Naïve Bayes classification model for isotopologue detection in LC-HRMS data

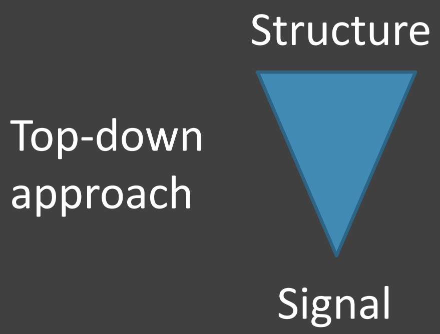
Denice van Herwerden, Jake O'Brien , Phil Choi, Kevin V. Thomas, Peter J. Schoenmakers, and Saer Samanipour

30 August 2022

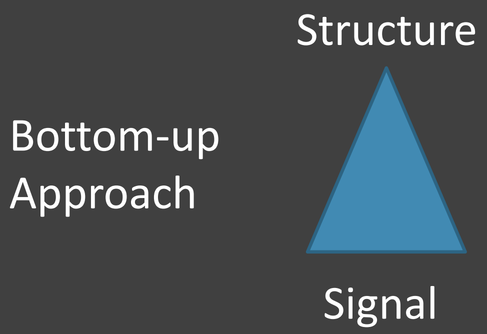
Monitoring

Target screening

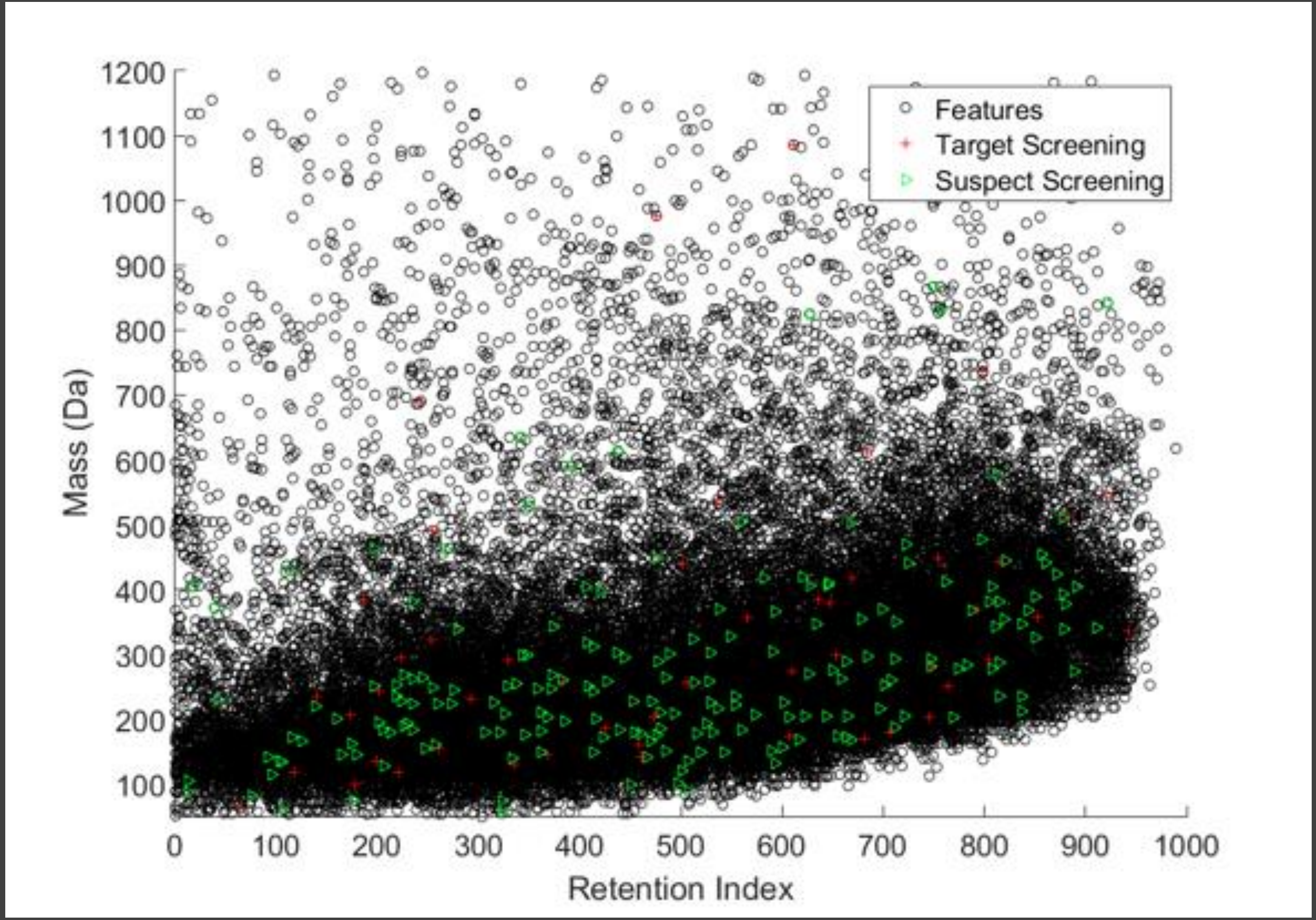
Suspect screening



Non-target screening



Non-targeted analysis (NTA)



Identify 'all' features

- Known & unknown compounds

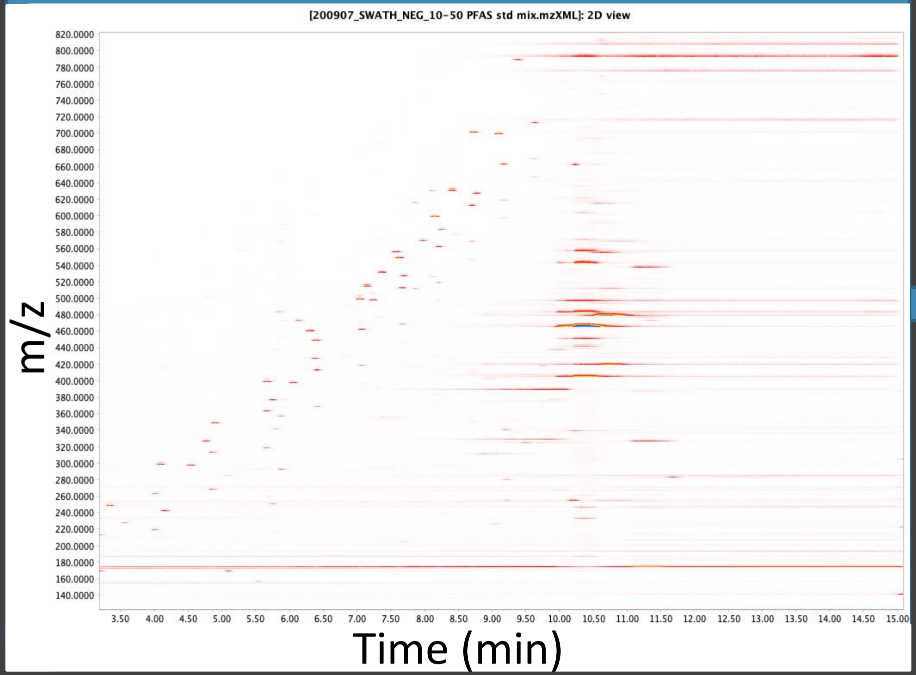
Structure

Bottom-up Approach

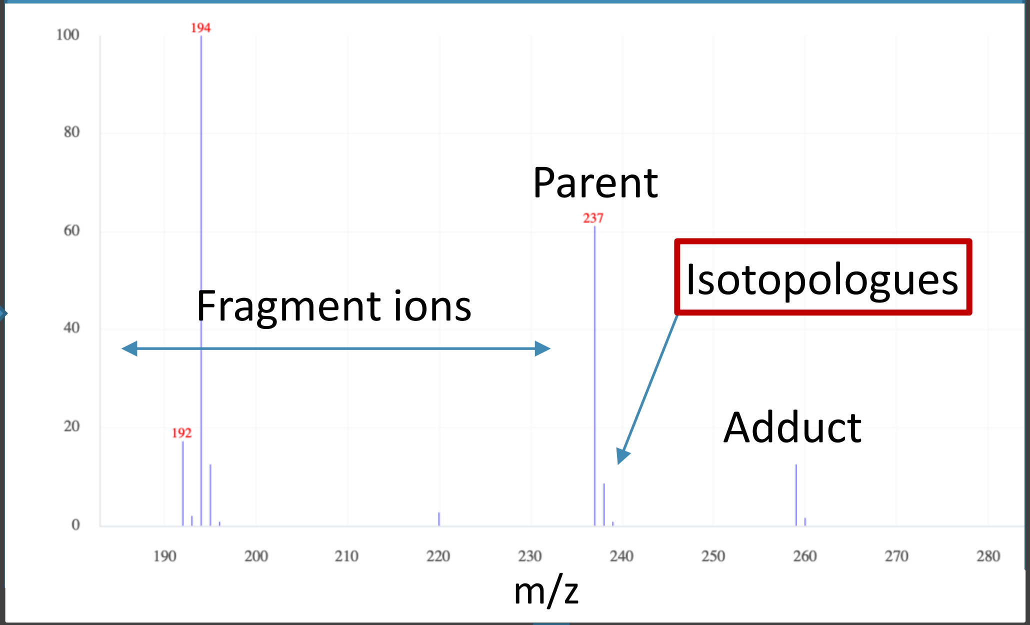
Signal

General NTA workflow

MS1 feature detection



Componentization



Identification

Current Isotopologue detection methods

PREDICTED ISOTOPE PATTERN

Requires:

- Information on the molecular formula

Limitation:

- NTA deals with known and unknown compounds
- Different formula can be assigned depending on the database

MASS DIFFERENCE OF 1.0033

Requires:

- Arbitrary mass tolerance

Limitation:

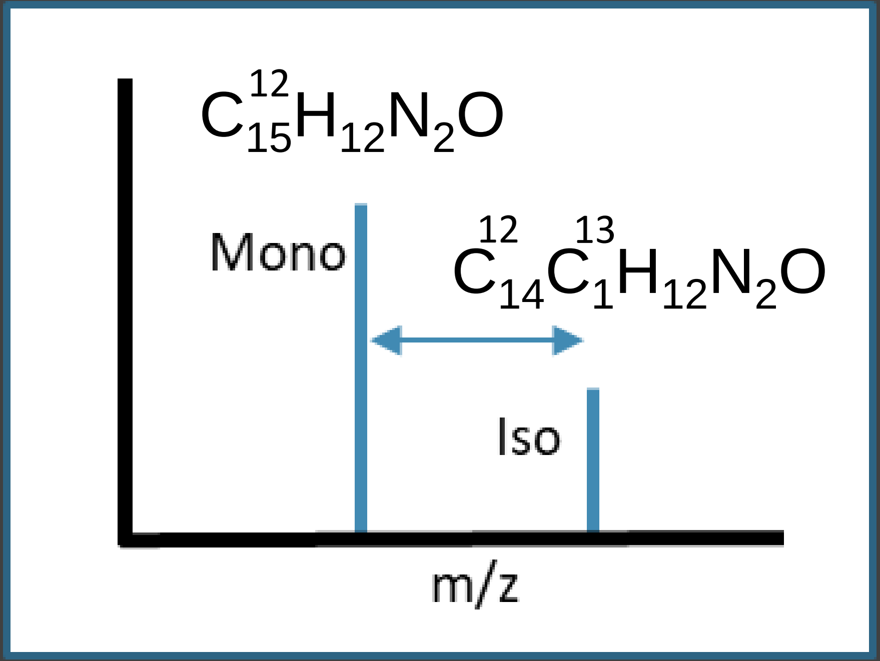
- Mass tolerance varies per instrument
- Can result in more false isotopologues when wrongly set

Isotopologue classification model

- No arbitrary thresholds needed
- No prior information required

Naïve Bayes isotopologue classification model

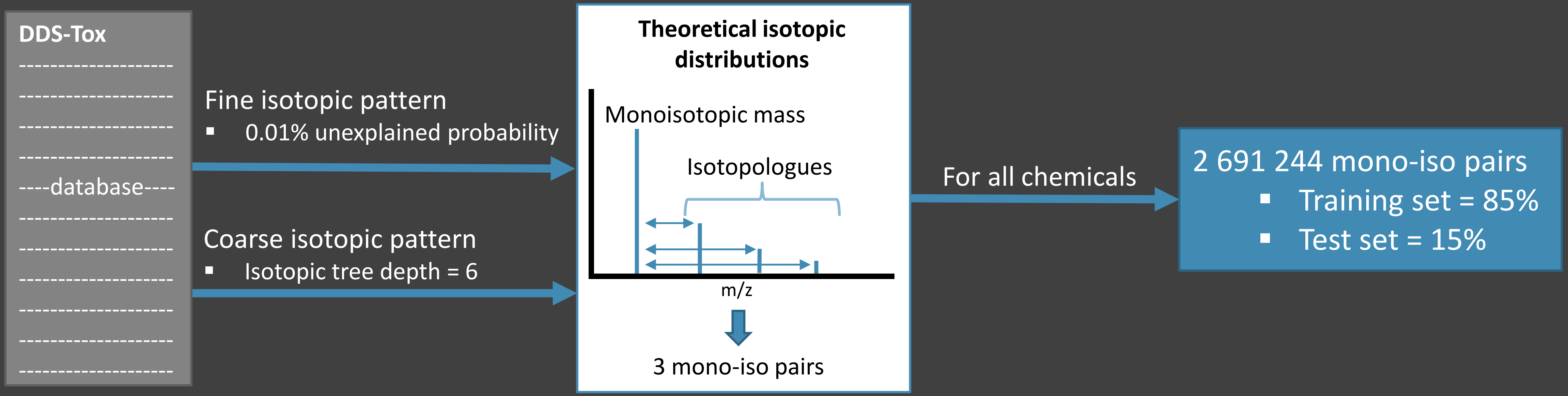
Theoretical assumption



Same molecular formula: $C_{15}H_{12}N_2O$

Same mass defect (MD)?

Theoretical isotopologue patterns



DDS-Tox

----database----

Fine isotopic pattern

- 0.01% unexplained probability

Coarse isotopic pattern

- Isotopic tree depth = 6

Theoretical isotopic distributions

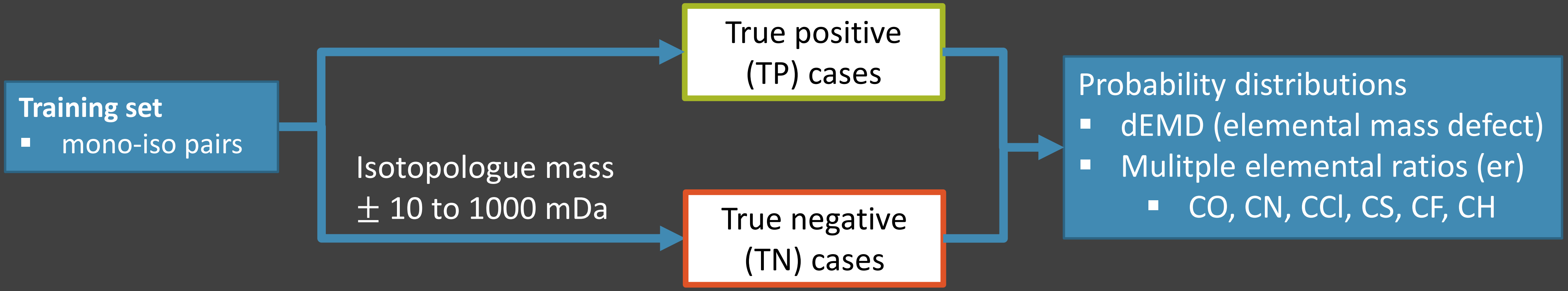
3 mono-iso pairs

For all chemicals

2 691 244 mono-iso pairs

- Training set = 85%
- Test set = 15%

Probabilistic classifier model setup



Training set

- mono-iso pairs

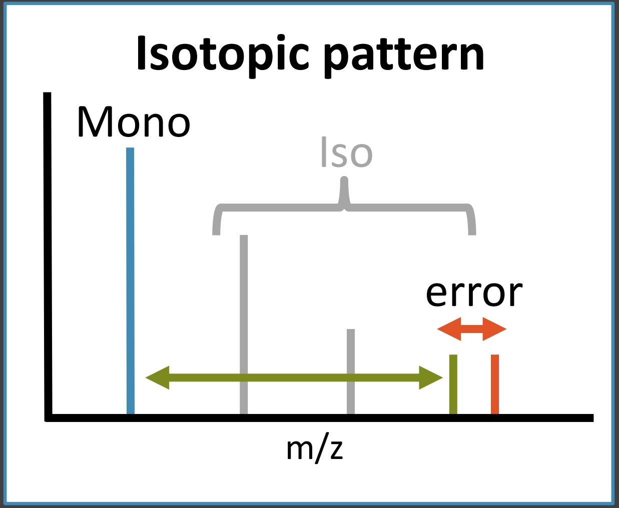
Isotopologue mass
± 10 to 1000 mDa

True positive (TP) cases

True negative (TN) cases

Probability distributions

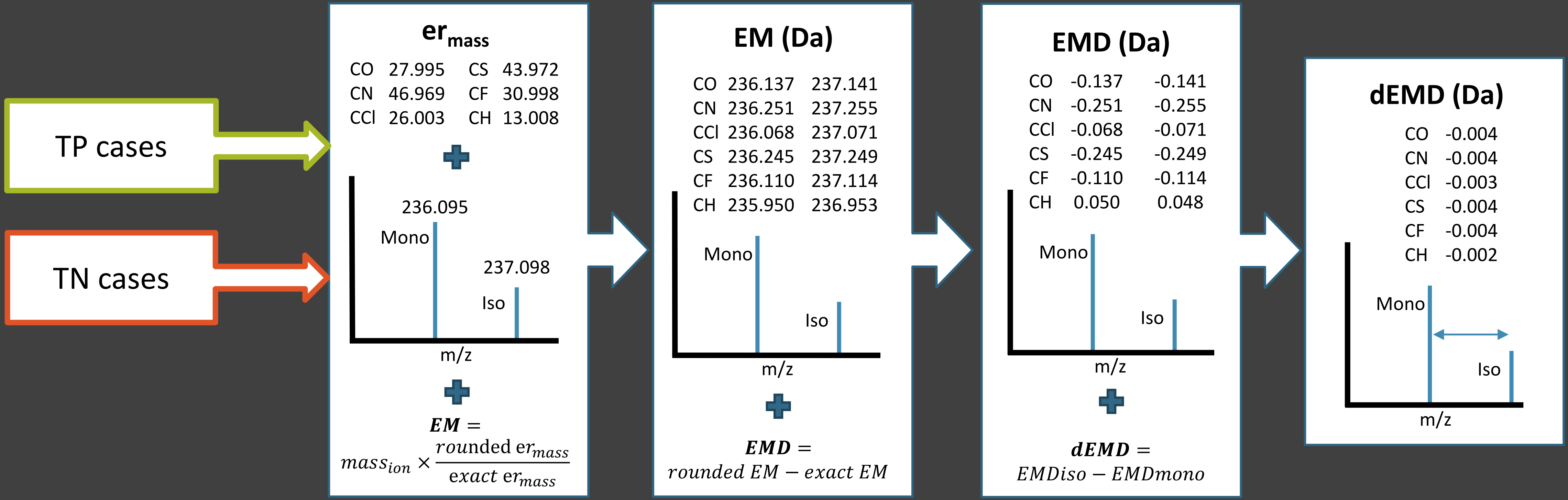
- dEMD (elemental mass defect)
- Mulitple elemental ratios (er)
    - CO, CN, CCl, CS, CF, CH



Isotopic pattern

dEMD calculation

Do $C^{12}_{15}H_{12}N_2O$ and $C^{12}_{14}C^{13}_1H_{12}N_2O$ have the same MD?



TP cases

TN cases

er$_{mass}$

CO	27.995	CS	43.972
CN	46.969	CF	30.998
CCl	26.003	CH	13.008

$$\boldsymbol{EM} = mass_{ion} \times \frac{\text{rounded } er_{mass}}{exact\ er_{mass}}$$

EM (Da)

CO	236.137	237.141
CN	236.251	237.255
CCl	236.068	237.071
CS	236.245	237.249
CF	236.110	237.114
CH	235.950	236.953

$$\boldsymbol{EMD} = rounded\ EM - exact\ EM$$

EMD (Da)

CO	-0.137	-0.141
CN	-0.251	-0.255
CCl	-0.068	-0.071
CS	-0.245	-0.249
CF	-0.110	-0.114
CH	0.050	0.048

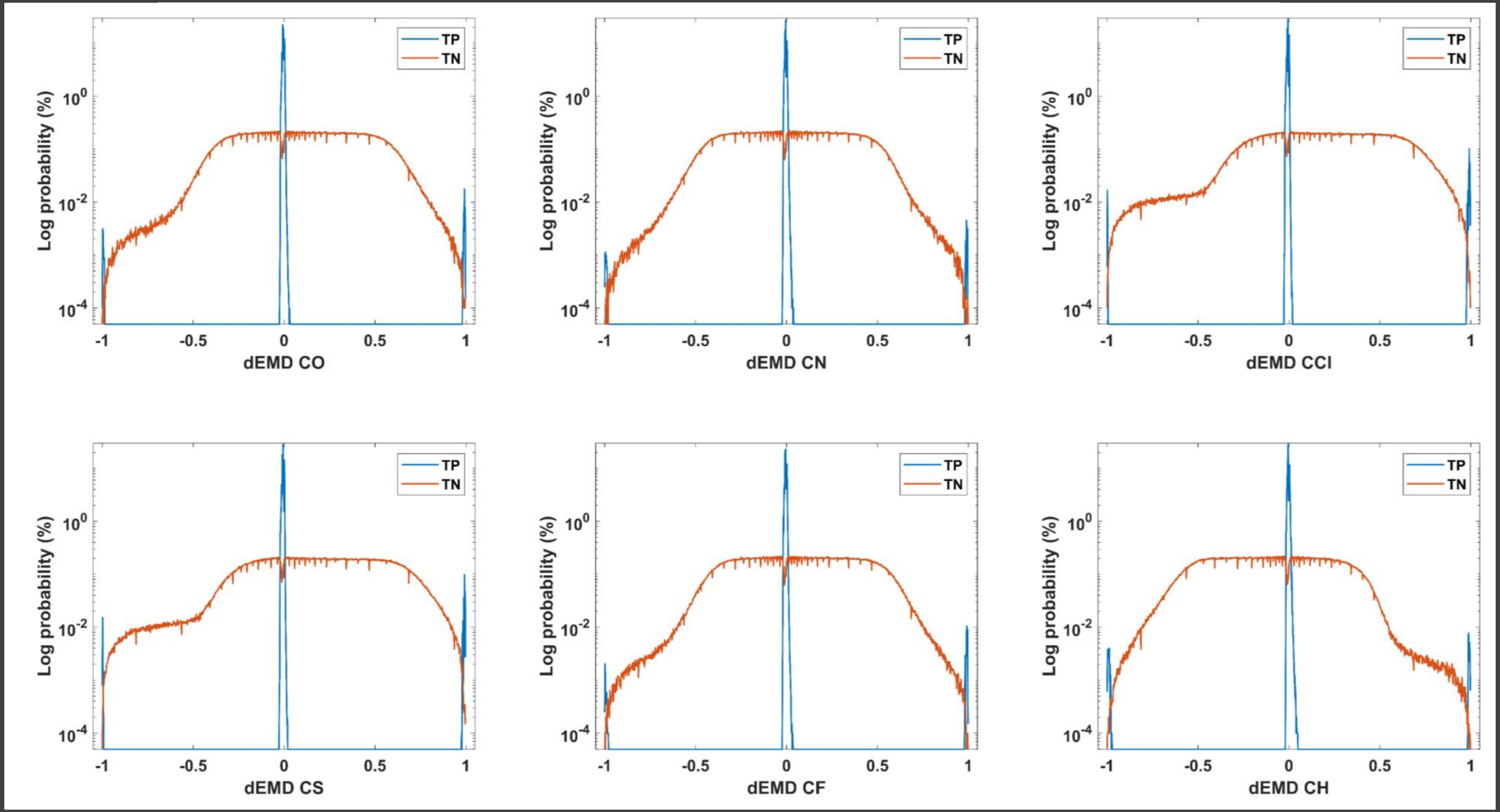
$$\boldsymbol{dEMD} = EMDiso - EMDmono$$

dEMD (Da)

CO	-0.004
CN	-0.004
CCl	-0.003
CS	-0.004
CF	-0.004
CH	-0.002

dEMD probability distributions

$$P(A|B) \propto \prod_{er} P(B_{er}|A)$$

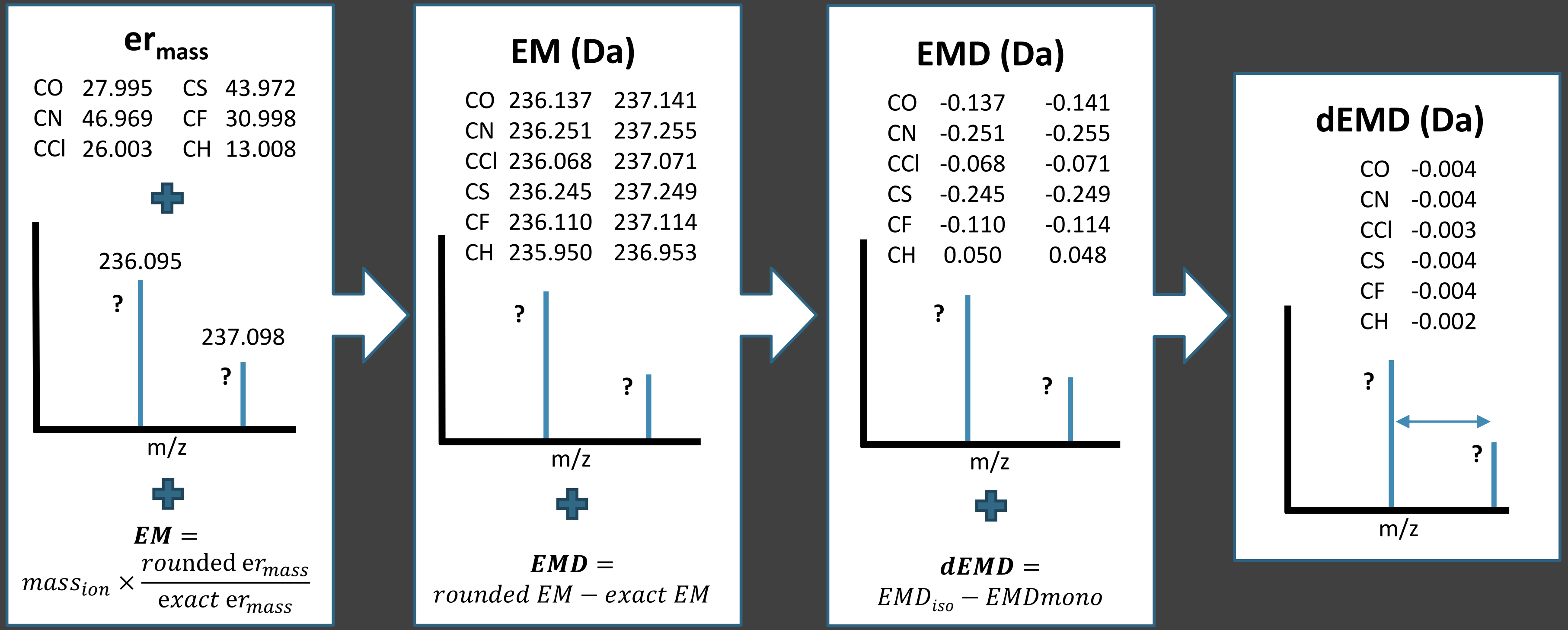


It is possible to differentiate between TP and TN isotopologues based on the dEMD values

Model usage

dEMD calculation - unknown

Do $C^{12}_{15}H_{12}N_2O$ and $C^{12}_{14}C^{13}_{1}H_{12}N_2O$ have the same MD? ✓



er_{mass}

CO	27.995	CS	43.972
CN	46.969	CF	30.998
CCl	26.003	CH	13.008

$$\boldsymbol{EM} = mass_{ion} \times \frac{rounded\ er_{mass}}{exact\ er_{mass}}$$

EM (Da)

CO	236.137	237.141
CN	236.251	237.255
CCl	236.068	237.071
CS	236.245	237.249
CF	236.110	237.114
CH	235.950	236.953

$$\boldsymbol{EMD} = rounded\ EM - exact\ EM$$

EMD (Da)

CO	-0.137	-0.141
CN	-0.251	-0.255
CCl	-0.068	-0.071
CS	-0.245	-0.249
CF	-0.110	-0.114
CH	0.050	0.048

$$\boldsymbol{dEMD} = EMD_{iso} - EMDmono$$

dEMD (Da)

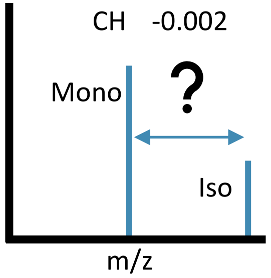
CO	-0.004
CN	-0.004
CCl	-0.003
CS	-0.004
CF	-0.004
CH	-0.002

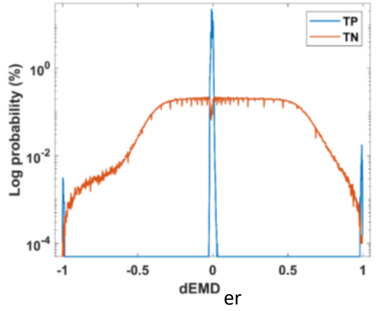
Isotopologue classification

$$P(A|B) \propto \prod_{er} P(B_{er}|A)$$

dEMD (Da)

CO	-0.004
CN	-0.004
CCl	-0.003
CS	-0.004
CF	-0.004
CH	-0.002





	pTP_{er} (%)	pTN_{er} (%)
CO	19.236	0.144
CN	10.065	0.157
CCl	27.169	0.131
CS	10.437	0.156
CF	23.039	0.140
CH	29.885	0.101

$$pTP = \prod pTP_{er}$$

$$= 3.78\text{e}{-05}$$

+

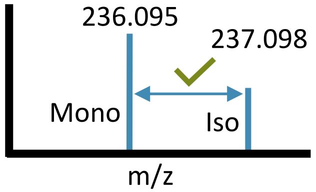
$$pTN = \prod pTN_{er}$$

$$= 6.52\text{e}{-17}$$

$$score_{EMD} = 1 - \frac{pTN}{pTP}$$

$$= 1 - 1.7256\text{e}{-12} \approx 1$$

$$score_{EMD} \geq \text{score threshhold}$$



Classification model performance

Classification model performance

Test set (15% cases)

- Obtained score$_{\text{EMD}}$ values

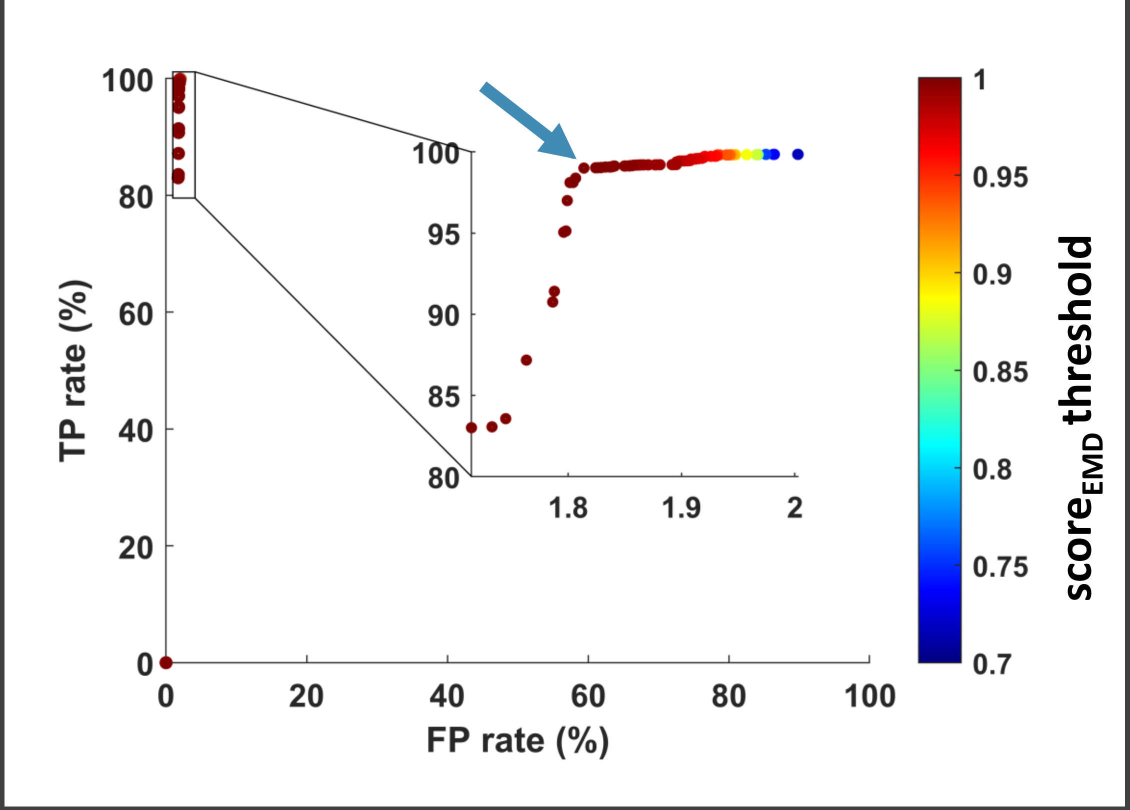
$$\textbf{T}\text{rue } \textbf{P}\text{ositive rate} = \frac{\mathrm{TP}}{\mathrm{FN+TP}} \cdot 100$$

$$\textbf{F}\text{alse } \textbf{P}\text{ositive rate} = \frac{\mathrm{FP}}{\mathrm{FP+T}N} \cdot 100$$

Optimal score$_{\text{EMD}}$ threhshold selected at 0.9997

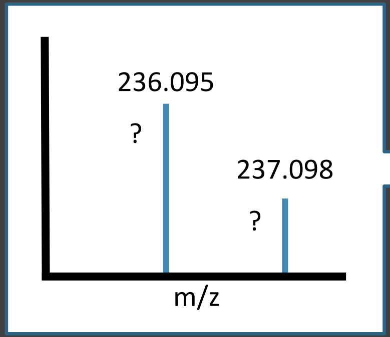
- TP$_{\text{rate}}$ = 99.0%
- FP$_{\text{rate}}$ = 1.8%

ROC curve



“In-house” mass difference method

Used by MZmine and CAMERA



$$remainder\left(\frac{mass_{iso} - mass_{mono}}{1.0033}\right)$$

$$remainder\left(\frac{237.098 - 236.095}{1.0033}\right) = 0.0003$$

Isotopologue detected when remainder < mass tolerance

Comparison with existing method

MASS DIFFERENCE METHOD

TP rate = 16.2 %   Mass tolerance = 0.1 mDa

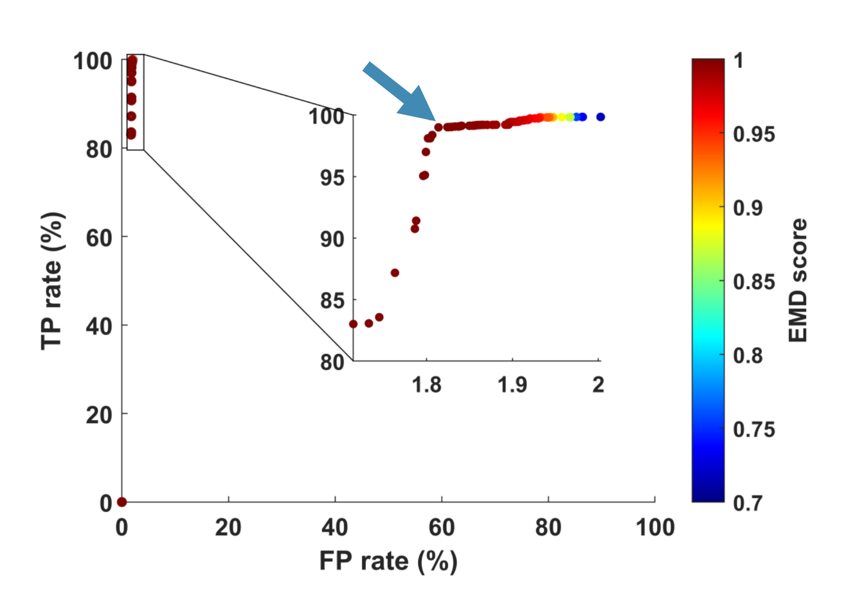
FP rate = 0.02 %

CLASSIFICATION MODEL

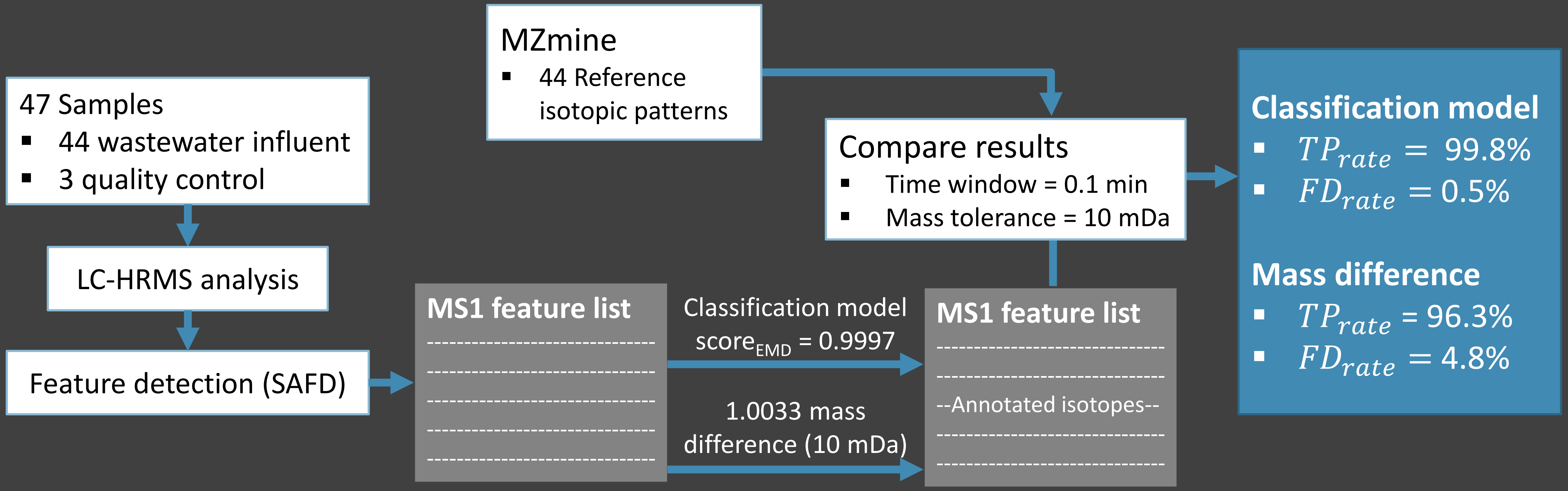
TP rate = 99.0 %   score = 0.9997

FP rate = 1.8 %

Classification model outperformed the “in-house” mass difference method for theoretical data



Performance for real samples



47 Samples

- 44 wastewater influent
- 3 quality control

LC-HRMS analysis

Feature detection (SAFD)

MZmine

- 44 Reference isotopic patterns

Compare results

- Time window = 0.1 min
- Mass tolerance = 10 mDa

MS1 feature list

Classification model score$_{EMD}$ = 0.9997

1.0033 mass difference (10 mDa)

MS1 feature list

--Annotated isotopes--

Classification model

- $TP_{rate} = 99.8\%$
- $FD_{rate} = 0.5\%$

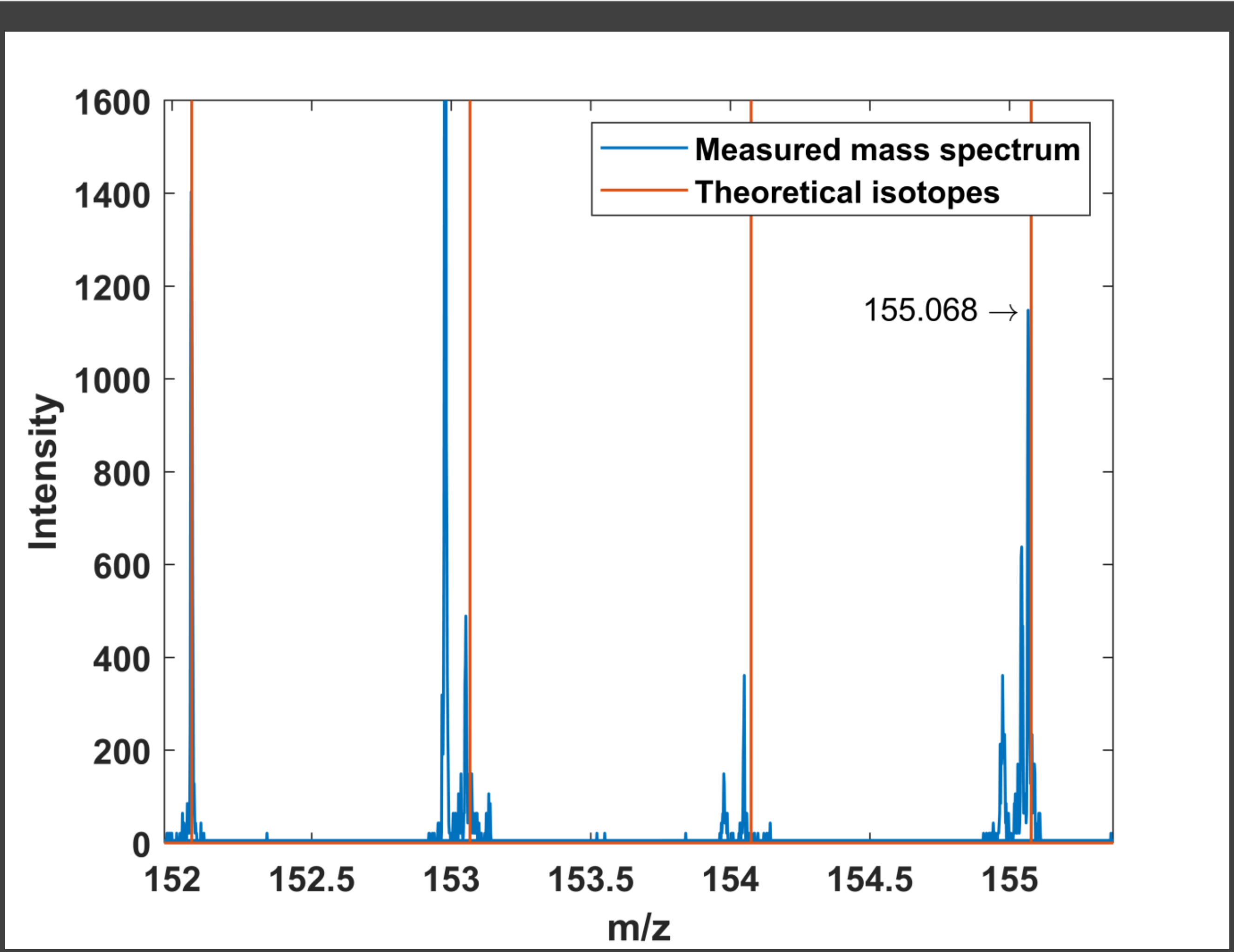
Mass difference

- $TP_{rate} = 96.3\%$
- $FD_{rate} = 4.8\%$

$$\text{True Positive rate} = \frac{\mathrm{FN}}{\mathrm{FN+TP}} \cdot 100$$

$$\text{False Detection rate} = \frac{\mathrm{FP}}{\mathrm{FP+TP}} \cdot 100$$

False detected cases



Preceding isotopologues (153 and 154) are either absent or have a lower intensity than the 155.068 peak

Conclusion & outlook

Naïve Bayes classification model based on elemental ratios can be used for the successful detection of isotopologues.

- Outperforming the state of the arts method
- Requiring no prior information on the molecular formula or an arbitrary threshold

Limitations

- Cannot distinguish between isotopologues coming from the same monoisotopic mass

Potentials

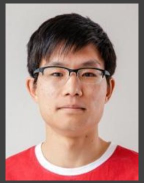
- Feature reduction for identification in NTA
- Assist in the correct molecular formula assignment

Environmental Modeling & Computational Mass Spectrometry

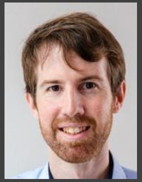
Saer Samanipour

Peter J. Schoenmakers

Kevin V. Thomas



Phil M. Choi



Jake W. O'Brien

Thank you!

Denice van Herwerden
d.vanherwerden@uva.nl

MAASTRICHT . THE NETHERLANDS

IMSC 2022

Paper

Algorithm on Bitbucket

UNIVERSITY OF AMSTERDAM

CASA

NVMS

CAST
Chemometrics and Advanced Separations Team