https://orcid.org/0000-0001-5844-8754

EPA United States Environmental Protection Agency

Development of models to predict physicochemical properties of PFAS

Todd Martin[1], Gabriel Sinclair[2], Christian Ramsland[2] and Antony Williams[1]*

[1] US EPA/ORD/CCTE
[2] ORAU

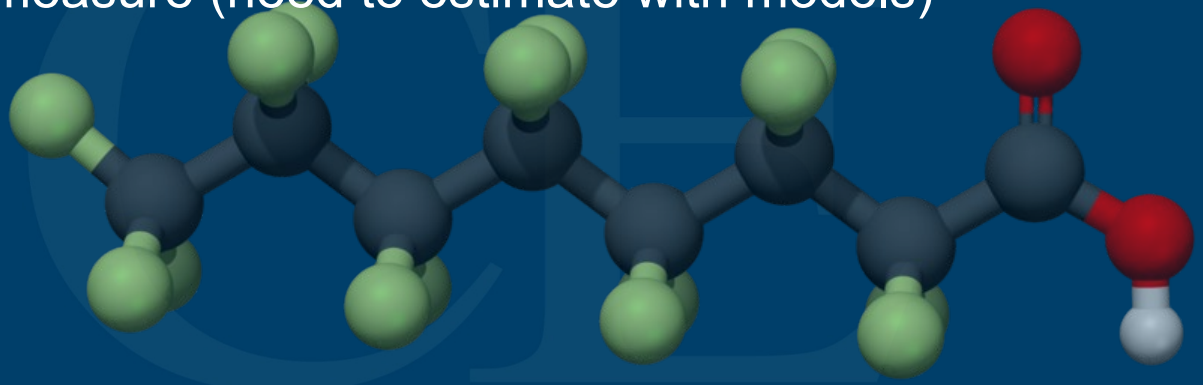
The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA.

August 19, 2021

Per- & Polyfluoroalkyl Substances

A structurally-diverse family of fluorinated chemicals with…

- Growing environmental concern
  - Detection in organisms
  - Potential toxic effects
  - Evidence of bioaccumulation/bioconcentration
  - Some PFAS are persistent and mobile in the environment and are difficult to remediate
- Widely varying physicochemical properties
  - Used characterize fate and transport
  - Hard to measure (need to estimate with models)



Case Study: Henry's Law Constant

- Henry's law states that the amount of dissolved gas in a liquid is proportional to its partial pressure above the liquid:

$$HLC = p / c_a \text{ in atm-m}^3\text{/mol}$$

  - Modeled values are –log10 (HLC atm-m^3/mol)
- Selected since **there are** a reasonable number of records for structure curation
- Goal: compare ability of global and local models to predict HLC of PFAS

Workflow

- Pull experimental records from database. Keep records if:
  - Have valid point estimate, property value, and units
  - Have valid experimental conditions (e.g. 20°C < T < 30 °C)
- Map records to DSSTox records based on name, CAS, smiles.
  - Yields DSSTOX ID numbers and curated smiles
- Omit based on smiles structure if:
  - Substance type = Mineral/Composite, Mixture/Formulation, Polymer
  - Substance has multiple organic fragments
  - Omit salts (depending on the endpoint)
- Convert smiles to QSAR Ready SMILES (standardizes tautomers)

Workflow, cont.

- Flatten QSAR records using first part of InChIKey (or canonical smiles)
  - Assumes property value is only a function of 2d connectivity
  - Median value is used for continuous endpoint
- Calculate ~800 T.E.S.T. (Toxicity Estimation Software Tool) descriptors from QSAR Ready SMILES (via web service)
- Create overall set from ID, property value, and descriptors
- Perform outlier detection (check original data)
- Find a “representative splitting” into training and prediction sets
- Create training and prediction set csv files from overall set and representative splitting

Sources of HLC data

Source	Description	#Records HLC	#Distinct CAS HLC
eChemPortal	REACH data	614	195
OChem	QSAR platform	2113	1349
OPERA	QSAR tool	686	686
PubChem	Website	535	535
Sander	HLC Website	6818	1950
ICF	Lit search for PFAS data	71	16

Data Set Creation

- 11,213 HLC total records gathered
- 10,261 records after filtering for appropriate experimental conditions & data
- 2,170 records after mapping in DSSTox
  - Mapping in DSSTOX only half completed
- 1,032 records after merging & filtering structures

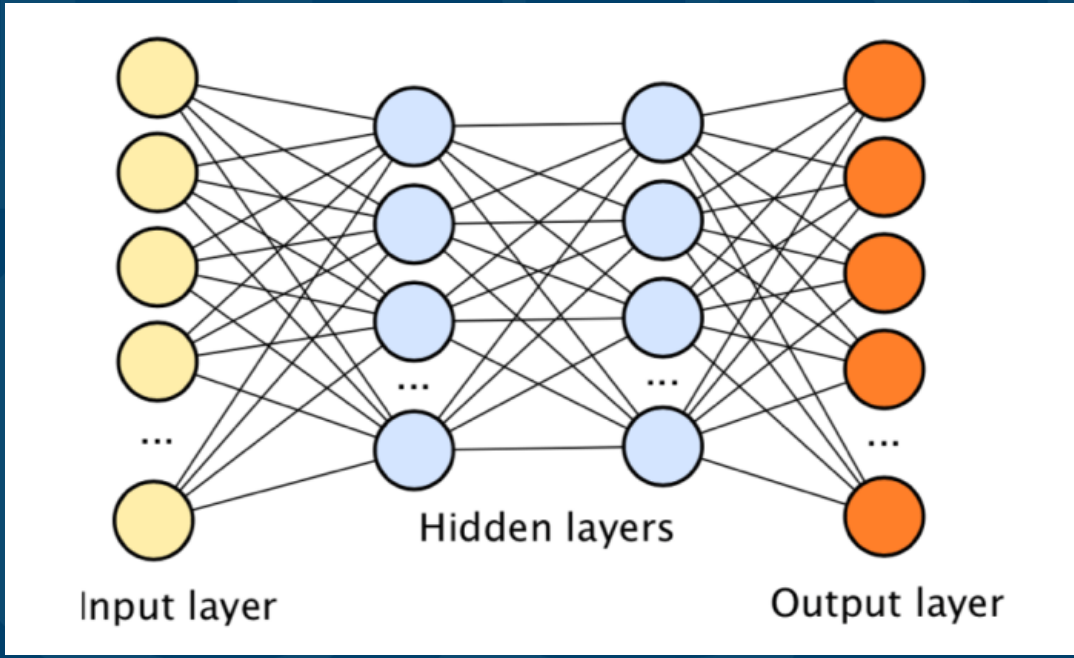
EPA United States Environmental Protection Agency

QSAR Methods

- **Python based QSAR methods**
  - RF - Random Forest
  - SVM – Support Vector Machine
  - DNN – Deep Neural Network
  - XGBoost – eXtreme Gradient Boosting
  - Consensus – average of above methods
- Easily implementable as web services for both model building and model prediction

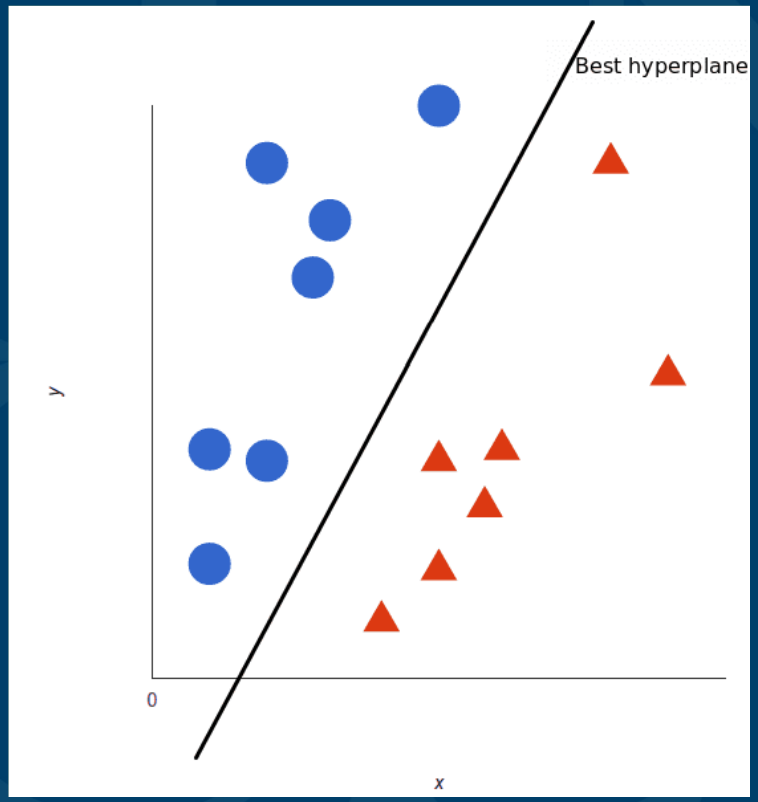
Deep Neural Network

- Deep learning approach using the Keras python package with TensorFlow backend
- Feedforward network with three hidden layer implementation
- Trained by adjusting the weights and biases of network nodes whenever a compound is classified correctly or incorrectly

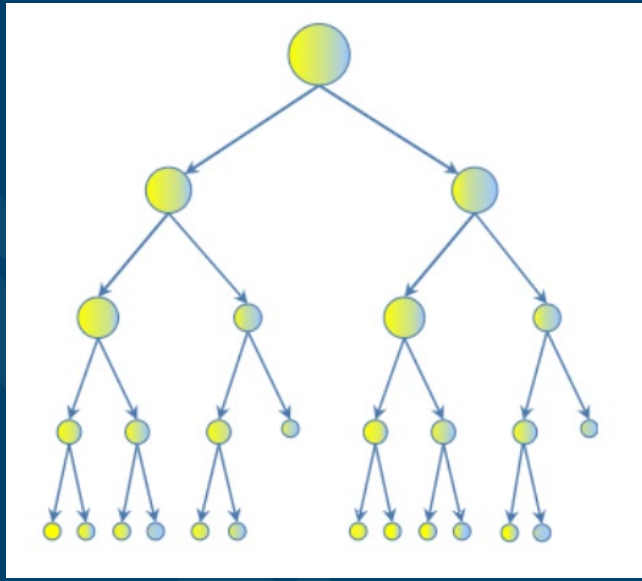


Support Vector Machines (SVM)

- SVM relies on the construction of hyperplanes between data belonging to two different classes.
- For continuous endpoints, SVR (support vector regression) is used

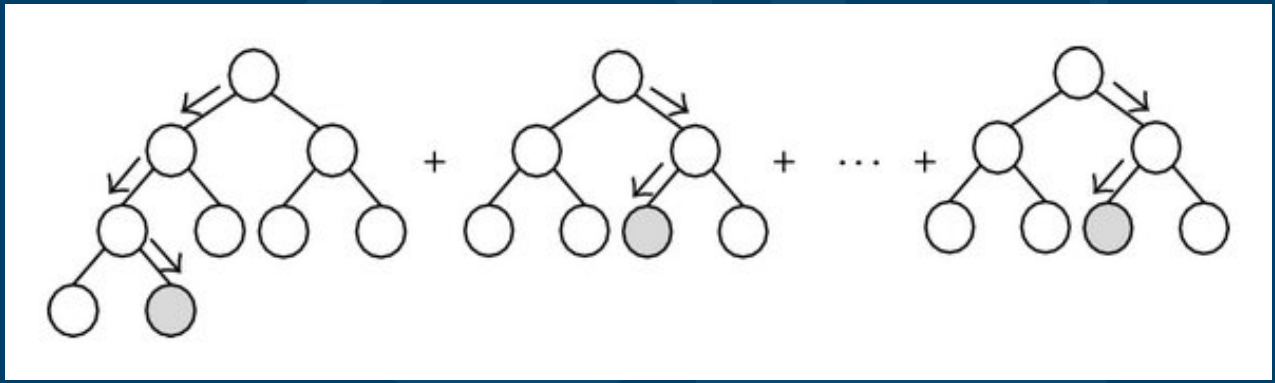


Random Forest and XGBoost



- Random forest is an ensemble decision tree approach to classification or regression

- XGBoost is decision-tree based method that adds new models to correct for mistakes made in previous models:



External validation

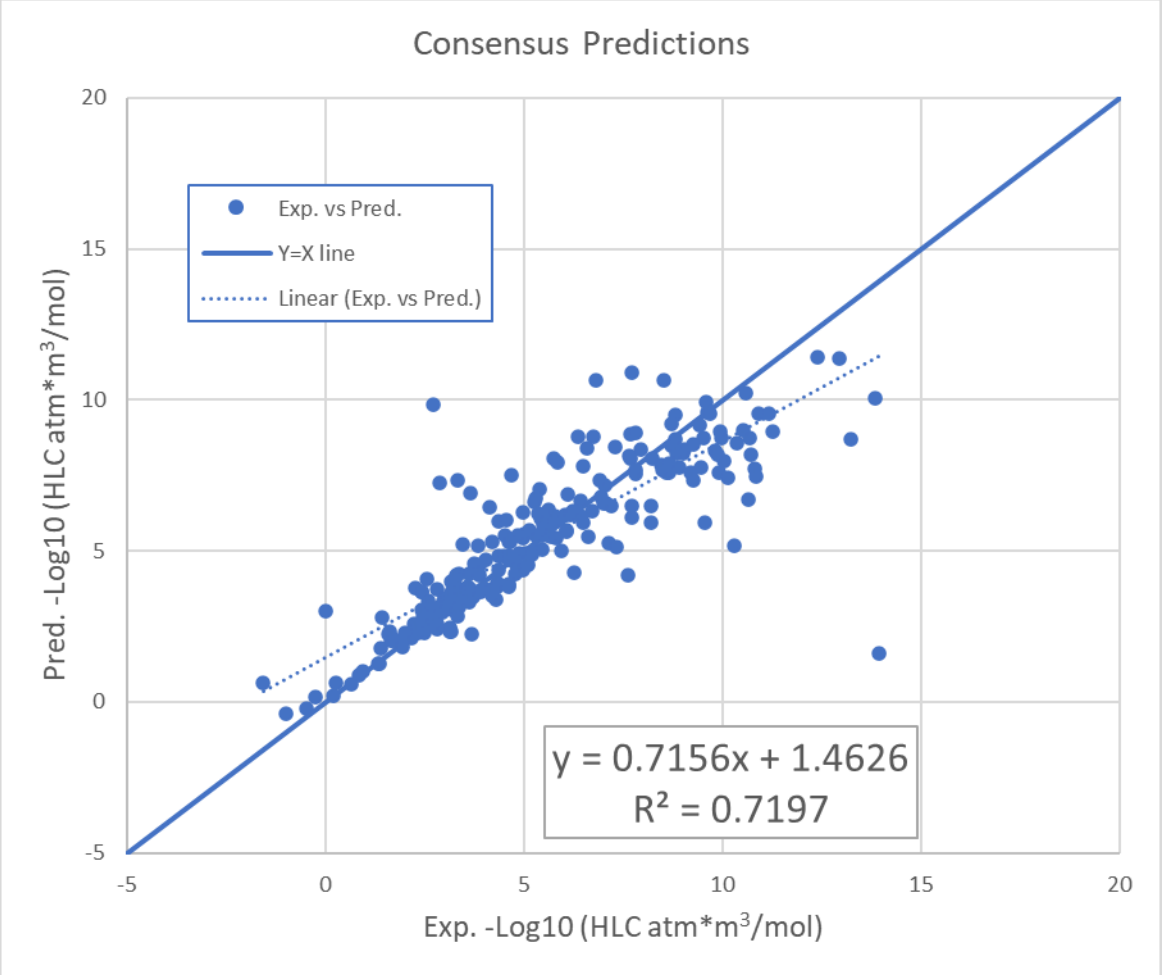
- Overall set was split into training and prediction set using random splitting
- Subsets of these sets were evaluated:

Training	Prediction
All (n=824)	All (n=207)
All (n=824)	Only PFAS (n=4)
All but PFAS (n=808)	Only PFAS (n=4)
Only PFAS (n=16)	Only PFAS (n=4)

Prediction results for global model (T=All, P=All)

Method	R^2	MAE
XGBoost 1.0	0.70	1.00
SVM 1.1	0.70	0.95
RF 1.1	0.70	1.03
DNN 1.8	0.65	1.00
Consensus 1.0	0.72*	0.93

*R^2 = 0.78 if outlier removed



Outlier

ID	Formula	exp	pred
DTXCID3043605	$C_{36}H_{74}$	13.94	1.619
DTXSID2060882	$C_{25}H_{52}$	-2.57	N/A

Home

Henry's Law Constants

Notes

References

Errata

Contact, Impressum, Acknowledgements

When referring to the compilation of Henry's Law Constants, please cite this publication:

R. Sander: *Compilation of Henry's law constants (version 4.0) for water as solvent*, Atmos. Chem. Phys., 15, 4399-4981 (2015), doi:10.5194/acp-15-4399-2015

Henry's Law Constants → Hydrocarbons (C, H) → Alkanes → hexatriacontane

FORMULA: $C_{36}H_{74}$
CAS RN: 630-06-8
STRUCTURE (FROM NIST):

InChIKey: YDLYQMBWCWFRAI-UHFFFAOYSA-N

H_{cp} [mol/(m³Pa)]	d ln H_{cp} / d (1/T) [K]	Reference	Type	Notes
8.6×10^{8}		Abraham 1984	V	

References

- Abraham, M. H.: Thermodynamics of solution of homologous series of solutes in water, J. Chem. Soc. Faraday Trans. 1, 80, 153-181, doi:10.1039/F19848000153, 1984.

Type

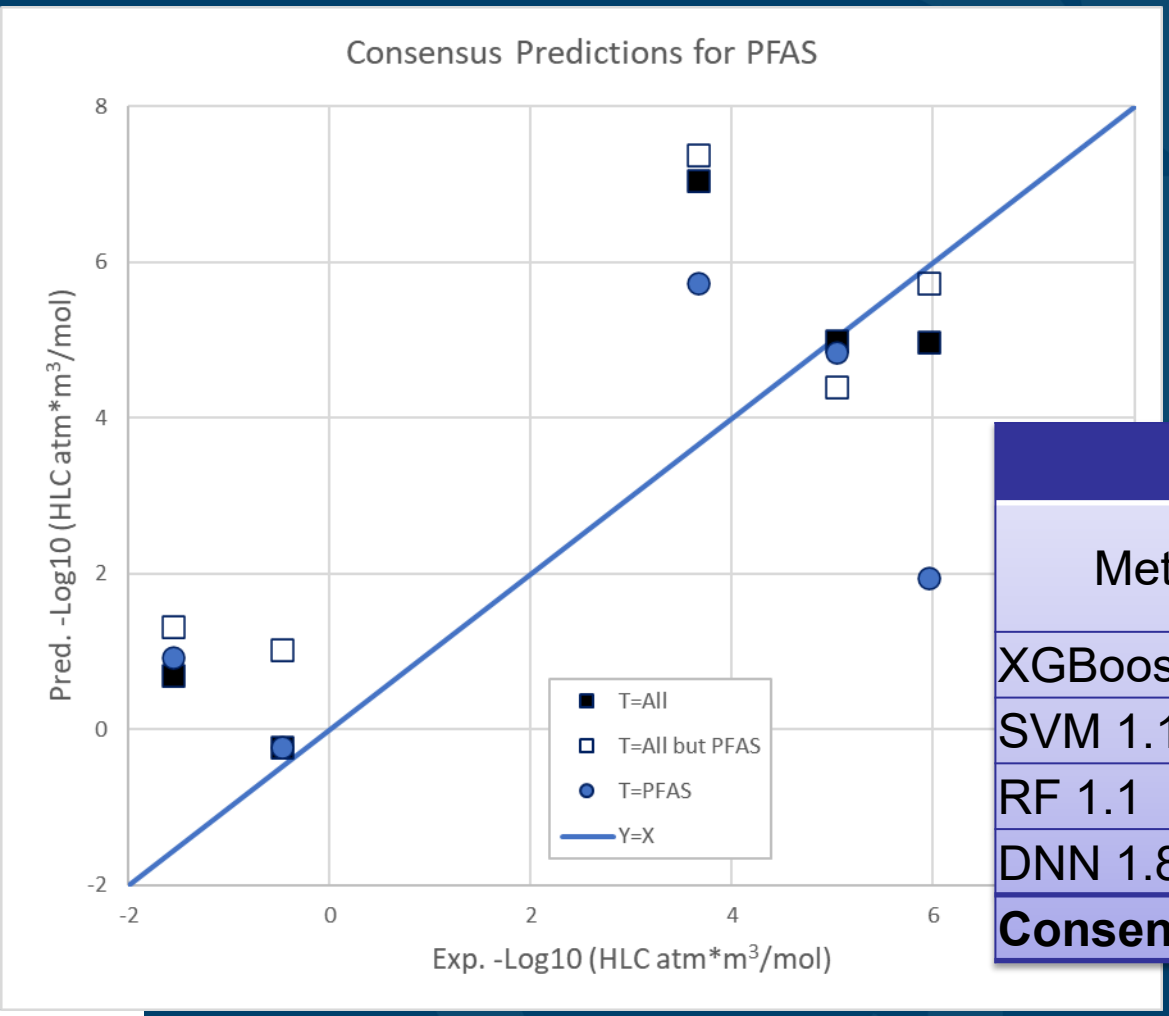
Table entries are sorted according to reliability of the data, listing the most reliable type first: L) literature review, M) measured, V) VP/AS = vapor pressure/aqueous solubility, R) recalculation, T) thermodynamical calculation, X) original paper not available, C) citation, Q) QSPR, E) estimate, ?) unknown, W) wrong. See Section 3.1 of Sander (2015) for further details.

Notes

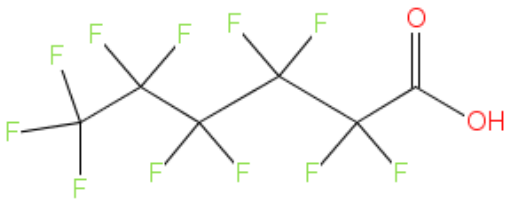
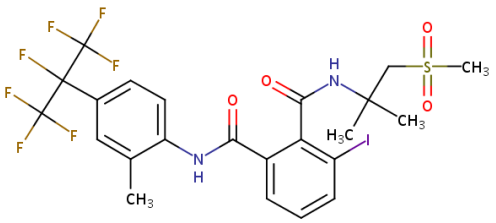
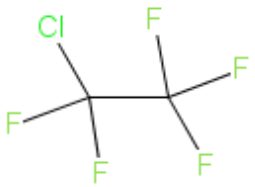
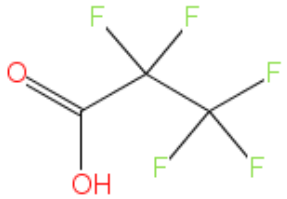
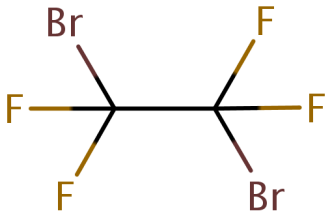
The numbers of the notes are the same as in Sander (2015). References cited in the notes can be found here.

➢http://satellite.mpic.de/henry/casrn/630-06-8

PFAS prediction results for global model vs local model



MAE in log units			
Method	T=All	T=All but PFAS	T=PFAS only
XGBoost 1.0	1.25	1.91	2.19
SVM 1.1	1.20	1.51	1.45
RF 1.1	1.93	3.38	2.03
DNN 1.8	1.42	1.93	1.79
Consensus 1.0	1.38	1.79	1.80

Chemical	Exp HLC	Pred Global	Pred Local
	5.04	4.98	4.85
	3.66	7.05	5.74
	-0.48	-0.24	-0.23
	5.95	4.97	1.95
	-1.56	0.69	0.93

Future work

- Add applicability domain measures
- Eliminate chemicals with high standard deviation for property value
- Finish structure curation and redo analysis

EPA
United States
Environmental Protection
Agency

Questions???

The views expressed in this presentation are those of the author and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency