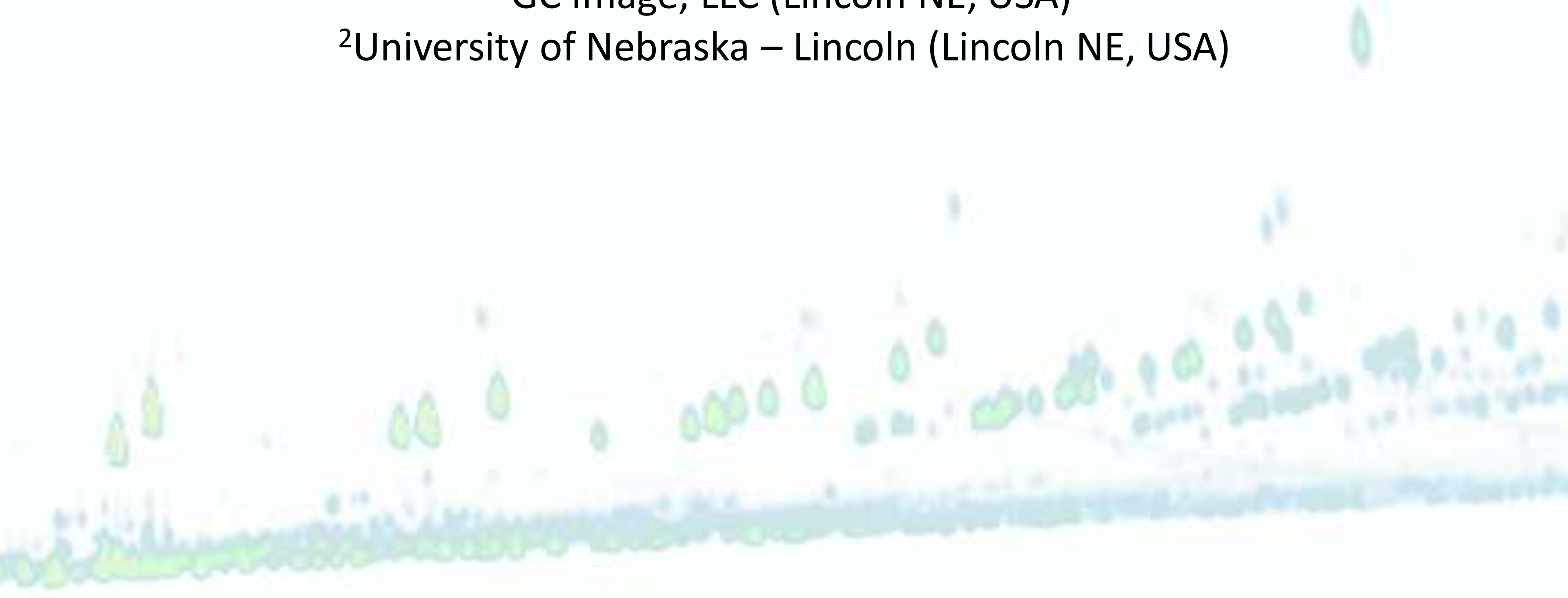
Advanced Software Tools for Plant Substances Analyses using Comprehensive Two-Dimensional Chromatography with High-Resolution Mass Spectrometry

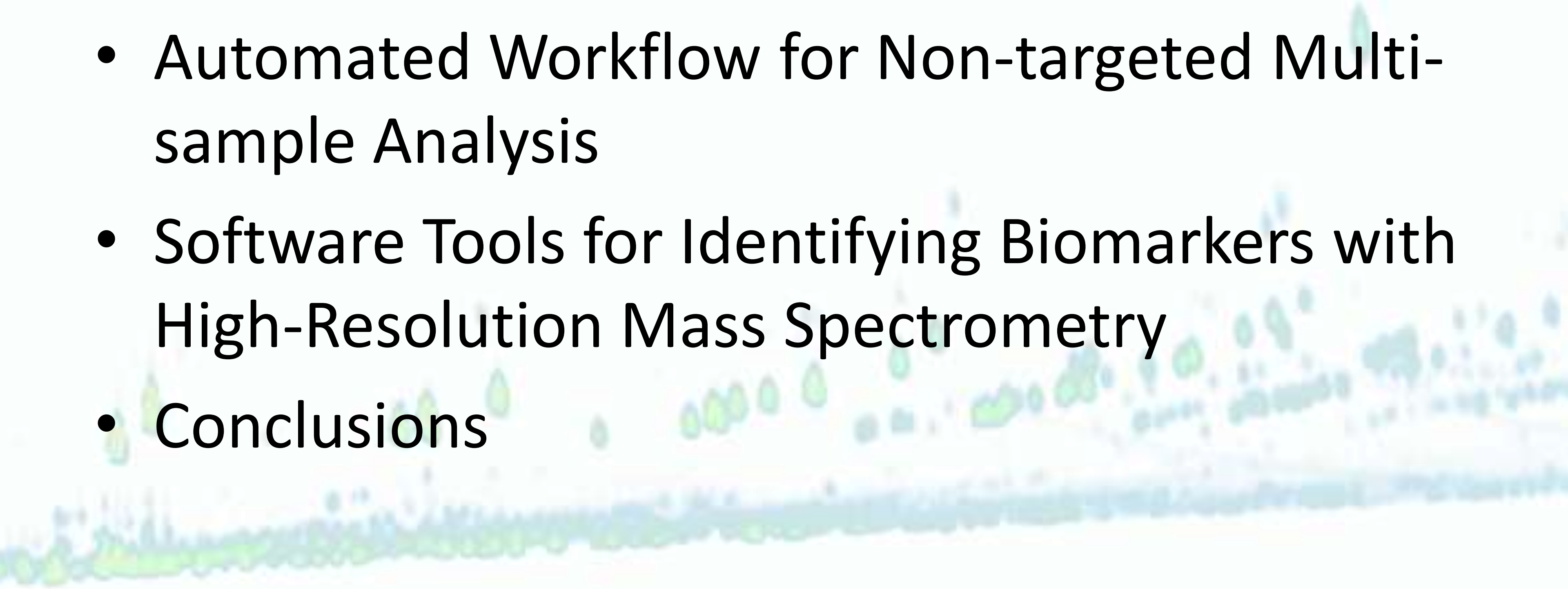
Qingping Tao[1], Stephen E. Reichenbach[1,2], Trevor S. Janke[1]

[1]GC Image, LLC (Lincoln NE, USA)
[2]University of Nebraska – Lincoln (Lincoln NE, USA)



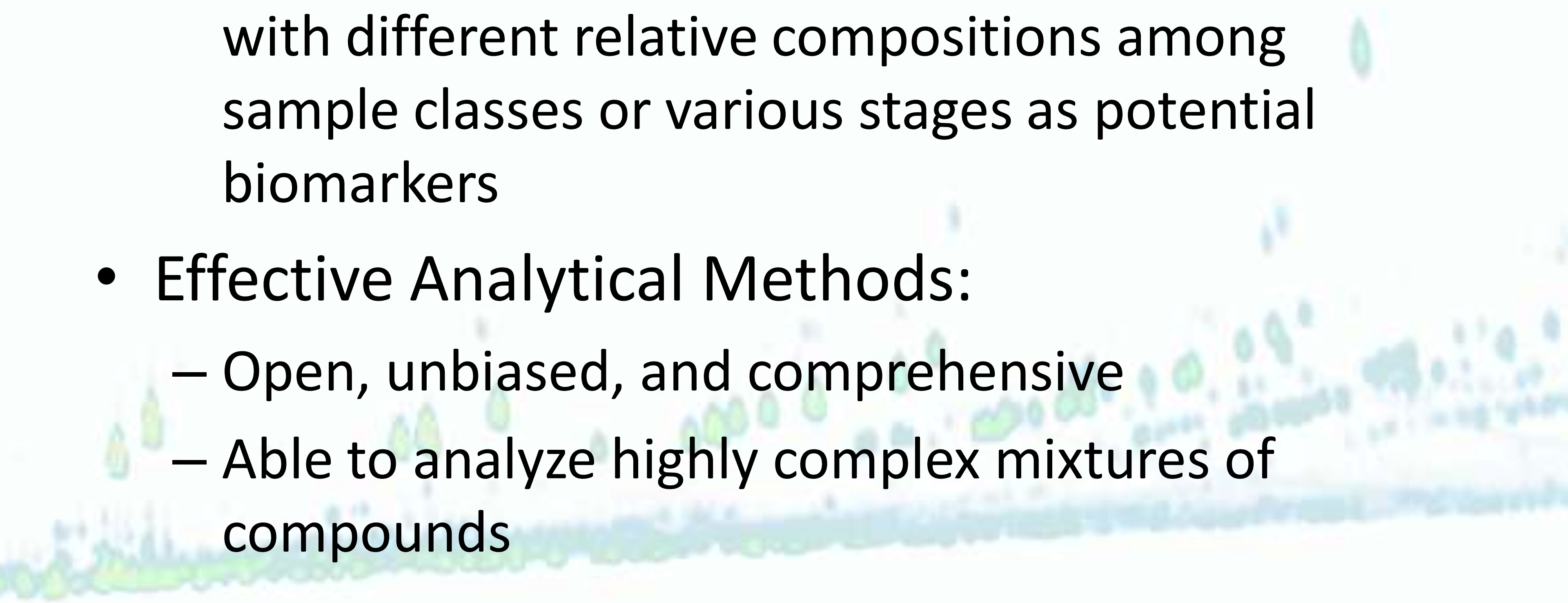
Outline

- Introduction
- Multi-Sample Analysis: Workflows and Challenges
- Automated Workflow for Non-targeted Multi-sample Analysis
- Software Tools for Identifying Biomarkers with High-Resolution Mass Spectrometry
- Conclusions



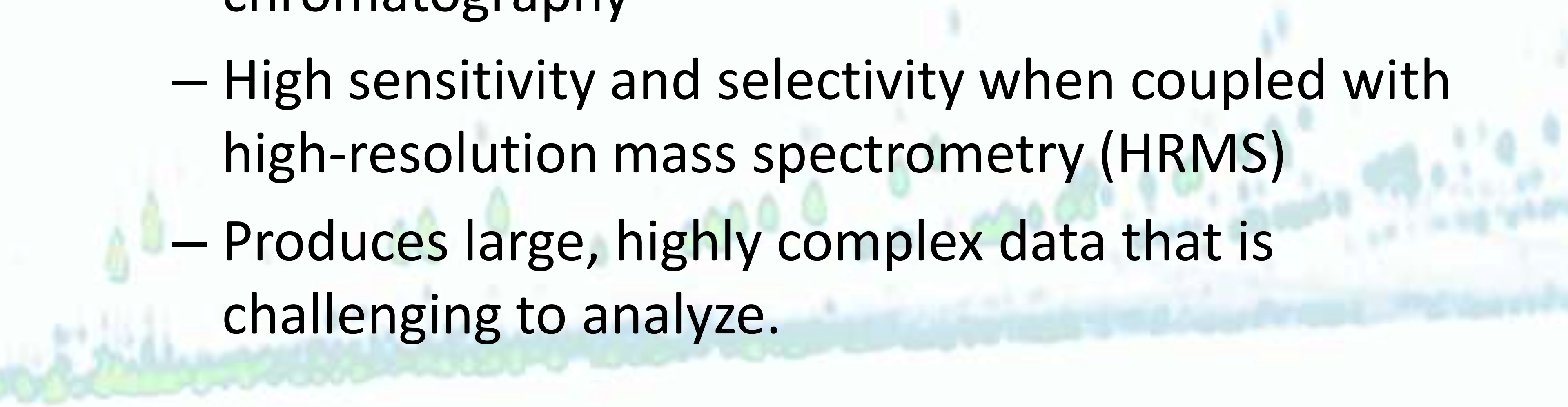
Introduction: Goals

- Goal: Understand the states or processes of biological systems of plants
  - Discover & document compounds/metabolites with different relative compositions among sample classes or various stages as potential biomarkers
- Effective Analytical Methods:
  - Open, unbiased, and comprehensive
  - Able to analyze highly complex mixtures of compounds



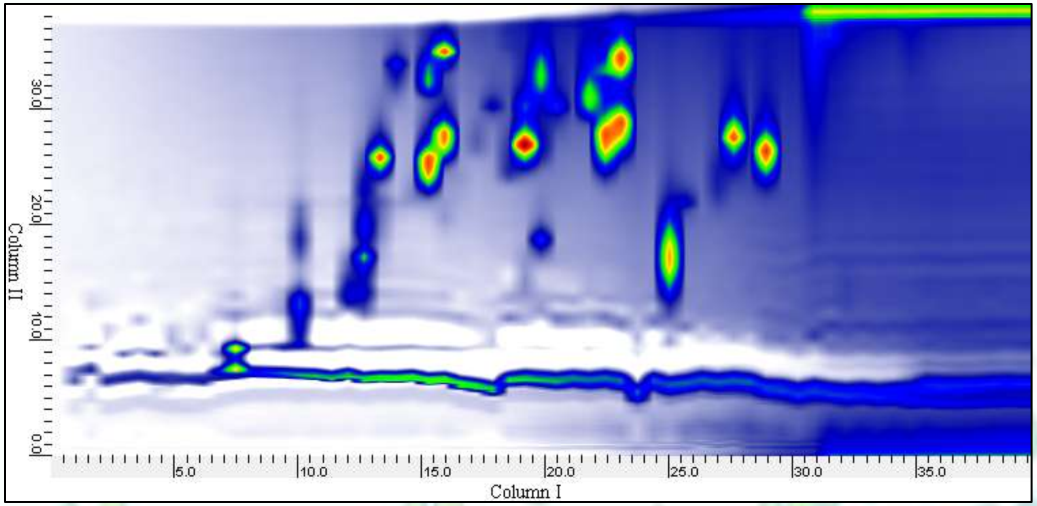
Introduction: Goals

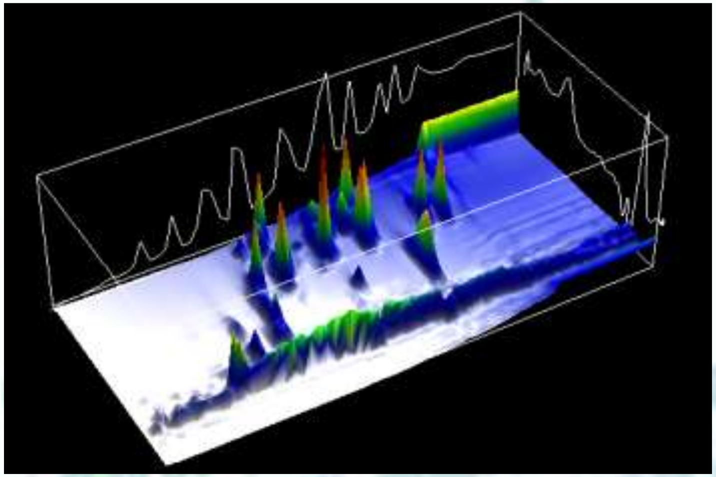
- Comprehensive two-dimensional gas and liquid chromatography (GCxGC and LCxLC)
  - Much greater separation capacity and signal-to-noise than traditional one-dimensional chromatography
  - High sensitivity and selectivity when coupled with high-resolution mass spectrometry (HRMS)
  - Produces large, highly complex data that is challenging to analyze.



Introduction: Two-Dimensional Chromatogram

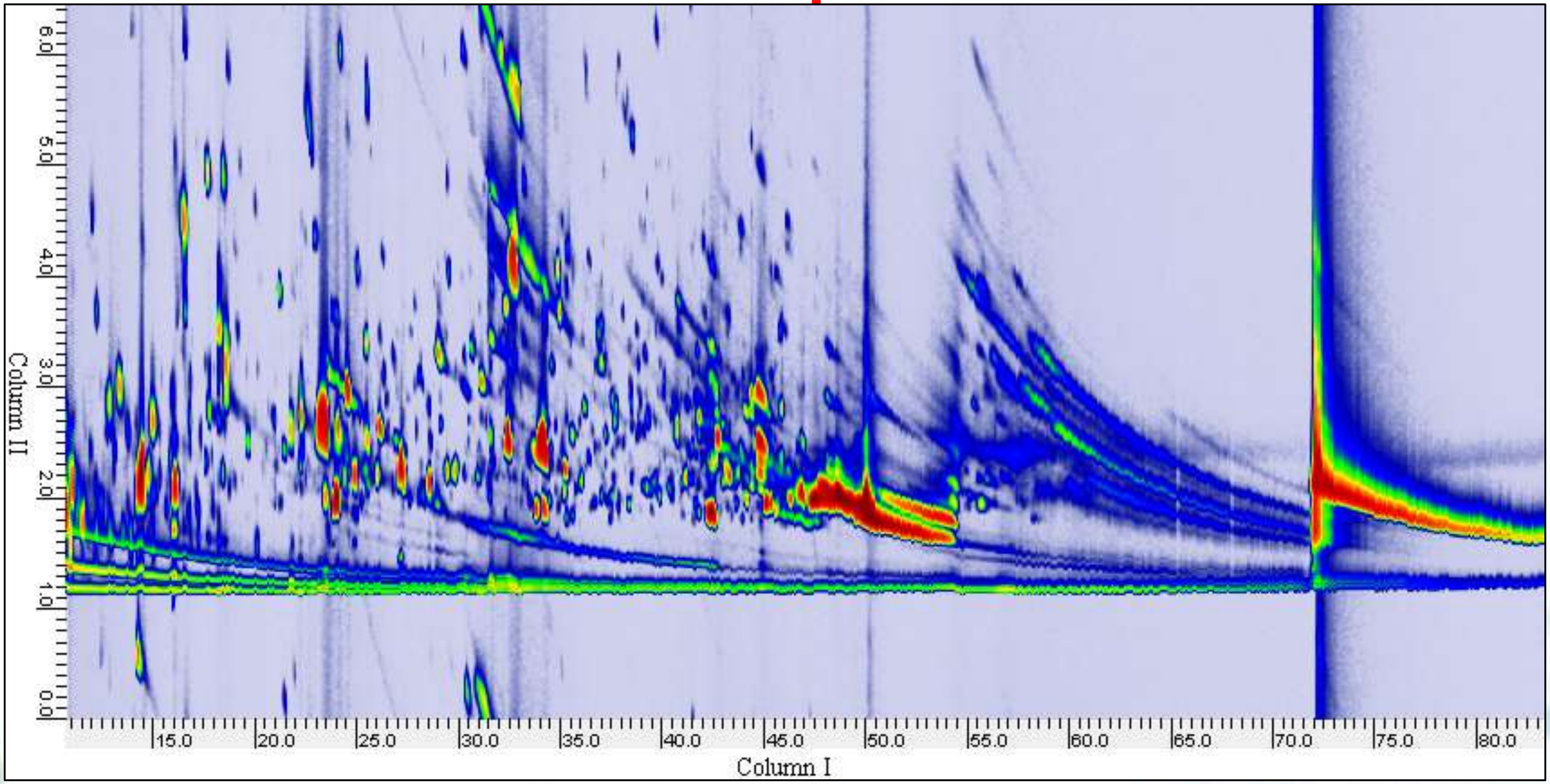
- Peaks are two-dimensional, with several Column II chromatograms cross each peak.





A LCxLC chromatogram of a standard mixture of polyphenolic compounds acquired by Agilent 1290 Infinity 2D-LC (S.E. Reichenbach and E. Naegele. Agilent Application Notes, 2013).

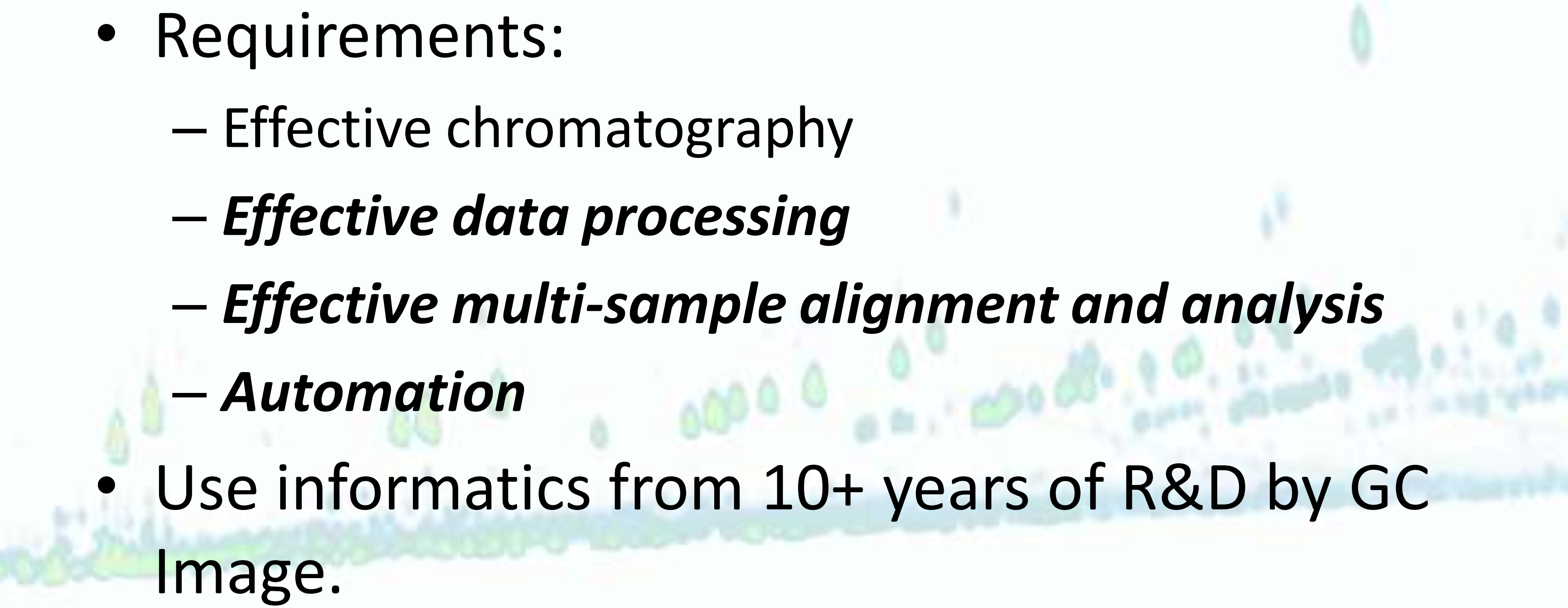
2D Chromatogram: Another Example



A GCxGC chromatogram of a wild type strain of rice blast fungus acquired by Agilent 7890B/ZOEX ZX2 thermal modulation system coupled with Agilent 7200 Q-TOF (Sofia Aronova *et al., ISCC* 2014)

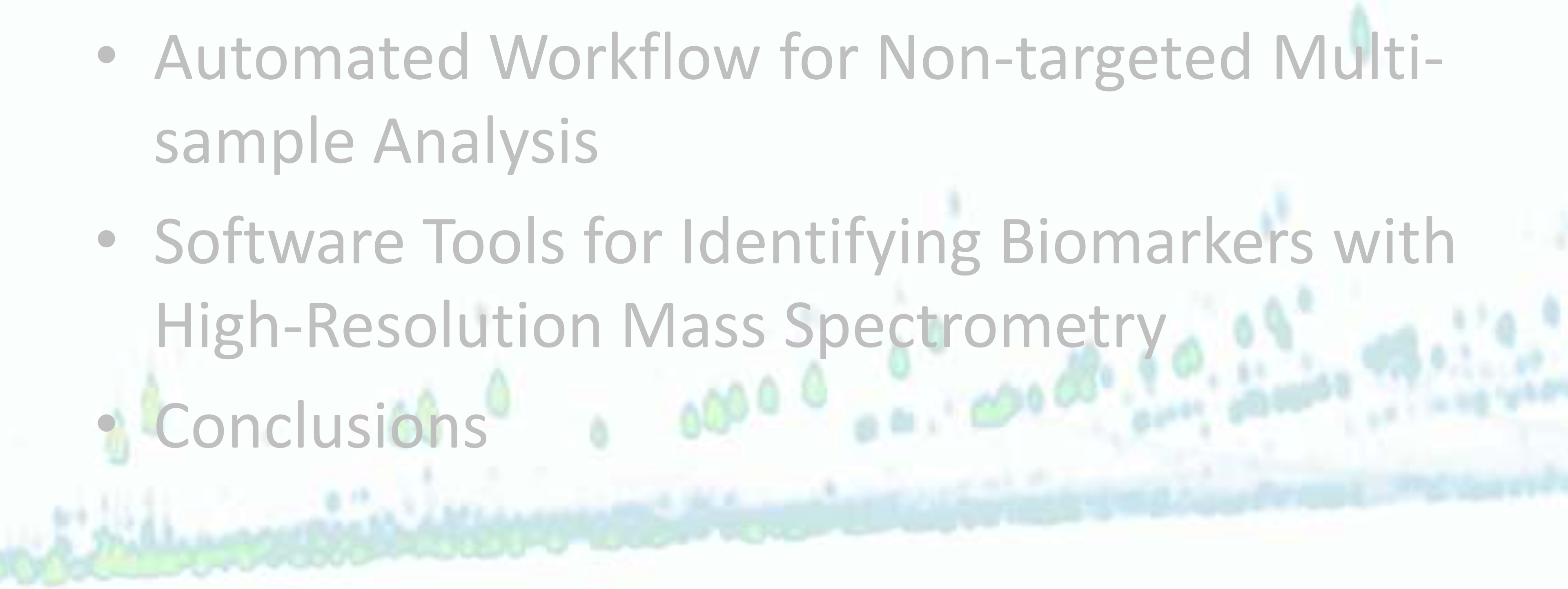
Introduction: Data Analysis

- Challenge: Comprehensive analysis of many compounds from multiple chromatograms
- Requirements:
  - Effective chromatography
  - ***Effective data processing***
  - ***Effective multi-sample alignment and analysis***
  - ***Automation***
- Use informatics from 10+ years of R&D by GC Image.



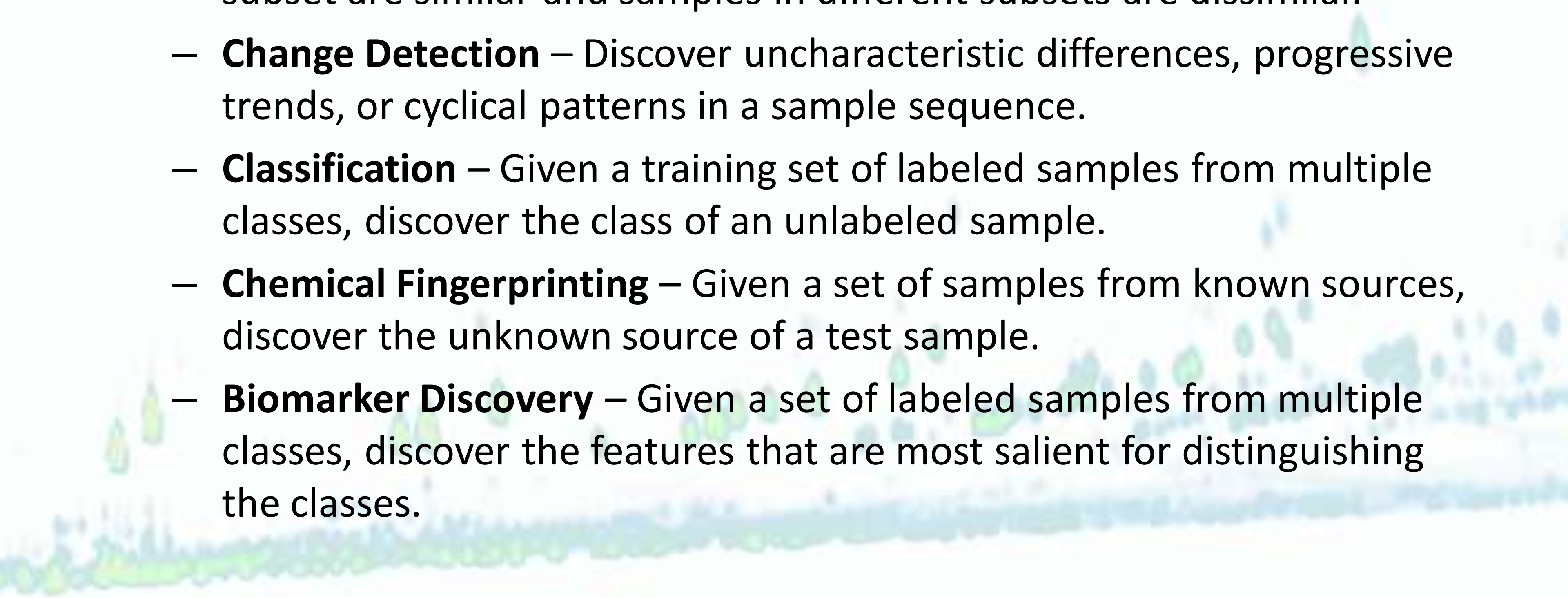
Outline

- Introduction
- Multi-Sample Analysis: Workflows and Challenges
- Automated Workflow for Non-targeted Multi-sample Analysis
- Software Tools for Identifying Biomarkers with High-Resolution Mass Spectrometry
- Conclusions

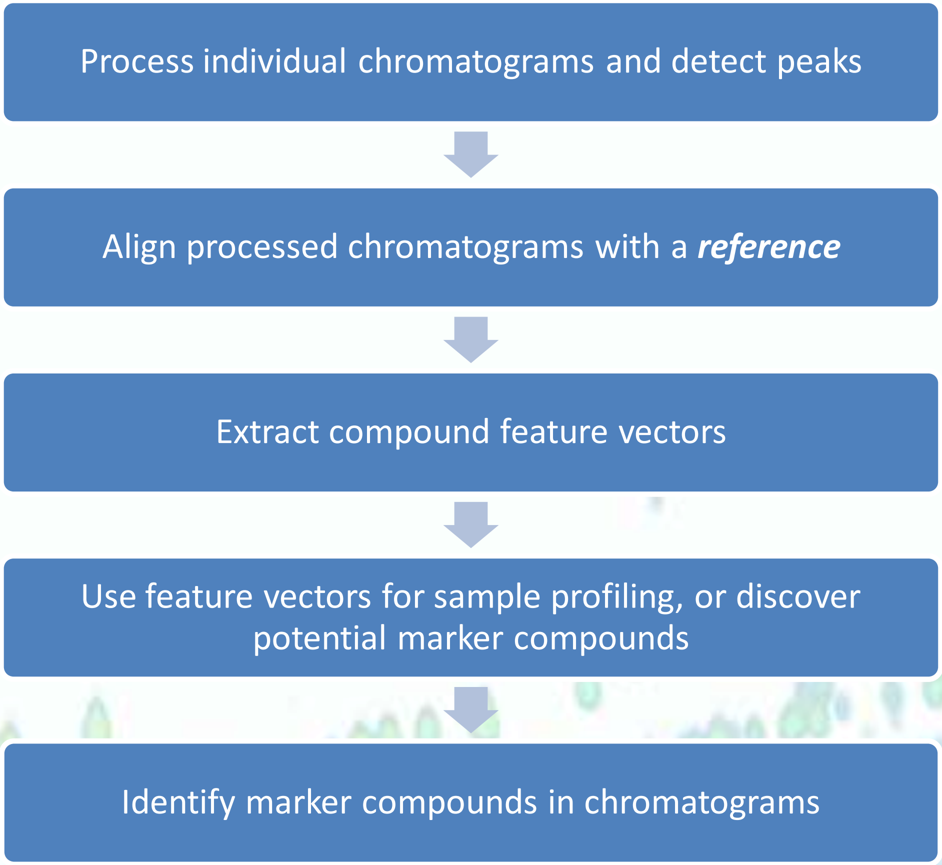


Multi-sample Analysis

- **Applications:**
  - **Clustering** – Discover sample subsets, such that samples in the same subset are similar and samples in different subsets are dissimilar.
  - **Change Detection** – Discover uncharacteristic differences, progressive trends, or cyclical patterns in a sample sequence.
  - **Classification** – Given a training set of labeled samples from multiple classes, discover the class of an unlabeled sample.
  - **Chemical Fingerprinting** – Given a set of samples from known sources, discover the unknown source of a test sample.
  - **Biomarker Discovery** – Given a set of labeled samples from multiple classes, discover the features that are most salient for distinguishing the classes.



Multi-Sample Analysis: Workflow



Process individual chromatograms and detect peaks

Align processed chromatograms with a ***reference***

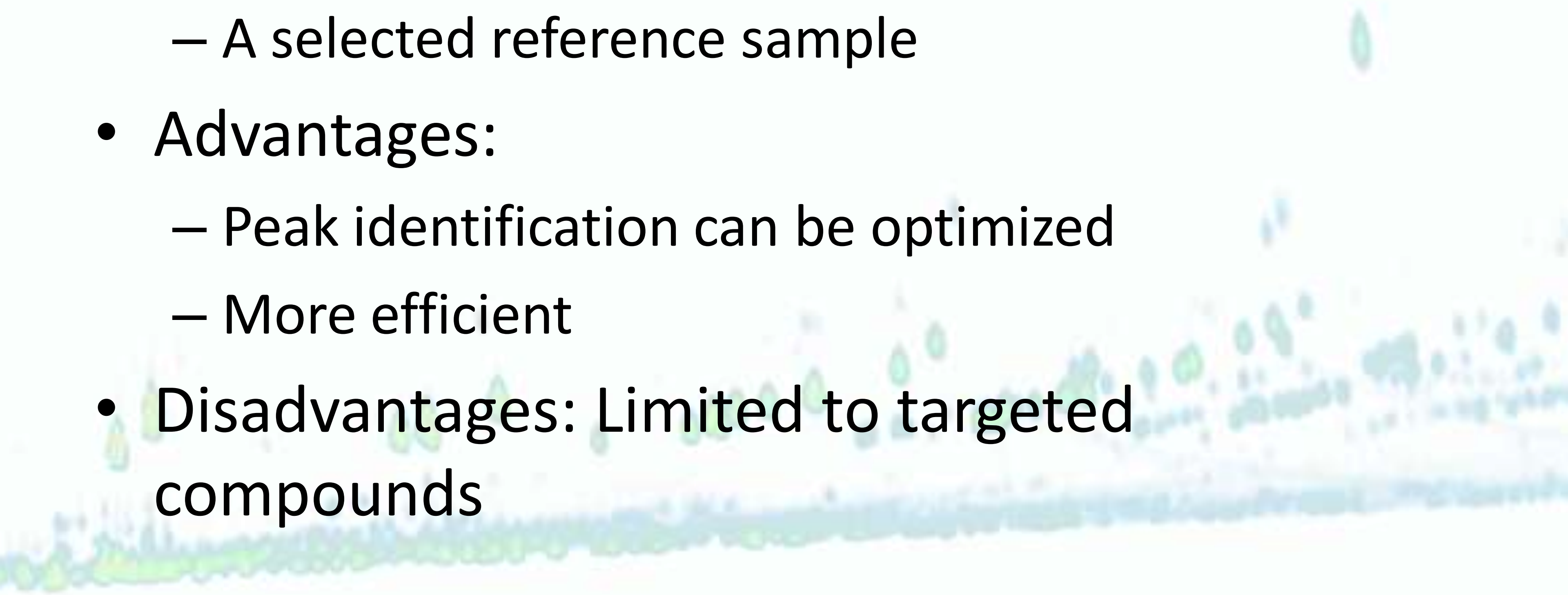
Extract compound feature vectors

Use feature vectors for sample profiling, or discover potential marker compounds

Identify marker compounds in chromatograms

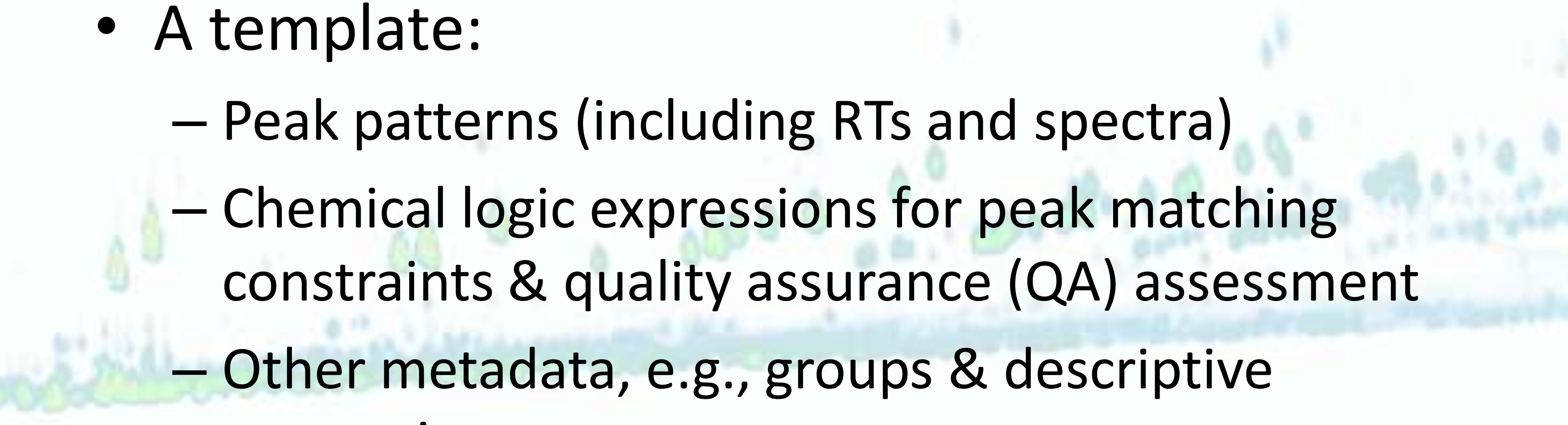
Multi-Sample Analysis: Targeted Analysis

- Reference
  - A standard mixture of known compounds
  - A selected reference sample
- Advantages:
  - Peak identification can be optimized
  - More efficient
- Disadvantages: Limited to targeted compounds



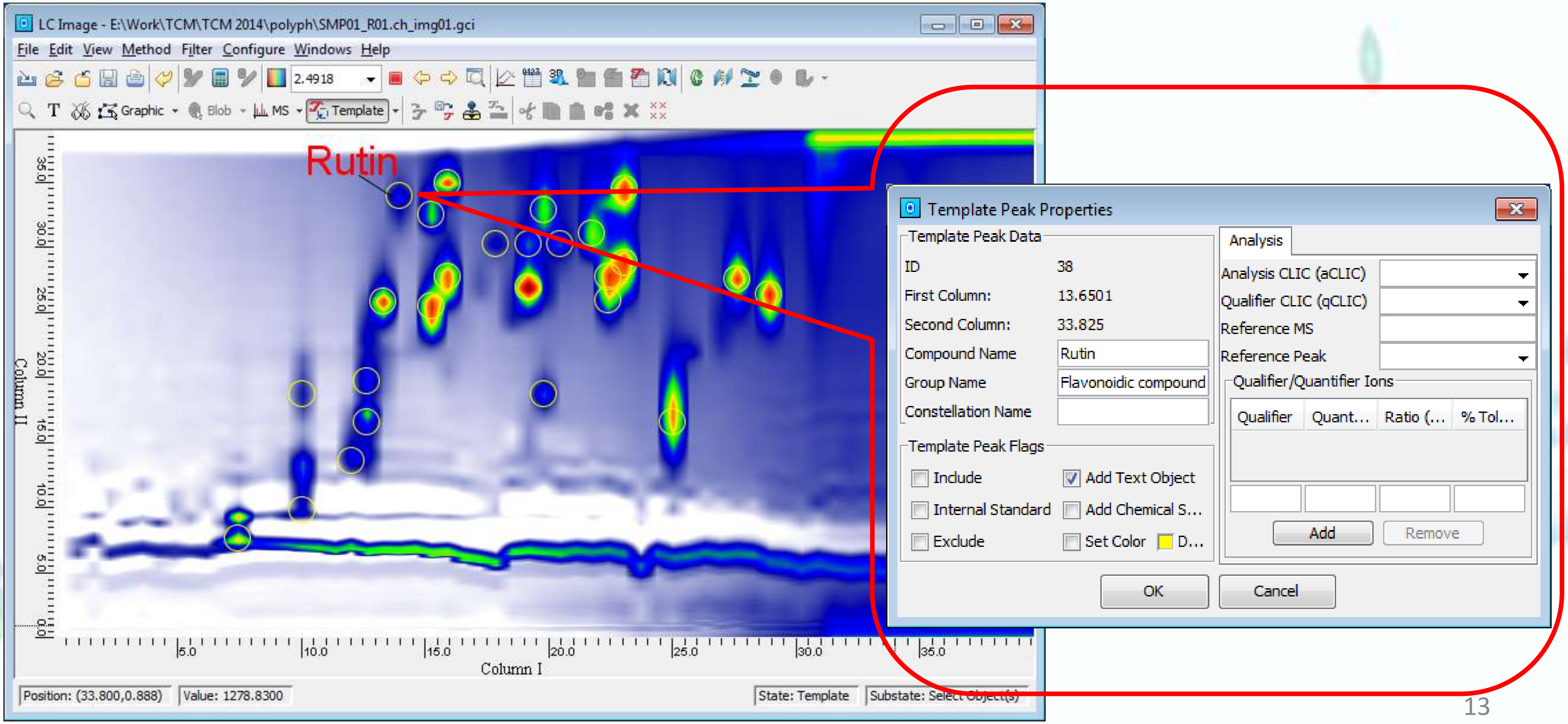
Targeted Analysis: Template Matching

- Template Matching from GC Image:
  - A powerful tool for automated identification
  - Use advanced pattern recognition to identify peak pattern in a new chromatogram
- A template:
  - Peak patterns (including RTs and spectra)
  - Chemical logic expressions for peak matching constraints & quality assurance (QA) assessment
  - Other metadata, e.g., groups & descriptive annotations



Template Matching: Example

- A template that contains 26 polyphenolic compounds from the standard mixture



LC Image - E:\Work\TCM\TCM 2014\polyph\SMP01_R01.ch_img01.gci

Position: (33.800,0.888)

Value: 1278.8300

State: Template

Substate: Select Object(s)

Template Peak Properties

Template Peak Data

ID	38
First Column:	13.6501
Second Column:	33.825
Compound Name	Rutin
Group Name	Flavonoidic compound
Constellation Name	

Template Peak Flags

☐ Include ☑ Add Text Object

☐ Internal Standard ☐ Add Chemical S...

☐ Exclude ☐ Set Color 🟨 D...

Analysis

Analysis CLIC (aCLIC)

Qualifier CLIC (qCLIC)

Reference MS

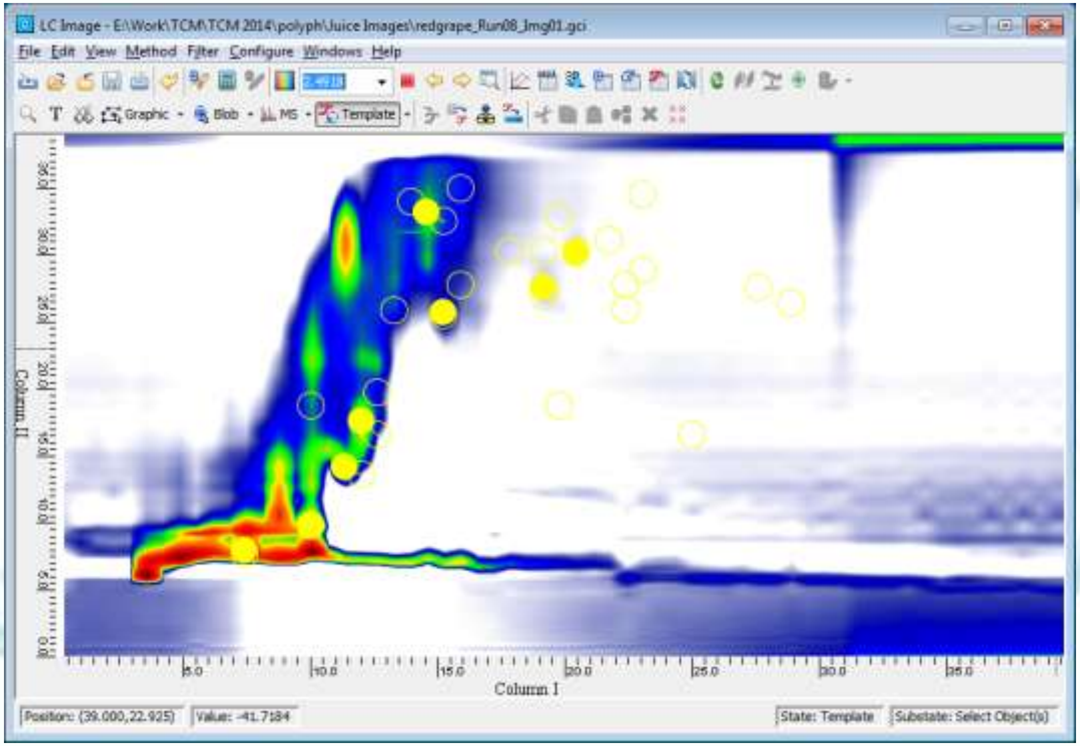
Reference Peak

Qualifier/Quantifier Ions

Qualifier	Quant...	Ratio (...	% Tol...

Template Matching: Example

- Match the template to a grape juice sample.
  - – Some peaks are matched shown as solid circle
  - – Others are unmatched:

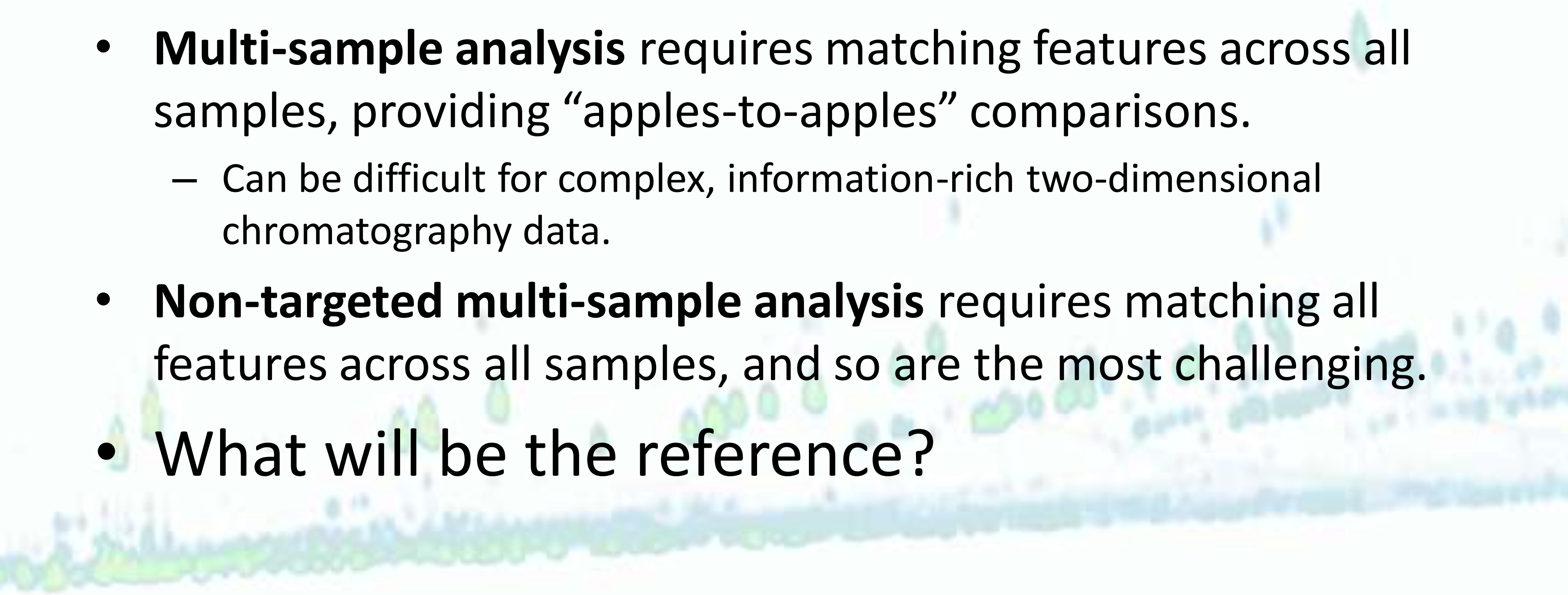

LC Image - E:\Work\TCM\TCM 2014\polyph\Juice Images\redgrape_Run08_Img01.gci
File Edit View Method Filter Configure Windows Help
Graphic
Blob
MS
Template
0.0
5.0
10.0
15.0
20.0
25.0
30.0
35.0
Column II
Column I

Position: (39.000,22.925) Value: -41.7184 State: Template Substate: Select Object(s)

- Not exist
- Undetected
  (Trace or co-eluted peaks)
- Unmatched
  (RT or spectral mismatch)
- May mismatch
  (nearer peak)

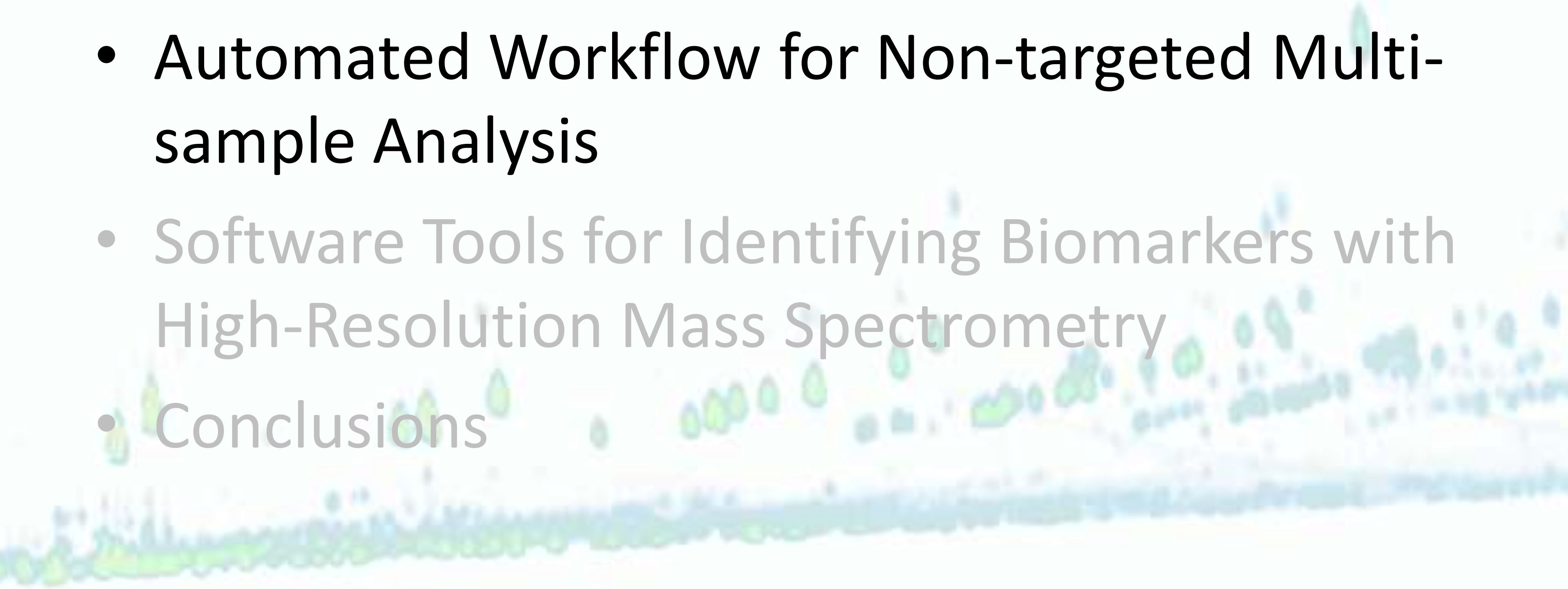
Multi-Sample Analysis: Non-targeted Analysis

- **Non-targeted analysis** requires features to characterize all compounds, not just targeted or selected compounds.
  - Characterize with retention times, intensity, mass spectrum.
- **Multi-sample analysis** requires matching features across all samples, providing “apples-to-apples” comparisons.
  - Can be difficult for complex, information-rich two-dimensional chromatography data.
- **Non-targeted multi-sample analysis** requires matching all features across all samples, and so are the most challenging.
- What will be the reference?



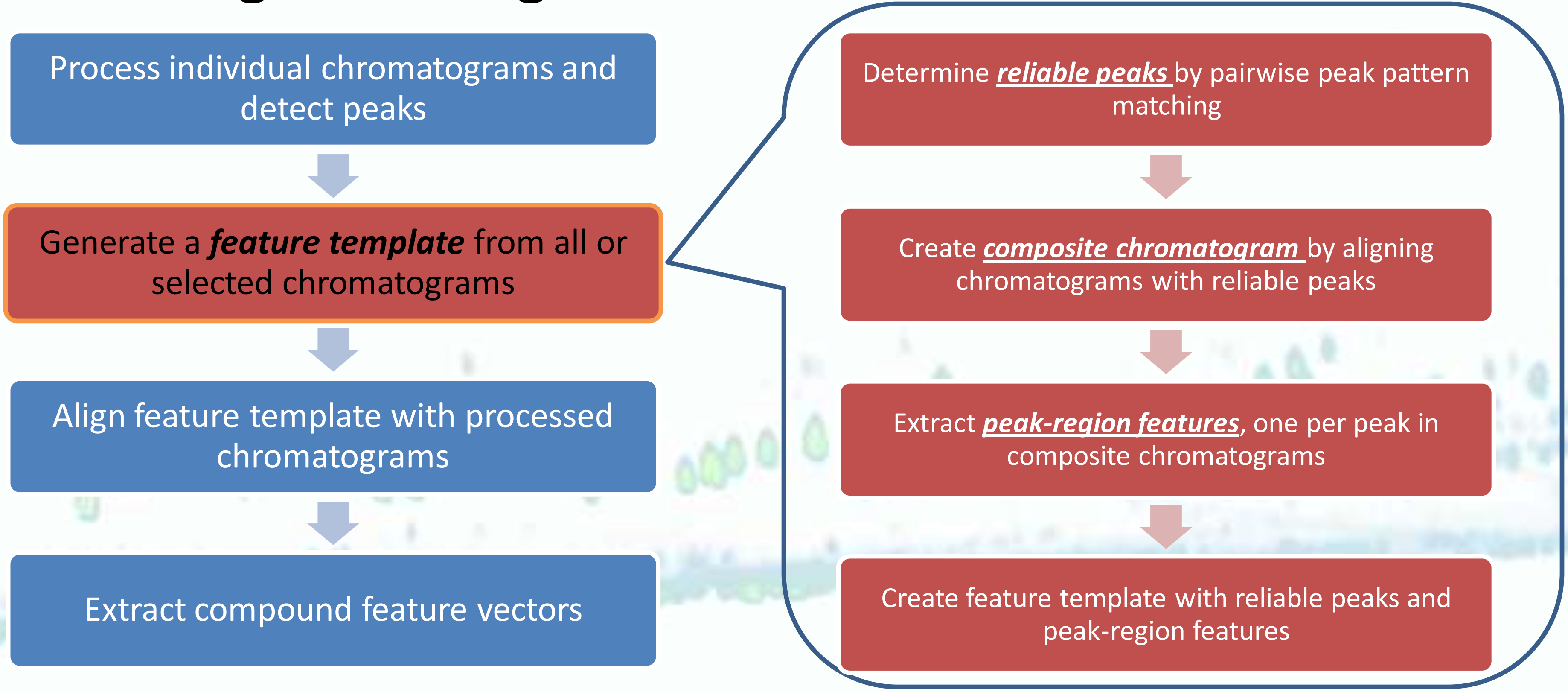
Outline

- Introduction
- Multi-Sample Analysis: Workflows and Challenges
- Automated Workflow for Non-targeted Multi-sample Analysis
- Software Tools for Identifying Biomarkers with High-Resolution Mass Spectrometry
- Conclusions



Non-targeted Multi-sample Analysis

- Automated workflow supported by GC Image's Image Investigator™ software:



Process individual chromatograms and detect peaks

Generate a ***feature template*** from all or selected chromatograms

Align feature template with processed chromatograms

Extract compound feature vectors

Determine ***reliable peaks*** by pairwise peak pattern matching

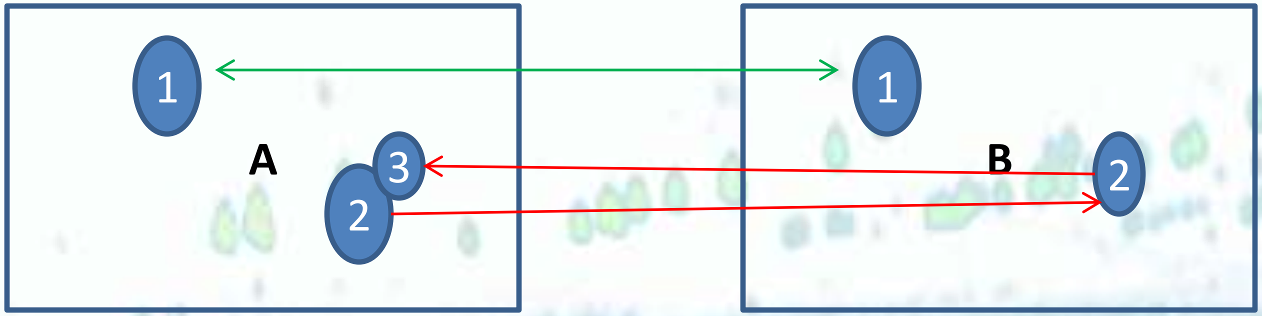
Create ***composite chromatogram*** by aligning chromatograms with reliable peaks

Extract ***peak-region features***, one per peak in composite chromatograms

Create feature template with reliable peaks and peak-region features

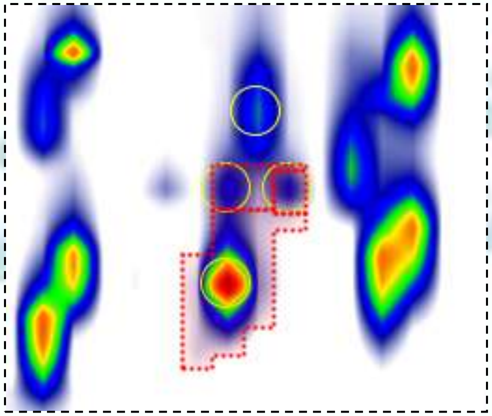
Feature Template: Reliable Peaks

- Reliable peaks are determined from the bidirectional pairwise matching of all possible pairs of chromatograms (Reichenbach *et al.*, *Anal Chem*, 85:4974, 2013).



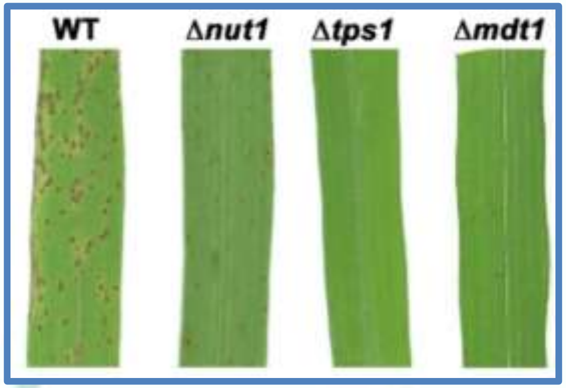
Feature Template: Peak-Region Features

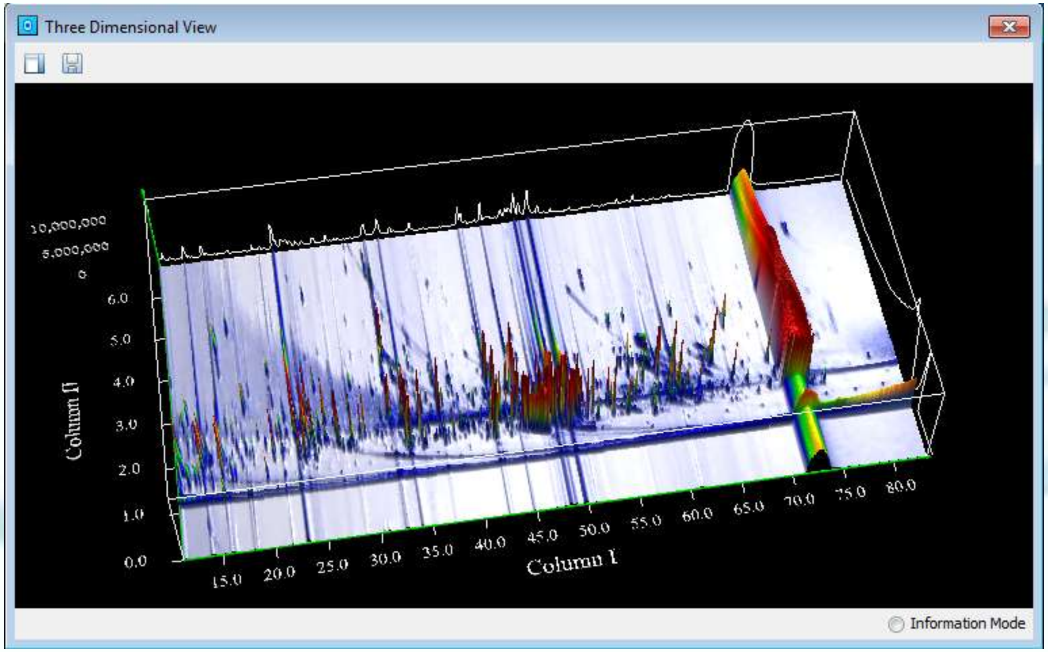
- Goal: Define a region for every peak in every chromatogram
  - **Composite Chromatogram:** All chromatograms are aligned and combined (e.g., by addition) to form a single composite chromatogram that is reflective of all of the constituents in all samples.
  - Peak-region features are delineated by peak detection in the composite chromatogram (Reichenbach et al., J Chromatogr A, 2012)
- Peak-Region features are comprehensive, accounting for every analyte, and feature matching is implicitly performed by the retention-time mapping.



Example Data

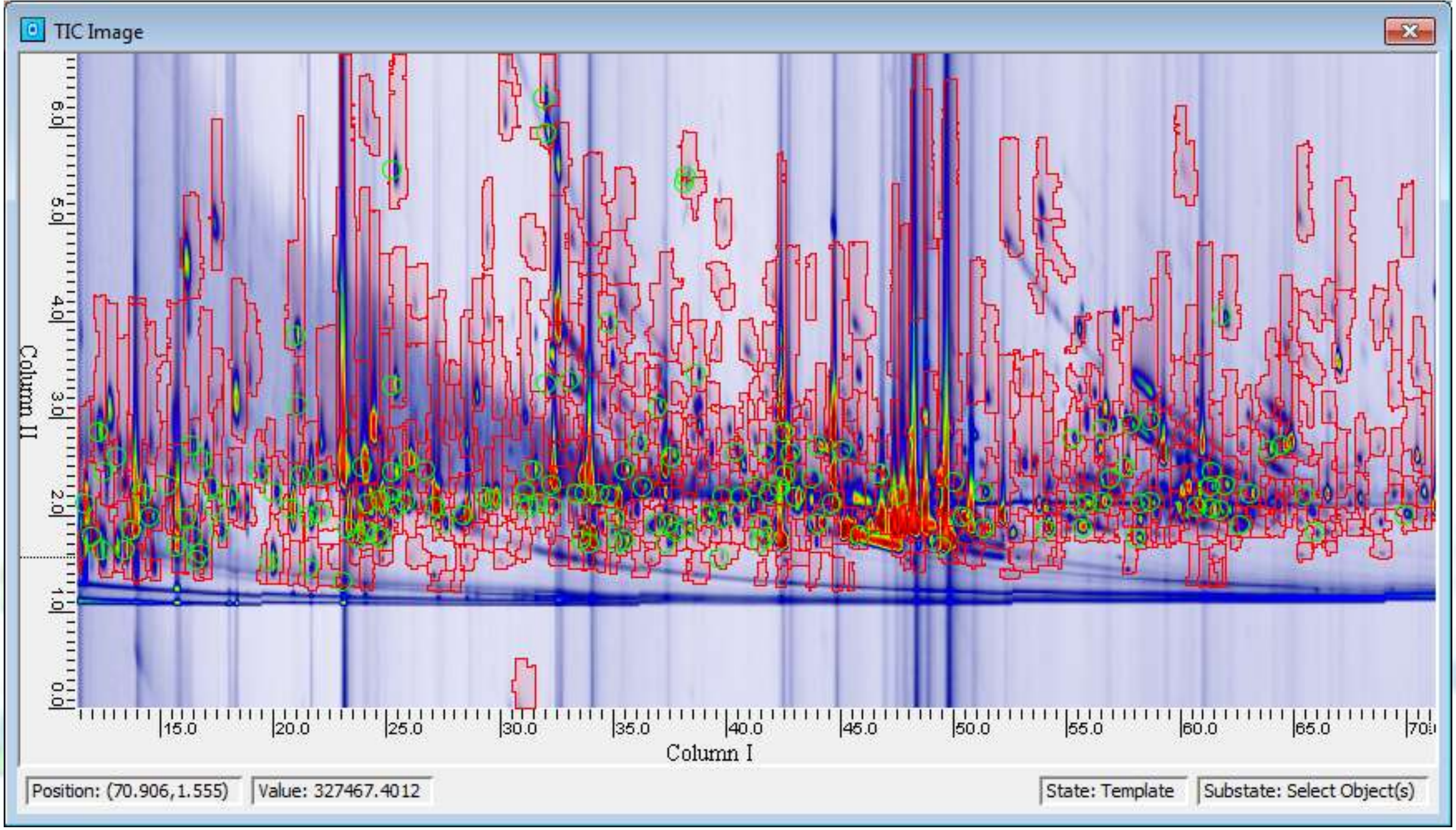
- Non-targeted analysis of 4 types of rice blast fungus *Magnaporthe oryzae* (Sofia Aronova *et al., ISCC* 2014)
  - Instrument: Agilent 7890B/ZOEX ZX2 thermal modulation system coupled with Agilent 7200 Q-TOF
  - Data analysis software: GC Image Pre-Release 2.5a0





Data Analysis: Feature Template

- The feature template with 159 reliable peaks & 572 peak-region features



Position: (70.906,1.555)

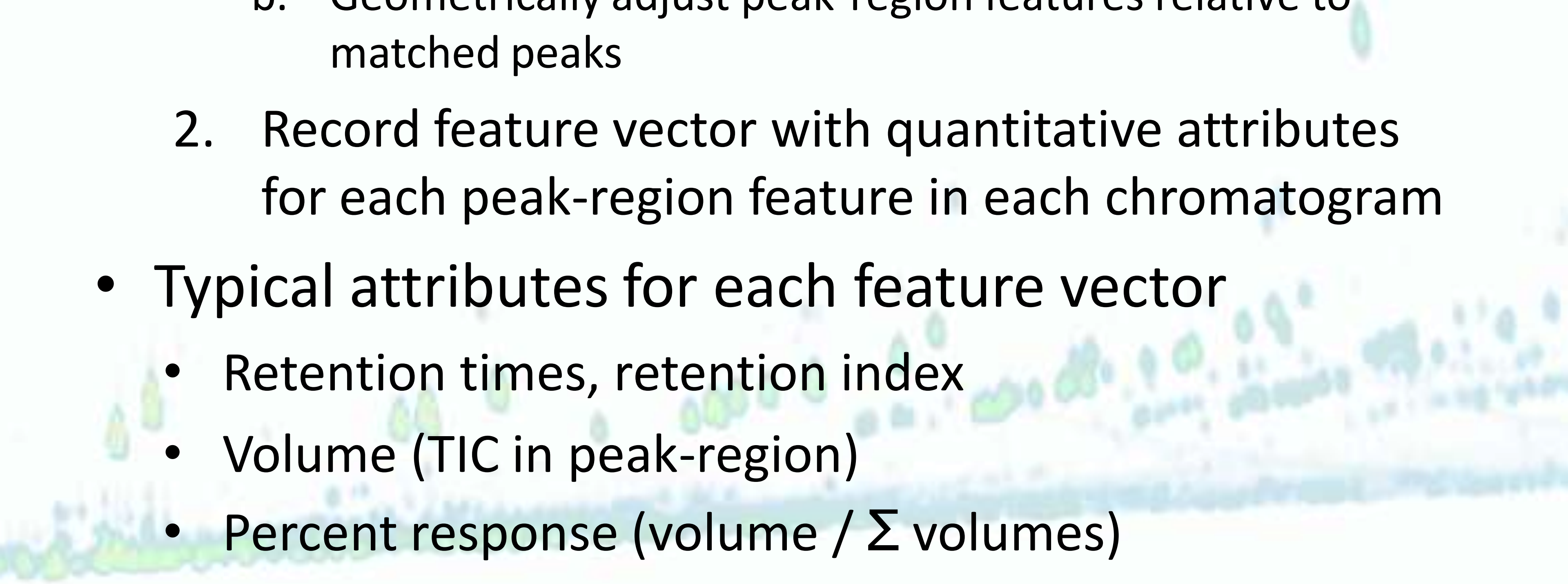
Value: 327467.4012

State: Template

Substate: Select Object(s)

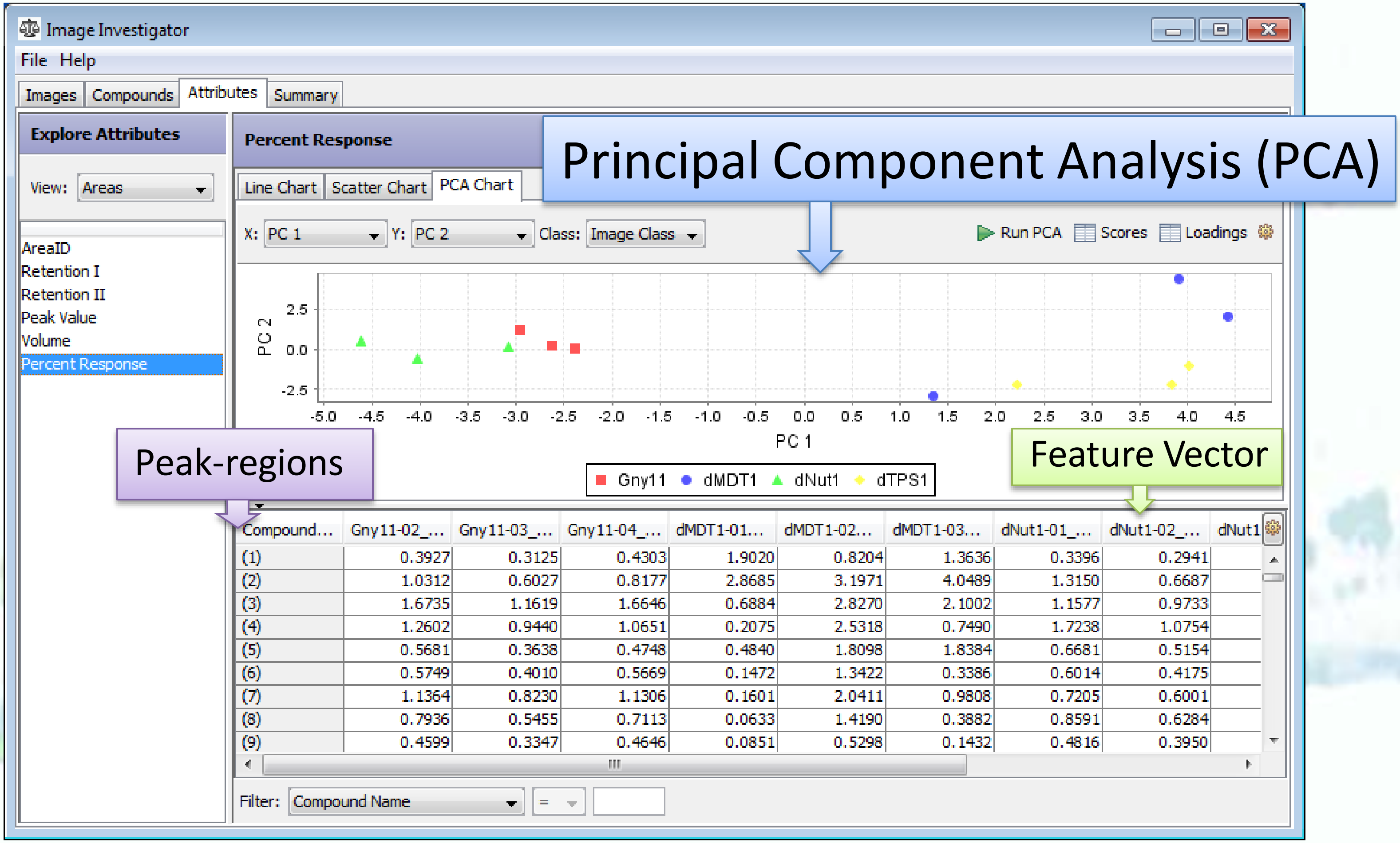
Data Analysis: Feature Measures

- Analyze all features in each chromatogram:
  1. Match feature template to each chromatogram
     a. Match pattern of reliable peaks to detected blobs
     b. Geometrically adjust peak-region features relative to matched peaks
  2. Record feature vector with quantitative attributes for each peak-region feature in each chromatogram
- Typical attributes for each feature vector
  - Retention times, retention index
  - Volume (TIC in peak-region)
  - Percent response (volume / Σ volumes)



Data Analysis: Feature Vectors

- Each chromatogram has a feature vector with attribute values of all peak-regions



View: Areas

X: PC 1   Y: PC 2   Class: Image Class

Compound...	Gny11-02_...	Gny11-03_...	Gny11-04_...	dMDT1-01...	dMDT1-02...	dMDT1-03...	dNut1-01_...	dNut1-02_...	dNut1
(1)	0.3927	0.3125	0.4303	1.9020	0.8204	1.3636	0.3396	0.2941	
(2)	1.0312	0.6027	0.8177	2.8685	3.1971	4.0489	1.3150	0.6687	
(3)	1.6735	1.1619	1.6646	0.6884	2.8270	2.1002	1.1577	0.9733	
(4)	1.2602	0.9440	1.0651	0.2075	2.5318	0.7490	1.7238	1.0754	
(5)	0.5681	0.3638	0.4748	0.4840	1.8098	1.8384	0.6681	0.5154	
(6)	0.5749	0.4010	0.5669	0.1472	1.3422	0.3386	0.6014	0.4175	
(7)	1.1364	0.8230	1.1306	0.1601	2.0411	0.9808	0.7205	0.6001	
(8)	0.7936	0.5455	0.7113	0.0633	1.4190	0.3882	0.8591	0.6284	
(9)	0.4599	0.3347	0.4646	0.0851	0.5298	0.1432	0.4816	0.3950	

Filter: Compound Name =

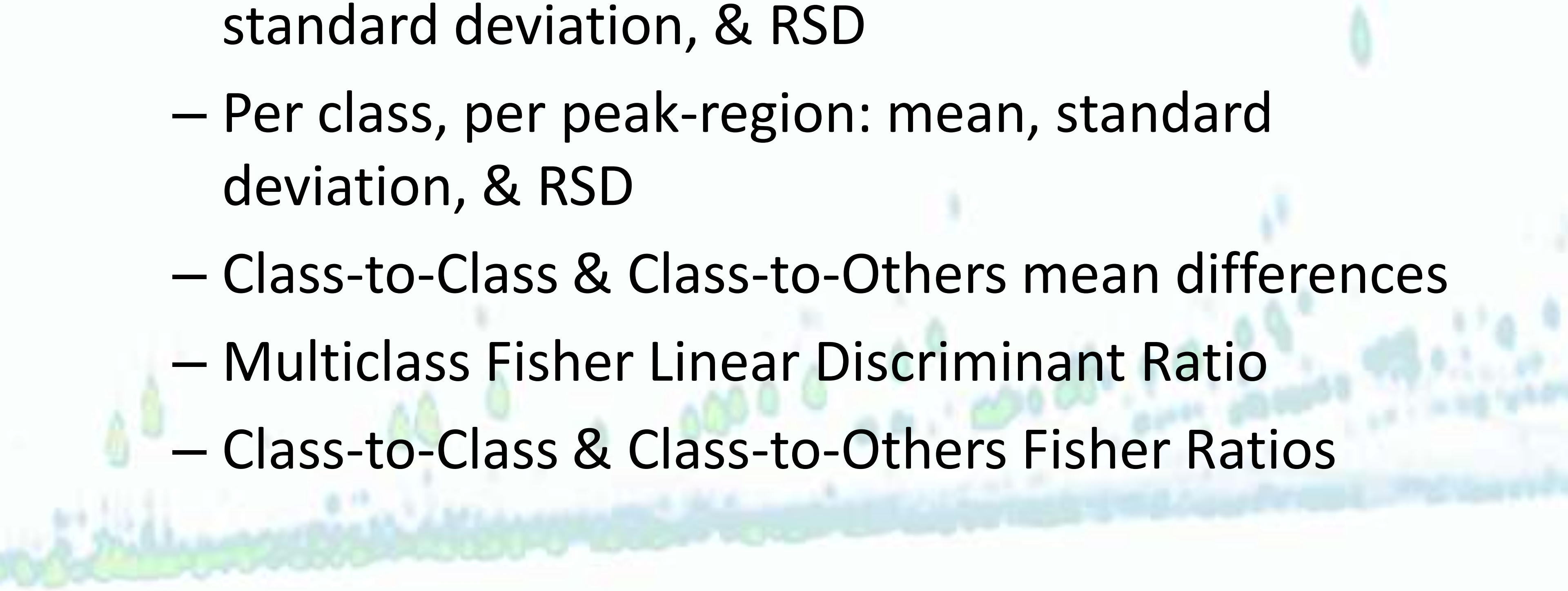
Principal Component Analysis (PCA)

Peak-regions

Feature Vector

Data Analysis: Feature Statistics

- Use attribute statistics (e.g., on previous slide) to select features of interest:
  - Per peak-region over all chromatograms: Mean, standard deviation, & RSD
  - Per class, per peak-region: mean, standard deviation, & RSD
  - Class-to-Class & Class-to-Others mean differences
  - Multiclass Fisher Linear Discriminant Ratio
  - Class-to-Class & Class-to-Others Fisher Ratios



Data Analysis: Feature Selection

- Potential biomarkers:
  - Large F value among all classes

$$F(x_1, \dots x_K) = \frac{\Sigma_i N_i (\mu_i - \mu)^2 / (K-1)}{\Sigma_{i,j} N_i (x_{i,j} - \mu_i)^2 / (N-K)}$$

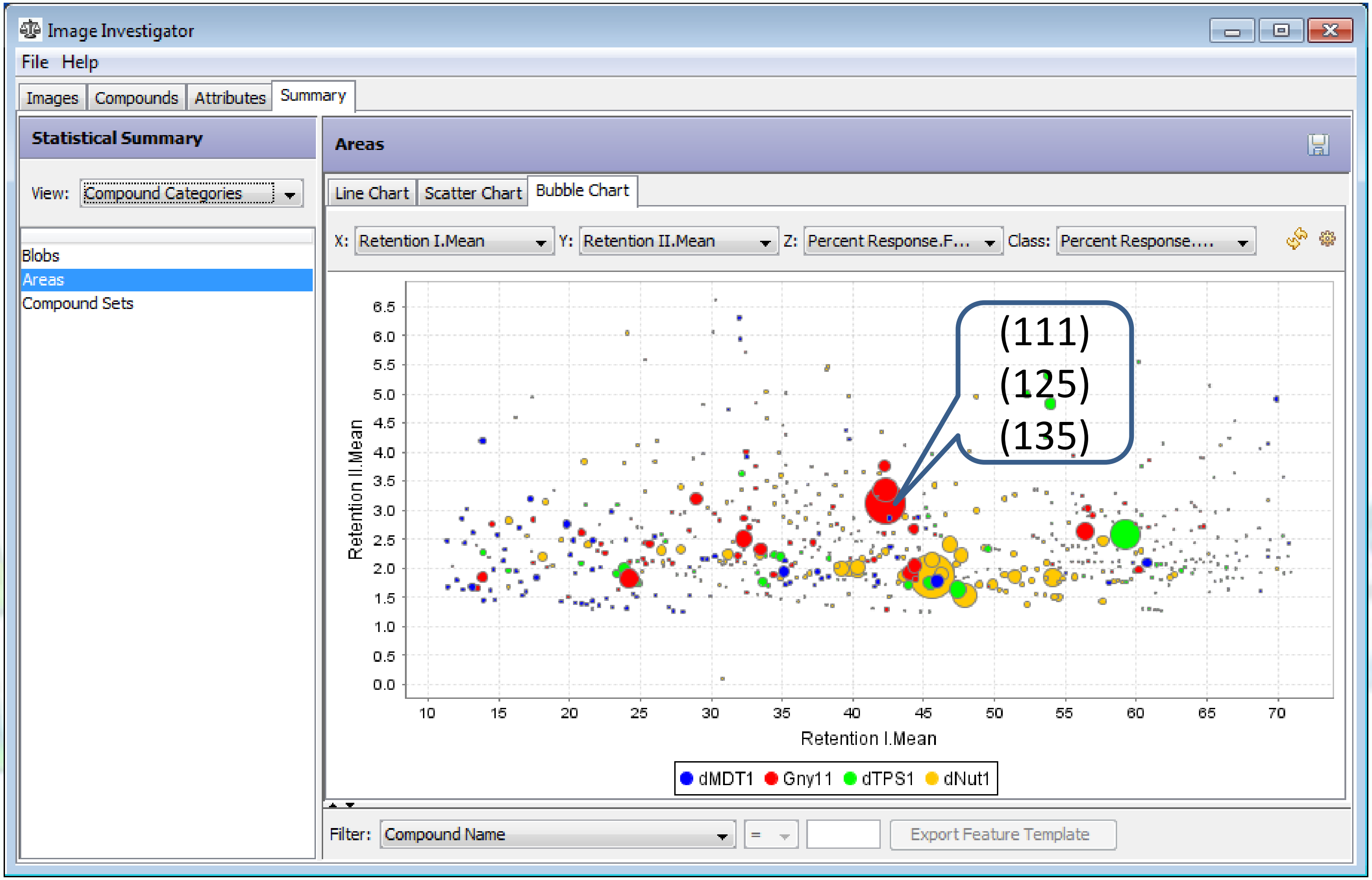
  - Larger Fisher ratio between one class and others

$$FDR(x_1, x_2) = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1^2 + \sigma_2^2)}$$

  - Both measures assess between-group variance against within-group variance.
- Analyze peak-regions that have large F values and Fisher ratios as prospective biomarker using HRMS

Data Analysis: Selected Features

- Several features with larger F value for all samples and larger Fisher ratios between one type against other types of samples



View: Compound Categories

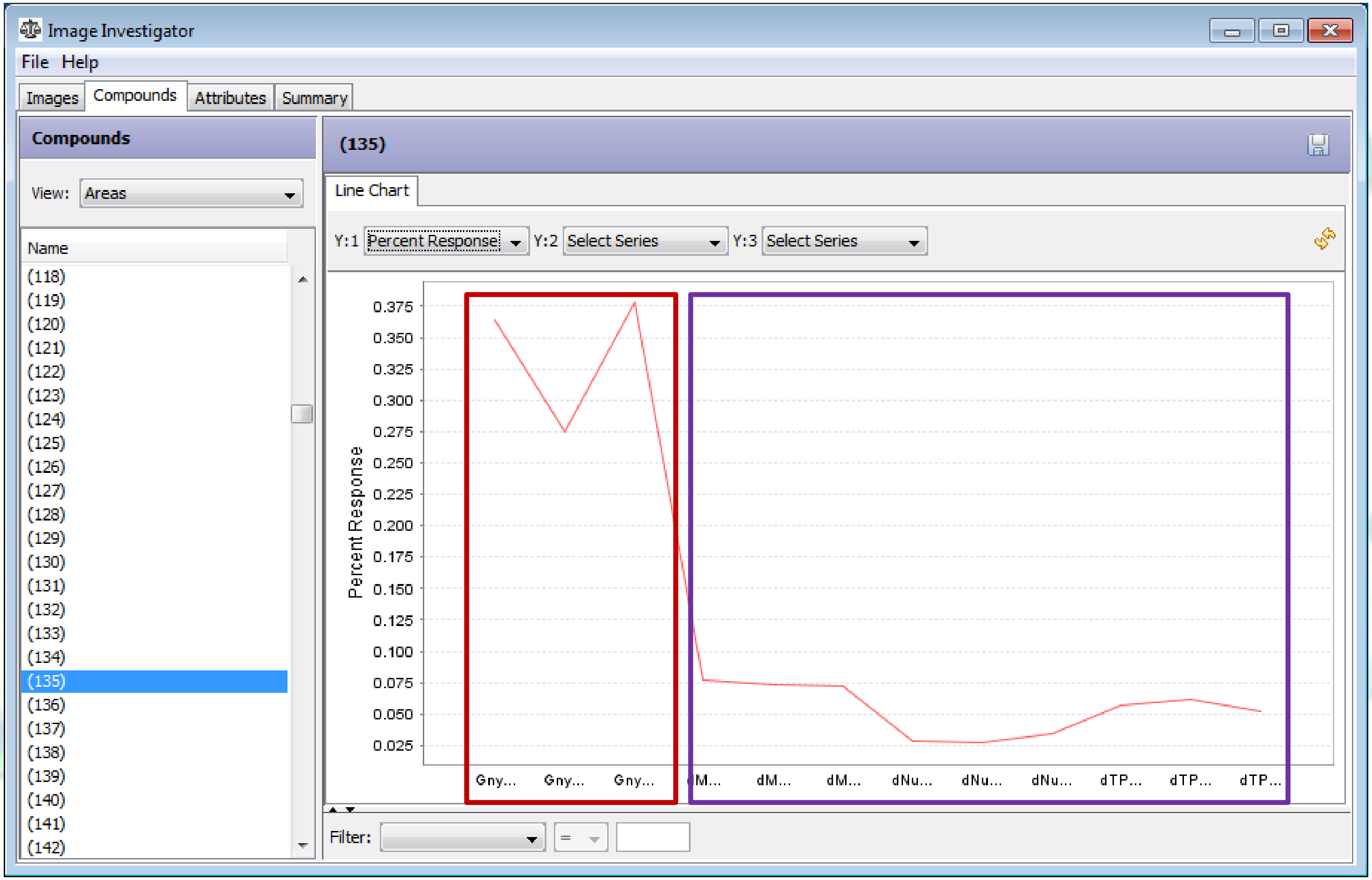
X: Retention I.Mean Y: Retention II.Mean Z: Percent Response.F... Class: Percent Response....

Filter: Compound Name =

(111)
(125)
(135)

Data Analysis: Selected Features

- Between-class differences in peak-region %response also indicates prospective biomarker

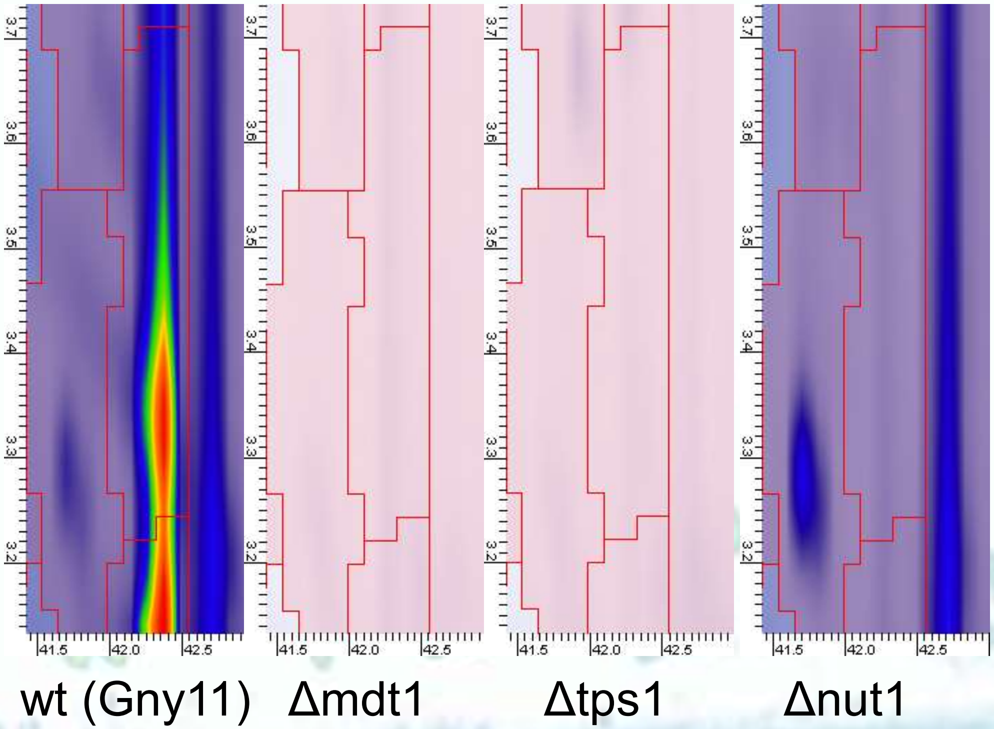


View: Areas

Y:1 Percent Response Y:2 Select Series Y:3 Select Series

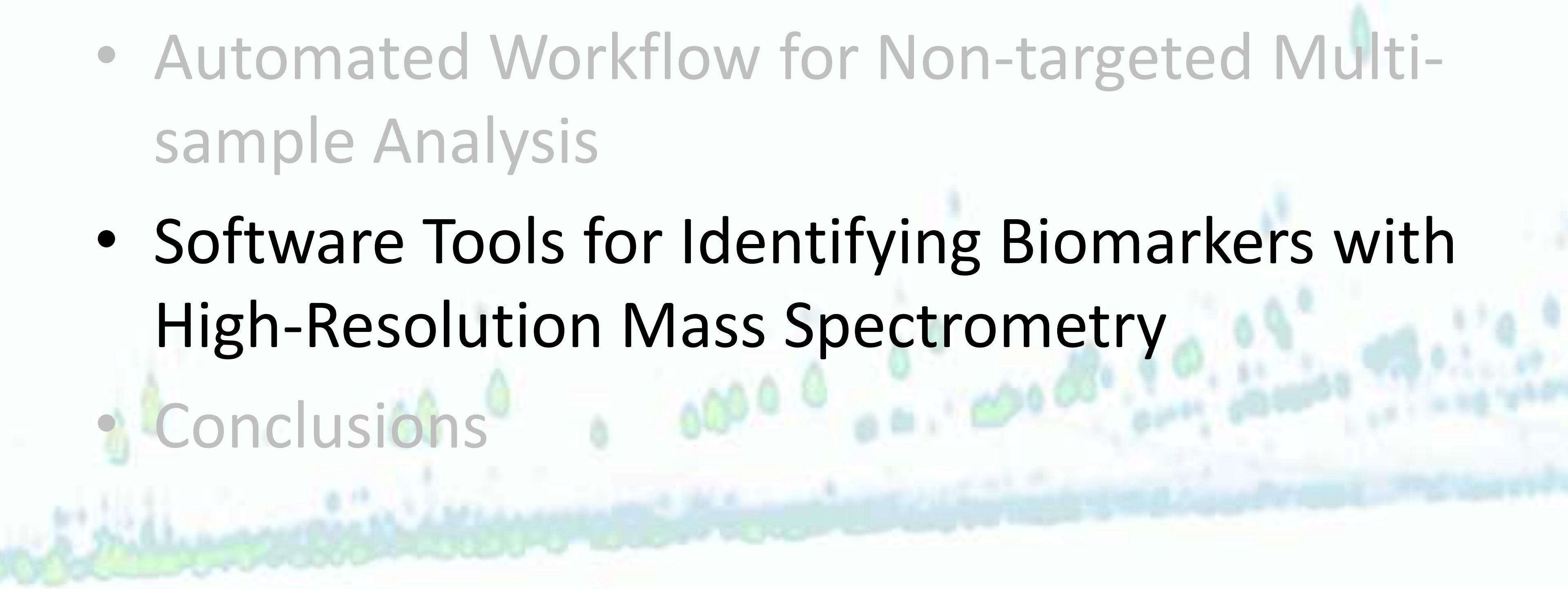
Data Analysis: Selected Features

- Chromatograms of Peak-Region (135) for the four classes:



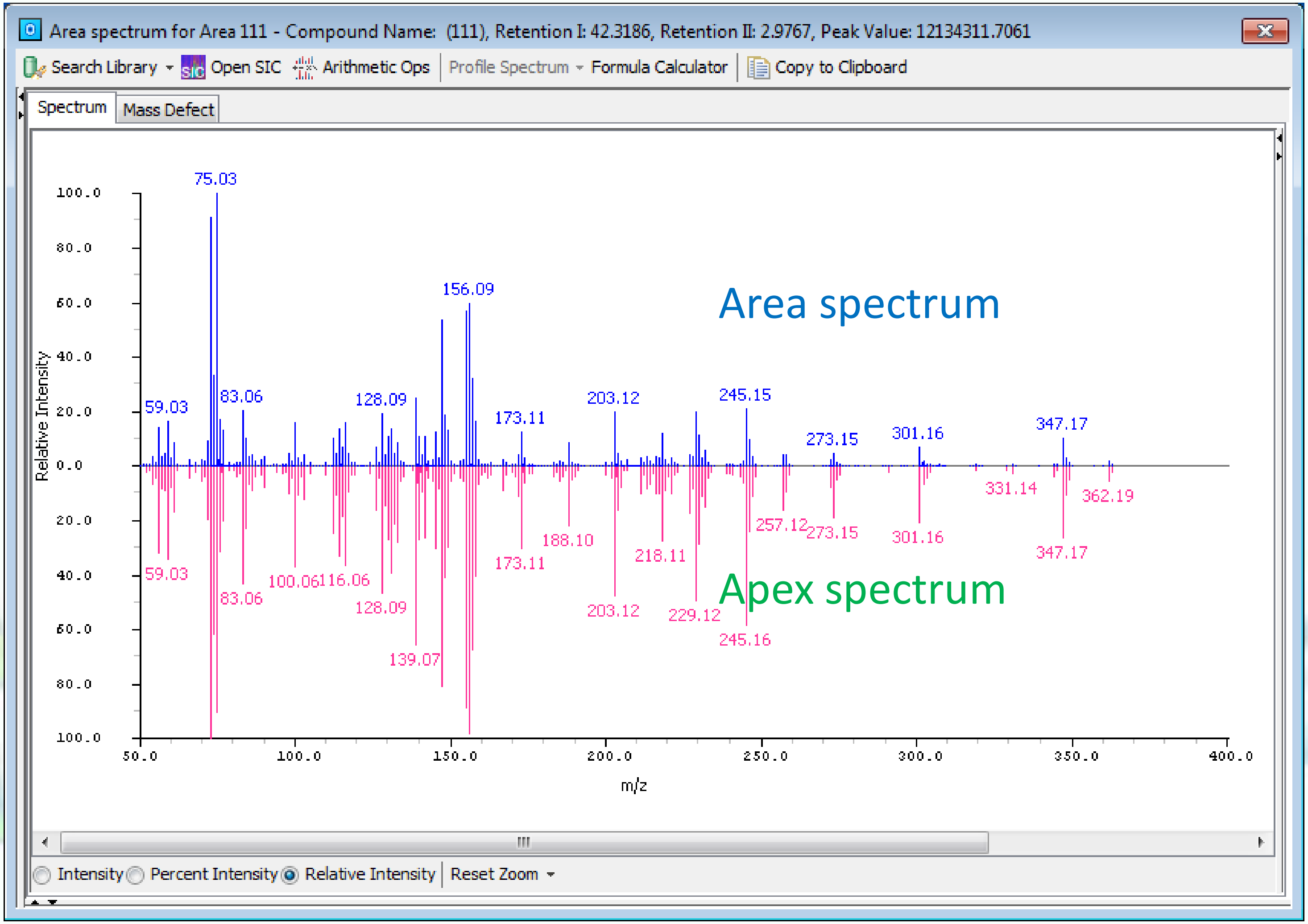
Outline

- Introduction
- Multi-Sample Analysis: Workflows and Challenges
- Automated Workflow for Non-targeted Multi-sample Analysis
- **Software Tools for Identifying Biomarkers with High-Resolution Mass Spectrometry**
- Conclusions



HRMS Analysis: Peak-Region (111)

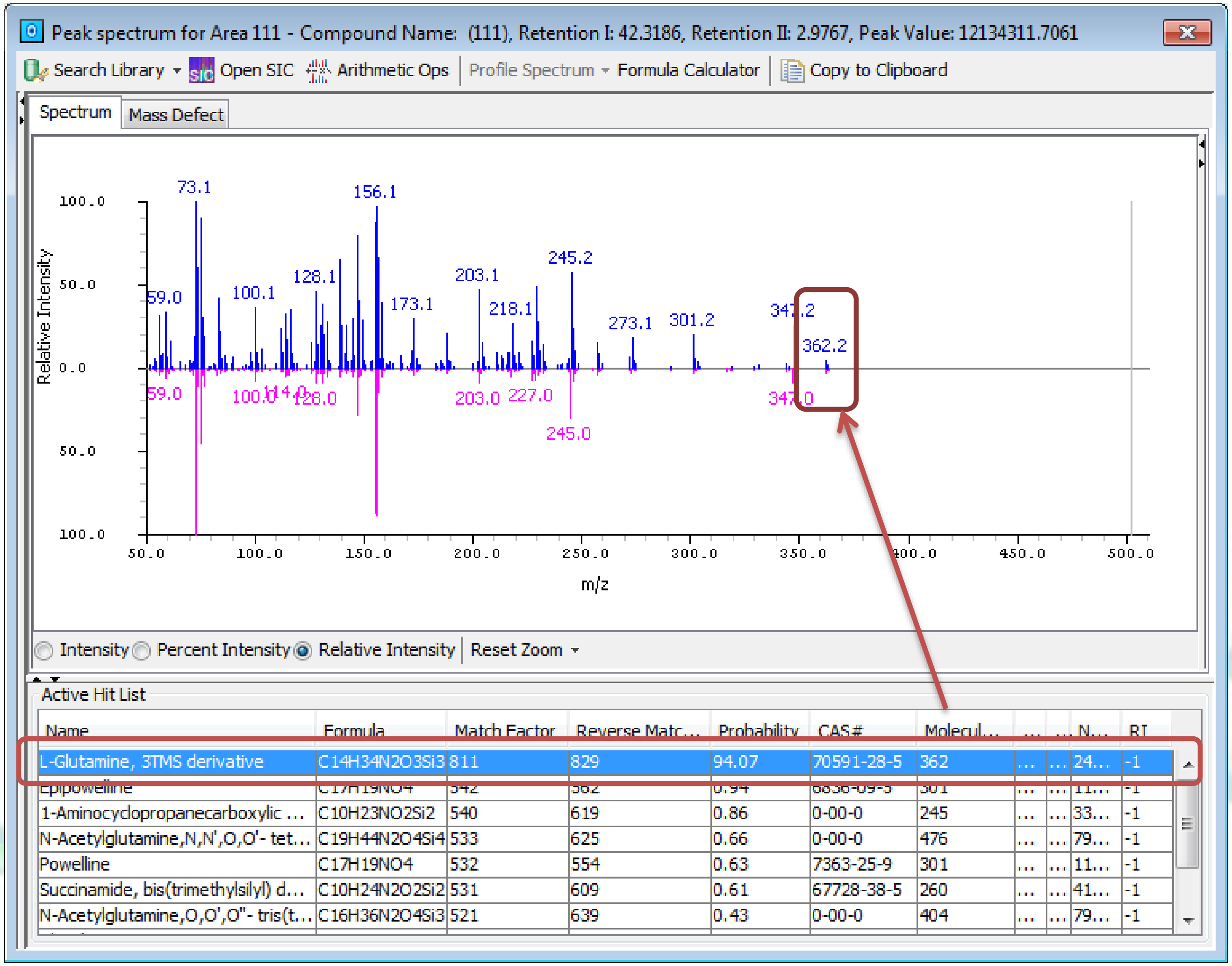
- Good spectral purity, e.g., Area & apex spectra



Area spectrum for Area 111 - Compound Name: (111), Retention I: 42.3186, Retention II: 2.9767, Peak Value: 12134311.7061

HRMS Analysis: Peak-Region (111)

- NIST14 MS match (811) & reverse match (829)



Peak spectrum for Area 111 - Compound Name: (111), Retention I: 42.3186, Retention II: 2.9767, Peak Value: 12134311.7061

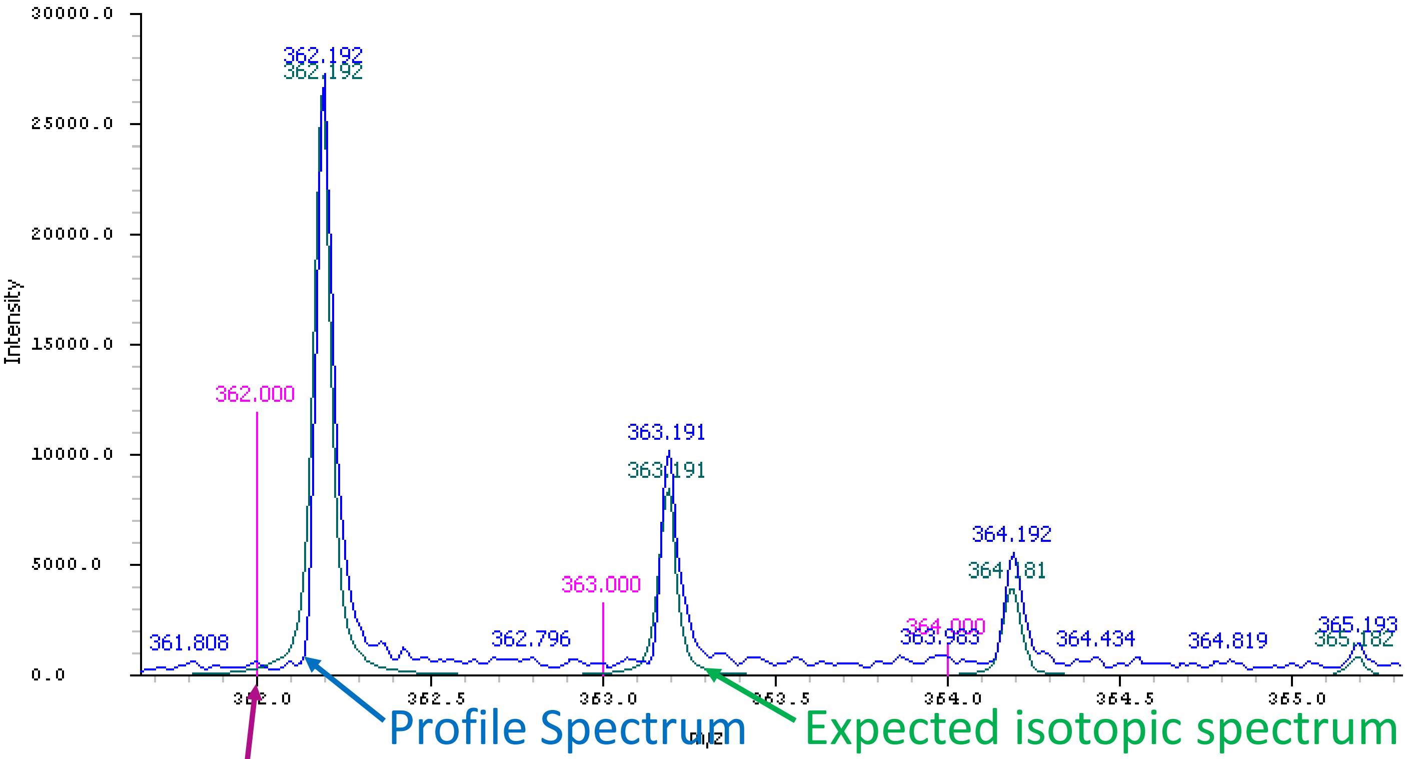
○ Intensity ○ Percent Intensity ● Relative Intensity

Active Hit List

Name	Formula	Match Factor	Reverse Matc...	Probability	CAS#	Molecul...	...	...	N...	RI
L-Glutamine, 3TMS derivative	C14H34N2O3Si3	811	829	94.07	70591-28-5	362	...	...	24...	-1
Epipowelline	C17H19NO4	542	562	0.94	6836-09-5	301	...	...	11...	-1
1-Aminocyclopropanecarboxylic ...	C10H23NO2Si2	540	619	0.86	0-00-0	245	...	...	33...	-1
N-Acetylglutamine,N,N',O,O'- tet...	C19H44N2O4Si4	533	625	0.66	0-00-0	476	...	...	79...	-1
Powelline	C17H19NO4	532	554	0.63	7363-25-9	301	...	...	11...	-1
Succinamide, bis(trimethylsilyl) d...	C10H24N2O2Si2	531	609	0.61	67728-38-5	260	...	...	41...	-1
N-Acetylglutamine,O,O',O''- tris(t...	C16H36N2O4Si3	521	639	0.43	0-00-0	404	...	...	79...	-1

HRMS Analysis: Peak-Region (111)

- Confirm the identification with exact mass
  - L-Glutamine, 3TMS derivative, $C_{14}H_{34}N_2O_3Si_3$



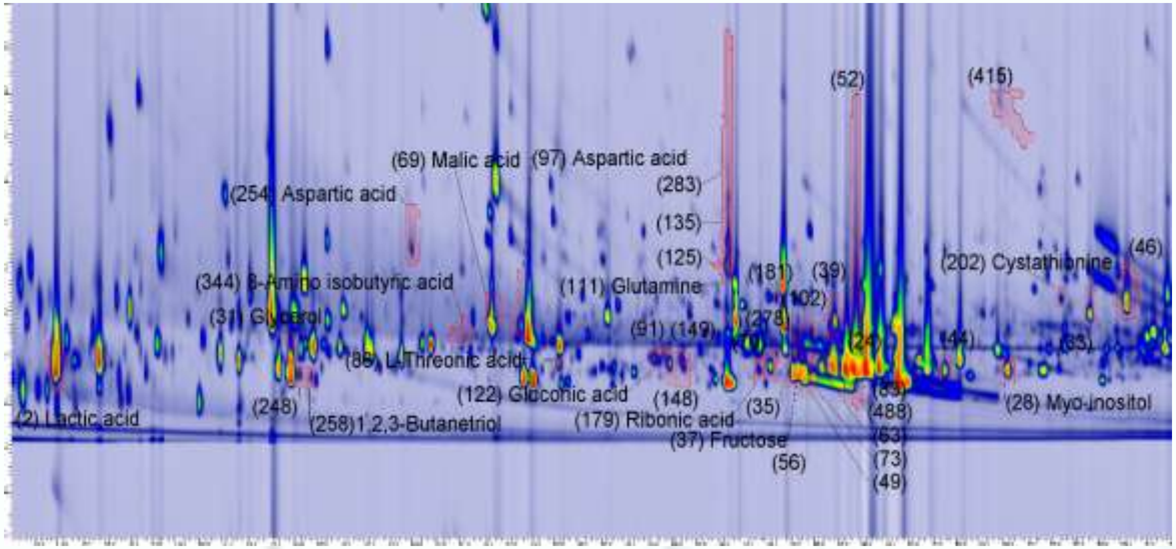
Profile Spectrum

Expected isotopic spectrum

Library Spectrum

HRMS Analysis: Peak-Region

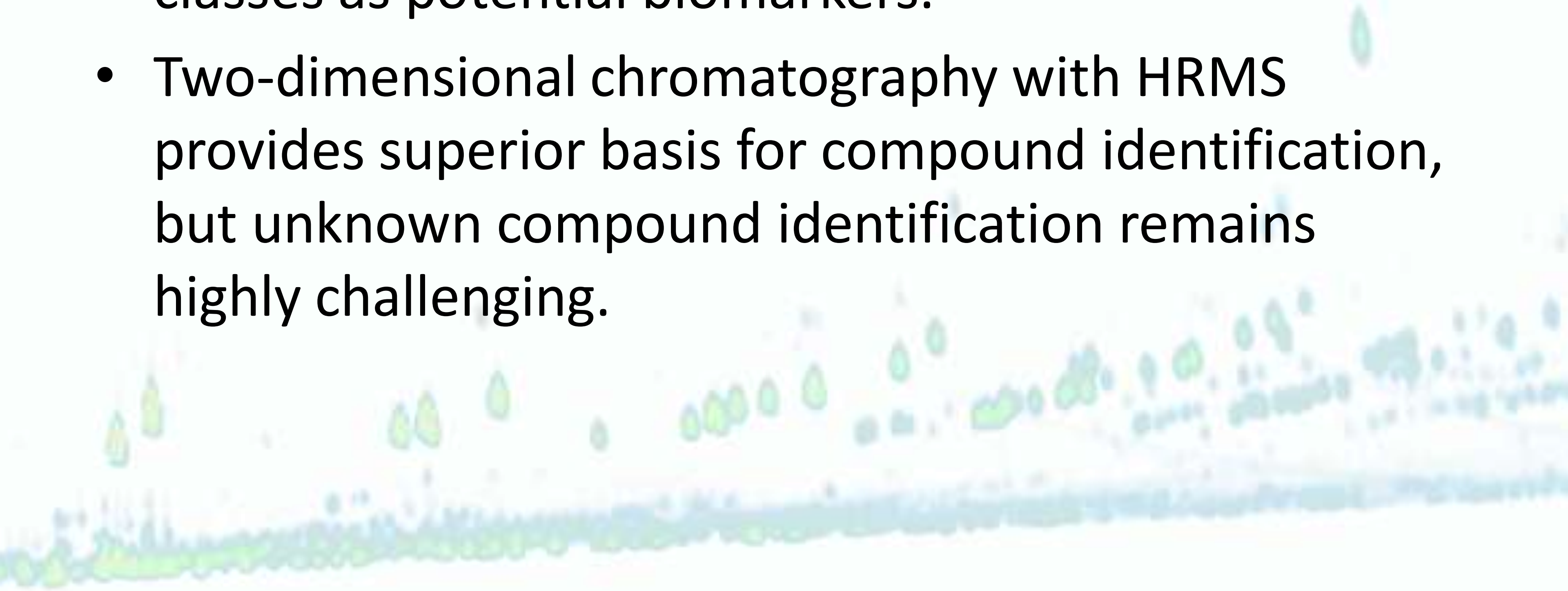
- Prospective peak-region biomarkers (with red polygons), labeled with ID and compound name (if identified).



- Not all compound identities can be determined
  - Not in NIST library (or other libraries, including Wiley 8E, Fiehn, Golm)
  - Large molecule with more complex elemental composition
  - Smaller concentration, so smaller ion peaks

Conclusions

- Advanced software tools with comprehensive peak-region feature analysis can effectively detect compounds that are highly differentiated between classes as potential biomarkers.
- Two-dimensional chromatography with HRMS provides superior basis for compound identification, but unknown compound identification remains highly challenging.



Acknowledgements

- Thanks to Agilent Technologies and Zoex Corporation for providing the example datasets.
- Thanks to Department of Plant Pathology - University of Nebraska at Lincoln for preparing samples for the rice fungus dataset.
- Part of the work was supported by NSF SBIR Phase II Award (1127264).
- Travel supported by Agilent Technologies.