Streamlining Environmental Data Analysis: A Deep Learning Approach

Edison Cerda , Tamas King, Oleksandr Sosnovshchenko, Charudatta Manwatkar, Flavio Avila, Yvonne Koch, Ruoji Luo

Agilent Technologies, Inc.

NEMC 2025
P-29

Agilent
Trusted Answers

Introduction

per- and polyfluoroalkyl substances (PFAS), a class of emerging persistent organic pollutants (POPs), are present at trace level not only in the environment (water, soil and air) but also in food. The quantitative analysis of PFAS is typically performed using liquid/gas chromatography-tandem mass spectrometry (LC-MS/MS, GC-MS/MS). However, even with these highly sensitive instruments, PFAS analysis remains challenging.

The data analysis of PFAS often involves time-consuming manual steps for the elimination of false positive or negative quantifier/qualifier peaks of the corresponding compound. This includes, but not limited to:

- Adjusting peaks from early eluting PFAS (e.g. PFBA, PFMPA),
- Combining partially or fully separated peaks of linear and branched isomers of some PFAS (e.g. PFOS, PFHxS), while accounting for variations in their ratio
- Removing false positive or negative peaks caused by matrix interferences or contamination.

In this work, a redesigned pipeline-originally for the GC/MS data-is adapted for the LC-MS/MS MRM mode data. Chemically relevant metadata, such as retention time shifts, quantifier-qualifier correlation, are considered during the design of the data preprocessing workflow. Data acquired for the analysis of PFAS in different environmental matrices following the EPA 1633 method, and from different LC-MS/MS instruments with varying sensitivities, are used for the model training and validation. A CNN and a transformer model are evaluated and their performance compared. Results show that both models perform well. When a trained deep learning (DL) model is deployed, data review time can be significantly reduced by eliminating most of the manual steps, as mentioned above, on a compound-by-compound base.

Figure 1. Infinity III 1290 and 6495D LC/TQ.

Experimental

Two different datasets are collected internally, one following the EPA 1633, and the other following AOAC Standard Method Performance Requirements (SMPR) 2023.003.[1,2]

During the model training period, samples are analyzed normally in MassHunter Quantitative Analysis software (12.1 update 2, Quant-My-Way UI, M version), following the conventional data analysis workflow in the software, as shown in Figure 2, left column. Data moves between the local PC and the infrastructure in the cloud. A schema of the components of the pipeline is shown in Figure 3. Various DL architectures have been adapted to handle LC-MS/MS MRM mode data.[3-7]

After the model is trained and deployed to the local environment, the user can start using the model prediction in the workflow (Figure 2, right column), minimizing manual adjustments for the loaded samples.

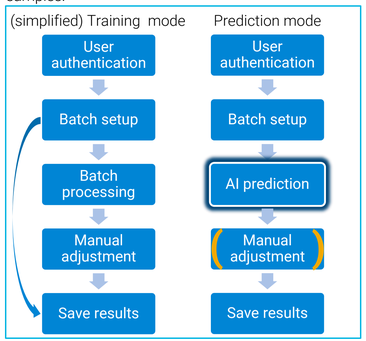

Figure 2. Training and prediction mode.

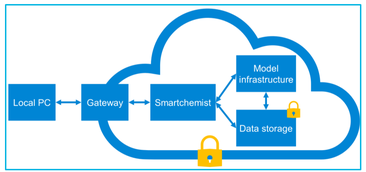

Figure 3. Simplified pipeline.

Results and Discussion

New Features (AI Flag & AI Confidence score)

The new AI prediction flag proposal enables users to easily identify how each peak was integrated: by the built-in integrator (Scenario I), by the AI model (Scenario II), or manually (Scenario III), as illustrated in Figure 4. Both the original MI flag and the AI prediction flag are displayed for each individual peak. In contrast, the AI Confidence score provides an overall assessment of the confidence at the compound level.

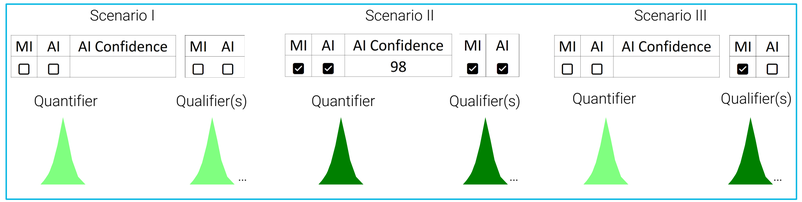

Figure 4. Schema of AI Confidence score and AI flag.

Model Prediction Accuracy

The hallmark of a well-trained machine learning (ML) model is its ability to provide reliable and accurate predictions consistently. Achieving this level of performance allows the model to streamline data analysis, reduce costs, and boost laboratory throughput. Figure 5 demonstrates an excellent example of how a properly trained model can handle tricky peak integrations, which otherwise usually need manual integration from the user after applying the integration from the built-in integrator of choice.

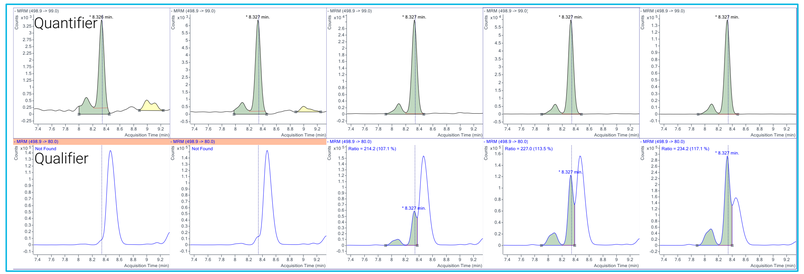

Figure 5. MRM chromatograms for the quantifier (top row) and qualifier (bottom row) transition of PFOS, with increasing concentration from left to right. AI Peak Integration results (marked as dark green) compared with the original results (dashed red line) from the built-in integrator.

Results and Discussion

Model performance

The results of the prediction time for a batch with 2, 5, 10, 25, 50 and 100 samples are summarized in Figure 6. The average prediction time per sample was about 5 to 6 seconds. The prediction time didn't include the data upload and updating integrations in MassHunter. The conventional approach requires on average 60 to120 seconds per sample (illustrated in Figure 6, orange area).

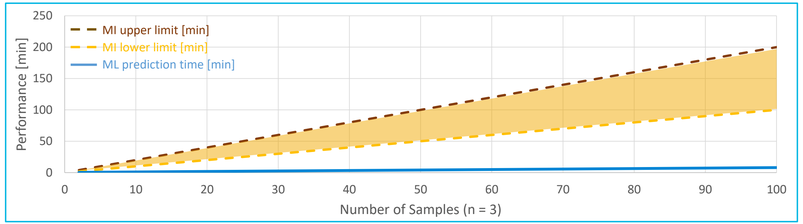

Figure 6. AI Peak Integration processing speed vs. number of samples per batch (n = 3). MI: manual integration, ML: machine learning.

Model Training and Validation

Figure 7 shows the positive trend in the Peak Screening Correctness Metrics during the training and validation phase. All metrices including F1 score, Positive Predictive Value (PPV) and Negative Predictive Value (NPV) were stabilized and above 0.95 after 20 epochs for the implemented model.

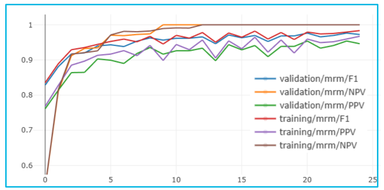

Figure 7. Schema of AI Confidence score and AI flag.

Conclusions

- A deep learning-assisted training and prediction procedure was successfully evaluated for the LC-MS/MS MRM PFAS data analysis, reducing manual integration to a minimum.
- The trained model can provide reproducible and reliable peak prediction.
- The following peak integration challenges can be handled by the trained model:
  - Early eluting PFAS with peak trailing/bordering
  - Linear and branched PFAS
  - False positive and negative peaks due to matrix interferences or contamination
- The data analysis time can be dramatically reduced using the trained model.

References

[1]EPA method 1633: Analysis of Per- and Polyfluoroalkyl Substances (PFAS) in Aqueous, Solid, Biosolids, and Tissue Samples by LC-MS/MS; 4th draft. United States Environmental Protection Agency, December 2024

[2]AOAC (2023) Standard Method Performance Requirements (SMPRs) for Per- and Polyfluoroalkyl Substances (PFAS) in Produce, Beverages, Dairy Products, Eggs, Seafood, Meat Products, and Feed (AOAC SMPR 2023.003).

[3]arXiv:1505.04597

[4]arXiv:1910.11162

[5]Perslev, M. et al. npj Digit. Med. 4, 72 (2021)

[6]arXiv:1803.01271

[7]arXiv:2105.15203

DE-008622