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

Azusa Kubota, Ayumi Kubo, Masaaki Ubukata [JEOL Ltd., Tokyo, Japan]

1. Introduction

- In GCxGC-MS, compound identification is commonly performed by searching the NIST database (DB) using electron ionization (EI) data.
- However, some compounds are not registered in the NIST DB.
- In addition, molecular formula estimation from molecular ions obtained by soft ionization often results in a large number of candidate formulas.
- To address these issues, we developed two machine learning (ML)–based methods for molecular formula recommendation and structural formula estimation.
- In this study, these methods are applied to the analysis of aroma compounds in spices using SPME-GCxGC-TOFMS.

2. Experiments

- **Sample:** 0.25 mg dried cardamom seeds

Dried cardamom

- **GC-HRTOFMS:** JMS-T2000GC (JEOL)
  - SPME Fiber: DVB/CAR/PDMS fiber (20 mm, 50/30 µm) at 50° C for 30 min.
  - GCxGC Thermal modulator: INSIGHT-Thermal (Sepsolve)
  - Modulation period: 6 s
  - Column: BPX5 30 m, 0.25 mm. 0.25 µm × Rxi-17Sil MS 3.4 m, 0.15 mm, 0.15µm
  - Ionization: Electron ionization (EI) and Field ionization (FI).
- **Data processing**: msFineAnalysis AI (JEOL)

Analysis flow

GCxGC-TOFMS measurement

- EI and FI accurate mass measurements by GCxGC-HRTOFMS

NIST DB search

- Accurate mass data were used to refine NIST DB search.

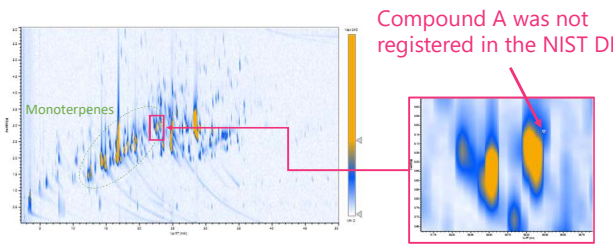
Not registered in the NIST DB

ML-based molecular formula recommendation and structural formula estimation

3. Results

3-1. GCxGC TICC

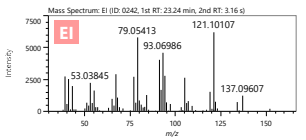
- A total of 518 compounds were detected by GCxGC analysis.
- Compound annotation was based on combined evidence from NIST DB searching, retention index, molecular ion and isotope information from FI, and accurate fragment masses from EI.

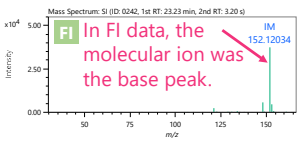


3-2. Why Compound A is not registered in the NIST DB

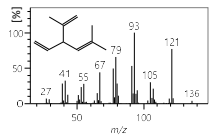
- The top candidate from NIST DB search was Santolina triene (M.F.: 781).
- However, molecular formula of this compound is $C_{10}H_{16}$.
- This does not match the elemental composition estimated from the measured FI mass spectrum.

Mass spectra





NIST DB search result



- Name: Santolina triene
- Formula: $C_{10}H_{16}$
- MW: 136
- Exact MW: 136.12520

Mismatch with the estimated elemental composition.

Elemental composition estimation results

Elemental Composition of IM (m/z: 152.12034)

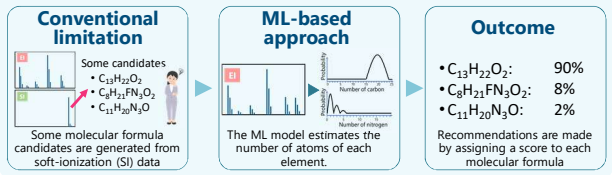
Formula	DBE	Calculated m/z	Mass Error [mDa]
C10 H16 O	3.0	152.11957	0.78
C8 H14 N3	3.5	152.11822	2.12

•Conditions: C: 0-50 units, H: 0-100 units, N: 0-5 units, O: 0-5 units, Mass Error: 2.5 mDa

Two candidates were obtained.

3-3. Molecular formula recommendation result

- **ML1:** An ML model was developed to calculate the probability distribution of each element.



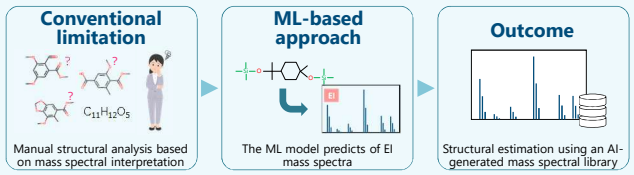
- IM Score shows molecular formula recommendation result.
- $C_{10}H_{16}O$ is considered to be the correct molecular formula.

Elemental Composition of IM (m/z: 152.12034)

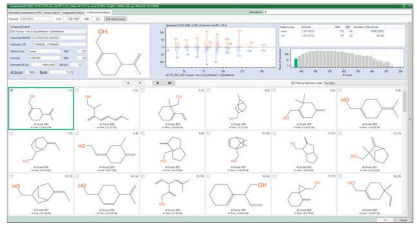
#	Formula	DBE	Calculated m/z	Mass Error [mDa]	Isotope Matching	Coverage	IM Score [%]
★ A01	C10 H16 O	3.0	152.11957	0.78	0.88	100	100
A02	C8 H14 N3	3.5	152.11822	2.12	0.81	80	0

3-4. Structural formula estimation result

- **ML2:** An ML model was developed to predict mass spectra from molecular structures.



- Structural formula candidates for terpenoids (e.g., monoterpene alcohols) were obtained.
- Our method enables rapid estimation of structural formulas.



4. Conclusion

- We reported two machine learning (ML)-based methods for molecular formula recommendation and structural formula estimation, and applied them to the analysis of aroma compounds.
- Our method is considered effective for rapid structural estimation of unknown compounds in GCxGC-MS measurements.