Poster Reprint

ASMS 2025
Poster number WP 113

De Novo PFAS Annotation and Classification Using Highly Accurate Formula Prediction and Kaufmann Algorithms Embedded in FluoroMatch Suite

Jeremy Koelmel[1]; Michael Kummer[2]; David Schiessel[2]; Olivier Chevallier[3]; David Godri[4]; Christian Klein[3]; Emma E Rennie[3]; Krystal J Godri Pollitt[1]

[1]Yale University, New Haven, CT;

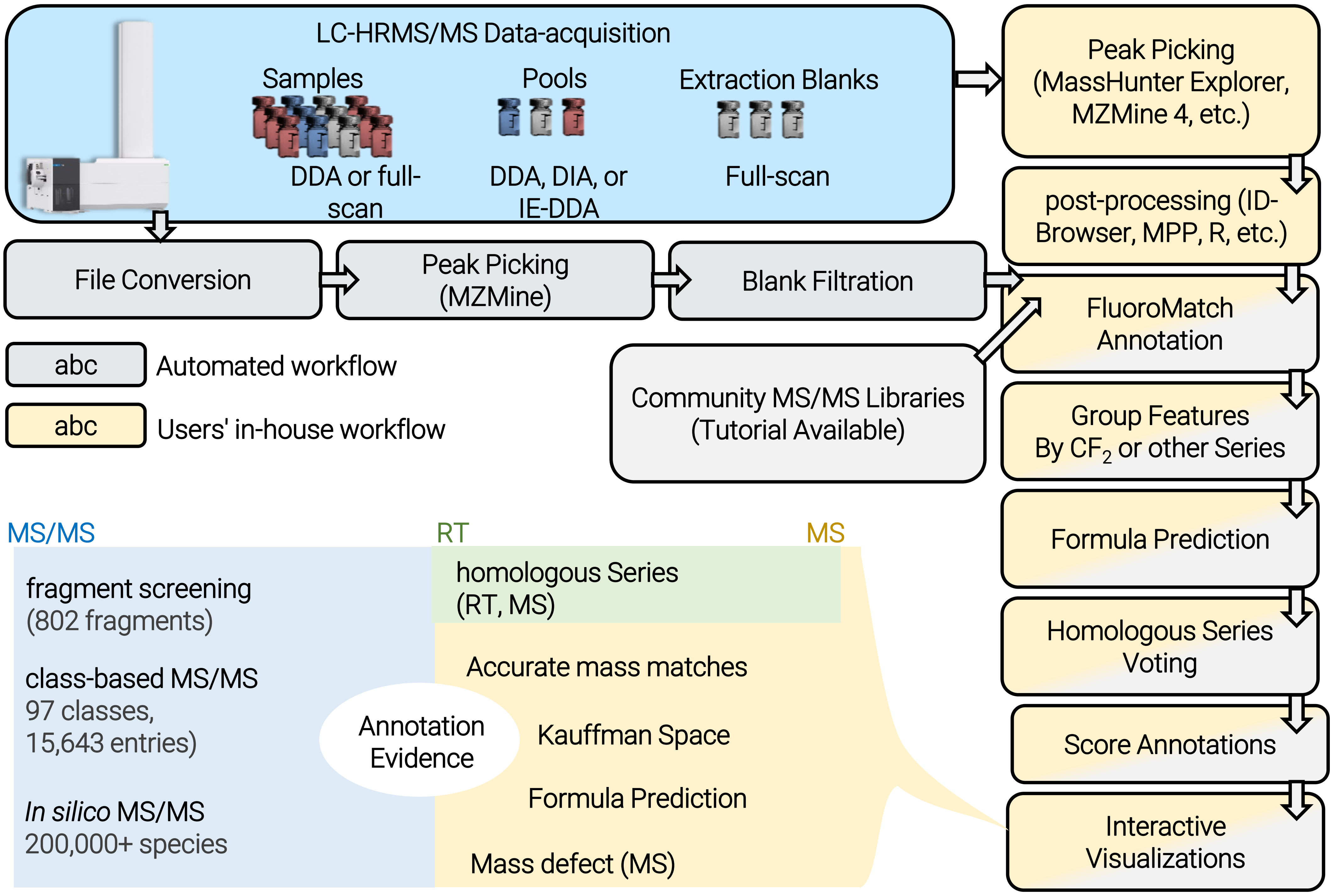
[2]Innovative Omics, Sarasota, FL;

[3]Agilent Technologies, Santa Clara, CA;

[4]3rd Floor Solutions, Caledon, ON

Overview: Intro, Methods, Results, and Conclusion

FluoroMatch Suite MS1 Only Algorithms: FluoroMatch Suite covers the entire non-targeted PFAS workflow **(Figure 1)** using MS/MS algorithms and the largest fragmentation database to date providing a <5% false positive rate. However, MS/MS coverage is often incomplete, and DIA methods often provide lower quality spectra after deconvolution. Therefore, we developed MS1 only algorithms for providing highly accurate formula predictions (**Figure 2**) and for classifying features as PFAS or not PFAS using Kaufmann analysis (**Figure 3**)
Validation: Algorithms were validated on NIST A and NIST C samples from the NIST PFAS interlaboratory study. NIST A consisted of known PFAS chemical standards, and NIST C consisted of an aqueous fire fighting foam (AFFF) contaminated soil sample. **Formula prediction with homologous series voting** (taking the most common repeating formula) **was extremely accurate for both samples with a 0% false positive rate**, and without homologous series voting the false positive rate was between 17-26% (**Table 1**). **Kaufmann Analysis was highly specific for classifying PFAS** with only 4% of non-PFAS being classified as PFAS with our Kaufmann based model, and only 6% of PFAS not be classified as PFAS according to our model for NIST C (**Table 2**).
Contact: **Jeremy.Koelmel@yale.edu with software questions, Krystal.Pollitt@yale.edu for research collaborations**
Download: **innovativeomics.com/software**



LC-HRMS/MS Data-acquisition

Samples
DDA or full-scan

Pools
DDA, DIA, or IE-DDA

Extraction Blanks
Full-scan

File Conversion

Peak Picking (MZMine)

Blank Filtration

abc Automated workflow

abc Users' in-house workflow

Community MS/MS Libraries (Tutorial Available)

Peak Picking (MassHunter Explorer, MZMine 4, etc.)

post-processing (ID-Browser, MPP, R, etc.)

FluoroMatch Annotation

Group Features By CF_2 or other Series

Formula Prediction

Homologous Series Voting

Score Annotations

Interactive Visualizations

MS/MS

fragment screening (802 fragments)

class-based MS/MS 97 classes, 15,643 entries)

In silico MS/MS 200,000+ species

RT

homologous Series (RT, MS)

MS

Accurate mass matches

Annotation Evidence

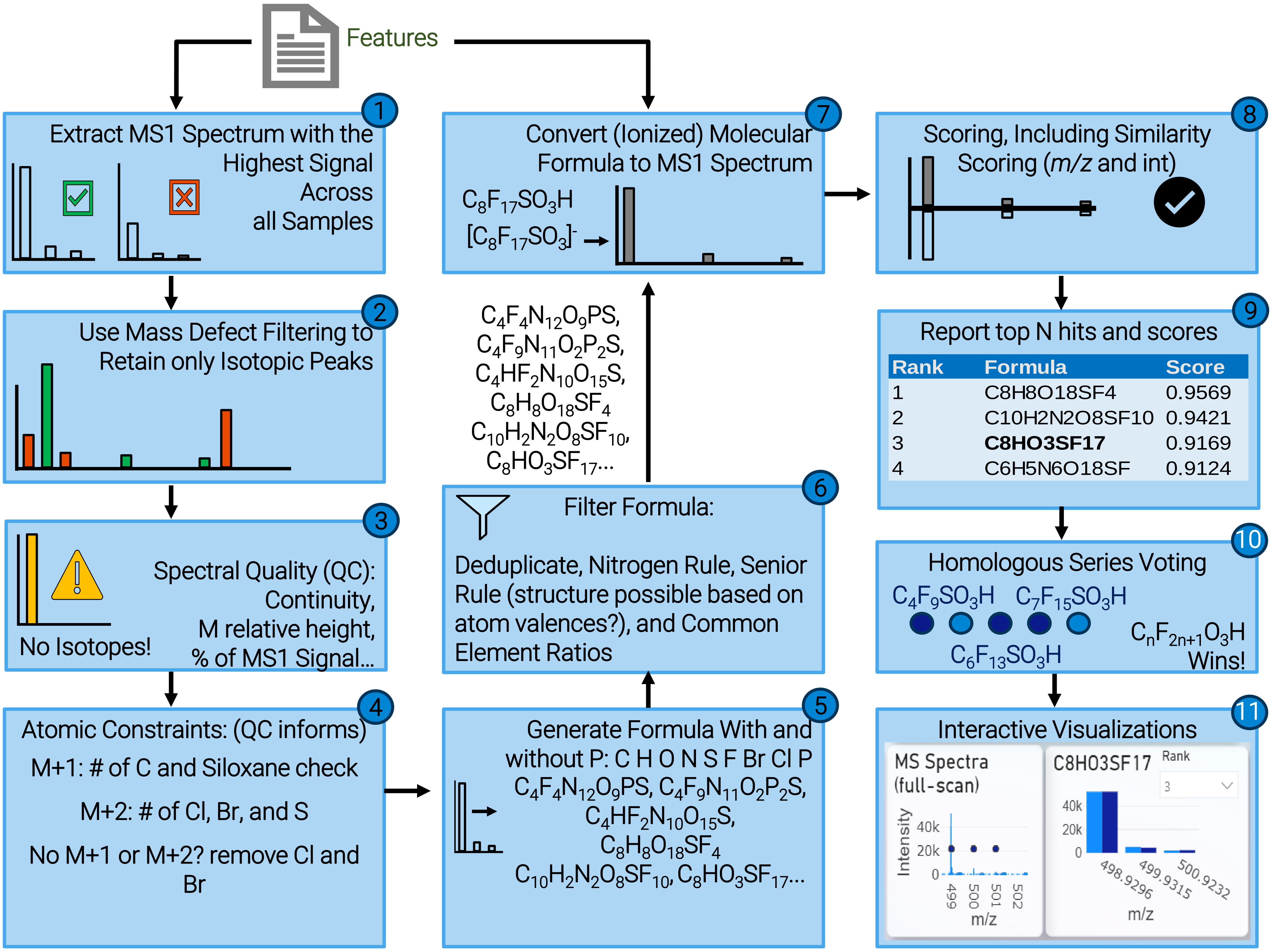
Kauffman Space

Formula Prediction

Mass defect (MS)

Figure 1. FluoroMatch Suite user and algorithm workflow: Input files accepted include all ions, data-dependent, intelligent data acquisition, and full-scan data from any vendor or mzML formatted file. After dragging files onto the software and choosing parameters (defaults provided for each vendor in a drop-down list), peak picking, blank filtering, annotation, homologous series detection, and PFAS classification is performed. Annotation consists of predicted MS/MS based-matching, rule-based MS/MS matching, accurate mass matching, and formula prediction. PFAS classification consists of Kaufmann analysis, homologous series detection, fragment screening, and mass defect filtering.

Formula Prediction Algorithm and Results



① Extract MS1 Spectrum with the Highest Signal Across all Samples

② Use Mass Defect Filtering to Retain only Isotopic Peaks

③ Spectral Quality (QC): Continuity, M relative height, % of MS1 Signal…

No Isotopes!

④ Atomic Constraints: (QC informs)

M+1: # of C and Siloxane check

M+2: # of Cl, Br, and S

No M+1 or M+2? remove Cl and Br

⑤ Generate Formula With and without P: C H O N S F Br Cl P

$C_4F_4N_{12}O_9PS$, $C_4F_9N_{11}O_2P_2S$, $C_4HF_2N_{10}O_{15}S$, $C_8H_8O_{18}SF_4$ $C_{10}H_2N_2O_8SF_{10}$, $C_8HO_3SF_{17}$…

⑥ Filter Formula:

Deduplicate, Nitrogen Rule, Senior Rule (structure possible based on atom valences?), and Common Element Ratios

$C_4F_4N_{12}O_9PS$, $C_4F_9N_{11}O_2P_2S$, $C_4HF_2N_{10}O_{15}S$, $C_8H_8O_{18}SF_4$ $C_{10}H_2N_2O_8SF_{10}$, $C_8HO_3SF_{17}$…

⑦ Convert (Ionized) Molecular Formula to MS1 Spectrum

$C_8F_{17}SO_3H$

$[C_8F_{17}SO_3]^- \rightarrow$

⑧ Scoring, Including Similarity Scoring (*m/z* and int)

⑨ Report top N hits and scores

Rank	Formula	Score
1	C8H8O18SF4	0.9569
2	C10H2N2O8SF10	0.9421
3	**C8HO3SF17**	0.9169
4	C6H5N6O18SF	0.9124

⑩ Homologous Series Voting

$C_4F_9SO_3H$ $C_7F_{15}SO_3H$

$C_6F_{13}SO_3H$

$C_nF_{2n+1}O_3H$ Wins!

⑪ Interactive Visualizations

Figure 2. Formula prediction algorithm: The formula prediction algorithm selects the most intense MS1 spectra and filters the spectra using mass defect to only retain ions of interest. Poor quality spectra are excluded from formula prediction. Atomic constraints are set based on isotopic pattern and resulting formula are filtered using various criteria (senior rule, common element ratios, nitrogen rule, etc.). The resulting list of formula (after adding MS/MS and database matches from FluoroMatch) are scored based on various criteria including *m/z* and intensity similarity. Presence of repeating formula motifs in homologous series are used to more accurately assign an entire series formula using "homologous series voting".

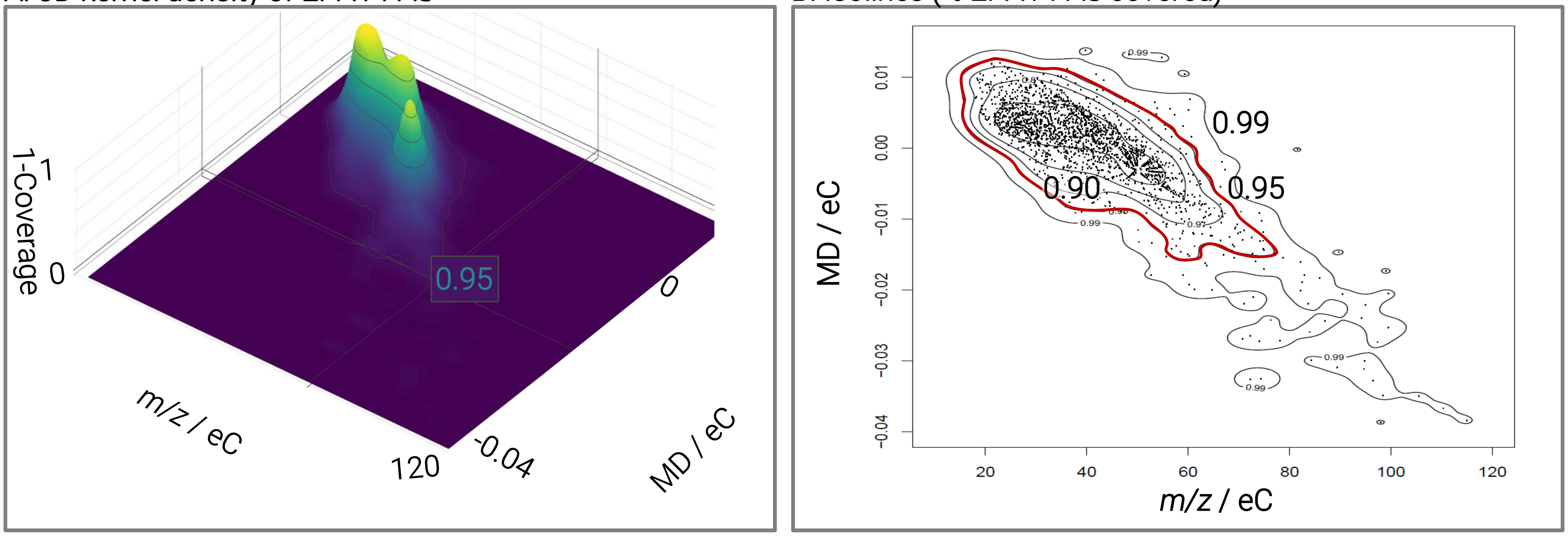
	n	False Positive Rate	False Positive Rate (With Homologous Series Voting)	False Negative Rate	False Negative Rate (With Homologous Series Voting)
NIST A	12	17%	0%	0%	17%
NIST C	151	26%	0%	30%	6%

Table 1. False positive and negative rates for formula prediction algorithm with and without homologous series voting for NIST A (chemical standards) and NIST C (AFFF contaminated soil). The false negative rate is the number of validated species which had no corresponding formula predicted (due to spectral quality issues, or in the case of homologous series voting, no consistent formula motif for the homologous series).

Kaufmann Analysis Algorithm and Results

A. 3D kernel density of EPA PFAS

B. Isolines (% EPA PFAS covered)



C. Kaufmann plot of confirmed PFAS

D. Kaufmann plot of unlikely PFAS

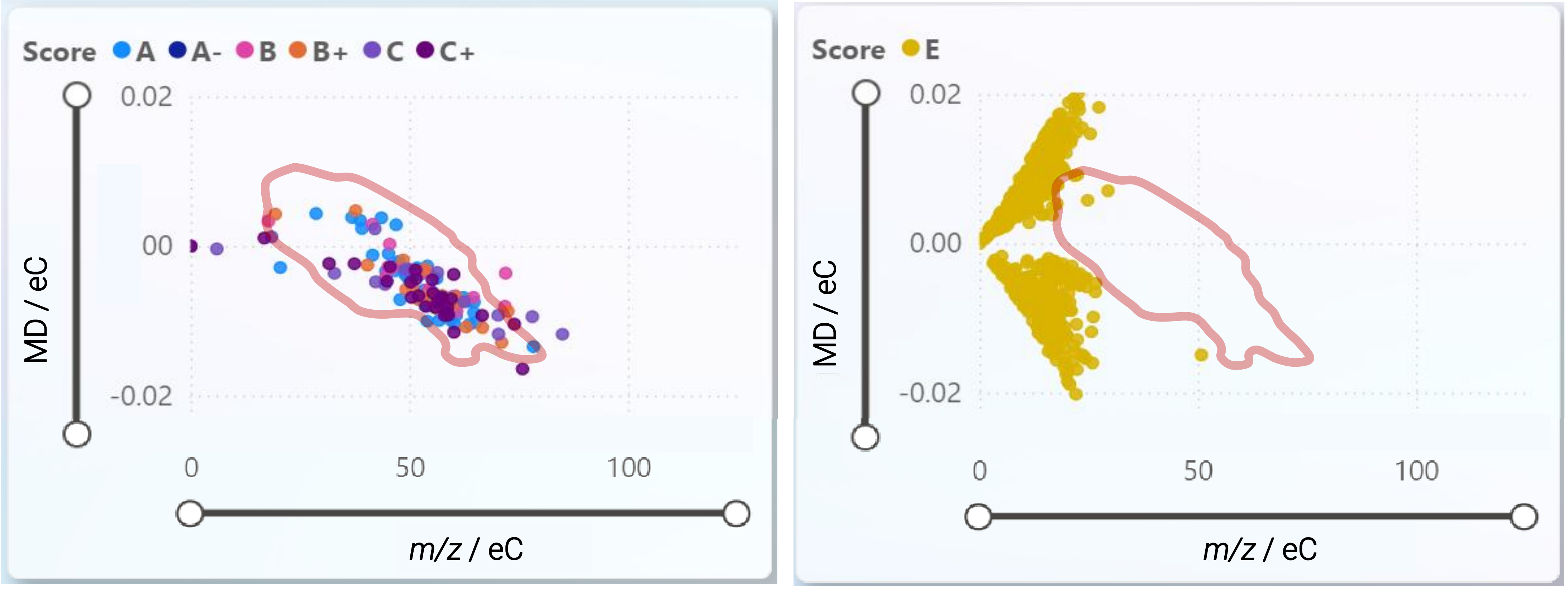


Figure 3. Kaufmann Classification: Kaufmann based filtering of NIST C using isolines established from kernel density plots of a known PFAS database to determine likely PFAS. A) Kaufmann kernel density plots using the EPA PFAS dataset (z axis converted to 1 - percent coverage). B) Isolines formed using the same EPA PFAS dataset. C) High confidence PFAS feature from NIST C with the 95% isoline overlay (red). D) Likely non-PFAS from NIST C with the 95% isoline overlay (red).

	False Positive Rate (non-PFAS < 0.96)	False Negative Rate (PFAS > 0.95)	Missing Estimations (eC = 0)
NIST A	24%	0%	0%
NIST A (n)	244	12	12
NIST C	4%	6%	15%
NIST C (n)	905	129	151

Table 2. False positive and negative rates for Kaufmann based PFAS classification model: The percent of unlikely PFAS (FluoroMatch score E) which were within the Kauffman 95% isoline were used to calculate the false positive rate: assigned PFAS based on Kaufmann analysis which were likely not true PFAS. The percentage of confirmed PFAS which were outside the Kaufmann 95% isoline were used to calculate the false negative rate: confirmed PFAS which were not assigned as likely PFAS by Kaufmann analysis. Kaufmann analysis could not be performed without an M+1 peak detected; in which case these features were considered as "missing estimations".

https://www.agilent.com/en/promotions/asms

DE-006597

Published in USA, May 15, 2025