The Multidimensional Chromatography Workshop 2025

Development of Unknown Compounds Analysis Method combining High-Resolution Mass Spectrometry, Soft Ionization Technique and AI Technology for Comprehensive 2-Dimensional Gas Chromatography

JEOL Ltd.
MS Business Unit, MS Application Group

Masaaki Ubukata, Ph. D.

Outlines

- GC/MS and GCxGC/MS Qualitative Analysis Workflow and Issue
- New Qualitative Analysis Workflow for Unknown Compounds
  - Integrated Analysis: Determine for the Molecular Formula
  - AI Structure Analysis: Elucidation for the Structure Formula
- GCxGC Data Handling on the Software and It's Applications
- Conclusion

GC/MS and GCxGC/MS Qualitative Analysis Workflow

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

TD

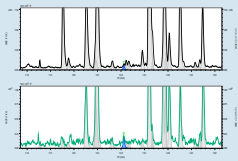
(株)ENVサイエンストレーディング

HS-SPME

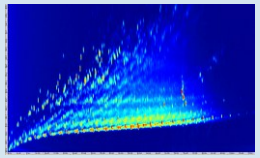
GC用オートサンプラ HT2850T

Introduction
of a sample

- GC/MS
- GCxGC/MS



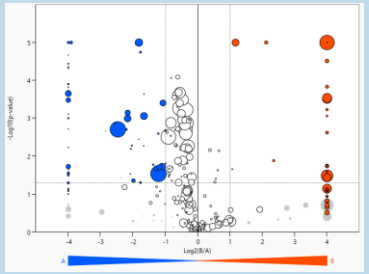
1D-GC



2D-GC

Separation
for compounds

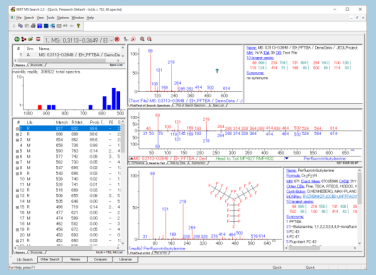
- Data Processing:
  Volcano plot, PCA



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

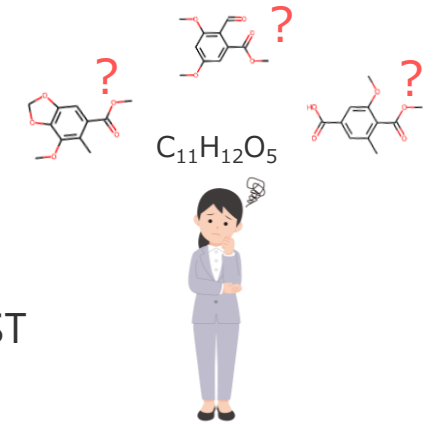
Identification
of specific compounds

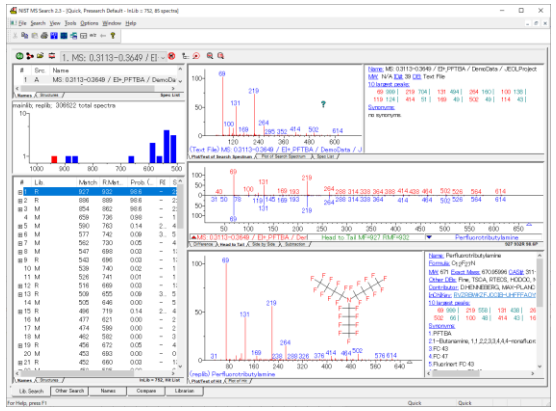
If a specific compound found after all the experiments is not listed in the commercially available EI mass spectral DB, how do you identify it?

Challenges in GC/MS and GCxGC/MS Qualitative Analysis

How to identify the unknown compounds

which is not registered in the database supervised NIST?





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

- EI mass spectrum database (DB) supervised 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 DB
- In many cases, they become unknown compounds
- Need time and knowledgeable for structure analysis

JEOL propose an automatic structure analysis solution
using the latest AI technology for real unknown compounds!

Our Goal: Automatic Structure Analysis Solution

Unknown Compounds

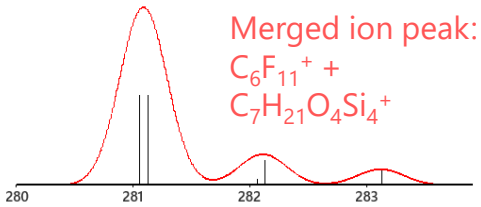
cannot be identified by NIST DB

Enormous amount of chemicals in the world!

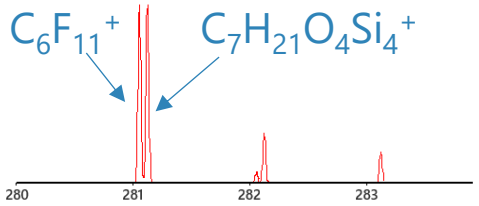
- The latest DB contains 347,100 compounds
- More than 100 million compounds in PubChem
- Almost all compounds are not listed in the DB

Composition Formula

Low-Res. mass spectrum



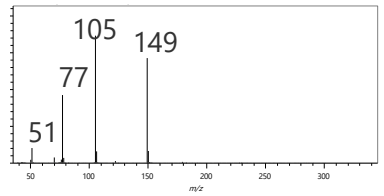
High-Res. mass spectrum



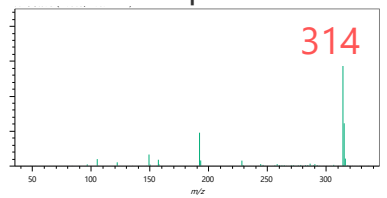
Only High-Resolution GC-TOFMS can be used to obtain composition formulas

Molecular Formula

EI mass spectrum



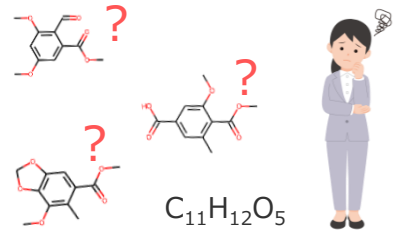
Soft ionization mass spectrum



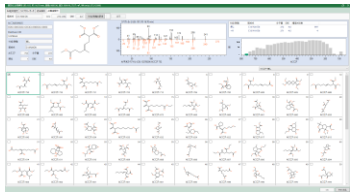
Soft ionization is necessary for determining the correct molecular formula

Structural Formula

By yourself manually



By software automatically



JEOL software automatically performs difficult structural analysis

Our Goal: Automatic Structure Analysis Solution

Unknown Compounds

cannot be identified by NIST DB

Enormous amount of chemicals in the world!

- The latest DB contains 347,100 compounds
- More than 100 million compounds in PubChem
- Almost all compounds are not listed in the DB

Composition Formula

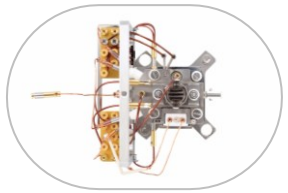
by accurate mass analysis

GC-TOFMS

- Mass resolving power: 30,000 @ m/z 614
- Mass accuracy: 1ppm (EI, internal calibration)
- Composition formula for all ionizations

Molecular Formula

by soft ionization



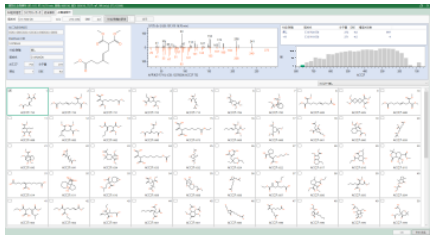
EI/FI comb. ion source

EI/PI comb. ion source

- Electron Ionization (EI), Chemical Ionization (CI), Photo-Ionization (PI), and Field Ionization (FI) for GC-MS
- Two combination ion sources: EI/FI and EI/PI
- Direct MS probe: DEP, DIP and Field Desorption (FD)

Structural Formula

by artificial intelligence



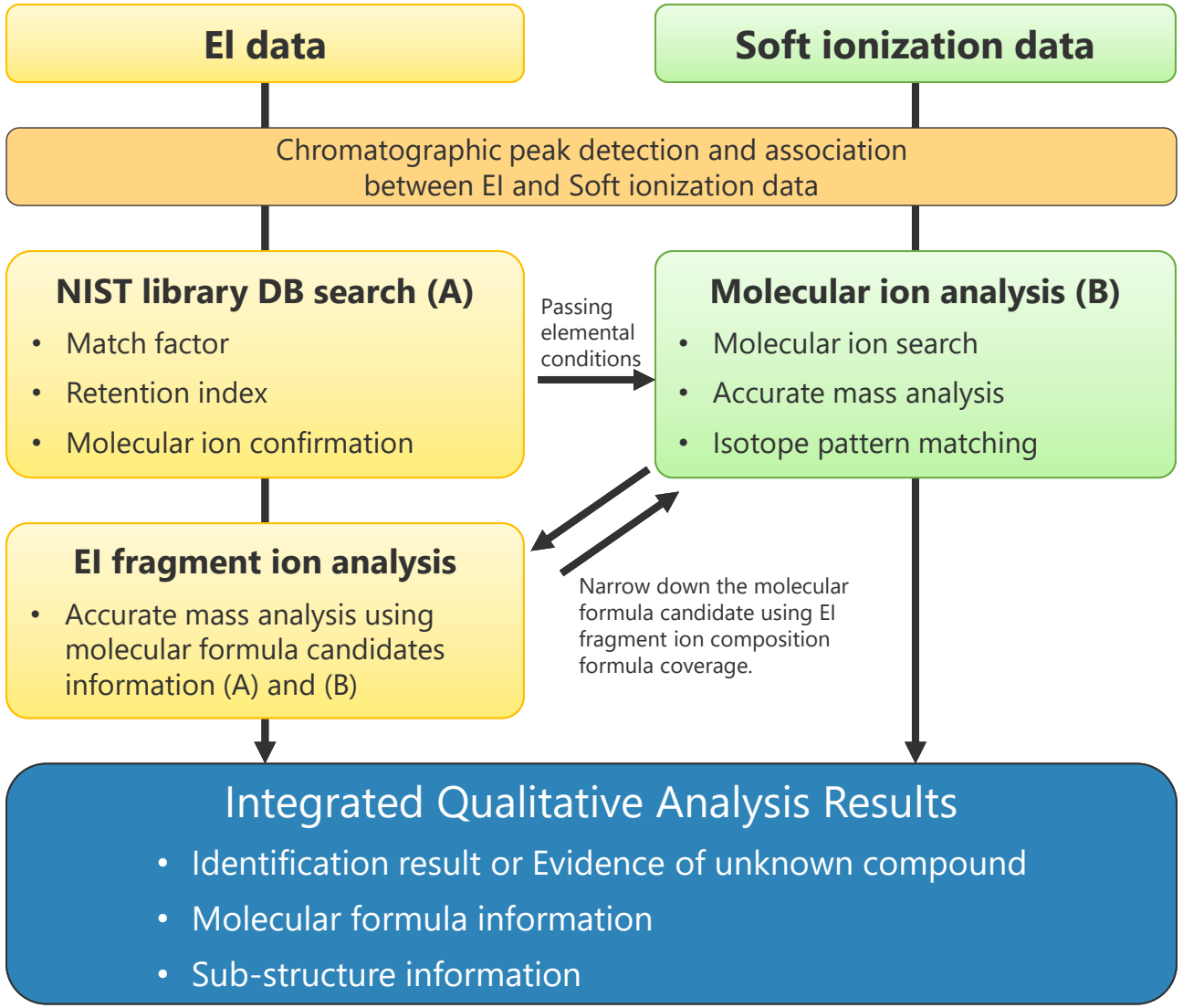
msFineAnalysis AI
(Qualitative software)

- Automated Structural Analysis Software
- AI structure analysis function
- Two data (EI and Soft Ionization) handling
- Deconvolution peak detection, Retention Index, Differential Analysis and so on

Outlines

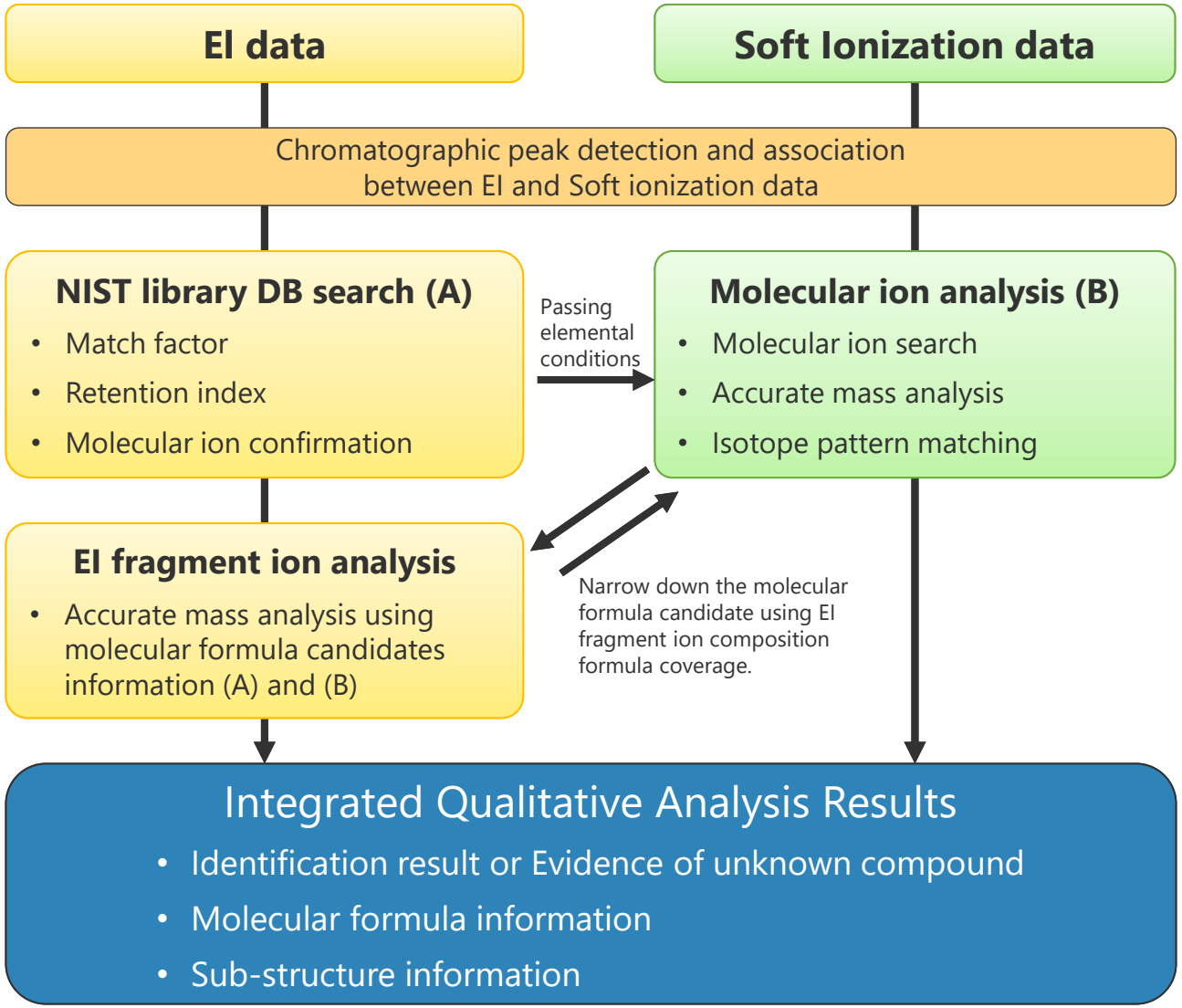
- GC/MS and GCxGC/MS Qualitative Analysis Workflow and Issue
- New Qualitative Analysis Workflow for Unknown Compounds
  - Integrated Qualitative Analysis: Determine for the Molecular Formula
  - AI Structure Analysis: Elucidation for the Structure Formula
- GCxGC Data Handling on the Software and It's Applications
- Conclusion

Integrated Qualitative Analysis: Determine for the Molecular Formula

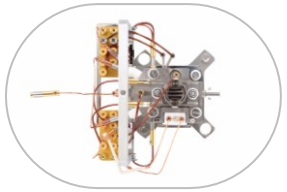


＊M. Ubukata et al, *Rapid Commun Mass Spectrom., 34 (**2020**). DOI: 10.1002/rcm.8820*

Integrated Qualitative Analysis: Determine for the Molecular Formula



GC-TOFMS



EI/FI comb. ion source

EI/PI comb. ion source

- GC-HRMS system and soft ionization techniques are necessary

Next step is elucidation for the structure formula of unknown compounds using AI technology!

Outlines

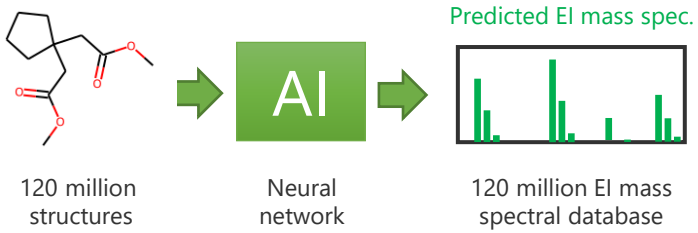
- GC/MS and GCxGC/MS Qualitative Analysis Workflow and Issue
- New Qualitative Analysis Workflow for Unknown Compounds
  - Integrated Qualitative Analysis: Determine for the Molecular Formula
  - AI Structure Analysis: Elucidation for the Structure Formula
- GCxGC Data Handling on the Software and It's Applications
- Conclusion

AI Structure Analysis: Elucidation for the Structure Formula

- JEOL uses a combination of **Deep learning** and **Machine learning** for unknown compounds analysis

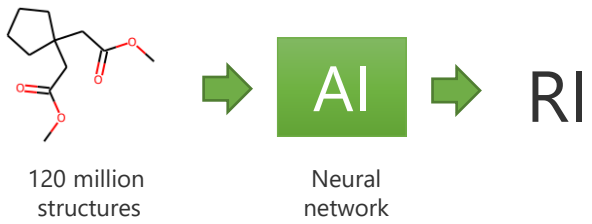
AI Library -Predicted EI Mass Spectral DB-

Mass Spectral Prediction



- Neural network
- Complex predictions can be done
- Predicts mass spectrum from structural formula
- 130 million predicted EI mass spectra

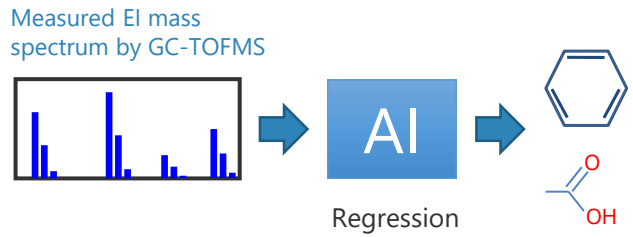
Retention Index Prediction



- Neural network
- Predicts RI value from structures
- All structural formulas in the AI library are assigned a predicted RI
- Used to narrow down candidate structural formulas

Substructure Analysis

Substructure Prediction

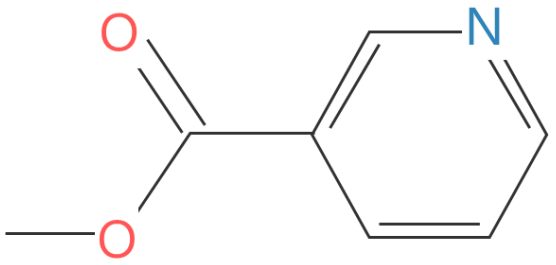


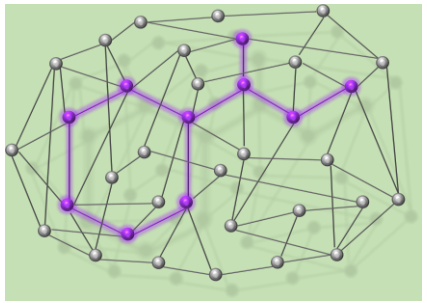
- Regression
- Limited prediction due to simple way
- Prediction of substructures from measured mass spectra
- 48 substructures can be predicted

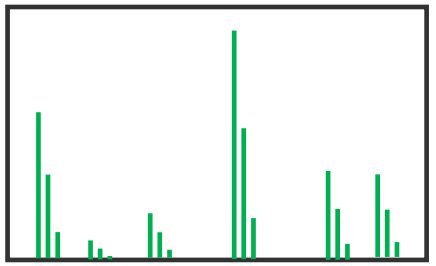
AI Library: Elucidation for the Structure Formula

- An AI model was developed to predict EI mass spectra and RI from chemical structures using deep learning.
- The AI model was then used to create 130 million EI mass spectral database for determining the structure candidates for unknown compounds.

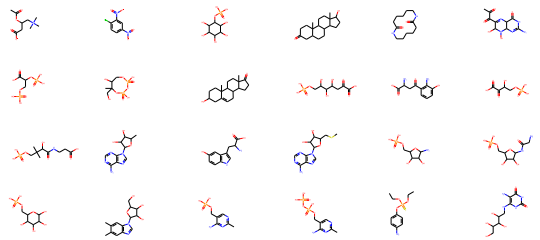
AI Model



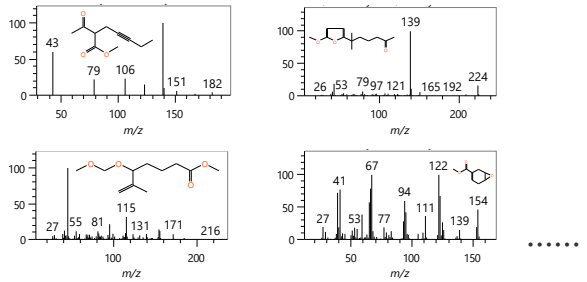




\+ RI



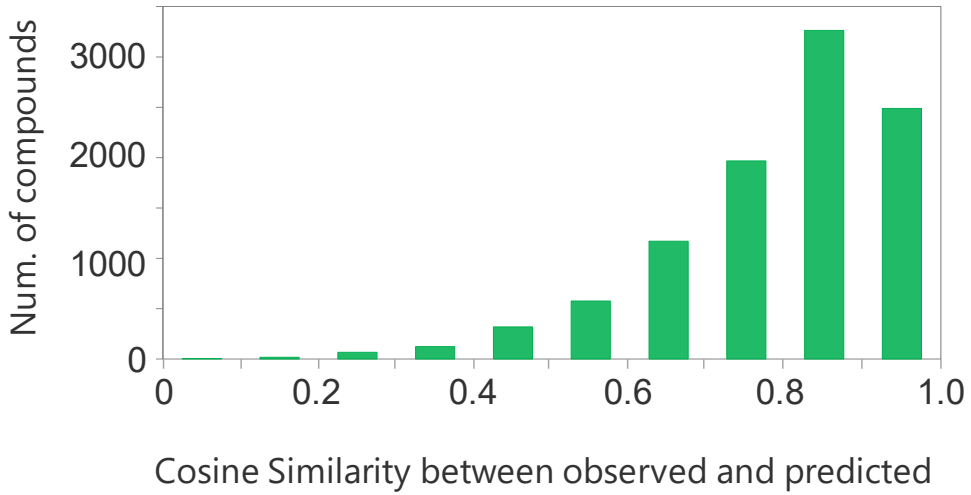
130 million structures
(PubChem DB, 2 *in-silico*)



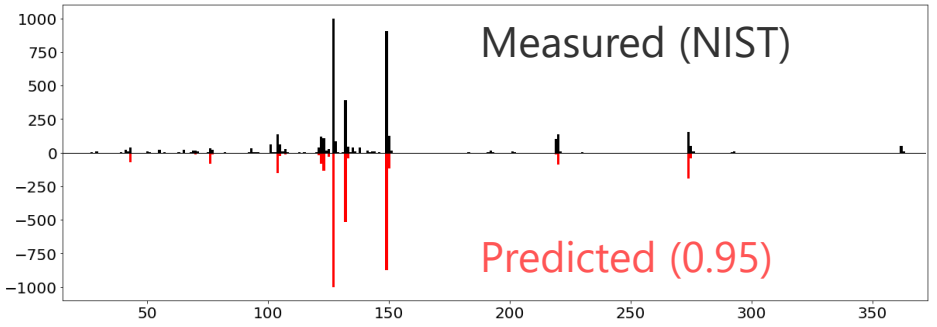
130 million EI mass spectral DB
with predicted RI

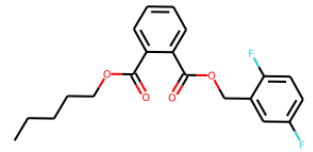
＊A. Kubo et al, *Mass Spectrometry*, **2023**, *12*, A0120.

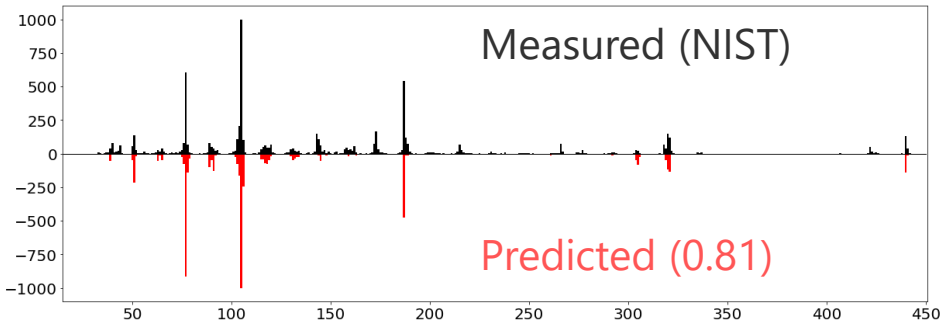
AI Model Performance: Cosine Similarity using Evaluation Data

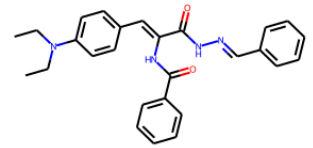


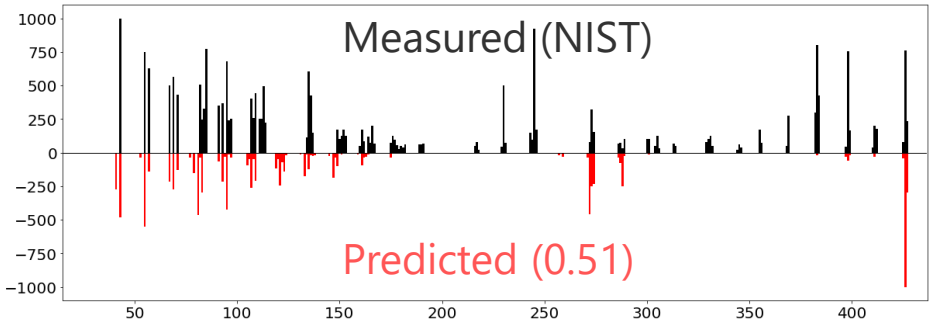
- Cosine similarity distribution for 10,000 compounds
- Averaged cosine similarity score: 0.80
- Comparison example of measured mass spectra and predicted mass spectra -->

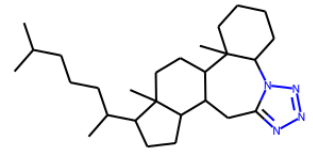




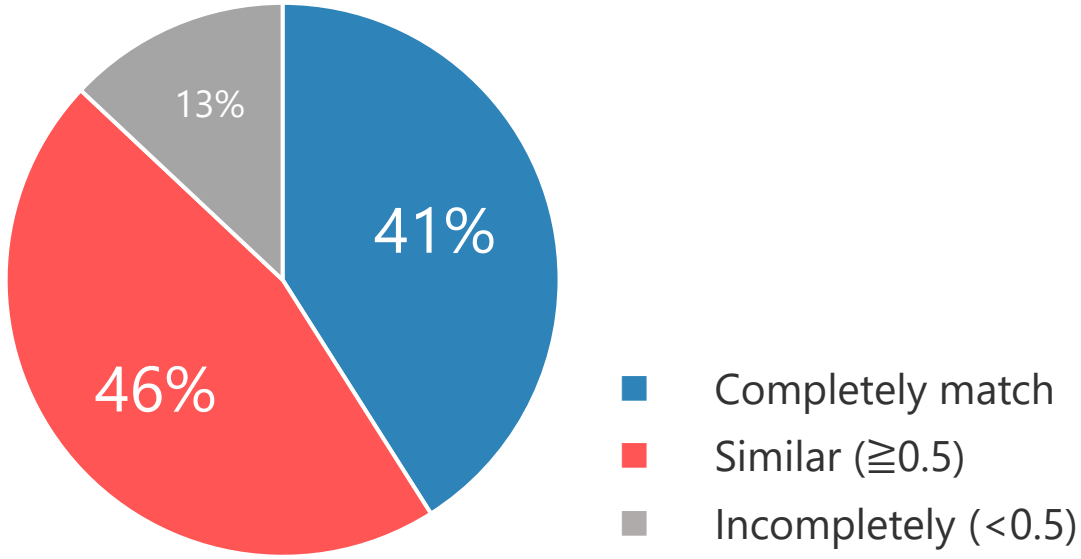






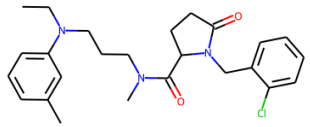
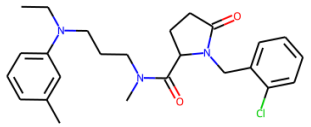
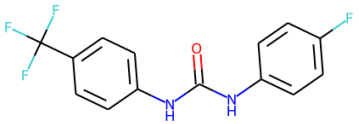
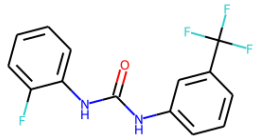
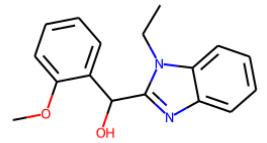
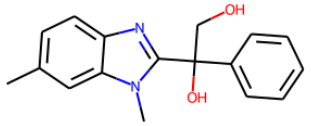
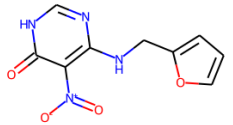


AI Model Performance: Accuracy using Evaluation Data



- Accuracy distribution for 10,000 compounds
- 41% were completely matching between with correct structure and top 1 similarity score.
- 46% had similar structure between with correct structure and top 1 similarity score.
- Totally 87% compounds were elucidated correct and similar structures.

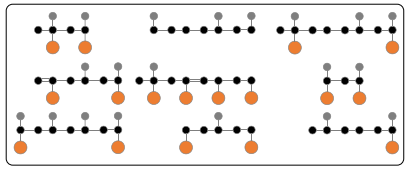
Example of the Accuracy result

Class	Correct Structure	Top 1 Similarity Score
Completely 41%		
Similar 46%		 (Similarly: 0.7)
		 (Similarly: 0.5)
Incompletely 13%		 (Similarly: 0.2)

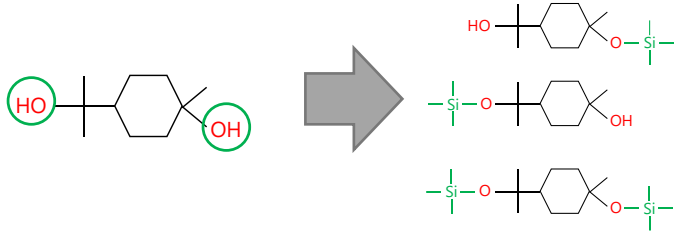
Output: 130 Million Predicted EI Mass Spectral DB “AI Library”

PubChem

- 100 million compounds
  - Less than *m/z* 1,000

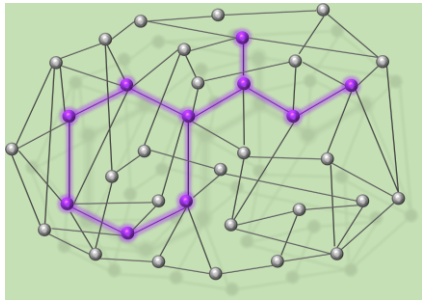


- 25 million compounds
  - Polymer pyrolyzates by *in-silico*
  - 49 homopolymer, 18 copolymer

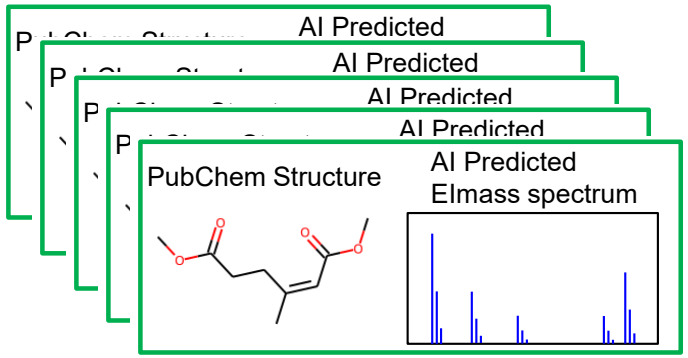


- 5 million compounds
  - TMS derivatives by *in-silico*
  - -OH, -NH2, -NH, -SH

AI Model



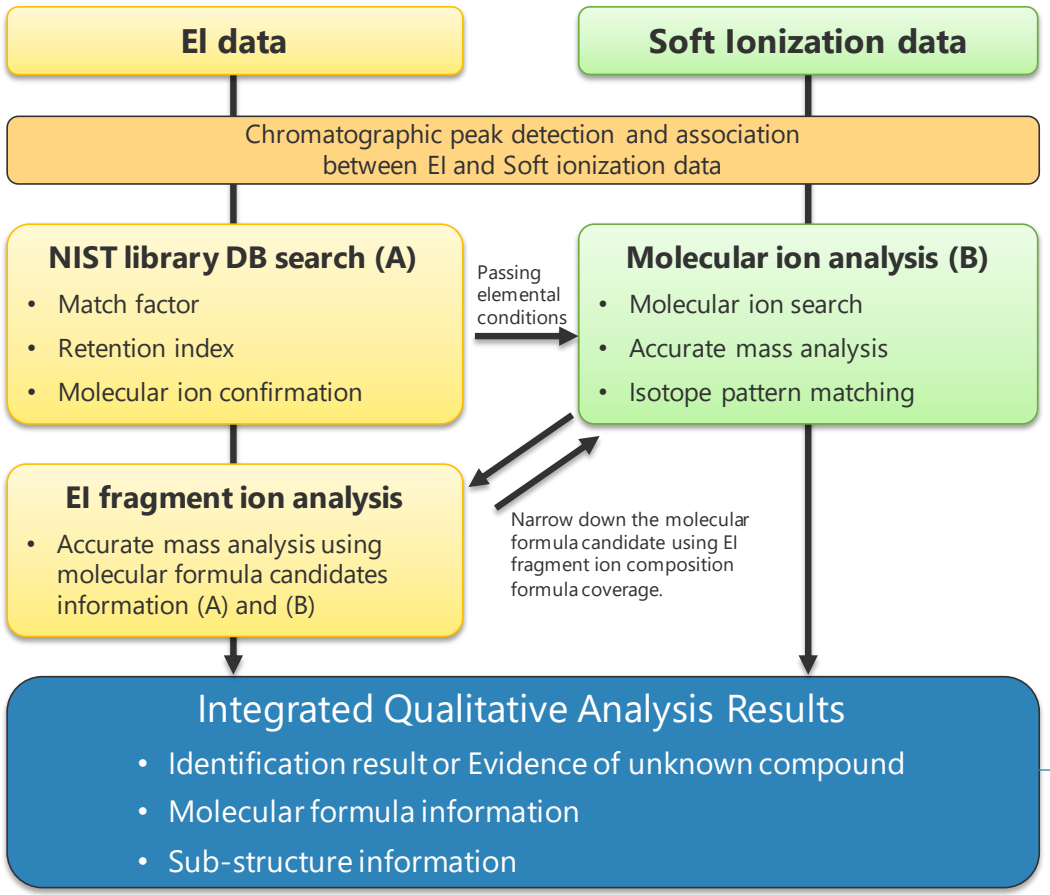
AI Library

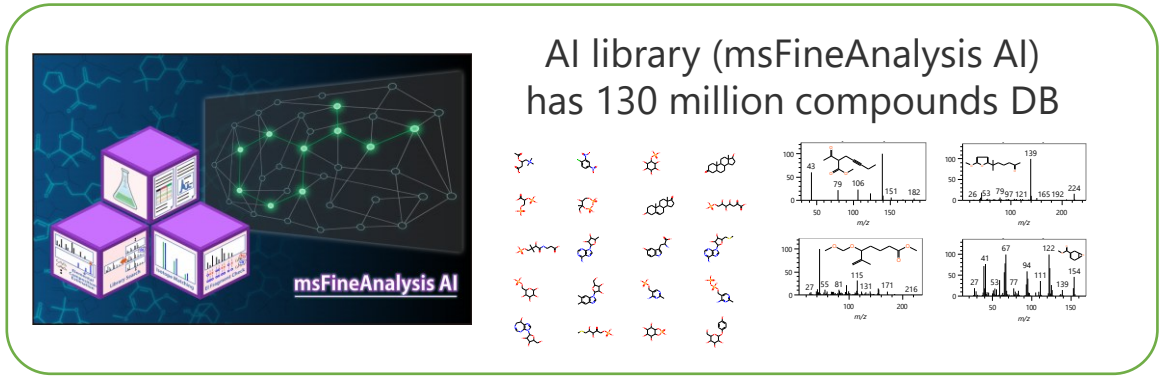


130 million predicted
EI mass spectral DB

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

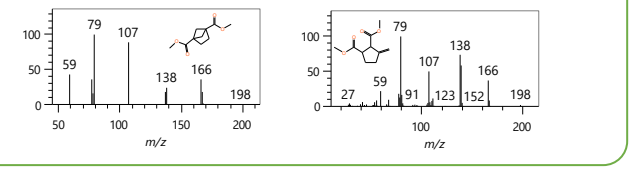
AI Structure Analysis Workflow after Integrated Qualitative Analysis

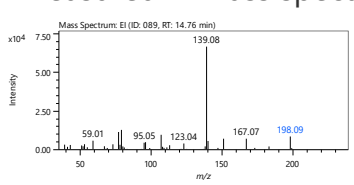


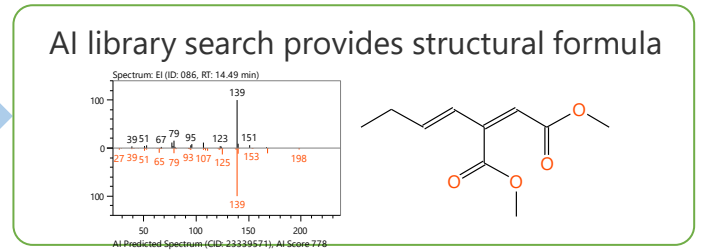


Molecular formula

determined by integrated analysis







- First, perform an integrated analysis
- Utilize molecular formula information determined by integrated analysis
- Non-registered compounds in the NIST DB become targets of AI structural analysis

Total Workflow for Unknown Compound Analysis

EI data

Soft Ionization data

Chromatographic peak detection and association between EI and Soft ionization data

NIST library DB search (A)
- Match factor
- Retention index
- Molecular ion confirmation

Passing elemental conditions

Molecular ion analysis (B)
- Molecular ion search
- Accurate mass analysis
- Isotope pattern matching

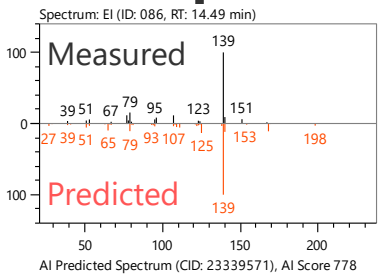
EI fragment ion analysis
- Accurate mass analysis using molecular formula candidates information (A) and (B)

Narrow down the molecular formula candidate using EI fragment ion composition formula coverage.

Integrated Qualitative Analysis Results
- Identification result or Evidence of unknown compound
- Molecular formula information
- Sub-structure information

Integrated Qualitative Analysis Results

AI library search
- 130 million compounds
- Narrow down the candidates by
  - molecular formula information
  - predicted RI (3rd AI)



Sub-structure analysis
- 48 sub-structures
- Supporting interpretation of structural analysis results from AI library

presence

absence

presence

Structure Analysis Results for unknowns
- Structure formula candidates
- Sub-structure prediction

GC/MS and GCxGC/MS Qualitative Analysis Workflow

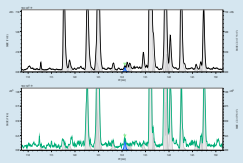
- Sample prep.
- Pre-treatment:
  HS, TD, Py…

TD

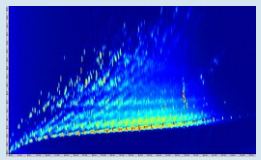
HS-SPME

(株)ENVサイエンストレーディング　GC用オートサンプラ　HT2850T

- GC/MS
- GCxGC/MS

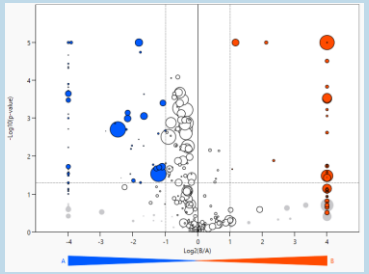


1D-GC



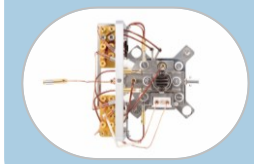
2D-GC

- Data Processing:
  Volcano plot, PCA

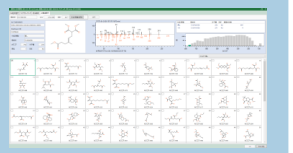


Volcano plot

- Qualitative Analysis:
  EI spectral DB search,
  Soft Ionization
  and AI technology!



EI/FI I. S.



msFineAnalysisAI

Introduction
of a sample

Separation
for compounds

Extraction
for specific compounds

Identification
of specific compounds

If a specific compound found after all the experiments is not listed in the commercially available EI mass spectral DB, how do you identify it?

We can identify it using soft ionization data and AI technology!

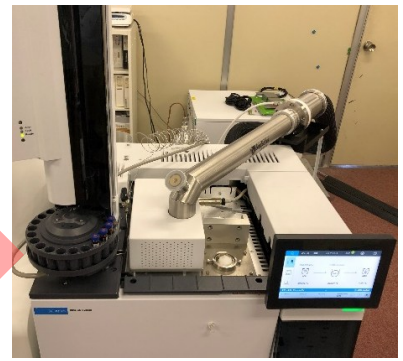
Outlines

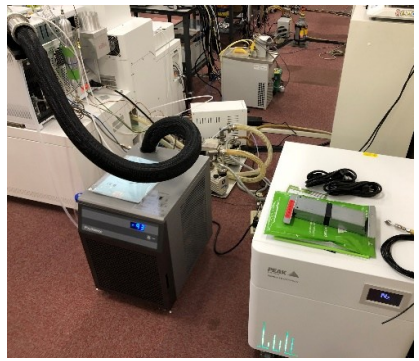
- GC/MS and GCxGC/MS Qualitative Analysis Workflow and Issue
- New Qualitative Analysis Workflow for Unknown Compounds
  - Integrated Analysis: Determine for the Molecular Formula
  - AI Structure Analysis: Elucidation for the Structure Formula
- GCxGC Data Handling on the Software and It's Applications
- Conclusion

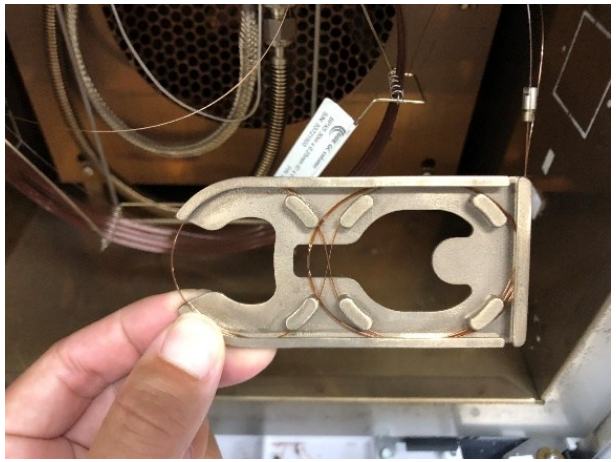
Petroleum Analysis and Metabolomics Analysis

GCxGC	
Instrument	INSIGHT-Thermal modulator (SepSolve)
Modulation	6 sec
GC-TOFMS	
Instrument	JMS-T2000GC (JEOL)
GC column	1st: BPX5, 30 m× 0.25 mm, 0.25 µm
	2nd: Rxi17-SilMS, 3.4 m× 0.15 mm, 0.15 µm
Inlet mode	Petroleum: Split 100:1, 300 °C Metabolomics: Split 5:1, 300 °C
Carrier gas	He, 1.2 mL/min (const.)
Oven prog.	50 °C (2 min) → 4 °C/min → 325 °C (10 min)
Ion source	EI/FI combination ion source
Ionization	EI: 70 eV, 300 µA
	FI: -10kV, 12 mA
Mass range	*m/z* 35-800
Acq. Speed	EI: 50 Hz, FI: 25 Hz

JMS-T2000GC (JEOL)



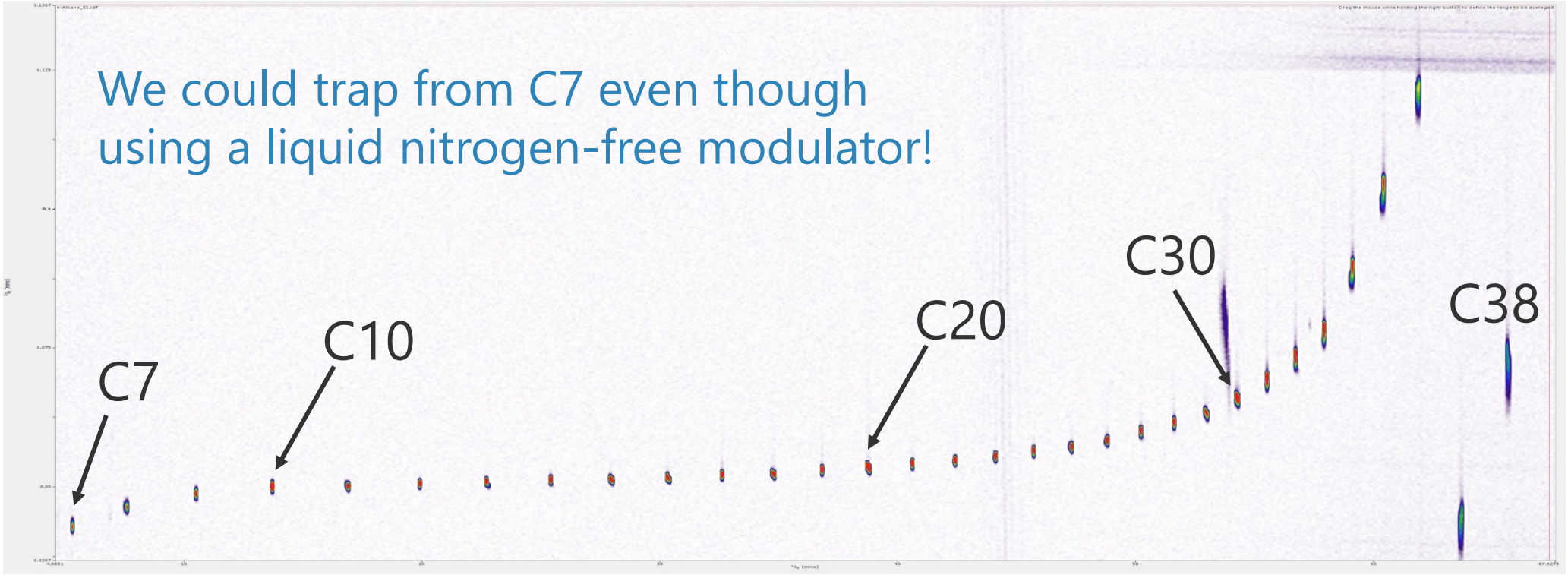




INSIGHT-Thermal modulator (SepSolve)

Thermal Modulator Performance

- Liquid nitrogen-free cryo system
- -100 °C at the chiller, but variable cooling gas flow rate
- We would like to use n-Alkane RT to generate measured RI for a GCxGC data



GCxGC Data Handling on the Software

Qualitative analysis result: Compounds name, CAS #/PubChem ID, Accurate mass analysis result, etc.

Total Workflow for Unknown Compound Analysis by GCxGC!

GCxGC/EI

GCxGC/Soft ionization

Chromatographic peak detection and association between EI and Soft ionization data

NIST library DB search (A)
- Match factor
- Retention index
- Molecular ion confirmation

Passing elemental conditions

Molecular ion analysis (B)
- Molecular ion search
- Accurate mass analysis
- Isotope pattern matching

EI fragment ion analysis
- Accurate mass analysis using molecular formula candidates information (A) and (B)

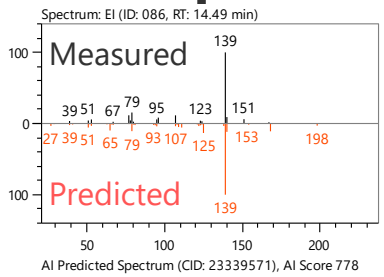
Narrow down the molecular formula candidate using EI fragment ion composition formula coverage.

Integrated Qualitative Analysis Results
- Identification result or Evidence of unknown compound
- Molecular formula information
- Sub-structure information

Integrated Qualitative Analysis Results

AI library search
- 130 million compounds
- Narrow down the candidates by
  - molecular formula information
  - predicted RI (3rd AI)

Sub-structure analysis
- 48 sub-structures
- Supporting interpretation of structural analysis results from AI library



presence

absence

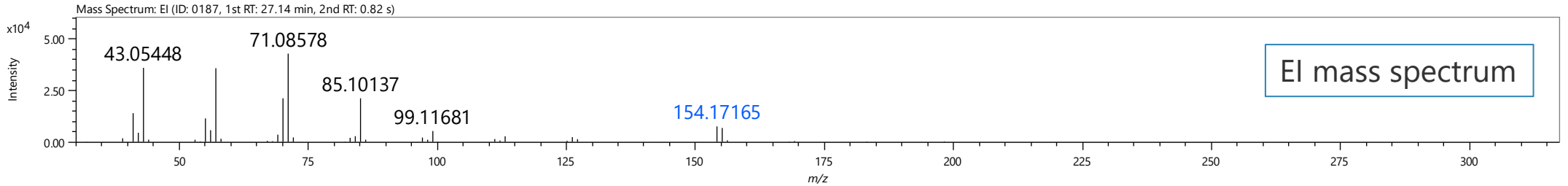
presence

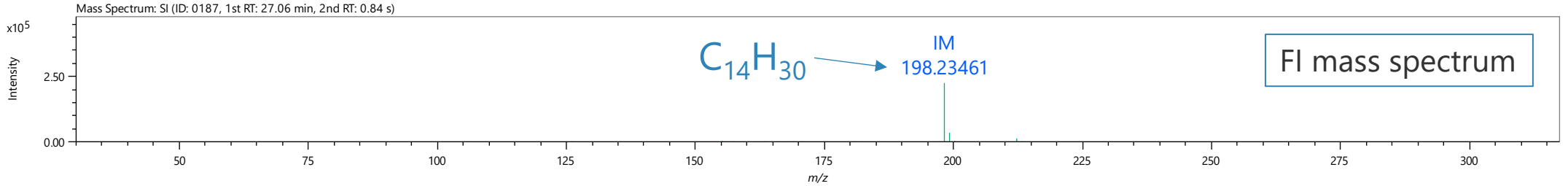
Structure Analysis Results for unknowns
- Structure formula candidates
- Sub-structure prediction

GCxGC Data Handling on the Software

Qualitative analysis result: Compounds name, CAS #/PubChem ID, Accurate mass analysis result, etc.

Integrated Qualitative Analysis Result of ID #187



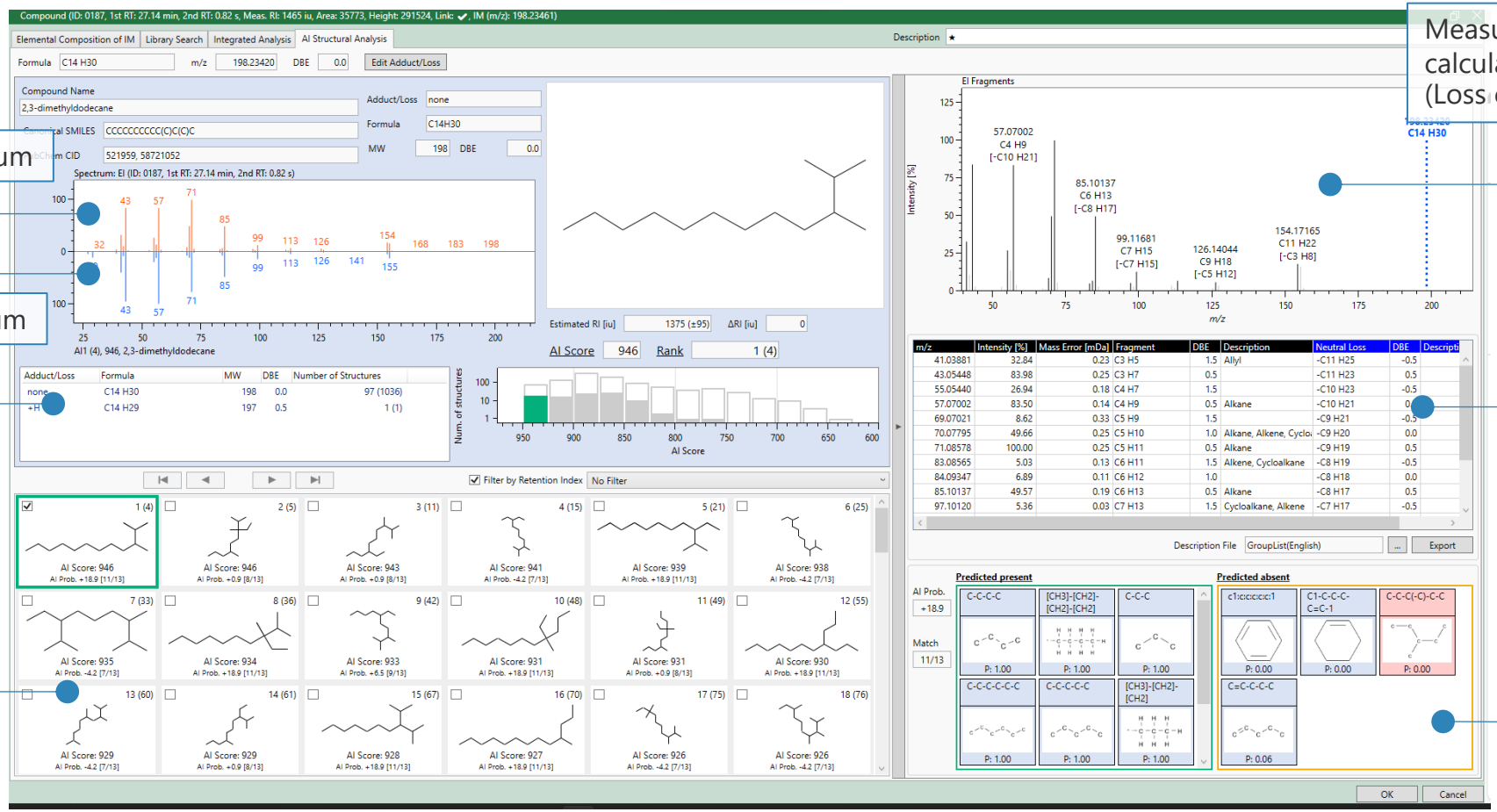


Elemental Composition of IM (m/z: 198.23461)							Integration	Library Search											
#	Formula	DBE	Calculated m/z	Mass Error [mDa]	Isotope Matching	Coverage	Adduct/ Loss	#	Library Name	CAS#	Lib.	Match Factor	Reverse Match Factor	Lib. RI [iu]	ΔRI [iu]	Formula	DBE	MW	EI Base Peak (Lib.)	Coverage
A01	C14 H30	0.0	198.23420	0.41	0.95	100														
								L03	Tetradecane, 5-methyl-	25117-32-2	mainlib	866	866	1453	12	C15 H32	0.0	212	43	100
								L06	Tridecane, 6-propyl-	55045-10-8	mainlib	861	861	1496	31	C16 H34	0.0	226	43	100
								L01	Tridecane, 4-methyl-	26730-12-1	mainlib	891	898	1359	106	C14 H30	0.0	198	43	100
								L02	2,3-Dimethyldodecane	6117-98-2	mainlib	869	874	1353	112	C14 H30	0.0	198	43	100
								L04	Undecane, 3-ethyl-	17312-58-2	mainlib	864	878	1260	205	C13 H28	0.0	184	57	100
								L05	Pentadecane, 6-methyl-	10105-38-1	mainlib	862	862	1547	82	C16 H34	0.0	226	57	100

- L03 and L06 compounds have higher match factor with NIST DB, but they have different molecular formula.
- L01, L02, L04 and L05 compounds have also higher match factor, but they have larger differences with measured RI.

- ID #187 is not listed in NIST DB. = This is unknown compound!
- We have obtained the information such as molecular formula, measured EI mass spectrum and EI fragment ion formulas for structure analysis by AI.

AI Structure Analysis Result of ID #187



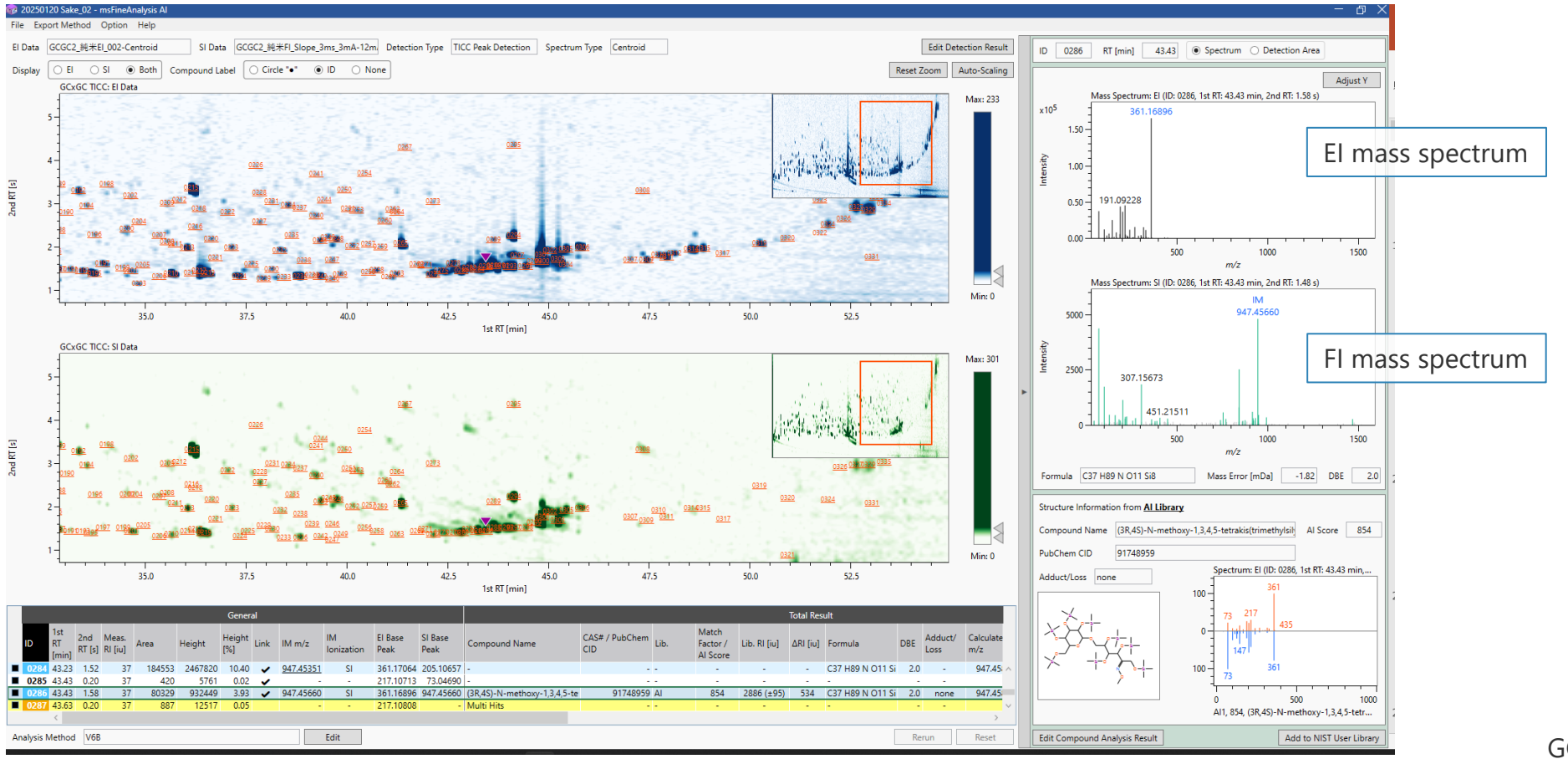
Structural formula Prediction by DL

- Displays ranked structural formulas in list.
- Selecting a structural formula, the information on the top renews.
- Below each structural formula is an AI score indicating the match percentage between structural formula and mass spectrum.

Substructure Prediction by ML

- Displays the information of predicted substructure
- Substructure predicted present on the left, predicted absent on the right.
- Those with blue background are the substructures that match the selected structural formulas. Those with red background are those that do not match.

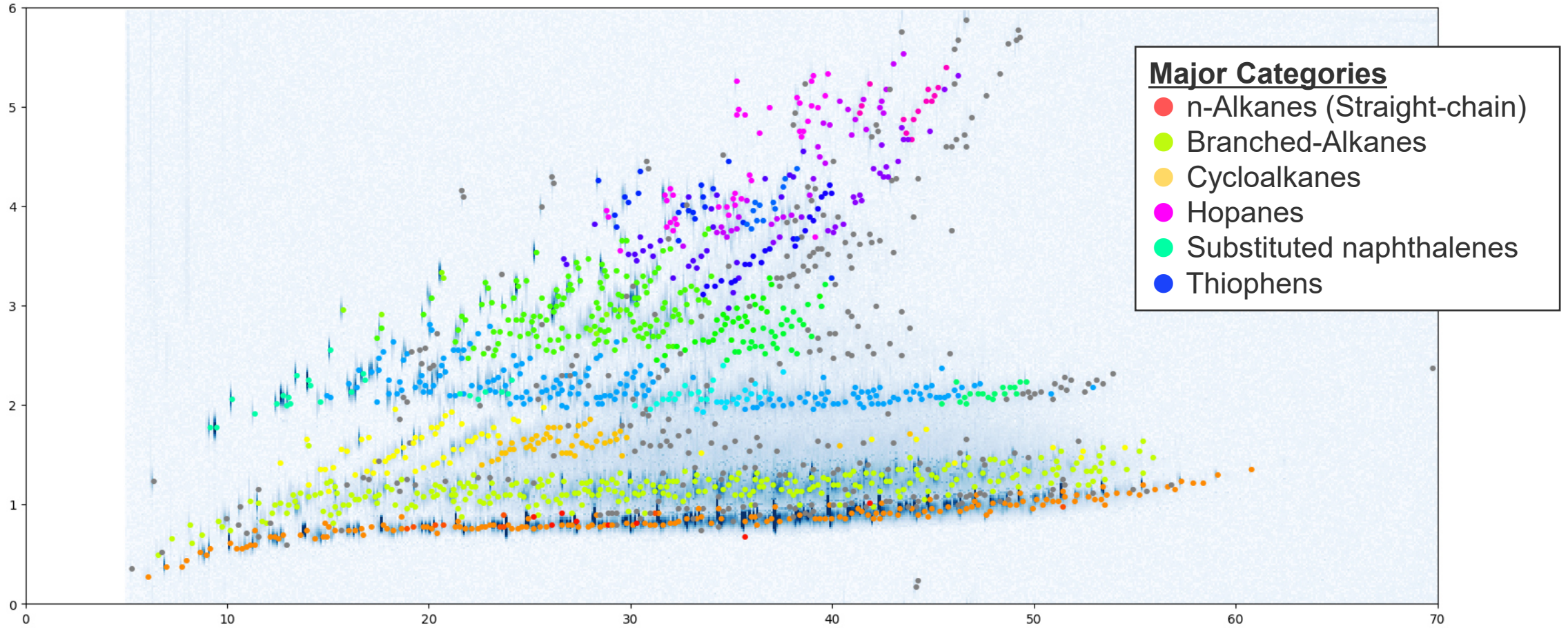
Metabolites Analysis in Japanese Sake



HS-SPME-GC-TOFMS
HS-SPME：2850T (HTA)
GC-TOFMS：JMS-T2000GC (JEOL)

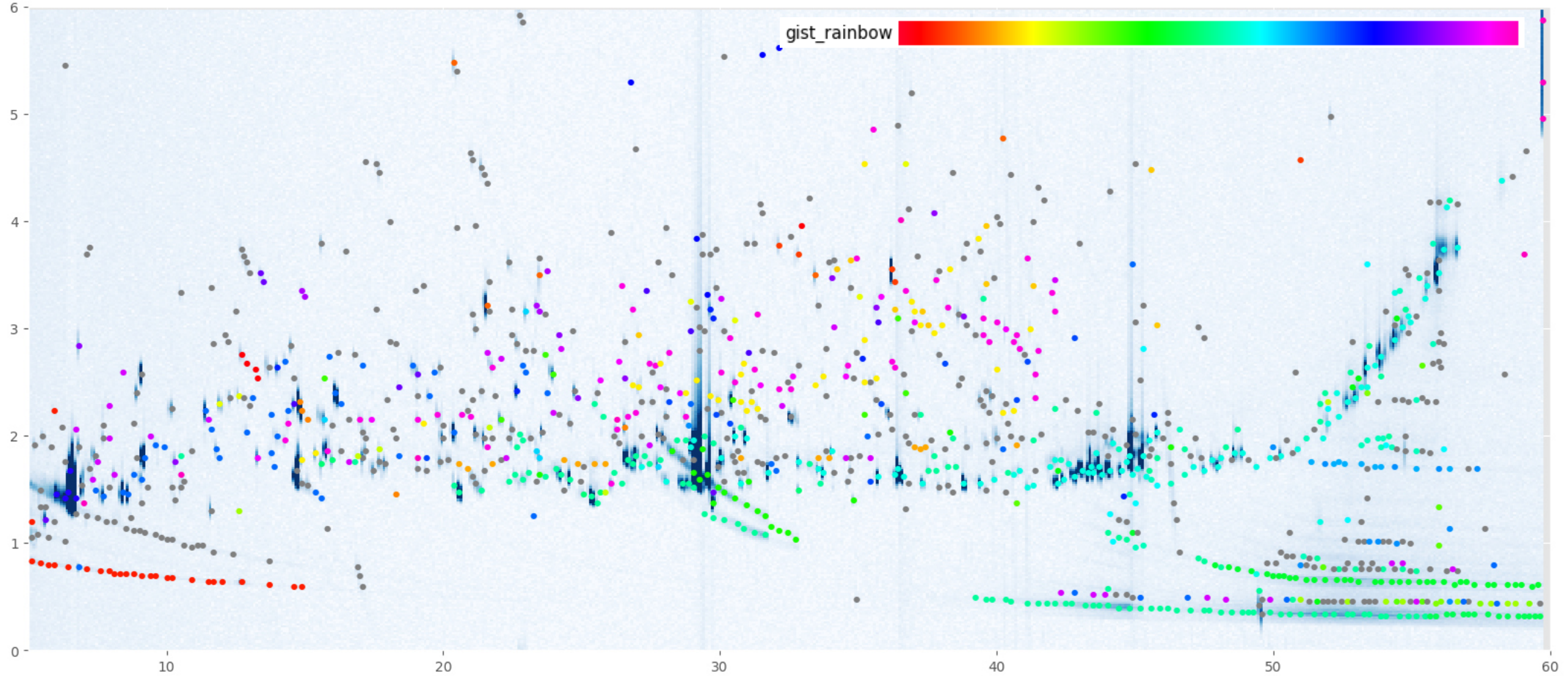
- In the FI mass spectrum, we could obtain the molecular ion at *m/z* 947.457 clearly.
- AI structure analysis automatically elucidated structure formula.
- Elucidated to be a TMS derivative of disaccharide structure based on AI structure analysis results.

Automatic Categorization using measured EI mass spectra



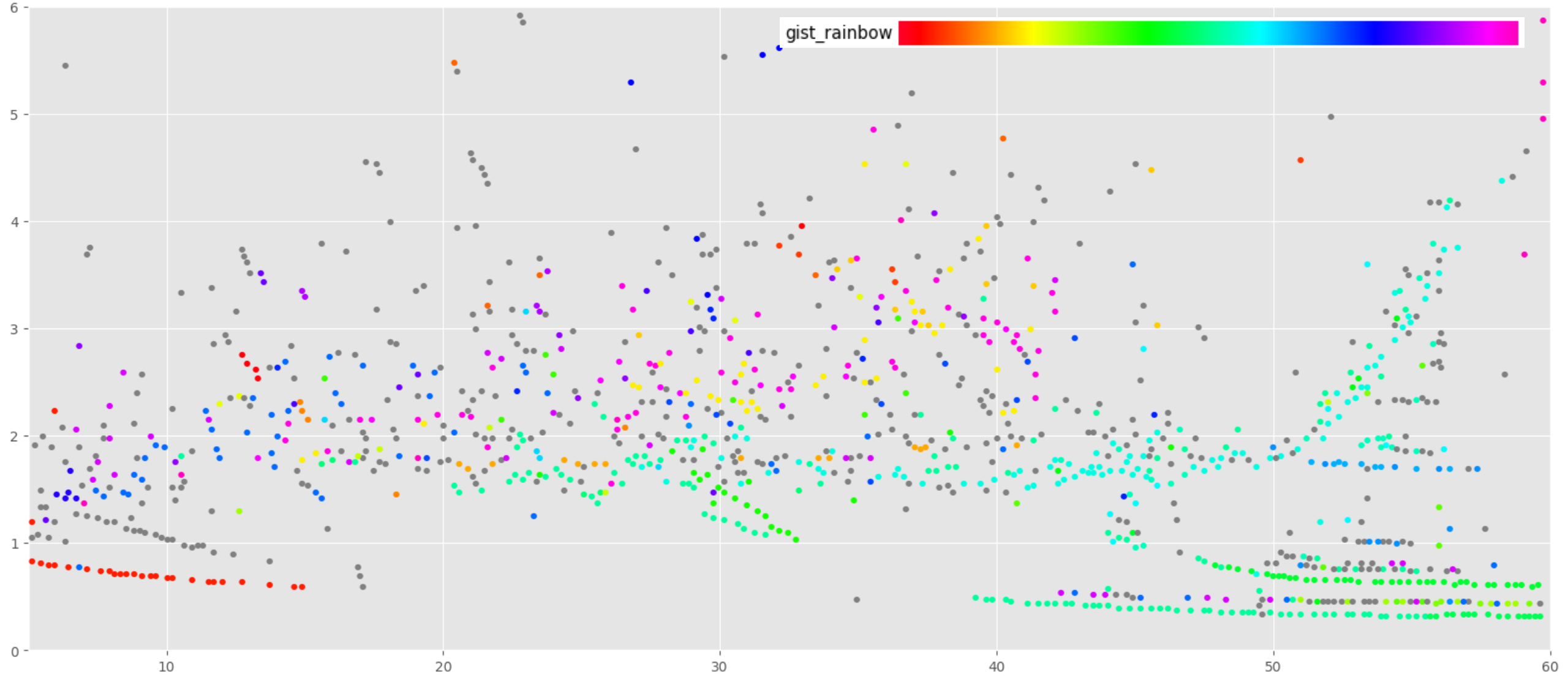
- Enabling us to clearly associate similar structures information from GCxGC chromatography separation and hierarchical clustering using EI mass spectrum.
- Useful information to determine the correct structure from the candidates by AI.

Automatic Categorization for Japanese Sake



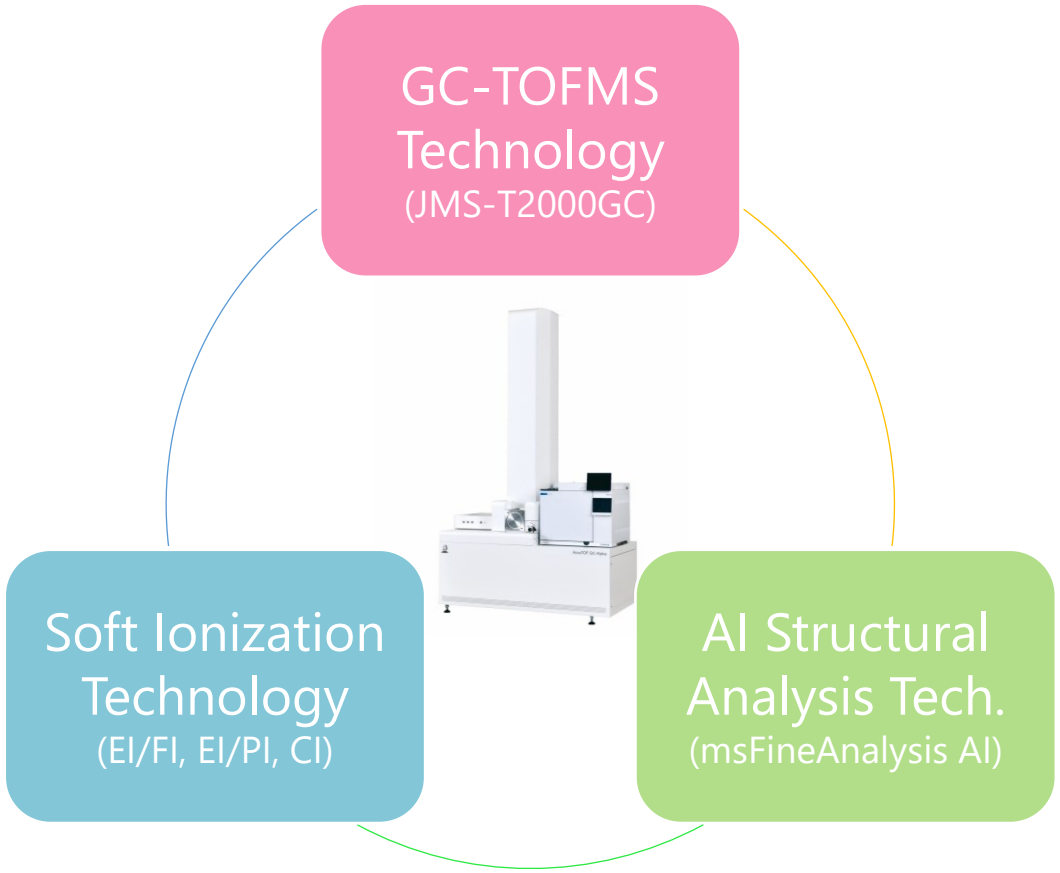
- Maximum number of detection: 1200
- 710 compounds were categorized in more than 20 groups

Automatic Categorization for Japanese Sake



- Maximum number of detection: 1200
- 710 compounds were categorized in more than 20 groups

Conclusion: Three Technologies Provide Structural Analysis Solution



- We made 130 million predicted EI mass spectra DB
  - 25 million *in silico* pyrolyzates library is useful for polymer analysis.
  - We also made 5 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.
- JEOL proposes this solution of the real unknown compounds analysis for all GC-MS fields.