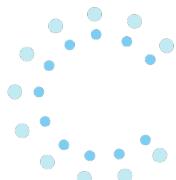




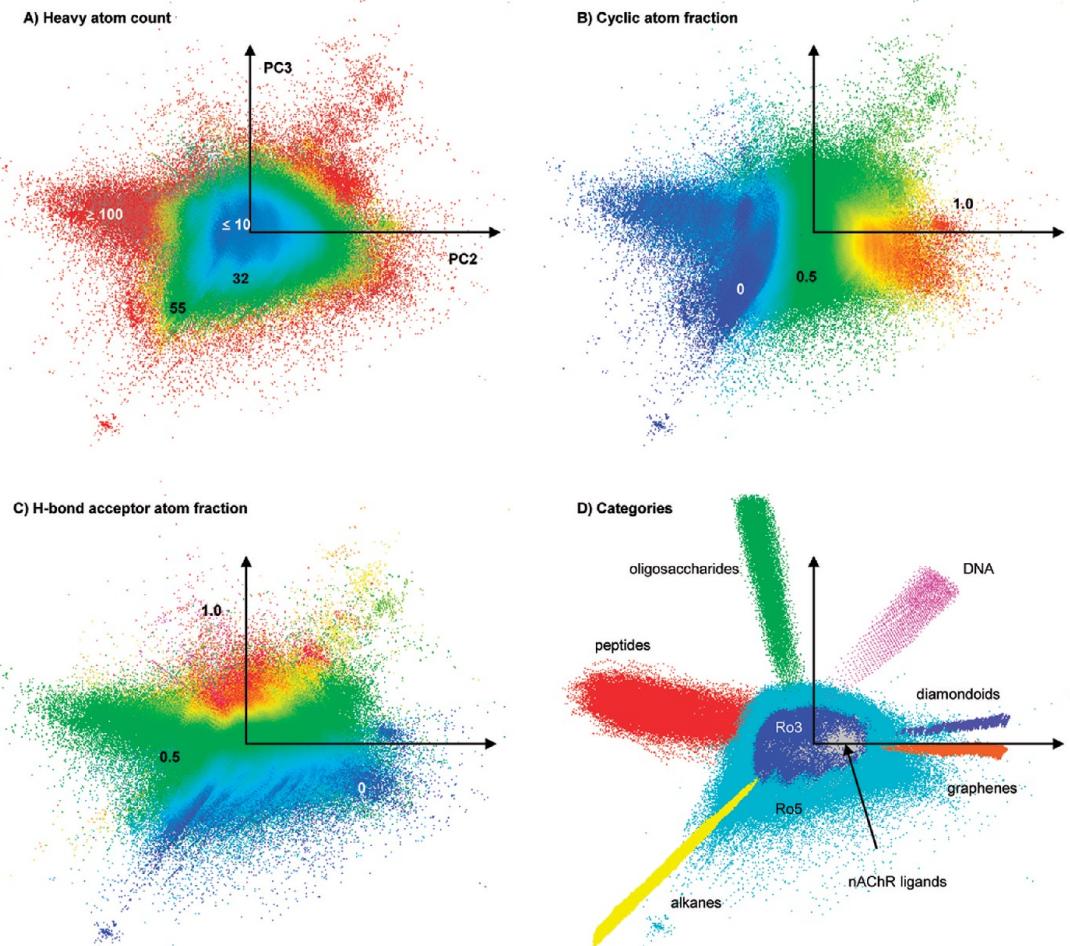
From descriptors to intrinsic fate and toxicity of chemicals: an alternative approach to chemical prioritization

Saer Samanipour, J. W. O'Brien, M. J. Reid, K. V. Thomas, and A. Praetorius



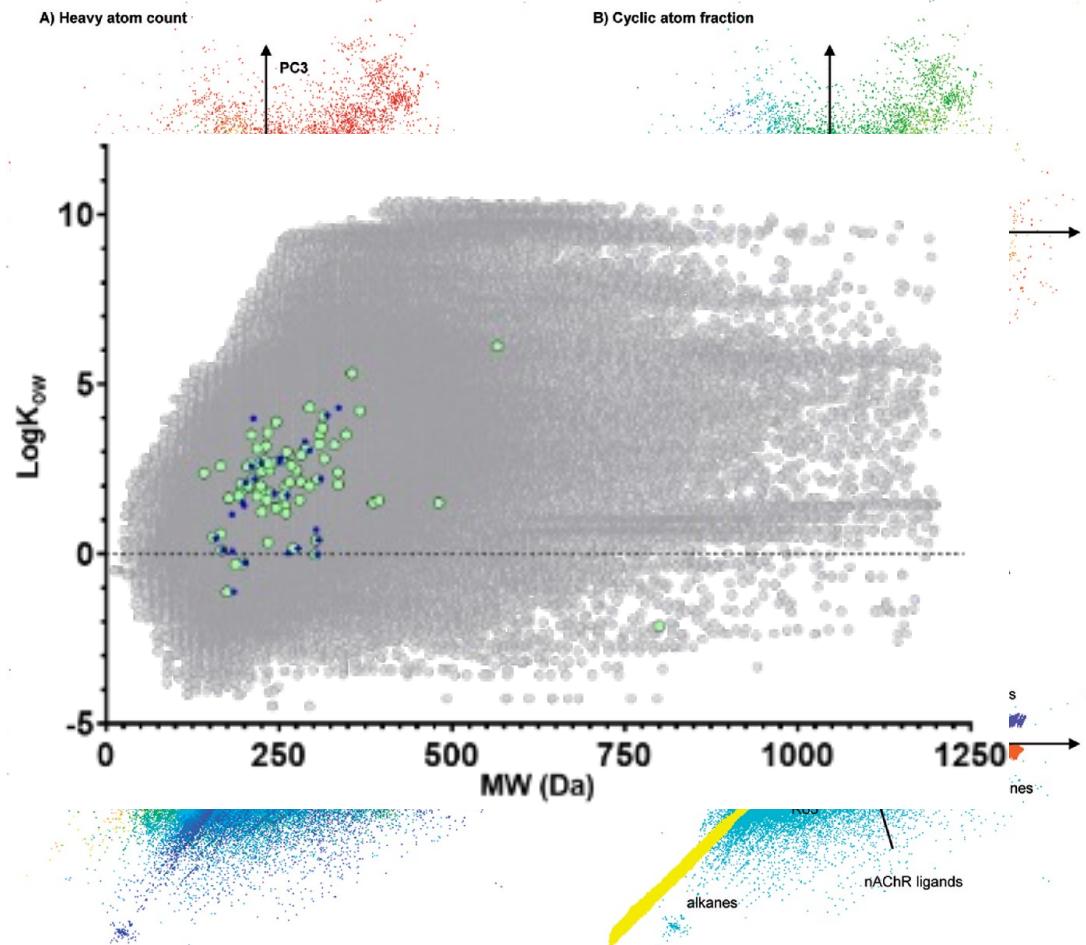


As environmental chemists, we aim at assessing and minimizing the health associated impact of chemical exposure.

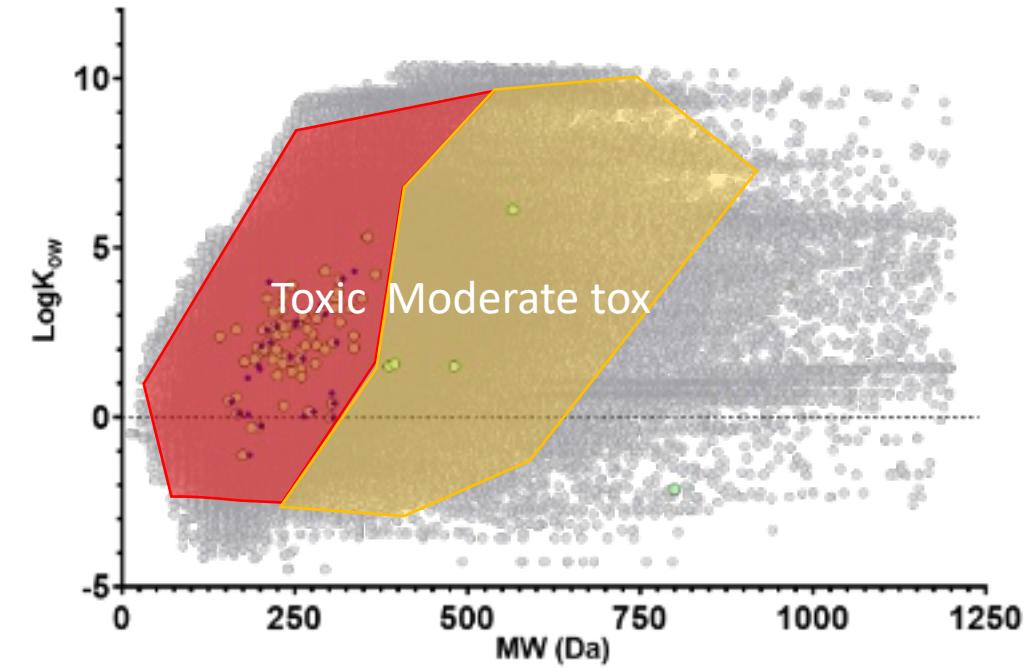
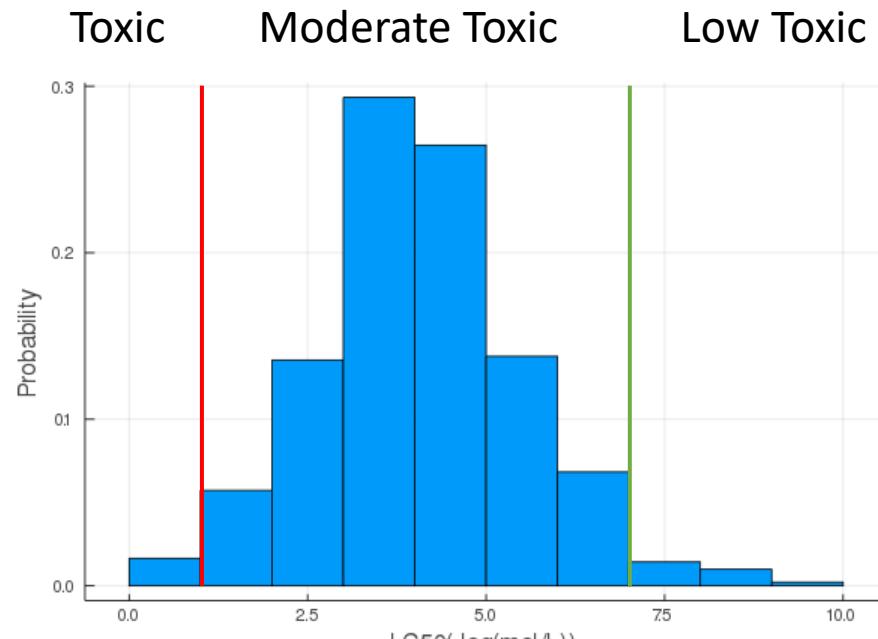


- There are $> 10^{60}$ possible structures with $M_w < 500$ Da.
- The physiochemical property range is too wide.
- Technologically we cannot cover this space.





- There are more than 800k known chemicals that are actively released into the environment.
- The transformation products are ignored.
- All natural chemicals were excluded.
- We have methods for less than 1%.

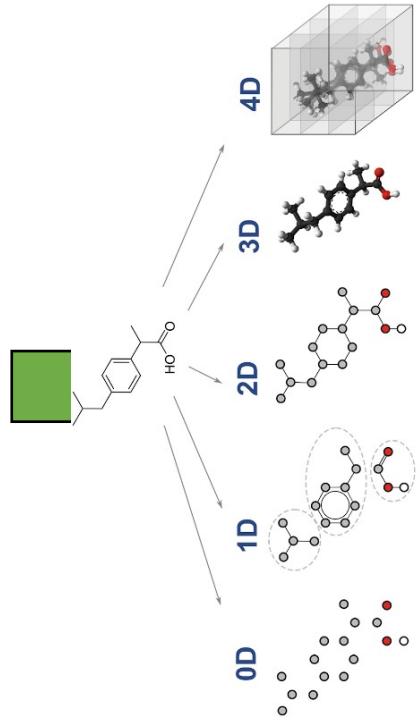


The combination of the measurements and expert knowledge will result in a set of criteria for chemical prioritization.

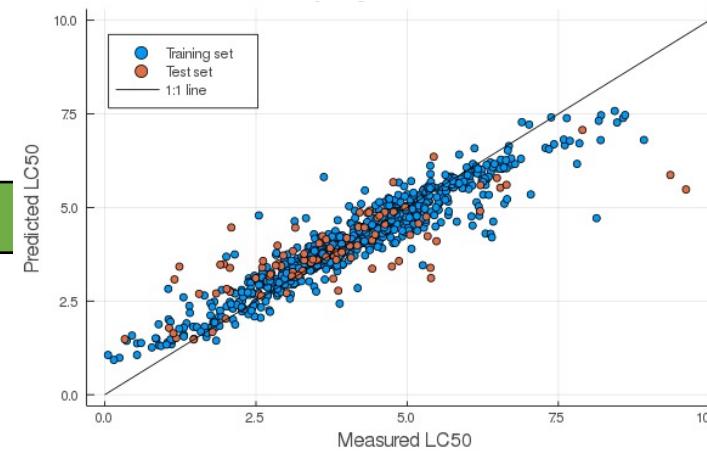




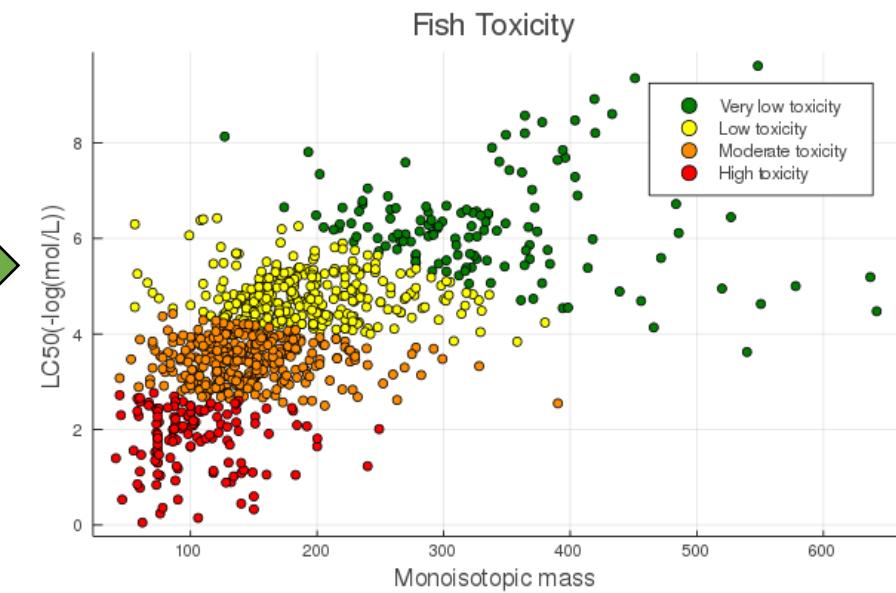
Molecular Descriptors



QSAR



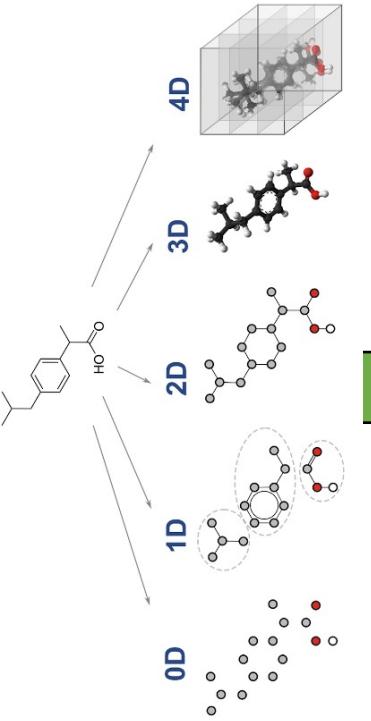
Tox Categories



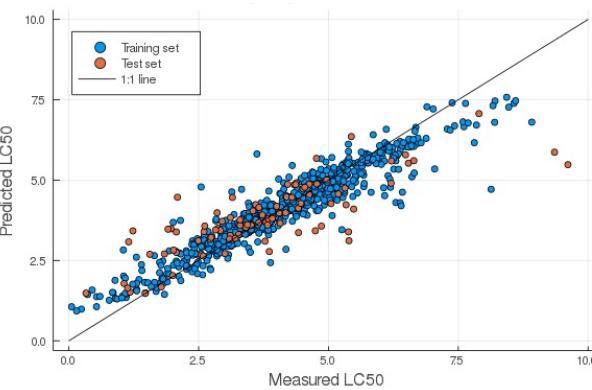
The molecular descriptors and QSAR models are used to predict toxicity of the chemicals. Those predicted toxicity values, then, are converted to toxicity categories.



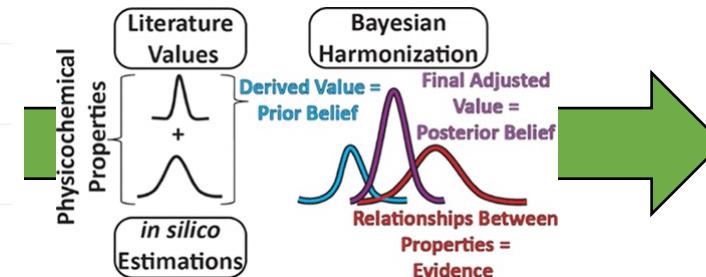
Molecular Descriptors



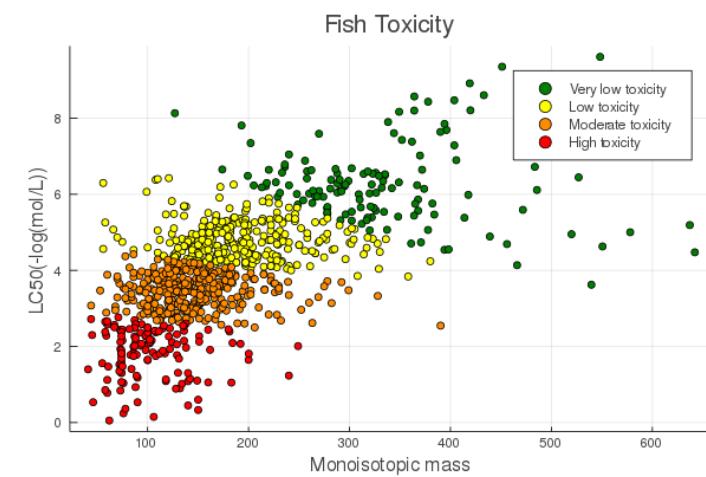
QSAR



Uncertainty



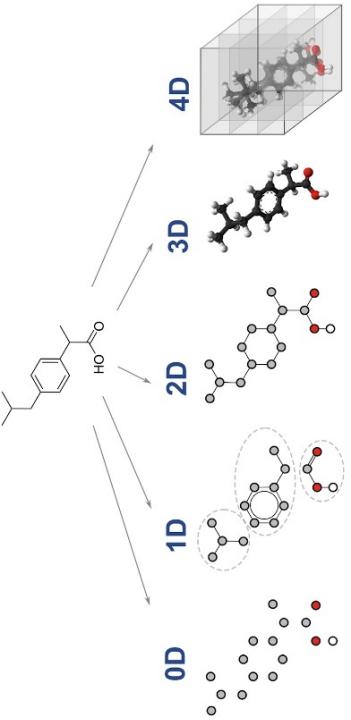
Tox Categories



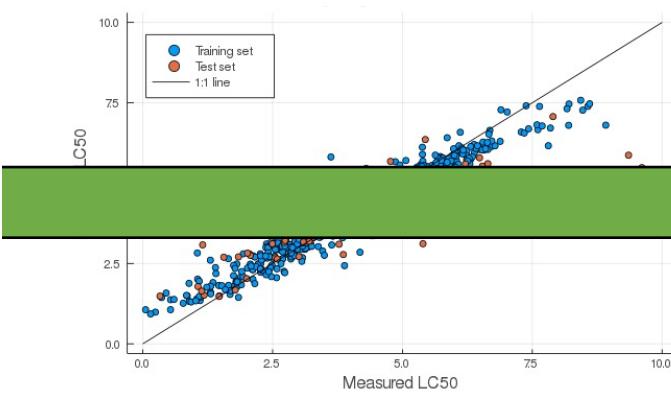
Recent studies have shown that the QSAR models depending on their training set and assumptions may have large levels of uncertainty.



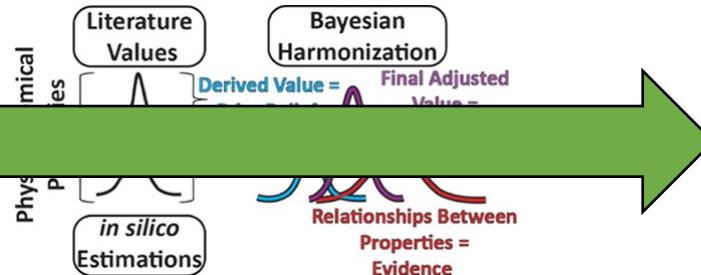
Molecular Descriptors



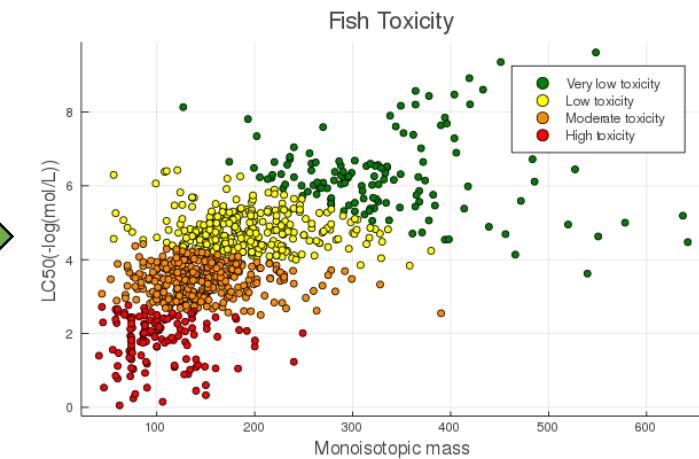
QSAR



Uncertainty



Tox Categories

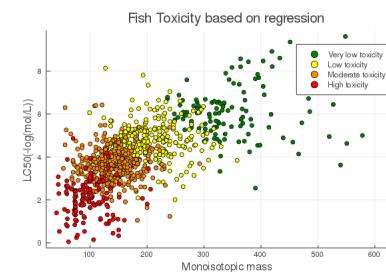
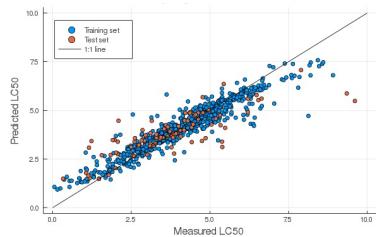
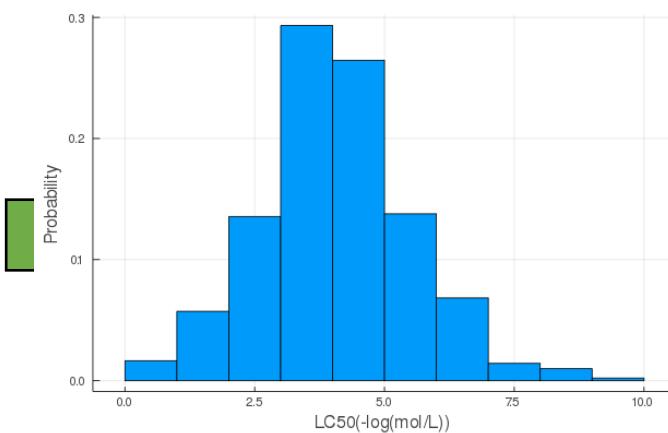


What if we go directly from descriptors to the toxicity categories?

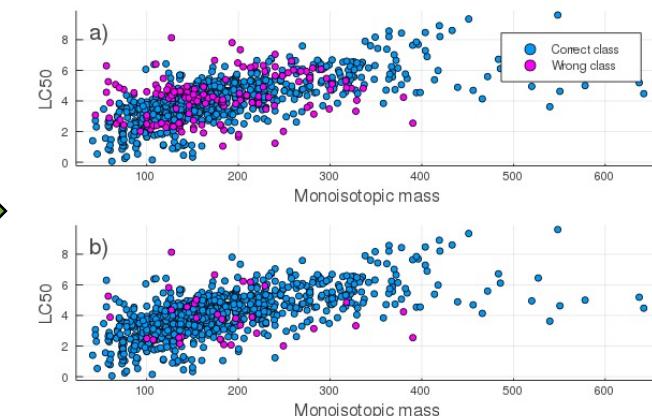


Conventional approach

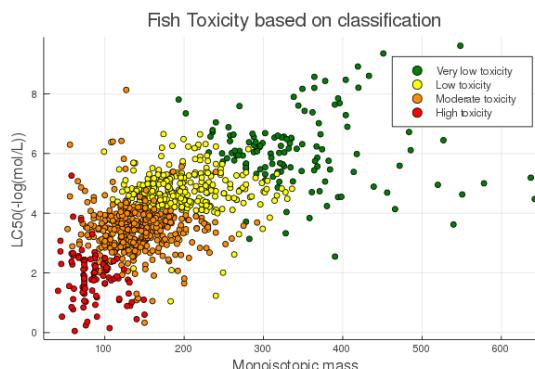
Training data
907 LC50 values



$$R^2_{\text{test}} = 0.75$$

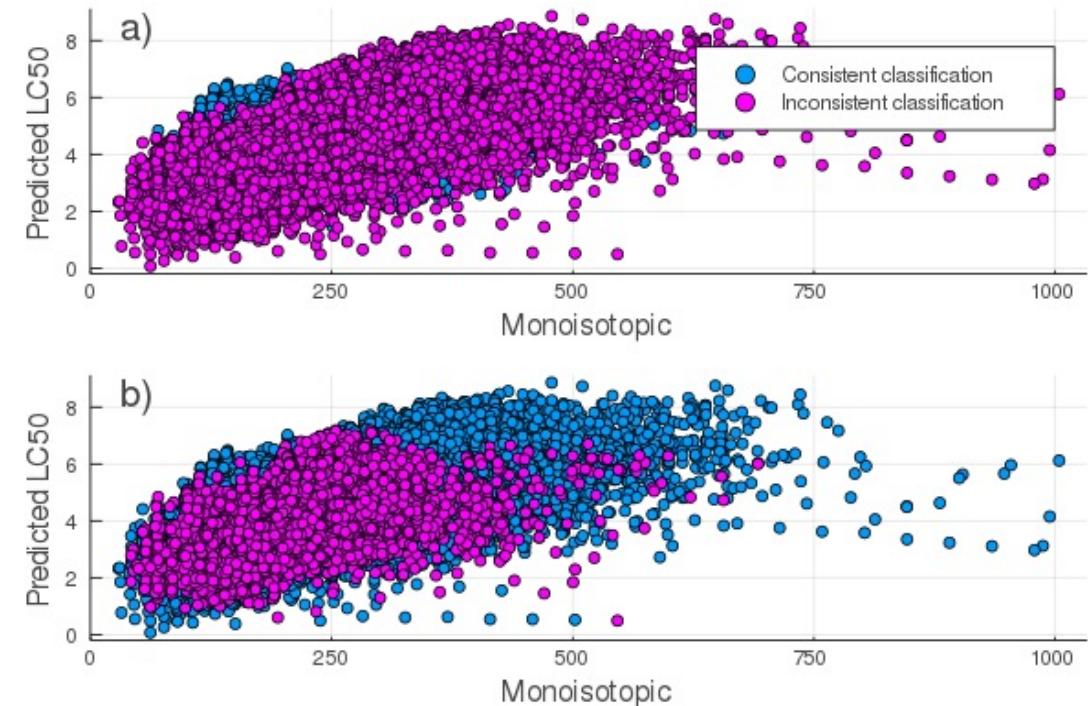
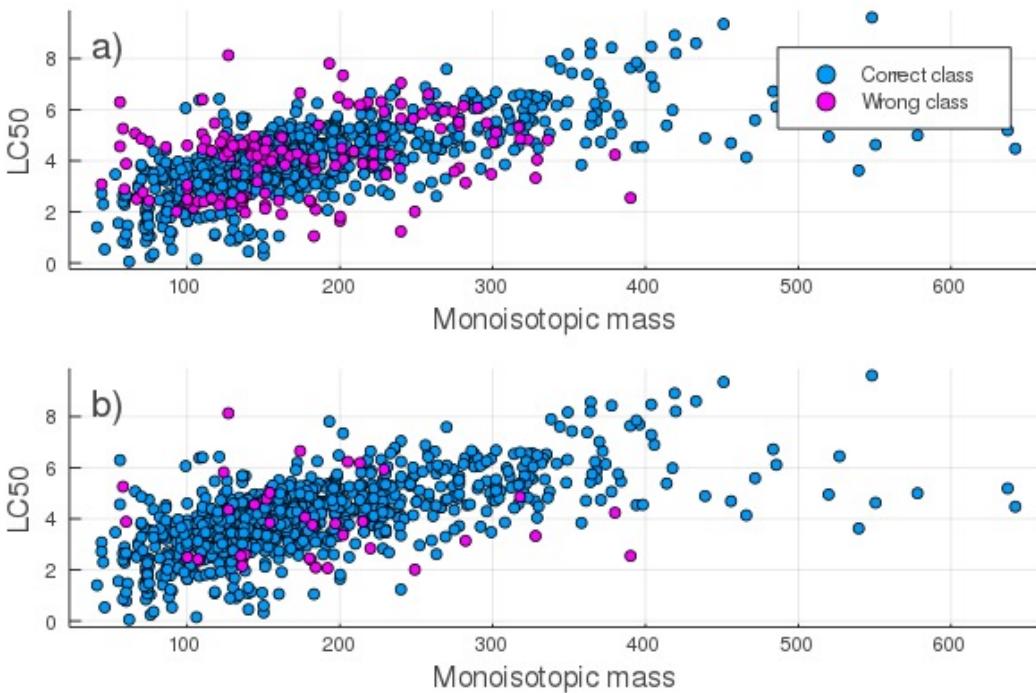


This work

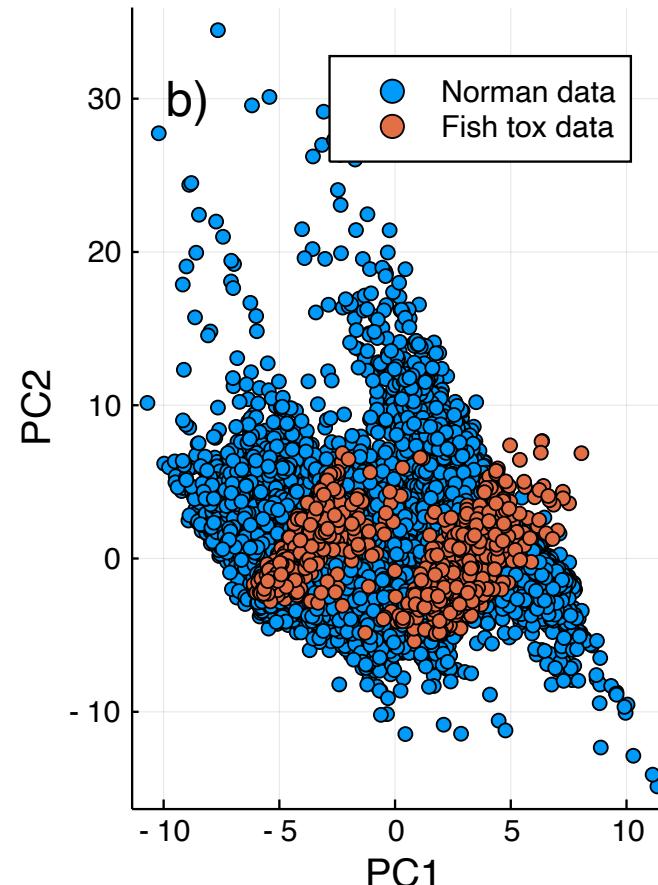
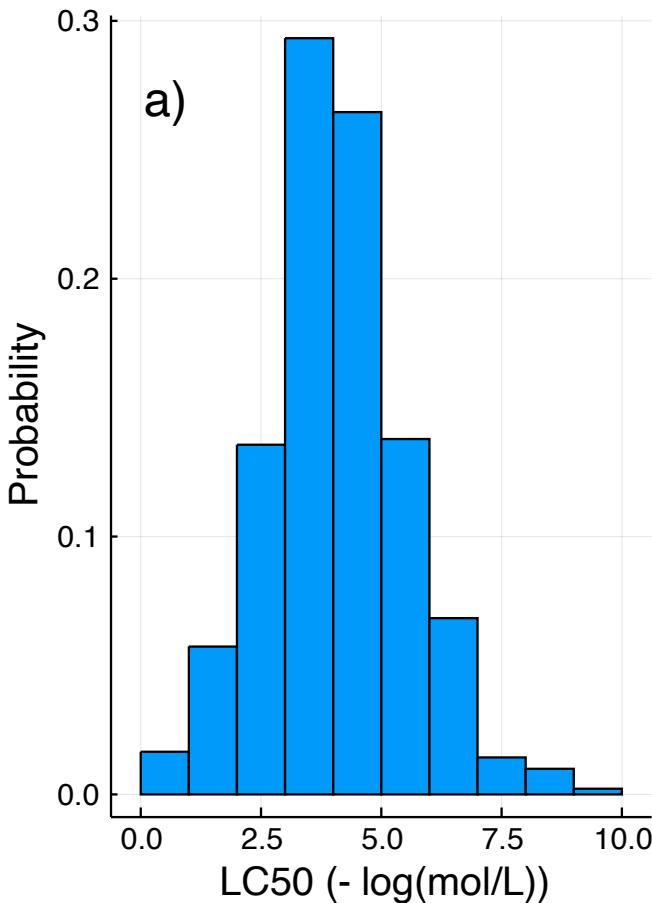


$$R^2_{\text{test}} = 0.85$$

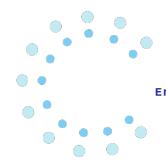


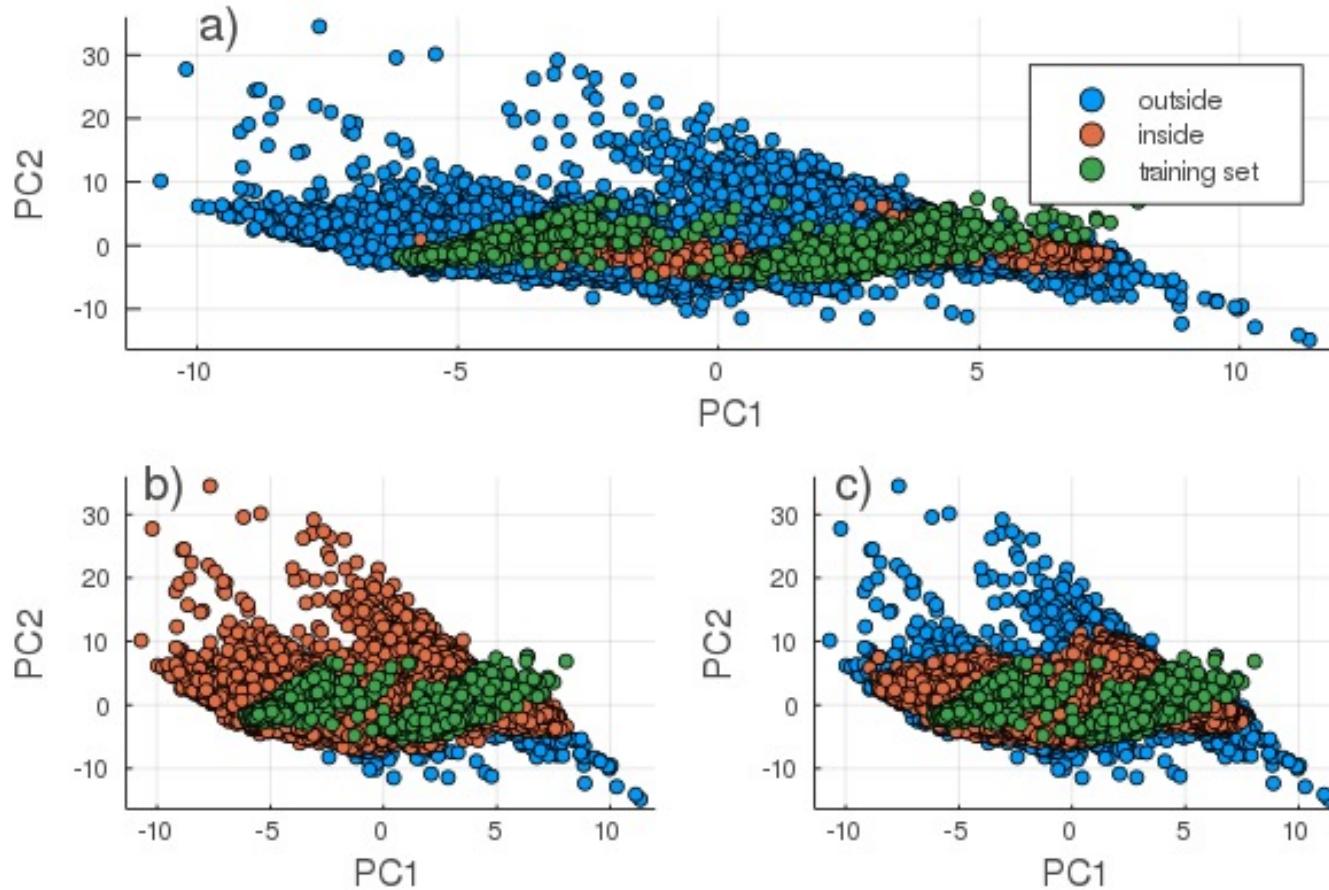


We observed 5 times increase in the correct tox categorization with direct classification model (35 vs 152 chemicals). For the Norman dataset our model was 3 times more accurate than the conventional approach.



The adequacy of the training set for modeling Norman dataset was assessed via Applicability Domain (AD) assessment. We used the leverage and bootstrapping as a means of these assessments.





The training set based applicability domain (AD) vs the model based AD were compared. The training set based strategy showed to be more effective in assessing the covered chemical space.





- A direct classification model was developed for prediction of the tox categories of organic pollutants.
- The classification model performed 5 times more accurately than the conventional approach.
- The applicability domains were further investigated.
- The training set based AD showed to be better in describing the covered chemical space.

- This strategy can be expanded to other fates and behavior (e.g. mobility).
- These categories can be predicted by looking at the HRMS spectra of the chemicals.
- The combination of these categories can be converted to hazard scores.



Thank you!

S.Samanipour@uva.nl

www.emcms.info



Code (public - FAIR)



Manuscript



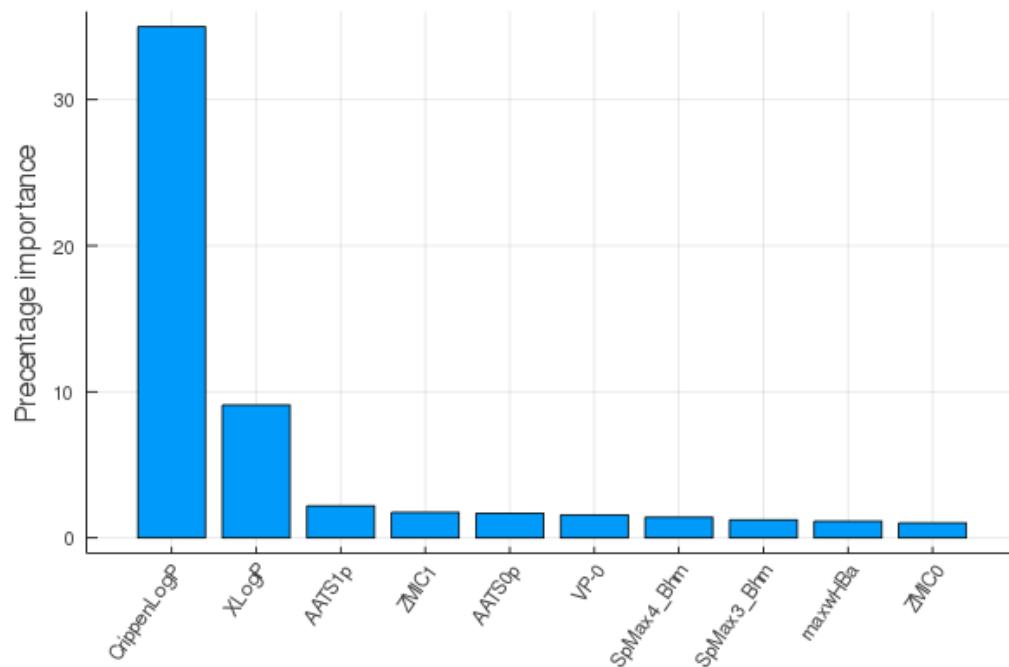


UNIVERSITY OF AMSTERDAM
Van 't Hoff Institute for Molecular Sciences

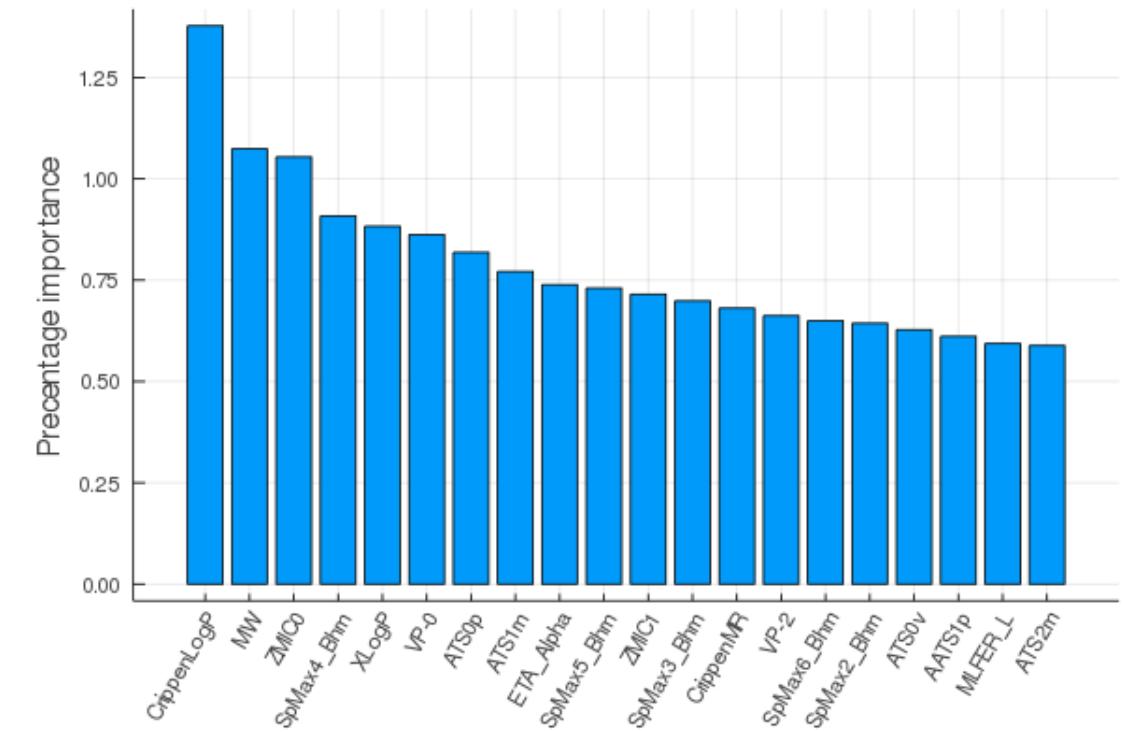


Environmental Modeling & computational Mass Spectrometry

Regression



Classification





LC50 values

