The Multidimensional Chromatography Workshop 2026

Structural elucidation using GCxGC-TOFMS and machine learning for unknown metabolites in HeLa cell

DAY 1 – TUESDAY January 13, 2026
1:50 - 2:10 PM, O-7

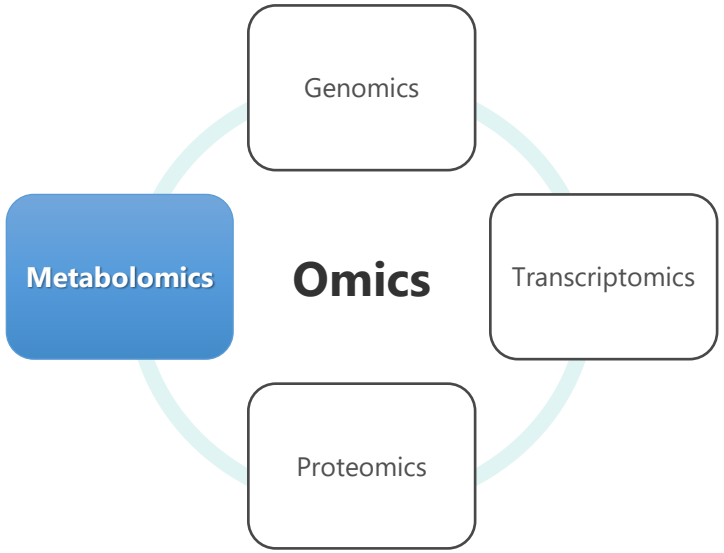
Masaaki Ubukata[1], Azusa Kubota[1],
Ayumi Kubo[1], Misaki Kurata[2], Hiroshi Tsugawa[2]

1. JEOL Ltd.
2. Tokyo University of Agriculture and Technology

What is Metabolomics?

What is Metabolomics?

- Metabolomics is the study of small chemical substances made in the body during life activities.
- It looks at many of these substances at the same time.
- It is the same as metabolome analysis.



What are Metabolites?

- Metabolites are small molecules made by living things inside the body.
- They can be divided into two groups: primary metabolites and secondary metabolites.

Primary Metabolites

- Primary metabolites are necessary for life activities.
- Examples: amino acids, sugars, and fatty acids.

Secondary Metabolites

- Secondary metabolites are unique to specific species.
- Examples: terpenoids and flavonoids.

GC-MS vs. LC-MS in Metabolomics

Feature	GC-MS (especially GC-TOFMS)	LC-MS
Analyte types	**Volatile and semi-volatile compounds**; stable derivatives	Broad polarity range, thermally labile compounds
Sample derivatization	Often required (e.g., **TMS derivatization**), but improves reproducibility	Usually not required
Chromatographic resolution	**High; excellent peak resolution** with capillary columns **GC x GC measurements**	Moderate; matrix effects common
Ionization variability	EI provides consistent ionization across compounds	Matrix effects can cause ion suppression
Spectral library support	**Extensive EI spectral libraries enable confident identification**	Limited; mainly exact mass
Reproducibility	**High; stable retention and fragmentation patterns**	Moderate; retention time drift across batches
Quantification accuracy	Highly reproducible; ideal for large cohort comparisons	Sensitive but often less reproducible
Structural elucidation	**EI fragmentation allows detailed structural interpretation**	Relies on MS/MS or database matching
Time-of-Flight (TOF) compatibility	GC-TOFMS offers high-speed, high-resolution acquisition	Common with QTOF-MS
Throughput	**High with fast GC**	High with short gradients
Data complexity	Cleaner spectra due to standardized EI fragmentation	Complex, especially with adducts and in-source fragments

GC-TOFMS offers highly reproducible and interpretable data, making it a powerful and underutilized tool in metabolomics, especially for non-targeted profiling with robust compound identification.

GC-MS Qualitative Analysis Workflow

- Sample prep.
- Pre-treatment:
  HS, TD, Py…

TD

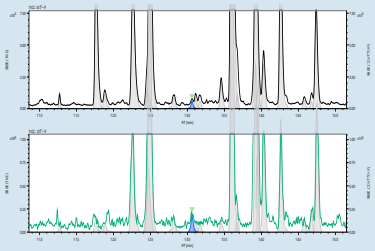
TD100-xr（MARKES）
Image provided by ENV Sciences Trading Co., Ltd.

HS-SPME

HT2850T (HTA)

Introduction
of a sample

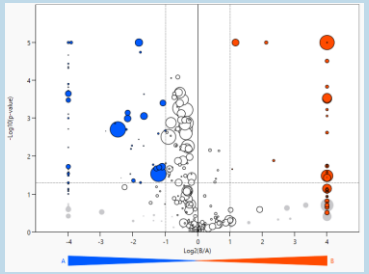
- GC/MS
- GCxGC/MS



GC/EI and GC/FI TICC

Separation & Detection
for compounds

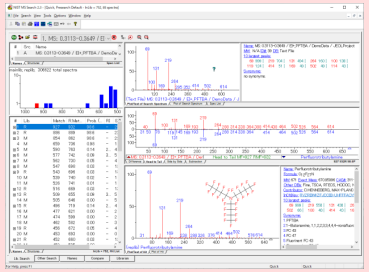
- Data Processing:
  Volcano plot, PCA



Volcano plot

Extraction
for specific compounds

- Qualitative Analysis:
  EI spectral DB search



The NIST Mass Spectral Search Program for the NIST/EPA/NIH Mass Spectral Library

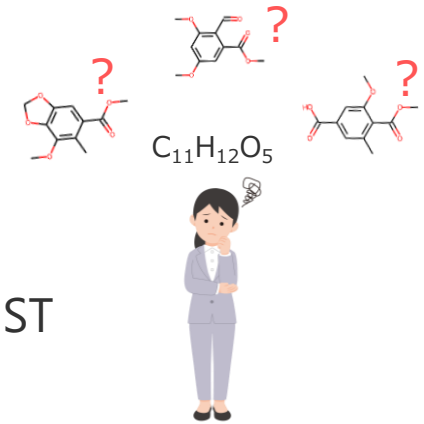
Identification
of specific compounds

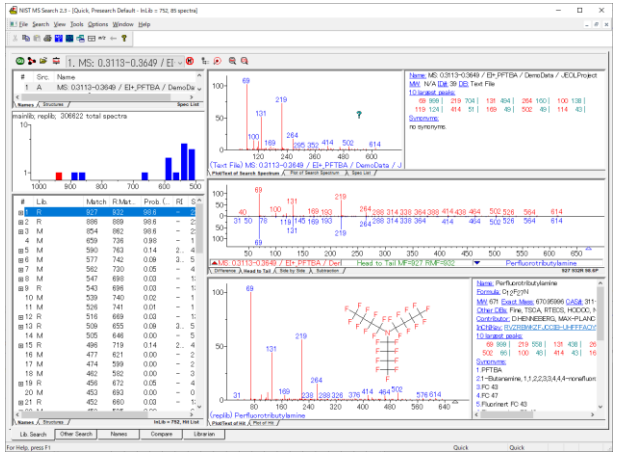
After having gone through all the steps above to find out that the compounds of your interest are not registered in the commercially available EI mass spectral database, how do you identify them?

Challenges in GC/MS Qualitative Analysis

How to identify the unknown compounds

which are not registered in the database developed by NIST?





[Source]
The NIST Mass Spectral Search Program for the NIST/EPA/NIH Mass Spectral Library

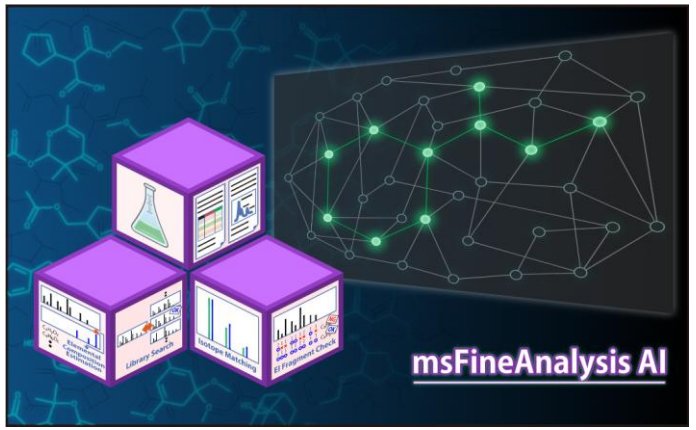
- EI mass spectrum database (DB) developed by NIST
- The latest DB contains 347,100 compounds
- More than 100 million compounds in PubChem
- Almost all compounds are not listed in the DB
- Only 0.3% of all compounds registered in the DB
- In many cases, they become “unknown compounds”
- Need time and knowledgeable for structure analysis

We made a predicted EI mass spectrum database by machine learning

Advanced Non-Targeted GC-MS Analysis Using AI-Based Structural Elucidation

Integrating GC-HRTOFMS and Machine Learning for Predictive EI Spectra Matching

GC-HRTOFMS:
JMS-T2000GC "AccuTOF™ GC-Alpha"



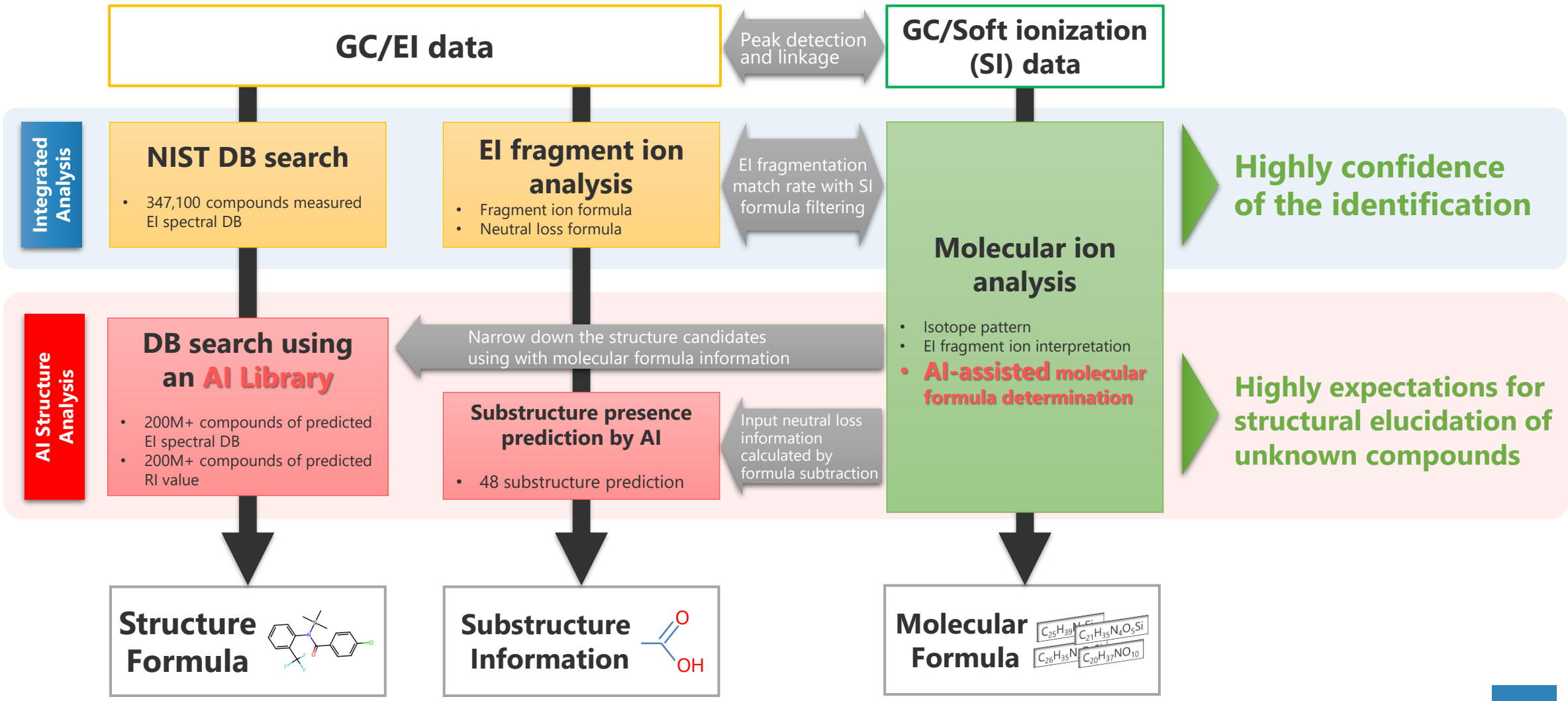
AI-Based Software:
"msFineAnalysisAI"

- Resolving power: >30,000 (FWHM @ *m/z* 614)
- Mass accuracy: ±1 ppm
- Ionization: EI, CI, PI, FI, and FD

- 200M+ compound predicted EI spectra by AI
- Combines EI and soft ionization data
- GC x GC data processing supported

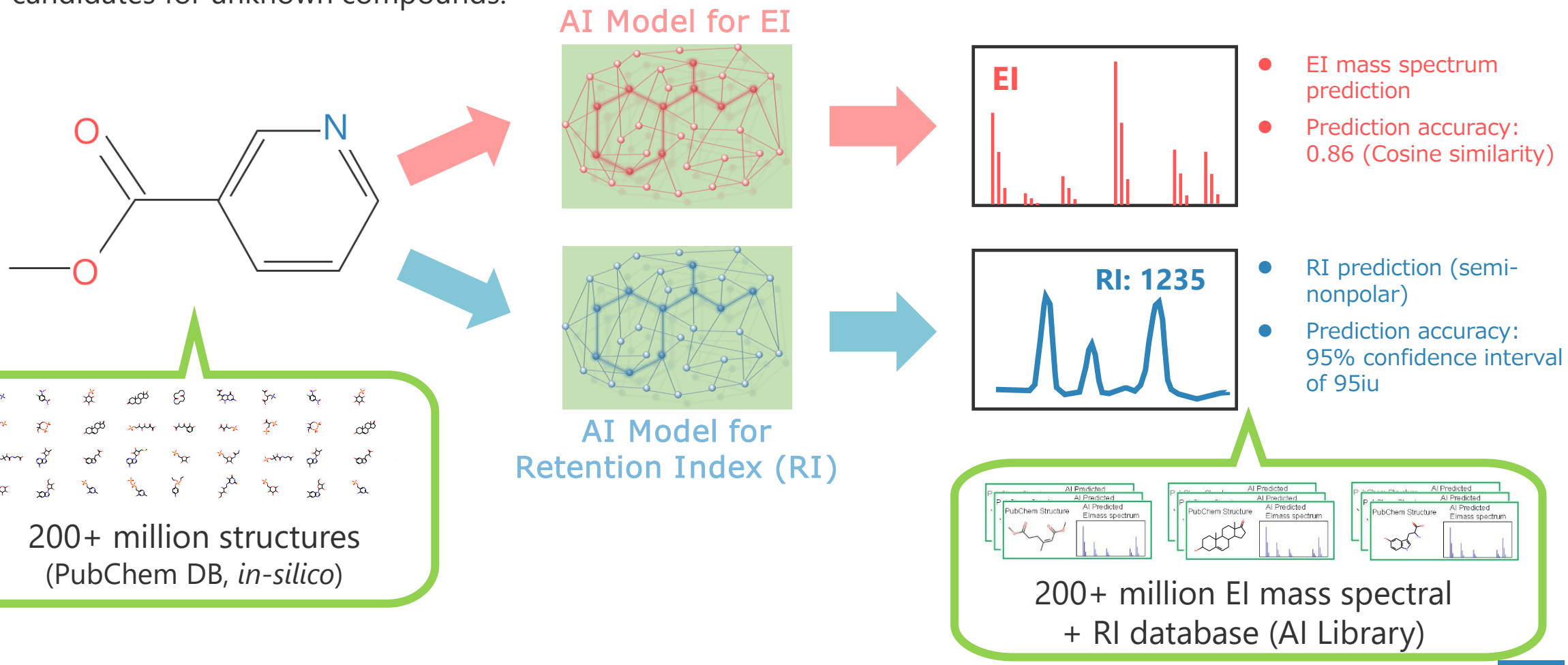
msFineAnalysis AI Workflow

- msFineAnalysis AI handle two GC/MS data, then to offer structure and molecular formula information



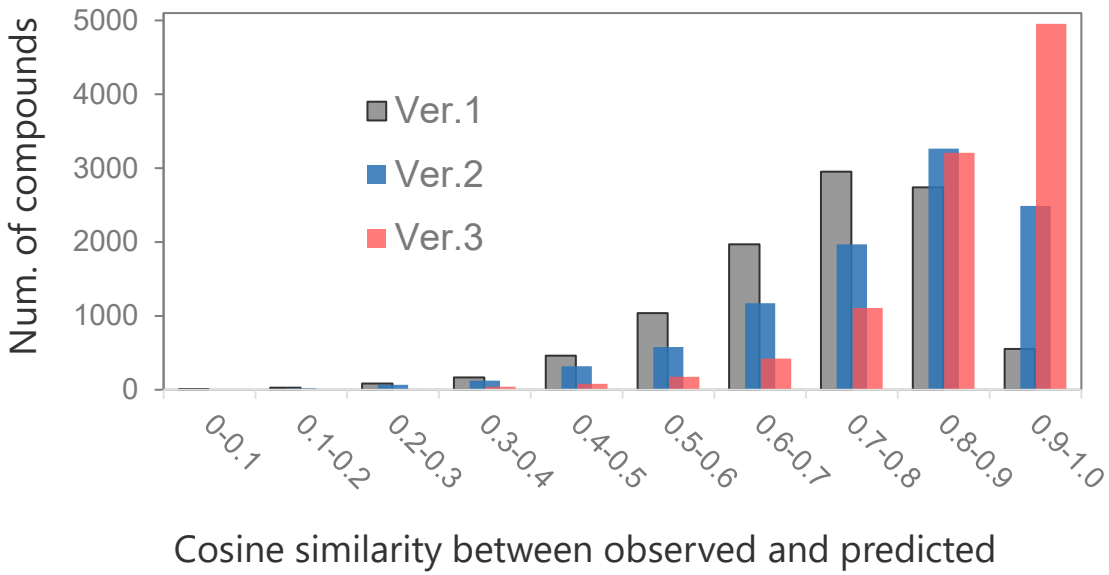
AI Model to Predict EI Mass Spectrum from Structural Formula

- An AI model was developed to predict EI mass spectra from chemical structures using machine learning.
- The AI model was then used to create over 200 million EI mass spectral database for determining the structure candidates for unknown compounds.

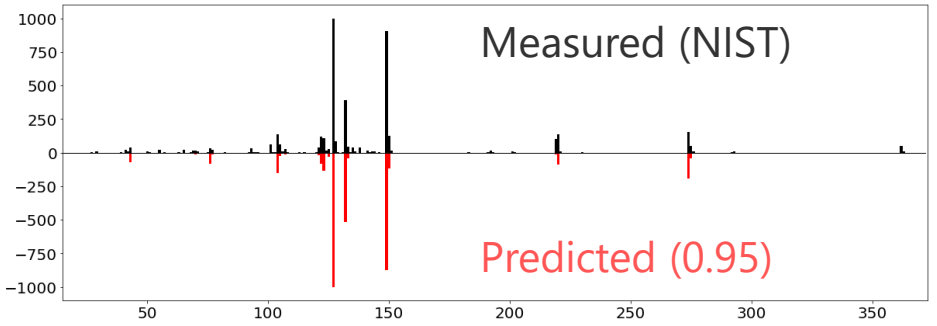


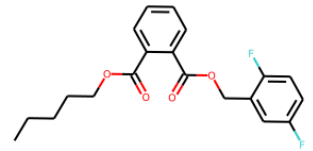
＊A. Kubo et al, *Mass Spectrometry*, **2023**, *12*, A0120.

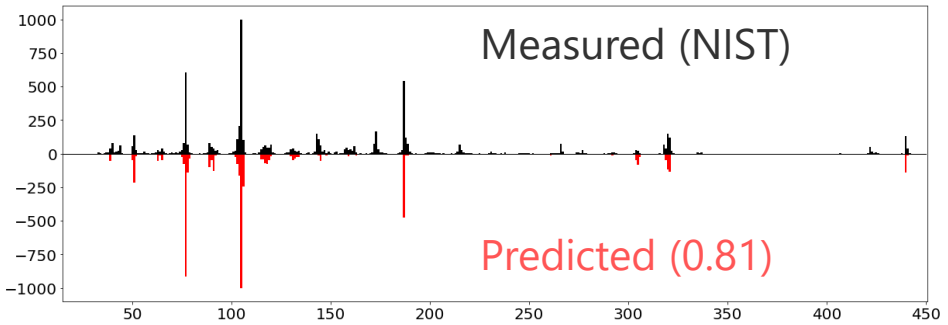
AI Model Performance: Cosine Similarity using Evaluation Data

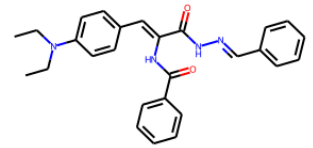


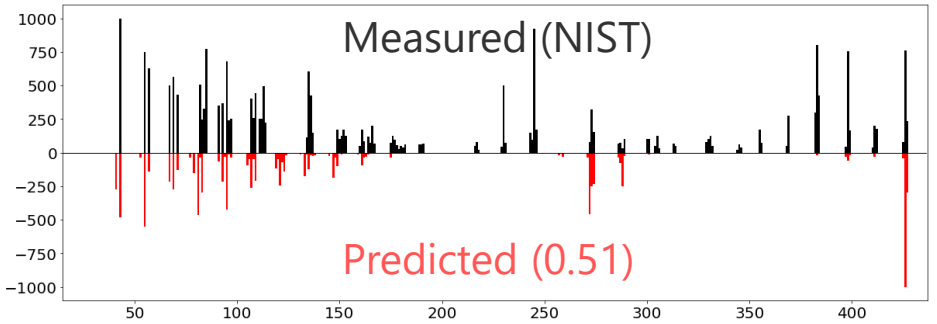
Cosine similarity between observed and predicted

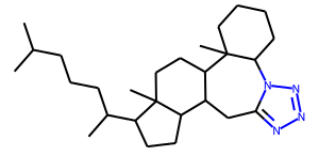












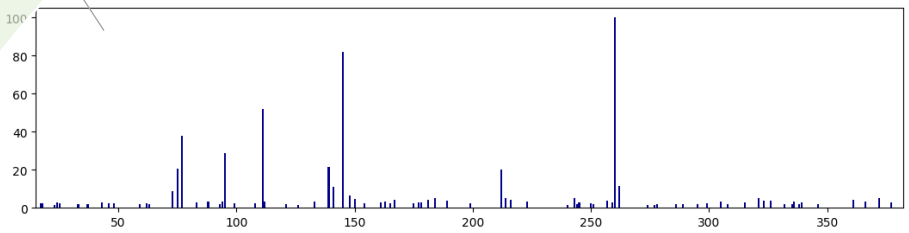
- Cosine similarity distribution for 10,000 known compounds, which were put aside for evaluation.
- Averaged cosine similarity score: 0.86
- Comparison examples of measured mass spectra and predicted mass spectra -->

Improving Prediction Accuracy

"Since its develop in 2022, the AI model has been updated twice, improving its predictive accuracy."

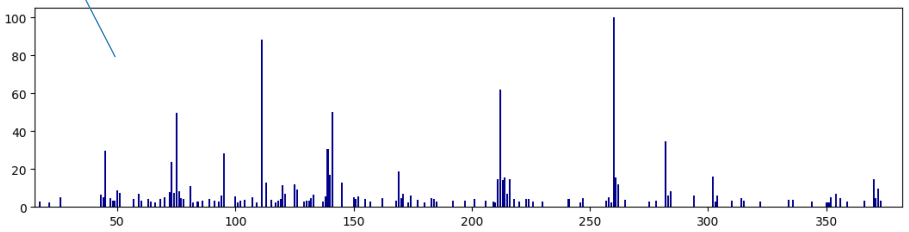
0.72
in 2022

Ver.1 model



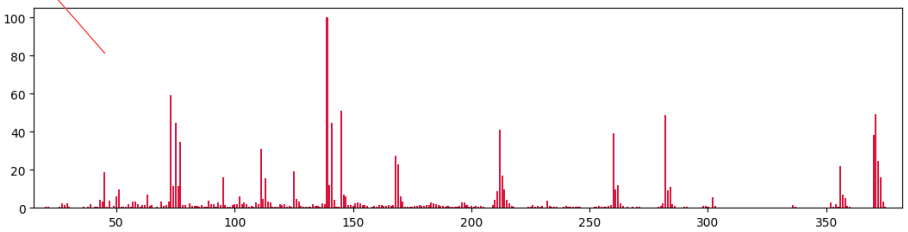
0.80
in 2024

Ver.2 model

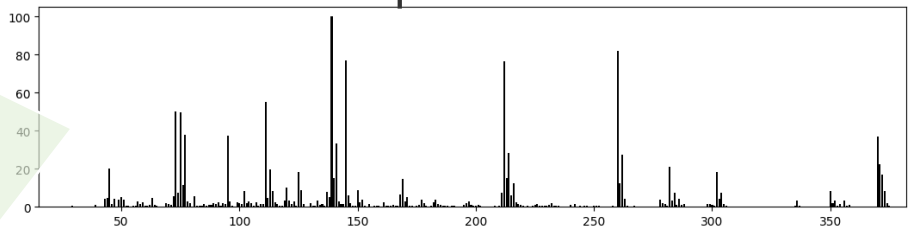


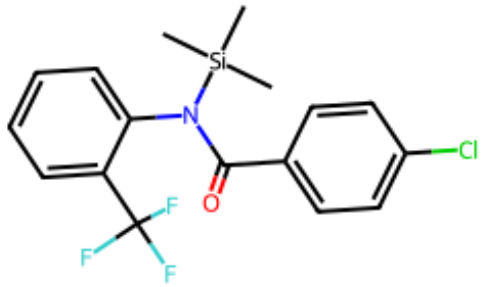
0.86
in 2025

This study, Ver.3 model



Correct Mass Spectrum

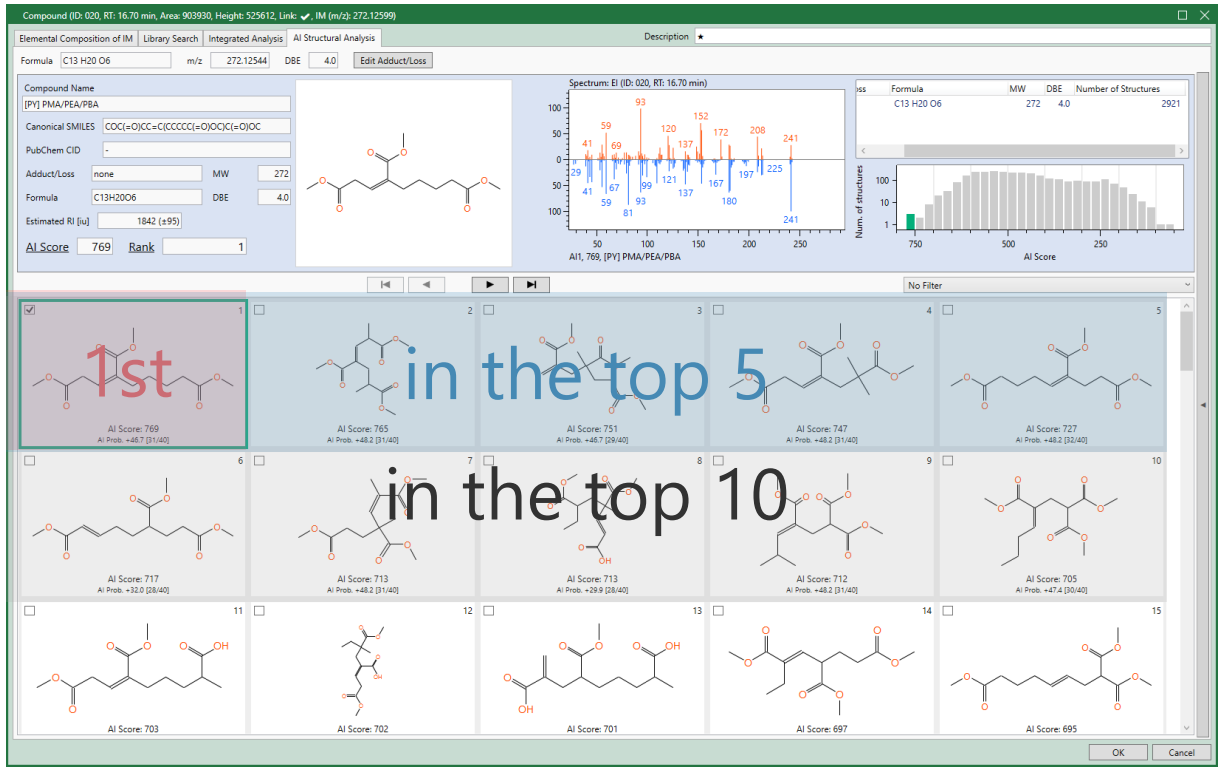




Predicted Mass Spectra from the structure

AI Model Performance: Structural Prediction Accuracy using Evaluation Data

AI Structure Analysis Window on the software (msFineAnalysis AI)



- $C_{13}H_{20}O_6$ isomers: 2,921 in the AI library
- A total of 2,921 structural formulas are available for viewing. Structural formulas are ordered by cosine similarity, starting from the top left (green box).

	Percentage of compounds (%)		
Correct structure is..	Ver.1	Ver.2	**Ver.3**
1st: Highest cosine similarity	22	42	**56**
in the top 5	62	72	**82**
in the top 10	72	80	**87**

- Ranking of correct structure for 10,000 known compounds, which were put aside for evaluation.
- The correct structure was elucidated as the top hit for 56% of the evaluation compounds.
- Highly useful structural information was obtained for 87% of the evaluation compounds.

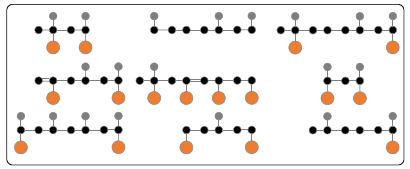
200+ Million Predicted EI Mass Spectral DB “AI Library”

PubChem

Image provided by NCBI/NLM

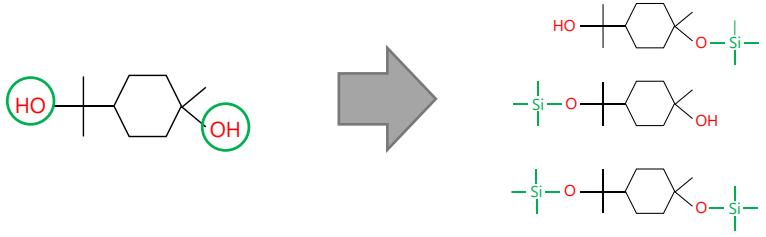
100 million compounds for General

- Less than 1,000 u



27 million compounds for Material

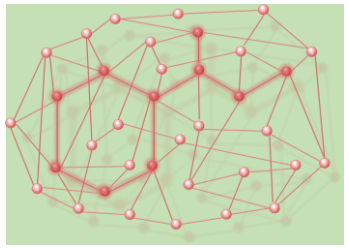
- Polymer pyrolyzates by *in-silico*
- 49 homopolymer, 18 copolymer



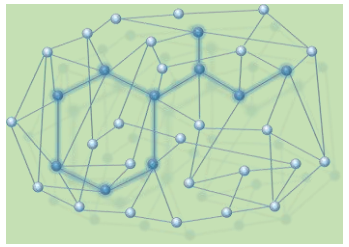
75 million compounds for Metabolomics

- TMS derivatives by *in-silico*
- Methoxime derivatives by *in-silico*

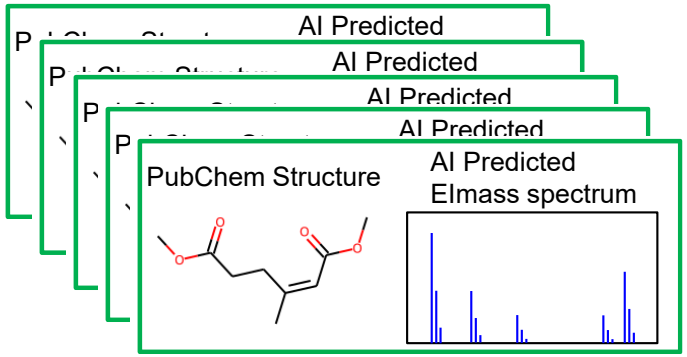
AI Model for EI



AI Model for RI



AI Library

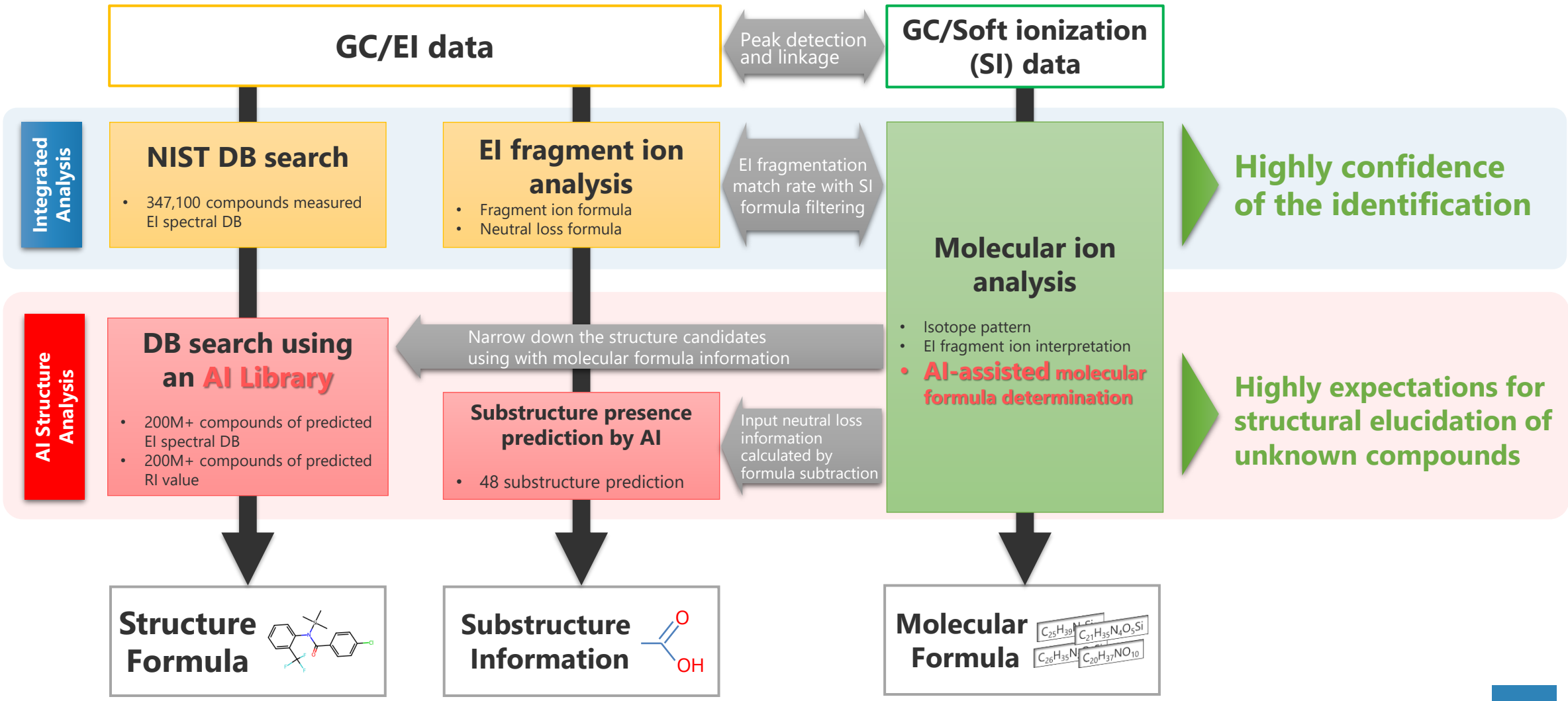


200+ million predicted
EI mass spectral DB

- Neural network
- Complex predictions can be made
- Predicts mass spectrum from the structural formula
- 200+ million predicted mass spectra

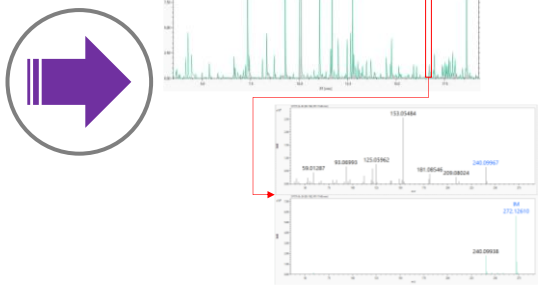
msFineAnalysis AI Workflow

- msFineAnalysis AI handle two GC/MS data, then to offer structure and molecular formula information



JMS-T2000GC × msFineAnalysis AI = New Structure Analysis Workflow

JMS-T2000GC x **msFineAnalysis AI**



$C_{13}H_{20}O_6$

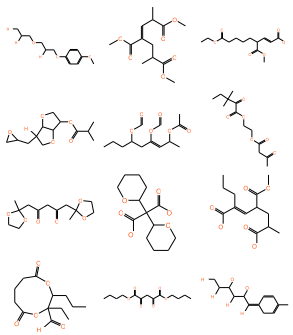
$C_{11}H_{18}N_3O_5$

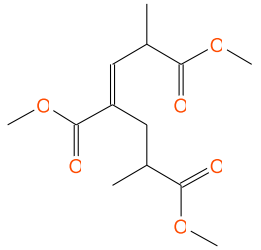
$C_{14}H_{16}N_4O_2$

$C_{16}H_{18}NO_3$

$C_9H_{16}N_6O_4$

AI library





Accurate mass and EI/soft ionization measurements by GC-HRMS

Compound detection by deconvolution

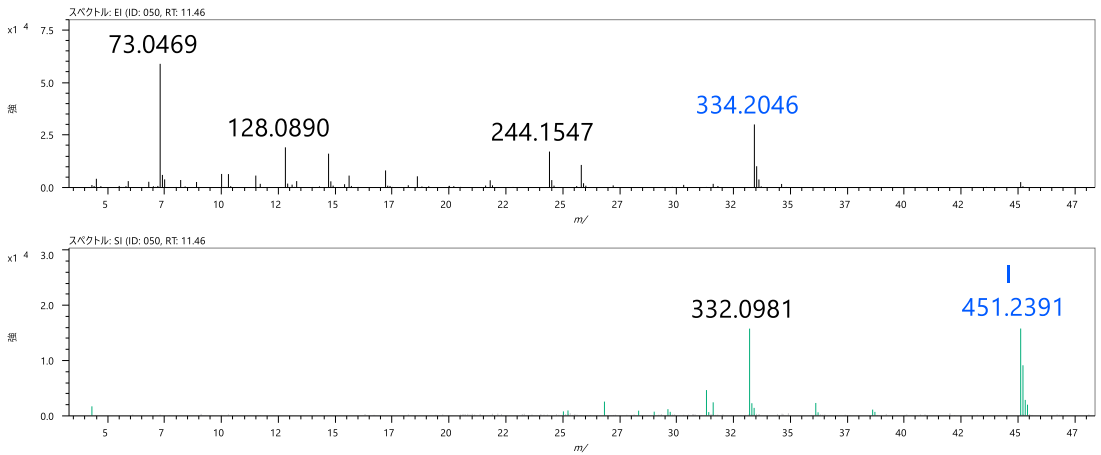
Determination of molecular formula from SI data

Narrow down the candidates using molecular formula information and predicted Retention index values by AI

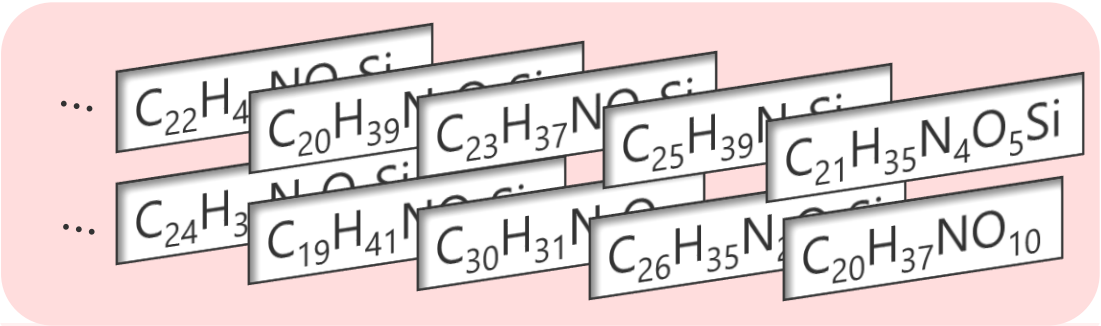
These isomers are then ranked using cosine similarity scores between the predicted and measured EI mass spectra.

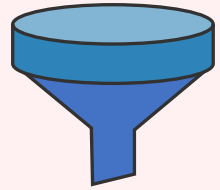
- Narrowing down candidates by molecular formula information before AI library search improves the speed and accuracy of analysis.
- Molecular formula information is also used in integrated analysis (checking NIST DB search results).
- It is important to obtain correct molecular formula information.

AI-assisted molecular formula determination

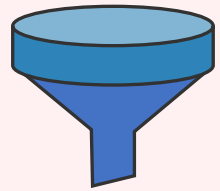


Elemantal Composition Estimation

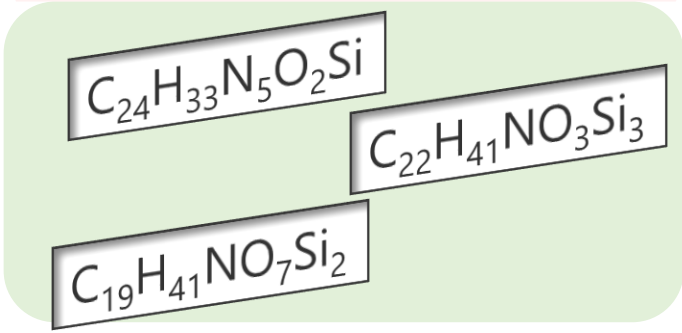




Filter1:
Isotopic Pattern

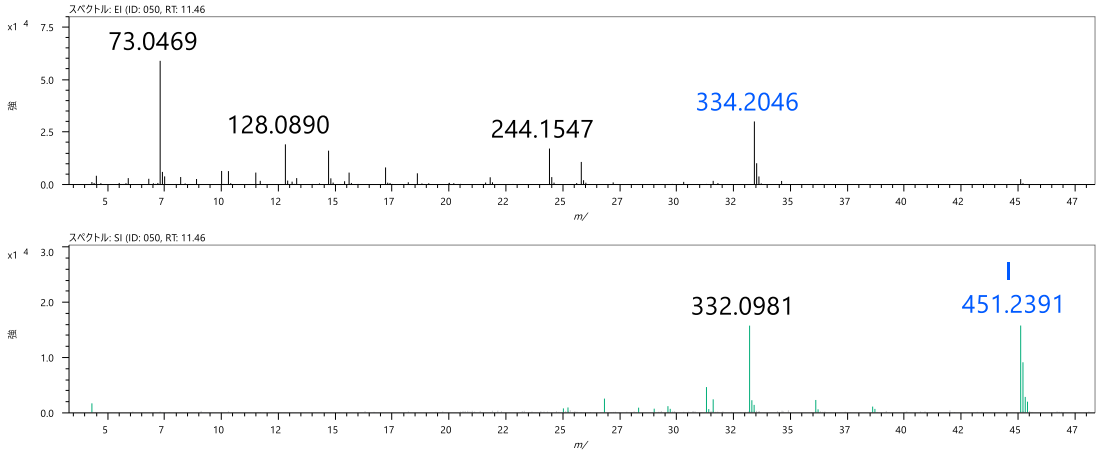


Filter2:
Accurate mass of
Fragment Ions



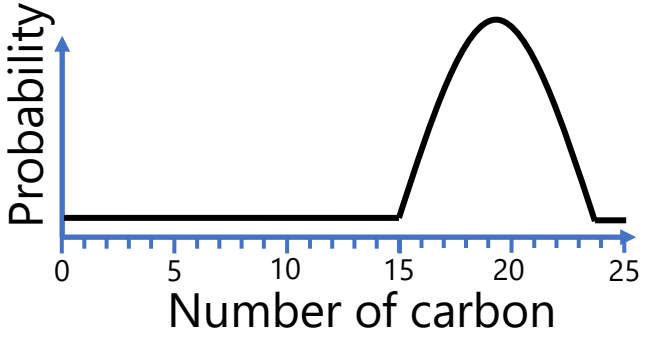
Three candidates are remained.....

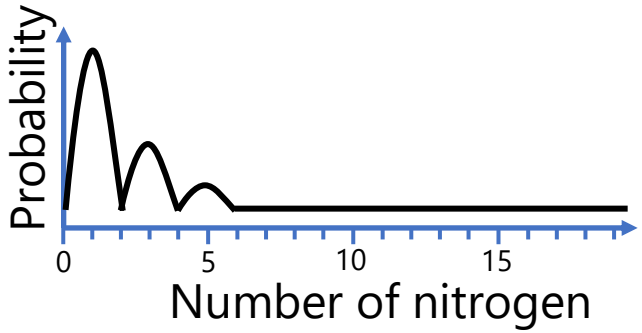
AI-assisted molecular formula determination



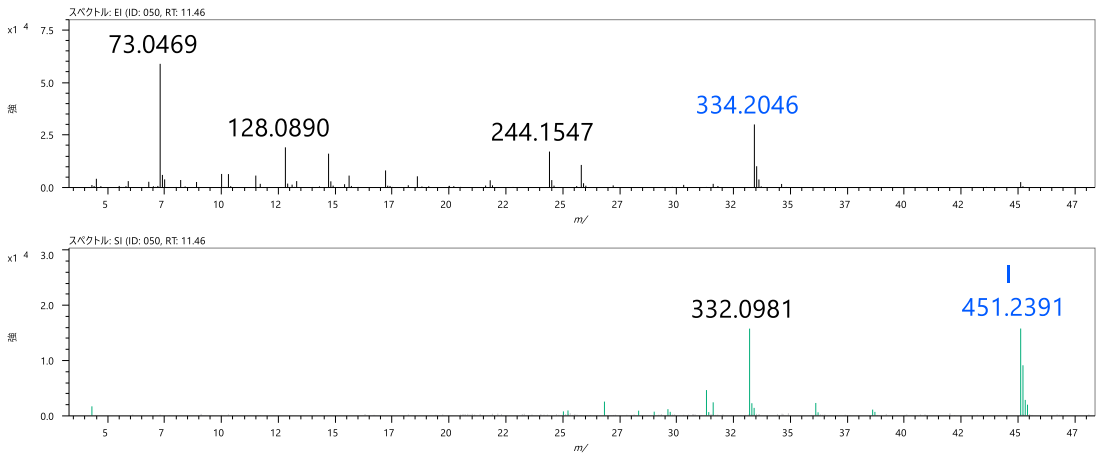
AI model

Probability distributions for the abundance of each elements



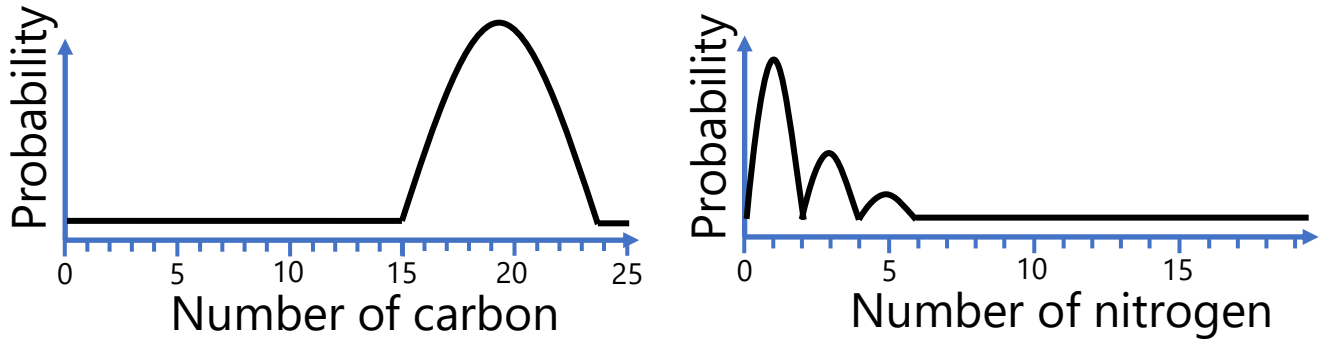


AI-assisted molecular formula determination



AI model

Probability distributions for the abundance of each elements



Recommend

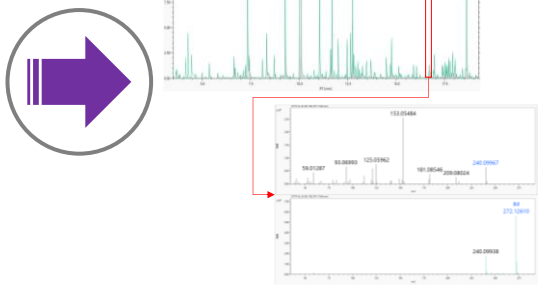
$C_{22}H_{41}NO_3Si_3$

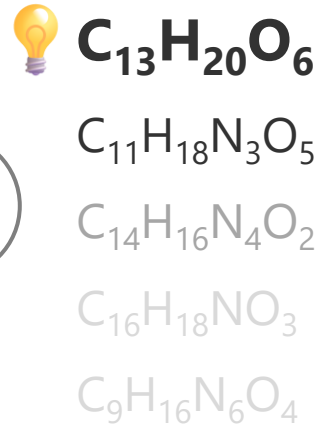
$C_{24}H_{33}N_5O_2Si$

$C_{19}H_{41}NO_7Si_2$

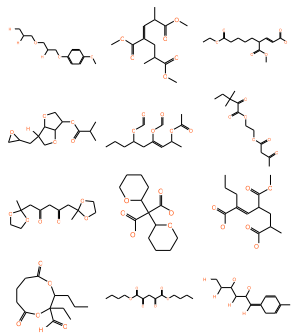
JMS-T2000GC × msFineAnalysis AI = New Structure Analysis Workflow

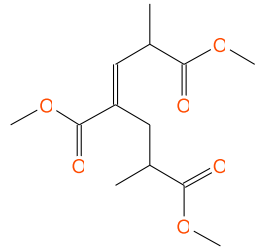
JMS-T2000GC x **msFineAnalysis AI**





AI library





Accurate mass and EI/soft ionization measurements by GC-HRMS

Compound detection by deconvolution

Determination of molecular formula from SI data

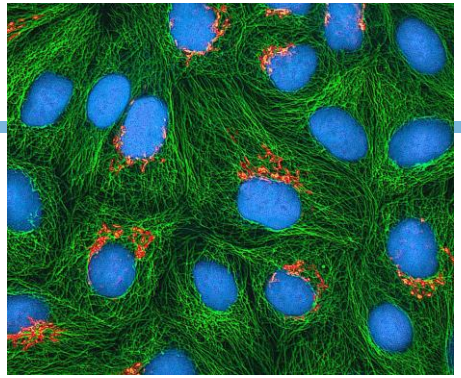
Narrow down the candidates using molecular formula information and predicted Retention index values by AI

These isomers are then ranked using cosine similarity scores between the predicted and measured EI mass spectra.

- Narrowing down candidates by molecular formula information before AI library search improves the speed and accuracy of analysis.
- Molecular formula information is also used in integrated analysis (checking NIST DB search results).
- It is important to obtain correct molecular formula information.

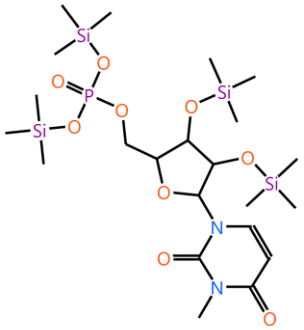
"Metabolomics Applications"

HeLa Cell Metabolome



HeLa cell
By NIH News release

- HeLa cell is cell strain from human cervical cancer.
- This cell is known as model organisms (non-organisms but of human origin) in human metabolomics.
- It is reported that there is compound in the HeLa cell not registered in the NIST DB (N-methyl-uridine monophosphate (N-methyl UMP))*.
- N-methyl-UMP is a valuable metabolomics marker because it reveals hidden RNA-modification–related metabolites that are not covered by standard databases.
- We tried metabolome analysis using JMS-T2000GC and msFineAnalsys AI.
- HeLa cell samples were obtained from Prof. H. Tsugawa, a KOL in the metabolomics field in Japan.



N-methyl UMP

Prof. Hiroshi Tsugawa
Tokyo University of Agriculture and Technology

*Z.Lai, H.Tshugawa, et al. (2017). Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics

HeLa Cell Metabolome

INSIGHT-Thermal modulator
SepSolve

+

JMS-T2000GC

GCxGC

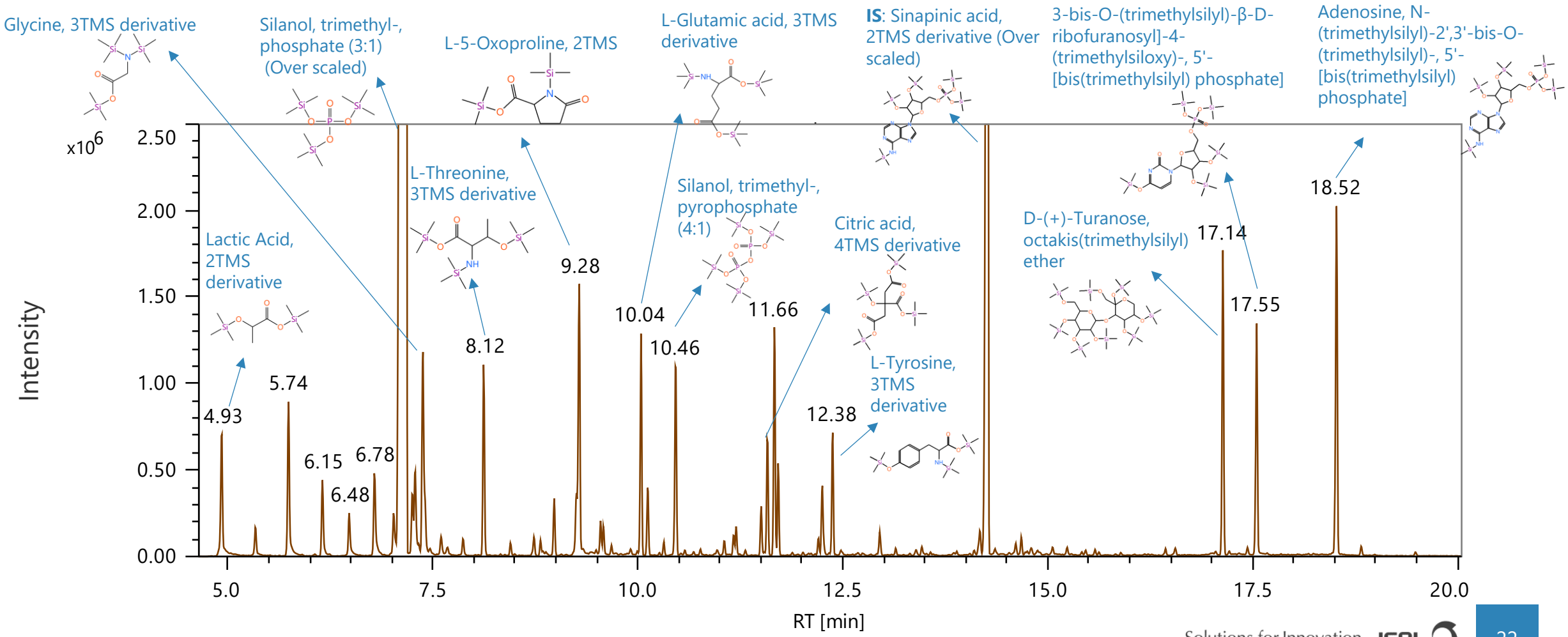
- Sample amount: 1 µL
- Column: 1st: BPX5 30 m× 0.25 mm, 0.25 µm,
  2nd: Rxi17-Sil MS 3.4 m× 0.15 mm, 0.15 µm
- Inlet: 250℃, Splitless
- Carrier gas flow: He, Constant Flow 1.2 mL/min
- Oven: 80℃ (2 min) → 5℃/min → 325℃ (9 min)

TOFMS

- Ion source: EI/FI combination ion source
- Ionization: EI+: 70 eV, 300 µA,
  FI+: -10 kV, 6 msec, 12 mA
- Recording interval: EI: 0.02 sec (50 Hz), FI: 0.04 sec (25 Hz)
- *m/z* range： *m/z* 33 – 800

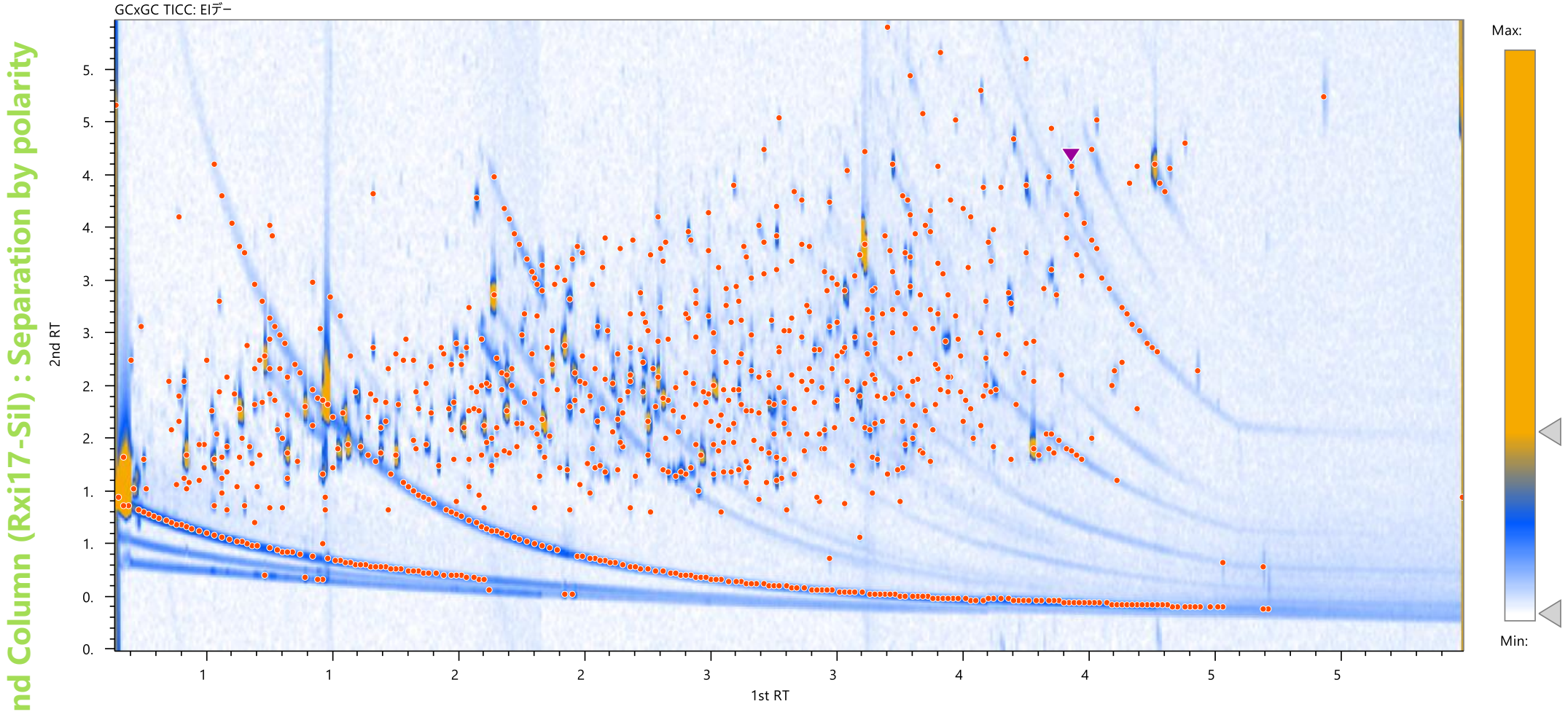
HeLa cell metabolites detection by 1D-GC

- Many metabolites were detected, including amino acids, phosphates, sugars, and organic acids.



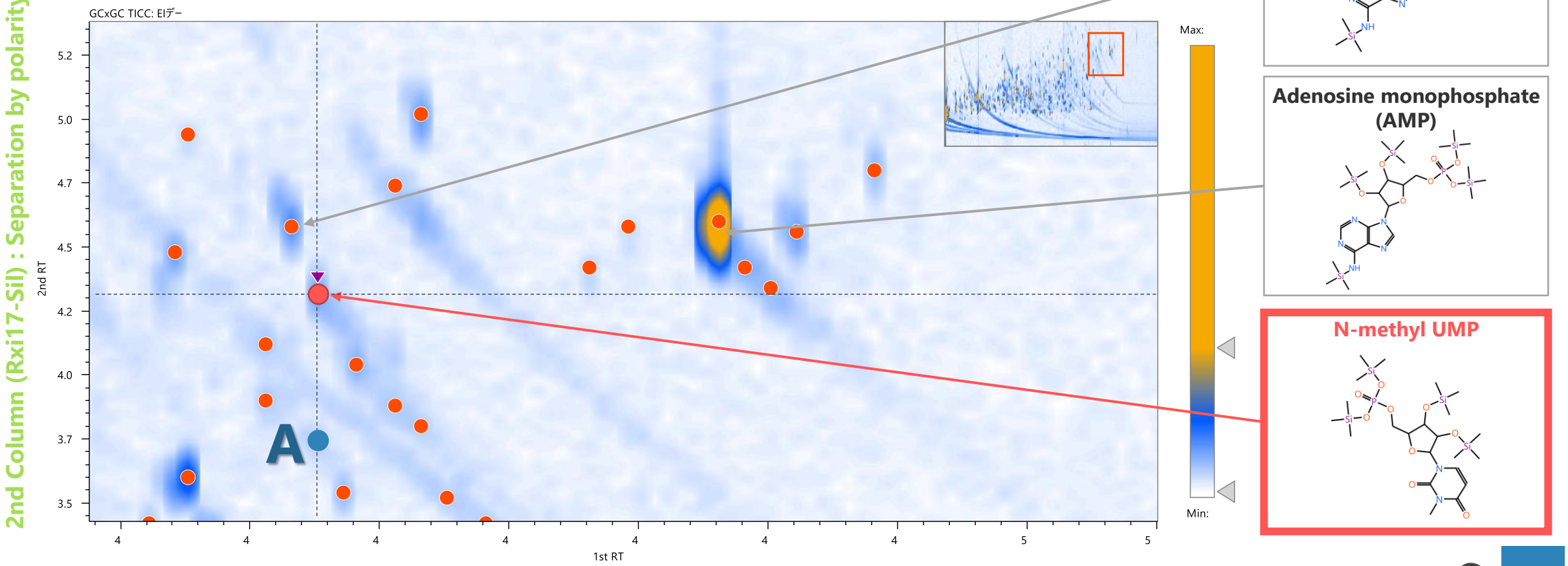
HeLa cell metabolites detection by 2D-GC

- Over 800 compounds were detected in 2D-GC TICC.



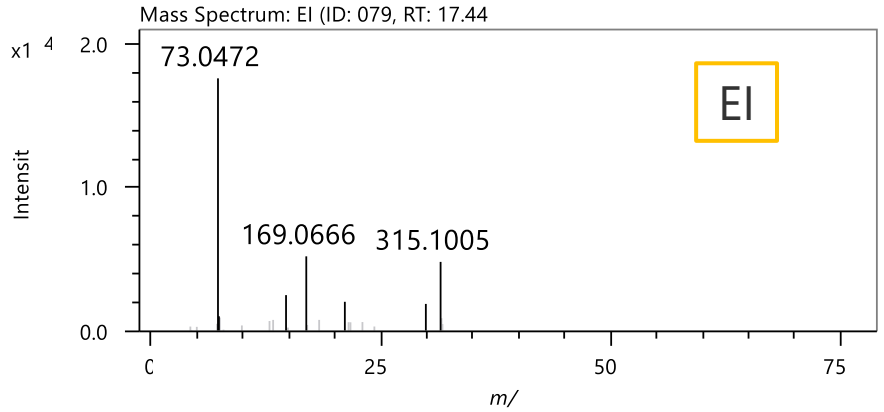
N-methyl UMP detection in 2D TICC

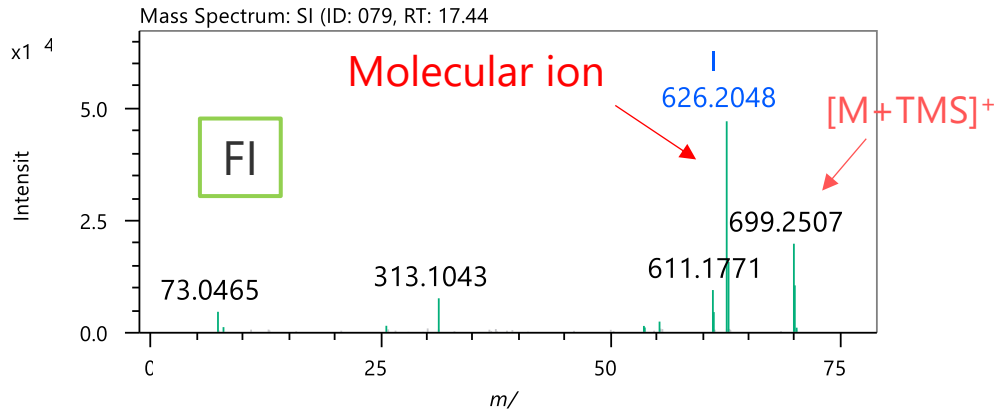
- N-methyl UMP could be separated by GCxGC from **"A"** which have same boiling point but different polarity.
- N-methyl UMP had close 2nd RT with nucleic acid-related compounds.



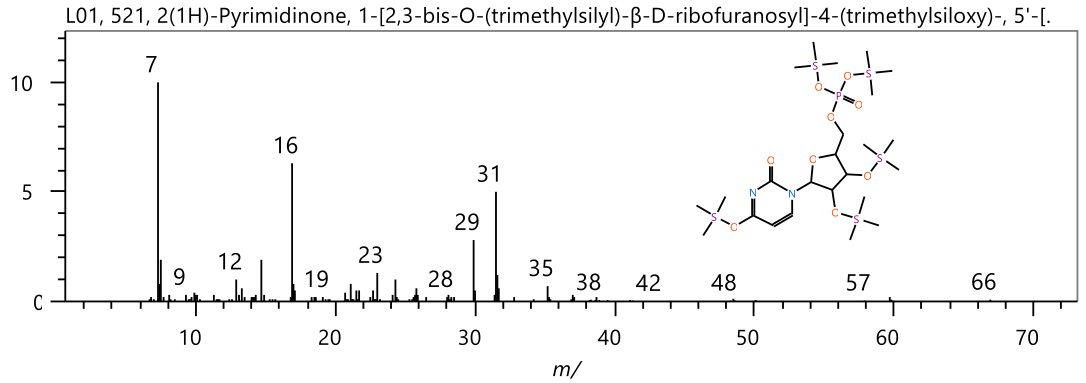
N-methyl UMP 4TMS mass spectra

- Mass spectra





- NIST DB search No.1 candidate (MW: 684)



- Integrated analysis result

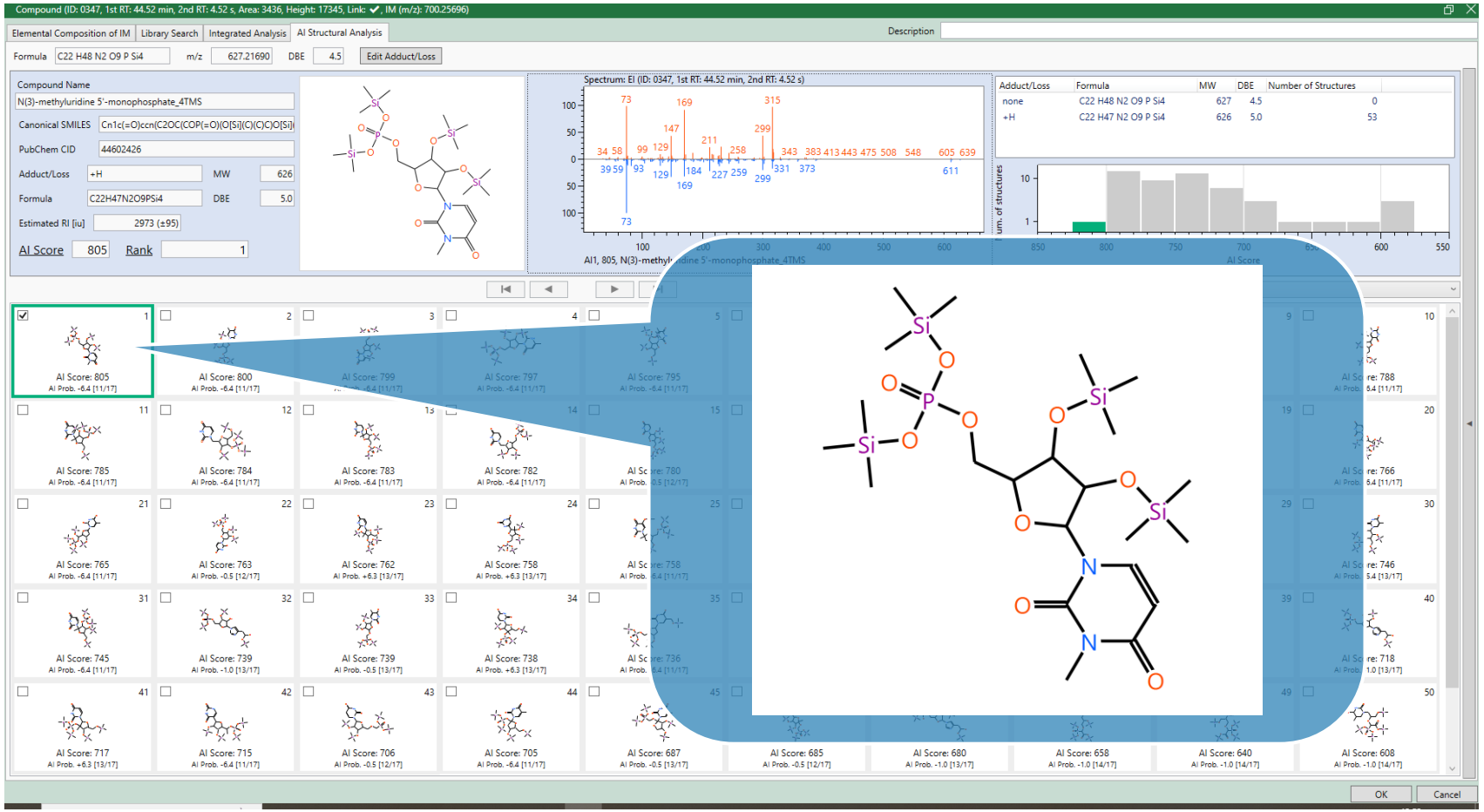
Elemental Composition of IM (m/z: 626.20489)						
#	Formula	DBE	Calculated m/z	Mass Error [mDa]	Isotope Matching	Coverage
A01	C22 H47 N2 O9 P Si4	5.0	626.20908	-0.90	0.61	100

After the integrated analysis:

- This EI mass spectrum is not registered NIST DB → Unknown compound
- Molecular formula is $C_{22}H_{47}N_2O_9PSi_4$ → Same with N-methyl UMP

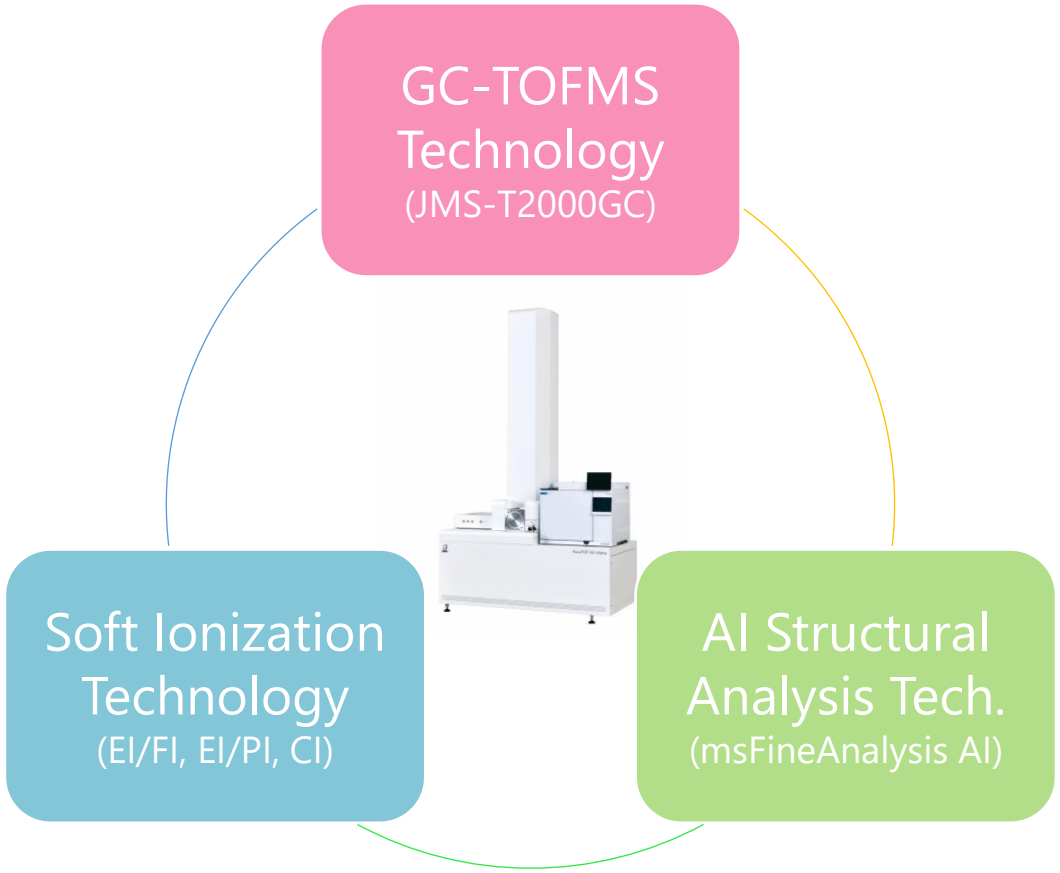
AI structure analysis result

- There were 53 compounds with the same molecular formula as N-methyl UMP.
- The structural formula of N-methyl UMP was the 1st candidate with an AI score of 805.



Conclusion

Three technologies for the advanced metabolomics



- Improved identification accuracy through NIST DB search of GC/EI data and accurate mass analysis.
- For unknown compounds, molecular formulas are obtained from GC-TOFMS and soft ionization data.
- Furthermore, structure estimation is performed using msFineAnalysis AI.
- We made +200 million predicted EI mass spectra DB
  - 27 million *in silico* pyrolyzates library is useful for polymer analysis.
  - 75 million *in silico* TMS derivatives library for metabolomics.
- It is important to combine molecular formula information to obtain the correct structure formula with AI library.
  - High Resolution Mass Spectrometry is necessary to obtain composition formula for both molecule and fragment ions.
  - Soft ionization is ideal to obtain molecule ion and protonated molecule.
- The combination of these three technologies is expected to be useful in advanced metabolomics.

JEOL Presentations at MDCW

#	Presentation	Day	Presenter	Session #
1	Structural elucidation using comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry and machine learning for unknown metabolites in HeLa cells	Day 1 1:30-1:50PM	Masaaki Ubukata	O-7
2	No more split ends? Flow-modulated GCxGC-QMS analysis without splitting off the GC flow	Day 3 9:00-9:20AM	Kirk Jensen	O-16
3	Analysis of aroma compounds in spices by comprehensive two-dimensional gas chromatography/time-of-flight mass spectrometry with machine learning-based structure elucidation and molecular formula estimation	Day 1 3:00-4:00PM	Azusa Kubota	P-9
4	What do we do with all that data? Complementary data processing methods for two-dimensional gas chromatography and mass spectrometry	Day 2 3:00-4:00PM	Robert Cody	P-11
5	The characterization of poly(1-butene) via pyrolytic conversion using comprehensive two-dimensional gas chromatography high-resolution time-of-flight mass spectrometer	Day 2 3:00-4:00PM	Bryan Katzenmeyer	P-24

Thank you for your cooperation and attention!