Can active learning improve the performance of computational mass spectrometry? Case study on the example of ionization efficiency predictions

Wei-Chieh Wang,[a] Nahid Amini,[b] Meelis Kull,[c] Anneli Kruve[a,d]

[a]*Department of Materials and Environmental Chemistry, Stockholm University, Svante Arrhenius väg 16, 114 18 Stockholm, Sweden*
[b] *Oriflame Cosmetics, Scientific Research & Technology, Fleminggatan 14, 112 26 Stockholm, Sweden*
[c]*Institute of Computer Science, University of Tartu, Narva mnt 18, 51009, Tartu, Estonia*
[d]*Department of Environmental Science, Stockholm University, Svante Arrhenius väg 8, 114 18 Stockholm, Sweden*

Quantifying chemicals detected with liquid chromatography high-resolution mass spectrometry (LC/HRMS) poses a considerable challenge, primarily due to the varied responses exhibited by different chemicals at the same concentration. Lack of analytical standards, especially for novel or unstable chemicals, limits the application of traditional calibration graph methods, leading to over 98% of the chemicals detected in environmental analyses remaining unquantified. In response to these challenges, machine learning (ML) models have been introduced to predict the responsiveness of chemicals in LC/HRMS and subsequently estimate concentrations. However, it has become a concern that these ML predictions may have significant errors. This discrepancy indicates a potential gap in the comprehensive training of these models to effectively associate differences in chemical structures with responsiveness. One plausible explanation is the insufficiency of diverse and informative data during the training phase, resulting in limited coverage of the desired chemical space. We advocate for an approach to address this limitation by concurrently exploring the chemical space while training ML models. This strategy is implemented through an active learning methodology that leverages uncertainty- and clustering-based algorithms. During each active learning iteration, uncertainty values are employed to identify and prioritize the most informative chemicals, while clustering ensures the maintenance of diversity within the selected chemicals. Additionally, leveraging the chemicals closest to the cluster centroids is designed to pinpoint the most representative ones. The density of the clusters is considered to strike a delicate balance between the two aforementioned indexes. Ultimately, we propose preferred proportions for the new targeted chemical space to reach the desired performance, which offers a practical solution to meet the evolving needs of chemical space exploration in LC/HRMS analyses.

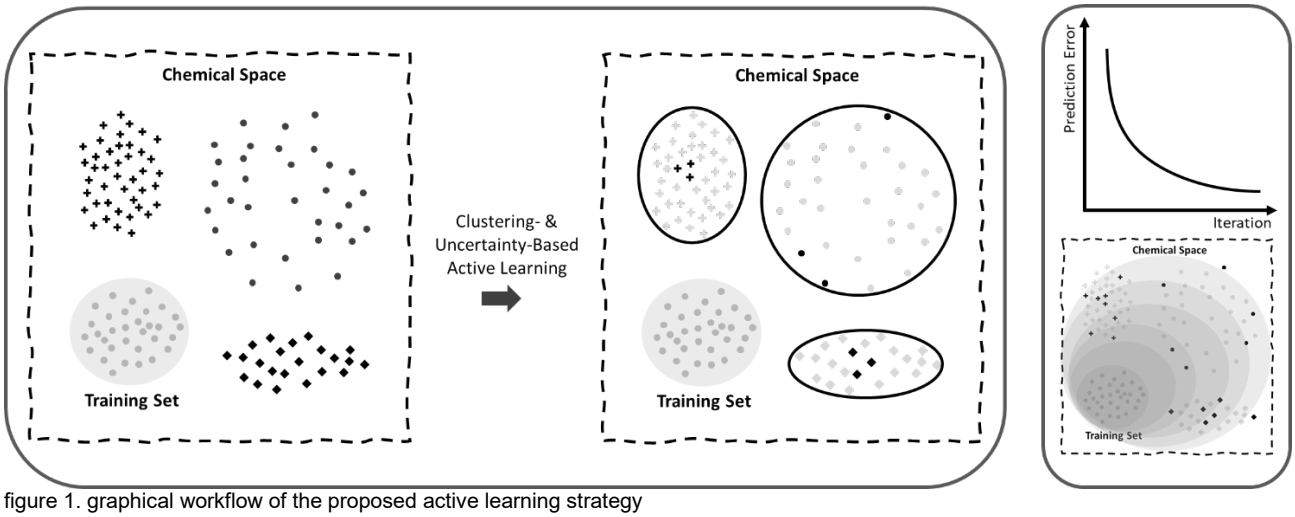

figure 1. graphical workflow of the proposed active learning strategy