A quantitative structure-activity relationship (QSAR) model to estimate serum half-lives of per- and polyfluoroalkyl substances (PFAS) in multiple species

Daniel Dawson[1], Chris Lau[1], Prachi Pradeep[2], Richard Judson[1], Rogelio Tornero-Velez[1], and John F. Wambaugh[1]

1. U.S. Environmental Protection Agency, Office of Research and Development
2. Oak Ridge Institutes for Science and Education

Daniel Dawson | dawson.daniel@epa.gov | 919-541-5662

United States Environmental Protection Agency

EPA at 50 — Progress for a Stronger Future

Trying to characterize the toxicokinetics of PFAS

- What are per- and polyfluoroalkyl substances (PFAS)?
  - Highly diverse group of anthropogenic molecules that substitute fluorine for hydrogen along carbon backbones
  - 1223 PFAS are currently in the TSCA inventory
  - 602 PFAS are known to be commercially active
  - Widely distributed in environmental substrates
  - Generally non-biodegradable
  - Often have relatively long half-lives in body tissues
- What do we want to do?
  - Describe and quantify toxicokinetics (TK) of PFAS
  - Extrapolate TK parameters from data-rich to data-limited substances
  - Prioritize chemicals for in-depth assessment based on predicted half-life and predicted exposure
- What's the problem?
  - TK parameters vary widely and non-allometrically across species
  - Some TK parameters vary by sex within species
  - Read-across methods and cross-species extrapolations are unreliable
- A potential solution: Machine Learning
  - Use data from *in vivo* testing of PFAS across multiple species and chemicals
  - Integrate multiple chemical and physiological characteristics
  - Produce fit-for-purpose predictions of TK parameters
  - Contribute to prioritization of chemicals for more in-depth evaluation

Putting it together: datasets and model assembly

- *In vivo* serum half-life (SHL) data
  - Curated dataset of values published in peer-reviewed literature[1-20]
  - 66 data points
    - 11 PFAS chemicals
    - 4 species (Human, Monkey, Rat, Mouse), sex-specific values when available
- Model predictors: capturing PFAS chemical/structural diversity
  - Considered 118 predictors in 6 general categories

❖ **Physio-chemical characteristics (OPERA)[21,22]**
❖ **Species-specific kidney characteristics[23,24,25]**
❖ **Similarity to endogenous compounds[26]**
❖ **Protein binding affinities[27,28]**
❖ **Critical micelle concentrations[29]**
❖ **Inclusion of ether bond**

- Predictors pruned in an iterative process
  1. Identified pairs of predictors with Spearman's ρ > 0.9
  2. Retained member of pair with better correlation with SHL
  - **29 predictors remaining for model construction**
- Machine learning method: Random Forest (RF)
  - Investigated two approaches with *caret* [30,31] package of R
    - RF Classification model with SHL binned into 3 bins
      - Fast (<1 day), Moderate (1 day – 1 year), and Slow (>1 year)
    - RF Regression Model
  - Both models fit with 10-fold cross validation, replicated 10 times

Evaluation of machine learning models

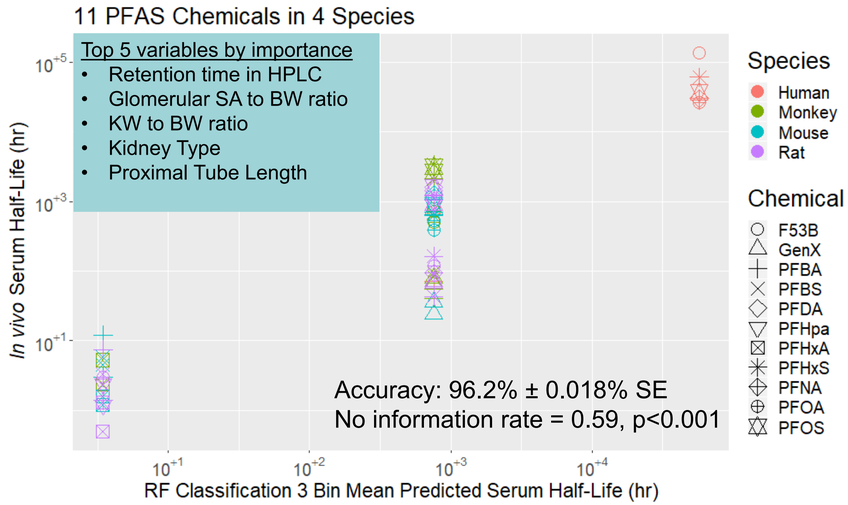

Figure 1. RF *classification* model results. Predicted (fast, moderate, or slow) vs *in vivo* SHL, shown by chemical and species.

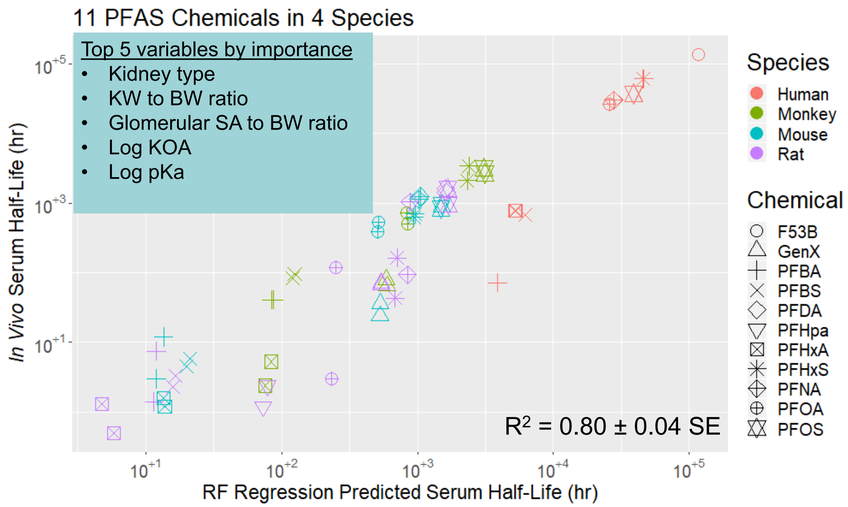

Figure 2. RF *regression* model results. Predicted vs *in vivo* SHL, shown by chemical and species.

Application to the DSSTox PFAS dataset

- Classification model used to predict SHL for 6648 PFAS and PFAS-like chemicals listed in the USEPA Distributed Structure-Searchable Toxicity (DSSTox) database

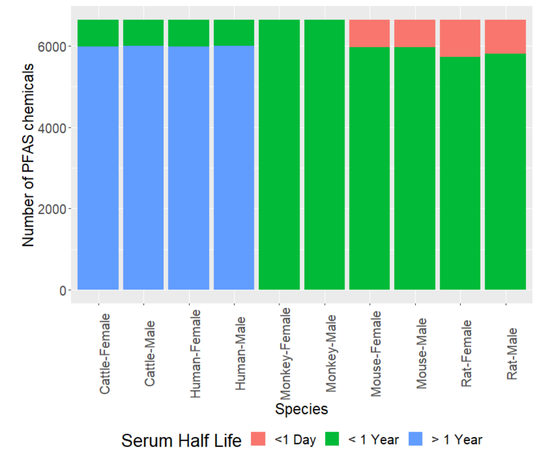

Figure 3. Classification of 6648 DSSTox chemicals by species (5) and sex into fast, moderate, or slow SHL categories. Based on kidney characteristics[23], model can potentially be extended to other species.

Conclusions and future directions

- Both classification (Fig 1) and regression (Fig 2) models are promising initial efforts to estimate an important TK parameter of PFAS chemicals using machine learning
- Though less precise, the classification model is highly accurate
- Application of classification model to DSSTox database suggests most PFAS may have long SHLs (> 1 year) in humans (Fig 3)
- Limitations
  - **Models are likely overfit (29 parameters with 66 obs)**
  - Regression model tends to overestimate length of SHL for quickly clearing chemicals (Fig 2)
    - Additional training data for such chemicals needed
  - Lack of independent test set due to data paucity
- Future work
  - Expansion of models with additional SHL data
    - Chemicals
    - Species
  - Development of evaluation sets of *in vitro* and *in vivo* data
  - Incorporation of predictions into TK modeling effort

Innovative Research for a Sustainable Future

*Numbered references available upon request