Semi-quantification

Semi-quantification: model development

All existing data was used to train IE prediction model for ESI-. For this, training and test sets of Thomas' data were joined. This model can be used for any semi-quantification of PFAS.

The code has been cleaned and organized in functions for smooth retraining if additional data is available for updating the model. The code and the model can be found from GitHub: https://github.com/kruvelab/PFOA_semi_quant

Semi-quantification: Homologous series vs ML model predictions

In addition to updating the IE prediction model, homologue series was tested for semi-quantification. For this, the following steps were taken:

1) Finding homologue series compounds
2) Computing the calibration graph for all of the chemicals
3) Assigning the calibration graph of the closest homologue to each analyte
4) Semi-quantification with the calibration graph of the homologue

This was repeated for all chemicals in the homologue series.

In addition the IE prediction model was used under comparable conditions. For this, following steps were taken:

1) Training IE predictive model based on all of the analytical standard, excluding the analyte
2) Predicting ionization efficiency for the analyte with the trained IE model
3) Predicting response factors from the predicted ionization efficiency by using RF vs IE plot of all analytical standards
4) Semi-quantification based on the predicted response factors

Finally, the results from the two approaches were compared visually and numerically.

Finding homologue series compounds

From Thomas' data, a dataset was generated, which contained all compounds that had at least one homologue within the dataset.

Assumption: Two compounds are considered homologues when their difference in molecular formula is CF_2.

This summary is made on the example of CF_2 homologues only, see below. Data for CF_2CF_2 homologues can be calculated if needed.

```
## # A tibble: 18 x 3
##    Compound Homologue pattern_CF2
##    <chr>    <chr>     <chr>
##  1 PFDA     PFNA      smaller
##  2 PFDA     PFUnDA    bigger
##  3 PFDoDA   PFTriDA   bigger
##  4 PFDoDA   PFUnDA    smaller
##  5 PFHpA    PFHxA     smaller
##  6 PFHpA    PFOA      bigger
##  7 PFHxA    PFHpA     bigger
##  8 PFHxA    PFPeA     smaller
##  9 PFNA     PFDA      bigger
## 10 PFNA     PFOA      smaller
## 11 PFOA     PFHpA     smaller
## 12 PFOA     PFNA      bigger
## 13 PFPeA    PFHxA     bigger
## 14 PFTeDA   PFTriDA   smaller
## 15 PFTriDA  PFDoDA    smaller
## 16 PFTriDA  PFTeDA    bigger
## 17 PFUnDA   PFDA      smaller
## 18 PFUnDA   PFDoDA    bigger
```

Semi-quantification with homologue

For each compound, calibration curve of a homologue was used for semi-quantification. If two homologues existed (bigger and smaller), quantification was done with both and both results are shown on the graph as well as taken into account in performance calculations.

The IE approach assumes that the intercept of the calibration graph is statistically insignificant or much smaller than the peak areas. Therefore, two different approaches were also used for the quantification with the homologue series.

First approach: Only slope (RF) was used to calculate concentrations (regression line was not forced to go though zero).

$(conc = area/slope_{homologue})$

Second approach: Both slope and intercept were used to calculate concentrations.

$(conc = (area - intercept_{homologue})/slope_{homologue})$

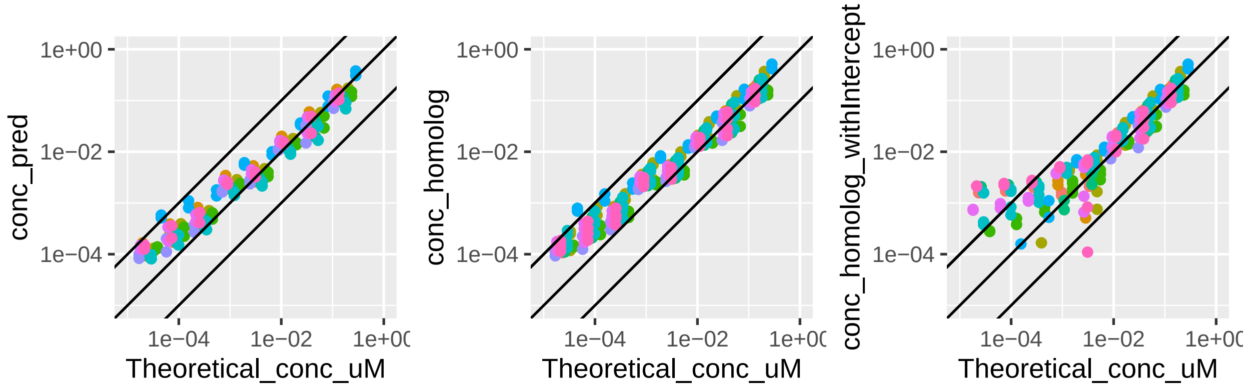
Predicting ionization efficiency and response factors for homologue series compounds + quantification

For each homologue series compound, the compound was removed from the training data and prediction model was trained (10 prediction models were trained in total). Then, the model was used to predict IE of the compound. The predicted ionization efficiency values were correlated with experimental response factors (that is slope) for all analytical standards, this correlation was used to predict a response factor for the analyte from the IE.

$(conc = area/slope_{predicted})$

Visual comparison

Comparing semi-quantification results from predicted slopes and homologue series compounds slopes with theoretical concentration. Ideal regression and ten-times error lines were added.



Results and interpretation

The error factor for IE and homologue series, first approach, based semi-quantification are very similar ranging from 2.3x to 3.0x. Homologue series based approach shows slightly higher error factor; therefore, it can be concluded that a single IE based semi-quantification can be used for all PFAS and homologue series based quantification is not necessary. See results below. Importantly, for both methods we see that the quantification is less accurate for lower concentration level. This probably arises from slightly reduced linearity at low concentration and/or importance of intercept at these low levels.

```
## # A tibble: 2 x 3
##   pattern error_IE error_homolog
##   <chr>      <dbl>         <dbl>
## 1 bigger      2.44          2.67
## 2 smaller     2.27          2.96
```