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

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 chemicals are currently in EPA's TSCA inventory
  - 602 PFAS chemicals are known to be commercially active
  - Generally non-biodegradable, and widely distributed in environmental media, humans and wildlife
- What do we want to do?
  - Characterize toxicokinetics (TK) of PFAS and extrapolate from data-rich to data-limited substances
  - Prioritize PFAS for in-depth assessment based on predicted half-life and 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
  - 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
  - Expert curated set of peer-reviewed literature[1-29]
  - Mean SHL values assembled through hierarchical process based on data availability
    - 89 datapoints
    - 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)[30,31]**
❖ **Species-specific kidney characteristics[32,33,34]**
❖ **Similarity to endogenous compounds[35]**
❖ **Protein binding affinities[36,37]**
❖ **Critical micelle concentrations[38]**
❖ **Inclusion of ether bond**

- Pruned down to **29 predictors** for model construction
- Machine learning method: Random Forest (RF)
  - RF Classification model with SHL segregated into 4 bins
    - Very Fast (<1 day), Fast (1 day – 1 week), Slow (1 week – 2 months) and Slow (>2 months)
  - 10-fold cross validation, replicated 10 times

Evaluation of QSAR Model

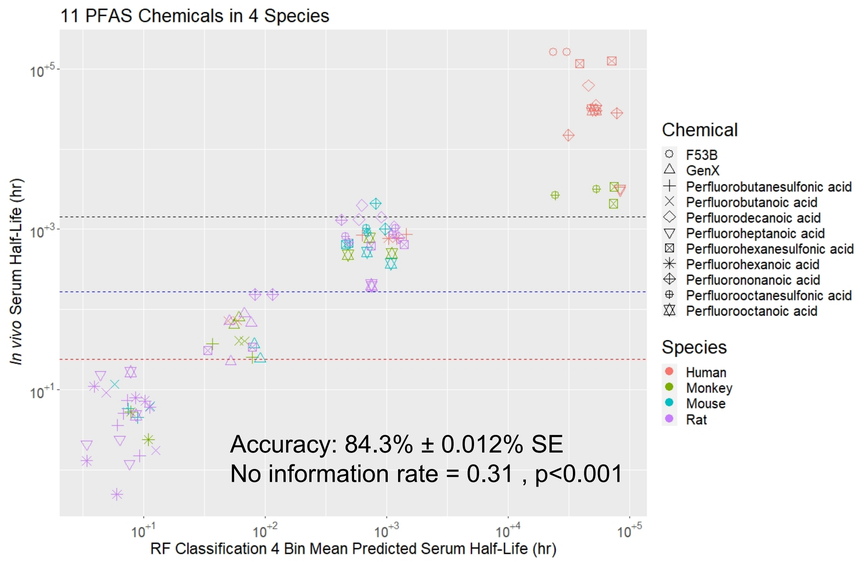

Figure 1. RF *classification* model results. Predicted (very fast, fast, moderate, or slow) vs *in vivo* SHL, shown by chemical and species.

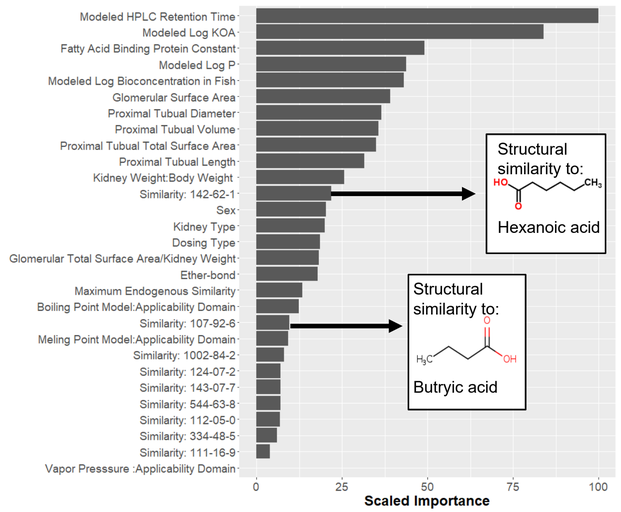

Figure 2. Scaled importance (change in accuracy) for all 29 model parameters. "Modeled" parameters are OpERA QSAR model predictions or applicability domains. Similarity values are based on compound structural similarity.

Application to the DSSTox PFAS dataset

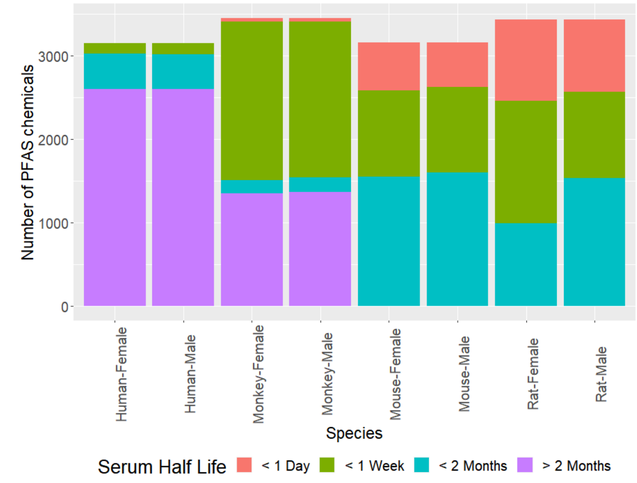

Figure 3. Classification of 3455/6648 chemicals of DSSTox chemicals by species (4) and sex into very fast, fast, moderate, or slow SHL categories. Chemicals outside of applicability domain[9] (3193) excluded from prediction.

Conclusions and future directions

- **Summary**
  - The model is less precise than a concurrently developed regression model but is still highly accurate
  - Top Predictors: HPLC retention time and Log KOA
  - Secondary: FABP constants > correlates of lipophilicity (Log P and Log BCF) > kidney characteristics > endogenous compound similarity
  - Application of model to DSSTox database suggests many PFAS may have long SHLs (> 2 months) in humans (Fig 3)
- **Limitations**
  - Lack of independent test set due to data paucity
  - Model is likely overfit (29 parameters with 89 obs)
  - However, RF methodology (collection of trees created over random subsets of data and parameters) is generally robust
- **Future work**
  - Refitting of models with additional SHL data and parameters
  - 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