toxics

MDPI

Article

A Machine Learning Model to Estimate Toxicokinetic Half-Lives of Per- and Polyfluoro-Alkyl Substances (PFAS) in Multiple Species

Daniel E. Dawson [1], Christopher Lau [2], Prachi Pradeep [1,3], Risa R. Sayre [1], Richard S. Judson [1], Rogelio Tornero-Velez [1] and John F. Wambaugh [1,*]

[1] U.S. Environmental Protection Agency, Office of Research and Development, Center for Computational Toxicology and Exposure, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711, USA
[2] U.S. Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment, 109 T.W. Alexander Drive, Research Triangle Park, NC 277011, USA
[3] Oak Ridge Institutes for Science and Education, Oak Ridge, TN 37830, USA
* Correspondence: wambaugh.john@epa.gov; Tel.: +1-919-541-7641

Abstract: Per- and polyfluoroalkyl substances (PFAS) are a diverse group of man-made chemicals that are commonly found in body tissues. The toxicokinetics of most PFAS are currently uncharacterized, but long half-lives ($t_{\frac{1}{2}}$) have been observed in some cases. Knowledge of chemical-specific $t_{\frac{1}{2}}$ is necessary for exposure reconstruction and extrapolation from toxicological studies. We used an ensemble machine learning method, random forest, to model the existing in vivo measured $t_{\frac{1}{2}}$ across four species (human, monkey, rat, mouse) and eleven PFAS. Mechanistically motivated descriptors were examined, including two types of surrogates for renal transporters: (1) physiological descriptors, including kidney geometry, for renal transporter expression and (2) structural similarity of defluorinated PFAS to endogenous chemicals for transporter affinity. We developed a classification model for $t_{\frac{1}{2}}$ (Bin 1: <12 h; Bin 2: <1 week; Bin 3: <2 months; Bin 4: >2 months). The model had an accuracy of 86.1% in contrast to 32.2% for a y-randomized null model. A total of 3890 compounds were within domain of the model, and $t_{\frac{1}{2}}$ was predicted using the bin medians: 4.9 h, 2.2 days, 33 days, and 3.3 years. For human $t_{\frac{1}{2}}$, 56% of PFAS were classified in Bin 4, 7% were classified in Bin 3, and 37% were classified in Bin 2. This model synthesizes the limited available data to allow tentative extrapolation and prioritization.

Keywords: perfluoro-alkyl substances; PFAS; half-life; machine learning model; toxicokinetics

Citation: Dawson, D.E.; Lau, C.; Pradeep, P.; Sayre, R.R.; Judson, R.S.; Tornero-Velez, R.; Wambaugh, J.F. A Machine Learning Model to Estimate Toxicokinetic Half-Lives of Per- and Polyfluoro-Alkyl Substances (PFAS) in Multiple Species. *Toxics* **2023**, *11*, 98. https://doi.org/10.3390/toxics11020098

Academic Editor: Maria João Rocha

Received: 12 December 2022
Revised: 9 January 2023
Accepted: 18 January 2023
Published: 20 January 2023

1. Introduction

Per- and polyfluoro-alkyl substances (PFAS) are a large and diverse class of organic chemicals in which all (per-) or some (poly-) carbon–hydrogen bonds have been replaced with carbon–fluorine bonds [1]. Since carbon–fluorine bonds are stronger, they help make PFAS resistant to metabolism and degradation [2]. PFAS have both hydrophobic and lipophobic properties, from which they derive both water- and stain-repellant properties, thereby providing some of their utility to industry and consumers [3]. The majority of PFAS have either a straight- or branched-chain alkane backbone, with one or more functional groups bonded to the terminal ends of the backbone [2,4]. Examples of commonly studied straight-chained PFAS include carboxylic acids (such as perfluorooctanoic acid/PFOA) and sulfonic acids (such as perfluorooctane sulfonic acid/PFOS). A branched PFAS of note is GenX (perfluoro-2-methyl-3-oxahexanoic acid) [5,6]. Even the relatively well-studied PFOA and PFOS have lesser studied branched isomers [7,8].

PFAS are commonly found in human tissues [1]. Chemical properties of PFAS, such as the propensity to bind to protein, contribute to significant partitioning in the liver, the kidney, and the blood [9,10]. PFAS are of significant public health concern, as exposure

has been associated with a growing list of pathologies in humans. Pathologies include endocrine system disorders, immunological disorders, fatty liver disease, cancers of the kidneys and testicles, and lower birth weight [11].

Due to the ubiquity of PFAS in body tissues, there is growing interest in characterizing the disposition of these chemicals within the body (that is, their toxicokinetics/TK) [12,13]. TK half-life ($t_{\frac{1}{2}}$) is the amount of time needed for 50% of the chemical to be eliminated from the body. $t_{\frac{1}{2}}$ is used to extrapolate from toxicological effects observed in animal species [14] and to understand human exposure [15–17]. Some PFAS (for example, PFOS) have been noted as having long half-lives (several years in humans). Widespread PFAS exposure from the environment and long half-lives result in the potential for bioaccumulation, as rates of uptake may exceed rates of excretion [18].

For typical organic chemicals, mathematical models exist for predicting properties related to human $t_{\frac{1}{2}}$ from chemical structure [19–22]. However, these approaches are expected to fail for some PFAS [23] due to the peculiarities of fluorous chemistry [24] and potential biological interactions [25–27]. The estimation of PFAS $t_{\frac{1}{2}}$ thus relies on either observational studies or extrapolation from animal species [11,28–30]. Typical extrapolation methods for TK parameters of PFAS are unreliable between species [14] and chemicals [27]. Efforts at extrapolating the measured PFAS $t_{\frac{1}{2}}$ across species are complicated by unusual and unpredictable variability [26]. The $t_{\frac{1}{2}}$ of perfluorohexanoic acid (PFHxA), for example, appears to scale allometrically (proportional to species weight) across mice, rats, monkeys, and humans [31]. In contrast, the $t_{\frac{1}{2}}$ of the PFOA ranges from a few hours in female rats, days in male rats, 30–130 days in mice and monkeys, respectively [32–34], to 2–4 years in humans [35–39]. This large variation for PFOA occurs despite its structural similarity to PFHxA.

Under current chemical risk assessment paradigms, animals such as rats, mice, and monkeys serve as models to obtain toxicological information for other species where experiments may not be conducted; that is, humans and endangered wildlife. As toxicity testing evolves to include new approach methodologies [40], this may be less true. However, it is well known from physiologically based toxicokinetic modeling that understanding what phenomena can and cannot be extrapolated between species will inform human chemical risk assessment [41–44]. Thus, a key goal for PFAS is understanding differences in elimination kinetics between species [27].

Lau et al. [11,28–30] have reviewed the literature on in vivo measured interspecies PFAS $t_{\frac{1}{2}}$ in 2007, 2012, 2015 and, most recently, in 2021. They have curated PFAS $t_{\frac{1}{2}}$ data for multiple species across eleven PFAS. Most of the measured data are for rodents. While some PFAS rapidly transform to one of these eleven PFAS in vivo, [45] there are many thousands more for which there are no data available [12]. This is, in part, because in vivo experiments are resource intensive [46,47]. Additionally, higher throughput toxicokinetic methods perform poorly for some PFAS due to a lack of data characterizing transporters [23]. For linear PFAS only, $t_{\frac{1}{2}}$ is observed to roughly increase with carbon chain length [36]. However, no systematic rules have been discerned for inter-species or inter-chemical extrapolation of PFAS $t_{\frac{1}{2}}$ in general. Instead, each chemical and species require new in vivo studies [14,26,48]. Interaction with transporters and protein binding have both been suggested as relevant mechanisms that might be accessible in vitro [25–27,32], but these again require species- and chemical-specific measurements that are generally unavailable. Additionally, $t_{\frac{1}{2}}$ varies with sex for some PFAS and species, with males typically having longer $t_{\frac{1}{2}}$ than females [1].

Given the failure of typical approaches for the inter-species or inter-chemical extrapolation of PFAS $t_{\frac{1}{2}}$, and the importance of this parameter for understanding the impact of these chemicals in the environment, a new approach is needed. Machine learning (ML) is an opportunity to use the available data to develop predictions for new chemical–species combinations. ML-based models of TK parameters can integrate multiple descriptors

into predictive models for chemical properties [20,21,49]. Ensemble ML-based methods, such as random forest, combine predictions from an assembly of models (for example, regression/classification trees) to improve the robustness of the predictions. Each model contributing to the ensemble is built from a subset of predictors and/or training data records. Such ensemble models have been shown to provide reasonably accurate predictions over a range of chemical properties when empirical data are unavailable [20,50]. ML has previously been applied to PFAS, including to identify efficient treatment and removal from water [51] and to prioritize groundwater testing [52]. These prior works also used a variety of different machine learning approaches, including neural networks, the method of random forests, and other classification algorithms [51,52]. ML-based models might organize existing PFAS $t_{\frac{1}{2}}$ data, categorize unmeasured PFAS, and identify the most impactful data needs for additional measurement. Since machine learning draws inferences from a data "training set", one key metric for evaluating performance is a comparison of the difference between an ML model built with the actual training set and a model built using a "y-randomized" training set [53]. In y-randomization, the outcome to be predicted (in this case, $t_{\frac{1}{2}}$) has been randomly swapped among the data. Y-randomization provides a baseline of how well a model might perform by chance.

In this study, we use the random forest method to develop a ML classification model for PFAS $t_{\frac{1}{2}}$. We first use Monte Carlo methods to supplement the Lau et al. [11,28–30] $t_{\frac{1}{2}}$ data set using TK studies not previously included. Given a small training set of eleven PFAS across four species, we aimed only to broadly classify PFAS chemical/species $t_{\frac{1}{2}}$ into four categories: less than 12 h, 12 h to 1 week, 1 week to 2 months, or greater than 2 months. A diverse array of 119 descriptors was considered by the ML as potential predictors. These descriptors were mechanistically motivated, including both chemical and physiological properties. In particular, the descriptor set included several potential surrogates for transporters. Feature elimination was used to ensure a parsimonious model. To assess coincidental associations between descriptors and predictions, the actual model was contrasted against models built using multiple training data randomization approaches. We applied the model to a large set (~6600) of PFAS, for which $t_{\frac{1}{2}}$ data are unavailable. Given the broad ranges of half-lives predicted by the model, for humans the model effectively predicts whether a given PFAS is more likely to be persistent. Those chemicals identified to likely be biologically persistent may pose an elevated risk. Finally, we use the predicted $t_{\frac{1}{2}}$ values and a simple TK model to predict whole body clearance and steady-state plasma concentrations in multiple species.

2. Materials and Methods

The major steps of the workflow for this study included training dataset assembly, predictor set assembly, model construction, and model application (Figure 1). Dataset assembly is described in brief below, and in detail in the Supplemental Information (S1.1, see S1_Dawson et al._ML PFAS_HL_101322.pdf). All analyses were performed using the freely available R statistical software platform v4.1.3 [54]. We used the following open-source tools ("packages") from the Comprehensive R Archive Network (https://cran.r-project.org/, accessed 20 September 2022): caret [55], classyfireR [56], corrplot [57], data.table [58], gdata [59], ggplot2 [60], httk [61], MLmetrics [62], OneR [63], openxlsx [64], purr [65], randomForest [66], readxl [67], scales [68], showtext [69], stringr [70], and tidyr [71]. All scripts and data are available at: https://github.com/USEPA/CompTox-PFASHalfLife (accessed 17 January 2023).

2.1. Dataset Assembly

2.1.1. PFAS Half-Life Data (Dependent Variable)

We modeled in vivo serum $t_{\frac{1}{2}}$ data for 11 PFAS using published data experimentally collected from 4 species. The literature base was assembled from the most recent curation of Lau et al. (2021) [11,28–30] and supplemented with studies not previously reviewed.

We intend models developed with these data to be preliminary attempts to classify the range of $t_{\frac{1}{2}}$ of PFAS. Of the 11 chemicals, 6 are straight-chain perfluoroalkyl carboxylic acids: perfluorobutanoic acid (PFBA, DTXSID4059916), perfluorohexanoic acid (PFHxA, DTXSID30318623031862), perfluoroheptanoic acid (PFHpA, DTXSID1037303), perfluorooctanoic acid (PFOA, DTXSID8031865), perfluorononanoic acid (PFNA, DTXSID8031863), and perfluorodecanoic acid (PFDA, DTXSID3031860); 3 chemicals are straight-chain perfluoroalkyl sulfonic acids: perfluorobutanesulfonic acid (PFBS, DTXSID5030030), perfluorohexanesulfonic acid (PFHxS, DTXSID7040150), perfluorooctanesulfonic acid (PFOS, DTXSID3031864). The 2 remaining chemicals, perfluoro-2-methyl-3-oxahexanoic acid (GenX, DTXSID70880215) and perfluoro (2-((6-chlorohexyl)oxy)ethanesulfonic acid (F-53B, DTXSID80892506), are branched perfluoroalkyl carboxylic acids and perfluoroalkyl sulfonic acids, respectively. See the Supplemental Information (S2.3, see S2_Dawson et al._ML PFAS_HL_101322.xlsx) for structural representations of each of the compounds. Chemicals and species were selected to have a range of data to inform extrapolation: species included humans, cynomolgus monkey (*Macaca fascicularis*), mouse (*Mus musculus*), and rat (*Rattus rattus*). Data from both sexes of each species were also included, as available.

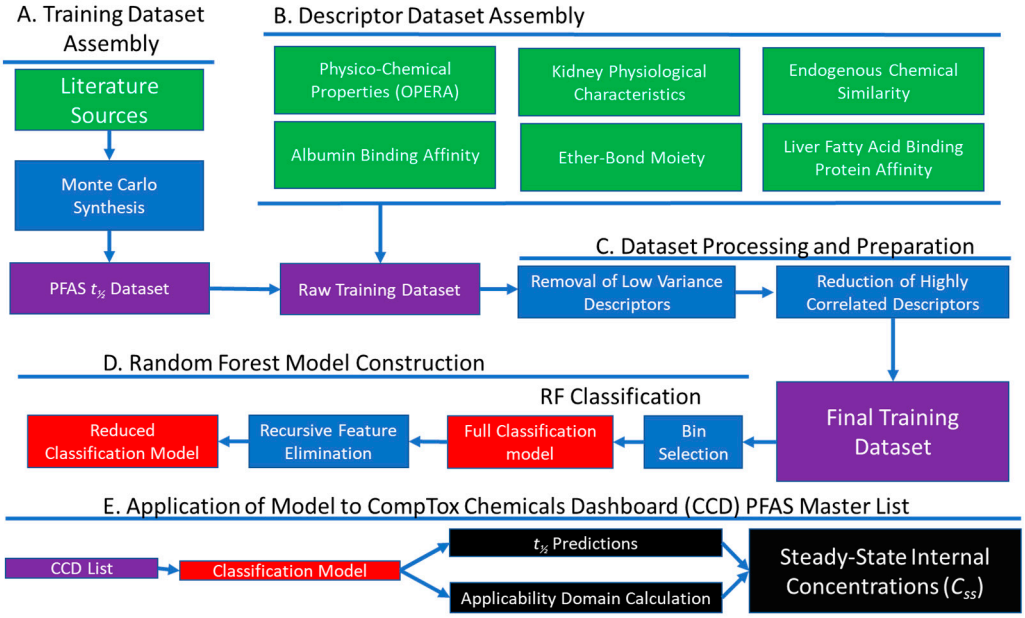

Figure 1. Scientific workflow including (**A**) Training Data Assembly, (**B**) Predictor Dataset Assembly, (**C**) Dataset Processing and Preparation, (**D**) Random Forest Model Construction, and (**E**) Application of the Models to the CCD PFAS list. Green boxes denote data sources, purple boxes denote assembled datasets, red boxes denote models, blue boxes data denote processing steps, black boxes denote model outputs, and arrows indicate flow between steps.

Lau et al. [11,28–30] provide point estimates and ranges synthesizing multiple sources into consensus estimates of chemical- and species-specific $t_{\frac{1}{2}}$. New peer-reviewed measurements were heterogeneously reported, including both measured and calculated mean $t_{\frac{1}{2}}$ values per species, sex, and chemical that were usually accompanied by measures of variance (standard deviation, standard error, or 95% confidence interval). A Monte Carlo approach generated random samples using standard errors (SE) as the bounds of reported values/ranges. See Supplemental Information (S1.1.1) for details.

Distributions were generated by randomly sampling N animals (N = the sample size used in each estimate) from within the SE bounds assigned to each measurement, storing these samples in a vector, and then repeating this process 100 times. Each contributing study was represented in the complete vector of sampled values, in proportion according to sample size. Lastly, we fit a distribution to all samples and used the mean of this distribution

as the $t_{\frac{1}{2}}$ value in our training set for the corresponding chemical/species/sex/dosing method. Distributions were fit using the R package fitdistrplus [72], and an appropriate distribution (between the normal, lognormal, gamma, and exponential) as chosen based on the lowest AIC score.

Data were aggregated across multiple sources into a final dataset with a single value of $t_{\frac{1}{2}}$ per chemical, species, sex, and dosing methodology; a total of 91 datapoints (Table 1). Of these, 50 were distinct measures by species and sex. See the Supplemental Information (S2) for the compiled processed dataset used for ML model construction.

Table 1. PFAS $t_{\frac{1}{2}}$ life estimates used in model construction (full data set is provided in Table S3). Data adopted from Fenton, Ducatman, Boobis, DeWitt, Lau, Ng, Smith and Roberts [11] was augmented by new studies wherever available. Values for chemical/species combinations that were not available were omitted from modeling, but values only available for one sex of a species were assumed to be same for both.

		Rat (*Rattus rattus*)			Mouse (*Mus musculus*)			Monkey (*Macaca fascicularis*)			Human (*Homo sapiens*)		
Chemical CAS/DTXSID	**Sex**	**Value**	**Unit**	**Ref.**	**Value**	**Unit**	**Ref.**	**Value**	**Unit**	**Ref.**	**Value**	**Unit**	**Ref.**
PFBS (C4) 375-73-5 DTXSID5030030	F M	1.5–7.4 3.6–5.0	Hours	[34,73,74]	4.5 5.8	Hours	[75]	1.1 1.6	Days	[73,74]	35 36	Days	[36,73]
PFHxS (C6) 355-46-4 DTXSID7040150	F M	1.3–1.4 26–27	Days	[34,76,77]	27 28	Days	[76]	87 140	Days	[76]	13 14	Years	[35–37,39]
PFOS (C8) 1763-23-1 DTXSID3031864	F M	28–43 34–36	Days	[32,34,77]	38 43	Days	[32]	110 130	Days	[32]	3.4 3.7	Years	[35–39]
PFBA (C4) 375-22-4 DTXSID4059916	F M	1.8 9.2	Hours	[78]	6.2 12	Hours	[78]	1.7	Days	[78]	3	Days	[78]
PFHxA (C6) 307-24-4 DTXSID3031862	F M	0.5–7.3 1.3–11	Hours	[74,79–81]				2.4 5.3	Hours	[74]	32	Days	[31]
PFHpA (C7) 375-85-9 DTXSID1037303	F M	1.2–2.1 1.5–2.4	Hours	[25,79]							140 130	Days	[35,36]
PFOA (C8) 335-67-1 DTXSID8031865	F M	1.7–4.8 8.1–8.5	Hours Days	[25,77,80,82]	16 22	Days	[83]	33 20–21	Days	[84]	3.5	Years	[35–37,85]
PFNA (C9) 375-95-1 DTXSID8031863	F M	6.4 3.3–5.5	Days	[25,86,87]	42 87	Days	[87]				1.7 3.2	Years	[35]
PFDA (C10) 335-76-2 DTXSID3031860	F M	45–59 55–83	Days	[25,80,86]							4 7.1	Years	[35]
F-53B 756426-58-1 DTXSID80892506	F M										18	Years	[88]
GenX 13252-13-6 DTXSID70880215	F M	0.9–2.8 3.0–3.7	Days	[89]	1.0 1.5	Days	[89]	3.3 2.7	Days	[89]	3.4	Days	[90]

2.1.2. Chemical and Species Descriptors (Independent Variables)

We assembled a set of 119 chemical and physiological descriptors as potential predictors of $t_{\frac{1}{2}}$ in ML models. These descriptors characterized either the structure of the chemical agent or the physiology of the animal species; please see full details in the Supplemental Information (S1.1.2). We use the term "predictor" for chemical descriptors that are identified as predictive by ML.

Physico-chemical descriptors (22 descriptors): Physico-chemical descriptors have been shown to characterize TK for organic chemicals present in pharmaceuticals, elsewhere in commerce, and the environment [20,21,91–93]. Here, 18 physico-chemical properties predicted by version 2.7 of the OPERA modeling platform [50] were used. We note that OPERA's training sets were recently updated to include additional PFAS data on LogP, water-solubility, vapor pressure, and melting point (https://github.com/kmansouri/OPERA/releases/tag/v2.7-beta2, accessed on 1 October 2021). In addition, some PFAS have been designed to include an ether bond to potentially facilitate more rapid metabolism [94]. To account for this, a binary descriptor (the ToxPrint Chemotype [95] "COC_alphatic")

was included, denoting the inclusion of an ether bond along the carbon backbone. Finally, average molecular mass and two chain length descriptors were included.

Transport/re-uptake analogs (88 descriptors): Although some PFAS are metabolically stable, they may still be subject to active cellular transport by the body, particularly if they are mistaken for endogenous, non-fluorinated analogs. For example, the long half-life of PFOA in humans has been attributed to reabsorption in the kidney by transporters for the endogenous caprylic acid [26,96]. Unfortunately, PFAS-specific transporter affinities [25,97] and species-specific data on variation on transporter ontogeny [98] are often unavailable. As surrogates for species- and chemical-specific data on the expression of relevant transporters, we examined two types of potential predictors:

Physiological descriptors including kidney structural features as surrogates for renal transporter expression (21 descriptors): The kidney is suspected to be a primary site of PFAS elimination and active transport (secretion/reabsorption) [96,99]. While the species- and chemical-dependent affinities for the transporters driving section/reabsorption are not typically known [26], they are expressed along the surface of the proximal tubule, and so geometry provides one available descriptor that might be correlated with clearance [100], in this case by limiting the surface area available for the expression of transporters. To capture the potential of physical aspects of the kidney as a surrogate for the amount of active transport, a suite of 21 kidney structure descriptors (for example kidney weight, number of nephrons, glomerular surface area) was assembled from Oliver [101] which reported these properties for rat, rabbit, dog, human, cattle, elephant, whale, horse, and chicken. Regressions were made on log-transformed body weight and these regressions were used to make predictions for mouse and monkey based upon body weights reported by Davies and Morris [102] (see GitHub file "CurrentScripts/1_PFAS_Dataset_building.R" for additional information). Overall species body weight was also included as a potential predictor, but was found to be heavily correlated by feature elimination (below).

The similarity of "Defluorinated" PFAS to Endogenous ligands as surrogates for transporter affinity (67 descriptors): As an additional surrogate of the impact of active transport on PFAS, we considered the structural similarity of defluorinated PFAS and a set of 894 endogenous compounds [103] that might be transporter substrates. We presume that structural similarity might result in exogenous chemicals serving as ligands for transporters of endogenous chemicals [104]. Several PFAS have similar non-fluorinated endogenous analogs; for example, caproic acid (that is, Hexanoic acid, CASRN:142-62-1, DTXSID7021607) may be a substrate for human peptide transporter 1 (PEPT1), which facilitates renal reabsorption of peptides in the proximal tubules of the kidney [105,106]. Caproic acid is structurally equivalent to perfluorohexanoic acid (CASRN: 307-24-4, DTXSID:3031862), with hydrogen atoms instead of fluorine atoms along its carbon backbone. To incorporate this information into a predictor dataset, we calculated molecular descriptors (PubChem and Morgan fingerprints) for PFAS in which each fluorine was replaced with hydrogen. Then, we calculated Tanimoto [107] scores (that is, Jaccard similarity) between the defluorinated PFAS and the endogenous compounds for each fingerprint. The subset of endogenous compounds with the highest and lowest similarity for each PFAS was then selected as potential predictors. In this subset, similarity values were discretized (>0.9 being similar (1), otherwise dissimilar (0)) and used as values for each predictor. Among the 11 structures, there were 65 endogenous ligands with at least one non-zero descriptor plus the two maximum values across all ligands for PubChem and Morgan.

Protein Binding (4 descriptors): PFAS bind to specific proteins in the liver and to albumin in serum, which likely influences clearance rates (and therefore $t_{\frac{1}{2}}$) [29]. To account for this, two experimentally available serum–albumin binding rate constants, K_a (M^{-1}) [26], and two binding rate dissociation constants to the liver fatty acid binding protein (L-FABP) [108] were added for a subset of PFAS where measurements had been made.

Categorical Descriptors (2 descriptors): We considered sex (male, female) and dosing type as indicated in the literature source documentation (intravenous, oral, other (epidemiological, via metabolite extrapolation)).

2.1.3. Descriptor Reduction

The total descriptor set (119) was reduced prior to modelling; see Supplemental Information (S1.1.3) for full details. First, we identified and eliminated low variance predictors—that is, those predictors that have nearly the same value for most chemicals—defined as predictors with standard deviation/mean < 0.05. Next, we eliminated highly (>0.9) correlated predictors using the "findCorrelation" function of the caret [55] package of R statistical analysis software. This resulted in 13 numeric descriptors plus the two categorical descriptors that were held out of the quantitative analysis. A summary of the 15 descriptors used is shown in Table 2. Prior to modeling, these were mean-centered and scaled by standard deviation.

Table 2. Summary of Descriptor Set Used. (A) All chemical descriptors used in model construction. * = indicates value is the mean, rather than median. This was used for binary descriptors with either a 1 or 0. For ether bond: 1 = present, 0 = not present; for endogenous similarity measures, 1 = similar (≥90% Tanimoto score), 0 = not similar. Endogenous ligand similarity was included as a surrogate for chemical-specific transporter data. (B) All physiological descriptors included in model by species. As additional surrogates for kidney transporter data, we focused on the geometry of the proximal tubule where they are expressed. Body and kidney weight (italicized) included here for reference but were identified as highly correlated with other features and eliminated by feature reduction for model building. (C) Categorical descriptors used.

A–Chemical Structure Descriptors

Parameter Type	Descriptor	Chemical Coverage (%)	Training Set Median	Training Set Min	Training Set Max
Protein binding	Albumin binding affinity constant (Mol^{-1})	45.45	2.84×10^5	2800	1.10×10^6
Physico-chemical	Average Mass (g/mol)	100	400.1	214	532
	Log Vapor Pressure (mmHg)		−2.07	−8.09	1.53
	Log Octanol: Air		4.16	3.46	6.33
	Log Octanol: Water		3.11	1.43	5.61
	Log Water Solubility (Mol/L at 25 °C)		−2.68	−4.9	−0.5
	Ether bond present		0.13 *	0	1
Endogenous Ligand Similarity	CAS 142-62-1	100	0.18 *	0	1
	CAS 107-92-6		0.088 *		
	CAS 111-16-0		0.066 *		

B–Physiological Descriptors

Species	Proximal tubule diameter (mm)	*Body Weight (kg)*	*Kidney Weight/Body Weight (g/kg)*	Glomerular Surface Area/Proximal Tubule Volume (1/mm)	Glomerular Surface Area/Kidney Weight (mm^2/kg)
Human	0.072	*70*	*2.23*	3.16	1.65
Monkey	0.062	*5*	*2.5*	2.13	2.04
Mouse	0.054	*0.02*	*8*	2.05	2.28
Rat	0.058	*0.24*	*2.92*	2.31	3.26

C–Categorical Descriptors

Sex	Female/Male
Dosing	intravenous, oral, other (epidemiological, via metabolite extrapolation)

2.2. Model Development

We used the R caret package [55] to iteratively call the randomForest package [66] to construct random forest [109] classification models of $t_{\frac{1}{2}}$ using all 15 independent descriptors. The classification approach was selected due to the limited size (91 data points) and scope (11 chemicals) of the training set. All models described below were fit using 10-fold cross validation with 10 repetitions at each step. We evaluated 3, 4, and 5 bin models. Bins were initially split into approximately equal proportions using the OneR package [63]. Bins were slightly adjusted towards whole number time increments. The distribution of data points into the bins was similar, ranging from 22.0 to 29.7%.

To further reduce overfitting, we used recursive feature elimination ("rfe" from caret [55]) to find the model with the highest accuracy with the fewest of the 15 descriptors in Table 2. Starting with the full 15 descriptor set, a series of models were built using sets of

progressively fewer descriptors, with the least "important" descriptor excluded from one series to the next. Predictor importance [109] was quantified as the percentage reduction of model accuracy resulting from permutation of that particular predictor.

2.3. Model Evaluation

Machine learning involves a set of data used to construct the model (a training set) and a second set of data used to evaluate the model (a test set). Our ability to evaluate models was limited, as insufficient data were available to formulate a test set. To partially evaluate the performance of the models, we employed y-randomization; in this case, y-randomization tests for false associations by randomly permuting the $t_{\frac{1}{2}}$ half-life categories, while keeping the descriptors the same. We then refit the model using the same methodology as for the training set. For each y-randomization approach we considered, we built ten models using ten different y-randomized data sets. To evaluate how the distribution of variance of $t_{\frac{1}{2}}$ values between species, between chemicals, and between chemicals and species influences model fitting, this process included $t_{\frac{1}{2}}$ values y-randomized in three ways. First, $t_{\frac{1}{2}}$ values were randomized across all species and chemicals. Next, $t_{\frac{1}{2}}$ values were randomized between species of the same chemical. Third, $t_{\frac{1}{2}}$ values were randomized between chemicals of the same species. Finally, we computed and compared model accuracies between the models constructed using the three types of y-randomized values and non-randomized $t_{\frac{1}{2}}$ values.

The prediction of error of the random forest models was characterized using out-of-bag (OOB) error—each decision tree of the random forest is constructed with a randomized subset (in-bag) of the available data and the data withheld from that tree's construction (OOB) are used as a test set to evaluate the performance of that tree. OOB error of the ensemble of trees (that is, the random forest) is the average OOB error across the ensemble. For a categorical (classification) model, a confusion matrix can be constructed in which each row of represents the instances of the correct class for the samples from a test set, and each column represents the predicted class for samples—a perfect predictor would only have values on the diagonal. For a random forest model constructed with R package randomForest, a confusion matrix is calculated using the OOB data only. Finally, for a categorical model, a "No Information Rate" is calculated as from the largest class percentage in the data set, representing the performance of a "model" in which all samples were predicted to be in the most commonly occurring class.

2.4. Model Application

2.4.1. Prediction of Half-Lives for Novel Chemicals and Species

The $t_{\frac{1}{2}}$ model was applied to the largest list of PFAS available from EPA's CompTox Chemicals Dashboard (CCD) [110] (https://comptox.epa.gov/dashboard/chemical-lists/pfasmaster, accessed on 1 January 2023). PFASMASTER is "a consolidated list of PFAS substances . . . of current interest to researchers and regulators worldwide" that includes PFAS from multiple EPA lists, the OECD New Comprehensive Global Database, KEMI Swedish Chemicals Agency Report, and the NORMAN Suspect List Exchange, among others. This is a list of 8163 PFAS compounds (as of August 2020) with structural information that is listed on the USEPA CompTox Chemicals Dashboard [110]. Predictor values for these compounds were assembled in a similar way to the training set. When a predictor value was unavailable for a chemical, average values were imputed from available data, resulting in some predictors being largely imputed from a small subset of available chemicals (for example, serum–albumin-binding coefficients). In addition, the model was applied to a new species, the domestic dog (*Canus domesticus*), to demonstrate its applicability to a novel species based on changing the model's kidney predictor values. The distribution of the predicted $t_{\frac{1}{2}}$ of the chemicals was plotted for both models for each species.

The applicability domain (AD) of the model was characterized using the methodology of Roy et al. [111]. This method considers whether the distribution of the scaled descriptors

of a novel chemical are captured within the distribution of the training chemical descriptors. Each chemical of the CCD PFAS list was described as either inside or outside the domain of the $t_{\frac{1}{2}}$ model by species. In addition, several predictors were chemical properties estimated with OPERA models, and thus had their own ADs. Thus, each chemical by species was further delineated by whether it was included in the domain of both the $t_{\frac{1}{2}}$ model and all underlying predictor models. We describe the intersection of "All Model ADs" as the "AM domain". Lastly, we used the chemical classification tool ClassyFire [56] to help characterize the predicted chemicals relative to the chemicals in the training set.

2.4.2. Prediction of Serum Concentration

Finally, we used $t_{\frac{1}{2}}$ predictions within a simple 1-compartment model framework to predict steady-state concentrations within the body following exposure. This process included first using predicted $t_{\frac{1}{2}}$ values to calculate elimination rate constants (k_{elim}, Equation (1), units of h^{-1}), which are then used to calculate whole body clearance rates (CL_{tot}, Equation (2), units of L/kg body weight/day) and whole-body, steady-state concentrations (C_{ss}, Equation (3), mg/L):

$$k_{elim} = \frac{\ln(2)}{t_{\frac{1}{2}}} \quad (1)$$

$$CL_{tot} = V_d \times k_{elim} \times 24 \quad (2)$$

$$C_{ss} = \frac{D}{CL_{tot}} \quad (3)$$

In Equation (2), the volume of distribution (V_d) can be defined as the volume needed to yield the concentration of a chemical observed in plasma [112]. To estimate V_d across chemicals and multiple species, we investigated developing models using the same process as for $t_{\frac{1}{2}}$; see Supporting Information (S1.2) for further details. In Equation (2), the factor of "24" allows CL_{tot} to be given in units of in L/kg body weight/day. In Equation (3), steady-state plasma concentration (C_{ss}) is calculated by assuming a constant dose rate (D) of 1 mg/kg body weight/day, which may be then used for reverse dosimetry in vitro–in vivo extrapolation [61]. Using this approach, we predicted steady-state concentrations C_{ss} for each species (including the inferred species, dog (*C. familiaris*)) for PFAS compounds for which QSAR-ready SMILES were available for descriptor calculations, and which fell into All ADs of the model.

3. Results and Discussion

3.1. Half-Life Model Optimization and Selection

Knowledge of chemical-specific $t_{\frac{1}{2}}$ is necessary for exposure reconstruction [15–17] and extrapolation from toxicological studies [14]. For PFOA, we found the TK $t_{\frac{1}{2}}$ scales only weakly across species with bodyweight (R^2 = 0.39). This scaling was on average even less for the other chemicals in our data set (R^2 = 0.26 overall). Instead, a total of 119 descriptors (including body weight) was considered for modeling $t_{\frac{1}{2}}$. The number of descriptors was reduced prior to modelling; see Supplemental Information (S1.1.3) for full details. First, correlation was used as a guide to identify 15 independent descriptors; for example, both body and kidney weight were identified as highly correlated with other physiological features and eliminated. For the 15 descriptors, listed in Table 2, models were constructed iteratively using subsets of the 15 descriptors. This recursive feature elimination process did not further reduce the number of predictors. That is, a model built using all 15 predictors was identified as optimal. We used ML to organize the available in vivo PFAS TK $t_{\frac{1}{2}}$ data into to three, four, or five bins using the predictors in Table 2.

The models had cross-validated accuracies of 82.2%, 86.1%, and 75.3% for three, four or five bins, respectively. Cohens's Kappa [113] was 0.731, 0.812, and 0.688, respectively.

Due to the slightly greater accuracy, the four-bin model was selected (Figure 2): 0–12 h, >12 h to 1 week, >1 week to 60 days, and >60 days. The four-bin model has an error rate of 11%. The misclassification events for the four-bin model were near the margins of the bins (Figure 2), and only occurred for rat for perfluorooctanoic acid (PFOA) and perfluorononanoic acid (PFNA), and perfluorodecanoic acid (PFDA) in mouse.

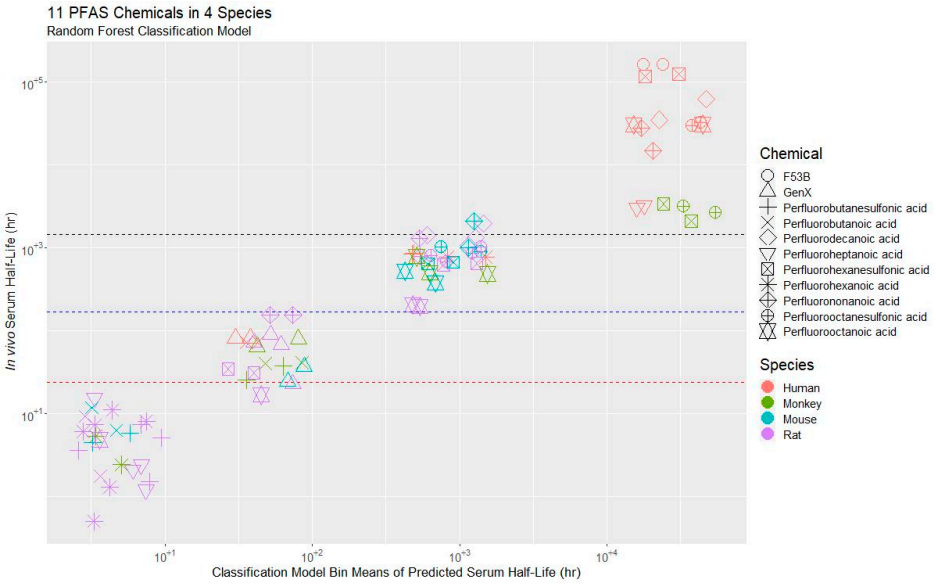

Figure 2. Values of $t_{\frac{1}{2}}$ of the training data (y-axis) vs. classification predictions by the RF Classification model using 15 predictors. Colors signify species, while shapes indicate different PFAS compounds. Bin margins (<12 h, 12 h^{-1} week, 1 $week^{-2}$ months, >2 months) are indicated as dotted lines. Note that observations have been jittered (that is, a small amount of random variation has been added) along the x-axis to increase readability.

Renal elimination includes three processes: glomerular filtration, proximal tubular secretion, and proximal tubular resorption [26]. The mechanistically motivated descriptors initially considered were selected to provide surrogates for PFAS-specific mechanisms of toxicokinetics, with an emphasis on potential renal resorption by the proximal tubules [96]. We do not know the species- and chemical-dependent affinities for the transporters driving section/reabsorption, nor the expression levels of the transporters. We do know that some transporters are expressed along the surface of the proximal tubule. Thus, we can assume that geometry might potentially be correlated with expression level. Similarity to the endogenous ligands of those transporters provides a potential correlate of affinity. The importance of predictors was estimated by the decrease in model performance when the predictor was randomized [109]. The five most important predictors (Table 3) were the average mass of the compound; OPERA model predictions for the logarithmic Octanol:Air partition coefficient and Vapor Pressure; and the kidney descriptors Glomerular Surface Area (SA):Kidney Weight Ratio and Proximal Tubule Diameter. In the case of average molecular mass, a recent review of $t_{\frac{1}{2}}$ data found that PFAS $t_{\frac{1}{2}}$ tends to increase with molecular weight in the same species included in this study [114]. This is consistent with previously observed increases in PFAS $t_{\frac{1}{2}}$ with increasing carbon chain length [26,27,36]. The belief that shorter chains result in faster excretion has prompted a drive to develop alternative chemicals with shorter carbon chains. For example, the chemical GenX is branched and has a shorter $t_{\frac{1}{2}}$ than straight chain PFAS, though without more data we do not know if this generalizes across PFAS.

Table 3. Predictor Importance [109] (percent reduction in accuracy) of all model predictors.

Parameter	Raw Accuracy Change	Scaled Accuracy Change
Average mass	9.49	100
Log Octanol:Air (OPERA)	7.02	73.3
Glomerular Surface Area (SA): Kidney Weight Ratio	6.32	65.6
Proximal Tubule Diameter	6.11	63.4
Log Vapor Pressure (OPERA)	4.86	49.7
Log Octanol:Water (OPERA)	4.37	44.4
Glomerular Surface Area: Proximal Tubule Volume Ratio	4.14	42.0
Log Water Solubility (OPERA)	3.72	37.4
Dosing Form	3.26	32.4
Albumin binding affinity	3.16	31.3
Ether Bond (COC)	2.56	24.8
Sex	2.14	20.2
Similarity to CAS 142-62-1	1.93	18.0
Similarity to CAS 107-92-6	0.61	3.63
Similarity to CAS 111-16-0	0.27	0

We found surrogates for active transport among the predictors. First, the kidney physiology predictors are likely proxies for both physical differences and species variation in the expression of transporters for PFAS. The kidney is a primary site of PFAS disposition and elimination for the body [96,99]. Previous work shows that anionic transporters play a key role in renal excretion and reabsorption of PFAS compounds [26]. Renal transporters reside on the membrane of the proximal tubules [26]. Importantly, proximal tubule structural features (length, surface area) were strongly correlated with body weight. Body weight was used to predict proximal tubule structural features for species for which data were not available (monkey and mouse). These results are, therefore, supportive of the need to further understand renal transporter activity for PFAS across species to better extrapolate to humans. Endogenous ligand similarity was the second type of surrogate for active transport that we considered. Three distinct endogenous ligands were identified after the others were eliminated based on correlation to these three. PFAS similarity to hexanoic acid (DTXSID7021607, CAS 142-62-1), butanoic acid (DTXSID8021515, CAS 107-92-6), and heptanedioic acid (DTXSID5021598, CAS 111-16-0) were considered as a surrogate for transporter affinity. Inclusion in our model indicates that the kidney transporters for which these compounds are ligands may be involved in PFAS $t_{\frac{1}{2}}$.

3.2. Model Evaluation

The aim of supervised machine learning is to identify patterns of descriptor values that predict how each entry in the training set has been "labelled". Here, we labelled each measured $t_{\frac{1}{2}}$ according to a broad bin (or category) spanning a range of times. To evaluate whether the patterns occur by chance, we used y-randomization. Additional models were constructed, following the same procedure as above but using ten y-randomized datasets. In a y-randomized dataset all descriptors were held the same, but the bins for the $t_{\frac{1}{2}}$ values were randomly permuted. The predictive performance of the ML model presented was compared to the performance of multiple y-randomized models. The non-randomized ML model accuracy (86.4%) was better than any of the models constructed with y-randomized data. A model using $t_{\frac{1}{2}}$ values randomized across all species-by-PFAS combinations had low predictive value (accuracy of 32.2 $\pm$ 13.3%).

y-Randomization showed that some variation in $t_{\frac{1}{2}}$ is accounted for by differences at the species and chemical level. The models for $t_{\frac{1}{2}}$ with training data randomized within species but not chemicals (that is, the chemicals were correct) had an accuracy of 36.8 $\pm$ 13.4%. The models where training data chemical identities were randomized, but not species, had an accuracy of 50.2 $\pm$ 15.6%. That is, species-specific data alone provide

information about the plausible values of $t_{\frac{1}{2}}$ of PFAS. However, the large improvement (86.4% vs. 50.2% accuracy) of the fully non-randomized model suggests that enough chemical-species TK interactions exist to justify combining chemical and species information together. The improvement of the full model over any randomized model indicates that the presented model for $t_{\frac{1}{2}}$ does not occur by chance.

The no information rate is an additional effective "null hypothesis" that we examined. The no information rate is the accuracy for a model that predicts all chemicals to be in the most common bin. The four-bin model has an accuracy of 86.4% compared to the no information rate of 27%. That is, the accuracy of the model presented here is an improvement over selecting the most commonly occurring bin. Since 64% of human $t_{\frac{1}{2}}$ falls into Bin 4 (the longest $t_{\frac{1}{2}}$), this provides a species-specific no information rate. The model accuracy (100% for humans) is greater than the human no information rate. The prevalence of predicted Bin 4 chemicals for humans across other PFAS (56%, as discussed in the following section) indicates fewer long $t_{\frac{1}{2}}$ PFAS than would be expected from the human observations alone.

3.3. Application of the Model to a PFAS Library

For each chemical–species prediction, the median half-life of the training data in each bin was used as the predicted $t_{\frac{1}{2}}$. For Bin 1 (<12 h) the median was 4.9 h; for Bin 2 (<1 week) 2.2 days; for Bin 3 (<2 months) 33 days; and for Bin 4 (>2 months) the median used was 3.3 years.

3.3.1. $t_{\frac{1}{2}}$ Predictions for CCD PFAS List

In Figure 3, we show predictions of $t_{\frac{1}{2}}$ across species and sex. Of the 8163 PFAS on the CCD PFAS master list, 6603 had sufficient information for model application. The applicability domain (AD) characterizes the range of chemicals for which we expect accurate predictions [115]. Using the method of Roy, Kar and Ambure [111], we found that the majority (63%) of these chemicals fall into the domain of the model. Across the four species, 4136 PFAS were within the AD (Figure 3A). For humans (over both sexes and dosing methods), 3890 chemicals were estimated to be within AD. Of these, 56% were classified in $t_{\frac{1}{2}}$ Bin 4, 7% were classified in Bin 3, and 37% were classified in Bin 2. We can further restrict predictions to only those chemicals within the ADs of the OPERA models (described as the AM domain; that is, intersection of All Model ADs). The AM domain further reduces the list to 2645 of the 6603 chemicals. For humans, a majority (47%) of this subset of chemicals were predicted to fall into Bin 4, followed by 45% in Bin 2 and 9% in Bin 3. Using the ClassyFire chemical structure ontology [56], the training set could be split into three classes: alkyl halides (9 chemicals), carboxylic acids and derivatives (GenX), and organic and sulfonic acids and derivatives (F-53B). A total of 921 of the PFAS were in these three classes (Figure 3B). For humans, a majority (60%) of this subset of chemicals were predicted to fall into Bin 4, followed by 34% in Bin 2 and 5% in Bin 3.

For chemicals in the domain, $t_{\frac{1}{2}}$ values tended to increase with relative body size. Mice (0.022 kg) and rats (0.225 kg) had more $t_{\frac{1}{2}}$ values in the two fastest bins. Monkeys (3.8 kg), humans (70 kg), and dogs (20 kg) tended to have $t_{\frac{1}{2}}$ values in the three slower bins (Figure 3A). When considering only those chemicals in the AM domain that also align with the three ClassyFire-based classes of the training set (Figure 3B), a similar pattern associated with body size emerges.

The differences in $t_{\frac{1}{2}}$ predictions between species are driven by those parameters in Table 3 that vary between species. From most to least important, these are Glomerular Surface Area (SA) to Kidney Weight Ratio, Proximal Tubule Diameter, and Glomerular Surface Area to Proximal Tubule Volume Ratio. We note that, while overall body weight was included as potential descriptor, it was eliminated during the variable selection process for being highly correlated with these more informative parameters. While these parameters

explicitly describe the geometry of the kidney nephron, they are also potential surrogates for multiple aspects of TK. Geometry impacts the flow through the nephron and extent of glomerular filtration, both of which can, in turn, impact the efficiency of clearance of PFAS from the blood. Additionally, since secretion/resorption transporters line the surface of the proximal tubule, the geometry of the proximal tubules (amount of surface area) provides an upper limit on the amount of transporter expression. Albumin binding affinity has the potential to vary between species [116], but species-specific data were not available for enough chemical–species combinations to be used as descriptors here. The number of chemicals within the domain of applicability in Figure 3 varies between species. This is because we calculated the domain of applicability as a function of both chemical and species descriptors, so that similarity to the training set depends on the specific PFAS-species combination.

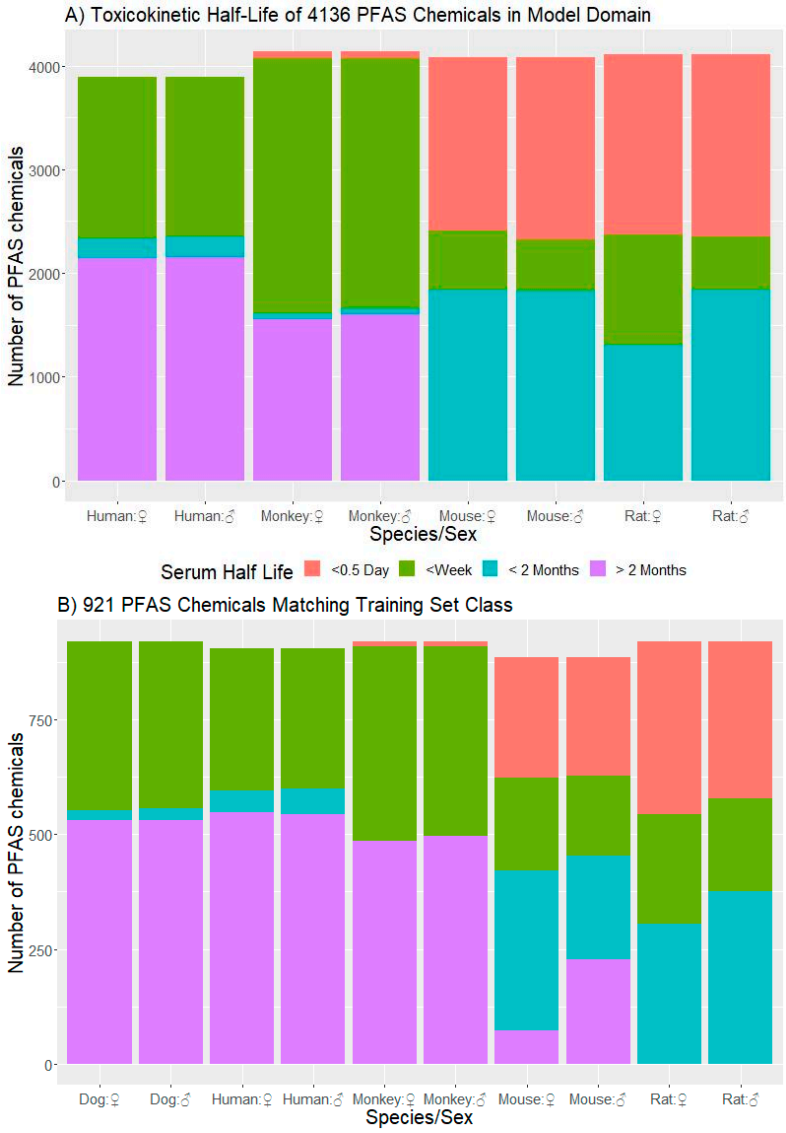

Figure 3. Distributions of predicted $t_{\frac{1}{2}}$ for (**A**) 4136 PFAS within the AD of the RF Classification model, and (**B**) 921 PFAS classified in the same 3 classes as the 11 training set chemicals via ClassyFire. Shown are the number of chemicals predicted to fall within half-life categories by sex (male = ♂, female = ♀) for 5 species. Bins are denoted by color, with pink $\leq$ 12 h, green = 12 h^{-1} week, blue = 1 $week^{-2}$ months, and purple $\geq$ 2 months.

For humans, no chemicals were predicted to fall within the fastest bin (<12 h). For Perfluoroundecanoic acid (PFUnDA, DTXSID8047553), Zhang, Beesoon, Zhu and Martin [35] observed a half-life of 12 years for men and a half-life of 4.5 years for women. Our model correctly predicted the longest bin (>60 days, median 3.32 years) for both sexes. PFUnDA was not included in our data set because a value was available for humans only.

The model predicts that two ether PFAS, Perfluoro-2,5-dimethyl-3,6-dioxanonanoic acid (DTXSID00892442) and Perfluoro(2-((6-chlorohexyl)oxy)ethanesulfonic acid (DTXSID80892506), are bioaccumulative, but that a third is not (Perfluoro-2-methyl-3-oxahexanoic acid, DTXSID70880215). These predictions for ether PFAS are consistent with fish bioconcentration factors for these three chemicals [117].

3.3.2. Prediction of Whole-Body Clearance and Steady-State Concentration

$t_{\frac{1}{2}}$ predictions were combined with an estimate of V_d to calculate steady-state concentrations (C_{ss}) using Equations (1)–(3). The application of our ML methodology did not support a model for V_d (see Supplemental Information S1.2). Instead, for all PFAS, we used the median across ~100 PFAS-by-species measurements (see Supplemental Information S2), 0.202 L/kg bodyweight. Based on the available kidney descriptors [101], we made predictions for a total of eight species (human, cattle, chicken, dog, horse, monkey, mouse, rabbit and rat) across chemicals falling into the AM domain. Clearance predictions in units of L/kg bodyweight/day are provided by column "CLtot.Lpkgbwpday" in Supplemental Information S3 (see Dawson et al._ML PFAS_HL_101722.zip). We anticipate that these predictions may be useful in cross-species extrapolation [118]. For humans, predictions for the chemicals PFOS and PFOA fell into same $t_{\frac{1}{2}}$ bin, corresponding to an average clearance of 1.15×10^{-4} L/kg BW/day. The 2016 EPA Drinking Water Health Advisories used 8.1×10^{-5} and 1.4×10^{-4} L/kg BW/day for these chemicals, respectively [42]. Those values were calculated using measured estimates of $t_{\frac{1}{2}}$ from exposed populations and similar values of V_d (0.23, 0.17 L/kg bw) [42]. Thus, model predictions for these chemicals fell reasonably close (that is, within an order of magnitude) of values calculated using measured data. PFOS and PFOA are the only two PFAS for which regulatory clearance estimates are available at the time of this analysis.

3.3.3. Domain of Applicability

Based on the range of properties of the training data (using the method of Roy, Kar and Ambure [111]), we found that 4136 PFAS were within the AD. Restricting predictions to only those chemicals whose properties were within the ADs of the OPERA predictors reduced this to 2645 PFAS. Alternatively, using the ClassyFire chemical structure ontology [56] restricted predictions to 921 PFAS. Expansion of the AD will require additional PFAS data both for $t_{\frac{1}{2}}$ and physico-chemical properties. It is hoped that both the $t_{\frac{1}{2}}$ model predictions and estimated AD can guide the selection of candidates' PFAS for additional testing.

Many PFAS are anions at physiologic pH. The distribution coefficient LogD characterizes the extent to which ionization impacts tissue partitioning. Initial work showed that LogD (as predicted by OPERA), which describes the distribution of substances as a function of lipophilicity and ionization state, was a predictor of $t_{\frac{1}{2}}$ [119]. Unfortunately, most of the PFAS without $t_{\frac{1}{2}}$ were calculated to not fall within the AD as calculated from the training set with respect to LogD. Thus, omitting LogD as a descriptor here slightly reduced the final model accuracy (from 87.2% to 86.4%), but increased the number of chemicals for which predictions could be made (from 1598 to 4136).

Though ionization has often been considered in drug development [120,121], the treatment of ionization equilibria has typically lagged for non-pharmaceutical chemicals [122]. Instead, success has been found considering other aspects of distribution [123–125]. The presence of molecular fluorines is thought to increase bioavailability through the modulation of ionization in medicinal chemistry [126]. Unfortunately, neither proprietary nor open-source ionization models include many PFAS in their training sets because the data

to do so do not yet exist. Additional measurements for PFAS with more varied LogD might enhance predictivity and provide an evaluation of whether this is an actual applicability domain issue. Other, similar issues are expected to be identified as the $t_{\frac{1}{2}}$ data are expanded. Ultimately, environmental decision makers may not have the luxury of waiting for more data, but might rather identify suitable chemical analogs [127]. It is hoped that this model provides a tentative tool for classifying PFAS TK $t_{\frac{1}{2}}$ on the basis of four bins of "analog" PFAS.

Using the ClassyFire chemical structure ontology, the total set of 6603 PFAS spanned 150 classes. As we calculated AD based on predictor values, the subset within the $t_{\frac{1}{2}}$ model domain spanned 149 classes, and the further subset within the AM domain spanned 121 classes. Alkyl halides made up the largest class of both subsets, with (14%). The other two training set classes, carboxylic acids and derivatives, and organic sulfonic acids and derivatives, made up 9% and 3%, respectively. As estimated from the predictors, there are diverse PFAS included in the $t_{\frac{1}{2}}$ model and AM domains, despite the narrow training diversity employed. This suggests that the predictors included were successful in capturing key drivers of $t_{\frac{1}{2}}$ variability. The most commonly occurring classes that were within the domain of the predictor values but that were not represented in the training set were organofluorides (13%), organooxygen compounds (11%) and fatty acyls (7.5%). These classes make good targets for future data collections. PFAS chemicals outside the AD included 44 classes, with the largest class (17% of chemicals) consisting of benzene and substituted derivatives.

See Supplemental Information S3 for model predictions, applicability domain status, ClassyFire classifications, and steady-state TK predictions for all CCD PFAS list chemicals for which sufficient information was available for model application. All the code to reproduce models and results is available from: https://github.com/USEPA/CompTox-PFASHalfLife (accessed 17 January 2023).

3.4. Model Limitations and Future Considerations

The knowledge of PFAS TK is essential for risk assessment of this large and important class of chemicals. Chemicals with longer $t_{\frac{1}{2}}$ may bioaccumulate, and thus may warrant closer regulatory scrutiny. The majority (56%) of PFAS were predicted to be in the longest $t_{\frac{1}{2}}$ category in humans. This study is an initial attempt to use ML to organize existing data to inform the TK of unmeasured PFAS. The accuracy (86.4%) of the ML developed here was far better than expected by chance (y-randomized accuracy was 32.2 $\pm$ 13.3%). While the constructed model was successful in describing the large variation in $t_{\frac{1}{2}}$ values of the training set (Figure 2) across species and chemicals, its development made use of most of the data available in the published literature. The training set consisted of only four species and 11 chemicals, and was dominated by alkyl halides; namely, perfluoro-carboxylic acids and perfluoro-sulfonic acids. The chemical structural space of the predicted chemicals within the AM domain was much more diverse than the training set. The distribution of $t_{\frac{1}{2}}$ was more heavily weighted toward faster values when chemicals were subset to contain only the three classes of chemicals found in the training set (Figure 3B). If the TK behavior of other classes of PFAS are significantly influenced by factors not captured by the included predictors, then predictions could be unreliable for those chemicals. These uncertainties can only be disentangled with additional data to evaluate this or similar models.

$t_{\frac{1}{2}}$ alone is insufficient to predict the TK of PFAS, including peak and time-integrated plasma concentrations (respectively, C_{max} and area under curve/AUC). Even the simplest approaches to TK modeling (that is, the one compartment empirical model) require the parameter V_d. Despite compiling a dataset of ~100 PFAS-by-species measurements of V_d, our ML model-building approach was unsuccessful (see Supplemental Information S1.2). In comparison to $t_{\frac{1}{2}}$, the compiled values for V_d varied relatively little. Median V_d values ranged across chemicals from 0.139 to 0.368 L/kg, and across species from 0.194 to 0.254 L/kg. Thus, our failure to build more compelling models for predicting inter-

chemical and -species differences in V_d is at least partially a function of the lack of variability among the data relative to the strong uncertainty. Notably, the uncertainty in the literature measurements just for PFOS in rat included V_d that ranged from 0.09 to 7.0 L/kg. This broad uncertainty confounded attempts to build a ML model. However, it possible to use the species-specific predictions provided in Supplemental Information S3 (see Dawson et al._ML PFAS_HL_101722.zip) to make TK predictions for PFAS (including C_{max} and AUC) using the median dataset value of V_d (0.201 L/kg), as we did in Section 3.3.2. In addition, some PFAS have the potential to transform in vivo to a variety of metabolites—which often include the 11 PFAS modeled here [128]. Thus, the development of data and models to predict metabolites that could be coupled with models of $t_{\frac{1}{2}}$ may greatly enhance our ability to predict the TK of PFAS.

Physico-chemical properties and albumin binding can both be measured. While OPERA has recently incorporated measurements for PFAS into its QSARs, additional measurements of PFAS albumin binding may be extremely useful both on their own and as training data for QSARs. Similarly, the critical micellular concentration (CMC) could be measured. CMC is a property that characterizes the aggregation of a chemical into micelles of like molecules, a process that essentially sequesters the chemical away from the rest of the body. For PFAS, the formation of a fluorous-phase of micelles is potentially irreversible [129], and might result in longer $t_{\frac{1}{2}}$ with decreasing CMC. Observed PFAS CMC tend to exceed 10 mM [130] and it remains unknown whether fluctuations or gradients ever lead to such concentrations physiologically. Regardless, experimental values for the CMC are not widely available and the only predictor for CMC available at the time of this writing was omitted for being based on proprietary descriptors [130]. Therefore, the creation of an open-source, verifiable model for CMC could provide an additional relevant PFAS descriptor for predicting TK.

Ample opportunity remains in both the experimental and epidemiological domains for researchers to generate the data with which to test these model predictions, as well as develop alternative models and descriptors. Figure 3 suggests that future in vivo TK studies in rodents might aim to investigate PFAS that are predicted to have different half-lives. This will allow the evaluation and refinement of ML approaches such as those developed here, as well as informing TK study design (for example, if measuring concentration changes over weeks is required). Taken as a synthesis of the available data on PFAS TK data, the prediction of our model might also help analysis of future TK studies by providing informative Bayesian priors [14,16,131]. Similarly, human studies investigating exposure changes (such as switching water supplies) might target unmeasured PFAS predicted to clear more rapidly. Finally, analysis of human biomonitoring data might qualitatively look for greater accumulation (that is, higher observed concentrations) of PFAS predicted to have long $t_{\frac{1}{2}}$ as opposed to those PFAS predicted to clear rapidly in humans.

We hope that the resources presented here will be used as a starting point by the broader scientific community to develop additional data and models for PFAS TK. The term "forever chemical" has been applied to some PFAS with regard to their persistence in the environment, bioaccumulation, and long human half-lives [132]. For humans, this preliminary model distinguishes between those PFAS with $t_{\frac{1}{2}}$ greater than two months and those that are eliminated much faster from the body. "Forever" lurks among those longer $t_{\frac{1}{2}}$ PFAS.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/toxics11020098/s1. File S1. S1_Dawson et al._ML PFAS_HL_101322.pdf [1,26,29,50,72,96,101,104,107,108,110,133–137], File S2. S2_Dawson et al._ML PFAS_HL_101322.xlsx. File S3. S3_Dawson et al._ML PFAS_HL_101722.zip.

Author Contributions: Conceptualization, J.F.W., C.L., R.T.-V., R.S.J. and D.E.D.; methodology, D.E.D., R.R.S., P.P. and J.F.W.; software, D.E.D., P.P., R.R.S. and J.F.W.; validation, D.E.D. and J.F.W.; data curation, C.L., D.E.D., R.R.S. and J.F.W.; writing—original draft preparation, D.E.D.; writing—review and editing, D.E.D., R.T.-V., C.L. and J.F.W.; visualization, D.E.D., R.T.-V. and J.F.W.; supervi-

sion, R.T.-V., R.S.J. and J.F.W.; project administration, J.F.W.; funding acquisition, J.F.W. All authors have read and agreed to the published version of the manuscript.

Funding: The United States Environmental Protection Agency (EPA) through its Office of Research and Development (ORD) funded the research described here.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: An R-markdown file allowing the application of the model to novel chemicals and species is available for download at the following git repository: https://github.com/USEPA/CompTox-PFASHalfLife (accessed 17 January 2023).

Acknowledgments: The authors thank Francesca Grisoni and Rocky Goldsmith for helpful early conversations on this project. We thank Barbara Wetmore and Katherine Phillips for their thorough U.S. EPA internal reviews of the manuscript. We thank Paul Kruse for providing expertise relating to ClassyFire.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. DeWitt, J.C. *Toxicological Effects of Perfluoroalkyl and Polyfluoroalkyl Substances*; Springer: Berlin/Heidelberg, Germany, 2015.
2. Buck, R.C.; Murphy, P.M.; Pabon, M. Chemistry, properties, and uses of commercial fluorinated surfactants. In *Polyfluorinated Chemicals and Transformation Products*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 1–24.
3. Rao, N.S.; Baker, B.E. Textile finishes and fluorosurfactants. In *Organofluorine Chemistry*; Springer: Berlin/Heidelberg, Germany, 1994; pp. 321–338.
4. Rayne, S.; Forest, K.; Friesen, K.J. Congener-specific numbering systems for the environmentally relevant C4 through C8 perfluorinated homologue groups of alkyl sulfonates, carboxylates, telomer alcohols, olefins, and acids, and their derivatives. *J. Environ. Sci. Health Part A* **2008**, *43*, 1391–1401. [CrossRef] [PubMed]
5. Wang, Z.; Cousins, I.T.; Scheringer, M.; Hungerbühler, K. Fluorinated alternatives to long-chain perfluoroalkyl carboxylic acids (PFCAs), perfluoroalkane sulfonic acids (PFSAs) and their potential precursors. *Environ. Int.* **2013**, *60*, 242–248. [CrossRef] [PubMed]
6. 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. *Environ. Int.* **2015**, *75*, 172–179. [CrossRef] [PubMed]
7. Schulz, K.; Silva, M.R.; Klaper, R. Distribution and effects of branched versus linear isomers of PFOA, PFOS, and PFHxS: A review of recent literature. *Sci. Total Environ.* **2020**, *733*, 139186. [CrossRef] [PubMed]
8. Loveless, S.E.; Finlay, C.; Everds, N.E.; Frame, S.R.; Gillies, P.J.; O'Connor, J.C.; Powley, C.R.; Kennedy, G.L. Comparative responses of rats and mice exposed to linear/branched, linear, or branched ammonium perfluorooctanoate (APFO). *Toxicology* **2006**, *220*, 203–217. [CrossRef] [PubMed]
9. Yang, D.; Han, J.; Hall, D.R.; Sun, J.; Fu, J.; Kutarna, S.; Houck, K.A.; LaLone, C.A.; Doering, J.A.; Ng, C.A. Nontarget screening of per-and polyfluoroalkyl substances binding to human liver fatty acid binding protein. *Environ. Sci. Sci. Sci. Technol.* **2020**, *54*, 5676–5686. [CrossRef]
10. Robuck, A.R.; McCord, J.P.; Strynar, M.J.; Cantwell, M.G.; Wiley, D.N.; Lohmann, R. Tissue-Specific Distribution of Legacy and Novel Per-and Polyfluoroalkyl Substances in Juvenile Seabirds. *Environ. Sci. Technol. Lett.* **2021**, *8*, 457–462. [CrossRef]
11. Fenton, S.E.; Ducatman, A.; Boobis, A.; DeWitt, J.C.; Lau, C.; Ng, C.; Smith, J.S.; Roberts, S.M. Per- and Polyfluoroalkyl Substance Toxicity and Human Health Review: Current State of Knowledge and Strategies for Informing Future Research. *Environ. Toxicol. Chem.* **2021**, *40*, 606–630. [CrossRef]
12. Patlewicz, G.; Richard, A.M.; Williams, A.J.; Grulke, C.M.; Sams, R.; Lambert, J.; Noyes, P.D.; DeVito, M.J.; Hines, R.N.; Strynar, M. A chemical category-based prioritization approach for selecting 75 per-and polyfluoroalkyl substances (PFAS) for tiered toxicity and toxicokinetic testing. *Environ. Health Perspect.* **2019**, *127*, 014501. [CrossRef]
13. Glaser, D.; Lamoureux, E.; Opdyke, D.; LaRoe, S.; Reidy, D.; Connolly, J. The impact of precursors on aquatic exposure assessment for PFAS: Insights from bioaccumulation modeling. *Integr. Environ. Assess. Manag.* **2021**, *17*, 705–715. [CrossRef]
14. Wambaugh, J.F.; Setzer, R.W.; Pitruzzello, A.M.; Liu, J.; Reif, D.M.; Kleinstreuer, N.C.; Wang, N.C.Y.; Sipes, N.; Martin, M.; Das, K. Dosimetric anchoring of in vivo and in vitro studies for perfluorooctanoate and perfluorooctanesulfonate. *Toxicol. Sci.* **2013**, *136*, 308–327. [CrossRef] [PubMed]
15. Egeghy, P.P.; Lorber, M. An assessment of the exposure of Americans to perfluorooctane sulfonate: A comparison of estimated intake with values inferred from NHANES data. *J. Expo. Sci. Environ. Epidemiol.* **2011**, *21*, 150–168. [CrossRef]
16. Chiu, W.A.; Lynch, M.T.; Lay, C.R.; Antezana, A.; Malek, P.; Sokolinski, S.; Rogers, R.D. Bayesian Estimation of Human Population Toxicokinetics of PFOA, PFOS, PFHxS, and PFNA from Studies of Contaminated Drinking Water. *Environ. Health Perspect.* **2022**, *130*, 127001. [CrossRef] [PubMed]

17. Lorber, M.; Egeghy, P.P. Simple intake and pharmacokinetic modeling to characterize exposure of Americans to perfluoroctanoic acid, PFOA. *Environ. Sci. Technol.* **2011**, *45*, 8006–8014. [CrossRef] [PubMed]
18. Arnot, J.A.; MacKay, D.; Webster, E.; Southwood, J.M. Screening level risk assessment model for chemical fate and effects in the environment. *Environ. Sci. Technol.* **2006**, *40*, 2316–2323. [CrossRef]
19. Arnot, J.A.; Brown, T.N.; Wania, F. Estimating screening-level organic chemical half-lives in humans. *Environ. Sci. Technol.* **2014**, *48*, 723–730. [CrossRef]
20. Dawson, D.E.; Ingle, B.L.; Phillips, K.A.; Nichols, J.W.; Wambaugh, J.F.; Tornero-Velez, R. Designing QSARs for Parameters of High-Throughput Toxicokinetic Models Using Open-Source Descriptors. *Environ. Sci. Technol.* **2021**, *55*, 6505–6517. [CrossRef]
21. Pradeep, P.; Patlewicz, G.; Pearce, R.; Wambaugh, J.; Wetmore, B.; Judson, R. Using chemical structure information to develop predictive models for in vitro toxicokinetic parameters to inform high-throughput risk-assessment. *Comput. Toxicol.* **2020**, *16*, 100136. [CrossRef]
22. Sipes, N.S.; Wambaugh, J.F.; Pearce, R.; Auerbach, S.S.; Wetmore, B.A.; Hsieh, J.-H.; Shapiro, A.J.; Svoboda, D.; DeVito, M.J.; Ferguson, S.S. An Intuitive Approach for Predicting Potential Human Health Risk with the Tox21 10k Library. *Environ. Sci. Technol.* **2017**, *51*, 10786–10796. [CrossRef]
23. Wambaugh, J.F.; Wetmore, B.A.; Pearce, R.; Strope, C.; Goldsmith, R.; Sluka, J.P.; Sedykh, A.; Tropsha, A.; Bosgra, S.; Shah, I. Toxicokinetic triage for environmental chemicals. *Toxicol. Sci.* **2015**, *147*, 55–67. [CrossRef]
24. Cametti, M.; Crousse, B.; Metrangolo, P.; Milani, R.; Resnati, G. The fluorous effect in biomolecular applications. *Chem. Soc. Rev.* **2012**, *41*, 31–42. [CrossRef] [PubMed]
25. Ohmori, K.; Kudo, N.; Katayama, K.; Kawashima, Y. Comparison of the toxicokinetics between perfluorocarboxylic acids with different carbon chain length. *Toxicology* **2003**, *184*, 135–140. [CrossRef] [PubMed]
26. Han, X.; Nabb, D.L.; Russell, M.H.; Kennedy, G.L.; Rickard, R.W. Renal Elimination of Perfluorocarboxylates (PFCAs). *Chem. Res. Toxicol.* **2012**, *25*, 35–46. [CrossRef] [PubMed]
27. Pizzurro, D.M.; Seeley, M.; Kerper, L.E.; Beck, B.D. Interspecies differences in perfluoroalkyl substances (PFAS) toxicokinetics and application to health-based criteria. *Regul. Toxicol. Pharmacol.* **2019**, *106*, 239–250. [CrossRef]
28. Lau, C. Perfluorinated Compounds. In *Molecular, Clinical and Environmental Toxicology: Volume 3: Environmental Toxicology*; Luch, A., Ed.; Springer: Basel, Switzerland, 2012; pp. 47–86.
29. Lau, C. Perfluorinated compounds: An overview. In *Toxicological Effects of Perfluoroalkyl and Polyfluoroalkyl Substances*; Humana Press: Cham, Switzerland, 2015; pp. 1–21.
30. 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*, 366–394. [CrossRef]
31. Russell, M.H.; Nilsson, H.; Buck, R.C. Elimination kinetics of perfluorohexanoic acid in humans and comparison with mouse, rat and monkey. *Chemosphere* **2013**, *93*, 2419–2425. [CrossRef]
32. Chang, S.-C.; Noker, P.E.; Gorman, G.S.; Gibson, S.J.; Hart, J.A.; Ehresman, D.J.; Butenhoff, J.L. Comparative pharmacokinetics of perfluorooctanesulfonate (PFOS) in rats, mice, and monkeys. *Reprod. Toxicol.* **2012**, *33*, 428–440. [CrossRef]
33. Teeguarden, J.G.; Tan, Y.-M.; Edwards, S.W.; Leonard, J.A.; Anderson, K.A.; Corley, R.A.; Kile, M.L.; Simonich, S.M.; Stone, D.; Tanguay, R.L.; et al. Completing the Link between Exposure Science and Toxicology for Improved Environmental Health Decision Making: The Aggregate Exposure Pathway Framework. *Environ. Sci. Technol.* **2016**, *50*, 4579–4586. [CrossRef]
34. Huang, M.; Dzierlenga, A.; Robinson, V.; Waidyanatha, S.; DeVito, M.; Eifrid, M.; Granville, C.; Gibbs, S.; Blystone, C. Toxicokinetics of perfluorobutane sulfonate (PFBS), perfluorohexane-1-sulphonic acid (PFHxS), and perfluorooctane sulfonic acid (PFOS) in male and female Hsd: Sprague Dawley SD rats after intravenous and gavage administration. *Toxicol. Rep.* **2019**, *6*, 645–655. [CrossRef]
35. Zhang, Y.; Beesoon, S.; Zhu, L.; Martin, J.W. Biomonitoring of perfluoroalkyl acids in human urine and estimates of biological half-life. *Environ. Sci. Technol.* **2013**, *47*, 10619–10627. [CrossRef]
36. Xu, Y.; Fletcher, T.; Pineda, D.; Lindh, C.H.; Nilsson, C.; Glynn, A.; Vogs, C.; Norström, K.; Lilja, K.; Jakobsson, K. Serum half-lives for short-and long-chain perfluoroalkyl acids after ceasing exposure from drinking water contaminated by firefighting foam. *Environ. Health Perspect.* **2020**, *128*, 077004. [CrossRef] [PubMed]
37. Worley, R.R.; Moore, S.M.; Tierney, B.C.; Ye, X.; Calafat, A.M.; Campbell, S.; Woudneh, M.B.; Fisher, J. Per-and polyfluoroalkyl substances in human serum and urine samples from a residentially exposed community. *Environ. Int.* **2017**, *106*, 135–143. [CrossRef] [PubMed]
38. 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**, *115*, 1298–1305. [CrossRef] [PubMed]
39. Li, Y.; Fletcher, T.; Mucs, D.; Scott, K.; Lindh, C.H.; Tallving, P.; Jakobsson, K. Half-lives of PFOS, PFHxS and PFOA after end of exposure to contaminated drinking water. *Occup. Environ. Med.* **2018**, *75*, 46–51. [CrossRef] [PubMed]
40. Krewski, D.; Andersen, M.E.; Tyshenko, M.G.; Krishnan, K.; Hartung, T.; Boekelheide, K.; Wambaugh, J.F.; Jones, D.; Whelan, M.; Thomas, R.; et al. Toxicity testing in the 21st century: Progress in the past decade and future perspectives. *Arch. Toxicol.* **2020**, *94*, 1–58. [CrossRef]

41. 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. *Environ. Int.* **2019**, *129*, 408–422. [CrossRef]
42. U.S. Environmental Protection Agency. PFOA Health Advisory; 2016. Available online: https://www.epa.gov/ (accessed on 1 January 2023).
43. Kenyon, E.M. Interspecies Extrapolation. In *Computational Toxicology: Volume I*; Reisfeld, B., Mayeno, A.N., Eds.; Humana Press: Totowa, NJ, USA, 2012; pp. 501–520.
44. Chiu, W.A.; Barton, H.A.; DeWoskin, R.S.; Schlosser, P.; Thompson, C.M.; Sonawane, B.; Lipscomb, J.C.; Krishnan, K. Evaluation of physiologically based pharmacokinetic models for use in risk assessment. *J. Appl. Toxicol.* **2007**, *27*, 218–237. [CrossRef]
45. Huang, M.; Robinson, V.; Waidyanatha, S.; Dzierlenga, A.; DeVito, M.; Eifrid, M.; Gibbs, S.; Blystone, C. Toxicokinetics of 8: 2 fluorotelomer alcohol (8: 2-FTOH) in male and female Hsd: Sprague Dawley SD rats after intravenous and gavage administration. *Toxicol. Rep.* **2019**, *6*, 924–932. [CrossRef]
46. Bell, S.M.; Chang, X.; Wambaugh, J.F.; Allen, D.G.; Bartels, M.; Brouwer, K.L.R.; Casey, W.M.; Choksi, N.; Ferguson, S.S.; Fraczkiewicz, G.; et al. In vitro to in vivo extrapolation for high throughput prioritization and decision making. *Toxicol. In Vitro* **2018**, *47*, 213–227. [CrossRef]
47. Hope, W.W.; Petraitis, V.; Walsh, T.J. Experimental design considerations in pharmacokinetic studies. In *ADME and Biopharmaceutical Properties*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2008; p. 1059.
48. Loccisano, A.E.; Campbell, J.L., Jr.; Andersen, M.E.; Clewell III, H.J. Evaluation and prediction of pharmacokinetics of PFOA and PFOS in the monkey and human using a PBPK model. *Regul. Toxicol. Pharmacol.* **2011**, *59*, 157–175. [CrossRef]
49. Chou, W.-C.; Lin, Z. Machine Learning and Artificial Intelligence in Physiologically Based Pharmacokinetic Modeling. *Toxicol. Sci.* **2022**, kfac101. [CrossRef] [PubMed]
50. Mansouri, K.; Grulke, C.M.; Judson, R.S.; Williams, A.J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J. Cheminformatics* **2018**, *10*, 10. [CrossRef] [PubMed]
51. Raza, A.; Bardhan, S.; Xu, L.; Yamijala, S.S.R.K.C.; Lian, C.; Kwon, H.; Wong, B.M. A Machine Learning Approach for Predicting Defluorination of Per- and Polyfluoroalkyl Substances (PFAS) for Their Efficient Treatment and Removal. *Environ. Sci. Technol. Lett.* **2019**, *6*, 624–629. [CrossRef]
52. George, S.; Dixit, A. A machine learning approach for prioritizing groundwater testing for per-and polyfluoroalkyl substances (PFAS). *J. Environ. Manag.* **2021**, *295*, 113359. [CrossRef] [PubMed]
53. Mitchell, J.B.O. Machine learning methods in chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, *4*, 468–481. [CrossRef]
54. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
55. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B. *Caret: Classification and Regression Training*, R package version 6.0-86; Astrophysics Source Code Library: Cambridge, MA, USA, 2020.
56. Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **2016**, *8*, 61. [CrossRef]
57. Fantke, P.; Chiu, W.A.; Aylward, L.; Judson, R.; Huang, L.; Jang, S.; Gouin, T.; Rhomberg, L.; Aurisano, N.; McKone, T.; et al. Exposure and toxicity characterization of chemical emissions and chemicals in products: Global recommendations and implementation in USEtox. *Int. J. Life Cycle Assess.* **2021**, *26*, 899–915. [CrossRef]
58. Dowle, M.; Srinivasan, A. *data.table: Extension of 'data.frame'*, R package version 1.14.2; R Foundation for Statistical Computing: Vienna, Austria, 2021.
59. Warnes, G.R.; Bolker, B.; Gorjanc, G.; Grothendieck, G.; Korosec, A.; Lumley, T.; MacQueen, D.; Magnusson, A.; Rogers, J. *Others. Gdata: Various R Programming Tools for Data Manipulation*, R package version 2.18.0.1; R Foundation for Statistical Computing: Vienna, Austria, 2022.
60. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
61. Pearce, R.G.; Setzer, R.W.; Strope, C.L.; Sipes, N.S.; Wambaugh, J.F. httk: R package for high-throughput toxicokinetics. *J. Stat. Softw.* **2017**, *79*, 1–26. [CrossRef]
62. Yan, Y. *MLmetrics: Machine Learning Evaluation Metrics*, R package version 1.1.1; R Foundation for Statistical Computing: Vienna, Austria, 2016.
63. von Jouanne-Diedrich, H. *OneR: One Rule Machine Learning Classification Algorithm with Enhancements*, R Package Version 2.2l; R Foundation for Statistical Computing: Vienna, Austria, 2017.
64. Schauberger, P.; Walker, A. *Openxlsx: Read, Write and Edit xlsx Files*, R package version 4.2.5; R Foundation for Statistical Computing: Vienna, Austria, 2021.
65. Henry, L.; Wickham, H. *Purrr: Functional Programming Tools*, R package version 0.3.4; R Foundation for Statistical Computing: Vienna, Austria, 2020.
66. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
67. Wickham, H.; Bryan, J. *Readxl: Read Excel Files*, R package version 1.4.1; R Foundation for Statistical Computing: Vienna, Austria, 2022.
68. Wickham, H.; Seidel, D. *Scales: Scale Functions for Visualization*, R package version 1.2.1; R Foundation for Statistical Computing: Vienna, Austria, 2022.

69. Qiu, Y. *Showtext: Using Fonts More Easily in R Graphs*, R package version 0.9-5; R Foundation for Statistical Computing: Vienna, Austria, 2022.
70. Wickham, H. *Stringr: Simple, Consistent Wrappers for Common String Operations*, R package version 1.4.1; R Foundation for Statistical Computing: Vienna, Austria, 2022.
71. Wickham, H.; Girlich, M. *Tidyr: Tidy Messy Data*, R package version 1.2.1; R Foundation for Statistical Computing: Vienna, Austria, 2022.
72. Delignette-Muller, M.L.; Dutang, C. fitdistrplus: An R Package for Fitting Distributions. *J. Stat. Softw.* **2015**, *64*, 1–34. [CrossRef]
73. Olsen, G.W.; Chang, S.-C.; Noker, P.E.; Gorman, G.S.; Ehresman, D.J.; Lieder, P.H.; Butenhoff, J.L. A comparison of the pharmacokinetics of perfluorobutanesulfonate (PFBS) in rats, monkeys, and humans. *Toxicology* **2009**, *256*, 65–74. [CrossRef] [PubMed]
74. Chengelis, C.P.; Kirkpatrick, J.B.; Myers, N.R.; Shinohara, M.; Stetson, P.L.; Sved, D.W. Comparison of the toxicokinetic behavior of perfluorohexanoic acid (PFHxA) and nonafluorobutane-1-sulfonic acid (PFBS) in cynomolgus monkeys and rats. *Reprod. Toxicol.* **2009**, *27*, 400–406. [CrossRef] [PubMed]
75. Lau, C.; Rumpler, J.; Das, K.P.; Wood, C.R.; Schmid, J.E.; Strynar, M.J.; Wambaugh, J.F. Pharmacokinetic profile of Perfluorobutane Sulfonate and activation of hepatic nuclear receptor target genes in mice. *Toxicology* **2020**, *441*, 152522. [CrossRef] [PubMed]
76. Sundström, M.; Chang, S.-C.; Noker, P.E.; Gorman, G.S.; Hart, J.A.; Ehresman, D.J.; Bergman, Å.; Butenhoff, J.L. Comparative pharmacokinetics of perfluorohexanesulfonate (PFHxS) in rats, mice, and monkeys. *Reprod. Toxicol.* **2012**, *33*, 441–451. [CrossRef] [PubMed]
77. 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 Chem. Toxicol.* **2016**, *97*, 243–255. [CrossRef]
78. Chang, S.-C.; Das, K.; Ehresman, D.J.; Ellefson, M.E.; Gorman, G.S.; Hart, J.A.; Noker, P.E.; Tan, Y.-M.; Lieder, P.H.; Lau, C. Comparative pharmacokinetics of perfluorobutyrate in rats, mice, monkeys, and humans and relevance to human exposure via drinking water. *Toxicol. Sci.* **2008**, *104*, 40–53. [CrossRef]
79. Kabadi, S.V.; Fisher, J.; Aungst, J.; Rice, P. Internal exposure-based pharmacokinetic evaluation of potential for biopersistence of 6: 2 fluorotelomer alcohol (FTOH) and its metabolites. *Food Chem. Toxicol.* **2018**, *112*, 375–382. [CrossRef]
80. Dzierlenga, A.L.; Robinson, V.G.; Waidyanatha, S.; DeVito, M.J.; Eifrid, M.A.; Gibbs, S.T.; Granville, C.A.; Blystone, C.R. Toxicokinetics of perfluorohexanoic acid (PFHxA), perfluorooctanoic acid (PFOA) and perfluorodecanoic acid (PFDA) in male and female Hsd: Sprague dawley SD rats following intravenous or gavage administration. *Xenobiotica* **2020**, *50*, 722–732. [CrossRef]
81. Gannon, S.A.; Johnson, T.; Nabb, D.L.; Serex, T.L.; Buck, R.C.; Loveless, S.E. Absorption, distribution, metabolism, and excretion of [1-14C]-perfluorohexanoate ([14C]-PFHx) in rats and mice. *Toxicology* **2011**, *283*, 55–62. [CrossRef]
82. 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. *J. Biochem. Toxicol.* **1991**, *6*, 83–92. [CrossRef]
83. Lou, I.; Wambaugh, J.F.; Lau, C.; Hanson, R.G.; Lindstrom, A.B.; Strynar, M.J.; Zehr, R.D.; Setzer, R.W.; Barton, H.A. Modeling single and repeated dose pharmacokinetics of PFOA in mice. *Toxicol. Sci.* **2009**, *107*, 331–341. [CrossRef] [PubMed]
84. Butenhoff, J.L.; Gaylor, D.W.; Moore, J.A.; Olsen, G.W.; Rodricks, J.; Mandel, J.H.; Zobel, L.R. Characterization of risk for general population exposure to perfluorooctanoate. *Regul. Toxicol. Pharmacol.* **2004**, *39*, 363–380. [CrossRef] [PubMed]
85. Bartell, S.M.; Calafat, A.M.; Lyu, C.; Kato, K.; Ryan, P.B.; Steenland, K. Rate of decline in serum PFOA concentrations after granular activated carbon filtration at two public water systems in Ohio and West Virginia. *Environ. Health Perspect.* **2010**, *118*, 222–228. [CrossRef] [PubMed]
86. Kim, S.-J.; Choi, E.-J.; Choi, G.-W.; Lee, Y.-B.; Cho, H.-Y. Exploring sex differences in human health risk assessment for PFNA and PFDA using a PBPK model. *Arch. Toxicol.* **2019**, *93*, 311–330. [CrossRef]
87. Tatum-Gibbs, K.; Wambaugh, J.F.; Das, K.P.; Zehr, R.D.; Strynar, M.J.; Lindstrom, A.B.; Delinsky, A.; Lau, C. Comparative pharmacokinetics of perfluorononanoic acid in rat and mouse. *Toxicology* **2011**, *281*, 48–55. [CrossRef]
88. Shi, Y.; Vestergren, R.; Xu, L.; Zhou, Z.; Li, C.; Liang, Y.; Cai, Y. Human exposure and elimination kinetics of chlorinated polyfluoroalkyl ether sulfonic acids (Cl-PFESAs). *Environ. Sci. Technol.* **2016**, *50*, 2396–2404. [CrossRef]
89. Gannon, S.A.; Fasano, W.J.; Mawn, M.P.; Nabb, D.L.; Buck, R.C.; Buxton, L.W.; Jepson, G.W.; Frame, S.R. Absorption, distribution, metabolism, excretion, and kinetics of 2, 3, 3, 3-tetrafluoro-2-(heptafluoropropoxy) propanoic acid ammonium salt following a single dose in rat, mouse, and cynomolgus monkey. *Toxicology* **2016**, *340*, 1–9. [CrossRef]
90. ECHA—The European Chemicals Agency. *Exposure Related Observations in Humans: Other Data*; E.C.A. 700-242-3; ECHA: Helsinki, Finland, 2021.
91. Zhu, X.-W.; Sedykh, A.; Zhu, H.; Liu, S.-S.; Tropsha, A. The use of pseudo-equilibrium constant affords improved QSAR models of human plasma protein binding. *Pharm. Res.* **2013**, *30*, 1790–1798. [CrossRef]
92. Ingle, B.L.; Veber, B.C.; Nichols, J.W.; Tornero-Velez, R. Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability. *J. Chem. Inf. Model.* **2016**, *56*, 2243–2252. [CrossRef]
93. Yun, Y.E.; Tornero-Velez, R.; Purucker, S.T.; Chang, D.T.; Edginton, A.N. Evaluation of Quantitative Structure Property Relationship Algorithms for Predicting Plasma Protein Binding in Humans. *Comput. Toxicol.* **2020**, *17*, 100142. [CrossRef]
94. Munoz, G.; Liu, J.; Vo Duy, S.; Sauvé, S. Analysis of F-53B, Gen-X, ADONA, and emerging fluoroalkylether substances in environmental and biomonitoring samples: A review. *Trends Environ. Anal. Chem.* **2019**, *23*, e00066. [CrossRef]

95. Yang, C.; Tarkhov, A.; Marusczyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C.H.; Schwoebel, J. New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. *J. Chem. Inf. Model.* **2015**, *55*, 510–528. [CrossRef]
96. 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*, 156–164. [CrossRef]
97. Cheng, X.; Klaassen, C.D. Critical Role of PPAR-α in Perfluorooctanoic Acid– and Perfluorodecanoic Acid–Induced Downregulation of Oatp Uptake Transporters in Mouse Livers. *Toxicol. Sci.* **2008**, *106*, 37–45. [CrossRef] [PubMed]
98. van Groen, B.D.; Nicolaï, J.; Kuik, A.C.; Van Cruchten, S.; van Peer, E.; Smits, A.; Schmidt, S.; de Wildt, S.N.; Allegaert, K.; De Schaepdrijver, L.; et al. Ontogeny of Hepatic Transporters and Drug-Metabolizing Enzymes in Humans and in Nonclinical Species. *Pharmacol. Rev.* **2021**, *73*, 597–678. [CrossRef] [PubMed]
99. Ferrari, F.; Orlando, A.; Ricci, Z.; Ronco, C. Persistent pollutants: Focus on perfluorinated compounds and kidney. *Curr. Opin. Crit. Care* **2019**, *25*, 539–549. [CrossRef] [PubMed]
100. Komiya, I. Urine flow-dependence and interspecies variation of the renal reabsorption of sulfanilamide. *J. Pharmacobiodyn.* **1987**, *10*, 1–7. [CrossRef] [PubMed]
101. Oliver, J. *Nephrons and Kidneys*; Hoeber: New York, NY, USA, 1968.
102. Davies, B.; Morris, T. Physiological parameters in laboratory animals and humans. *Pharm. Res.* **1993**, *10*, 1093–1095. [CrossRef]
103. Rappaport, S.M.; Barupal, D.K.; Wishart, D.; Vineis, P.; Scalbert, A. The blood exposome and its role in discovering causes of disease. *Environ. Health Perspect.* **2014**, *122*, 769. [CrossRef]
104. O'Hagan, S.; Kell, D.B. Understanding the foundations of the structural similarities between marketed drugs and endogenous human metabolites. *Front. Pharmacol.* **2015**, *6*, 105. [CrossRef]
105. Brodin, B.; Nielsen, C.U.; Steffansen, B.; Frøkjær, S. Transport of Peptidomimetic Drugs by the Intestinal Di/tri-peptide Transporter, PepT1. *Pharmacol. Toxicol.* **2002**, *90*, 285–296. [CrossRef] [PubMed]
106. Tramonti, G.; Xie, P.; Wallner, E.I.; Danesh, F.R.; Kanwar, Y.S. Expression and functional characteristics of tubular transporters: P-glycoprotein, PEPT1, and PEPT2 in renal mass reduction and diabetes. *Am. J. Physiol.-Ren. Physiol.* **2006**, *291*, F972–F980. [CrossRef] [PubMed]
107. Tanimoto, T.T. *Elementary Mathematical Theory of Classification and Prediction*; International Business Machines Corp: Armonk, NY, USA, 1958.
108. Zhang, L.; Ren, X.-M.; Guo, L.-H. Structure-based investigation on the interaction of perfluorinated compounds with human liver fatty acid binding protein. *Environ. Sci. Technol.* **2013**, *47*, 11293–11301. [CrossRef] [PubMed]
109. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
110. Williams, A.J.; Grulke, C.M.; Edwards, J.; McEachran, A.D.; Mansouri, K.; Baker, N.C.; Patlewicz, G.; Shah, I.; Wambaugh, J.F.; Judson, R.S. The CompTox Chemistry Dashboard: A community data resource for environmental chemistry. *J. Cheminform.* **2017**, *9*, 61. [CrossRef]
111. Roy, K.; Kar, S.; Ambure, P. On a simple approach for determining applicability domain of QSAR models. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 22–29. [CrossRef]
112. Greenblatt, D.J. Volume of distribution–Again. *Clin. Pharmacol. Drug Dev.* **2014**, *3*, 419–420. [CrossRef]
113. Kvålseth, T.O. Note on Cohen's kappa. *Psychological reports* **1989**, *65*, 223–226. [CrossRef]
114. Hofer, T.; Myhre, O.; Peltola-Thies, J.; Hirmann, D. Analysis of elimination half-lives in MamTKDB 1.0 related to bioaccumulation: Requirement of repeated administration and blood plasma values underrepresent tissues. *Environ. Int.* **2021**, *155*, 106592. [CrossRef]
115. Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504. [CrossRef]
116. 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*, 775–781. [CrossRef] [PubMed]
117. Pan, Y.; Zhang, H.; Cui, Q.; Sheng, N.; Yeung, L.W.Y.; Guo, Y.; Sun, Y.; Dai, J. First Report on the Occurrence and Bioaccumulation of Hexafluoropropylene Oxide Trimer Acid: An Emerging Concern. *Environ. Sci. Technol.* **2017**, *51*, 9553–9560. [CrossRef] [PubMed]
118. Wetmore, B.A. Quantitative in vitro-to-in vivo extrapolation in a high-throughput environment. *Toxicology* **2015**, *332*, 94–101. [CrossRef] [PubMed]
119. Mansouri, K.; Cariello, N.F.; Korotcov, A.; Tkachenko, V.; Grulke, C.M.; Sprankle, C.S.; Allen, D.; Casey, W.M.; Kleinstreuer, N.C.; Williams, A.J. Open-source QSAR models for pKa prediction using multiple machine learning approaches. *J. Cheminformatics* **2019**, *11*, 1–20. [CrossRef]
120. Rodgers, T.; Leahy, D.; Rowland, M. Physiologically based pharmacokinetic modeling 1: Predicting the tissue distribution of moderate-to-strong bases. *J. Pharm. Sci.* **2005**, *94*, 1259–1276. [CrossRef]
121. Peyret, T.; Poulin, P.; Krishnan, K. A unified algorithm for predicting partition coefficients for PBPK modeling of drugs and environmental chemicals. *Toxicol. Appl. Pharmacol.* **2010**, *249*, 197–207. [CrossRef]
122. Strope, C.L.; Mansouri, K.; Clewell, H.J.; Rabinowitz, J.R.; Stevens, C.; Wambaugh, J.F. High-throughput in-silico prediction of ionization equilibria for pharmacokinetic modeling. *Sci. Total Environ.* **2018**, *615*, 150–160. [CrossRef]

123. Arnot, J.A.; Brown, T.N.; Wania, F.; Breivik, K.; McLachlan, M.S. Prioritizing Chemicals and Data Requirements for Screening-Level Exposure and Risk Assessment. *Environ. Health Perspect.* **2012**, *120*, 1565–1570. [CrossRef]
124. Wetmore, B.A.; Wambaugh, J.F.; Ferguson, S.S.; Sochaski, M.A.; Rotroff, D.M.; Freeman, K.; Clewell, H.J., 3rd; Dix, D.J.; Andersen, M.E.; Houck, K.A.; et al. Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. *Toxicol. Sci.* **2012**, *125*, 157–174. [CrossRef]
125. Armitage, J.M.; Wania, F.; Arnot, J.A. Application of mass balance models and the chemical activity concept to facilitate the use of in vitro toxicity data for risk assessment. *Environ. Sci. Technol.* **2014**, *48*, 9770–9779. [CrossRef]
126. Purser, S.; Moore, P.R.; Swallow, S.; Gouverneur, V. Fluorine in medicinal chemistry. *Chem. Soc. Rev.* **2008**, *37*, 320–330. [CrossRef] [PubMed]
127. Wang, N.C.Y.; Zhao, Q.J.; Wesselkamper, S.C.; Lambert, J.C.; Petersen, D.; Hess-Wilson, J.K. Application of computational toxicological approaches in human health risk assessment. *I. A tiered surrogate approach. Regul. Toxicol. Pharmacol.* **2012**, *63*, 10–19. [CrossRef] [PubMed]
128. Tal, T.; Vogs, C. Invited Perspective: PFAS Bioconcentration and Biotransformation in Early Life Stage Zebrafish and Its Implications for Human Health Protection. *Environ. Health Perspect.* **2021**, *129*, 071304. [CrossRef] [PubMed]
129. Riess, J.G. Fluorous micro-and nanophases with a biomedical perspective. *Tetrahedron.* **2002**, *58*, 4113–4131. [CrossRef]
130. Bhhatarai, B.; Gramatica, P. Prediction of aqueous solubility, vapor pressure and critical micelle concentration for aquatic partitioning of perfluorinated chemicals. *Environ. Sci. Technol.* **2011**, *45*, 8120–8128. [CrossRef] [PubMed]
131. Wambaugh, J.F.; Barton, H.A.; Setzer, R.W. Comparing models for perfluorooctanoic acid pharmacokinetics using Bayesian analysis. *J. Pharmacokinet. Pharmacodyn.* **2008**, *35*, 683–712. [CrossRef]
132. Langenbach, B.; Wilson, M. Per- and Polyfluoroalkyl Substances (PFAS): Significance and Considerations within the Regulatory Framework of the USA. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11142. [CrossRef] [PubMed]
133. Jian, J.-M.; Chen, D.; Han, F.-J.; Guo, Y.; Zeng, L.; Lu, X.; Wang, F. A short review on human exposure to and tissue distribution of per-and polyfluoroalkyl substances (PFASs). *Sci. Total Environ.* **2018**, *636*, 1058–1069. [CrossRef]
134. Maurya, H.; Kumar, T.; Kumar, S. Anatomical and physiological similarities of kidney in different experimental animals used for basic studies. *J. Clin. Exp. Nephrol.* **2018**, *3*, 9. [CrossRef]
135. Mandikian, D.; Figueroa, I.; Oldendorp, A.; Rafidi, H.; Ulufatu, S.; Schweiger, M.G.; Couch, J.A.; Dybdal, N.; Joseph, S.B.; Prabhu, S. Tissue physiology of cynomolgus monkeys: Cross-species comparison and implications for translational pharmacology. *AAPS J.* **2018**, *20*, 1–13. [CrossRef]
136. Bolton, E.E.; Wang, Y.; Thiessen, P.A.; Bryant, S.H. PubChem: Integrated platform of small molecules and biological activities. In *Annual Reports in Computational Chemistry*; Elsevier: Amsterdam, The Netherlands, 2008; Volume 4, pp. 217–241.
137. Morgan, H.L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107–113. [CrossRef]