 243-255.

(59) Kudo, N.; Sakai, A.; Mitsumoto, A.; Hibino, Y.; Tsuda, T.; Kawashima, Y. Tissue distribution and hepatic subcellular distribution of perfluorooctanoic acid at low dose are different from those at high dose in rats. *Biological and Pharmaceutical Bulletin* **2007,** *30* (8), 1535-1540.

(60) Martin, J. W.; Mabury, S. A.; Solomon, K. R.; Muir, D. C. Bioconcentration and tissue distribution of perfluorinated acids in rainbow trout (Oncorhynchus mykiss). *Environ. Toxicol. Chem.* **2003,** *22* (1), 196-204.

(61) Conder, J. M.; Hoke, R. A.; Wolf, W. d.; Russell, M. H.; Buck, R. C. Are PFCAs bioaccumulative? A critical review and comparison with regulatory criteria and persistent lipophilic compounds. *Environmental science & technology* **2008,** *42* (4), 995-1003.

(62) Han, X.; Nabb, D. L.; Russell, M. H.; Kennedy, G. L.; Rickard, R. W. Renal elimination of perfluorocarboxylates (PFCAs). *Chem. Res. Toxicol.* **2011,** *25* (1), 35-46.

(63) Olsen, G. W.; Burris, J. M.; Ehresman, D. J.; Froehlich, J. W.; Seacat, A. M.; Butenhoff, J. L.; Zobel, L. R. Half-life of serum elimination of perfluorooctanesulfonate, perfluorohexanesulfonate, and perfluorooctanoate in retired fluorochemical production workers. *Environ. Health Perspect.* **2007**, 1298-1305.

(64) Kennedy, G. L.; Butenhoff, J. L.; Olsen, G. W.; O'Connor, J. C.; Seacat, A. M.; Perkins, R. G.; Biegel, L. B.; Murphy, S. R.; Farrar, D. G. The toxicology of perfluorooctanoate. *Crit. Rev. Toxicol.* **2004,** *34* (4), 351-384.

(65) Naomi, K.; Kawashima, Y. Toxicity and toxicokinetics of perfluorooctanoic acid in humans and animals. *The Journal of toxicological sciences* **2003,** *28* (2), 49-57.

(66) Hebert, P. C.; MacManus-Spencer, L. A. Development of a fluorescence model for the binding of medium-to long-chain perfluoroalkyl acids to human serum albumin through a mechanistic evaluation of spectroscopic evidence. *Analytical chemistry* **2010,** *82* (15), 6463-6471.

(67) Luebker, D. J.; Hansen, K. J.; Bass, N. M.; Butenhoff, J. L.; Seacat, A. M. Interactions of flurochemicals with rat liver fatty acid-binding protein. *Toxicology* **2002,** *176* (3), 175-185.

(68) Weaver, Y. M.; Ehresman, D. J.; Butenhoff, J. L.; Hagenbuch, B. Roles of rat renal organic anion transporters in transporting perfluorinated carboxylates with different chain lengths. *Toxicol. Sci.* **2009**, kfp275.

(69) Chou, W.-C.; Lin, Z. Bayesian evaluation of a physiologically based pharmacokinetic (PBPK) model for perfluorooctane sulfonate (PFOS) to characterize the interspecies uncertainty between mice, rats, monkeys, and humans: Development and performance verification. *Environment international* **2019,** *129*, 408-422.

(70) Levitt, D. G. The pharmacokinetics of the interstitial space in humans. *BMC Clin. Pharmacol.* **2003,** *3* (1), 3.

(71) Ellmerer, M.; Schaupp, L.; Brunner, G. A.; Sendlhofer, G.; Wutte, A.; Wach, P.; Pieber, T. R. Measurement of interstitial albumin in human skeletal muscle and adipose tissue by open-flow microperfusion. *American Journal of Physiology-Endocrinology and Metabolism* **2000,** *278* (2), E352-E356.

(72) Johnson, J. D.; Gibson, S. J.; Ober, R. E. Cholestyramine-enhanced fecal elimination of carbon-14 in rats after administration of ammonium [14C] perfluorooctanoate or potassium [14C] perfluorooctanesulfonate. *Fundam. Appl. Toxicol.* **1984,** *4* (6), 972-976.

(73) Faber, K. N.; Müller, M.; Jansen, P. L. Drug transport proteins in the liver. *Adv. Drug Del. Rev.* **2003,** *55* (1), 107-124.

(74) Zhao, W.; Zitzow, J. D.; Ehresman, D. J.; Chang, S.-C.; Butenhoff, J. L.; Forster, J.; Hagenbuch, B. Na+/taurocholate cotransporting polypeptide and apical sodium-dependent bile acid transporter are involved in the disposition of perfluoroalkyl sulfonates in humans and rats. *Toxicol. Sci.* **2015**, kfv102.

(75) Han, X.; Hinderliter, P. M.; Snow, T. A.; Jepson, G. W. Binding of Perfluorooctanoic Acid to Rat Liver ‑ form and Kidney ‑ form α 2u ‑ Globulins. *Drug Chem. Toxicol.* **2004,** *27* (4), 341-360.

(76) Maatman, R. G.; van de Westerlo, E. M.; Van Kuppevelt, T.; Veerkamp, J. H. Molecular identification of the liver-and the heart-type fatty acid-binding proteins in human and rat kidney. Use of the reverse transcriptase polymerase chain reaction. *Biochemical Journal* **1992,** *288* (1), 285-290.

(77) Kimura, H.; Odani, S.; Nishi, S.; Sato, H.; Arakawa, M.; Ono, T. Primary structure and cellular distribution of two fatty acid-binding proteins in adult rat kidneys. *Journal of Biological Chemistry* **1991,** *266* (9), 5963-5972.

(78) Crone, C. The permeability of capillaries in various organs as determined by use of the 'indicator diffusion'method. *Acta physiologica scandinavica* **1963,** *58* (4), 292-305.

(79) Gad, S., ADME and Biopharmaceutical Properties. In NJ: 2008.

(80) Weaver, Y. M.; Ehresman, D. J.; Butenhoff, J. L.; Hagenbuch, B. Roles of rat renal organic anion transporters in transporting perfluorinated carboxylates with different chain lengths. *Toxicol. Sci.* **2009,** *113* (2), 305-314.

(81) Cheng, W.; Ng, C. A. A Permeability-Limited Physiologically Based Pharmacokinetic (PBPK) Model for Perfluorooctanoic acid (PFOA) in Male Rats. *Environmental Science & Technology* **2017**.

(82) Ng, C. A.; Hungerbühler, K. Bioconcentration of perfluorinated alkyl acids: how important is specific binding? *Environmental science & technology* **2013,** *47* (13), 7214-7223.

(83) Zhao, W.; Zitzow, J. D.; Ehresman, D. J.; Chang, S.-C.; Butenhoff, J. L.; Forster, J.; Hagenbuch, B. Na+/taurocholate cotransporting polypeptide and apical sodium-dependent bile acid transporter are involved in the disposition of perfluoroalkyl sulfonates in humans and rats. *Toxicol. Sci.* **2015,** *146* (2), 363-373.

(84) Ebert, A.; Allendorf, F.; Berger, U.; Goss, K.-U.; Ulrich, N. Membrane/water partitioning and permeabilities of perfluoroalkyl acids and four of their alternatives and the effects on toxicokinetic behavior. *Environmental science & technology* **2020,** *54* (8), 5051-5061.

(85) Bernillon, P.; Bois, F. Y. Statistical issues in toxicokinetic modeling: a Bayesian perspective. *Environmental Health Perspectives* **2000**, 883-893.

(86) MacLeod, M.; Fraser, A. J.; Mackay, D. Evaluating and expressing the propagation of uncertainty in chemical fate and bioaccumulation models. *Environmental Toxicology and Chemistry: An International Journal* **2002,** *21* (4), 700-709.

(87) Hack, C. E.; Chiu, W. A.; Zhao, Q. J.; Clewell, H. J. Bayesian population analysis of a harmonized physiologically based pharmacokinetic model of trichloroethylene and its metabolites. *Regulatory Toxicology and Pharmacology* **2006,** *46* (1), 63-83.

(88) Haario, H.; Laine, M.; Mira, A.; Saksman, E. DRAM: efficient adaptive MCMC. *StCom* **2006,** *16* (4), 339-354.

(89) Roy, V. Convergence diagnostics for Markov chain Monte Carlo. *Annual Review of Statistics and Its Application* **2020,** *7*, 387-412.

(90) Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B., *Bayesian data analysis*. CRC press: 2013.

(91) Trudel, D.; Horowitz, L.; Wormuth, M.; Scheringer, M.; Cousins, I. T.; Hungerbühler, K. Estimating consumer exposure to PFOS and PFOA. *Risk Anal.* **2008,** *28* (2), 251-269.

(92) Soetaert, K.; Petzoldt, T. Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME. *Journal of statistical software* **2010,** *33* (3), 1-28.

(93) Yang, Y.; Xu, X.; Georgopoulos, P. G. A Bayesian population PBPK model for multiroute chloroform exposure. *J. Expo. Sci. Environ. Epidemiol.* **2010,** *20* (4), 326-341.

(94) Ohmori, K.; Kudo, N.; Katayama, K.; Kawashima, Y. Comparison of the toxicokinetics between perfluorocarboxylic acids with different carbon chain length. *Toxicology* **2003,** *184* (2-3), 135-140.

(95) Harada, K.; Inoue, K.; Morikawa, A.; Yoshinaga, T.; Saito, N.; Koizumi, A. Renal clearance of perfluorooctane sulfonate and perfluorooctanoate in humans and their species-specific excretion. *Environmental research* **2005,** *99* (2), 253-261.

(96) Benskin, J. P.; De Silva, A. O.; Martin, L. J.; Arsenault, G.; McCrindle, R.; Riddell, N.; Mabury, S. A.; Martin, J. W. Disposition of perfluorinated acid isomers in sprague ‐ dawley rats; Part 1: Single dose. *Environmental Toxicology and Chemistry: An International Journal* **2009,** *28* (3), 542-554.

(97) Wambaugh, J. F.; Barton, H. A.; Setzer, R. W. Comparing models for perfluorooctanoic acid pharmacokinetics using Bayesian analysis. *J. Pharmacokinet. Pharmacodyn.* **2008,** *35* (6), 683-712.

(98) Wu, L.-L.; Gao, H.-W.; Gao, N.-Y.; Chen, F.-F.; Chen, L. Interaction of perfluorooctanoic acid with human serum albumin. *BMC structural biology* **2009,** *9* (1), 1-7.

(99) Qin, P.; Liu, R.; Pan, X.; Fang, X.; Mou, Y. Impact of carbon chain length on binding of perfluoroalkyl acids to bovine serum albumin determined by spectroscopic methods. *Journal of agricultural and food chemistry* **2010,** *58* (9), 5561-5567.

(100) Armitage, J. M.; Arnot, J. A.; Wania, F.; Mackay, D. Development and evaluation of a mechanistic bioconcentration model for ionogenic organic chemicals in fish. *Environmental toxicology and chemistry* **2013,** *32* (1), 115-128.

(101) D'Agostino, L. A.; Mabury, S. A. Identification of novel fluorinated surfactants in aqueous film forming foams and commercial surfactant concentrates. *Environmental science & technology* **2013,** *48* (1), 121-129.

(102) Wang, S.; Huang, J.; Yang, Y.; Hui, Y.; Ge, Y.; Larssen, T.; Yu, G.; Deng, S.; Wang, B.; Harman, C. First report of a Chinese PFOS alternative overlooked for 30 years: its toxicity, persistence, and presence in the environment. *Environmental science & technology* **2013,** *47* (18), 10163-10170.

(103) Zhao, P.; Xia, X.; Dong, J.; Xia, N.; Jiang, X.; Li, Y.; Zhu, Y. Short-and long-chain perfluoroalkyl substances in the water, suspended particulate matter, and surface sediment of a turbid river. *Science of the Total Environment* **2016,** *568*, 57-65.

(104) Wen, W.; Xia, X.; Hu, D.; Zhou, D.; Wang, H.; Zhai, Y.; Lin, H. Long-Chain Perfluoroalkyl acids (PFAAs) Affect the Bioconcentration and Tissue Distribution of Short-Chain PFAAs in Zebrafish (Danio rerio). *Environmental Science & Technology* **2017,** *51* (21), 12358-12368.

(105) Lam, J. C.; Lyu, J.; Kwok, K. Y.; Lam, P. K. Perfluoroalkyl substances (PFASs) in marine mammals from the South China Sea and their temporal changes 2002–2014: Concern for alternatives of PFOS? *Environmental science & technology* **2016,** *50* (13), 6728-6736.

(106) Shi, Y.; Vestergren, R.; Zhou, Z.; Song, X.; Xu, L.; Liang, Y.; Cai, Y. Tissue distribution and whole body burden of the chlorinated polyfluoroalkyl ether sulfonic acid F-53B in crucian carp (Carassius carassius): Evidence for a highly bioaccumulative contaminant of emerging concern. *Environmental science & technology* **2015,** *49* (24), 14156-14165.

(107) Yao, Y.; Zhao, Y.; Sun, H.; Chang, S.; Zhu, L.; Alder, A. C.; Kannan, K. Per-and Polyfluoroalkyl Substances (PFASs) in Indoor Air and Dust from Homes and Various Microenvironments in China: Implications for Human Exposure. **2018**.

(108) Yeung, L. W.; Stadey, C.; Mabury, S. A. Simultaneous analysis of perfluoroalkyl and polyfluoroalkyl substances including ultrashort-chain C2 and C3 compounds in rain and river

water samples by ultra performance convergence chromatography. *Journal of Chromatography A* **2017,** *1522*, 78-85.

(109) Sun, M.; Arevalo, E.; Strynar, M.; Lindstrom, A.; Richardson, M.; Kearns, B.; Pickett, A.; Smith, C.; Knappe, D. R. Legacy and emerging perfluoroalkyl substances are important drinking water contaminants in the Cape Fear River Watershed of North Carolina. *Environmental science & technology letters* **2016,** *3* (12), 415-419.

(110) Chen, F.; Liu, H.; Sun, H.; Pan, P.; Li, Y.; Li, D.; Hou, T. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein–protein binding free energies and re-rank binding poses generated by protein–protein docking. *Physical Chemistry Chemical Physics* **2016,** *18* (32), 22129-22139.

(111) Hou, T.; Wang, J.; Li, Y.; Wang, W. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model.* **2010,** *51* (1), 69-82.

(112) Rastelli, G.; Rio, A. D.; Degliesposti, G.; Sgobba, M. Fast and accurate predictions of binding free energies using MM‐PBSA and MM‐GBSA. *Journal of computational chemistry* **2010,** *31* (4), 797-810.

(113) Sun, H.; Li, Y.; Tian, S.; Xu, L.; Hou, T. Assessing the performance of MM/PBSA and MM/GBSA methods. 4. Accuracies of MM/PBSA and MM/GBSA methodologies evaluated by various simulation protocols using PDBbind data set. *Physical Chemistry Chemical Physics* **2014,** *16* (31), 16719-16729.

(114) Morris, G. M.; Lim-Wilby, M., Molecular docking. In *Molecular modeling of proteins*, Springer: 2008; pp 365-382.

(115) Zhang, L.; Ren, X.-M.; Guo, L.-H. Structure-based investigation on the interaction of perfluorinated compounds with human liver fatty acid binding protein. *Environmental science & technology* **2013,** *47* (19), 11293-11301.

(116) Sheng, N.; Cui, R.; Wang, J.; Guo, Y.; Wang, J.; Dai, J. Cytotoxicity of novel fluorinated alternatives to long-chain perfluoroalkyl substances to human liver cell line and their binding capacity to human liver fatty acid binding protein. *Archives of toxicology* **2018,** *92* (1), 359-369.

(117) Sharma, A.; Sharma, A. Fatty acid induced remodeling within the human liver fatty acid-binding protein. *Journal of Biological Chemistry* **2011,** *286* (36), 31924-31928.

(118) Thompson, J.; Winter, N.; Terwey, D.; Bratt, J.; Banaszak, L. The Crystal Structure of the Liver Fatty Acid-binding Protein A COMPLEX WITH TWO BOUND OLEATES. *Journal of Biological Chemistry* **1997,** *272* (11), 7140-7150.

(119) Ng, C. A.; Hungerbuehler, K. Exploring the use of molecular docking to identify bioaccumulative perfluorinated alkyl acids (PFAAs). *Environmental science & technology* **2015,** *49* (20), 12306-12314.

(120) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **2015,** *10* (6), 845.

(121) Zhang, J.; Begum, A.; Brännström, K.; Grundström, C.; Iakovleva, I.; Olofsson, A.; Sauer-Eriksson, A. E.; Andersson, P. L. Structure-based virtual screening protocol for in silico identification of potential thyroid disrupting chemicals targeting transthyretin. *Environmental science & technology* **2016,** *50* (21), 11984-11993.

(122) Luo, Z.; Shi, X.; Hu, Q.; Zhao, B.; Huang, M. Structural evidence of perfluorooctane sulfonate transport by human serum albumin. *Chem. Res. Toxicol.* **2012,** *25* (5), 990-992.

(123) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **2012,** *4* (1), 17.

(124) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* **2010,** *31* (2), 455-461.

(125) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **2009,** *30* (16), 2785-2791.

(126) Prevedouros, K.; Cousins, I. T.; Buck, R. C.; Korzeniowski, S. H. Sources, fate and transport of perfluorocarboxylates. *Environmental science & technology* **2006,** *40* (1), 32-44.

(127) Gomis, M. I.; Wang, Z.; Scheringer, M.; Cousins, I. T. A modeling assessment of the physicochemical properties and environmental fate of emerging and novel per-and polyfluoroalkyl substances. *Science of the Total Environment* **2015,** *505*, 981-991.

(128) Schrodinger, LLC, The PyMOL Molecular Graphics System, Version 1.8. In 2015.

(129) Case, D. A.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham Iii, T.; Darden, T.; Duke, R.; Gohlke, H. Amber 14. **2014**.

(130) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **2006,** *65* (3), 712-725.

(131) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **2004,** *25* (9), 1157-1174.

(132) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high - quality atomic charges. AM1 - BCC model: II. Parameterization and validation. *Journal of computational chemistry* **2002,** *23* (16), 1623-1641.

(133) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics* **1993,** *98* (12), 10089-10092.

(134) Le Grand, S.; Götz, A. W.; Walker, R. C. SPFP: Speed without compromise—A mixed precision model for GPU accelerated molecular dynamics simulations. *Computer Physics Communications* **2013,** *184* (2), 374-380.

(135) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *Journal of computational chemistry* **2005,** *26* (16), 1668-1688.

(136) Genheden, S.; Ryde, U. How to obtain statistically converged MM/GBSA results. *Journal of computational chemistry* **2010,** *31* (4), 837-846.

(137) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000,** *33* (12), 889-897.

(138) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate− DNA helices. *Journal of the American Chemical Society* **1998,** *120* (37), 9401-9409.

(139) Homeyer, N.; Gohlke, H. Free energy calculations by the molecular mechanics Poisson− Boltzmann surface area method. *Mol. Inform.* **2012,** *31* (2), 114-122.

(140) Hadden, J. A.; Tessier, M. B.; Fadda, E.; Woods, R. J. Calculating Binding Free Energies for Protein–Carbohydrate Complexes. *Glycoinformatics* **2015**, 431-465.

(141) Poland, D. Statistical Mechanics (McQuarrie, Donald A.). *Journal of Chemical Education* **1977,** *54* (10), A428.

(142) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein–protein binding by binding free energy calculation and free energy decomposition for the Ras–Raf and Ras–RalGDS complexes. *Journal of molecular biology* **2003,** *330* (4), 891-913.

(143) Metz, A.; Pfleger, C.; Kopitz, H.; Pfeiffer-Marek, S.; Baringhaus, K.-H.; Gohlke, H. Hot spots and transient pockets: predicting the determinants of small-molecule binding to a protein–protein interface. *J. Chem. Inf. Model.* **2011,** *52* (1), 120-133.

(144) Caldwell, G. W.; Yan, Z., Isothermal titration calorimetry characterization of drug-binding energetics to blood proteins. In *Optimization in Drug Discovery*, Springer: 2004; pp 123-149.

(145) Kastritis, P. L.; Bonvin, A. M. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *J R Soc Interface* **2013,** *10* (79), 20120835.

(146) Cheng, W.; Ng, C. A. Predicting Relative Protein Affinity of Novel Per-and Polyfluoroalkyl Substances (PFASs) by An Efficient Molecular Dynamics Approach. *Environmental science & technology* **2018,** *52* (14), 7972-7980.

(147) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015,** *521* (7553), 436.

(148) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Frontiers in Environmental Science* **2016,** *3,* 80.

(149) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions. *arXiv preprint arXiv:1406.1231* **2014**.

(150) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **2010,** *50* (5), 742-754.

(151) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018,** *9* (2), 513-530.

(152) Brusseau, M. L. The influence of molecular structure on the adsorption of PFAS to fluid-fluid interfaces: Using QSPR to predict interfacial adsorption coefficients. *Water research* **2019,** *152,* 148-158.

(153) Kar, S.; Sepúlveda, M. S.; Roy, K.; Leszczynski, J. Endocrine-disrupting activity of per-and polyfluoroalkyl substances: Exploring combined approaches of ligand and structure based modeling. *Chemosphere* **2017,** *184,* 514-523.

(154) Kovarich, S.; Papa, E.; Li, J.; Gramatica, P. QSAR classification models for the screening of the endocrine-disrupting activity of perfluorinated compounds. *SAR and QSAR in Environmental Research* **2012,** *23* (3-4), 207-220.

(155) Bhhatarai, B.; Gramatica, P. Oral LD 50 toxicity modeling and prediction of per-and polyfluorinated chemicals on rat and mouse. *Mol. Divers.* **2011,** *15* (2), 467-476.

(156) Bhhatarai, B.; Gramatica, P. Per-and polyfluoro toxicity (LC50 inhalation) study in rat and mouse using QSAR modeling. *Chem. Res. Toxicol.* **2010,** *23* (3), 528-539.

(157) Hoover, G.; Kar, S.; Guffey, S.; Leszczynski, J.; Sepúlveda, M. S. In vitro and in silico modeling of perfluoroalkyl substances mixture toxicity in an amphibian fibroblast cell line. *Chemosphere* **2019,** *233,* 25-33.

(158) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072* **2015**.

(159) Sperandei, S. Understanding logistic regression analysis. *Biochemia medica: Biochemia medica* **2014,** *24* (1), 12-18.

(160) Breiman, L. Random forests. *MLear* **2001,** *45* (1), 5-32.

(161) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low data drug discovery with one-shot learning. *ACS central science* **2017,** *3* (4), 283-293.

(162) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design* **2016,** *30* (8), 595-608.

(163) Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem bioassay: 2014 update. *Nucleic acids research* **2013,** *42* (D1), D1075-D1082.

(164) Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009,** *49* (2), 169-184.

(165) Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational modeling of β-secretase 1 (BACE-1) inhibitors using ligand based approaches. *J. Chem. Inf. Model.* **2016,** *56* (10), 1936-1949.

(166) Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian approach to in silico blood-brain barrier penetration modeling. *J. Chem. Inf. Model.* **2012,** *52* (6), 1686-1697.

(167) Challenge, T. https://tripod.nih.gov/tox21/challenge/. **2019**.

(168) Cummings, J. L.; Morstorf, T.; Zhong, K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res. Ther.* **2014,** *6* (4), 37.

(169) Wong, S. C.; Gatt, A.; Stamatescu, V.; McDonnell, M. D. In *Understanding data augmentation for classification: when to warp?*, 2016 international conference on digital image computing: techniques and applications (DICTA), 2016; IEEE: 2016; pp 1-6.

(170) Cortes, C.; Vapnik, V. Support-vector networks. *MLear* **1995,** *20* (3), 273-297.

(171) Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics* **2000,** *28* (2), 337-407.

(172) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *Journal of computer-aided molecular design* **2008,** *22* (3-4), 133-139.

(173) Snoek, J.; Larochelle, H.; Adams, R. P. In *Practical bayesian optimization of machine learning algorithms*, Advances in neural information processing systems, 2012; 2012; pp 2951-2959.

(174) Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **2012,** *17* (5), 4791-4810.

(175) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008,** *48* (9), 1733-1746.

(176) Jiménez, J.; Ginebra, J. pyGPGO: Bayesian Optimization for Python. *The Journal of Open Source Software* **2017,** *2*, 431.

(177) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. In *Neural message passing for quantum chemistry*, Proceedings of the 34th International Conference on Machine Learning-Volume 70, 2017; JMLR. org: 2017; pp 1263-1272.

(178) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. In *Convolutional networks on graphs for learning molecular fingerprints*, Adv. Neural Inf. Process. Syst., 2015; 2015; pp 2224-2232.

(179) Serpionov, G. V.; Alexandrov, A. I.; Ter-Avanesyan, M. D. Distinct mechanisms of mutant huntingtin toxicity in different yeast strains. *Yeast Research* **2017,** *17* (1), fow102.

(180) Huang, Z.; Xie, H.; Wang, R.; Sun, Z. Retinoid-related orphan receptor γt is a potential therapeutic target for controlling inflammatory autoimmunity. *Expert Opin. Ther. Targets* **2007,** *11* (6), 737-743.

(181) Casciello, F.; Windloch, K.; Gannon, F.; Lee, J. S. Functional role of G9a histone methyltransferase in cancer. *Front. Immunol.* **2015,** *6*, 487.

(182) Donnelly, D. The structure and function of the glucagon ‐ like peptide ‐ 1 receptor and its ligands. *British journal of pharmacology* **2012,** *166* (1), 27-41.

(183) Shabalina, I. G.; Kalinovich, A. V.; Cannon, B.; Nedergaard, J. Metabolically inert perfluorinated fatty acids directly activate uncoupling protein 1 in brown-fat mitochondria. *Archives of toxicology* **2016,** *90* (5), 1117-1128.

(184) Su, P. C.; Johnson, M. E. Evaluating thermodynamic integration performance of the new amber molecular dynamics package and assess potential halogen bonds of enoyl ‐ ACP reductase (FabI) benzimidazole inhibitors. *Journal of computational chemistry* **2016,** *37* (9), 836-847.

(185) Ridgeway, G.; Madigan, D. A sequential Monte Carlo method for Bayesian analysis of massive datasets. *Data Mining and Knowledge Discovery* **2003,** *7* (3), 301-319.

(186) Kantas, N.; Doucet, A.; Singh, S. S.; Maciejowski, J. M. An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. *IFAC Proceedings Volumes* **2009,** *42* (10), 774-785.

(187) Russell, S.; Norvig, P. Artificial intelligence: a modern approach. **2002**.

(188) Richard, A. M.; Huang, R.; Waidyanatha, S.; Shinn, P.; Collins, B. J.; Thillainadarajah, I.; Grulke, C. M.; Williams, A. J.; Lougee, R. R.; Judson, R. S. The Tox21 10K Compound Library: Collaborative Chemistry Advancing Toxicology. *Chem. Res. Toxicol.* **2020**.

(189) Breen, M.; Ring, C. L.; Kreutz, A.; Goldsmith, M.-R.; Wambaugh, J. F. High-throughput PBTK models for in vitro to in vivo extrapolation. *Expert Opin. Drug Metab. Toxicol.* **2021,** (just-accepted).

(190) Luechtefeld, T.; Rowlands, C.; Hartung, T. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicology research* **2018,** *7* (5), 732-744.

(191) Brown, R. P.; Delp, M. D.; Lindstedt, S. L.; Rhomberg, L. R.; Beliles, R. P. Physiological parameter values for physiologically based pharmacokinetic models. *Toxicology and industrial health* **1997,** *13* (4), 407-484.

(192) Davies, B.; Morris, T. Physiological parameters in laboratory animals and humans. *Pharm. Res.* **1993,** *10* (7), 1093-1095.

(193) Blouin, A.; Bolender, R. P.; Weibel, E. R. Distribution of organelles and membranes between hepatocytes and nonhepatocytes in the rat liver parenchyma. A stereological study. *The Journal of cell biology* **1977,** *72* (2), 441-455.

(194) Larson, M.; Sjöquist, M.; Wolgast, M. Renal interstitial volume of the rat kidney. *Acta physiologica scandinavica* **1984,** *120* (2), 297-304.

(195) Barratt, T. M.; Walser, M. Extracellular fluid in individual tissues and in whole animals: the distribution of radiosulfate and radiobromide. *The Journal of clinical investigation* **1969,** *48* (1), 56-66.

(196) EVERETT, N. B.; Simmons, B.; Lasher, E. P. Distribution of blood (Fe59) and plasma (I131) volumes of rats determined by liquid nitrogen freezing. *Circulation research* **1956,** *4* (4), 419-424.

(197) McConnell, E. L.; Basit, A. W.; Murdan, S. Measurements of rat and mouse gastrointestinal pH, fluid and lymphoid tissue, and implications for in ‐ vivo experiments. *Journal of Pharmacy and Pharmacology* **2008,** *60* (1), 63-70.

(198) Munger, K.; Baylis, C. Sex differences in renal hemodynamics in rats. *American Journal of Physiology-Renal Physiology* **1988,** *254* (2), F223-F231.

(199) Biewald, N.; Billmeier, J. Blood volume and extracellular space (ECS) of the whole body and some organs of the rat. *Experientia* **1978,** *34* (3), 412-413.

(200) Kirkman, H.; Stowell, R. Renal filtration surface in the albino rat. *The Anatomical Record* **1942,** *82* (3), 373-391.

(201) DeSesso, J.; Jacobson, C. Anatomical and physiological parameters affecting gastrointestinal absorption in humans and rats. *Food and chemical toxicology* **2001,** *39* (3), 209-228.

(202) Arthur, S.; Green, R. Fluid reabsorption by the proximal convoluted tubule of the kidney in lactating rats. *The Journal of physiology* **1986,** *371* (1), 267-275.

(203) Bonvalet, J.; de Rouffignac, C. Distribution of ferrocyanide along the proximal tubular lumen of the rat kidney: its implications upon hydrodynamics. *The Journal of physiology* **1981,** *318* (1), 85-98.

(204) Frazier, K. S.; Seely, J. C.; Hard, G. C.; Betton, G.; Burnett, R.; Nakatsuji, S.; Nishikawa, A.; Durchfeld-Meyer, B.; Bube, A. Proliferative and nonproliferative lesions of the rat and mouse urinary system. *Toxicologic Pathology* **2012,** *40* (4_suppl), 14S-86S.

(205) Kanauchi, O.; Agata, K.; Fushiki, T. Mechanism for the increased defecation and jejunum mucosal protein content in rats by feeding germinated barley foodstuff. *Bioscience, biotechnology, and biochemistry* **1997,** *61* (3), 443-448.

(206) Cavigelli, S.; Monfort, S.; Whitney, T.; Mechref, Y.; Novotny, M.; McClintock, M. Frequent serial fecal corticoid measures from rats reflect circadian and ovarian corticosterone rhythms. *Journal of Endocrinology* **2005,** *184* (1), 153-163.

(207) Crisman, T. S.; Claffey, K. P.; Saouaf, R.; Hanspal, J.; Brecher, P. Measurement of rat heart fatty acid binding protein by ELISA. Tissue distribution, developmental changes and subcellular distribution. *J. Mol. Cell. Cardiol.* **1987,** *19* (5), 423-431.

(208) Levitt, D. G. The pharmacokinetics of the interstitial space in humans. *BMC Clin. Pharmacol.* **2003,** *3* (1), 1-29.

(209) Ockner, R.; Manning, J.; Kane, J. Fatty acid binding protein. Isolation from rat liver, characterization, and immunochemical quantification. *Journal of Biological Chemistry* **1982,** *257* (13), 7872-7878.

(210) Peters Jr, T. The Biosynthesis of Rat Serum Albumin: II. INTRACELLULAR PHENOMENA IN THE SECRETION OF NEWLY FORMED ALBUMIN. *Journal of Biological Chemistry* **1962,** *237* (4), 1186-1189.