

# Integrated Spectroscopy

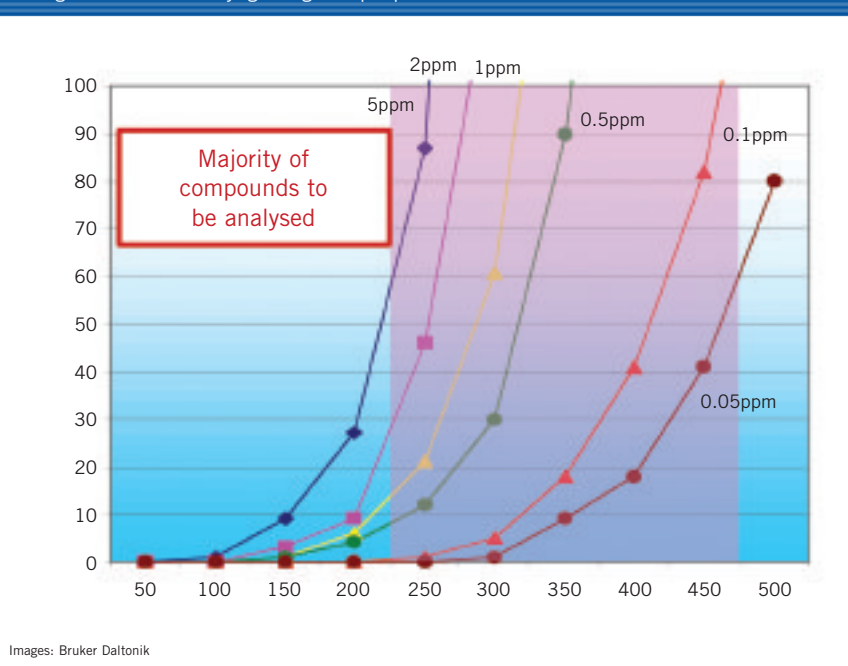
Ulrich Braumann at Bruker BioSpin and Herbert Thiele at Bruker Daltonik analyse structure verification of small molecules using integrated MS spectrometry and NMR spectroscopy

In service laboratories in the pharma industry today, chemists synthesise huge numbers of different compounds on the basis of parallel synthesis and combinatorial chemistry, which must be verified. One main aspect is the confirmation of chemical identity and information about molecular formulae to identify possibly present impurities. Most often, a screening approach is applied in order to obtain a rough estimation on the purity, concentration and identity of the synthesised products. This flags up samples below a certain purity or those with too many impurities or samples that are not what they are considered to be. It is often impossible to use all samples at the same time; therefore most of the samples are stored in libraries for later use. Undertaking quality control on a regular basis is mandatory in order to check for possible degradation reactions.

The main application areas for quality control and confirmation of molecular ID is synthetic chemistry (medicinal chemistry, core facility, organic chemistry and pharma NCE). Another broad field of application is the identification of small molecules, such as metabolite ID, natural products and pharma impurity.

All of these issues are handled using different spectroscopic techniques (such as NMR, LC-MS and X-ray) and by manual inspection of all the spectra types generated. This activity demands an expensive reservoir of human experts with vast knowledge in spectral interpretation; a talent that is becoming very rare nowadays. It also is very cost-intensive, which is why the quality assurance of larger libraries was often neglected in the past, and only rudimentary analysis was performed on select samples. An automated approach to structure verification based on the integration of different information rich spectroscopic methods should be the method of choice.

Figure 1: Dependency of the number of compounds obtained for a given mass accuracy ignoring isotopic peak ratios



## COMPLETE MOLECULAR CONFIDENCE

Complete molecular confidence (CMC) is a concept for a fully-integrated NMR/LC-MS based solution optimally complemented by X-ray spectroscopy, supporting molecular formulae

determination and automated structure verification for small molecules and natural products.

This new concept explores the synergies of the major analytical techniques LC-MS, NMR and also X-ray. CMC incorporates

## Maximum certainty in small molecule identification requires cutting-edge performance from the MS instrument: A resolution of, typically, 15,000-20,000 at a high acquisition speed of 20 spectra/s is mandatory to cope with ultra fast chromatography systems.

complementary analytical techniques: mass spectrometry for molecular formula determination and nuclear magnetic resonance for structure verification and elucidation. CMC assists the analysis on data combined from the information-rich spectroscopic techniques and will finally output the molecular profile in the form of a compact report. This details metrics for the quality of the fit of the molecular formula and structure, as well as any other sum formula/structure candidates. It delivers a probability for the verification of the proposed molecular structure, approximate purity information and quantity of the sample.

### FROM MASS SPECTRUM TO MOLECULAR FORMULA

This signals a new dimension in compound verification. Molecular formulas can be calculated directly from the MS spectrum. Because mass spectra do not automatically convey elemental information, data analysis tools are necessary to extract the information inherent in MS spectra and provide molecular formula candidates. The software program typically combines measured mass information and expected mass accuracy with valence and electronic configurations to produce a list of potential candidates near the measured mass. Mass accuracy is an essential parameter to limit the number of potential candidates. Figure 1 (page 58) shows the plot of the number of possible sum formulae generated by a generic formula generator using maximum composition (C200H400N50O50S5P5) against the compound mass.

Due to the high number of possible combinations, additional constraints for formula generation are needed to restrict the number of solutions. Basic chemical knowledge can supply boundary constraints for formula generation. Some of the constraints can be derived directly

from 1D-NMR measurements. The integral of non-overlapping signals can give information about the number (min/max) of aliphatic hydrogen atoms and the number of olefinic and/or aromatic double bonds.

### THE UNIQUE ESI-TOF TECHNOLOGY

Ideally, an instrument that can provide both high mass accuracy and stable, true isotopic pattern (TIP) information could provide greater information content. This technical concept allows for a two-dimensional analytical method: a combination of accurate mass determination with the analysis of the isotopic distribution. Combining both complementary sets of information is essential to find the correct formula for the elemental composition. Time-of-flight instruments are usually the best choice for molecular formula determination. This is especially true for Electrospray-Q-TOF-MS instruments (such as the micrOTOF-Q), which use linear ion counting to determine the known natural isotopic ratios.

Maximum certainty in small molecule identification requires cutting-edge performance from the MS instrument: A resolution of, typically, 15,000-20,000 at a high acquisition speed of 20 spectra/s is mandatory in order to cope with ultra fast chromatography systems. Mass resolution and mass accuracy have to be maintained in all scan modes and speeds – in MS as well as in MS/MS. The outstanding dynamic range is related to the fast repetition rate of the TOF (5,000-20,000 Hz) and the adequate analogue-to-digital-conversion (ADC) technique. This allows exact mass determination over the whole dynamic range as it is

not compromised by dead-time effects found in the more common time-to-digital-conversion (TDC).

The ADC technique combined with high-resolution TOF-MS is also important for the accuracy of the relative intensities of the isotopic peaks. This accuracy is required for the success of the true isotopic pattern (TIP) matching strategy. This can be a severe problem for TDC based TOF-MS, because the intensity of isotopes following high abundant isotopes is often reduced by the dead-time of the detector. In MS/MS spectra generated with the micrOTOF-Q II, the isotopic pattern information and the accuracy is also retained in the fragment ions. Sum formula proposals can be made for the fragment ions in the same way as for

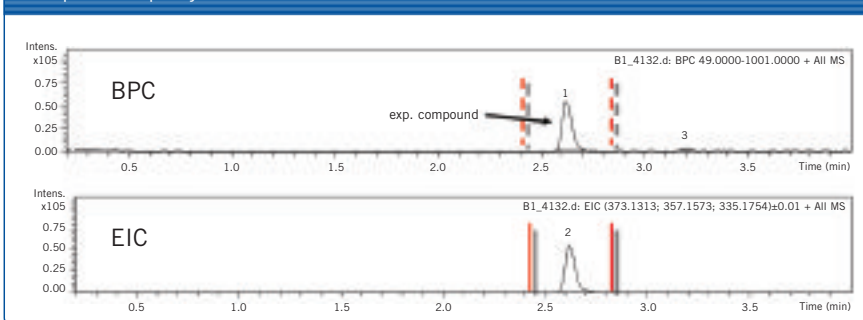
**Figure 2:** Automated LC-MS measurement with micrOTOF-Q II, for maximum certainty in small molecule identification



**Figure 3:** Generate molecular formula parameters

Charge	Tolerance	Sigma limit	Electron conf.	Nitrogen rule	Calibration
+1	8ppm	0.1	Even	True	True
<b>Expected formula:</b>		C21H22N2O2			<b>Adduct(s):</b> H, Na, K

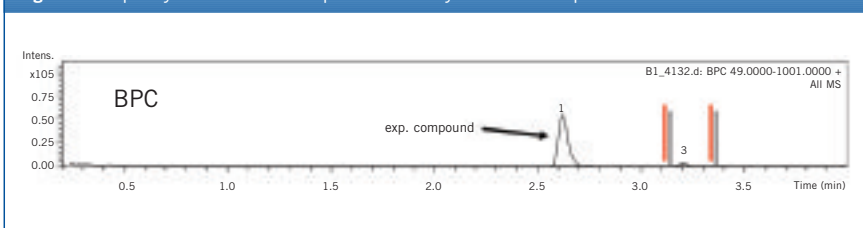
**Figure 4:** Peak matching between base peak and extracted ion chromatogram is required for purity calculation



**Figure 5:** Confirm formula results

#	Meas. m/z	Theo. m/z	Lerrl [ppm]	Sigma	Formula	Adduct	Purity (BPC)
2	335.1752	335.1754	0.6	0.010	C21 H23 N2 O2	M+H	95.8%

**Figure 6:** Impurity identification is performed only on unknown peaks



MS spectra, which adds a third level of confidence.

### ISOTOPIC PATTERN ANALYSIS – SCORING OF FORMULA CANDIDATES

Mass accuracy is not enough to reduce the number of possible hits in molecular formula generation. The SmartFormula approach considers the isotopic pattern distribution for MS spectra. After generation of a list of all possible formulae for a selected mass of an LC-MS peak, the measured isotopic pattern is compared with the theoretical isotopic pattern – resulting in a statistical match factor, the sigma factor ( $\sigma$ ). The sigma factor reported is simply the statistical variance between the measured and theoretical isotopic profile based on the intensity values of the peaks in the pattern. This comparison is done for all the generated molecular formulae. The sigma value is then used to put the formula candidates in order of rank.

### ENHANCED PROBABILITY-BASED CONCEPT

An isotopic pattern is described by three characteristic properties: the mass position of the peaks in the pattern; the peak intensities; and the peak distances within the pattern. Because of the precise isotopic patterns delivered by the MS instrument, it is possible to combine the mass position, the distance between the isotope peaks and their intensity into an integrated scoring with higher confidence for the individual hits. This procedure results in a scoring value which can be used to reduce the list of hypothetical formulas based on a true quality criterion, reflecting the quality of the matching property of the true isotopic profile for the individual molecular formula.

In many cases there are several molecular formulae with well-matched properties. It is not sufficient to consider only one hit with the best scoring factor; there are a few hits satisfying the overall criteria. The

overall ranking of the formulae candidates has to be extended into a probability-based scoring concept, modelling the distributions of mass accuracy and true isotopic pattern matching for a number of possible candidates. Considering all of the generated formula candidates using a Bayesian statistical modelling of the deviations, a score value with a range of 0 to 100 can be derived.

### PRECISION IN FORMULA GENERATION: TRUE ISOTOPIC PATTERN OF FRAGMENTS

Several techniques have been developed which use the information from MS/MS spectra as an additional criterion for reducing the number of possible formulae for the precursor ion. These methods sum up the potential formulae for product ion and the neutral loss to establish the identity of the precursor ion.

In contrast to these algorithms a new approach utilises accurate measured mass and additionally accurate measured isotopic pattern (see note). It could generate a confident list of formulae simultaneously for the precursor ion and all fragment ions. In contrast to the already known algorithms, both candidate lists are already drastically pruned, thus reducing the time to evaluate all possible relationships between the potential precursor formulae and the related product ions.

### FROM MOLECULAR FORMULA DETERMINATION TO AUTOMATED STRUCTURE VERIFICATION

A typical task for an analytical service laboratory in the pharmaceutical industry is to answer three typical questions:

- Did I really synthesise the correct compound?
- What is the compound purity level?
- What are the impurities?

These can be answered using an automatic compound verification concept. Using this technique, the chemist only needs to provide the molecular formula of the anticipated compound. As Figure 3 shows, the expected formula is C21H22N2O2 (Strychnine) allowing for the adducts H, Na and K.

**Figure 7: Impurity formula results**

FormulaMin: C1 H0		FormulaMax: C100 H200 N50 O50 Cl10 S10 Na K			
#	Meas. m/z	Theo. m/z	Lerrl [ppm]	Sigma	Formula
3	229.1408	229.1410	1.2	0.008	C10 H22 Na O4
		229.1408	0	0.011	C8 H17 N6 O2

## MS-BASED COMPOUND CONFIRMATION

### Step 1: Confirm Formula

The extracted ion chromatogram (EIC) is created for the masses of the expected formula plus possible adducts as defined in the 'operation method'. A chromatographic peak finder determines the peak boundaries and an averaged MS spectrum is generated over the whole peak.

The expected compound is verified by the algorithm mentioned above. As a result, the parameters  $\text{err}[\text{ppm}]$  (that is, the mass deviation) and  $\text{sigma}$  (the fit of the theoretical and measured isotopic pattern) indicate the quality of the compound confirmation process. As a rule of thumb, the parameter  $\text{err}[\text{ppm}]$  should be less than 2ppm and the parameter  $\text{sigma}$  ( $\sigma$ ) should be less than 0.02 to indicate a good confirmation of the expected molecular formula. If present, common adducts such as Na and K are listed – adducts are confirmed and reported as well.

### Step 2: Purity Calculation

In the next step the chromatographic peak purity is calculated. Therefore we have to calculate the fraction of the peak area of the expected compound with regard to the total peak area. The expected compound peak in the base peak chromatogram (BPC) is determined by retention time matching with the corresponding EIC peak. For the shown example the calculated purity for the confirmed molecular formula  $\text{C}_{12}\text{H}_{23}\text{N}_2\text{O}_2$  is 95.8 per cent.

As neither MS nor UV(VIS) detectors are able to detect all chemical compounds, a purity calculation based on UV or BPC can only be partially correct. This is especially true if only one ionisation method is involved. In general, the UV(VIS) trace might yield additional information for structure verification compared to MS and NMR. However, this may be more important for a structure elucidation concept where there is

no *a priori* guess as to the detected compound.

### Step 3: Impurity Identification

In this step the algorithm operates for all peaks in the BPC except the peak assigned to the expected compound. For one or more of the most intensive MS peaks in the averaged spectrum over the chromatographic peaks, the model generates possible molecular formulae, which can be discriminated by the quality parameters of the algorithm [ $\text{ppm}/\sigma$ ]. Although mass accuracy is good for both formulae the ranking, based on the  $\text{sigma}$  value, returns the correct formula as the top hit (in this special case the sodium adduct of a common plasticiser).

LC-MS is a routine analysis technique that can quickly deliver precise and reliable information for the identification of compounds with high sensitivity. In addition, it directly detects the presence of atoms such as oxygen, sulphur, nitrogen, chlorine that are typically not the subject

of a direct NMR analysis due to their inherent low NMR sensitivity.

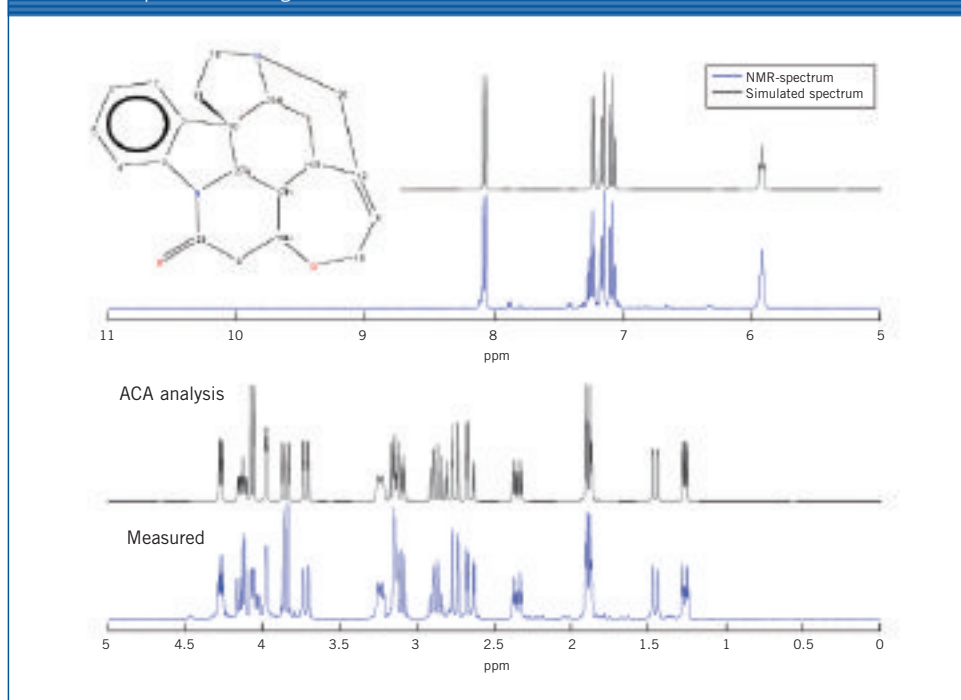
However, important aspects of the task of CMC are not covered due to the intrinsic properties of an LC-MS analysis. These include differentiation of isomers and the differences in ionisation efficiency and matrix effects that make any kind of quantification and thus purity control without specific reference compounds impossible. Here the NMR provides a complementary technique that generates a wealth of structural information. NMR is a fully quantitative method – its signal intensities correlate directly to the amount of protons of all components and thus yield the concentration as well as an estimation for the purity of a solution. The same principle can be used to determine the water content in DMSO samples that, for example, originate from typical liquid-store repositories in pharmaceutical industry.

## NMR MEASUREMENT FOR QUALITY CONTROL?

Traditional containers for NMR samples are, typically, 7"-long NMR tubes with 5mm diameter that need to be positioned in spinners which are then one-by-one put on a sample changer with a capacity in the

**Figure 8: Structure verification by NMR**

**Figure 9:** Comparison of the measured  $^1\text{H}$  NMR with the predicted and iterated spectrum of the given molecule



order of 100 samples. One solution is 100mm long NMR tubes in well plate format that not only allow for the simple filling and preparation of the NMR tubes by a commercial liquid handler, but also reduce the manual transfer into the NMR to one single action for a complete set of samples.

#### VERIFICATION BASED ON 1D AND 2D NMR DATA

The first step in the verification of a compound is the acquisition and evaluation of a  $^1\text{H}$  spectrum. For this analysis a  $^1\text{H}$  spectrum is predicted for the proposed structure based on the supplied MOL or SDF file using a chemical shift and spectral prediction tool. In the subsequent automatic consistency analysis (ACA), the predicted and experimentally acquired spectra are compared and, in an iterative process, the predicted spectrum is adapted to the experimental data. The likelihood that the experimental data correspond to the proposed structure is reported. This is based on the number of changes that were applied to the predicted

spectrum and the similarity of iterated and experimental spectrum with respect to peak position, integrals and coupling constants. Signals arising from the solvent are also taken into account.

This analysis also works for more complex spectra, which can be shown in the example of strychnine. However, the limited chemical shift range of  $^1\text{H}$  data for larger or more complex molecules can lead to an overlap of signals. Conclusions, therefore, become increasingly uncertain. For this purpose, typically, an HSCQ spectrum is acquired. This separates out the signals in a second dimension and, in addition, provides larger spectral dispersion through the  $^{13}\text{C}$  information. In contrast to traditional  $^{13}\text{C}$  spectra, an HSQC experiment produces quick access to proton carrying  $^{13}\text{C}$  resonances with low sample concentration, thus making it the experiment of choice for screening methods.

The analysis of the 2D file is also automated. The predicted ranges and experimental data are overlaid and the detected signals are assigned to the

predicted ranges. The result is a matching factor that describes the similarity of the experimental and theoretical data.

#### QUANTIFICATION AND PURITY

In contrast to LC-MS, the signal integral in an NMR spectrum correlates directly to the amount of detected compound as long as a few simple experimental parameters are respected. The only additional sample information required is the number of protons that contribute to the NMR signal used for the quantification.

With a proposed structure, this information is available through the predicted NMR spectra from the PERCH

routine. Here an assignment of signal and structure element is automatically provided. Even without a structure, the NMR spectrum by itself provides, in most cases, enough information to identify at least the structure element of one signal group. This is already sufficient for the quantification.

The CMC concept includes a solution package for computer-assisted compound quantification. This solution package includes all the necessary parameters for the setup, the acquisition, the processing and the analysis of the NMR spectra. It supports the use of samples of minute quantity stored in normal (that is, non-deuterated) DMSO, as it is most commonly used in compound libraries in pharma industry.

For a data interpretation, a human-logic-emulation based interpretation wizard is supplied, which analyses a 1D NMR spectrum in a similar way as a human spectroscopist would do it. This wizard automatically determines the water content in these typical liquid-store samples and assists the user on the data interpretation for

compound quantification with artificial intelligence. The wizard suggests proton numbers for certain signals in the spectrum and thus determines concentrations from parameter-free spectra interpretation without the need to know the chemical structure. The analysis-assistant's user interface allows one to browse through, check or modify these suggested results in a very simple way. At the end of such an analysis run, a comprehensive report is generated on the complete set of analysed compounds.

The instrument parameter required for the absolute quantification is one NMR signal integral of a known amount of one compound as an independent reference. In traditional analysis, this is accomplished with an internal reference that is added to each sample solution during preparation or simply during the transfer into the NMR tube. However, state-of-the-art software and NMR electronics allow, after initial calibration, the measured integrals of the NMR signals to be taken without a further reference signal as a measure for the sample amount. This initial calibration is also an integrated part of the above mentioned solution package. The

#### About the authors



Dr Ulrich Braumann studied chemistry at the University of Tübingen at the Institute of Organic Chemistry. He finished his studies at this university with his doctoral thesis in 1995. He started working at Bruker in 1995 in the field of LC-NMR where he is now responsible for application field flow-NMR as Product Manager. Email: [ulrich.braumann@bruker-biospin.de](mailto:ulrich.braumann@bruker-biospin.de)



Dr Herbert Thiele was educated at the Institute of Organic Chemistry, TH Aachen, where he gained his diploma and wrote his doctoral thesis. He started out in 1977 as a teacher of Chemistry and Mathematics at the Fachschule des Heeres für Technik/Aachen. He joined Bruker in 1978 and has held various management positions, as well as being made an Honorary Professor in Analytical Chemistry at the University of Bremen. In 2002 he was appointed Director of Bioinformatics. Email: [ht@bdal.de](mailto:ht@bdal.de)

only precondition is that each NMR experiment is correctly prepared (tuning/matching and 90° pulse determination), which is nevertheless possible in a fully automated system.

A byproduct of the identification and quantification is the purity. From the prediction of the NMR spectrum the integral regions of the NMR signal of the compound of interest are known. A relative comparison of this integral and the integral of the total NMR spectrum can be used as a measure for purity.

to the quantity of the impurities. However, as long as the number of protons contained in the main compound and in the observed impurities is in the same order or magnitude, this ratio provides a very good estimate of the purity level. It should be mentioned that the information required for the purity and quantification can be derived from the same spectrum already acquired for the identification of the compound. No further time is required in order to generate this information.

**Figure 10:** Molecular Profile™: all results are presented in the form of a standardised sample quality report



Of course the integral of the unwanted signals does not directly correlate

#### ADDITIONAL TECHNIQUES

These include a fully automated bench-top X-ray diffractometer for structure determination of small molecules. Single crystal crystallography provides accurate and precise measurements of molecular dimensions in a way that no other science can begin to approach. Small molecule crystallographers study compounds which are of chemical and biological interest – new synthetic chemicals, catalysts, pharmaceuticals, natural products and many more.

**Note:** SmartFormula 3D, Bruker