DreaMS-PFAS: A Novel Transformer-Based Approach for Early Detection of PFAS Contamination in Drinking Water and Environmental Samples

Ramanathan, Hamsini (School: Seattle Academy of Arts and Sciences)

Per- and polyfluoroalkyl substances (PFAS) are synthetic chemicals with strong carbon-fluorine bonds that resist breakdown in the environment and are associated with serious health effects, including cancer, immune suppression, and liver damage. Approximately 165 million Americans live in communities where drinking water has tested positive for PFAS contamination, highlighting the need for scalable PFAS detection methods. Existing PFAS detection techniques depend on either targeted screening, which is limited to select known PFAS, or non-targeted screening (NTS), which requires labor-intensive expert annotation and cannot detect emerging PFAS. In this work, DreaMS, a pre-trained transformer-based model for tandem mass spectrometry data, was fine-tuned on a large dataset (737,876 spectra) combining MassSpecGym, NIST20, and NIST-PFAS by appending a binary classification head to detect known and emerging PFAS. The resulting model, DreaMS-PFAS, achieved 93.8% precision and 89% recall (5 runs) on a molecule-level Murcko scaffold validation split, exceeding rule-based baselines by 4× in recall (PFAScreen recall: 22.1%). Model predictions are automated and run in ~60 ms per spectrum costing ~$3–$4 per 100,000 spectra on a single NVIDIA GPU. DreaMS-PFAS identified nine high-confidence putative PFAS hits from 327,130 spectra in a public MassIVE dataset of reef and marine organism samples from Moorea, French Polynesia that were not detected by PFAScreen. All nine candidates showed negative mass defects with four exhibiting characteristic PFAS fragmentation patterns, suggesting novel PFAS hits. These results demonstrate that DreaMS-PFAS can enable rapid, large-scale PFAS screening at low cost, prioritizing high-confidence candidates for laboratory validation.

Awards Won:
Fourth Award of $600