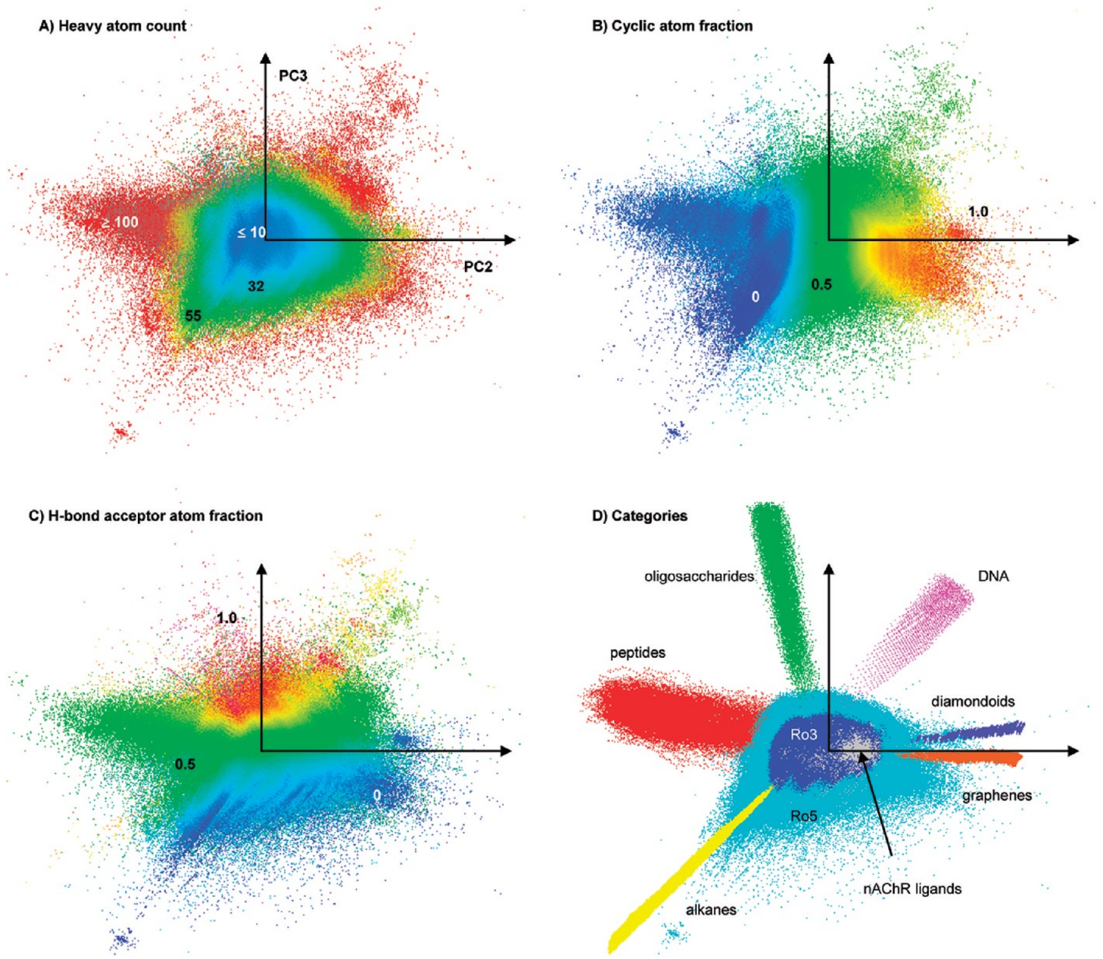
University of Amsterdam
Van ‘t Hoff Institute for Molecular Sciences

An in-silico coverage evaluation of RPLC-MS chemical space via two different retention indices scales

Saer Samanipour, A. Nikolopoulos, D. van Herwerden, J. W. O’Brien, L. Barron, and K. V. Thomas

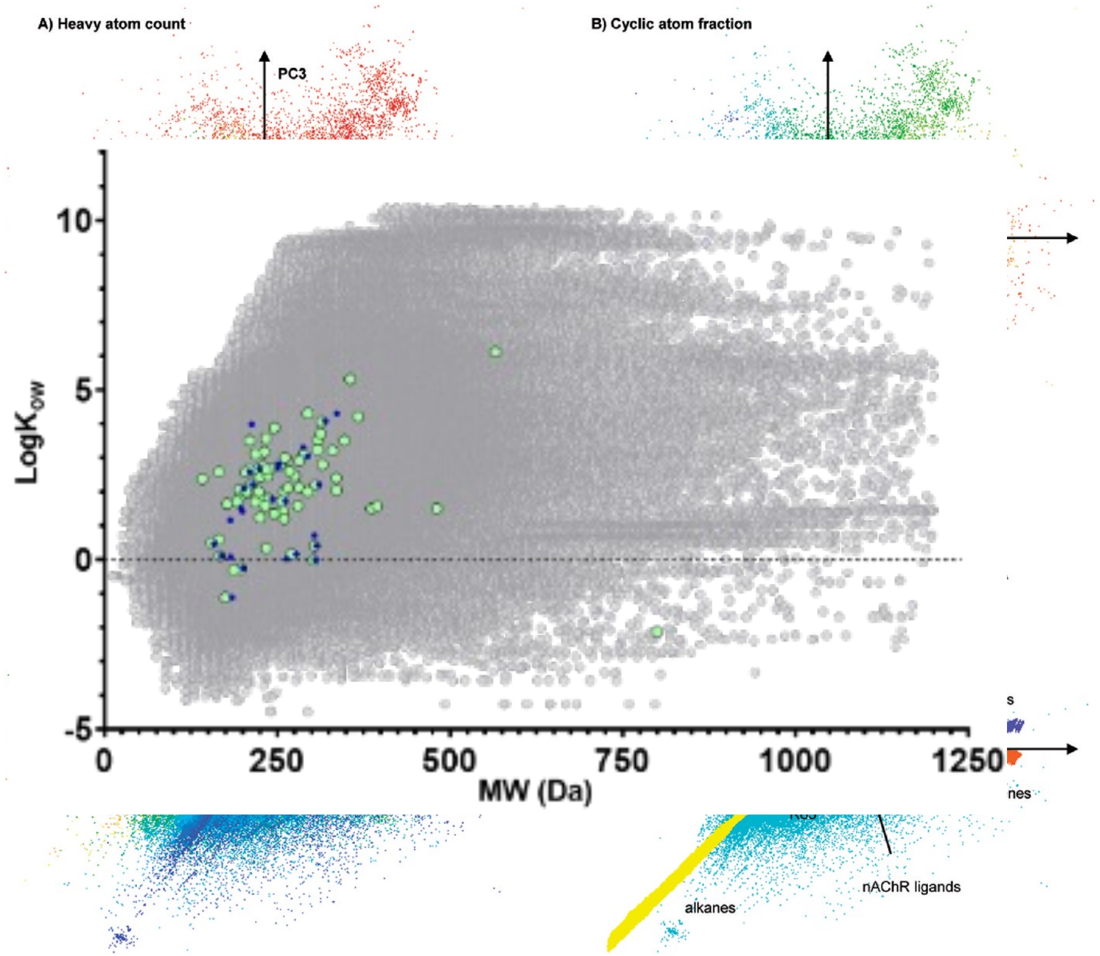
Environmental Modeling & Computational Mass Spectrometry

The University of Queensland
Australia
Create Change



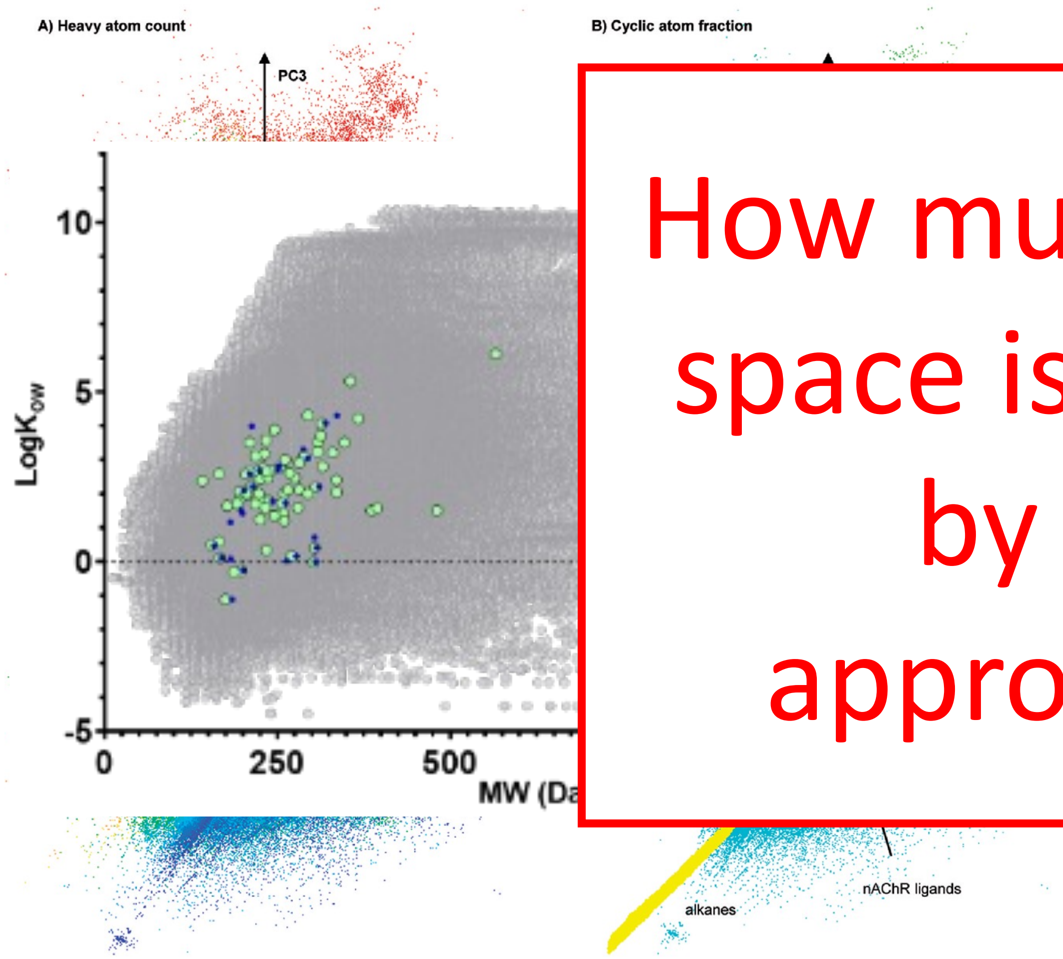
- There are > 10^{60} possible structures with Mw < 500 Da.
- The physiochemical property range is too wide.
- Technologically we cannot cover this space.

ACS Chem. Neurosci. 2012, 3, 9, 649–657



- 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%.

Schulze, others, and ***Samanipour****, TRAC, 2020* and McEachran et al. Anal Bioanal Chem (2017) 409:1729–1735.



How much of this space is covered by NTA approaches?

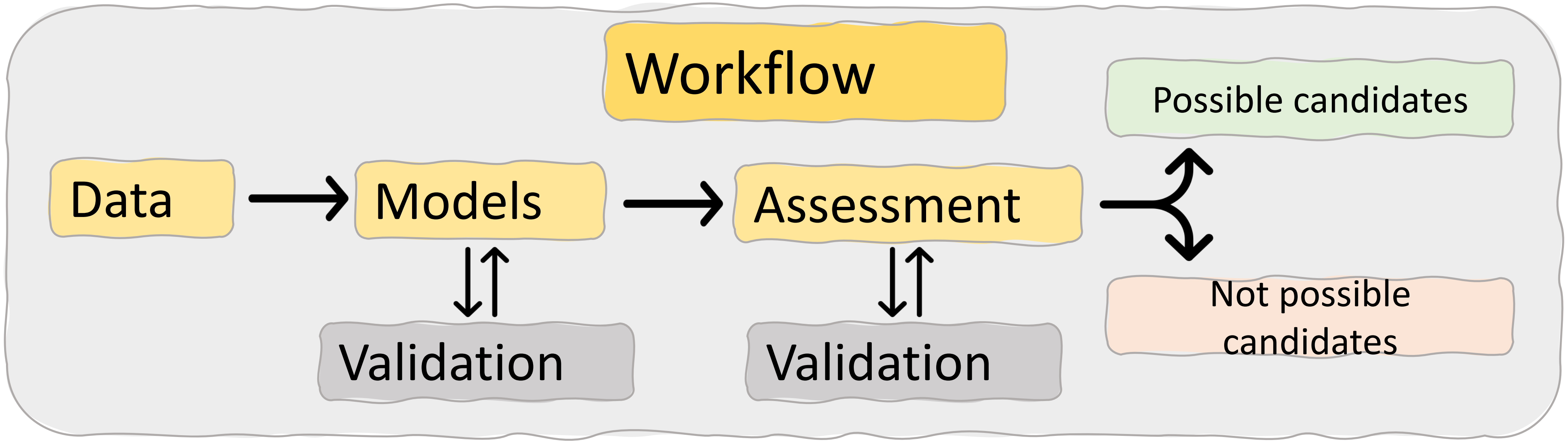
nore than 800k known
hat are actively released
vironment.
rmation products are

chemicals were

ethods for less than 1%.

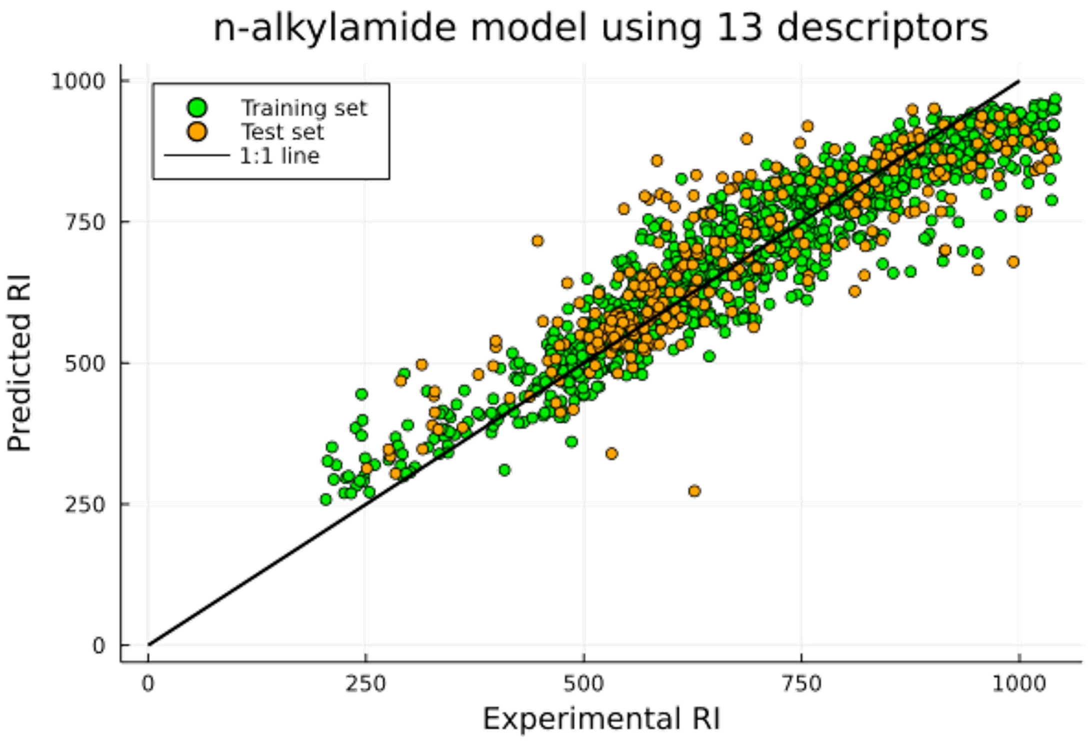
*Schulze, others, and **Samanipour**, TRAC, 2020* and McEachran et al. Anal Bioanal Chem (2017) 409:1729–1735.

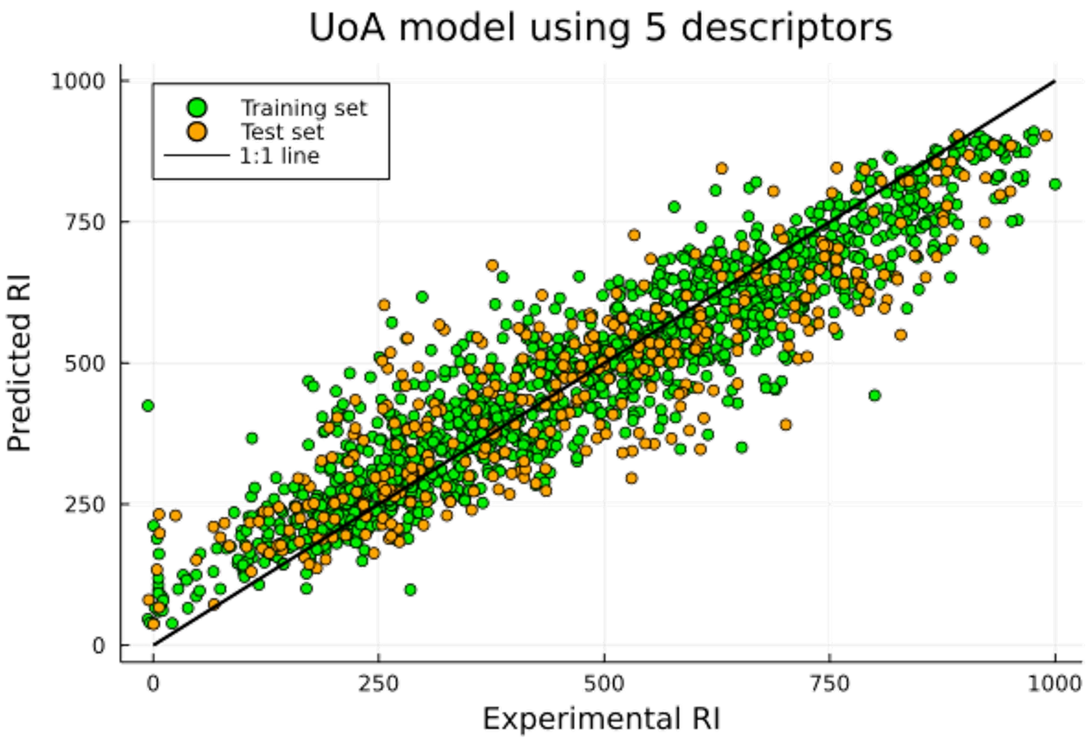
To filter out the chemicals with very low or very high retention on C18 columns.



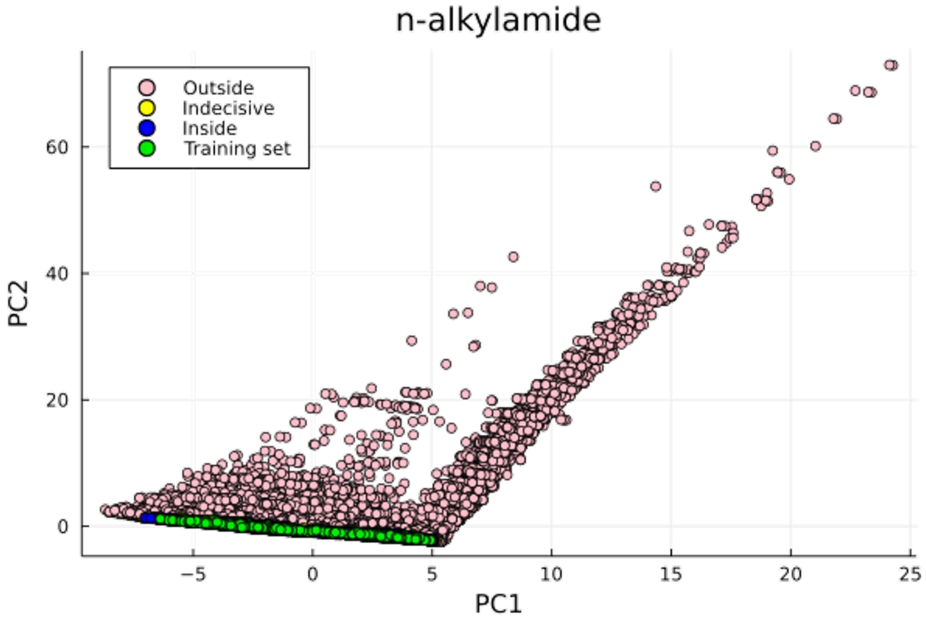
To filter out the chemicals with very low or very high retention on C18 columns.

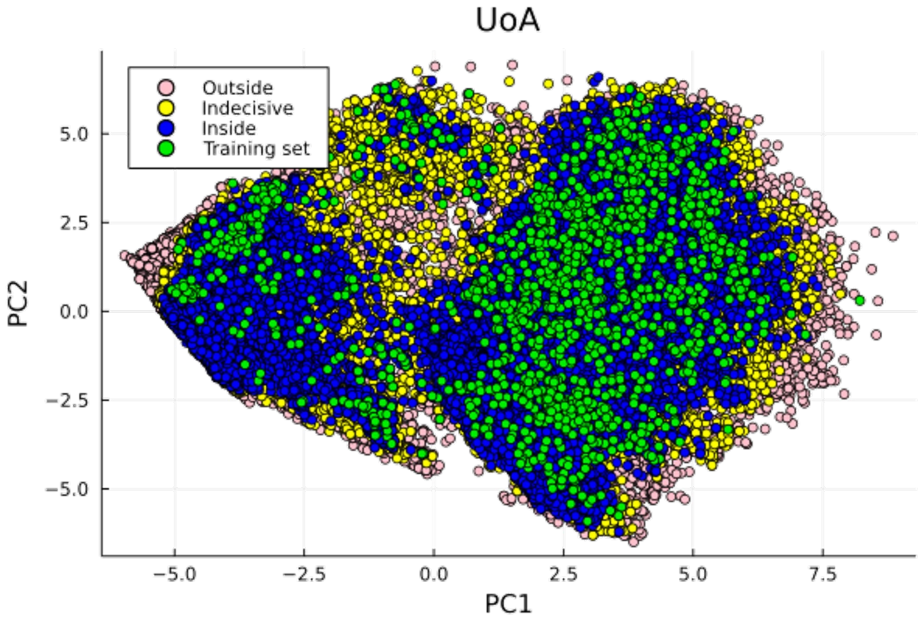
- We built two models,
- We validated those models with a set of true positives and negatives,
- We applied the two models to the SusDat as a filtering strategy.


n-alkylamide model using 13 descriptors


UoA model using 5 descriptors

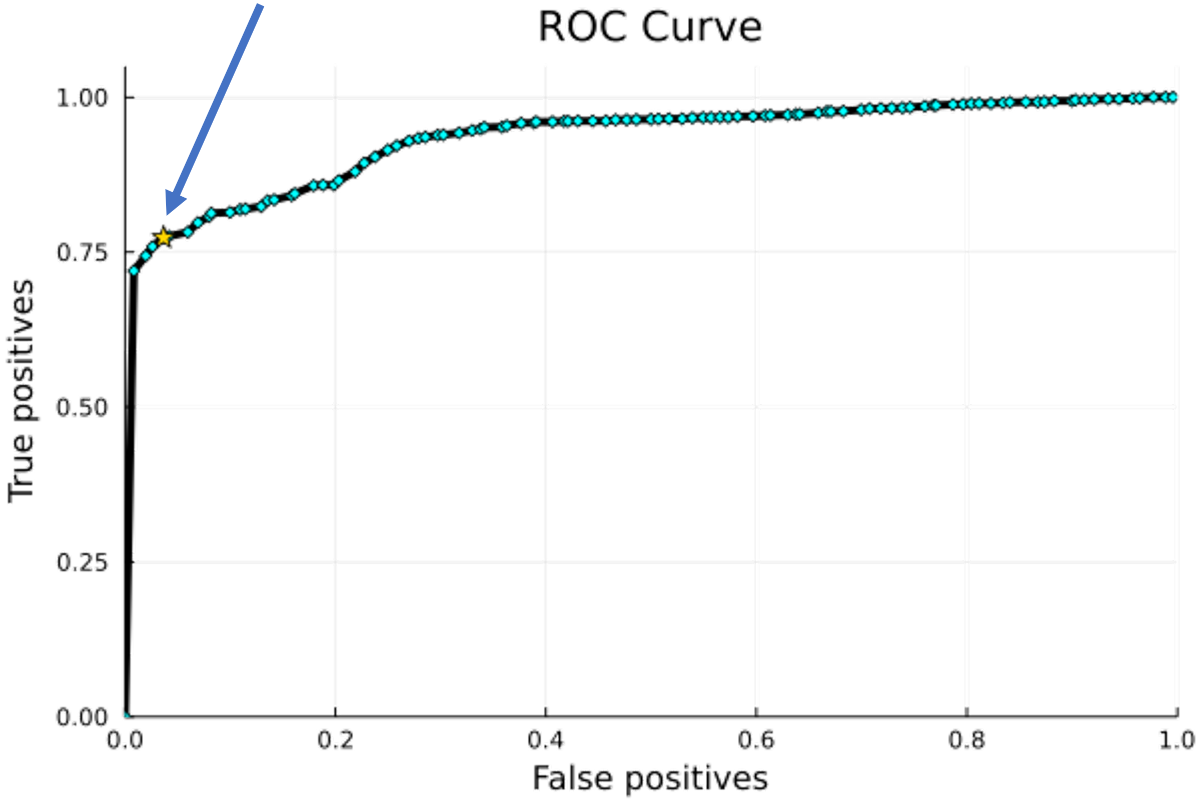
Two detention indices prediction models were developed, one based on amide scale and the second one based on the UoA scale.





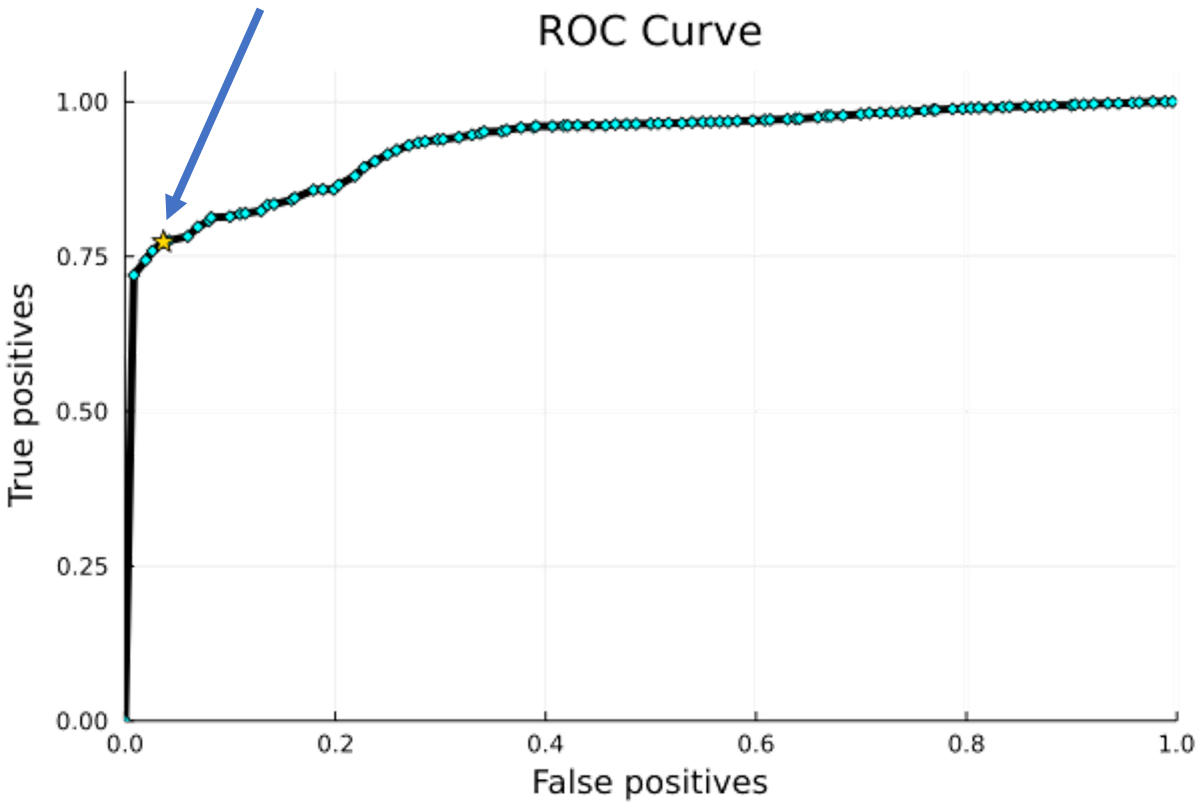
We used the applicability domain as a first step towards filtering the chemicals that may not fit into our C18 column space.

Optimal point



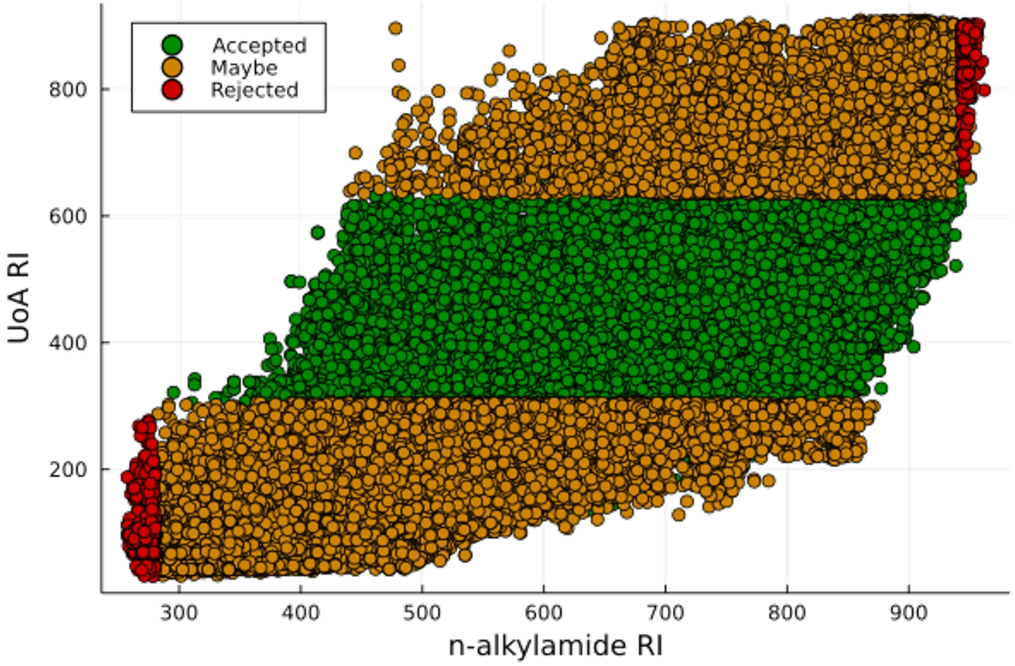
- 2.2 k chemicals from MassBank => true positives
- 9 k chemicals from SusDat => true negatives
  - logP > 6 or logP < -4
- These were divided in validation and test sets (80/20).

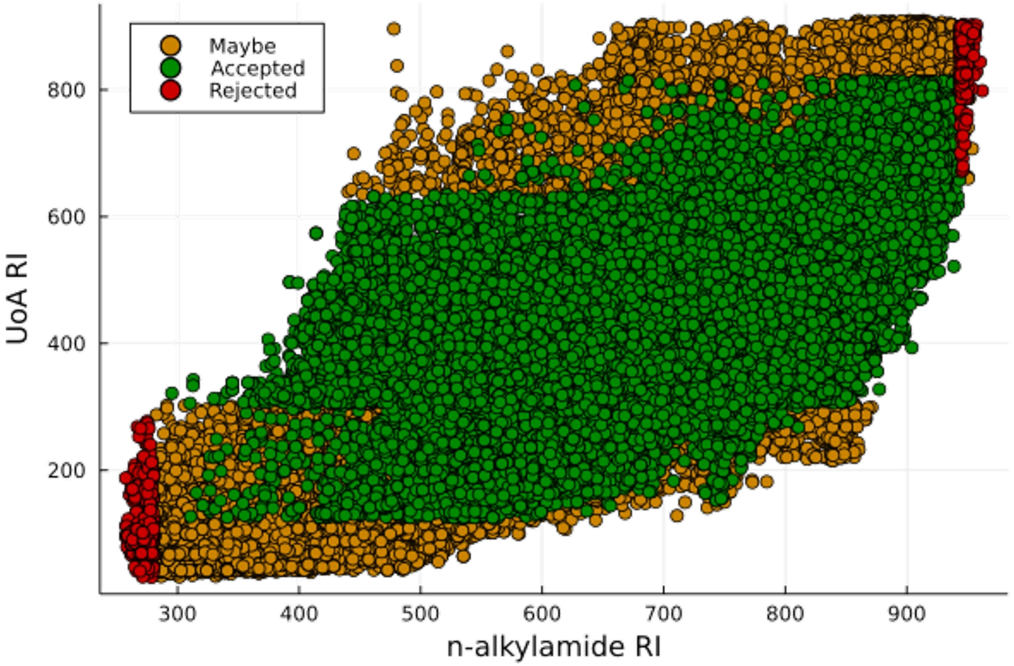
Optimal point



- 2.2 k chemicals from MassBank => true positives
- 9 k chemicals from SusDat => true negatives
  - logP > 9 or logP < -4
- These were divided in validation and test sets (70/30).

Under the optimal conditions our model showed to have an accuracy of 92.4%.





Roughly around 50% of chemicals in SusDat do not fit into the C18 column space. Other chromatographic methods are needed to cover the missing space.

- Two RI prediction models were developed.
- The combination of applicability domain and the model errors were used for the extrapolation assessment.
- Our models when combined together accurately (92.4%) predicted the usability of C18 for the analysis.
- When applied to SusDat around 50% of the chemicals did not fit the C18 chemical space.

- This approach can be used for a-priori filtration of the chemical databases.
- This would increase the speed of queries and their relevance.
- Highly indicative of the need for additional chromatographic approaches.
- The ionization efficiency may be an additional filtering tool.

www.emcms.info

Thank you!

Code (public - FAIR)

Models with reduced number of variables

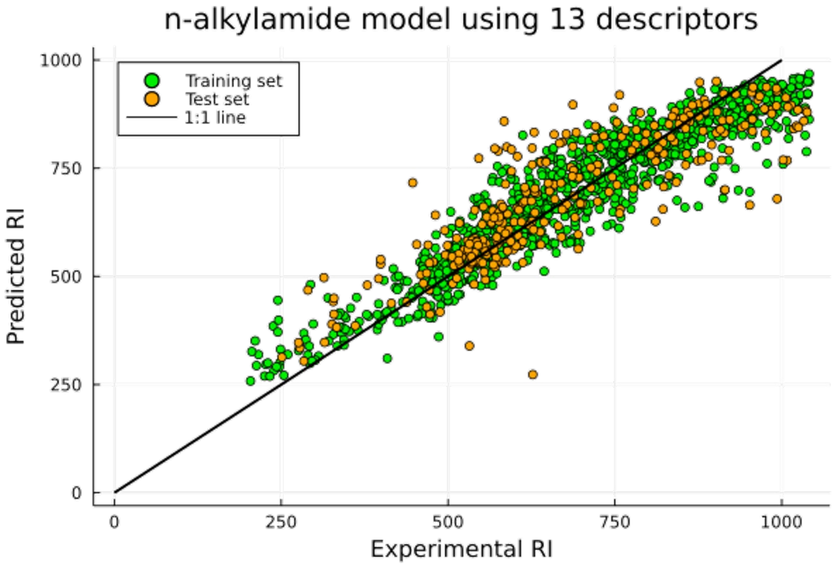
However, using 1170 variables is computationally expensive, what if we use fewer descriptors?

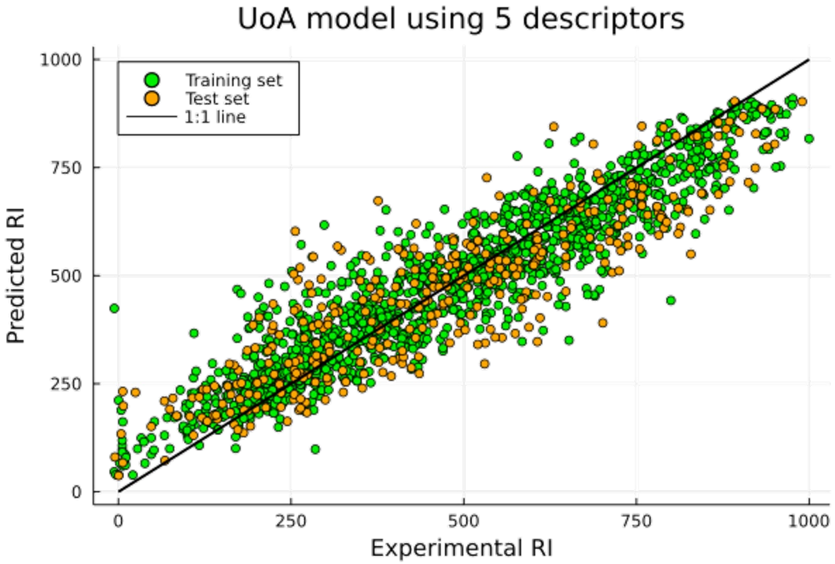
Most important descriptors were found based on average variance reduction

New models were created with a small number of (the most important) descriptors

	n-alkylamide	UoA
R^2	0.73	0.77

We have two models able to predict accurately and fast the retention indices!


n-alkylamide model using 13 descriptors


UoA model using 5 descriptors

Model hyperparameters optimization

Optimization

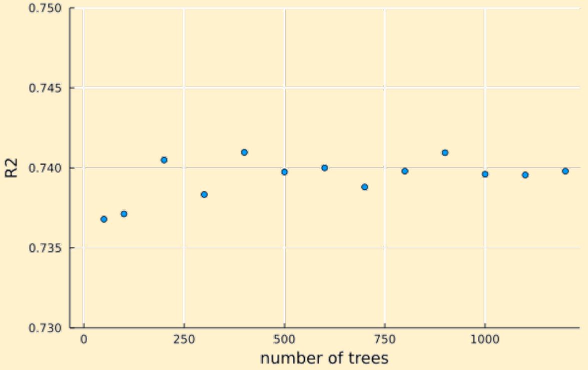
The hyperparameters were optimized to maximize accuracy and avoid overfitting

	n-alkylamide	UoA
No. of trees	400	500
Sample per leaf	4	4

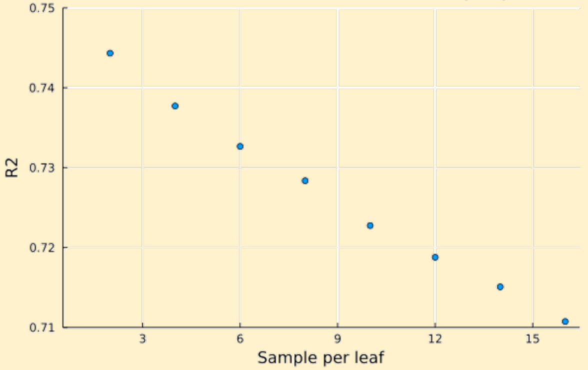
We have two models able to predict accurately the retention indices!

n-alkylamide dataset

Cross validation score for different number of trees

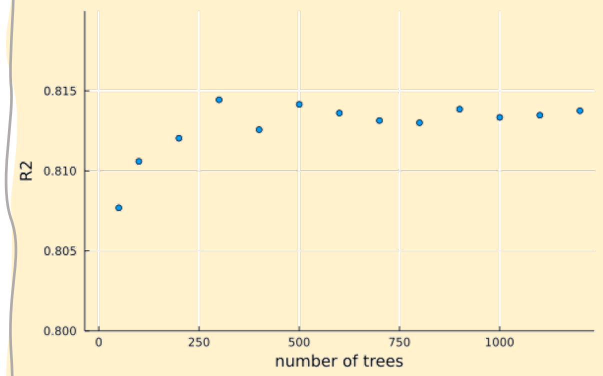


Cross validation score for different sample per leaf

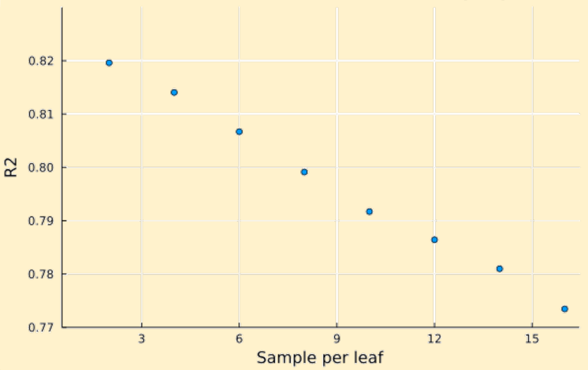


UoA dataset

Cross validation score for different number of trees



Cross validation score for different sample per leaf



Descriptors acquisition

Training datasets

- n-alkylamide: 1488 compounds and their experimental r_i
- UoA: 1816 compounds and their experimental r_i

Dataset used for application

- SusDat NORMAN: 95115 compounds

PaDEL was used to calculate the descriptors of these datasets

But can we use all 2756 descriptors provided by PaDEL?