A Machine Learning Model for PFAS Toxicokinetic Half-Life

John Wambaugh

Center for Computational Toxicology and Exposure
Office of Research and Development
U.S. Environmental Protection Agency

Computational Toxicology and Exposure Communities of Practice

Thursday, August 24, 2023

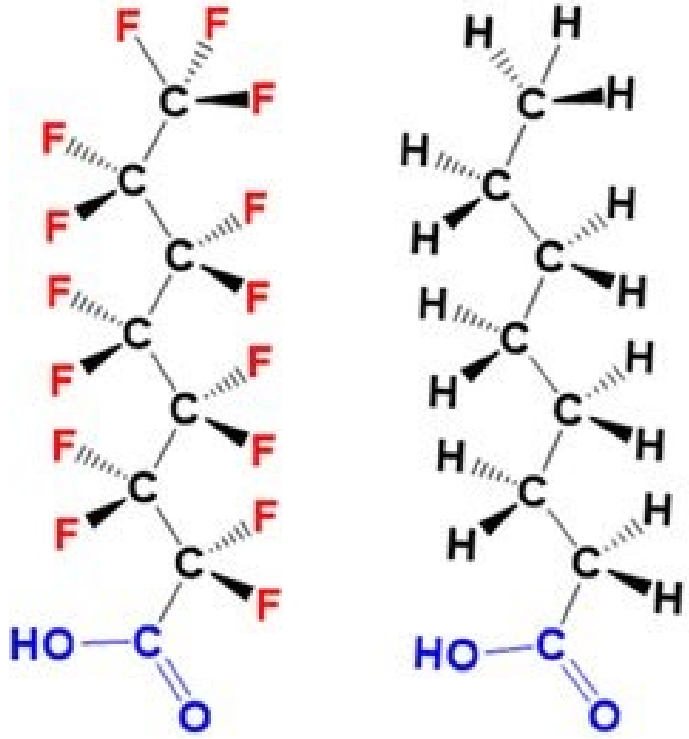
Slides are publicly available
Please send questions to **wambaugh.john@epa.gov**

http://orcid.org/0000-0002-4024-534X

Per- and polyfluoro-alkyl substances (PFAS)

- 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 (DeWitt, 2015)
- Since carbon–fluorine bonds are stronger, they help make PFAS resistant to metabolism and degradation (Buck et al., 2012)
- PFAS are commonly found in human tissues (DeWitt, 2015)

Schwidetzky, et al. (2021)



Perfluorooctanoic Acid

Octanoic Acid

Toxicokinetic Half-Life ($t_{1/2}$) for PFAS

- PFAS are commonly found in human tissues (DeWitt, 2015)
- Toxicokinetic (TK) half-life ($t_{1/2}$) is the amount of time needed for 50% of the chemical to be eliminated from the body.
- $t_{1/2}$ is used to extrapolate from toxicological effects observed in animal species (Wambaugh et al., 2013) and to understand human exposure (Egeghy et al., 2011; Chiu et al., 2022)
- Some PFAS have been noted as having long $t_{1/2}$ (several years in humans)

Issues with PFAS TK Half-Lives

- Typical extrapolation methods for TK parameters of PFAS are unreliable between species and chemicals (Wambaugh et al., 2013; Pizzurro et al., 2019)

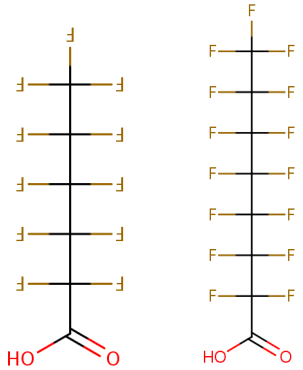
- PFAS have both hydrophobic and lipophobic properties (Rao et al., 1994)
  - For non-PFAS many TK properties are scaled by octanol:water ratio – may not work here
  - Only a dozen PFAS with human measured half-life

- The $t_{½}$ of perfluorohexanoic acid (PFHxA), for example, appears to scale allometrically (proportional to species weight) across mice, rats, monkeys, and humans (Russell et al., 2013)

- In contrast, the $t_{½}$ of the perfluorooctanoic acid (PFOA) spans:
  - a few hours in female rats
  - days in male rats
  - 30–130 days in mice and monkeys
  - 2–4 years in humans

- This large variation for PFOA occurs despite its structural similarity to PFHxA.

PFHxA PFOA



Half-Lives and Exposure

- Knowledge of chemical-specific $t_{½}$ is necessary for relating environmental concentrations of PFAS with concentrations in the body tissues

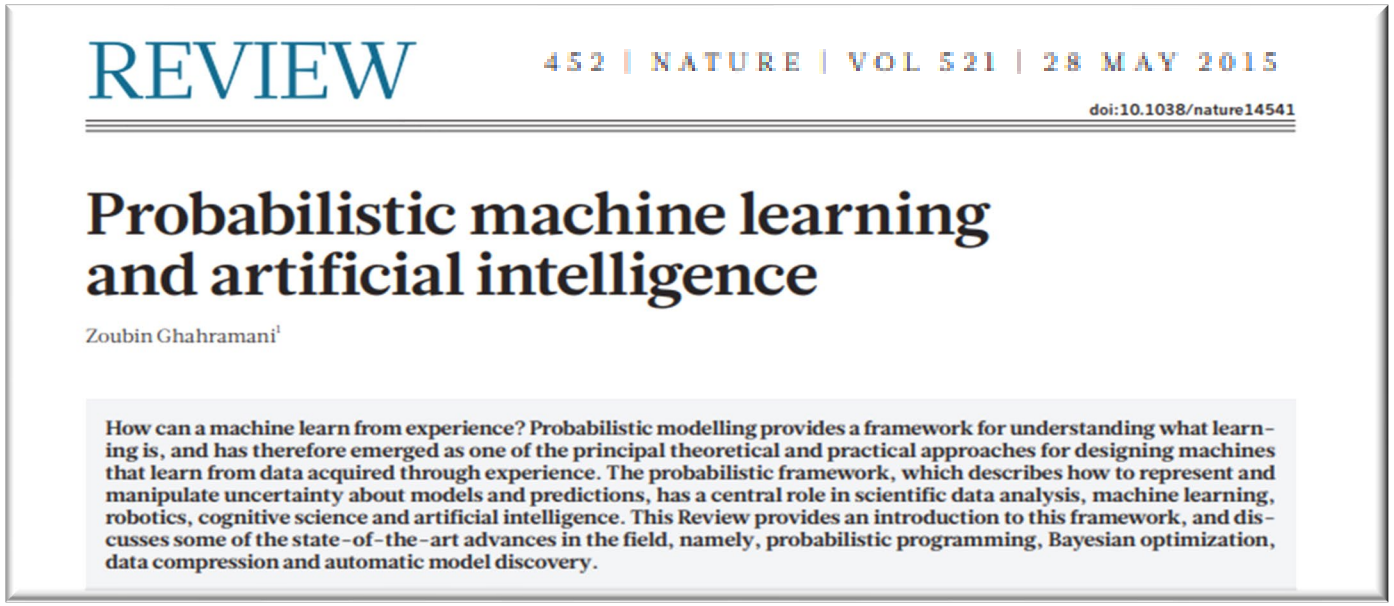
- Using $t_{½}$ and an estimate of how the chemical distributes within the body can:
  1) Predict blood PFAS levels from known external exposures, or
  2) Estimate external exposures from known blood PFAS levels
  (This is an empirical one compartment TK model)

- 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 (Arnot et al., 2006)

- Given the failure of typical approaches for the inter-species or inter-chemical extrapolation of PFAS $t_{½,}$ and the importance of this parameter for understanding the impact of these chemicals in the environment, a new approach is needed.

Machine Learning: A Subset of Artificial Intelligence

"…machine learning can be thought of as inferring plausible models to explain observed data."



REVIEW

452 | NATURE | VOL 521 | 28 MAY 2015

doi:10.1038/nature14541

Probabilistic machine learning and artificial intelligence

Zoubin Ghahramani[1]

How can a machine learn from experience? Probabilistic modelling provides a framework for understanding what learning is, and has therefore emerged as one of the principal theoretical and practical approaches for designing machines that learn from data acquired through experience. The probabilistic framework, which describes how to represent and manipulate uncertainty about models and predictions, has a central role in scientific data analysis, machine learning, robotics, cognitive science and artificial intelligence. This Review provides an introduction to this framework, and discusses some of the state-of-the-art advances in the field, namely, probabilistic programming, Bayesian optimization, data compression and automatic model discovery.

At the EPA we are applying publicly available machine learning algorithms to bridge data gaps and draw inferences from complex data sets.

Machine Learning Overview

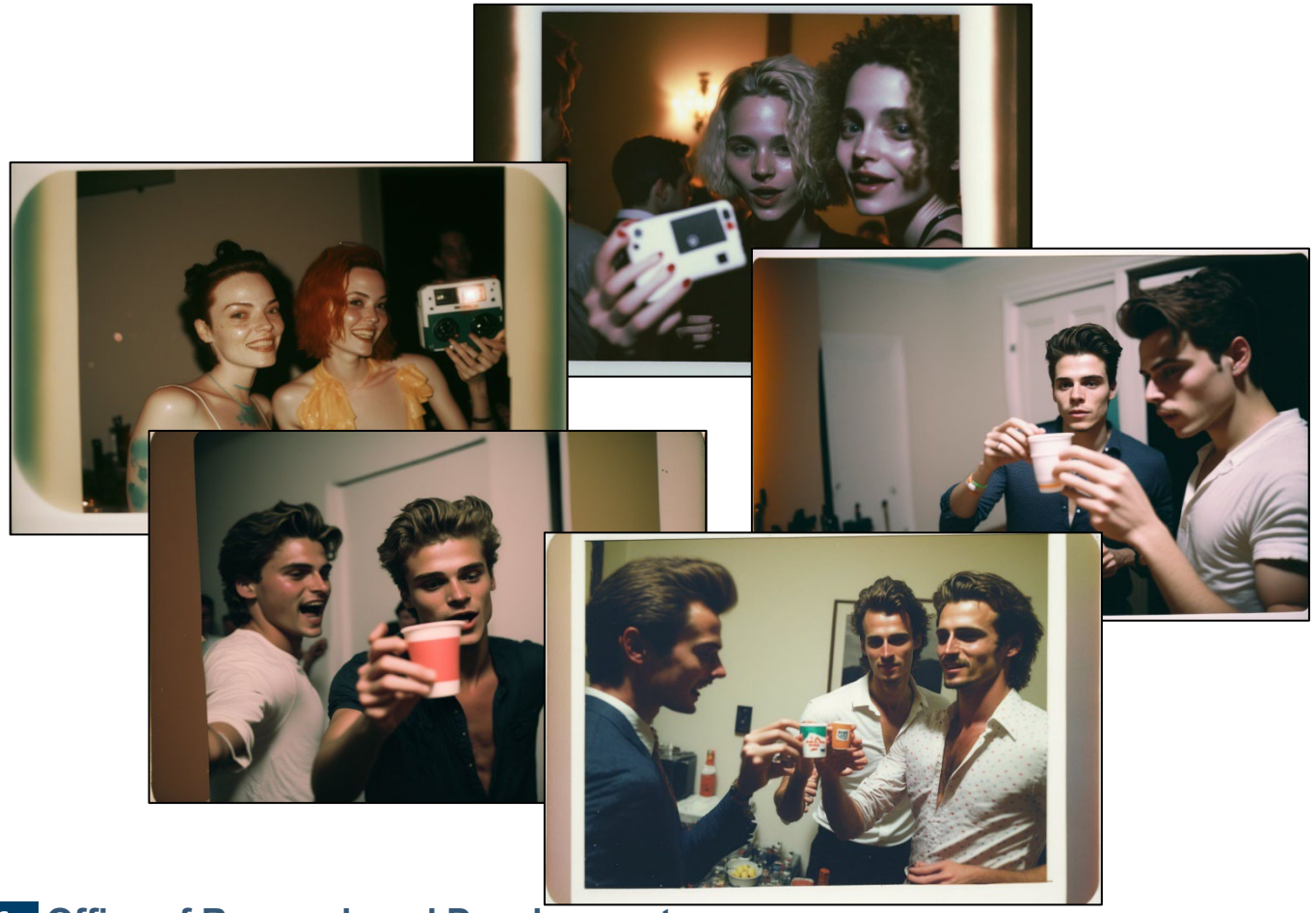
- Machine learning may be more easy to use for categorical predictions

Machine learning image generator prompted for:
“young people at party”

Machine Learning Overview

- Machine learning may be more easy to use for categorical predictions

Machine learning image generator prompted for:
"young people at party"

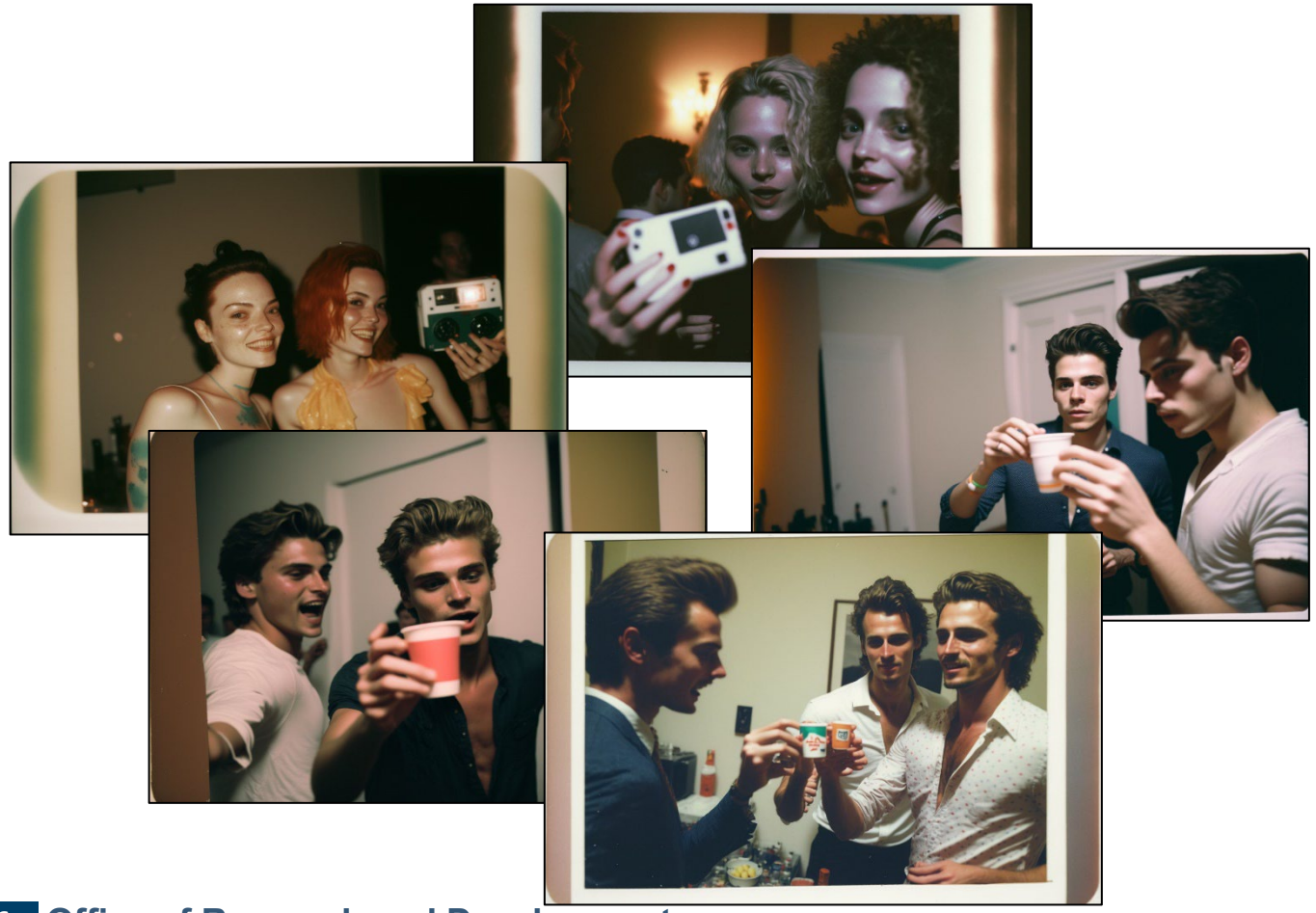


How many fingers do these generated people have?

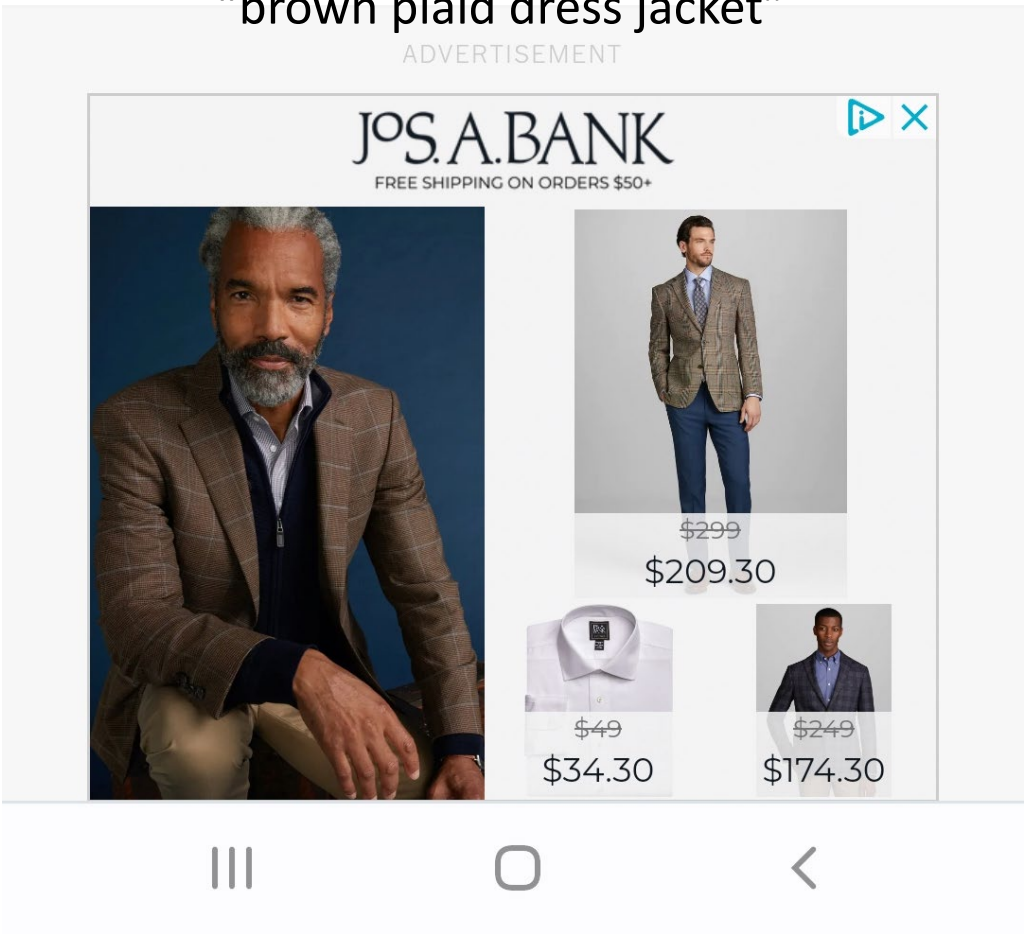
Machine Learning Overview

- Machine learning may be more easy to use for categorical predictions

Machine learning image generator prompted for:
“young people at party”



Advertisements that show up browsing web after searching for
“brown plaid dress jacket”



Machine Learning Overview

- There are many different machine learning technologies, most require some sort of **training set**

- In **supervised machine learning**, there is **labeled training data**: examples annotated with **descriptors**

Overview of Supervised Machine Learning

- Let's focus on **supervised machine learning**, where there is **labeled training data**
- labeled examples are annotated with descriptors

Example	Class
1	**Shirt**
2	**Shirt**
3	**Shirt**
4	**Shirt**
5	**Pants**
6	**Pants**
7	**Pants**
8	**Pants**

Examples are **labeled**

Training Set

Overview of Supervised Machine Learning

- Let's focus on supervised machine learning, where there is labeled training data
- labeled **examples** are annotated with descriptors

Example	Class
1	**Shirt**
2	**Shirt**
3	**Shirt**
4	**Shirt**
5	**Pants**
6	**Pants**
7	**Pants**
8	**Pants**

Examples are **labeled**

Training Set

Overview of Supervised Machine Learning

- Let's focus on supervised machine learning, where there is labeled training data
- labeled examples are annotated with **descriptors**

Example	Class	Color	Buttons	Stripes
1	**Shirt**	Blue	1	0
2	**Shirt**	Red	8	0
3	**Shirt**	Blue	8	1
4	**Shirt**	Green	0	0
5	**Pants**	Khaki	1	0
6	**Pants**	Blue	1	0
7	**Pants**	Black	1	0
8	**Pants**	Blue	4	0

Descriptors

Shirts

Pants

Training Set

Overview of Supervised Machine Learning

- **To train a machine learning model we make choices about what descriptors to include**
- Sometimes the descriptors we want are unavailable
- Further, it is possible that some (or all!) of the available descriptors are not relevant

Example	Class	Color	Buttons	Stripes
1	**Shirt**	Blue	1	0
2	**Shirt**	Red	8	0
3	**Shirt**	Blue	8	1
4	**Shirt**	Green	0	0
5	**Pants**	Khaki	1	0
6	**Pants**	Blue	1	0
7	**Pants**	Black	1	0
8	**Pants**	Blue	4	0

Descriptors

Shirts

Pants

Training Set

Overview of Supervised Machine Learning

- It is possible that some (or all!) of the available descriptors are not relevant
- **Machine learning methods identify the descriptors and values that help make the best predictions**

Example	Class	Color	Buttons	Stripes	Holes	Pockets
1	**Shirt**	Blue	1	0	4	0
2	**Shirt**	Red	8	0	4	0
3	**Shirt**	Blue	8	1	4	0
4	**Shirt**	Green	0	0	4	0
5	**Pants**	Khaki	1	0	3	2
6	**Pants**	Blue	1	0	3	2
7	**Pants**	Black	1	0	3	2
8	**Pants**	Blue	4	0	3	2

Descriptors

Shirts

Pants

Training Set

Overview of Supervised Machine Learning

- It is possible that some (or all!) of the available descriptors are not relevant
- **Machine learning methods identify the descriptors and values that help make the best predictions**

Example	Class	Color	Buttons	Stripes	Holes	Pockets
1	**Shirt**	Blue	1	0	4	0
2	**Shirt**	Red	8	0	4	0
3	**Shirt**	Blue	8	1	4	0
4	**Shirt**	Green	0	0	4	0
5	**Pants**	Khaki	1	0	3	2
6	**Pants**	Blue	1	0	3	2
7	**Pants**	Black	1	0	3	2
8	**Pants**	Blue	4	0	3	2

Descriptors

Shirts

Pants

Training Set

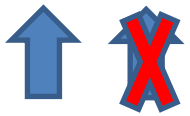
These descriptors both distinguish pants from shirts

Overview of Supervised Machine Learning

- Machine learning methods identify the descriptors and values that help make the best predictions
- However, models may be overfit to their training set – so it's important to **check with external data**

Example	Class	Color	Buttons	Stripes	Holes	Pockets
1	**Shirt**	Blue	1	0	4	0
2	**Shirt**	Red	8	0	4	0
3	**Shirt**	Blue	8	1	4	0
4	**Shirt**	Green	0	0	4	0
5	**Pants**	Khaki	1	0	3	2
6	**Pants**	Blue	1	0	3	2
7	**Pants**	Black	1	0	3	2
8	**Pants**	Blue	4	0	3	2

Descriptors



Shirts

Pants

Training Set

Overview of Supervised Machine Learning

- Finally, sometimes (often), we do not have enough examples of one category or another to build a training set
- Hard to tell a helpful descriptor from an irrelevant descriptor

Example	Class	Color	Buttons	Stripes	Holes	Pockets
1	**Shirt**	Blue	1	0	4	0
2	**Shirt**	Red	8	0	4	0
3	**Shirt**	Blue	8	1	4	0
4	**Shirt**	Green	0	0	4	0
5	**Pants**	Khaki	1	0	3	2

Descriptors

Shirts

Pants

Training Set

Overview of Supervised Machine Learning

- Might end up with a model that always picks dominant category (everything is a shirt would be 80% accurate)*

Example	Class	Color	Buttons	Stripes	Holes	Pockets
1	**Shirt**	Blue	1	0	4	0
2	**Shirt**	Red	8	0	4	0
3	**Shirt**	Blue	8	1	4	0
4	**Shirt**	Green	0	0	4	0
5	**Pants**	Khaki	1	0	3	2

Descriptors

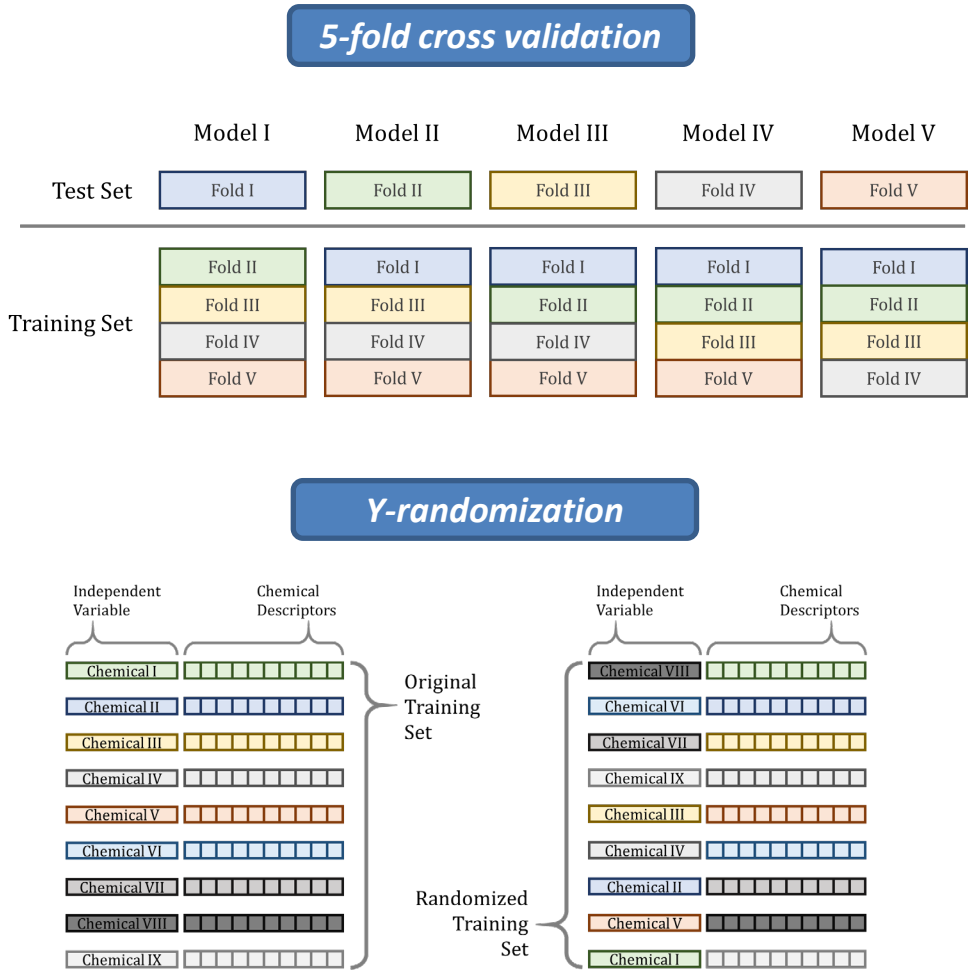
Shirts

Pants

Training Set

*The “no information rate” is an effective “null hypothesis” – it is the accuracy for a model that predicts all chemicals to be in the most common bin.

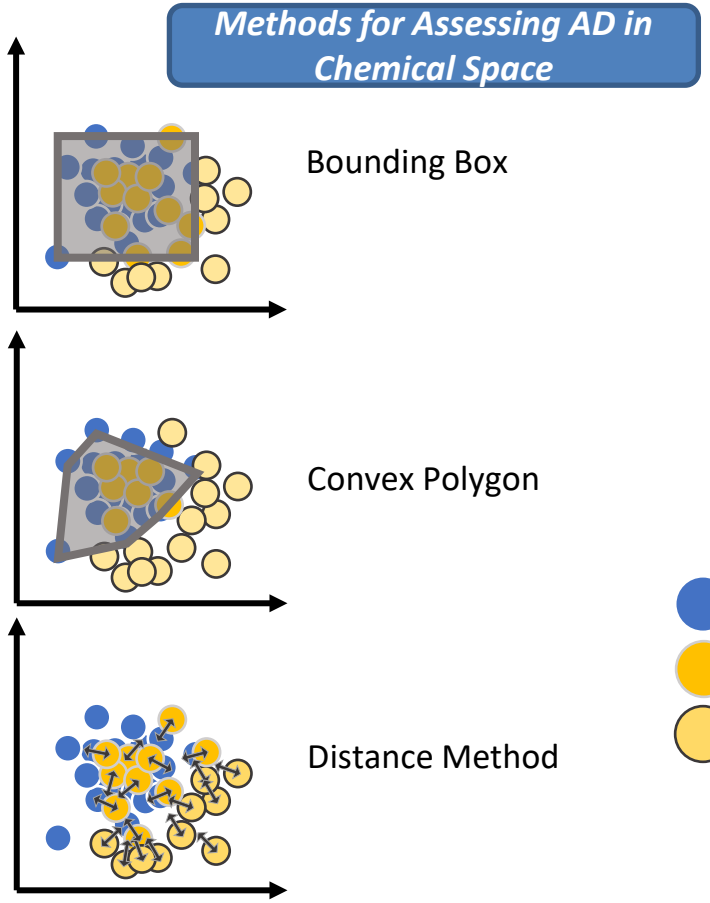
Model Evaluation and Applicability Domain



- QSAR/Machine learning best-practices include an emphasis on model validation and the need to define model applicability domain (AD) in the chemistry space (Tropsha and Golbraikh, 2007)
- Evaluation approaches:
  - 5-fold cross validation (build the model 5 times withholding a different subset of the data each time for testing)
  - Y-randomization (build the model using randomized target assignment to descriptors - does the true model outperform the randomized version?)
  - Evaluation with true external training sets

Figure from Katherine Phillips

Model Evaluation and Applicability Domain



As in Sahigara et al., Molecules (2012)

- Knowledge of the applicability domain (AD) is required for assessing confidence in predictions for a new chemicals and quantifying the utility of additional data

- We estimate AD of the model using the methodology of Roy et al. (2015)

Training Set
External Set – Inside AD
External Set – Outside AD

- Chemical space is defined by the values of the descriptors included in the model – the closer the values of the descriptors for a new chemical are to the training set, the more likely it is to be in domain

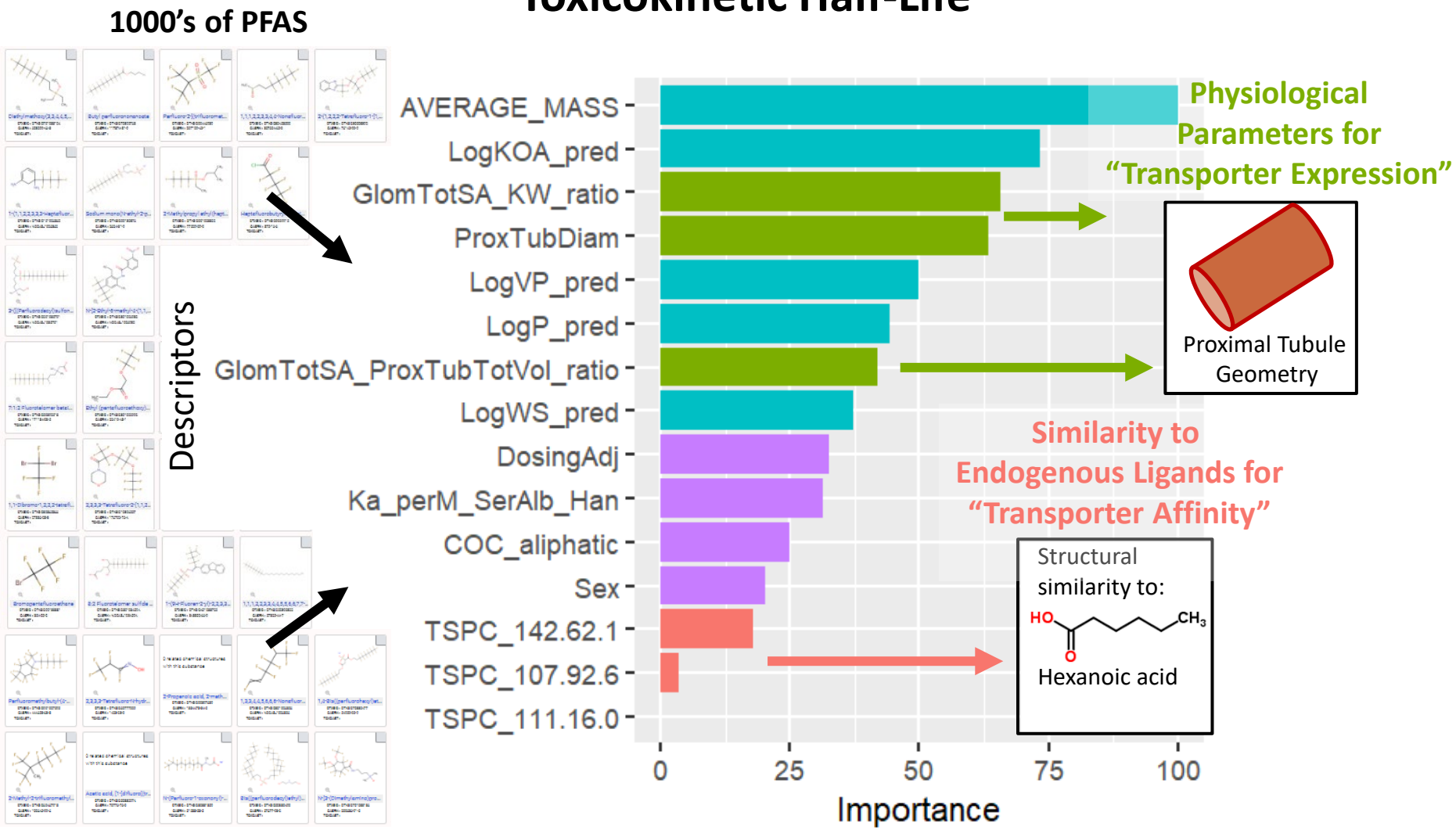
Figure from Katherine Phillips

Dawson et al. (2023)

Machine Learning for PFAS Toxicokinetic Half-Life

toxics

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



PFAS Half-Life Training Set

Data compiled by Lau et al. (2007, 2012, 2015, 2021) and updated for Dawson et al. (2023)

		Humans (Homo sapiens)		
Chemical	Sex	Value	Unit	Ref.
PFBS (C4) 375-73-5 DTXSID5030030	F	35	Days	{Olsen, 2009; Xu, 2020}
	M	36		
PFHxS (C6) 355-46-4 DTXSID7040150	F	13	Yrs	{Zhang, 2013; Worley, 2017; Li, 2018; Xu, 2020}
	M	14		
PFOS (C8) 1763-23-1 DTXSID3031864	F	3.4	Yrs	{Zhang, 2013; Xu, 2020; Worley, 2017; Olsen, 2007; Li, 2018}
	M	3.7		
PFBA (C4) 375-22-4 DTXSID4059916	F	3	Days	{Chang, 2008}
	M			
PFHxA (C6) 307-24-4 DTXSID3031862	F	32	Days	{Russell, 2013}
	M			
PFHpA (C7) 375-85-9 DTXSID1037303	F	140	Days	{Zhang, 2013; Xu, 2020}
	M	130		
PFOA (C8) 335-67-1 DTXSID8031865	F	3.5	Yrs	{Zhang, 2013; Xu, 2020; Worley, 2017; Bartell, 2010}
	M			
PFNA (C9) 375-95-1 DTXSID8031863	F	1.7	Yrs	{Zhang, 2013}
	M	3.2		
PFDA (C10) 335-76-2 DTXSID3031860	F	4	Yrs	{Zhang, 2013}
	M	7.1		
F-53B 756426-58-1 DTXSID80892506	F	18	Yrs	{Shi, 2016}
	M			
GenX 13252-13-6 DTXSID70880215	F	3.4	Days	{ECHA, 2021}
	M			

- Human half-lives for PFAS range from days to years
- Only slight sex differences observed
- 11 chemicals -- not enough data to build a machine learning model
- What if we include data for other species?

PFAS Half-Life Training Set

Data compiled by Lau et al. (2007, 2012, 2015, 2021) and updated for Dawson et al. (2023)

		Rat (Rattus rattus)			Mouse (Mus musculus)			Monkey (Macaca fascicularis)			Humans (Homo sapiens)		
Chemical	**Sex**	**Value**	**Unit**	**Ref.**	**Value**	**Unit**	**Ref.**	**Value**	**Unit**	**Ref.**	**Value**	**Unit**	**Ref.**
PFBS (C4) **375-73-5** **DTXSID5030030**	F	1.5-7.4	Hrs	{Olsen, 2009; Chengelis, 2009; Huang, 2019}	4.5	Hrs	{Lau, 2020}	1.1	Days	{Olsen, 2009; Chengelis, 2009}	35	Days	{Olsen, 2009; Xu, 2020}
	M	3.6-5.0			5.8			1.6			36		
PFHxS (C6) **355-46-4** **DTXSID7040150**	F	1.3-1.4	Days	{Sundstrom, 2012; Kim, 2016; Huang, 2019}	27	Days	{Sundstrom, 2012}	87	Days	{Sundstrom, 2012}	13	Yrs	{Zhang, 2013; Worley, 2017; Li, 2018; Xu, 2020}
	M	26-27			28			140			14		
PFOS (C8) **1763-23-1** **DTXSID3031864**	F	28-43	Days	{Kim, 2016; Huang, 2019; Chang, 2012}	38	Days	{Chang, 2012}	110	Days	{Chang, 2012}	3.4	Yrs	{Zhang, 2013; Xu, 2020; Worley, 2017; Olsen, 2007; Li, 2018}
	M	34-36			43			130			3.7		
PFBA (C4) **375-22-4** **DTXSID4059916**	F	1.8	Hrs	{Chang, 2008}	6.2	Hrs	{Chang, 2008}	1.7	Days	{Chang, 2008}	3	Days	{Chang, 2008}
	M	9.2			12								
PFHxA (C6) **307-24-4** **DTXSID3031862**	F	0.5-7.3	Hrs	{kabadi, 2018; Dzierlenga, 2020; Gannon, 2011; Chengelis, 2009}				2.4	Hours	{Chengelis, 2009}	32	Days	{Russell, 2013}
	M	1.3-11						5.3					
PFHpA (C7) **375-85-9** **DTXSID1037303**	F	1.2-2.1	Hrs	{Ohmori, 2003; Kabadi, 2018}							140	Days	{Zhang, 2013; Xu, 2020}
	M	1.5-.24									130		
PFOA (C8) **335-67-1** **DTXSID8031865**	F	1.7-4.8	Hrs	{Vanden Heuvel, 1991; Ohmori, 2003; Kim, 2016; Dzierlenga, 2020}	16	Days	{Lou, 2009}	33	Days	{Butenhoff, 2004}	3.5	Yrs	{Zhang, 2013; Xu, 2020; Worley, 2017; Bartell, 2010}
	M	8.1-8.5	Days		22			20-21					
PFNA (C9) **375-95-1** **DTXSID8031863**	F	6.4	Days	{Kim, 2019; Tatum, 2011; Ohmori, 2003}	42	Days	{Tatum, 2011}				1.7	Yrs	{Zhang, 2013}
	M	3.3-5.5			87						3.2		
PFDA (C10) **335-76-2** **DTXSID3031860**	F	45-59	Days	{Ohmori, 2003; Kim, 2019; Dzierlenga, 2020}							4	Yrs	{Zhang, 2013}
	M	55-83									7.1		
F-53B **756426-58-1** **DTXSID80892506**	F										18	Yrs	{Shi, 2016}
	M												
GenX **13252-13-6** **DTXSID70880215**	F	0.9-2.8	Days	{Gannon, 2016}	1.0	Days	{Gannon, 2016}	3.3	Days	{Gannon, 2016}	3.4	Days	{ECHA, 2021}
	M	3.0-3.7			1.5			2.7					

Supervised Machine Learning Model

- **Machine learning methods identify the descriptors that make the best predictions**

Example	Class	Color	Buttons	Stripes	Holes	Pockets
1	**Shirt**	Blue	1	0	4	0
2	**Shirt**	Red	8	0	4	0
3	**Shirt**	Blue	8	1	4	0
4	**Shirt**	Green	0	0	4	0
5	**Pants**	Khaki	1	0	3	2
6	**Pants**	Blue	1	0	3	2
7	**Pants**	Black	1	0	3	2
8	**Pants**	Blue	4	0	3	2

Descriptors

Shirts

Pants

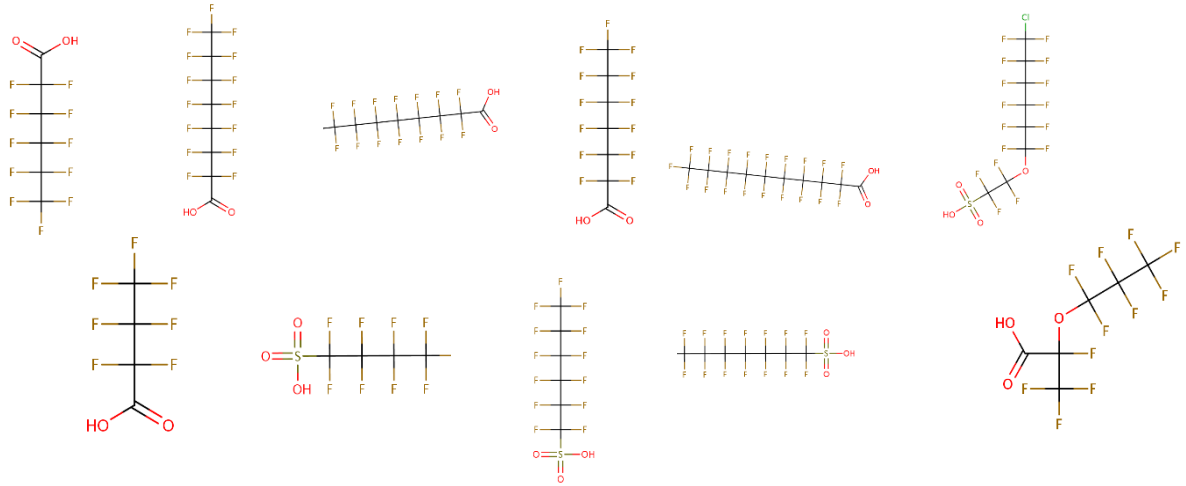
Training Set

Supervised Machine Learning Model

- **Machine learning methods identify the descriptors that make the best predictions**

Example	PFAS	Species	Half-Life	Chemical Structure	Physiology	Categorical
1	**PFHxA**	Human	Slow	#	#	#
2	**PFOA**	Human	Very Slow	#	#	#

Descriptors



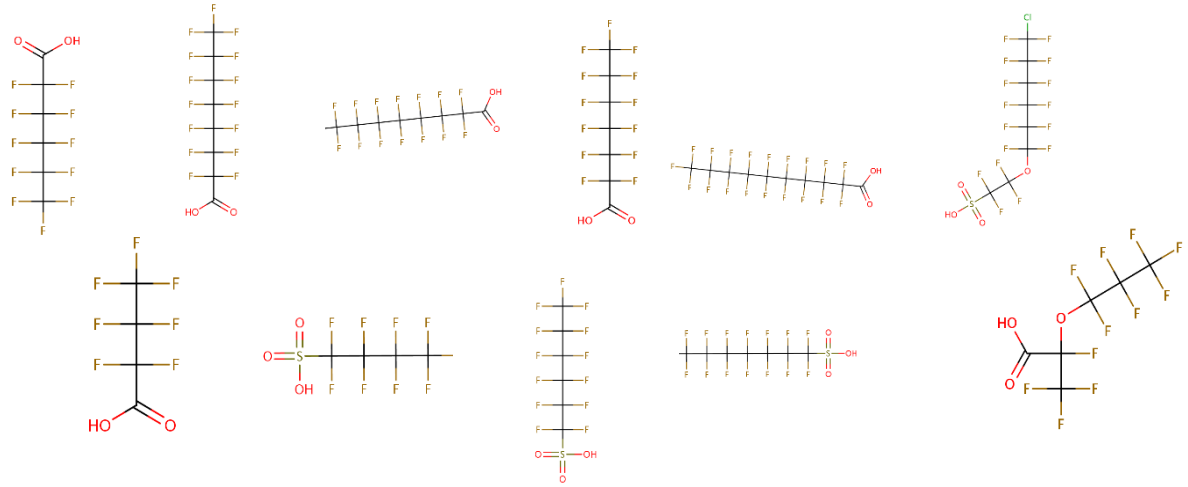
We only have half-life measurements for ~dozen PFAS

Supervised Machine Learning Model

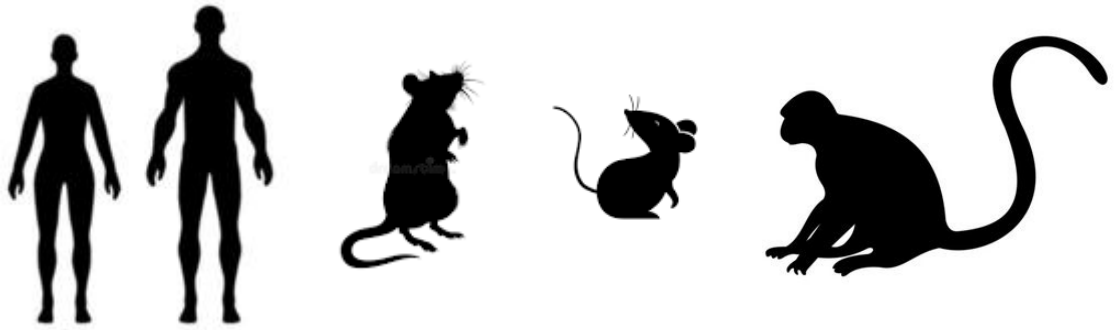
- **Machine learning methods identify the descriptors that make the best predictions**

Example	PFAS	Species	Half-Life	Chemical Structure	Physiology	Categorical
1	**PFHxA**	Human	Slow	#	#	#
2	**PFOA**	Human	Very Slow	#	#	#
3	**PFBS**	Mouse	Very Fast	#	#	#
4	**PFOS**	Mouse	Slow	#	#	#
5	**PFHxA**	Rat	Very Fast	#	#	#
6	**PFOA**	Rat	Fast	#	#	#
7	**PFBS**	Monkey	Fast	#	#	#
8	**PFOS**	Monkey	Very Slow	#	#	#

Descriptors



Let's use huge interspecies variability to our advantage



EPA United States Environmental Protection Agency

Descriptors

- We assembled a set of 119 chemical and physiological (species) descriptors as potential predictors of $t_{½}$ in ML models

Chemical Structure Descriptors:

- Protein Binding (4 descriptors): serum albumin and liver fatty acid binding protein
- Physico-chemical descriptors (22 descriptors)
- Transport/re-uptake analogs:
  - Similarity of "Defluorinated" PFAS to Endogenous ligands as surrogates for transporter affinity (67 descriptors)

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	2.84E+05	2800	1.10E+06
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	Glomerular Surface Area / Kidney Weight
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)

Descriptors

- We assembled a set of 119 chemical and physiological (species) descriptors as potential predictors of $t_{½}$ in ML models

Physiological Descriptors:

- Transport/re-uptake analogs:
  - Physiological descriptors including kidney structural features as surrogates for renal transporter expression (21 descriptors)
  - Body weight initially considered but eliminated for being too correlated with other descriptors

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	2.84E+05	2800	1.10E+06
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	Glomerular Surface Area / Kidney Weight
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)

Descriptors

- We assembled a set of 119 chemical and physiological (species) descriptors as potential predictors of $t_{½}$ in ML models

Categorical Descriptors:

- Sex and route of dose administration

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	2.84E+05	2800	1.10E+06
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

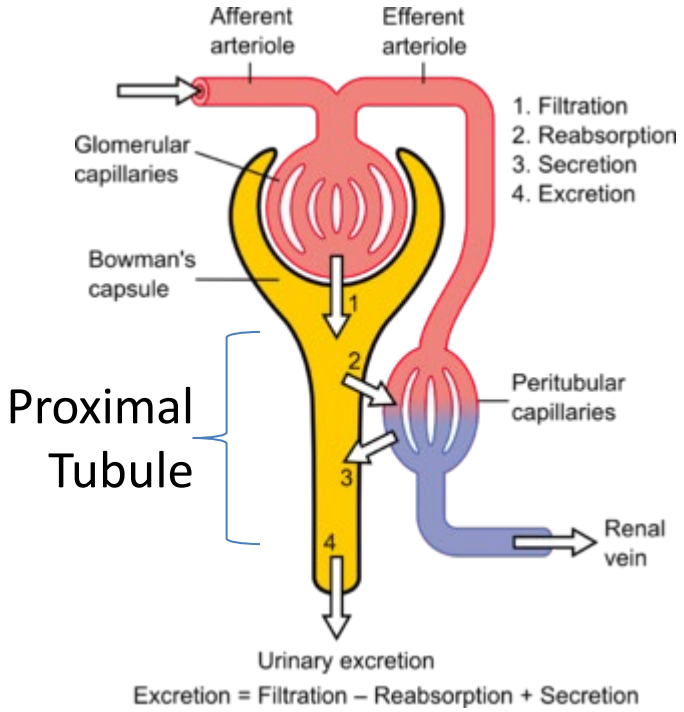
Species	Proximal tubule diameter (mm)	*Body Weight (kg)*	*Kidney Weight / Body Weight (g/kg)*	Glomerular Surface Area / Proximal Tubule Volume	Glomerular Surface Area / Kidney Weight
Human	0.072	*70*	2.23	3.16	1.65
Monkey	0.062	*5*	2.5	2.13	2.04
Mouse	0.054	*0.02*	8	2.05	2.28
Rat	0.058	*0.24*	2.92	2.31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

C - Categorical Descriptors

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

Why Might Transporter Surrogates Work for PFAS TK?

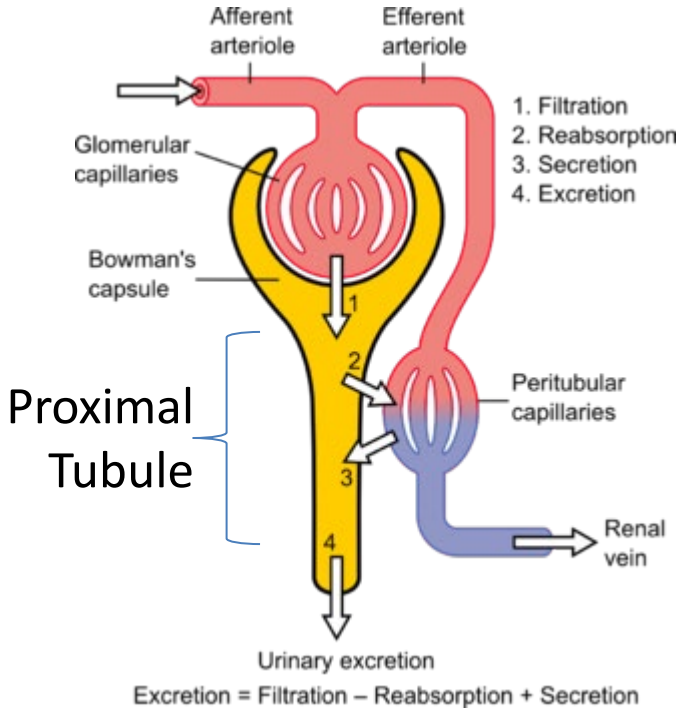
- After glomerular filtration from plasma into the lumen of the proximal tubule, chemicals are subject to active secretion to and absorption from the lumen by the cells that make up the surface of the proximal tubule
- Ohmori et al. (2003) hypothesized that some PFAS are substrates for reabsorption by the kidney tubules, perhaps because of their similarity to nutrient rich fatty acids (PFOA for example is caprylic acid with hydrogens replaced by fluorines).
- Expression of some fatty acid transporters is modulated by sex hormones
- Different PFAS may variously have greater affinity for different transporters
- Different species may have varying expression levels
- Generally do not know affinity as a function of PFAS, transporter, and species



Kidney Physiology (Wikipedia)

Why Might Transporter Surrogates Work for PFAS TK?

- **Generally do not know affinity as a function of PFAS, transporter, and species**
- As a **surrogate for transporter expression**:
  - We do know how the geometry (shape, surface area, volume) of the proximal tubules varies between species (Oliver, 1968)
- As a **surrogate for transporter affinity** we can also calculate how similar each PFAS is to endogenous (naturally present) chemicals:
  - We assume that transporters are more likely to act on endogenous chemicals
  - Compared PFAS to 894 endogenous chemicals from Rappaport et al. (2014)
  - Replaced all fluorines on each PFAS with hydrogens and then calculated structural similarity with Tanimoto (1958) scores



Kidney Physiology (Wikipedia)

Model Building

- We used method of random forests to construct a machine learning model (Brieman, 2001)
- We pared the original set of descriptors down to 15 through elimination of correlated or unchanging descriptors
- We used recursive feature elimination to balance accuracy with subsets of these descriptors

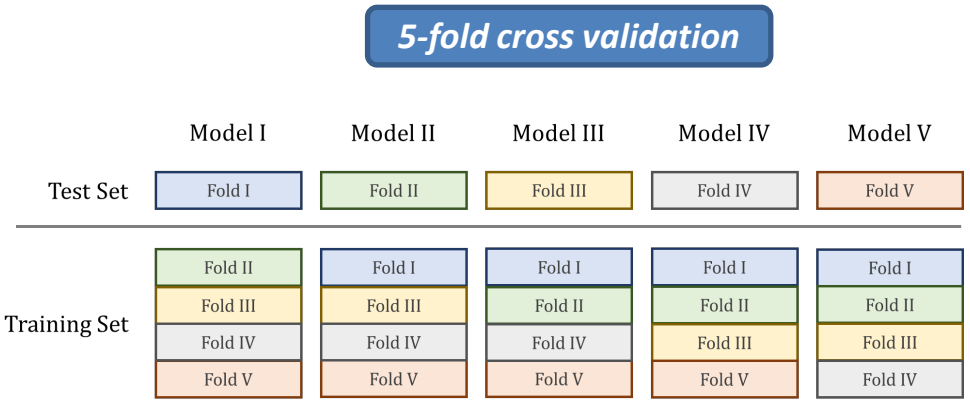


Figure from Katherine Phillips

- We used cross-validation to determine optimal number of half-life bins
  - Cross-validated accuracies of 82.2%, 86.1%, and 75.3% for three, four or five bins

Model Performance

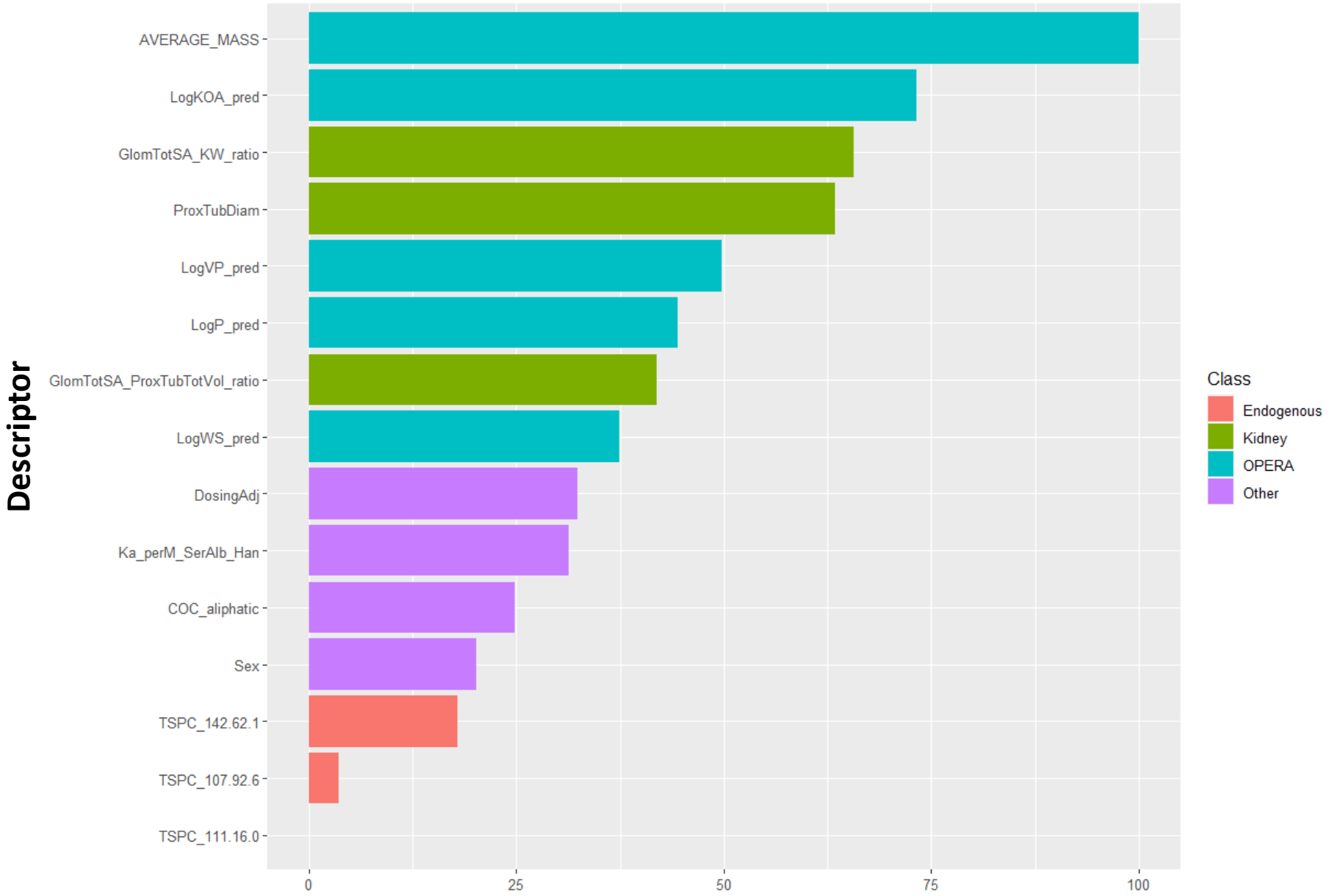
- The four-bin model was selected – chemicals were grouped into half-life bins: 0–12 h, >12 h to 1 week, >1 week to 60 days, and >60 days.

- The four-bin model has an accuracy of 86.4% compared to the no information rate of 27%.

- The non-randomized ML model accuracy (86.4%) was better than any of the models constructed with y-randomized data:
  - A model using $t_{½}$ values randomized across all species-by-PFAS combinations had low predictive value (accuracy of 32.2 ± 13.3%)
  - The models for $t_{½}$ with training data randomized within species but not chemicals (that is, the chemicals were correct) had an accuracy of 36.8 ± 13.4%.
  - The models where training data chemical identities were randomized, but not species, had an accuracy of 50.2 ± 15.6%..

Descriptors

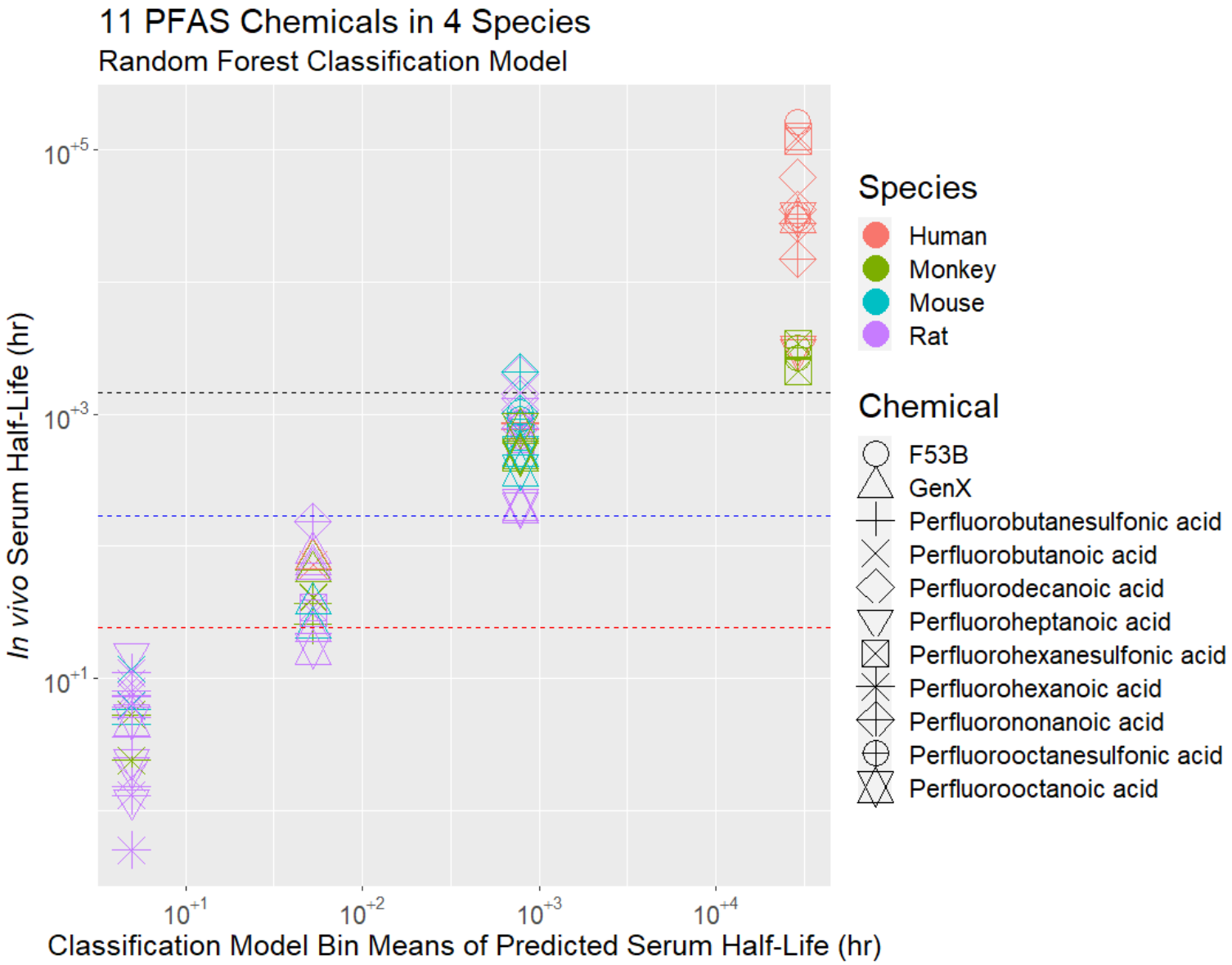
- We found surrogates for active transport among the predictors.
- The kidney physiology predictors are proxies for both physical differences and species variation in the expression of transporters
- PFAS similarity to endogenous hexanoic, butanoic, and heptanedioic acids were considered as surrogates for transporter affinity



Importance metric from Archer and Kimes (2008)

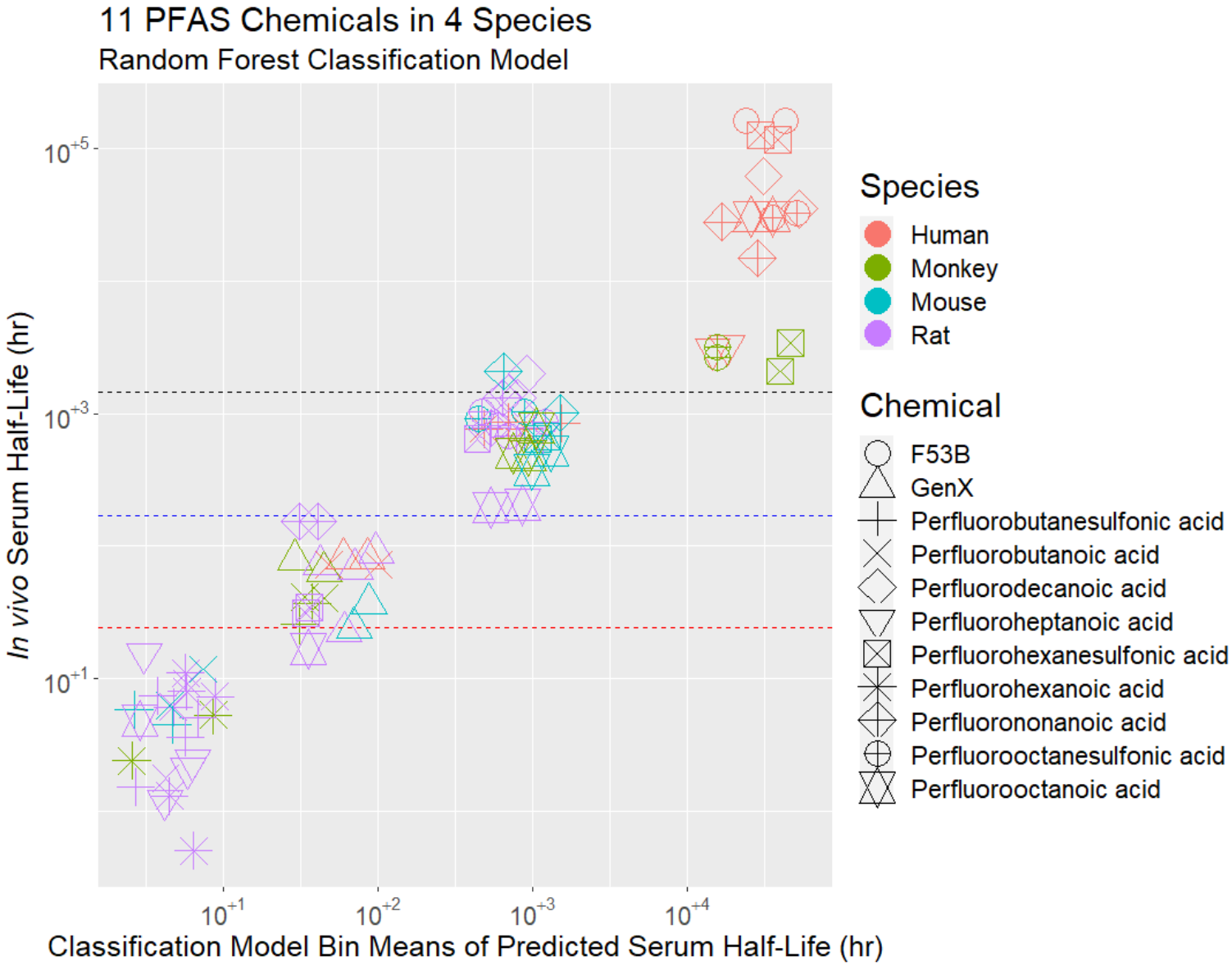
Descriptors

- Values of $t_{½}$ of the training data (y-axis) vs. classification predictions by the RF Classification model using 15 predictors
- Accuracy of 86.4% compared to the no information rate of 27% and y-randomization accuracy of 32.2 ± 13.3%



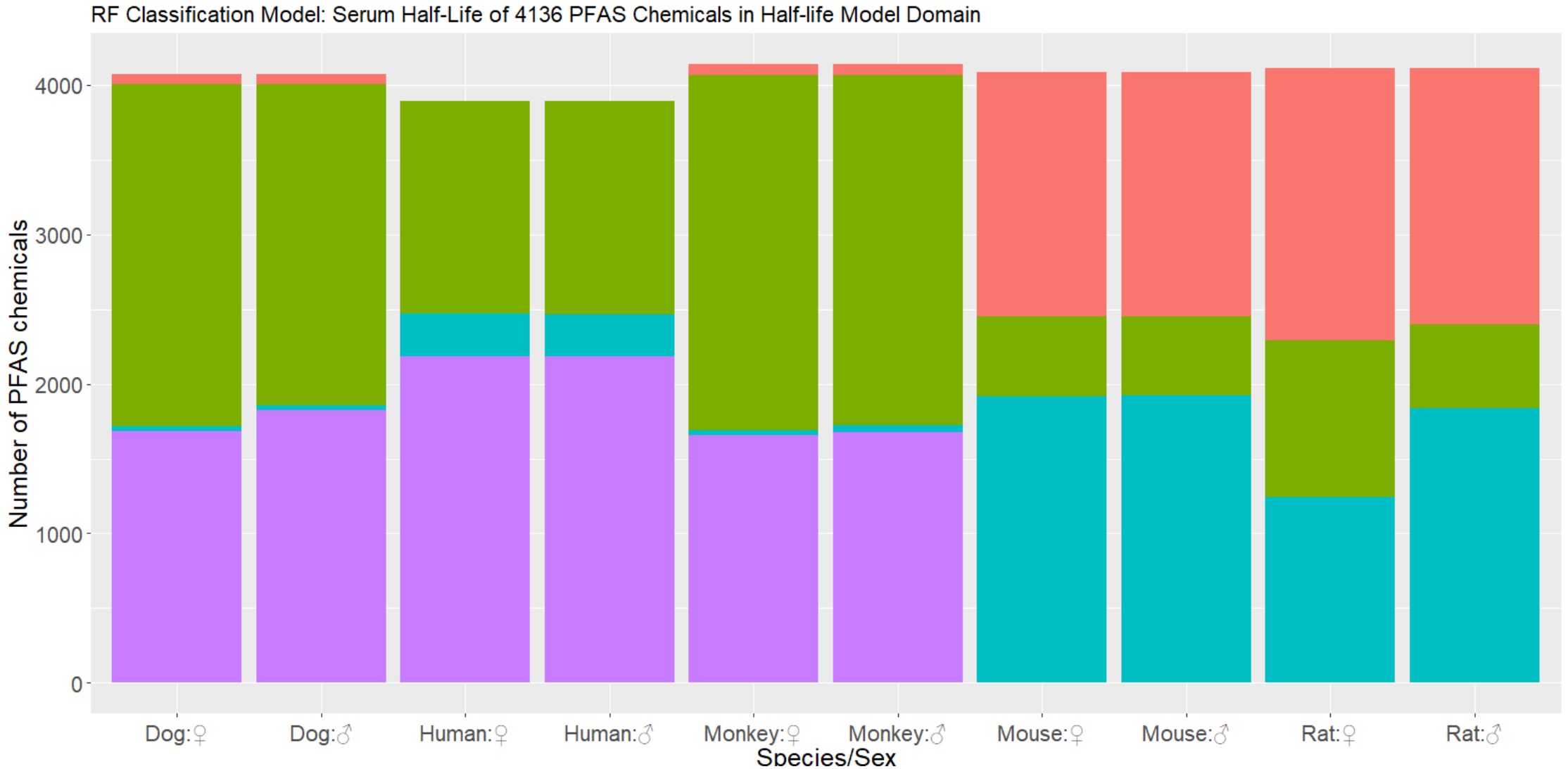
Descriptors

- Values of $t_{½}$ of the training data (*y*-axis) vs. classification predictions by the RF Classification model using 15 predictors.
- Note that observations have been jittered (that is, a small amount of random variation has been added) along the *x*-axis to increase readability.



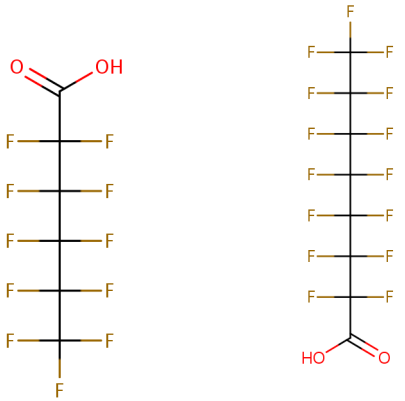
Predictions for PFAS in Model Domain

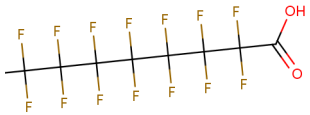
Out of 8163 PFAS on list https://comptox.epa.gov/dashboard/chemical-lists/pfasmaster

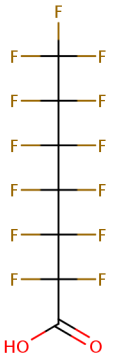


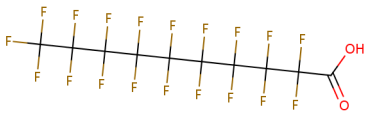
Domain of Applicability

- We calculated domain of applicability using method of Roy et al. (2015) based on descriptor properties
- However, the training set only included three classes: alkyl halides (9 chemicals), carboxylic acids and derivatives (GenX), and organic and sulfonic acids and derivatives (F-53B) (ClassyFire , Djoumbou Feunang et al., 2016)

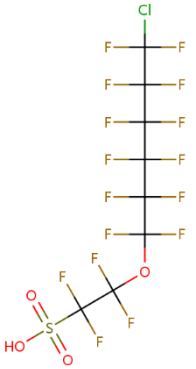




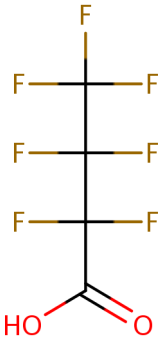


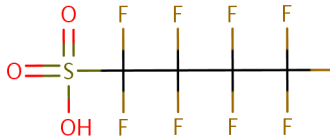


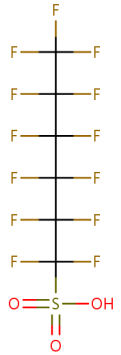
alkyl halides

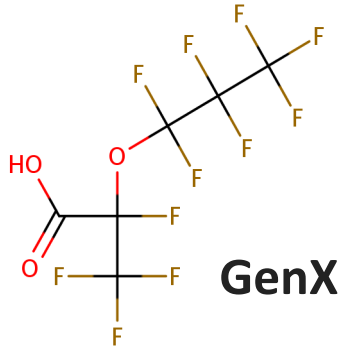


F-53B







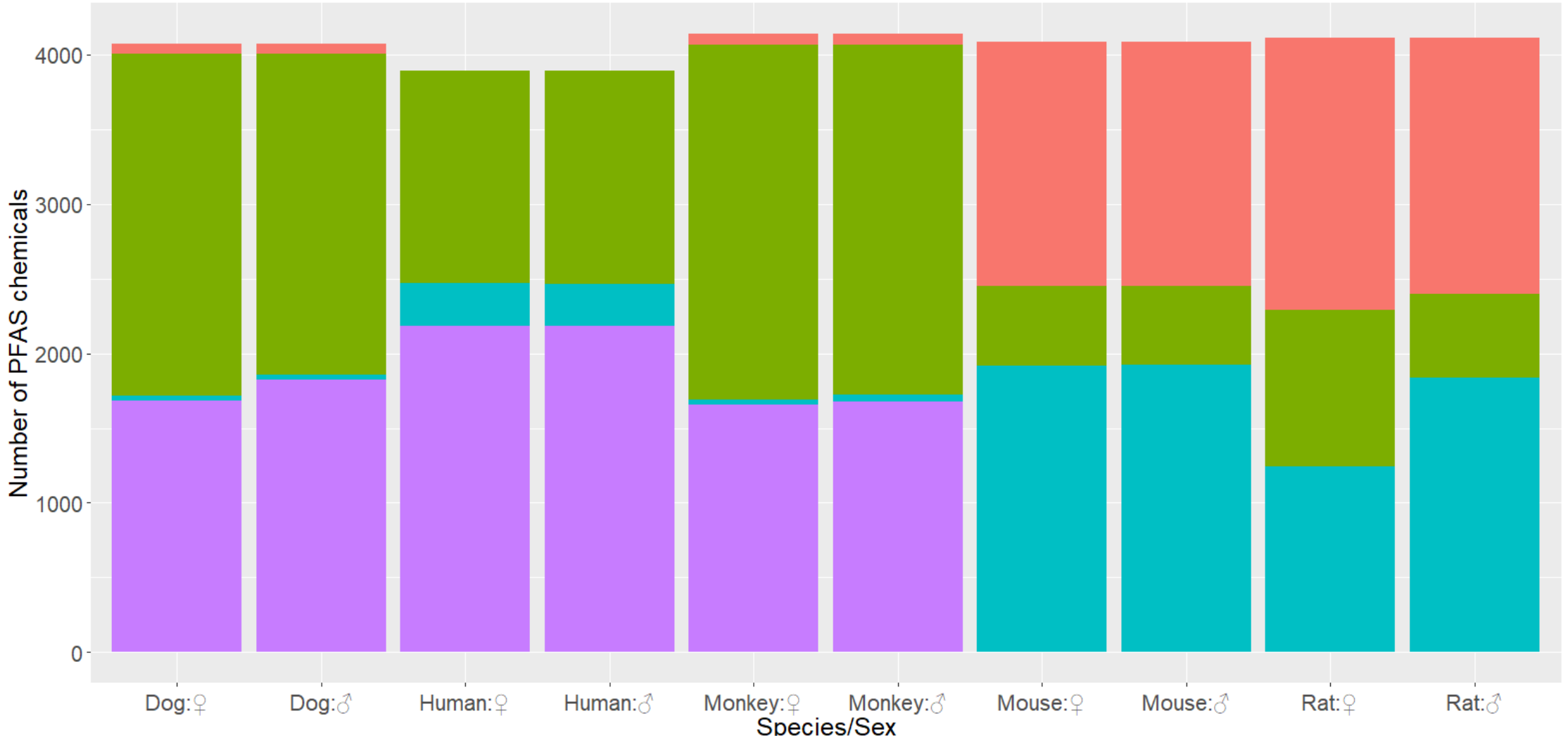


GenX

Predictions for PFAS in Model Domain

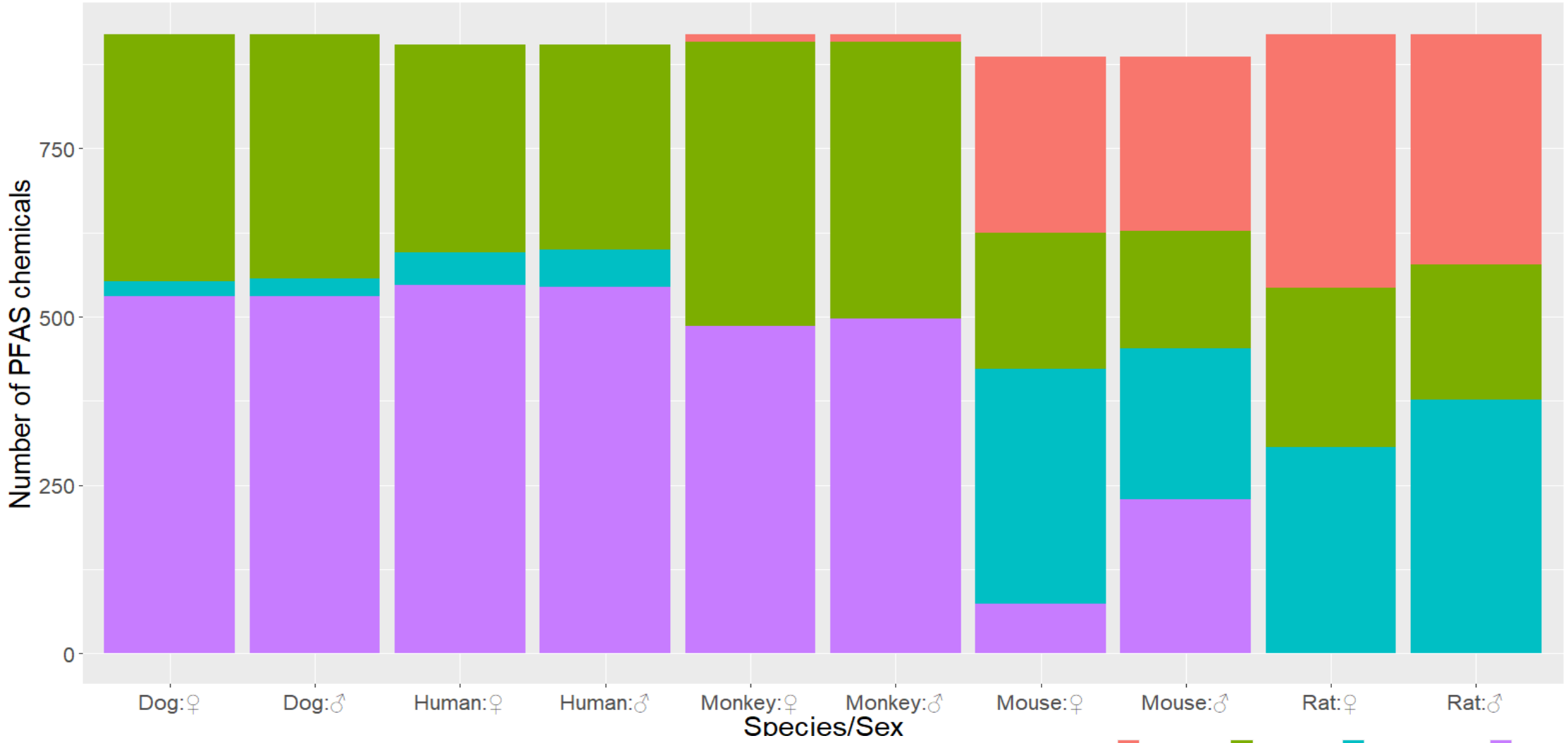
Out of 8163 PFAS on list https://comptox.epa.gov/dashboard/chemical-lists/pfasmaster

4136 PFAS in Roy et al. (2015) Applicability Domain (Without Consideration of Chemical Class)



Predictions for PFAS Matching Training Set Classes

921 PFAS both in Roy et al. (2015) Applicability Domain and Also Matching Chemical Classes from Training Set

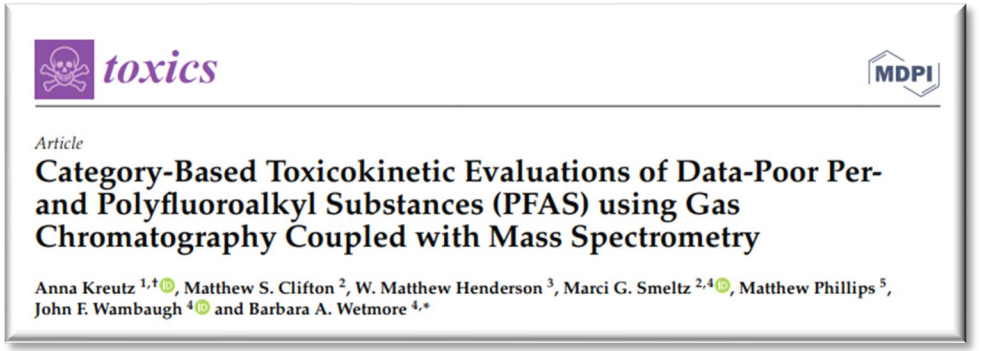


Limitations

- Model development (the training set) included most of the data available
  - Methods were cross-validated, but new data are needed for evaluation
- The training set consisted of only four species and 11 chemicals, and was
  - Training set dominated by alkyl halides
- The chemicals in need of half-life predictions were from classes that were much more diverse than the training set
  - TK behavior of other classes of PFAS could be influenced by factors not captured by the predictors included in the model
  - Uncertainties would be best characterized with additional data for model evaluation
  - Future *in vivo* TK studies in rodents might investigate PFAS that are predicted to have differing half-lives

PFAS in R Package "httk"

- Previously high throughput methods for TK (HTTK) have not been well-matched to PFAS
  - Inappropriateness of logP for prediction
  - Lack for transporters
- New PFAS-specific HTTK data (protein binding and metabolism) have been generated for ~120 PFAS
  - Kreutz et al. (2023)
  - Smeltz et al. (2023)
  - Crizer et al. (in preparation)
- New PFAS-specific correction for membrane affinity has been added
- New function *parameterize_1comppfas()* has been added to retrieve pre-computed Dawson et al. (2023) predictions
  - Simple one compartment model
  - Includes transporter-like effects



New version of "httk" will be released alongside submission of Crizer et al. manuscript

https://CRAN.R-project.org/package=httk

Implications

- A machine learning (ML) model for PFAS half-life allowstoxicokinetic (TK) predictions for ~900-4000 PFAS with no other data
  - We are relying on a empirical, one compartment TK model
  - ML predicts the half-life bin (very slow/slow/fast/very fast) based on species and PFAS, and we then use the median training data in each bin as the predicted half-life
  - Because an ML could not be built for volume of distribution (Vd), we choose to use the median dataset value of V_d = 0.201 L/kg for all PFAS and species
  - Model building scripts and predictions available at: https://github.com/USEPA/CompTox-PFASHalfLife
  - Upcoming version of R package "httk" will include Dawson et al. (2023) predictions
- Chemicals with longer $t_{½}$ may bioaccumulate and thus may warrant closer scrutiny
- The majority (56%) of PFAS were predicted to be in the longest $t_{½}$ category in humans

Summary

- Our work is based upon expert curation of publicly available organism half-life information for multiple PFAS compounds, in multiple species
- We have created a robust machine learning model of PFAS half-life in humans and other species, projecting half-life categories (very slow/slow/fast/very fast) for 4000+ PFAS depending on sex, dose route, and species
- We have incorporated an estimated applicability domain into the model
- Inclusion of kidney biology as species descriptors provided mechanistic intuition and a potential framework to consider future information on transporters.

Please send any questions to: wambaugh.john@epa.gov

EPA United States Environmental Protection Agency

Acknowledgements

- Lead author Dan Dawson (now at Integral Consulting)
- Co-authors Chris Lau, Prachi Pradeep (now at Bundesinstitut für Risikobewertung), Risa Sayre, Richard Judson, and Mike Tornero-Velez
- *In vitro* TK collaborators Barbara Wetmore, Anna Kreutz, Marci Smeltz, David Crizer, and Michael Devito
- Katherine Phillips and Kristin Isaacs for machine learning expertise and example figures
- Paul Kruse for help with ClassyFire
- Francesca Grisoni (Eindhoven University of Technology) and Rocky Goldsmith (Congruence Therapeutics) for helpful early conversations

United States Environmental Protection Agency

References

Read the Dawson et al. (2023) paper at: https://doi.org/10.3390/toxics11020098

- Archer, Kellie J., and Ryan V. Kimes. "Empirical characterization of random forest variable importance measures." Computational Statistics & Data Analysis 52.4 (2008): 2249-2260.
- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Arnot, J.A. et al. Screening level risk assessment model for chemical fate and effects in the environment. Environ. Sci. Technol. 2006, 40, 2316–2323.Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
- Buck, R.C.; et al. Chemistry, properties, and uses of commercial fluorinated surfactants. In Polyfluorinated Chemicals and Transformation Products; Springer: Berlin/Heidelberg, Germany, 2012; pp. 1–24.
- Chiu, W.A et al. Bayesian Estimation of Human Population Toxicokinetics of PFOA, PFOS, PFHxS, and PFNA from Studies of Contaminated Drinking Water. Environ. Health Perspectives. 2022, 130, 127001.
- Dawson, Daniel E., et al. "A Machine Learning Model to Estimate Toxicokinetic Half-Lives of Per-and Polyfluoro-Alkyl Substances (PFAS) in Multiple Species." Toxics 11.2 (2023): 98.
- DeWitt, J.C. Toxicological Effects of Perfluoroalkyl and Polyfluoroalkyl Substances; Springer: Berlin/Heidelberg, Germany, 2015.
- Djoumbou Feunang, Y. et al. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. J. Cheminform. 2016, 8, 61.
- Egeghy, P.P. and 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.
- Fenton, S.E. et al. 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.
- Ghahramani, Zoubin. "Probabilistic machine learning and artificial intelligence." Nature 521.7553 (2015): 452-459.
- Kreutz, Anna, et al. "Category-Based Toxicokinetic Evaluations of Data-Poor Per-and Polyfluoroalkyl Substances (PFAS) using Gas Chromatography Coupled with Mass Spectrometry." Toxics 11.5 (2023): 463.
- Lau, C. et al. J. Perfluoroalkyl Acids: A Review of Monitoring and Toxicological Findings. Toxicol. Sci. 2007, 99, 366–394.
- Lau, C. Perfluorinated Compounds. In Molecular, Clinical and Environmental Toxicology: Volume 3: Environmental Toxicology; Luch, A., Ed.; Springer: Basel, Switzerland, 2012; pp. 47–86.
- Lau, C. Perfluorinated compounds: An overview. In Toxicological Effects of Perfluoroalkyl and Polyfluoroalkyl Substances; Humana Press, Cham, Switzerland, 2015; pp. 1–21.
- 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.
- Pizzurro, D.M. et al. Interspecies differences in perfluoroalkyl substances (PFAS) toxicokinetics and application to health-based criteria. Regul. Toxicol. Pharmacol. 2019, 106, 239–250.
- Rao, N.S. and Baker, B.E. Textile finishes and fluorosurfactants. In Organofluorine Chemistry; Springer: Berlin/Heidelberg, Germany, 1994; pp. 321–338.
- Roy, K. et al. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. 2015, 145, 22–29.
- Russell, M.H. et al. Elimination kinetics of perfluorohexanoic acid in humans and comparison with mouse, rat and monkey. Chemosphere 2013, 93, 2419–2425.
- Schwidetzky, Ralph, et al. "Ice Nucleation Activity of Perfluorinated Organic Acids." The Journal of Physical Chemistry Letters 12.13 (2021): 3431-3435.
- Smeltz, Marci, John F. Wambaugh, and Barbara A. Wetmore. "Plasma Protein Binding Evaluations of Per-and Polyfluoroalkyl Substances for Category-Based Toxicokinetic Assessment." Chemical Research in Toxicology (2023).
- Tropsha, Alexander, and Alexander Golbraikh. "Predictive QSAR modeling workflow, model applicability domains, and virtual screening." Current pharmaceutical design 13.34 (2007): 3494-3504.
- Wambaugh, J.F et al. Dosimetric anchoring of in vivo and in vitro studies for perfluorooctanoate and perfluorooctanesulfonate. Toxicol. Sci. 2013, 136, 308–327.

Please send any questions to: wambaugh.john@epa.gov