Unsupervised learning for candidate structure prioritization based on retention time prediction

Gordian Sandberg[a], Helen Sepman[a,b], Isak Samsten[c] & Anneli Kruve[a,b]

[a] Department of Materials and Environmental Chemistry, [b] Department of Environmental Science, [c] Department of Computer and Systems Science, University of Stockholm

Introduction

Wastewater contains a multitude of different chemicals that arise from agricultural use, personal care products, industry and natural sources.

These potentially toxic chemicals should be identified to find an appropriate wastewater treatment and considered for regulation.

For identification rapid and reliable separation and data acquisition is necessary. The combination of High Performance Liquid Chromatography (HPLC)

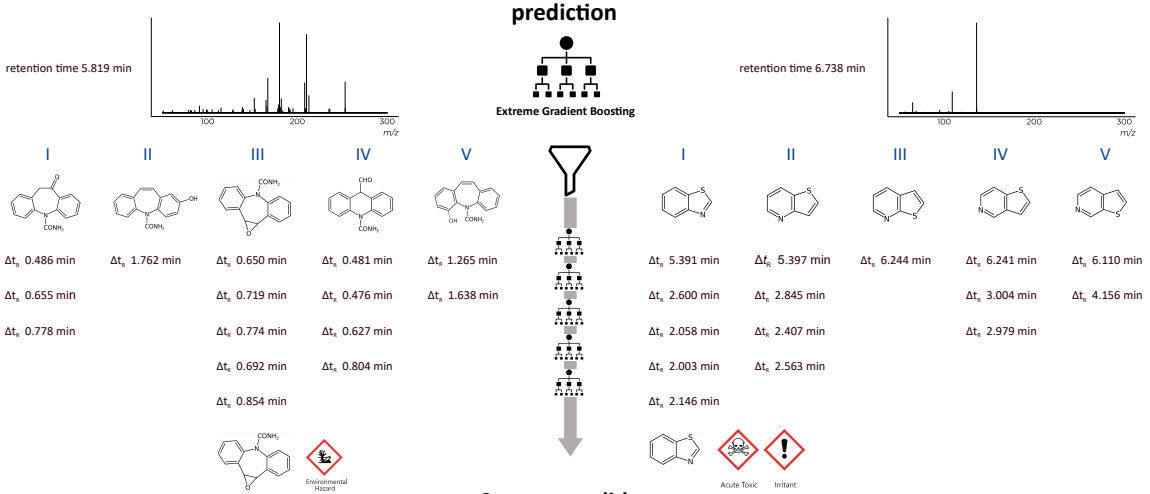
with an Electrospray Ionisation (ESI) High Resolution Mass Spectrometer (HRMS) provides thousands of LC/ESI/HRMS^2 features.

These LC/HRMS^2 features can be structurally interpreted by MS^2 scorers. The state of the art program is SIRIUS+CSI:FingerID.[1] Ranked candidate structures for each LC/HRMS^2 feature are produced.

The idea of the following workflow is, to make use of the readily available retention time as orthogonal information to rerank the Top 5 candidate structures that are suggested for each LC/HRMS^2 feature by predicting their retention times and comparing with the measured retention time.

Workflow

Molecular descriptors calculated with PaDEL from 2D structures, ranging from atom and element counts to molecular volumes and hydrophobicity, were calculated. Highly correlated molecular descriptors (>75%) and molecular descriptors with more than ten missing values are removed.
For each LC/HRMS^2 feature the measured retention time and a collection of ranked candidate structures is extracted. The performance of the method is evaluated on spiked chemicals, matched with the candidate structures based on retention time and MS^2 spectra.



Method

The number of candidate structures obtained from LC/HRMS^2 is reduced by predicting the retention times of individual candidates based on the PaDEL descriptors. The predicted retention times for candidate structures are compared with the measured retention time of the LC/HRMS^2 feature.

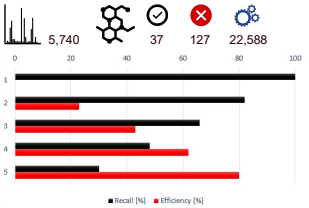
$$\Delta t_R = |\text{predicted } t_R - \text{measured } t_R|$$

The compound with the highest Δt_R (per LC/HRMS^2 feature) is removed and the retention time prediction is repeated upon the remaining candidates. Thereby, the quality of the candidate dataset improves, assuming that the correct candidate structure for an LC/HRMS^2 feature is among the suggested candidates. For model training the Top 5 ranked structures (I-V) per LC/HRMS^2 feature from SIRIUS+CSI:FingerID are used.

Results

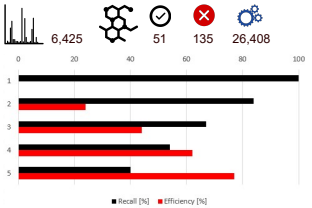
Workflow and method were developed on a dataset of candidate structures from two different wastewater plants and validated on a dataset of candidate structures from three different wastewater plants. Retention time prediction models were trained on 5,740 LC/HRMS^2 features with 22,588 candidate structures (SIRIUS+CSI:FingerID Top 5 ranked structures) for the method development and on 6,452 LC/HRMS^2 features with 26,408 candidate structures for the validation.
The method development dataset contained 37 spiked chemicals with 127 incorrect candidates for performance evaluation, the validation dataset contained 51 spiked chemicals with 135 incorrect candidates.

Method development



The performance evaluation on the candidates of the spiked chemicals is twofold: while the recall measures the percentage of correct candidates retained over five prediction-removal cycles, the efficiency measures the percentage of incorrect candidates removed.

Method validation



Test and validation set show very similar recall and efficiency over the first three prediction-removal cycles. The maximum difference is 2% here. In the 4th and 5th cycle the larger validation set shows higher performance in the recall, retaining 40% instead of 30% of correct candidate structures.
Validation on larger datasets is upcoming. Nontheless, the method exhibits a promising separation power of correct and incorrect candidate structures.

Kruve lab

Funded by the
Wallenberg Initiative Material Science for Sustainability

[1] Dührkop, K. et al., *Nat. Methods* **2019**, 16 (4), 299–302.
[2] Yap, C. W.; *Comput. Chem.* **2011**, 32, 1466–1474.