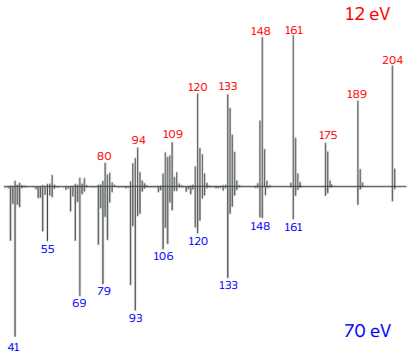
White Paper 041
May 2021

Technical note: Improving discovery workflows using Tandem Ionisation® data



This technical note demonstrates how Tandem Ionisation (TI) data can be used to improve discovery workflows by reducing the frequency of false positives and describes how this is achieved using an untargeted, tile-based approach in ChromCompare+ software to find the significant class-based differences between GC and GC×GC chromatograms.

Contents

Why do we need discovery workflows?

In discovery (or untargeted) workflows, we typically don't know what compounds are important and need to extract as much information from the data as possible in order to draw meaningful conclusions. In recent years, there has been increased demand for improved non-target workflows across a diverse range of applications – from discovery-based analysis for novel biomarkers of disease to the detection of food fraud, where many of the possible adulterants remain unknown.

SepSolve Analytical Ltd
T: +44 (0)1733 669222 (UK) +1 888-379-3835 (USA)
E: hello@sepsolve.com
www.sepsolve.com

What is tile-based analysis?

The novel chemometrics platform ChromCompare+ was developed to tackle the challenges associated with untargeted analysis by automatically uncovering the significant differences between sample classes. In chemometrics, sample classes are simply the categories used to classify samples – for example, in food authenticity studies, we may have 'genuine' versus 'fraudulent' sample classes.

The raw data is aligned and imported directly into ChromCompare+ using the innovative workflow described in Figure 1. This approach divides the chromatogram into small tiles, allowing every m/z channel to be compared for every section of each chromatogram in the dataset. These sections overlap each other to reduce the risk of missing important details.

All of the raw data is extracted, meaning that hundreds of thousands of individual 'features' are found. Each feature represents the summed intensity of a specific m/z in a specific tile and the feature name relates to these details. It is important to note that while Figure 1 shows an example for GC×GC, the same workflow can also be applied to 1D GC data.

This approach is automated, reducing the need for laborious pre-processing steps (such as integration and identification), thereby accelerating discovery workflows.

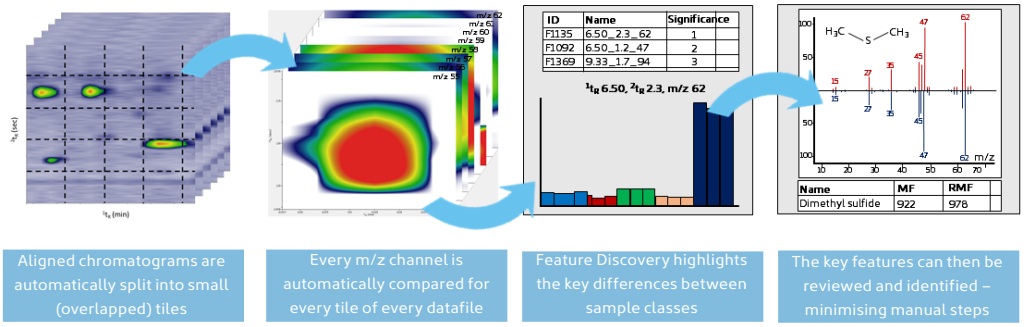


Figure 1

Overview of the untargeted, tile-based workflow in ChromCompare+. Each feature is named based on the details of the tile it was discovered in, i.e., a feature named '52.350_2.300_62' is for m/z 62 at $^{1}t_{R}$ = 52.350 min and $^{2}t_{R}$ = 2.3 seconds.

How do we know what features are significant?

Obviously, if over 100,000 features are extracted per datafile, we must then find a way to reduce this list to those that best differentiate the sample classes. In ChromCompare+, Feature Discovery is used for this purpose.

Our Feature Discovery algorithm is proprietary, but it uses a multivariate method to consider covariance between features; in other words, it will try to select features that exhibit different class behaviour. In contrast, in univariate methods, there is no consideration given to the covariance between features, so the second feature selected is often not useful because it (often) has the same class behaviour as the first feature selected. Univariate algorithms are known for their efficiency, making them popular choices in processing large datasets where

many multivariate algorithms fail (due to lack of speed). Even though the proprietary Feature Discovery in ChromCompare+ is multivariate, it still works efficiently for hundreds of thousands of features, making it well-suited to GC and GC×GC datasets.

Even with this sophisticated algorithm, however, some features are selected which are not true discriminators of the classes. Such features are called 'false positives'.

Why do we get false positives?

False positives are inevitable – they will exist in every dataset. For example, if we look at 100,000 'features' using randomly generated numbers and a confidence limit of 5%, we would expect to see approximately 5000 false positives distributed over the tiles more or less equally. The Fisher ratio (F-ratio), a univariate statistic, is frequently used to measure the discriminating power of a feature. In short, the F-ratio is the ratio of 'between-class' variance to 'within-class' variance, as seen in the equation below for a two-class case. As this ratio increases, it becomes increasingly unlikely that the difference between the class means can be explained by random chance.

$$F - ratio = n_1 \frac{(\bar{x}_1 - \bar{x}_2)^2}{S_1^2 + S_2^2}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means for class 1 and class 2, while S_1^2 and S_2^2 are the variances and n_1 is the number of samples in each class.

For example, if we have two classes and 10 samples per class and we desire a significance level of 1%, we find a critical value (f_{crit}) of 8.29. If our calculated F-ratio is equal to or greater than this, then it can be considered significant (only a 1% chance that the samples are drawn from the same population). But what if another feature has an F-ratio of 6.1? While this feature is not significant at the 1% confidence level, it is significant at the 5% level (f_{crit} = 4.41). How do we know which features are true differences and which are false positives?

In the set of 100,000 random features, if we look at three replicates (Figure 2, left) the top four features (as ranked by F-ratio) appear to be promising differentiators between the two classes (f_{crit} = 74.14 at a 0.1% significance level). However, when we expand our analysis to the full 10 replicates per class (Figure 2, right), there is no believable class differentiation and none of the F-ratios is now above the critical value of 8.29. True positives are highly unlikely to behave in this manner as the number of samples increases. Hence, one strategy to reduce the incidence of false positives is to employ larger sample sizes with lower significance levels (higher f_{crit} values).

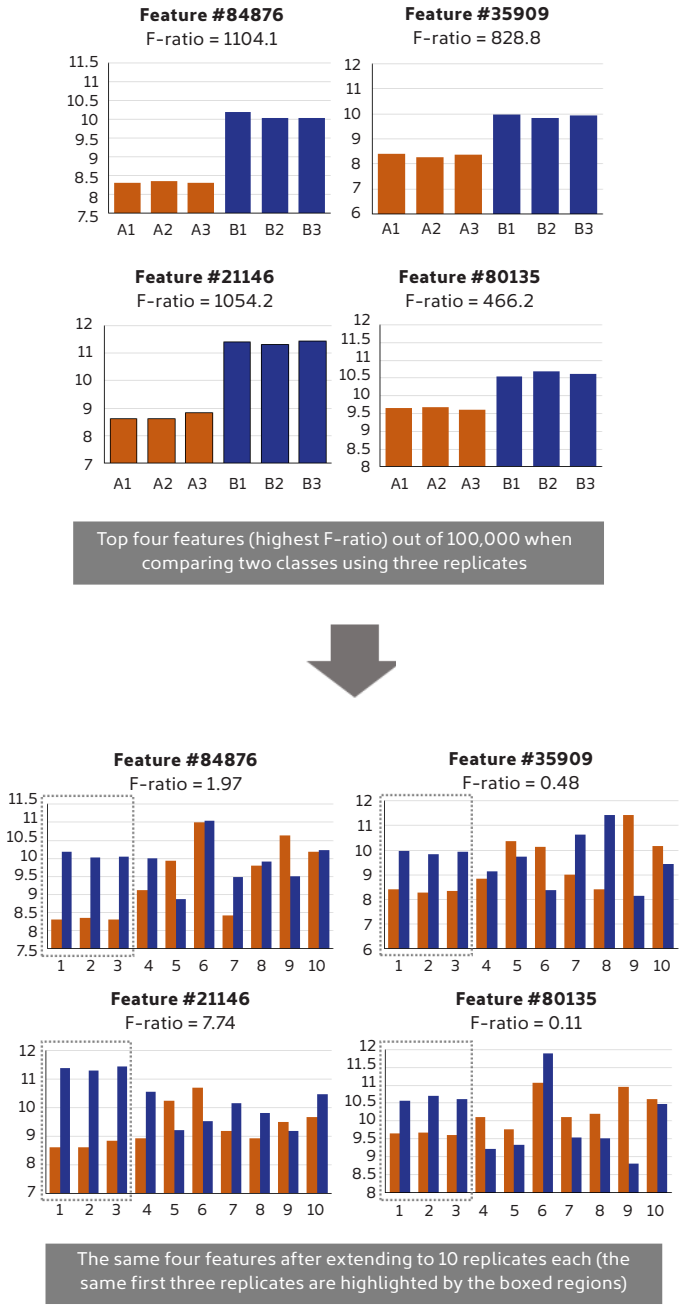


Figure 2

Comparison of F-ratios when using three and 10 replicates across a set of 100,000 randomly generated features.

Therefore, we recommend running more samples per class as part of good experimental design, but sample availability and time restrictions may mean that this is not always possible. To combat this, a minimum intensity threshold can be set to filter out weak features that are unlikely to be true positives. Additionally, a frequency threshold can be used to ensure the feature is found across a number of samples and doesn't represent a contaminant in one or two samples. Now, we introduce a breakthrough in discovery workflows, using the power of our award-winning Tandem Ionisation to reduce the rate of false positives in untargeted comparisons.

How can Tandem Ionisation help?

Tandem Ionisation using the BenchTOF2™ mass spectrometers simultaneously acquires both hard and soft EI ionisation for complementary chemical information[1] with no added analysis time (Figure 3).

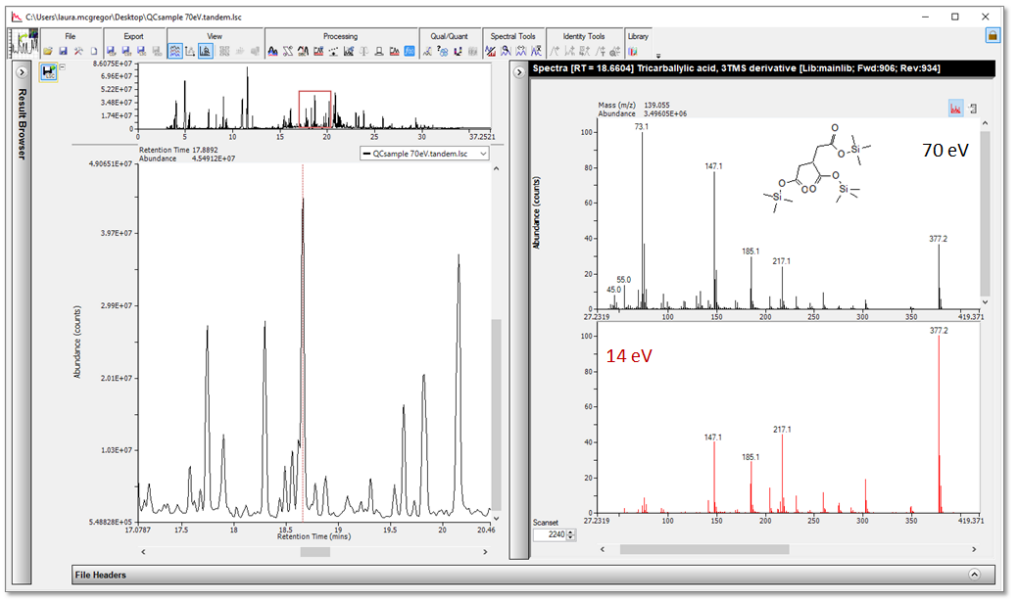


Figure 3

A tandem data file acquired using BenchTOF2-TI™, with both 70 eV and 14 eV EI data blocks represented within a single file.

As can be seen in Figure 3, the spectra corresponding to the different ionisation conditions are similar but clearly distinct. Less obvious is that the noise profiles for the corresponding mass channels are also different. So, the data from the two different ionisation modes provide different F-ratios for the same mass/tile combination. By filtering out features that are only significant in one ionisation mode but not the other, we can dramatically reduce the incidence of false positives. For example, with 100,000 features in each ionisation mode and at a 5% significance level, we expect 5000 false positives spread across, say, 250 tiles or 20 false positives per tile (random data). However, when we require positives to be confirmed in the soft EI dataset, we reduce the number of false positives by a factor of 20 to 250, or one per tile on average. The example shown in Figure 4 illustrates this point, using a randomly generated dataset of 100,000 features once again. When using a single dataset (e.g., 70 eV), 18–20 hits per tile are commonplace, but when a confirmatory tandem dataset (e.g., 70/16 eV) is used, the number of 'confirmed hits' per tile is dramatically reduced with no instances of five or more 'confirmed' hits per tile.

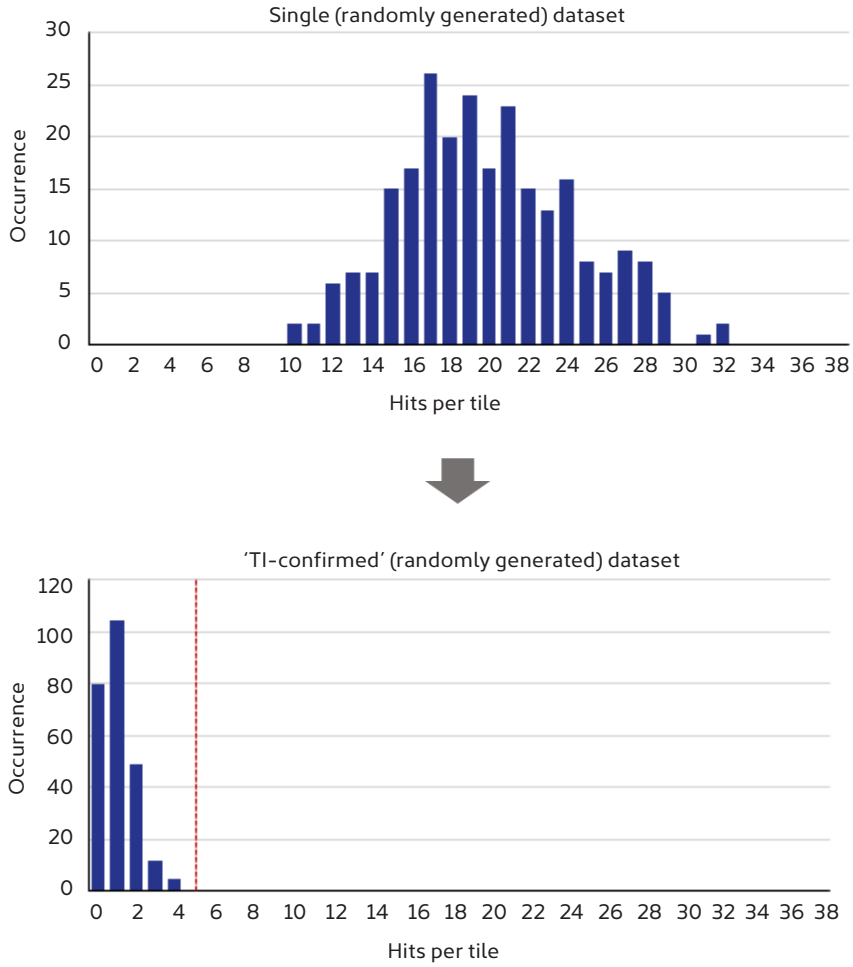


Figure 4

Illustration of the expected number of false positive hits per tile in a single randomly generated dataset (e.g., 70 eV, top) and that of a tandem 'confirmed' dataset (e.g., 70/16 eV, bottom).

Feature Discovery in ChromCompare+ can be used to consider both sets of MS data in a single workflow (Figure 5) and can be set to require five or more TI-confirmed positives per tile for a tile to be considered further. In this way, tiles without true differentiators can be set aside, minimising review time and improving confidence in results.

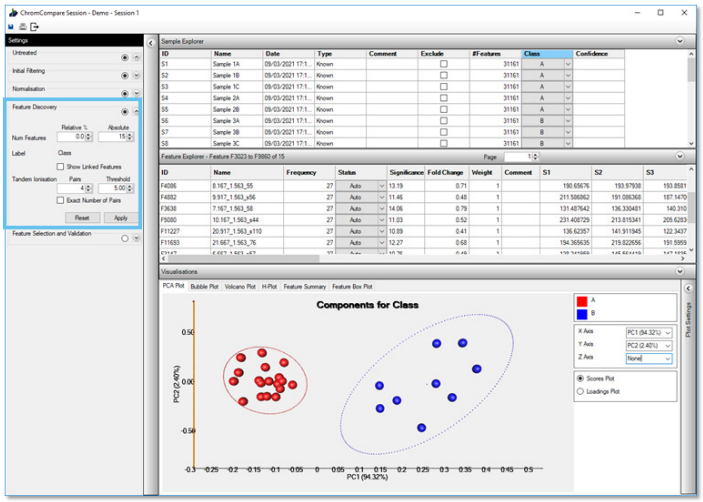


Figure 5

ChromCompare+ project window with the boxed region highlighting the Tandem Ionisation (TI) filter for Feature Discovery.

SepSolve Analytical Ltd
T: +44 (0)1733 669222 (UK) +1 888-379-3835 (USA)
E: hello@sepsolve.com

Conclusions

- Fully untargeted data analysis using the tile-based approach in ChromCompare+ automatically finds the significant differences between complex 1D and 2D GC chromatograms.
- Feature Discovery uses a multivariate method to consider covariance between features and finds those that best differentiate multiple sample classes.
- False positives will exist in every dataset, so it is important to optimise workflows to reduce their frequency (e.g., by analysis of replicates).
- Tandem Ionisation (TI) by BenchTOF2 mass spectrometers provides complementary hard and soft ionisation data in a single analysis.
- Tandem data can be utilised in streamlined workflows to confirm positive hits, thereby reducing the rate of false positives, minimising review and increasing confidence in results.
- An improved discovery of subtle differences is shown, since true positives with a lower F-ratio are uncovered more easily by suppression of false positives.

For more information on this application, or any of the techniques or products used, please contact SepSolve.

References

[1] C.E. Freye, N.R. Moore and R.E. Synovec, Enhancing the chemical selectivity in discovery-based analysis with tandem ionization time-of-flight mass spectrometry detection for comprehensive two-dimensional gas chromatography, *Journal of Chromatography A*, 2018, 1537: 99–108.

D0080_1_010521

SepSolve Analytical Ltd
T: +44 (0)1733 669222 (UK) +1 888-379-3835 (USA)
E: hello@sepsolve.com

www.sepsolve.com