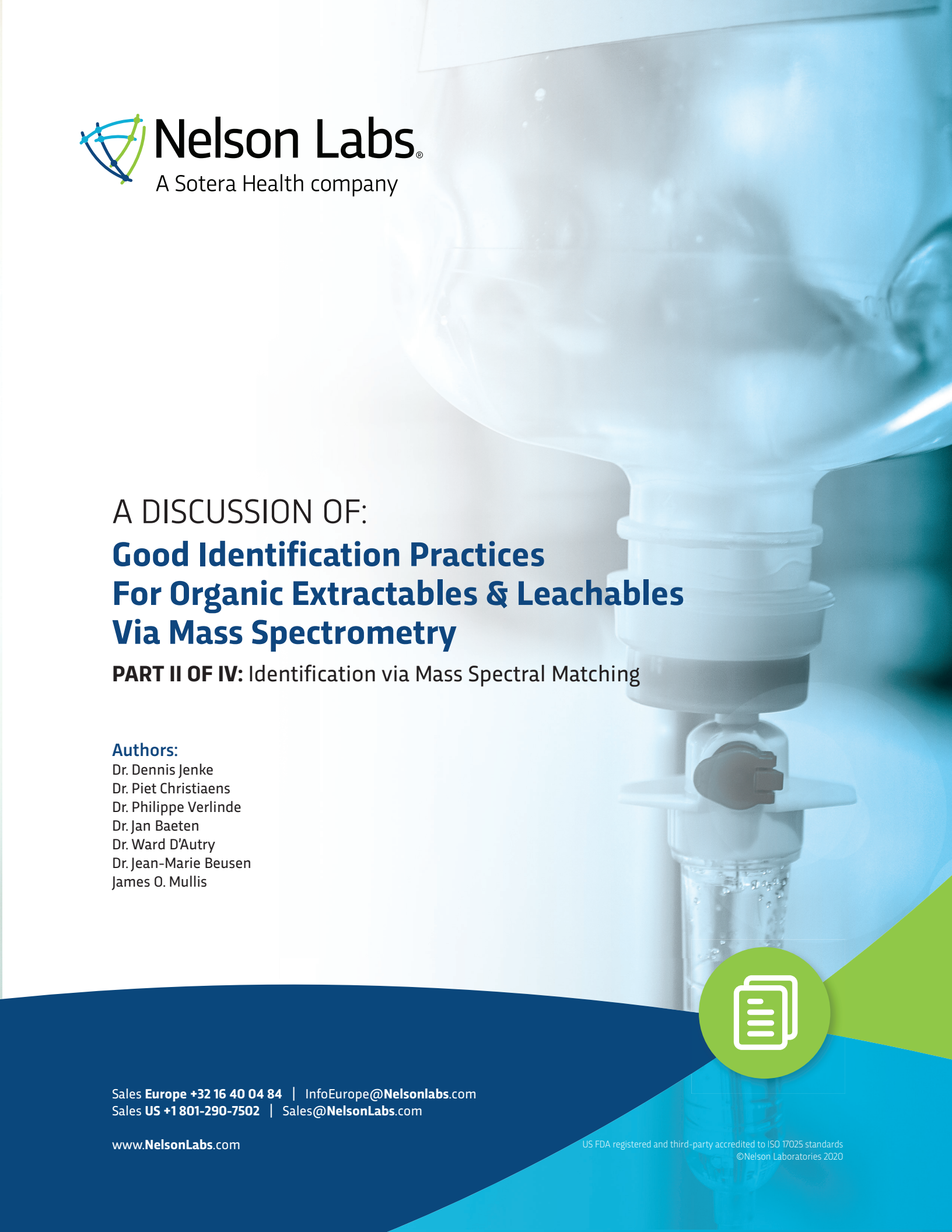
Nelson Labs®
A Sotera Health company



A DISCUSSION OF:

Good Identification Practices For Organic Extractables & Leachables Via Mass Spectrometry

PART II OF IV: Identification via Mass Spectral Matching

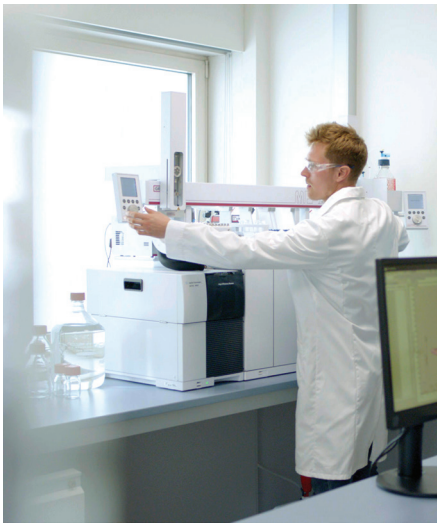
Authors:
Dr. Dennis Jenke
Dr. Piet Christiaens
Dr. Philippe Verlinde
Dr. Jan Baeten
Dr. Ward D'Autry
Dr. Jean-Marie Beusen
James O. Mullis

Sales **Europe +32 16 40 04 84** | InfoEurope@**Nelsonlabs**.com
Sales **US +1 801-290-7502** | Sales@**NelsonLabs**.com

www.**NelsonLabs**.com

PART 2: IDENTIFICATION VIA MASS SPECTRAL MATCHING

OPENING THOUGHTS



Identification of extractables and leachables is a critical aspect of reporting these substances for toxicological safety risk assessment, as it is these substances' identities that establishes their inherent toxicity. Nelson Labs has generated a series of white papers focusing on the aspect of identification and, more specifically, on the process by which mass spectral data and other supporting evidence are used to secure, judge, and justify complete and correct identities for all relevant extractables or leachables surfaced by the chromatographic screening analyses. Part 1 of this series, published previously, introduced the concept of identification, establishing its critical role in safety assessment. Part 1 described in general the various means of securing identities, discussed the concept of identification classes and the importance of confidence in identification, and delineated the identification process via an identification decision tree *(see Figure 6 in Part 1 of this series on Good Identification Practices: Identification Classes, Process and Practices)*. In Part 2 of this series, we will discuss the process of securing a compound's identity via mass spectral matching.

INTRODUCTION TO MASS SPECTRAL MATCHING

The process of mass spectral matching is exactly what you would expect from its name. Screening chromatographic analysis of either an extract, a drug substance, or a drug product produces chromatographic peaks associated with extractables or leachables. When mass spectrometry is employed as the chromatographic detection method, each peak has an associated mass spectrum characteristic of the analyte responsible for the peak. This test spectrum can be compared to a compiled library of reference mass spectra generated via the analysis of authentic reference standards. An acceptable feature-by-feature match between test and reference spectra would suggest that the analyte that produced the test spectrum and the reference compound that produced the reference spectrum are one and the same. In this way, the analyte has been identified via mass spectral matching.

When reliable and up-to-date libraries containing mass spectra of relevant organic compounds are available, mass spectral matching is the most commonly employed and efficient means of establishing a tentative identity for a compound. General discussions of mass spectral matching and interpretation are contained in numerous publications and references as outlined in the Annex of this document *("Concepts supporting the interpretation of identifications using mass spectral matching library results")*.

Mass spectral matching is, in essence, the search for those library spectra that are similar to the test spectrum of the compound of interest. The primary outcome of library matching is a list of "hits" to those library spectra that have some level of similarity to the test spectrum. In most commercially available libraries "hits" are further delineated with the hits' critical identifying information, names, chemical structures, CAS numbers, and possibility other information. Each "hit" is typically accompanied with a numerical value *(e.g. match factors, probability scores, etc.)* that is obtained by an established matching algorithm, which indicates the degree to which the hit's mass spectrum corresponds to the test spectrum. The simplest interpretation of the "hit list", is that the hit with the highest match score is taken as the identity of the compound of interest. Even if the top hit is not taken as the compound's identity, it is generally "assumed" that the proper identity for the compound of interest is among the highly ranked reference compounds.

"Although the concept is intuitive and straightforward, it is not without its challenges. For example, this identification strategy can only be successful if the mass spectrum of the compound of interest is actually present in the library. If the mass spectrum of the compound is missing in the spectral library, the identification strategy via mass spectral matching cannot possibly lead to the proper identification of the compound of interest."

If the compound's mass spectrum is present in the library, the best-case matching outcome will be that the library spectrum with the highest ranked hit was, in fact, produced by the compound. While it would be desirable if this were always the case, there is absolutely no guarantee that this will be the case in all circumstances. Whether the right match is the best hit depends upon the selected mass spectral match criteria and the similarity between the analytical conditions used to produce the test and reference spectra.

Because the match score alone does not reliably produce the right identity every time, there is a level of uncertainty in a tentative identity secured in this way. However, greater certainty in the identification can be secured if an experienced mass spectrometrist reviews the available spectral information and is able to substantiate the identification.

RELEVANCE OF EXTERNAL MASS SPECTRAL LIBRARIES FOR IDENTIFICATION

An ideal mass spectral library for the identification of extractables and leachables would:

1. Contain as many of the possible extractables and leachables as possible *(increasing the likelihood of securing a match)*
2. Contain no substances that are not extractables or leachables *(decreasing the possibility of an incorrect match)*
3. Be well controlled and maintained
4. Be constantly and routinely updated *(in a controlled manner)*
5. Be peer-reviewed
6. Contain spectra that are secured under standardized analytical conditions
7. Have scientifically validated search algorithms and scientifically vetted means of establishing the match factor
8. Have a user-friendly output that supports review, interpretation, and assessment
9. Be compatible with all data platforms used in all generally available commercial instrumentation
10. Be universally available, accessible and employed

Considering point #1, it is likely that many extractables and leachables will be included in the most commonly used commercial spectral libraries *(e.g. NIST, Wiley)* because these extractables and leachables are themselves commonly encountered organic compounds such as residual solvents, monomers, processing aids, and commonly used additives *(anti-oxidants, plasticizers, slip agents, acid scavengers, nucleating agents, curing agents, and others).* However, there is also a large population of extractables and leachables whose mass spectra will not likely be included in the most commonly available commercial mass spectral libraries. These compounds are

infrequently encountered degradation and reaction products formed either in the production process of the test item or during the period of time that the test item is in contact with the "communicating" entity *(for example, a drug product stored in its container closure system over its shelf-life)*. Such infrequently encountered substances arise by a variety of mechanisms; for example, via oxidation reactions of the polymer or its additives or impurities, hydrolysis reactions, sterilization degradation *(cleavage, cyclization…),* interactions between polymer additives/impurities during the formation process, oligomer formation and reactions, etc. Although a number of these degradation compounds may be known to the industry – and their mass spectra may be represented in internal, closely held mass spectral libraries – they are less likely to have been reported and incorporated into the commercial databases.

"...extractables and leachables that are in the commercial databases are the "low hanging fruit" as they have probably already been encountered in numerous studies and are therefore relatively well-known. It is the fruit that is more difficult to reach, the rarely or never-before-encountered compounds with complex and largely unrecognizable structures (and thus spectra) that will be the most challenging compounds to identify."

Moreover, those extractables and leachables that are in the commercial databases are the "low hanging fruit" as they have probably already been encountered in numerous studies and are therefore relatively well-known. It is the fruit that is more difficult to reach, the rarely or never-before-encountered compounds with complex and largely unrecognizable structures *(and thus spectra)* that will be the most challenging compounds to identify. If these compounds are not in the commercial libraries, some other means will be required for securing their identities.

The two most commonly used hyphenated chromatographic methods for the screening of organic extractables or leachables are gas chromatography – mass spectrometry *(GC/MS)* and liquid chromatography – mass spectrometry *(LC/MS).* Mass spectral matching is a particularly powerful tool for securing identities in GC/MS for the simple reason that the operating conditions for the mass spectrometer in GC/MS have been standardized *(e.g. electron impact (EI) mass spectra that were recorded at an ionisation energy of 70 eV)*. Because of the highly standardized mass spectrometric data acquisition parameters, EI mass spectra are very reproducible across different GC/MS platforms and reproducible across test systems and commercial libraries. Well-known and well-controlled large commercial reference libraries such as the NIST/EPA/NIH Mass Spectral Library and the Wiley Registry of Mass Spectra are widely used tools to facilitate spectral matching of GC/MS data. Such databases fulfil most of the ideal requirements listed previously with the notable exception that the databases contain many more compounds that are not extractables or leachables than they contain compounds that are extractables and leachables, thus increasing the likelihood that false and generally impossible identifications are secured.

While external mass spectral libraries are routinely used for mass spectral matching as a first means to identify a compound in GC/MS, this is not the case for LC/MS. The fact that either in-source fragmentation spectra or multi-stage MS (MS^n) fragmentation spectra are necessary for mass spectral matching in LC/MS, adds to the complexity of the identification process. As there are no standard ionization and fragmentation settings for LC/MS^n detection and because the ionization can greatly be influenced by the chromatographic conditions, no commercial libraries are available that can readily and reliably be used to perform a useful first pass mass spectral matching for each and every LC/MS instrument platform. Although NIST and some instrument vendors have started to create collision-induced dissociation *(CID)* based MS^n spectral libraries that can serve as a resource to scientists who seek to establish a compound's identity in LC/MS, one should be careful when using these data as the mass spectra were acquired with a certain and specific combination of instrumental detector settings. Wrong selection of the precursor ion or deviations between the experimental conditions or instrument type used to collect the test and the library spectrum may produce aberrations between the experimental mass spectrum and the library mass spectrum, complicating the identification process and leading to lower confidence in its outcome. As a consequence, mass spectral matching is not widely applied as a routine practice for securing a tentative identity via LC/MS. Rather, LC/MS matches should be considered to be supporting information for identifications made by other methods, for example, *de novo* structure elucidation using mass spectral interpretation by experienced mass spectrometrists.

"As there are no standard ionization and fragmentation settings for LC/MSN detection and because the ionization can greatly be influenced by the chromatographic conditions, no commercial libraries are available that can readily and reliably be used to perform a useful first pass mass spectral matching for each and every LC/MS instrument platform."

EVALUATION OF MASS SPECTRAL MATCHING RESULTS

As noted previously, mass spectral matching can result in the tentative identification of detected compounds, producing the minimum level of information suitable for subsequent toxicological assessments. However, the use of undisciplined or unsubstantiated mass spectral matching is problematic. This is because the decision that the match is the correct identity is often based solely on the calculated similarity values *(i.e. match factors)* without subsequent critical review of the spectra. A high correlation between an experimental spectrum and a library spectrum does not necessarily mean that the identification is unequivocal. Moreover, unilaterally choosing the highest ranked hit as the reported identity is problematic at best and has been established to generate false positives in many cases [1].

Due to the complexities of mass spectral interpretation, specifying "objective" criteria for a mass spectral matching identification strategy is not straightforward. Match factors *(MF)* are calculated values which indicate the similarity of two spectra by comparing individual m/z values and their

corresponding intensities *(see Annex).* In practice, however, there is no "universal" match factor (MF) threshold value that exclusively establishes that the corresponding match identity represents the true identity of a compound. The underlying reason for this is the varying degree of spectral uniqueness among the universe of chemical compounds. Certain compounds may have a rather unique spectrum and are thus more likely to be correctly identified, while others may have a spectrum that very closely resembles the spectra of many other compounds, making identification a challenge. Nevertheless, the "goodness" of a mass spectral match factor can be correlated with the probability that the match factor has suggested the right compound. The lower a mass spectral match factor, the lower the quality of the fit and the more mass spectral interpretation efforts are necessary to justify an identification decision that is solely based on mass spectral matching. These efforts may include, for instance, inspection of a mirror image of both experimental and library spectrum to reveal the presence of additional or missing m/z values in either spectrum. Additionally, when using the NIST MS search software, the probability score and *In Lib* score can be evaluated *(see Annex).* The former represents the relative probability that any matching spectrum in the hit list is correct, while the latter is a measure of the probability that the spectrum of the compound being searched is effectively contained in the library *(see Annex).* Hits with an MF below 700 are generally associated with a very low probability that the identification is correct. Nevertheless, identifications based on an MF below 700 have been reported by testing laboratories, particularly in the case when an MF below 700 is the highest ranked, or the only match.

"... there is no "universal" match factor (MF) threshold value that exclusively establishes that the corresponding match identity represents the true identity of a compound."

The uncertain nature of identification by mass spectral matching leads to the conclusion that mass spectral matching based identifications always require a close examination by a mass spectrometrist. Also, these identifications need to remain TENTATIVE unless they are corroborated by additional evidence, such as information obtained through analysing the authentic standard *(mass spectrum and retention time)* or additional supporting documentation (which will be explained in Part 4 of this series on Good Identification Practices, *"Additional Evidences supporting Higher Level Identifications")*. Various approaches can be used to review mass spectral matching results *(see Annex)* and the complexity of the review could vary upon the quality and the number of the returned match results from the search as illustrated in the following examples.

"... the "goodness" of a mass spectral match factor can be correlated with the probability that the match factor has suggested the right compound. The lower a mass spectral match factor, the lower the quality of the fit and the more mass spectral interpretation efforts are necessary to justify an identification decision that is solely based on mass..."

Various maximum MF values are used by different software vendors. In the examples below, MFs are expressed relative to a maximum value of 999. Refer to the Annex of this document for a detailed description of the match factor and the reversed match factor.

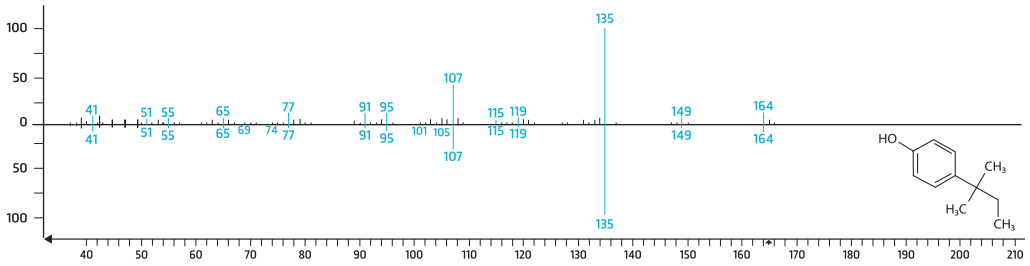
EXAMPLE 1: CORRECT IDENTIFICATION FOR BEST HIT (MF > 900); GC/MS

The top-3 hit list for a compound of interest is presented in Table 1 with the respective MS spectra being displayed in Figure 1. All 3 hits have relatively high match factors. The highest ranked candidate has a high match factor of 931 and visual review of the mass spectra shows an almost perfect mirror image match, without missing characteristic ions, in either the experimental or library mass spectrum *(A).* This is in contrast to the two lower ranked hits, where image match is not nearly perfect. The second ranked spectrum contains additional peaks at m/z 177, 191 206 which are not detected in the experimental spectrum *(B).* The third ranked spectrum contains an additional peak at m/z 150 and lacks peaks at m/z 149 and 164 compared to the experimental spectrum *(C).* These observations lead to the conclusion that the top hit, p-tert-pentylphenol, has a high probability of being the correct tentative identity for the compound. In addition, comparison of the unknown's retention index *(1394)* to the experimental retention indices present in the library further corroborates the identification and could allow upgrading the identification to the confident level. *(for more background information on the relevancy of the RI, see Part 4 of this series of documents: "Additional Evidences supporting Higher Level Identifications")*

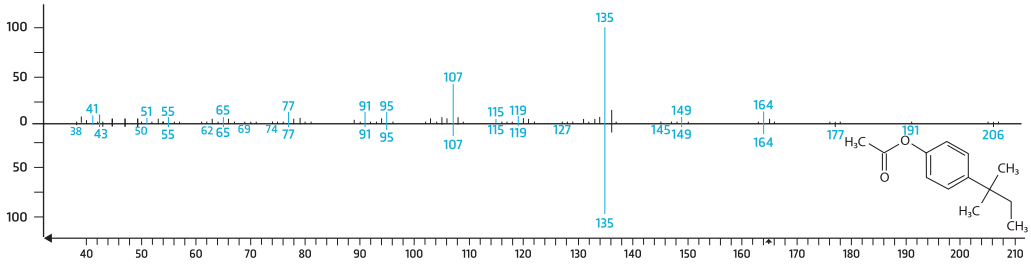
Rank	Candidate	Match	R.Match	Prob(%)	Retention Index
1	p-tert-pentylphenol	931	931	81.2	1400 ± 4 (*n* = 5) *
2	p-tert-pentylphenol acetate	825	825	5.45	1502 (*n* = 1) *
3	p-tert-butylphenol	816	852	3.96	1295 ± 3 (*n* = 9) *

In Lib Score = 312 (*) Experimental Retention Index in library for relevant stationary phase (median ± deviation (# of entries))

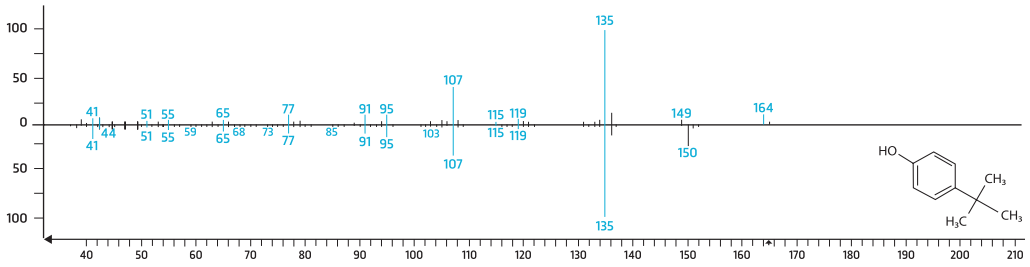
Table 1. *Top 3 ranked hit list for a mass spectral matching example with an excellent (> 900) match score and a high probability of securing a correct tentative identity for a compound after review of the hit list and spectra (Figure 1) by an expert mass spectrometrist (mass spectral matching using NIST MS Search v2.3). Experimental RI = 1394.*



A: Rank 1 candidate spectrum



B: Rank 2 candidate spectrum



C: Rank 3 candidate spectrum

Figure 1. *Mirror mass spectrum plots of an unknown (top) and its match factor based top three ranked identification candidates where the Rank 1 candidate represents the correct identification (Table 1).*

EXAMPLE 2: CORRECT IDENTIFICATION FOR BEST HIT (800 < MF < 900); HS-GC/MS

This example shows a top-5 hit list for a second compound of interest *(Table 2)* where only the best hit has a match factor above 800. Visual inspection of the spectra *(Figure 2)* shows a good mirror plot for the best ranked hit; the only marked difference is the relative intensity of the peak clusters at m/z 249 and m/z 265. The lower ranked hits, on the contrary, show clearly deviating features such as additional or missing m/z values and very different relative intensities. In addition, Table 2 shows that the probability score of the best hit is very high and markedly different from the lower ranked hits. Therefore, the compound can be tentatively identified as octamethyl cyclotetrasiloxane with a high degree of confidence.

Rank	Candidate	Match	R.Match	Prob(%)	Retention Index*
1	Octamethyl cyclotetrasiloxane	829	842	92.7	
2	1,1,3,3,5,5,7,7-Octamethyl-7-(2-methylpropoxy)tetrasiloxan-1-ol	757	757	4.04	
3	3-Ethoxy-1,1,1,5,5,5-hexamethyl-3-(trimethylsiloxy)trisiloxane	730	738	1.18	
4	3-Butoxy-1,1,1,5,5,5-hexamethyl-3-(trimethylsiloxy)trisiloxane	726	739	1.0	
5	2,6-Dihydroxyacetophenone, 2TMS derivative	697	697	0.28	

In Lib Score = 237

(*) No retention indices available for low-medium polarity stationary phase with which data were acquired.

Table 2. *Top 5 ranked hit list for a mass spectral matching example with a good (800-900) match score and a high probability of securing a correct tentative identity for a compound after review of the hit list and spectra (Figure 2) by an expert mass spectrometrist (mass spectral matching using NIST MS Search v2.3).*

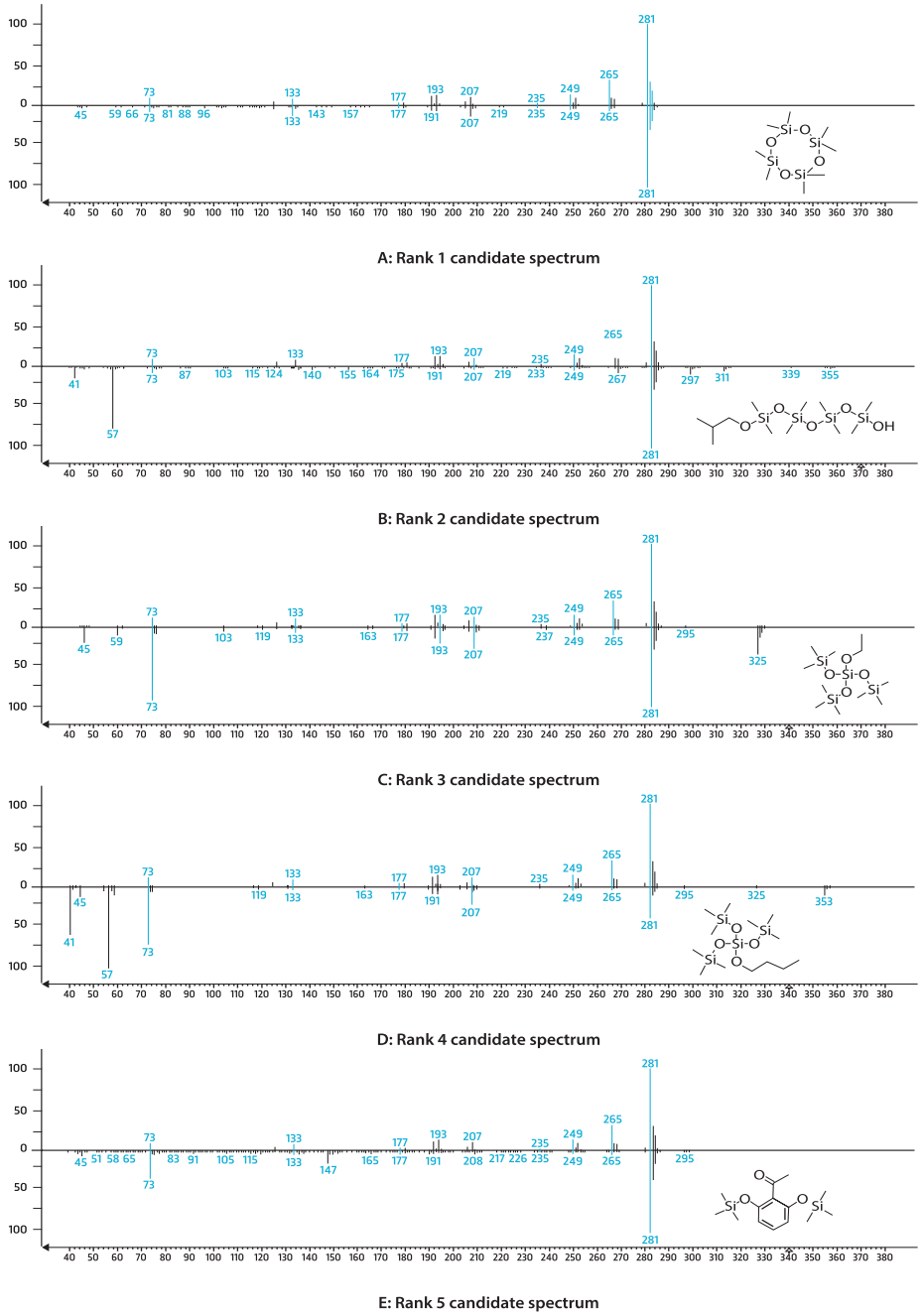


Figure 2. *Mirror mass spectrum plots of an unknown (top) and its match factor based top five ranked identification candidates where the Rank 1 candidate represents the correct identification (Table 2).*

EXAMPLE 3: INCORRECT IDENTIFICATION FOR BEST HIT (800 < MF < 900); GC/MS

Table 3 shows another example for a third compound of interest where the highest MF value is approximately 800. Upon cursory examination, this top hit seems to provide a good and acceptable match. Mirror plots for the five best hits *(Figure 3)*, however, show that none the library spectra are very good matches to the experimental spectrum. Either there are a number of characteristic ions which are present in the experimental spectrum and not in the library spectrum, or vice versa. For instance, peaks at m/z 42, 55 and 73 in the best hit library spectrum *(Figure 3A)* are missing in the experimental spectrum which indicates that this identity is incorrect. The same decision can be made for the other hits *(Figure 3B-D)* because they either lack characteristic ions in the unknown spectrum or have deviating relative intensities.

Moreover, the In Lib value of the mass spectral search *(Table 3)* suggests that the compound being searched is not present in the mass spectral library and the unknown's retention index *(916)* does not correspond to any retention index in the hit list *(for more information on retention index (RI) considerations, see Part 4 of this series of documents: "Additional Evidences supporting Higher Level Identifications")*. Consequently, further mass spectral interpretation efforts by an expert are necessary to identify the compound of interest (*see Part 3 of this series: "Identification by Mass Spectral Interpretation").* Such an evaluation of the spectrum according to the principles described in Part 3 of this series of documents on Good Identification Practices *("Identification via Mass Spectral Interpretation")* would reveal that the concerned compound is a silicon containing molecule, which would be considered to be a PARTIAL identification.

Rank	Candidate	Match	R.Match	Prob(%)	Retention Index
1	Methoxy-phenyl-oxime	804	822	85.3	1301 ± 382 °
2	Cyclopentyl 4-ethylbenzoate	698	710	5.73	1713 ± 201 °
3	Sec-butyl 4-ethylbenzoate	671	677	1.68	1507 ± 201 °
4	Cyclohexyl 4-ethylbenzoate	668	679	1.49	1833 ± 201 °
5	Isobutyl 4-ethylbenzoate	652	678	0.85	1507 ± 201 °

In Lib Score = 135

(°) Estimated Retention Index in library for relevant stationary phase (estimated value ± 95% Confidence interval)

Table 3. *Top 5 ranked hit list of an example for a flawed mass spectral matching exercise with a moderate and low match scores and a low probability of securing a correct tentative identity for a compound after review of the hit list and spectra (Figure 3) by an expert mass spectrometrist without additional information (mass spectral matching using NIST MS Search v2.3). Experimental RI = 916.*

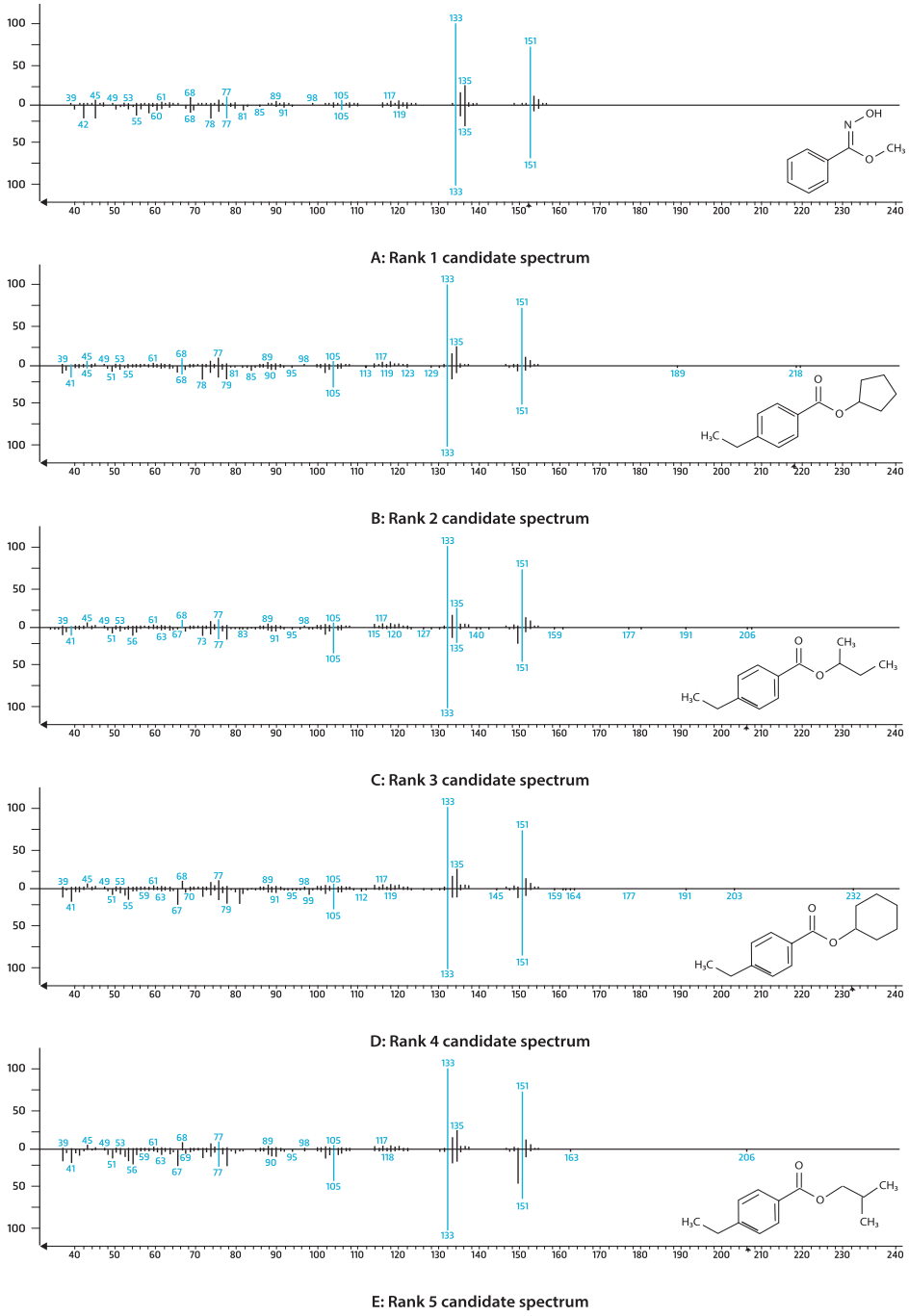


Figure 3. *Mirror mass spectrum plots of an unknown (top) and its match factor based top five ranked identification candidates where none of the identification candidates is correct (Table 3).*

EXAMPLE 4: FALSE POSITIVE IDENTIFICATION FOR BEST HIT (700 < MF < 800)

Identification based on mass spectral matching becomes even more difficult when the quality of the match factors deteriorates further, as reflected in even lower MF values. An example of this is shown in Table 4 and the associated Figure 4, where the MF values are between 750 and 700 for the five best ranked hits. A visual inspection, performed by an experienced mass spectrometrist, would reveal that none of the candidate library spectra fit the experimental spectrum of the compound of interest. Consequently, additional efforts in mass spectral interpretation are essential to secure the correct identity of this compound.

Rank	Candidate	Match	R.Match	Prob(%)	Retention Index
1	4a,8a-Dimethyloctahydro-2(1H)-napthalenone	743	743	17.5	
2	1-Methyl-1-cyclododecene	723	723	7.93	1387 ± 5 (n = 2) *
3	Tetrahydroionyl acetate	713	713	5.62	
4	Neophytadiene	710	775	4.97	1837 ± 5 (n = 19) *
5	(2,2-Dimethylcyclopentyl)cyclohexane	709	712	4.77	

In Lib Score = 733

(*) Experimental Retention Index in library for relevant stationary phase (median ± deviation (# of entries))

Table 4. *Top 5 ranked hit list of an example for a flawed mass spectral matching exercise with a moderate and low match scores and a low probability of securing a correct tentative identity for a compound after review of the hit list and spectra (Figure 4) by an expert mass spectrometrist without additional information (mass spectral matching using NIST MS Search v2.3). Experimental RI = 1178.*

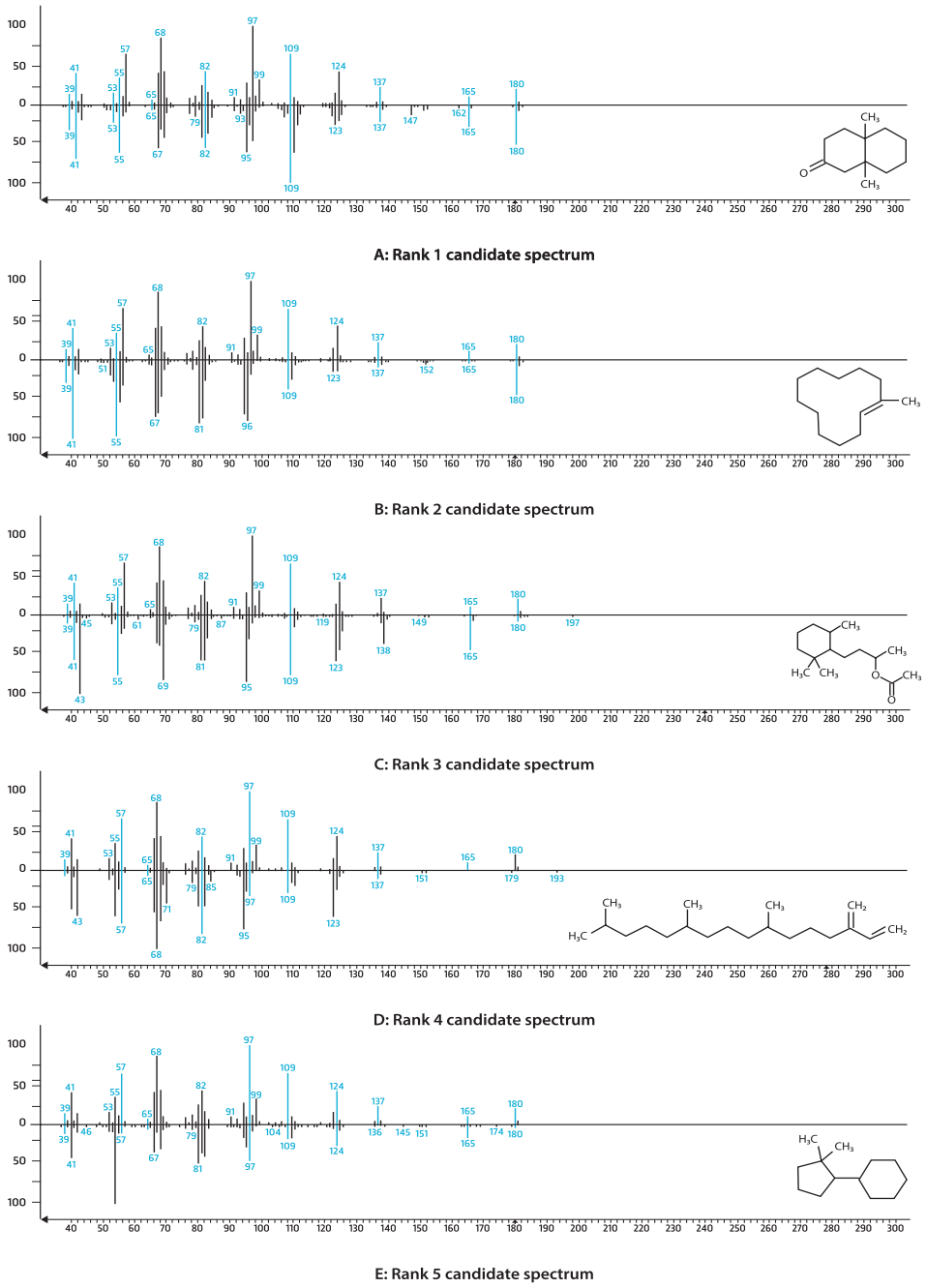


Figure 4. *Mirror mass spectrum plots of an unknown (top) and its match factor based top five ranked identification candidates where none are the correct.*

CONCLUSION:

Once a mass spectrum has been obtained, the process of identification begins in one of two ways, spectral matching or structure elucidation. Considering spectral matching, it is noted that this technique is most productive when it is applied to GC/MS data as large, well-maintained and standardized commercial libraries of spectra exist. Nevertheless, even in this circumstance, spectral matching is not an exact science and the following recommendations are made to assist in securing the right identity:

- It should be stressed that there is no mass spectral match factor *(MF)* that unequivocally guarantees that the correct identity of the compounds has been determined, based upon the MF alone. It is clear that the exclusive reliance on mass spectral match factors without any expert review cannot robustly and routinely lead to correct identifications. Reporting the highest ranked hit as an analyte's tentative identity is an all-to-common error and matching based on absolute MF thresholds can still lead to false positive identifications. Therefore, the practice of reporting the identity of an analyte of interest as the compound with the highest match score by default is strongly discouraged. Any proposed identity should be verified by an expert mass spectrometrist, which may require that the identification hypothesis is corroborated by additional interpretation efforts or information.

- An expert should always visually evaluate the spectra of match candidates in the mass spectral matching's hit list, regardless of the quality of the MF. This evaluation serves as a means to compare the mass spectra of the target mass spectrum with the library spectrum to verify the resemblance of all mass fragments in both mass spectra.

- It is consequently considered as good and necessary practice that tentative identifications based on mass spectral matching are always substantiated by comparative spectrum plots such as mirror plots as such comparative data allows one to visually confirm the quality of, and increase the confidence in, the fit of the matched spectra .

- The lower the MF, the more intense the mass spectral interpretation exercise will need to be to secure the identity of the compound based solely upon the merits of its mass spectrum.

- While MF's above 80% *(or 800 depending on the scoring scale)* may lead –after a careful mass spectral evaluation – to an unequivocal identification, the probability of securing the right identity via mass spectral matching decreases quickly below the above stated values. When the MF's are further deteriorating, e.g. below 70% *(or 700)*, the probability of correctly identifying a compound primarily based upon MF is extremely low. In that case the most likely outcome will be that the compound remains Unidentified, although it is possible that the match is sufficient to substantiate and support a PARTIAL identification.

- Identifications of organic compounds based solely upon the practice of mass spectral matching should be considered as TENTATIVE identifications, as the identification is a "one-dimensional" identification where the one piece of evidence is its mass spectrum. The Identification Class can be augmented by acquiring information obtained through analysing the authentic standard *(mass spectrum and retention time: CONFIRMED IDENTITY)* or through additional supporting documentation as explained in Part 4 of this series on Good Identification Practices *("Additional Evidences supporting Higher Level Identifications").*

MOVING FORWARD

In the first two Parts of this Series, the importance of proper, correct, and confident identification as a cornerstone of efficient and effective toxicological safety risk assessment of organic extractables and leachables has been established. The various idiosyncrasies of identification have been considered and an identification process and identification classification have been delineated. Moreover, the identification technique of spectral matching has been discussed in detail, with the recommendation that no matter how good the match is via numerical metrics *(e.g., match score)*, the highest match score does not always provide the correct identification and that any identification based on matching should be reviewed by a trained mass spectroscopist and substantiated by corroborating data such as retention index match.

In Part 3 of this series, a second important means of securing an organic compound's identity, mass spectral interpretation, will be addressed.

ANNEX: Concepts supporting the interpretation of identifications using mass spectral library search results:

Detection and discrimination of analyte signals (spectra) for identification

The process of discovering, identifying, and quantifying organic extractables in extracts *(or organic leachables in drug products)* involves the analysis of the extract using compatible and orthogonal hyphenated chromatographic techniques, typically gas chromatography/mass spectrometry *(GC/MS)* and high-performance liquid chromatography/mass spectrometry *(LC/MS)* [2]. These hyphenated techniques yield at least two-dimensional compound specific information for analytes present in an extract in the form of chromatograms which can be used to make inferences regarding the analyte's chemical structure based on chromatographic retention behaviour and mass spectral data.

Chromatographic selectivity is one of the determining parameters in terms of the reliability of the discovery and identification processes. When chromatographic selectivity increases, peak co-elution decreases, and the discriminating power of the analysis method improves. However, peak co-elution cannot generally be avoided, and it becomes a challenge to resolve peak responses sufficiently so that useful, uncompromised responses can be obtained for the coeluting analytes. It is obvious that unreliable identifications arise when the determined mass spectra of the detected analytes are compromised due to spectral contamination. Typical sources of spectral contamination include ion signals from coeluting compounds, column bleed, solvent tailing, or even electronic noise.

As visual inspection of complex chromatograms is an ineffective means of resolving chromatographic peaks and their associated mass spectra, application of data processing techniques that can account for such phenomena are necessary. When determining the mass spectra that serve as the basis for identification either through library matching and/or mass spectral interpretation, background subtraction should be the absolute minimum approach. To support a higher quality of identification, however, application of a scrutinous deconvolution-based approach is more preferable as it delivers better quality mass spectra *(i.e. free of interferences)* and thus reduces the risk of misidentification.

Deconvolution is the process of computationally extracting analyte signals from a complex mass chromatogram, resulting in the elimination of background noise and the spectral separation of co-eluting compounds. An example is given in Figure 5.

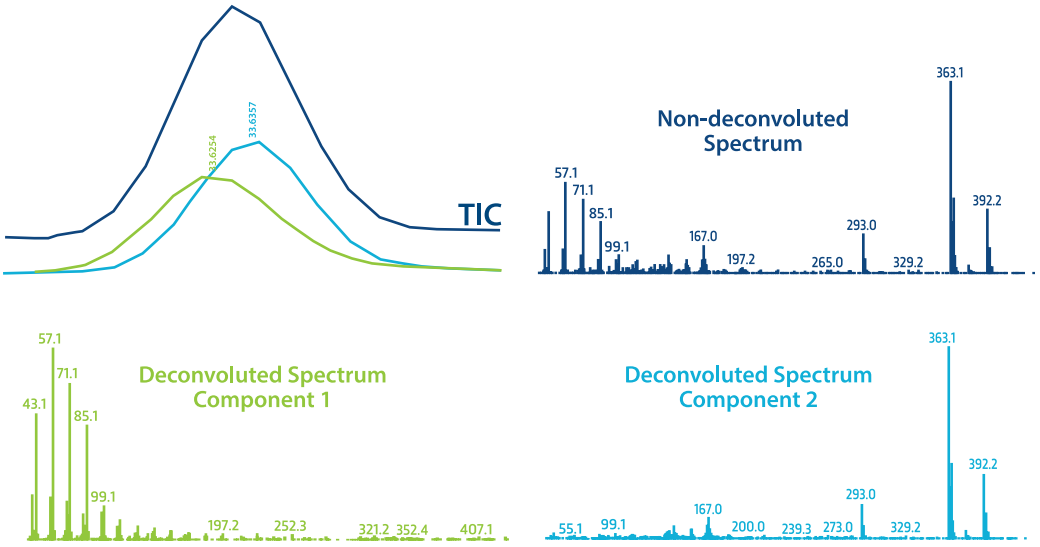


Figure 5. *Deconvolution example of a single signal peak detected in a Total Ion Current (TIC) chromatogram that is composed of two closely eluting components.*

The deconvolution process involves multiple steps such as noise analysis, peak shape analysis *(ions belonging to the same peak should have the same apex and peak shape),* and the assembly of a deconvoluted spectrum [1]. It should be noted that deconvolution algorithms and related parameters can differ among various software platforms and therefore it is possible to generate slightly different deconvoluted spectra for the same raw data.

To facilitate identification, most instrument software platforms enable the combination of deconvolution with a mass spectral library search, *i.e.* comparison of the resulting spectra against collection(s) of reference mass spectra included in either public, commercial, or user defined libraries using matching algorithms. The result is an *(indexed)* hit list containing the most similar spectra present in the searched libraries. Spectral library search programs are very powerful supporting tools for the identification of analytes in non-targeted analysis and are a well-established strategy that has long been applied in the reporting and identification of detected *(non-target)* analytes in environmental analysis [10]. Relying solely on mass spectral matching to secure an identity, however, is an inappropriate interpretation practice, as more incorrect results are probably reported because of sole reliance on mass spectral library search programs than have been reported due to all other types of errors that can occur in mass spectral data [3]. A high correlation *(high match index or probability of match)* between an unknown spectrum and a library spectrum does not necessarily mean that the unknown has been identified unequivocally. Additionally, the match with the highest match score is not

always the correct identity. The criteria needed to identify an unknown by chromatography mass spectrometry must include a visual comparison of the unknown and library spectrum as documented by an expert mass spectrometrist and may demand that the identification hypothesis is corroborated by additional information such as its expected retention behaviour, etc [4].

PRINCIPLES UNDERLYING THE INTERPRETATION OF MASS SPECTRAL LIBRARY SEARCH RESULTS:

A number of mass spectral libraries are available for matching, with the NIST/EPA/NIH Mass Spectral Library and the Wiley Registry of Mass Spectra being the largest. The continuous evolution and development of the NIST/EPA/NIH Mass Spectral Library since the early 1970's makes this mass spectral search program *(NIST MS Search Program)* the most widely distributed and used.

Despite the availability of large collections of reference spectra, it is important to consider that most methods that involve the confirmation of target analytes by mass spectrometry require the use of a library of mass spectra that were obtained on the specific instruments used for analysis. Taking this into account, it is recommended that E&L laboratories establish their own mass spectral library of spectra using reference standards of the analytes of interest to minimize the ambiguity of identifications. This is of utmost importance since continuous proprietary evolutions in the material manufacturing business imply that an important part of the chemical space that is relevant for the E&L field is not being disclosed to the public and hence is not covered by the previously mentioned commercial libraries.

Mass spectral libraries and library search software have historically provided mainly search possibilities for GC/MS EI *(electron ionization)* spectral data, whereas the availability of reference LC-MS/MS spectra has steadily grown over the last two decades. The NIST 2017 library for example contains EI spectra for 267376 chemical compounds, compared to MS/MS spectra for 13808 chemical compounds. With the recent rise of modern, accurate, mass-high-resolution instrumentation capable of generating multi-stage MS(n) mass spectra, identification of unknowns in non-targeted screening is requiring new analytical data processing strategies to take full advantage of the additional level of information present in this type of data [5].

Library search algorithms:

Several algorithms have been developed to express the degree of similarity between the spectrum of an unknown analyte and the spectrum of a reference compound as a *match factor*. In most modern software platforms primarily two search algorithms are used: the PBM or *probability based matching,* and the INCOS or *cosine / dot-product* algorithm *(or its derivatives).*

Probability based matching algorithm:

Originally developed by McLafferty, the PBM algorithm uses a "weighting" and "reverse search". The "weighting" in PBM is used to determine the importance of peaks based on masses and abundances and is used in the pre-search. If a sample and library spectrum do not have the same base peak, it is very likely that that library spectrum will be excluded. The probability that particular abundances will occur follows a log-normal distribution. The probability of most mass

values also varies in predictable manner. The larger molecular fragments tend to decompose into smaller fragments; so, the probability of higher m/z values decreases by a factor of two approximately every 130 mass units. This weighting system is used for indexing the "Important Peak Index of the Registry of Mass Spectral Data".

The second feature of PBM, "reverse searching", treats peaks in the submitted spectrum and not in the library spectrum as if they are from another compound. This is valuable in identification of the spectrum of more than one compound. The PBM algorithm compares the submitted spectrum against the Important Peak Index. Spectra found by this comparison are then evaluated against the spectra in the entire library *(or a condensed version of it).*

It should be noted that because of the way that the "Important Peak Index" is developed, the PBM search algorithm is limited in its performance when small libraries are considered.

Dot-Product (INCOS) and NIST MS Search algorithms

The Dot-Product *(INCOS)* library search system uses a preprocessing step of matching the eight most intense peaks in the submitted spectrum with a set of the 16 most intense peaks in the library spectrum. Subsequently a search is performed against the 50 most chemically significant peaks in the spectra retrieved by the pre-search. The search applies a weighting of the intensity, which considers that peaks of higher mass are more significant than those of lower mass. The comparison is done relative to adjusted abundances based on the square root of the observed peak intensity times the m/z value.

The NIST algorithms are similar to the INCOS algorithm with improvements such as the pre-search as well as the main search include peak intensity scaling and ion mass weighting to increase the significance of lower intensity and higher mass peaks respectively. In the latest version *(2017 NIST MS Search v.2.3)* three search algorithms are implemented, the Normal-Identity [1] [6] the Simple-Similarity [6] and Hybrid-Similarity [7]. The user has control of different factors that can affect the search result *(such as whether or not neutral loss logic is used).* Identity search is designed to find exact matches of the compound that produced the submitted spectrum and therefore presumes that the unknown compound is represented in the reference library. In other words, only experimental variability prevents a perfect match of the unknown and reference spectrum. However, when it is expected that a compound is not present in the reference library or when a compound cannot be identified by the Identity Search, the Similarity Search is optimized to find compounds that exhibit a "similar" spectrum to the submitted spectrum to support the mass spectrometrist in inferring structural information of the unknown. It should be noted that the pre-processing algorithms for the Similarity Search are similar to the Normal Identity Search except that scaling and maximum mass peaks are not used.

Interpretation of *(NIST MS Search)* mass spectral library search results

Terminology

When using the NIST Identity Search, spectrum search results in a hit list are summarized using four numeric descriptors: the *Match Factor* value, the *Reverse Match Factor* value, the *Probability* value and the *InLib probability* value. Maximal values for the descriptors represent a perfect match and are 1000, 100, or 1 depending on the data system.

The *Match Factor* for the unknown and the library spectrum assumes that the former originates from a single compound and uses all peaks in both spectra for spectral similarity determination, in other words, it is a direct match of peak m/z values and relative intensities *(pure spectrum match factor).*

The *Reverse Match Factor* for the unknown and the library spectrum assumes that the former spectrum can be contaminated by "impurities". In its calculation, peaks in the unknown spectrum that are missing in the library spectrum are disregarded *(impure spectrum match factor).* The *Reverse Match Factor* consequently enables the identification of multiple compounds represented by a single spectrum. The closeness of the *Match* and *Reverse Match factor* values should consequently be considered as a measure of the 'purity' of the similarity.

The *Probability* value describes the likelihood of the unknown and reference spectrum being from the same compound based on all the matches found during the search. It is derived if the compound is represented by a spectrum in the libraries and uses the differences between adjacent hits in the hit list to determine the relative probability that any hit in the list is correct. This value is derived from an analysis of the results of searching the NIST/EPA/NIH Main Library with a set of replicate spectra *(given in the Replicates Library)*. The relative probability of each of the hits requires only the difference values because the total probability of the compound being in the searched libraries is assumed to be one [14]. When the best hit has a high *Match Factor* value *(>900)* and the next hit has a much lower value *(e.g. 800 or less),* the *Probability* value means that the probability of the compound being correctly identified is very large and that the probability of the compound being in the searched library is also large. When the *Probability* is high, it means that – apart from the hit – there are no other good matching mass spectra in the library which makes the hit 'unique' and obviously increases the likelihood of a correct tentative identification. When the *Probability* is low, it means that there are other good matching mass spectra present in the library, which makes it difficult to pick the best hit. This is typical when isomers exist *(e.g. xylenes)*. Caveat: this descriptor assumes that the target molecule is present in the library, which is in reality a false hypothesis to start from!
As its name suggests, the *InLib probability* value indicates the probability that the searched compound is present in the searched libraries and is meant as a guidepost. Generally, any positive value is acceptable. Values greater than approximately 300 usually mean that the spectrum is nearly unique. Negative values below 200 are generally a warning that the spectrum is not identified. Note that negative values will occur when there are many compounds with similar spectra. In these cases, the difference between the Match Factors for different spectra is very small, and the search cannot be assured of providing the correct, unique answer. Usually in these cases, especially when *Match Factors* are high, this test will provide very good guidance on the structure of the molecule [12].

Interpretation of mass spectral library search results:

Conjoint evaluation of all the above descriptors, with a visual comparison of the unknown and library spectra of candidates in the ranked hit list, is essential in putting forward a tentative identification for the unknown spectrum. Evaluation of the *(pseudo)* molecular ion *(if detectable)* and important characteristic fragment ions in both unknown and reference spectrum should minimally be part of the visual spectral evaluation and corroborate the identification hypothesis.

The combination of high and nearly corresponding *Match* and *Reverse Match Factor* values, are synonym to high spectral similarity between the spectra and point out that one should take the identity of the corresponding hit into account. Such values alone, however, do not warrant the trueness of the identification. In the case of multiple hits with similar high *Match Factor* values, for example, low *Probability* values can be expected and a conclusive tentative identification cannot be inferred without taking additional information into consideration.

When the target spectrum would be a coelution of two compounds, the *Reverse Match Factor* can be high *(the reference spectrum is fully present in the target spectrum)* but the *Match Factor* value would be lower *(the target spectrum is only partially present in the reference spectrum).* In this case, it might be possible, by subtracting the mass spectrum of the first reference spectrum from the target spectrum, to perform a second search on the remaining spectrum to *(partially or tentatively)* identify the second *(co-eluting)* compound. When the target spectrum, however, is part of a reference spectrum *(e.g. sharing the same structural backbone),* the *Reverse Match Factor* will be low since the additional ions of the reference spectrum are absent in the target spectrum.

It is well recognized that mass spectral search is not failure proof as investigations [6] [8] have pointed out that, in general, the first ranked match can be a false identification in 20-25% of the cases. **A non-scrutinized assignment of the upper result(s) of a spectral search against the NIST or Wiley libraries, or both, as tentative identification for unknown spectra may obviously result in unreliable identifications**. To discriminate between true and false positive identifications, additional information such as the retention behaviour *(retention index, see Part 4: Additional Evidences Supporting Higher Level Identifications),* material related to prior information, or results from orthogonal methods performed on the sample should be taken into account.

BIBLIOGRAPHY

[1] S. Stein, "Chemical substructure identification by mass spectral library searching, " *Journal of the American Society for Mass Spectrometry,* vol. 6, pp. 644-655, 1995.

[2] V. Sica, K. Krivos, D. Kiehl, C. Pulliam, I. Henry and T. Baker, "The role of mass spectrometry and related techniques in the analysis of extractable and leachable chemicals," *Mass Spectrometry Reviews,* 2019.

[3] J. Watson and O. Sparkman, Introduction to mass spectrometry: instrumentation, applications and strategies for data interpretation, 4th ed., Chichester: John Wiley and Sons, 2007.

[4] R. Smith, Understanding mass spectra: A basic approach, 2nd ed., Hoboken, NJ: John Wiley and Sons, 2004.

[5] B. Milman, Chemical identification and its Quality Assurance, Heidelberg: Springer-Verlag, 2011.

[6] T. Kind and O. Fiehn, "Advances in structure elucidation of small molecules using mass spectrometry, "*Bioanalytical Reviews,* vol. 2, no. 1-4, pp. 23-60, 2010.

[7] S. Stein, "An integrated method for spectrum extraction and compound identification from gas chromatography / mass spectromtery data," *Journal of the American Society for Mass Spectrometry,* vol. 10, pp. 770-781, 1999.

[8] S. Stein and W. Wallace, *NIST Standard Reference Database 1A - User's Guide NIST/EPA/NIH Mass Spectral Library (NIST17) and NIS Mass Spectral Search Program (version 2.3),* Gaithersburg : U.S. Department of Commerce - National Institute of Standards and Technology, 2017.

[9] A. Moorthy, W. Wallace, A. Kearsley, D. Tchekhovskoi and S. Stein, "Combining fragment-ion and neutral-loss matching during mass spectral library searching: A new general purpose algorithm applicable to illicit drug identification," *Analytical Chemistry,* vol. 89, no. 24, pp. 13261-16268, 2017.

[10] T. Kind and F. O, "Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry," *BMC Bioinformatics,* vol. 8, p. 105, 2007.

[11] R. Hites and K. Jobs, "Is nontargeted screening reproducible?," *Environmental science and technology,* vol. 52, no. 21, pp. 11975-11976, 2018.

[12] S. Stein and D. Scott, "Optimization and testing of mass spectral library search algorithms for compound identification," *Journal of the American Society for Mass Spectrometry,* vol. 5, pp. 859-566, 1994.