# New peak detection (Pyco) and isotope grouping (Prism) algorithms for an improved compound detection workflow

Yovany Cordero Hernandez, Pedro Navarro, Konstantin Ayzikov, Martin Strohalm, James Dillon, Alexander Tiegel, Ute Comberg, Ralf Tautenhahn, Alexander Makarov, Christoph Henrich, Thermo Fisher Scientific, 355 River Oaks Parkway, San Jose, CA, 95134

## Abstract

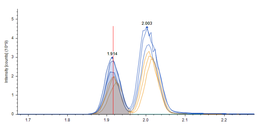
A reliable peak detection algorithm is key for any compound detection workflow. Here we introduce the new detection workflow in Thermo Scientific™ Compound Discoverer™ 3.3 software, based on a novel parameter-free peak detection algorithm, Pyco, and a redesigned isotope grouping algorithm, Prism. We show that this new approach works more consistently on various chromatography methods, as peak integration does not rely on restrictive peak model assumptions, such as Gaussian shape or symmetry. Using peak quality metrics, we corroborate that both identification and quantification results have been significantly improved. We were able to reduce false positive compounds by avoiding over-splitting of peaks. The new detection workflow is not only more precise, but also more efficient, reducing the overall computational time for processing.

## Introduction

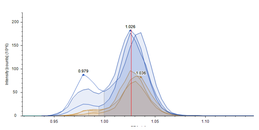
### Challenges in peak detection

- Overlapping peaks/deconvolution
- Baseline detection and removal
- Accurate peak integration
- Peak-model-based detection and -integration vs. model-free

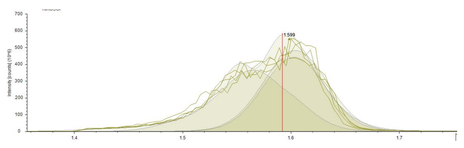
**Baseline separated peaks**



**Partially overlapping peaks**

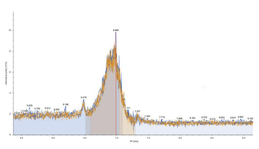


**Problems with model-based peak integration**

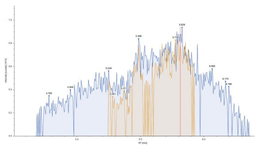


- Sensitivity vs. Specificity
- “Real” peak or just noise ?
- How to measure peak quality?

**High baseline, noisy peak**



**Noisy peak, low signal**



## Compound Detection workflow

Compound Detection in Compound Discoverer software is performed in four steps.

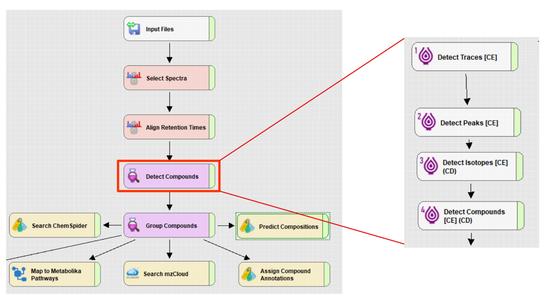
a. Trace Generation and -filtering

b. Detection of chromatographic peaks

c. Detection of isotopic patterns

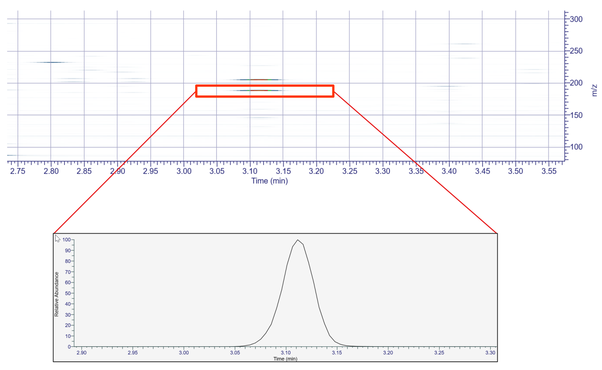
d. Grouping of adducts and fragments

**Compound Detection workflow in Compound Discoverer software**




## Trace Detection

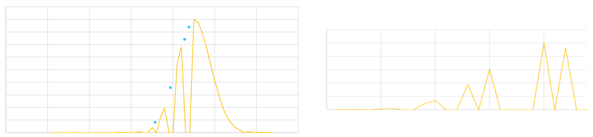
**Detection of mass traces in the *m/z* – RT domain.**



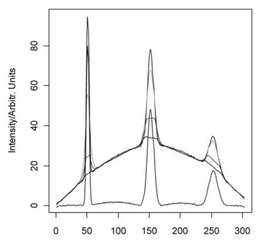

**Trace quality filtering**

Gaps in traces are interpolated when gaps are found between two non-zero values and the gap is no longer than two missing points.

Trace is rejected when gap ratio is too high or entropy too low.



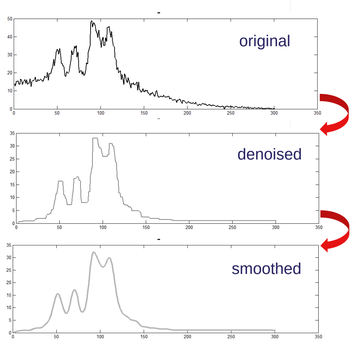
**Baseline removal using AirPLS**




Baseline removal using Adaptive iteratively reweighted Penalized Least Squares[1]

**Additional step to provide better baseline correction compared to previous model**

**Trace Denoising and Smoothing**

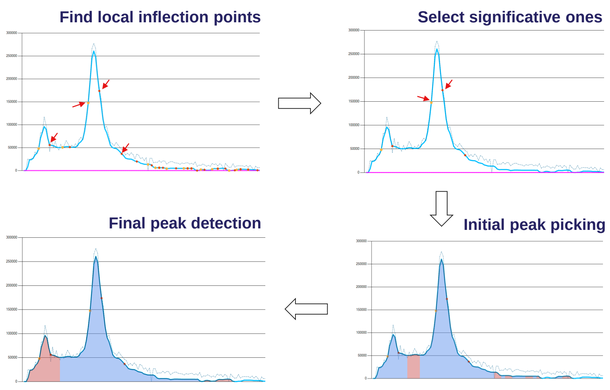



Total Variation Denoising (TVD). Denoising without signal distortion. Parameters are estimated from an FT image of a trace.

Smoothing via convolution with Gaussian window

## Peak Detection

**Peak Detection algorithm in images**

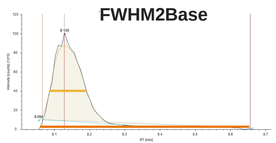



**Peak Quality Factors[2]**

**Jaggedness**

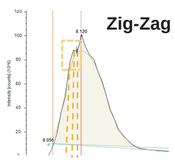
Captures shape quality by calculating the number of changes in direction over length of intensity vector.

**FWHM2Base**



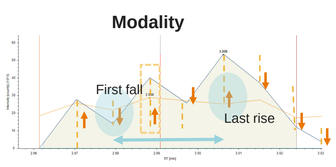
Assesses separation of peaks by measuring peak width at half-maximum vs. peak width at base.

**Zig-Zag Index**



Captures shape quality by measuring the normalized variance between a point and its neighbors on either side.

**Modality**

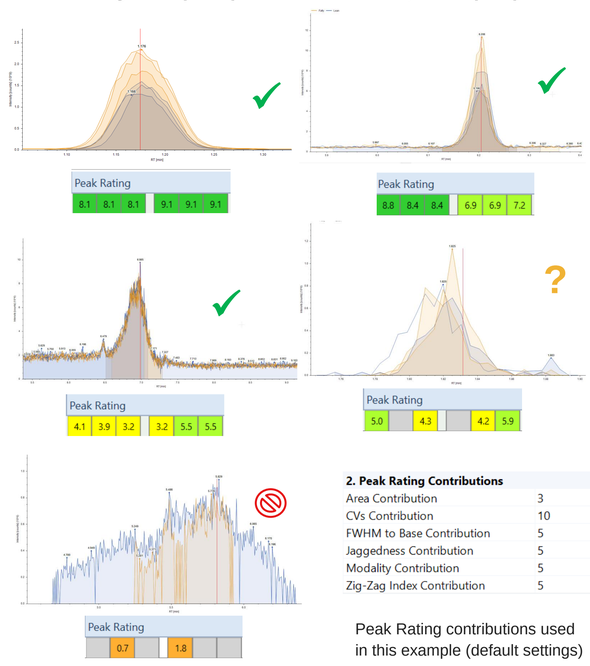



Measures the first unexpected change in direction of intensity to detect splitting and integration of multiple peaks.

**Final Peak Rating**

Peak Rating values in Compound Discoverer software are calculated based on the four different Peak Quality Factors, but also take into account CV and relative peak area.

**Peak Rating examples (ZDF demo dataset, 6 samples)**



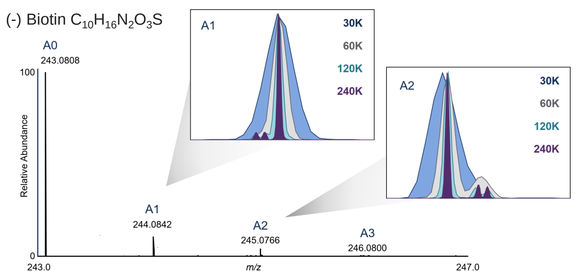

| 2. Peak Rating Contributions | |
|---|---|
| Area Contribution | 3 |
| CVs Contribution | 10 |
| FWHM to Base Contribution | 5 |
| Jaggedness Contribution | 5 |
| Modality Contribution | 5 |
| Zig-Zag Index Contribution | 5 |

Peak Rating contributions used in this example (default settings)

## Isotope Detection & Grouping

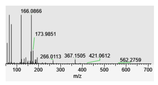
Prism (**P**attern **R**ecognition of **I**sotopes in **S**mall **M**olecules)

- Fast isotope pattern detection tuned for small molecule data.
- Looks for isotopes by mass shifts relative to an $A_0$ mass.
- Supports detection of the following isotopic elements: carbon, hydrogen, oxygen, nitrogen, sulfur, bromine, chlorine

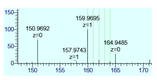



**Isotope detection using PRISM**

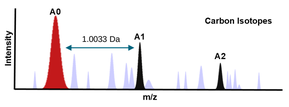

**Step 1.**
Select a chromatographic peak and set that as the $A_0$ mass.



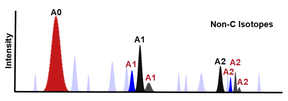

**Step 2.**
Load mass spectrum at chromatographic peak apex.



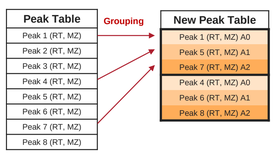

**Step 3.**
Select window around the $A_0$ mass.




**Step 4.**
Look for carbon isotopes that match the mass offset and expected intensity of current $A_0$ mass.

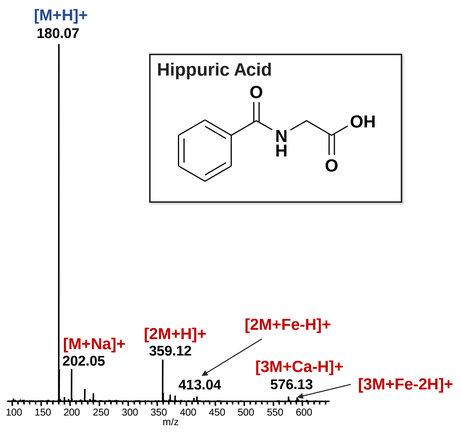



**Step 5.**
Look for non-carbon isotopes that match the mass offset and expected intensity of the current $A_0$ mass.

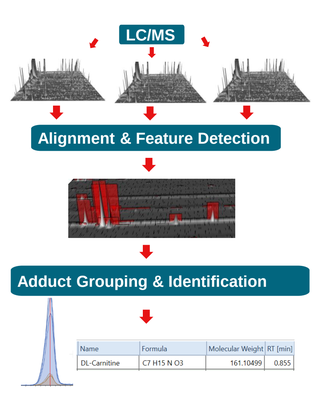



**Step 6.**
Compare list of detected isotopes to list of detected chromatographic peaks. Select matching chromatographic peaks and group with the $A_0$ peak.

## Adduct Grouping




## Data Reduction Process




**100,000,000**
**Raw Data Points**
(thousands of spectra in each file)

**100,000 Features**
e.g. m/z 162.1120 @ 0.88 min
m/z 163.1155 @ 0.89 min

**1,000 Compounds**
e.g. Carnitine, MW 161.10 @ 0.85 min

## Results

### Improved sensitivity and runtime

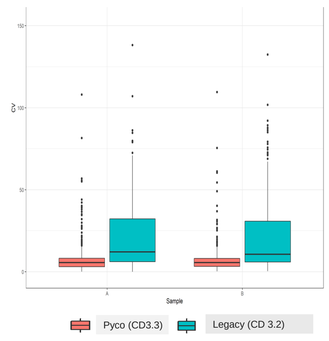
- New peak detection is faster
- New peak detection is more sensitive: detects more compounds and more ions per compound
- Result tables can be much larger → filter based on Peak Rating

| Compound Discoverer Software version | Intensity threshold | Peak Rating Threshold | Number of Compounds after Group Compounds, unfiltered | Number of Compounds after Group Compounds, filtered by Peak Rating in Compound Discoverer Software | Runtime \| Detect Compounds |
|---|---|---|---|---|---|
| 3.2 | 10,000 | *n/a* | 59,958 | Not possible to filter by Peak Rating | 17h 3min |
| **3.3** | 10,000 | **Peak Rating ≥ 4** in at least 3 samples | **104,332** | **19,931** | **4h 27min** |

121 samples (soy sauce). Processed on a HP Z840, 2 x (Intel Xeon CPU E5-2667 @3.2 GHz, 8 cores), 64GB RAM, purchased in 2017.

### Quantification Reproducibility

- Evaluation using data published by Li et al.[3], 1100 spiked compounds at known concentrations
- Publication contains comparison of free software tools and a much older version of Compound Discoverer software (v2.1, current version is 3.3)
- CVs calculated with compounds detected in at least 2 raw files




- Pyco shows a significantly reduced variability compared to the legacy algorithm in Compound Discoverer software
- Pyco’s variability is on par with the best results from other software compared in the paper

## Conclusions

- Increased sensitivity reduces chances of missing compounds
- Peak Rating is a flexible mechanism to filter out low quality compounds and adds an additional metric for evaluation
- New peak detection contributes to improved quantification
- New detection workflow improves workflow run time

## References

1. *Z.-M. Zhang, S. Chen, and Y.-Z. Liang*, Baseline correction using adaptive iteratively reweighted penalized least squares. Analyst 135 (5), 1138-1146 (2010).

2. Adapted from *Pandey et al.*, Metabolomics (2020), MetaClean: a machine learning-based classifier for reduced false positive peak detection in untargeted LC–MS metabolomics data

3. *Li et al.*, Anal Chim Acta. (2018), Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection.

## Trademarks/licensing



Learn more at www.thermofisher.com/compounddiscoverer