Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants

Nikiforos Alygizakis [1,2 §*], Francois Lestremau [3,4 §], Pablo Gago-Ferrero [5 §], Rubén Gil-Solsona[5], Katarzyna Arturi [6], Juliane Hollender [6,7], Emma L. Schymanski[8], Valeria Dulio[4], Jaroslav Slobodnik [2], Nikolaos S. Thomaidis [1*]

[1] Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Greece

[2] Environmental Institute, Okružná 784/42, Koš, 97241, Slovakia

[3] Hydrosciences Montpellier, Univ. Montpellier, IMT Mines Ales, IRD, CNRS, Ales, France

[4] Institut National de l'Environnement Industriel et des Risques (INERIS), Parc ALATA BP2, 60550, Verneuil en Halatte, France

[5] Institute of Environmental Assessment and Water Research (IDAEA) Severo Ochoa Excellence Center, Spanish Council for Scientific Research (CSIC), Jordi Girona 18-26, E-08034, Barcelona, Spain

[6] Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland

[7] Institute of Biogeochemistry and Pollutant Dynamics (IBP), ETH Zurich, 8092 Zurich, Switzerland

[8] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Avenue du Swing 6, L-4367 Belvaux, Luxembourg

*Corresponding authors

Nikiforos Alygizakis
Environmental Institute, Okružná 784/42,
972 41, Koš, Slovak Republic
E-mail: alygizakis@ei.sk
Phone: +421 919183401

Nikolaos S. Thomaidis
Laboratory of Analytical Chemistry, Department of Chemistry, National and Kapodistrian University of Athens, Panepistimiopolis
Zografou, 15771 Athens, Greece
E-mail: ntho@chem.uoa.gr
Phone: +30 210 727 4317

[§] These authors contributed equally to this work

For submission to: **TrAC Trends in Analytical Chemistry**

Highlights

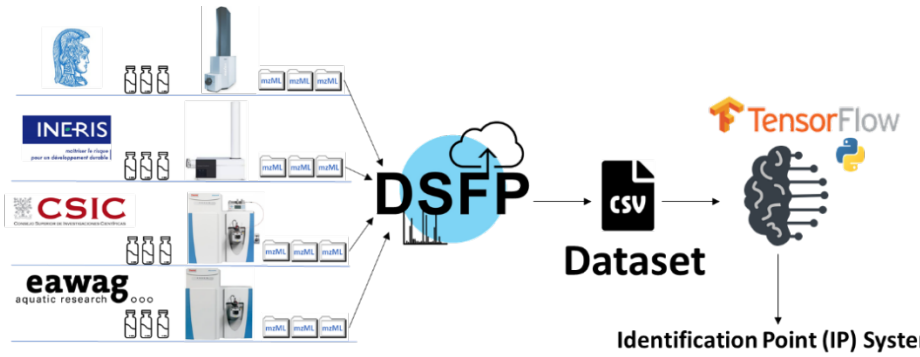
1. A model was developed to classify identifications as reliable and unreliable
2. Machine learning provided insight for the weights of the most informative parameters
3. Identification confidence was influenced mostly by fragmentation and isotopic fit
4. An identification point (IP) system scaled from 0 to 1 was proposed and applied
5. The IP system was connected with the widely used identification confidence levels

Abstract

Non-target screening (NTS) methods are rapidly gaining in popularity, empowering researchers to search for an ever-increasing number of chemicals. Given this possibility, communicating the confidence of identification in an automated, concise and unambiguous manner is becoming increasingly important. In this study, we compiled several pieces of evidence necessary for communicating NTS identification confidence and developed a machine learning approach for classification of the identifications as reliable and unreliable. The machine learning approach was trained using data generated by four laboratories equipped with different instrumentation. The model discarded substances with insufficient identification evidence efficiently, while revealing the relevance of different parameters for identification. Based on these results, a harmonized IP-based system is proposed. This new NTS-oriented system is compatible with the currently widely used five level system. It increases the precision in reporting and the reproducibility of current approaches via the inclusion of evidence scores, while being suitable for automation.

Keywords: Identification point (IP) system, suspect screening, non-target screening, communication of identification confidence, retrospective screening, high-resolution mass spectrometry

Graphical abstract



Identification Levels	IP Score
Level 1	>0.75-1.00
Level 2	>0.60-0.75
Level 3	0.50-0.60
Level 4	0.20-<0.50
Level 5	0.00-0.20

1. Introduction

The global universe of chemicals is very complex and includes hundreds of thousands of substances in commercial use **[1-3].** In recent years, advances in high resolution mass spectrometry (HRMS) have revolutionized our ability to measure organic chemicals in a wide variety of matrices, expanding the analytical window and rapidly increasing the popularity of suspect and non-target analysis (NTS) **[4, 5]**. These approaches are currently widely used for the tentative identification of a large and still increasing number of potential contaminants, especially polar and semi-polar ones, as well as many endogenous compounds in different organisms **[6, 7]**. Chemical studies often result in large lists of tentatively identified substances **[8, 9]**. This has created the need to communicate the confidence in the identification in a way that reflects all the evidence available **[10]**. This is essential for a consistent advancement in the fields that rely on the analysis of organic substances at trace level, including environmental chemistry **[11]**.

Currently, in the last step of a target or suspect HRMS screening, the analyst is obliged to spend a significant amount of time evaluating all proposed identifications case by case **[1, 12]**. The analyst relies on orthogonal analytical evidence (chromatographic retention behavior, isotopic profile, MS fragments, among others) and other additional metadata (e.g., number of patents, literature references) **[13, 14]**. Nevertheless, in the end, expert judgement is required to assign the given identifications a certain level of confidence. This manual evaluation is time-consuming and lacks reproducibility, while the time required is increasingly moving beyond the realms of manual efforts due to the sheer numbers of screened compounds and samples **[12, 15]**. So far, most environmental studies report the confidence based on hierarchical degrees of confidence **[10]**, ranging from Level 5 (exact mass), Level 4 (unequivocal molecular formula), Level 3 (tentative structure), Level 2a and 2b (probable structure) through to Level 1 (confirmed identification). In many cases, while the aforementioned levels are certainly useful (as is evident from their widespread and increasing adoption), it is still difficult to communicate the evidence associated with the assigned identification confidence level in a concise and unambiguous manner. Early attempts to include identification evidence via identification points (IPs) described in the Commission Decision 2002/657/EC were already implemented in the first NORMAN Collaborative Trial on non-target screening in 2013/14 **[16]**. Recently, this approach was also applied to communicate the confidence in the identification of analytes for target analysis **[17]**. This IP system considers retention time, mass accuracy, isotopic fit and fragmentation, taking advantage of the capacities of the HRMS instruments, but it is not yet explicitly implemented as a standard for non-target screening (NTS) **[16, 18]**. Other recent efforts include the integration of automated level system functionality in patRoon – where users can adjust the requirements **[19]** and specific guidance released by the per- and polyfluoroalkyl substance (PFAS) community **[11].** A complementary system that allows the community to understand the identification evidence associated with a reported compound identification in a rapid, concise and reproducible manner is necessary. A system based upon identification points (IPs) and thus compatible between target and non-targeted approaches would be a valuable addition to the field.

There is an urgent need to automate the evaluation process and create a more reproducible and harmonized approach **[20]**, due to the number of chemicals (or features; hereafter "chemicals" for the purpose of this manuscript) involved in NTS. Machine learning models are well suited to these tasks. Ideally, such a model should produce a score to assist in the reporting, limiting the amount of manual

work required by the analyst, but present sufficient information to enable quick and efficient manual quality control. This allows a focus of efforts on the most challenging cases of greatest importance to the study outcomes. One of the drawbacks of this approach is that machine learning models must be trained individually for each instrument and analytical strategy used by the laboratories for optimal performance. The large variety of instruments and data acquisition methods further complicates the situation and highlights the need for harmonization of data treatment **[21]**. To create such informative machine learning models, it is critical to identify the most informative parameters using domain knowledge. Once such models are built, these provide deeper insights into the importance of the parameters involved and can eventually be used to propose an easy-to-follow generic IP system, automatable and applicable under any instrumental and data acquisition conditions.

This article takes a close look at the challenges in harmonizing the NTS identifications, focusing on liquid chromatography mass spectrometry (HRMS/MS). An interpretable machine learning approach for classification of NTS identification confidence was developed, capable of automatically discarding substances with insufficient evidence for reliable identification. The described approach can be implemented by any laboratory performing NTS analysis. It provides clear benefits in terms of accurately describing the evidence associated with identified substances. Moreover, it progresses towards the development of automatic prioritization schemes for the management of chemicals. An IP-based system is proposed for the communication of evidence accompanying identification confidence based on the results obtained here, the insights gained by this exercise and the participation in NORMAN NTS collaborative trials e.g. **[16, 22]** and other ongoing trials. While developed on LC-ESI-MS/MS, it is applicable to any soft ionization technique (e.g., GC-APCI-HRMS/MS and GC-CI-HRMS/MS), given that they produce the molecular ion and considerably less fragment ions. This new NTS-oriented system is compatible and comparable with target analysis and adds more precision and reproducibility to current approaches, while being suitable for automation – a key necessity required for high throughput NTS screening.

2. Parameters/Evidence used for NTS identification

NTS identification of polar and semi-polar organic chemicals is based on the available information, commonly generated by LC-HRMS/MS systems. Several pieces of evidence provide information about the identity of a compound. However, not all are equally relevant or even available in all cases. While some information is critical and always available (e.g., mass accuracy), other information increases the degree of confidence to a lesser extent and are not as essential. Likewise, not all pieces of evidence lead to a concise measurable parameter that can be directly transformed into IPs.

This section describes the parameters that should be considered in an objective, concise and potentially automatized IP-based system and discusses their possible role in the harmonization of NTS identifications as well as their automation potential. The parameters are divided into those that should be considered by any consistent IP-based system and others that would add additional confidence but where the implementation is more challenging.

2.1 Essential Parameters/Evidence for NTS Identification Confidence:

1. **Mass accuracy:** The accurate mass of an ion is the mass experimentally determined (and recalibrated with a reference mass standard if applicable) in the mass spectrometer**.**
   This is the parameter upon which HRMS identifications rely and is the starting point in any identification, either to match a target, check the potential presence of a suspect, to perform exact mass searches, or to assign molecular formulas in non-target studies. The parameter mass deviation between the measured (accurate) and theoretical (exact) masses should be below the acceptable threshold according to the instrument manufacturer (for most of the instruments <5 ppm at *m/z* 200 and/or <2 mDa; modern instruments or internal calibration can achieve < 2 ppm) and should be verified with regular calibration. The confidence increases with lower mass deviation.

2. **Retention Time (RT) information:** Retention time plausibility is a requirement to reach a certain identification confidence. Many RT prediction models have been developed in the literature and have proven to improve suspect and non-target screening **[23-25]**. There is an increasing need for comparable and harmonized RT in LC-HRMS/MS among different laboratories. In this regard, flexible and system independent unified retention time indices (RTI) can help improve the automation of NTS approaches by reducing the number of false positives in a first screening step. For GC-(HR)MS, the n-alkanes mixture is most commonly used for retention indexing and calculation of the Kovat's index **[26]**, which is the established protocol in the NIST mass spectral library. For LC-MS, one such RTI method is based on carefully selected calibrants that can be easily used and applied under any liquid chromatographic conditions **[27]**.

3. **Isotopic fit**: The isotopic pattern that forms in the mass spectrum by the separation of the various isotopes of the atoms present in a molecule is used to increase the confidence in the element and molecular formula assignment. Although it is certainly a useful parameter (especially for halogenated molecules and other molecules with distinct isotopic patterns), in many cases when working at trace levels the intensity of the isotopic peaks is so low that it cannot be observed or can deviate substantially from the theoretical pattern. Therefore, a less accurate isotopic fit for low intensity masses should not be used as a strong argument to discard candidates during identification. It is quite frequent phenomenon that the lack of isotopic fit results in false non-detections, impacting drastically automated evaluations. Isotopic patterns can also be used to recognize the presence of certain elements, such that this information can be used without necessarily strictly restricting the identification efforts to a specific molecular formula.
   In the evaluation of isotopic fit, it is important to consider the importance of the isolation window in data dependent data: If it is above 1 Da, isotopic peaks can appear in the MS/MS, which can be helpful to identify heteroatoms, but may result in unwanted interferences in the spectrum. Wide isolation windows can be beneficial for matrix-free samples such as drinking water. However, a conservative choice of isolation window below 1 Da is preferable for more complex samples such as biological or wastewater samples, which suffer from matrix interferences.

4. **Number of fragments ions / Presence of qualifier fragment ions:** Compound identification requires the measurement of MS/MS spectra for individually selected precursors [data

dependent acquisition (DDA)] or simultaneously for all precursor ions (data independent acquisition (DIA)). The number of fragments constitutes critical information for the reliability of a given identification. However, not all fragments provide the same level of diagnostic information, as some fragments are very common to many chemicals, while others are very specific to only a certain chemical or class of chemicals. The absence of a qualifier fragment ion for a given chemical (e.g 68.9958 corresponding to $-CF_3$ for perfluorinated compounds) can be an exclusion criterion. Other more common fragments (such as 77.95736, for $[SO_3]^-$, 95.960697 for $[HPO_4]^+$ or a low mass CHON fragment) are less informative and should have less influence on the degree of confidence of the identification. An important aspect is that low mass fragments can have high variations in mass accuracy due to being at the lower end of instrument detection ranges. Establishing a cut off for a minimum number of matching fragments can help automation. For example, cases where less than two experimental fragments are detected can be automatically flagged. In this manner a binary variable (TRUE, FALSE) can be obtained. Then, the analyst should be cautious with the identification and manual inspection may be required. Three main aspects must be evaluated: the fragmentation potential (total number of fragments), number of relevant fragments, and presence/absence of those. It is worth considering detected fragments between different chromatographic runs within the same batch. Chemicals detected with high intensity in a chromatogram will often exhibit a clearer fragmentation pattern (including a higher number of fragments and consistent ratios between them) than the same substances detected in lower intensity in other chromatograms within a batch. Fragments that match those present in spectral libraries obtained in an experimental manner (e.g. MassBank **[28]**, MoNA **[29]**, mzCloud **[30]**) provide more confidence than those predicted *in silico*. It is worth noting that there are many different *in silico* prediction tools such as CSI:FingerID **[31]**, CFM-ID **[32]**, MetFrag **[33]**, MAGMa **[34]** and other approaches, the performance of which has not been thoroughly analyzed within DSFP.

5. **Presence of MS/MS spectra from DDA**: Different acquisition modes provide different degrees of confidence in fragment ion assignment. DDA data increases the confidence of the assigned fragments since the chances that they are generated from the parent compound are higher. Therefore, those fragments should provide more IPs than those obtained with DIA.

6. **Presence of heteroatoms in fragments (if available) and plausibility of their molecular formulas:** It is important to assess the molecular formula assignment of the fragments, which should agree with the formula of the compound. The presence of heteroatoms in each structure facilitates its identification. The presence of these heteroatoms in the associated fragment ions (many times even with a distinctive isotopic pattern if the isolation window is >1 Da) provides important evidence. Despite the ongoing efforts, HRMS libraries with appropriate molecular formula annotations for fragments have not been widely implemented. While the situation is improving, improving the automatic extractability of such information would greatly facilitate automated interpretation.

2.2 Additional Parameters/Evidence for NTS Identification Confidence:

7. **Presence of adduct ions**: The presence of related adduct ions, although not always available, can help increase the certainty of the neutral exact mass calculated from the precursor ion. Therefore, the detection of adducts can help to avoid focusing on neutral masses calculated from the incorrect adduct (*e.g.*, incorrect assumption of $[M+H]^+$ for a $[M+NH_4]^+$ signal) or *in source* fragments, both of which are common for example in electrospray ionization. There are many clustering approaches such as nontarget **[35]** and RAMClustR **[36]** among others, that can help with automation.

8. **Fragment ratio at least between quantifier and qualifier ions**:
The ratios between the detected MS/MS fragments for a given chemical in LC-HRMS/MS analysis should remain constant (within a given tolerance) for the same/ equivalent collision energy, in an analogous manner the ratio of intensities between transitions used in quantification via selected reaction monitoring mode (SRM). The evaluation of these ratios can significantly increase the degree of confidence of the identifications in ambiguous situations. The variation of the fragmentation ratio under different collision energies can also be informative. Unlike GC-MS libraries, the lack of standardization of the collision energy of the LC-HRMS libraries prevents the automatization of the fragment ratio at this stage.

9. **Mass of fragments:** Fragment ions with higher mass can provide more specific structural information than lower mass fragments. Fragments with lower masses suffer from more interference, particularly when high collision energies are used. This weighting approach has been applied successfully by the software of NIST. Low mass fragments also tend to represent common substructures present in many structures. While this provides some structural evidence, this can apply to many possible candidates.

10. **Additional dimensions to the data:** The dimension of the available data can be increased by the addition of separation methods. In this category, one of the most promising developments is ion mobility separation (IMS). IMS separates ionized compounds based on their charge, shape and size, facilitating the removal of co-eluting isomeric/isobaric species **[37]**. Therefore, it helps to obtain cleaner mass spectra (facilitating data interpretation), while also providing information about the collision cross section of the molecule, thus providing additional evidence. The drift times provided by IMS are expressed as collision cross-section (CCS) values and may further contribute to delineating database hits and confirming structure identification. CCS is a robust measurement suitable for use as an additional parameter in NTS identification, where available. Its importance will increase as the number of instruments with IMS on the market increase and becomes available to the laboratories, along with efforts to include CCS values in open resources **[37, 38]**. Other efforts to increase the information available for identification include the use of different chromatographies, ionizations and even sample preparation methods but their detailed explanation goes beyond the objective of this study.

3. Automated allocation of identification evidence using machine learning

3.1 Implementation of Parameters

The essential parameters for NTS identification confidence (Section 2.1) were used to build classifiers able to differentiate between the availability of sufficient or insufficient evidence for confident identification. To achieve this, the batch screening functionality of NORMAN Digital Sample Freezing Platform (DSFP) **[20]** was upgraded to output the following scores:

1) mass accuracy (mz_{score}),
2) RT index information (RTI_{score}),
3) isotopic fit ($IsoFit_{score}$),
4) number of fragments ions considering both DIA and DDA ($Fragment_{score}$),
5) presence of MS/MS spectra from DDA as a TRUE/FALSE variable (DDA_{score}),
6) fit of molecular formula of fragments ($FitMolForm_{score}$) and
7) spectral similarity ($SpecSimil_{score}$).

mz_{score}, RTI_{score} and $Fragment_{score}$ compare experimentally measured values (*exp*) with theoretically calculated (*theor)* or predicted (*pred*) values and are given from the equations presented in **Table 1**.

Table 1*. Equations for the calculation of mz_{score}, RTI_{score} and $Fragment_{score}$. The subscript abbreviation exp indicates experimental value, theor indicates theoretical value, pred indicates predicted value.*

Equation	Equation number
$\mathrm{mz_{score}} = 1 - \frac{\mathrm{abs}(\mathrm{mz_{exp}} - \mathrm{mz_{theor}})}{\min(\mathrm{mz}_{exp}, \mathrm{mz}_{theor})} * \frac{10^6}{tolerated\ accuracy\ in\ ppm}$	eq. 1
$\mathrm{RTI_{score}} = 1 - \frac{\mathrm{abs}(\mathrm{RTI_{exp}} - \mathrm{RTI_{pred}})}{1000}$	eq. 2
$\mathrm{Fragment_{score}} = \frac{\mathrm{number\_of\_uniques(matched\ fragment\ ions\ in\ DIA\ \cup\ matched\ fragments\ in\ DDA)}}{\mathrm{total\ number\ of\ fragments\ in\ the\ library}}$	eq. 3

The $IsoFit_{score}$ and $FitMolForm_{score}$ were defined based on MOLGEN-MS/MS **[39, 40].** DDA_{score} is a binary variable indicating whether data-dependent HRMS/MS scan is available. $SpecSimil_{score}$ was calculated based on OrgMassSpecR package **[41]**. Where experimental HRMS/MS is not available, $SpecSimil_{score}$ =0. If an experimental mass spectrum is not available (e.g., because there is no record in MassBank), the match with the CFM-ID (v. 4.0) *in-silico* predicted mass spectrum is considered **[32]**. All scores range from 0 to 1.

3.2 Experimental / Measurement data

Measurements from four organizations (the National and Kapodistrian University of Athens (UoA), the French National Institute for Industrial Environment and Risks (INERIS), the Institute of Environmental Assessment and Water Research (IDAEA-CSIC)) and the Swiss Federal Institute of Aquatic Science

Technology (Eawag) were used to generate the dataset used here. The organization performed analysis using the following HRMS instruments: the quadrupole time of flight (Q-TOF) mass analyser maXis Impact by Bruker, the 6550 iFunnel Q-TOF by Agilent Technologies, the Q-Exactive™ Orbitrap and Q-Exactive™ Plus Orbitrap by Thermo Fischer Scientific, respectively.

The dataset of UoA included 18 mixtures of substances, containing in total 383 individual reference standards at final concentration 50 ng mL^{-1}. The mixtures were organized based on the chemical class of the substances (e.g., separate mixtures of pesticides, pharmaceuticals, industrial chemicals etc.). These mixtures were injected on an Acclaim™ RSLC C18 column (2.1x100 mm, 2.2 µm; Thermo Fischer Scientific) coupled to a LC-ESI-QTOF from Bruker using DIA and DDA (5-most abundant precursors per scan) according to instrumental settings presented in detail elsewhere **[17].**

The dataset of INERIS included in total 91 pesticides, which were prepared at concentrations of 1, 10 and 50 ng mL^{-1}. The reference standards were organized in four different mixtures. The mixtures were separated by a ZORBAX® SB-Aq (1.8 um, 2.1x150 mm; Agilent Technologies) column and were detected by an Agilent 6550 iFunnel QTOF. The samples were analysed using DIA acquisition according to instrumental settings presented in detail elsewhere **[42]**.

The dataset of IDAEA-CSIC contained 21 pesticides in one mix, 83 compounds of various classes in another mix and 129 compounds of various classes in another mix (all at concentration 50 ng mL^{-1}). The samples were separated using a Cortecs C18 column (2.1x100 mm, 2.7µm; Waters), preceded by a guard column of the same packaging material and were detected using a Q-Exactive™ Orbitrap mass analyser (Thermo Fisher Scientific). Instrumental details can be found in the respective publications **[43, 44]**.

The dataset of Eawag was created using groundwater samples spiked with in total 519 compounds at two concentration levels (10 and 100 ng L^{-1})**.** Separation was achieved on an Atlantis® T3 column (3 µm, 3.0 x 150 mm; Waters) and the detection on a Q-Exactive™ Plus Orbitrap mass analyser (Thermo Fisher Scientific) with electrospray ionization. The samples were analysed using DDA acquisition according to instrumental setup described elsewhere **[45]**.

Detailed information on the instrumental setups and acquisitions can be found in **Table S1**.

3.3 Establishment of the Machine Learning Model

3.3.1 Dataset generation

The data of all participants was uploaded to the NORMAN DSFP using the established contribution procedure and was screened using the batch-mode utility **[20]**. The NTS workflow has been validated and explained in detail elsewhere **[20]**. Briefly, the workflow uses the centWave algorithm for peak picking **[46]** with previously optimized ppm and peakwidth parameters through the IPO R-package **[47]**. Optimized peak-picking parameters can be found in **Table S2**. The peak picking workflow searches for consecutive masses within a mass error threshold forming peak shape in chromatographic dimension. The next step is componentization, which is a procedure for grouping peaks coming from the same compound (e.g., adducts, isotopic peaks). Componentization is accomplished with the nontarget R package **[35]**.

The aim of the screening was to generate a dataset with examples of successful and unsuccessful identifications. Here, unsuccessful identifications originate from the pick-up of signals in samples with acceptable mass accuracy and plausible retention time index. The generated dataset included in total 1424 instances (rows) after the exclusion of substances (< 1%) that were not detected in the chromatographic data due to analytical reasons (either low concentration or insufficient sensitivity). The detected substances were accompanied with the individual scores from categories 1 to 7 (described previously in section 3.1). The generated dataset is provided in the **supplementary excel file.** The column "Spiked" is the label (response variable) and indicates whether a compound was spiked in the samples or not.

3.3.2 Machine learning

This dataset was used to create the following classifiers: decision tree (DT), support vector machine (SVM), logistic regression (LR), gaussian Naive Bayes (NB), random forest (RF), k-nearest neighbors (kNN). More complex ensemble methods (e.g., XGBoost) were not used for modeling. Modeling was performed using the scikit-learn python package **[48]**. The script and calculations are available at https://github.com/nalygizakis/IPscore.

The performance of the classifiers was tested using 10-fold cross validation and default parameters **[48]**. RF outperformed the other classification models for this specific modeling task (**Figure 1a**). Given that the training and evaluation sets were unbalanced (not equal instances per class), the overall macro-averaged F1 score was used as the evaluation metric of the accuracy. The macro-averaged F1 score is calculated by taking the arithmetic mean of all the per-class F1 scores. The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean. Satisfactory accuracy was achieved for kNN and SVM, whereas similar but lower F1 score was observed for DT, LR and NB.

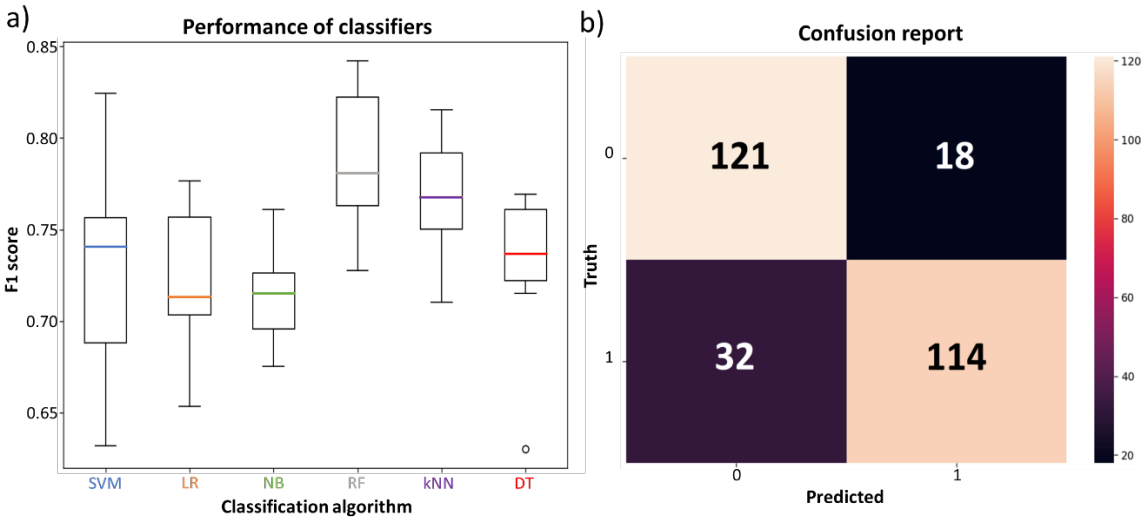


***Figure 1a**. Performance of various classification models using 10-fold cross-validation. Abbreviations: support vector machine (SVM), logistic regression (LR), gaussian Naive Bayes (NB), random forest (RF), k-nearest neighbors (kNN), and decision tree classifier (DT), **1b**. Confusion report for the optimized random forest model in the training set. The model yielded accuracy 79.2%. In total, 235 instances were classified correctly (121+114) and 50 instances were classified incorrectly.*

RF was selected for further optimization of the hyperparameters, as it showed the best performance. The following parameter grid was investigated:

- Number of estimators: 40 values from linear space 10 to 1000
- Maximum depth: 40 values from linear space 2 to 50
- Minimum samples split: 20 values from linear space 1 to 50
- Minimum samples leaf: 20 values from linear space 1 to 50
- Bootstrap: parameters: 'True' and 'False'
- Maximum features: parameters: 'auto', 'log2', 'sqrt'

After a 1-hour, six-core experiment on an Intel® Core i9-10885H CPU, the optimized parameters were: 873 for number of estimators, 50 for maximum depth, 3 for minimum samples split, 3 for minimum samples leaf, 'True' for bootstrap and 'log2' for maximum features. The optimized RF model after hyperparameter tuning provided accuracy of 79.2% in the test set (**Figure 1b**). In total, 235 instances/compounds were classified correctly (121+114) and 50 instances/compounds were classified incorrectly.

3.3.3 Importance of parameters

The parameter importance ranking of the optimized RF model is presented in **Table 2**. As shown in Table 2, $Fragment_{score}$ proved to be the most decisive parameter for the discrimination of the identifications. It is important to note that $Fragment_{score}$ considers the number of unique fragments detected in both DDA and DIA (where both are available). One reason mass accuracy was not ranked high was that it is also used indirectly in the parameter $Fragment_{score}$. Moreover, the way that negative hits were defined diminishes the possible importance of mz_{score} and to a lesser extent RTI_{score}. mz_{score} proved less important because exact masses are not unique parameters and the negative hits used in the study are per definition within the defined mass tolerance. Since the fragments capture additional complementary information, they ended up with higher relevance and this made mz_{score} alone less relevant. Finally, DDA_{score} proved to be highly correlated with FitMolFormscore (r=0.75) thus it was excluded from the evaluation.

Results from the machine learning approach showed that the number and the quality of the fragments are the important parameters for a reliable identification. Isotopic fit also proved to play an important role. RTI, mass accuracy and spectral similarity scores were ranked lower, but provided additional meaningful information for the classifier. Based on the outcomes of the implemented approach and the insights gained by the exercise, the next section details a simplified IP-based system for the communication of identification confidence.

***Table 2**. Parameter importance of the optimized RF model. The scores* $Fragment_{score}$, $FitMolForm_{score}$ *and* $SpecSimil_{score}$ *transfer the spectral information (purple background).* $IsoFit_{score}$, RTI_{score}, *and* mz_{score} *were colored with green, yellow and orange background, respectively. These colors were applied to all graphical elements.*

Score	Importance of parameters
$Fragment_{score}$	0.225
$IsoFit_{score}$	0.209
$FitMolForm_{score}$	0.173
RTI_{score}	0.162
mz_{score}	0.141
$SpecSimil_{score}$	0.090

The IP Score system proved helpful. However, it is difficult to be implemented for every laboratory, since it is unreasonable to expect all laboratories to establish their own machine learning-based system. Furthermore, In order to bring non-target screening at regulatory level, there is a clear need for the generation of a harmonized identification scoring. This identification scoring system must allow communication of the identification confidence in an automated, concise and unambiguous manner that reflects all the available evidence. Reproducibility and transparency in confidence communication will open up possibilities to develop novel prioritization schemes for the management of chemicals. Therefore, the machine learning approach was used as the basis for the proposal of the IP system described in section 4. The IP system is based on a combination of the results gained within this exercise, intuition and common knowledge, which may be difficult to implement with machine learning.

4. Proposed Identification Points (IP) system in target & non-target HRMS analysis

In this section, an IP system is proposed to help in the harmonization of HRMS-based identifications for target and non-target screening. This system aims at being simple and easy to use, with only objective criteria as outlined above. The maximum score of an identification can reach 1.00 for target screening and 0.75 for suspect and non-target screening. The purchase of reference standard for the confirmation of the identification (i.e. target analysis) is mandatory to achieve the highest IP score of 1.00. The fact that the system scales from 0 to 1 is important to communicate the identification confidence to non-experts. It can transfer the information immediately to non-experts and can help implement and embed non-target screening into future regulatory frameworks in an easily interpretable manner.

Accuracy below 2 mDa / 5 ppm for the precursor ion was regarded as mandatory. Only for target screening, a retention time match with a reference standard (± 0.2 min in target screening) the IP is increased by 0.40 points. The ± 0.2 min decision was based on the decision of European Commission 2002/657/EC **[49]** and the fact that robustness of the LC systems has greatly improved during the last decades. For non-target screening where retention time match is not available, retention time index (RTI) is used. In case of RTI match (typically ± 20% in suspect/non-target screening) the IP is increased by 0.15 points (decision based on **Table 2**). The tolerance on RTI depends on the structure of the suspected molecule, the QSRR model and the RTI system that is used. The number of IPs can increase

by 0.20, in case of excellent isotopic pattern fits match (decision based on **Table 2**). Fragmentation information can increase the IP by a tital of 0.40 (experimental spectra available) and by a total of 0.20 (*in-silico* spectra available). This decision was based on $Fragment_{score}$, $SpecSimil_{score}$ and partially on $FitMolForm_{score}$ (**Table 2**), because $FitMolForm_{score}$ does not explicitly correspond to fragmentation. *In-silico* fragmentation score is not considered in cases where meaningful experimental fragmentation is available. The 0.40 points due to fragmentation match with experimental spectra are split: 0.20 points in case of match of the most abundant fragment and 0.20 with the remaining fragments. A penalty of -0.10 points is applied in case of a compound with poor fragmentation (≤2 fragments). Finally, a penalty of -0.10 points is applied in case there is no recorded data-dependent scan with clear isolation and fragmentation of the precursor ion. This penalty relates to the fact that DIA suffers from matrix interferences. Introduction of additional separation dimensions (e.g. ion mobility) or other advanced acquisition types (e.g. SWATH MS) can make DIA acquisition more efficient and this penalty could thus be eliminated. However, this aspect has not been thoroughly investigated yet.

***Table 3**. Proposed Identification Point (IP) system in target and non-target HRMS analysis*

Requirements	**Identification Points (IP) earned**
Precursor ion (Accuracy < 2 mDa / 5 ppm, R>15000)	mandatory
Retention time ± 0.2 min (only applicable in target)	0.40
Predicted Retention time index (only applicable in suspect where retention time match is not available, validated approach with provided uncertainty)	0.15
Isotopic fit (at least one isotope: abundance and accuracy of M+1, M+2,…)	0.20
Most intense **experimental** fragment ion	0.20
All other **experimental** fragment ions Number of experimental fragments normalized to the total number of fragments in the library	0.20
The "All other **experimental** fragment ions" score is penalized if the number of other **experimental** fragments present in the database is 2 or less	-0.10
In silico predicted fragment ions in case experimental fragments are not available Number of experimental fragments normalized to the total number of fragments in the library max number of fragments in library=10 most intense	0.20
Only DIA	-0.10

Overall, to avoid subjective evaluations, the use of software to calculate the isotopic fit is advised. The use of a single software (either vendor or open source) for a given case-study is highly encouraged. The reason for this recommendation is that there are various methodologies to calculate isotopic fit (e.g., dot product and overlap percentage). In this way, unbiased identification evaluations can be

achieved in a flexible manner. The IP value can be increased by the determination of previously known fragment ions with accurate mass at the same RT (i.e., target screening). For in house method comparison, the same system and instrumental conditions applying proper quality controls to ensure RT accuracy and MS/MS spectra consistency should be used. An attempt to associate the IP system (**Table 3**) with the widely used identification levels **[10]** is presented in **Table 4**. Level 1 (confirmed identification) requires IP score higher than 0.75. Identifications of level 2 (probable structure) require IP score from 0.60 to 0.75, whereas level 3 (tentative identification) requires score higher than or equal to 0.50 and less or equal to 0.60. To claim a Level 4 (unequivocal molecular formula) identification, the score should be below 0.5 and higher or equal to 0.2. All identifications that receive below 0.20 IP can be presented as level 5 (exact mass) identifications.

Table 4. Connection of the identification levels **[10]** with the IP score proposed in this study.

Identification level	IP Score
1	>0.75-1.00
2	>0.60-0.75
3	0.50-0.60
4	>0.20-<0.50
5	0.00-0.20

4.1 Application of IP score in target screening

The first example (**Figure 2a**) shows an ideal target identification: the analysis of oxazepam in surface water. In this case, a good peak for the precursor ion (*m/z*: 287.0582) was determined at the exact RT, along with a good isotopic profile (very clear with the presence of one Cl atom) and qualifier fragments at the same RT, reaching 1.0 IP, which translates to level 1 (**Table 4**).

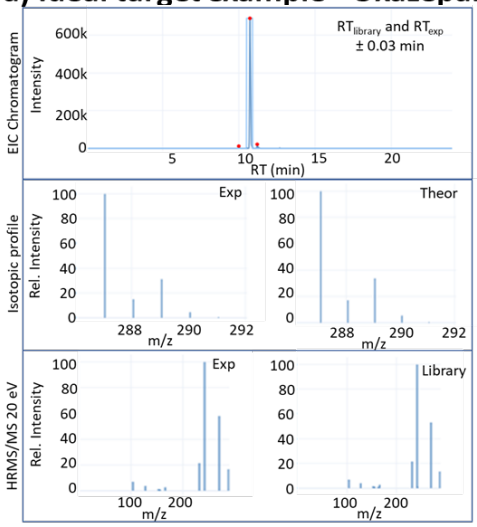
Since target analysis does not always lead to such clear IP identification, the second example (**Figure 2b**) shows the target identification of tramadol in the wastewater from the national French campaign **[50]**. In this case, the precursor ion (*m/z*: 264.1958) was determined with an acceptable RT (± 0.2 min) and isotopic fit, reaching 0.60 IP. Only one qualifier ion (the most intense) could be determined, adding 0.20 IP to finally reach a score of 0.80 IP. The score is penalized by 0.10 because the acquisition has been performed in DIA, reaching to 0.70 IP, corresponding to a level 2 ranking. It would have qualified as level 1 (score >0.75) if DDA acquisition had been performed.

A third example given in **Figure 2c** shows the determination of perfluorohexanesulfonic acid (PFHxS), which received just 0.60 IP, due to the lack of fragmentation of PFHxS. The lowest IP for target compounds was set to 0.60 IP (**Table 4**). The lower IP shows clearly that the identification has a lower confidence despite the matching reference standard. This information is often not provided for target analysis. This example does not qualify for level 1, but instead is given a Level 3.

Several other examples of the application of the IP system are provided for both target and suspect/non-target screening in the following sections and in the SI (**Table S3** for target screening and **Table S4** for non-target screening). **Table S3** provides 11 additional target screening examples. More specifically, it provides 1) an example with maximum possible score, 2) an ideal target screening example, 3) an acceptable target example, 4) a target example with isotopic fit but without fragments,

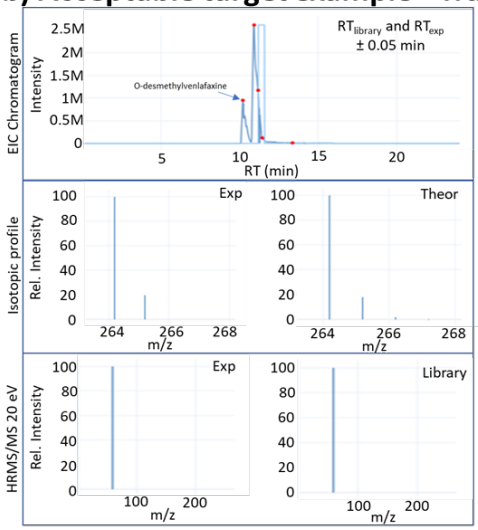
5) an ideal target example in DIA, 6) another target example in DIA, 7) a poor target example in DIA, 8) a target example without isotopic fit and fragments, 9) a target example with no isotopic fit, 10) a target example with no isotopic fit and no other experimental fragments, and 11) a target example without retention time but isotopic fit and fragments. The examples of **Table S3** match the IP to the well-established identification levels **[10]**.

a) Ideal target example - Oxazepam



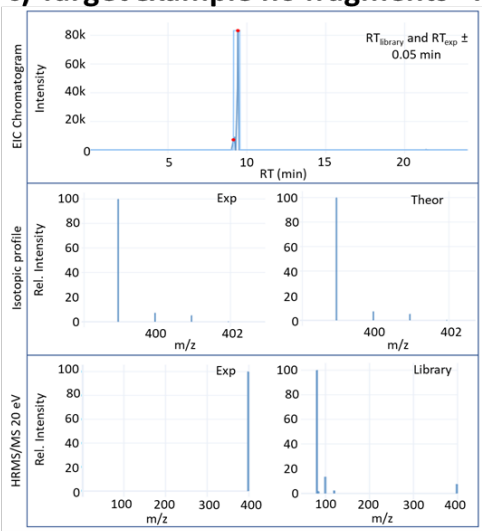
Ideal target example	
Substance	**Oxazepam**
Precursor ion (accurate mass)	mandatory
Retention time (only applicable in target)	0.40
Predicted Retention time index (only applicable in suspect where retention time match is not available, validated approach with provided uncertainty)	/
Isotopic fit	0.20
Most intense **experimental** fragment ion	0.20
All other **experimental** fragment ions Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	0.20
The "All other experimental fragment ions" score is penalized if the number of other experimental fragments present in the database is 2 or less	/
In-silico predicted fragment ions in case experimental fragments are not available Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	/
Only DIA	/
Total Score	1.00
Confidence level (Schymanski et al., 2014)	1

b) Acceptable target example - Tramadol



Acceptable target example	
Substance	**Tramadol**
Precursor ion (accurate mass)	mandatory
Retention time (only applicable in target)	0.40
Predicted Retention time index (only applicable in suspect where retention time match is not available, validated approach with provided uncertainty)	/
Isotopic fit	0.20
Most intense **experimental** fragment ion	0.20
All other **experimental** fragment ions Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	0
The "All other experimental fragment ions" score is penalized if the number of other experimental fragments present in the database is 2 or less	/
In-silico predicted fragment ions in case experimental fragments are not available Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	/
Only DIA	/
Total Score	0.80
Confidence level (Schymanski et al., 2014)	1

c) Target example no fragments - PFHxS



Target example with isotopic fit but without fragments	
Substance	**PFHxS**
Precursor ion (accurate mass)	Mandatory
Retention time (only applicable in target)	0.40
Predicted Retention time index (only applicable in suspect where retention time match is not available, validated approach with provided uncertainty)	/
Isotopic fit	0.20
Most intense **experimental** fragment ion	0
All other **experimental** fragment ions Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	0
The "All other experimental fragment ions" score is penalized if the number of other experimental fragments present in the database is 2 or less	/
In-silico predicted fragment ions in case experimental fragments are not available Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	/
Only DIA	/
Total Score	0.60
Confidence level (Schymanski et al., 2014)	3

Figure 2*. Target examples for IP identification: a) Oxazepam (DDA acquisition of surface water sample); b) tramadol (DIA acquisition of effluent wastewater - the compound is frequently confused with O-desmethyl-venlafaxine, which is the first peak shown in the chromatogram), c) PFHxS example with 0.60 IP evidence.*

4.2 Application of IP score in suspect screening

In suspect screening, identifications are more challenging given the lack of reference standards. Thus, the maximum score in a suspect identification is 0.55 IP for *in silico* predicted fragments and 0.75 for experimental fragments. The identification of the accurate mass of the parent ion with a plausible RT via a predicted RTI provides 0.15 IP. Isotopic fit can provide an additional 0.20 IP. While the presence of heteroatoms may provide additional meaning to isotopic fit, this is not reflected in the IPs to avoid additional complexity in the scheme. The presence of all fragments included in a good quality library can lead to a maximum of 0.40 IP. However, penalties in the score are applied if (i) only DIA data is available (-0.10), and (ii) the database for other experimental fragments (apart the most intense ion) includes two or less fragments (-0.10 IP).

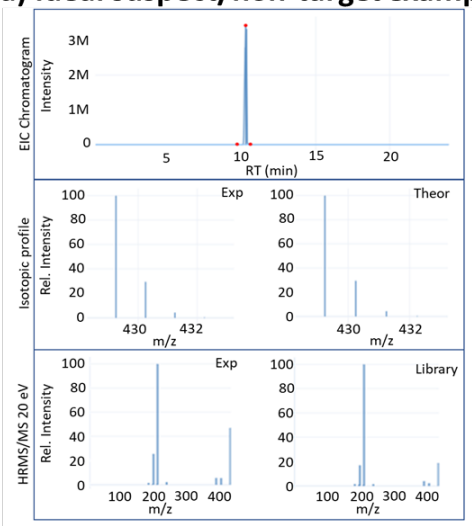
Figure 3a shows an example of the suspect identification of irbesartan. In this case, an intense and well-shaped peak was detected for the precursor ion (*m/z*: 429.2397) at a plausible RT according to the RT prediction model and excellent isotopic fit, obtaining 0.35 IP. The seven fragments included in the library were detected in the experimental spectra, providing additional 0.40 IP up to a total score of 0.75 IP, leading to a level 2 identification.

A second example of suspect screening with a slightly lower score is given in **Figure 3b**, showing the identification of triethyl phosphate (TEP). A score of 0.18 IP (out of 0.20 IP) was assigned for the isotopic fit, while the RTI within acceptable range (0.15 IP). To avoid subjective evaluations, the vendor software (Agilent MassHunter® Workstation Software) was used to calculate the isotopic fit, which was found to be 0.18 IP. In this case the three fragments present in the library were also detected (0.40 IP). However, given that a penalty is applied since only 2 other experimental fragments (apart the most intense one) were present, the identification ended up with a score of 0.63 IP, corresponding to level 2.

In the final example, less confidence was achieved in the case of nordiazepam (**Figure 3c**). The precursor ion was found at a plausible RT and good isotopic fit, indicating the presence of heteroatoms. The most intense fragment was detected (+0.20). Moreover, 5 of the 10 other fragments present in the library were detected, providing +0.10 IP, but since only DIA data was available (-0.10 IP), this led to a total score of 0.55 IP and a level 3 identification.

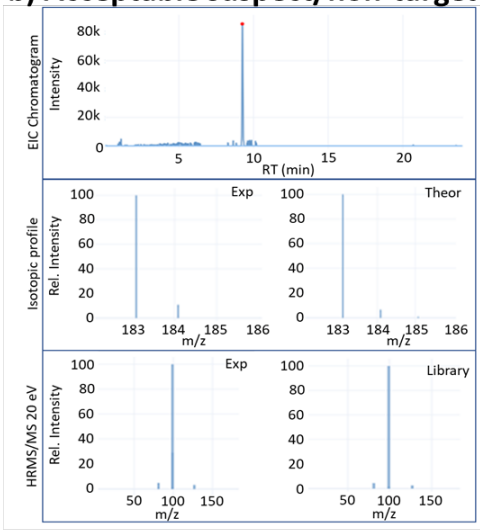
Table S4 provides 13 additional suspect/non-target examples. More specifically, it provides 1) an example with the maximum possible score, 2) an ideal non-target example, 3) an acceptable non-target example in DIA, 4) an example with partial fragment match in DIA, 5) an example with partial fragment match in DDA, 6) an example with partial isotopic fit, 7) an example with partial isotopic fit and partial fragment match, 8) an example without fragments, 9) an example without isotopic fit, 10) an example with only predicted RTI match, 11) an example without predicted retention index but ideal match for other scores, 12) an ideal example with match for the most intense fragment only, and 13) an ideal example with match for predicted fragments. The examples of **Table S4** match the IP to the well-established identification levels **[10]**.

a) Ideal suspect/non-target example - Irbesartan



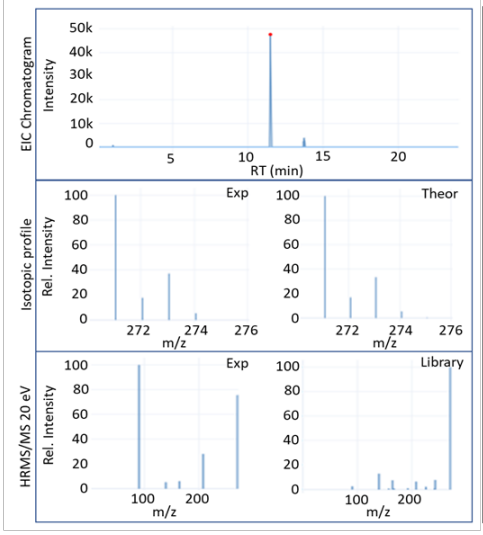
Ideal non-target example	
Substance	**Irbesartan**
Precursor ion (accurate mass)	mandatory
Retention time (only applicable in target)	/
Predicted Retention time index (only applicable in suspect where retention time match is not available, validated approach with provided uncertainty)	0.15
Isotopic fit	0.20
Most intense **experimental** fragment ion	0.20
All other **experimental** fragment ions Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	0.20
The "All other experimental fragment ions" score is penalized if the number of other experimental fragments present in the database is 2 or less	/
In-silico predicted fragment ions in case experimental fragments are not available Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	/
Only DIA	/
Total Score	0.75
Confidence level (Schymanski et al., 2014)	2

b) Acceptable suspect/non-target example - Triethyl phosphate (TEP)



Acceptable non-target example	
Substance	**TEP**
Precursor ion (accurate mass)	Mandatory
Retention time (only applicable in target)	/
Predicted Retention time index (only applicable in suspect where retention time match is not available, validated approach with provided uncertainty)	0.15
Isotopic fit	0.18
Most intense **experimental** fragment ion	0.20
All other **experimental** fragment ions Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	0.20
The "All other experimental fragment ions" score is penalized if the number of other experimental fragments present in the database is 2 or less	-0.10
In-silico predicted fragment ions in case experimental fragments are not available Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	/
Only DIA	/
Total Score	0.63
Confidence level (Schymanski et al., 2014)	2

c) Suspect/non-target example partial fragment match - Nordiazepam



Non-target example with partial fragment match - DIA	
Substance	**Nordiazepam**
Precursor ion (accurate mass)	Mandatory
Retention time (only applicable in target)	/
Predicted Retention time index (only applicable in suspect where retention time match is not available, validated approach with provided uncertainty)	0.15
Isotopic fit	0.20
Most intense **experimental** fragment ion	0.20
All other **experimental** fragment ions Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	0.10
The "All other experimental fragment ions" score is penalized if the number of other experimental fragments present in the database is 2 or less	/
In-silico predicted fragment ions in case experimental fragments are not available Number of experimental fragments normalized to the total number of fragments in the library (mass accuracy)	/
Only DIA	-0.10
Total Score	0.55
Confidence level (Schymanski et al., 2014)	3

Figure 3*. Suspect examples for IP identification: a) Irbesartan (DIA acquisition of wastewater sample); b) Nordiazepam (DIA acquisition of wastewater sample); triethyl phosphate (TEP) (DIA acquisition) in effluent wastewater sample*

4.3 Consideration of analysis of samples batch

In the case where several samples are analyzed by batch, the same substances can be determined in different samples at various levels/scores, depending notably on the intensities obtained. For instance, a substance analyzed via target screening and present at a high intensity in a sample of this batch would provide a maximum score of 1.0, corresponding to a level 1 identification. The same substance with a lower intensity in a different sample could potentially end up with a reduced score for isotopic fit and fragmentation score (score down to 0.60 for example leading to a level 3 rank). If there is sufficient evidence to indicate that it is indeed the same substance (notably by similar experimental retention times), then the latter case can be elevated to the level of the best scoring within the batch, here at level 1 instead of level 3. Overall, contemporary LC systems have robust retention time that should not shift more than 2.5% **[49]**. This means that for a chromatographic run of 1200 seconds (20 minutes), the maximum acceptable RT shift is 30 seconds. This consideration can be implemented with the requirement that the samples have been analyzed within the same batch and that LC system operates as expected. Given these restrictions, this operation can be automated.

5. Perspective: Towards a harmonized identification scoring system for NTS

Machine learning approaches can help in creating reproducible decisions on the evidence surrounding the confidence of identification. A higher degree of automation and the reduction of manual decisions will improve the reproducibility of NTS identification efforts and empower high throughput screening efforts. In this regard, the use of advanced models aimed to mimic/reproduce expert decisions will reduce the time need for a human to validate identification results, as the evidence can be presented clearly for quick confirmation. To ensure trust in machine driven data treatment, robust validation processes coupled with specific QA/QC procedures should be developed on large sample datasets to ensure the validity of the results. Based on the experience gained in this study, conducted with the results obtained by four laboratories with wide expertise in NTA, a scoring system is proposed that provides a simplified and harmonized approach for presenting the evidence associated with an identification. It aims at improving reproducibility and facilitating the communication of the evidence associated with identification based on objective criteria.

The design of the scoring system is based on current data extraction capabilities, both in terms of algorithmic and instrumentation limits. The proposal described in the present paper can serve as a basis that can and should be further improved and adapted to new technological and conceptual opportunities. A representative example can be found in the use of CCS values (both experimental and predicted), which have proven effective in confirming structure identification **[37]**. The use of CCS could be introduced into the scheme presented here once its use becomes more widespread in the majority of NTS laboratories, and thus when sufficient data is available for implementing the approach as described here.

A wide use of the scoring system by different users following their specific approaches with large data sets will help define the important pieces of evidence more precisely and improve the prediction accuracy. The system described and assessed here on a wide range of selected cases will be implemented in the NORMAN DSFP. This will enable a large-scale community validation and will help

determine whether the proposed system is ready to become a basis to support identification confidence communication in a reproducible and transparent manner.

Funding

PGF acknowledges his Ramon y Cajal fellowship (RYC2019-027913-I) from the AEI-MICI. ELS is supported by the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

Contributions

- **Nikiforos Alygizakis:** Writing original draft preparation, formal analysis, machine-learning, software development, review and editing
- **Francois Lestremau:** Writing original draft preparation, formal analysis, data contributor, method validation, review and editing
- **Pablo Gago-Ferrero:** Writing original draft preparation, formal analysis, data contributor, method validation, review and editing
- **Rubén Gil-Solsona:** Data contributor, Review and editing
- **Katarzyna Arturi:** Evaluation of machine-learning approaches, Review and editing
- **Juliane Hollender:** Data contributor, Review and editing
- **Emma L. Schymanski:** Formal analysis, Review and editing
- **Valeria Dulio:** Investigation, Review and editing
- **Jaroslav Slobodnik:** Investigation, Review and editing
- **Nikolaos S. Thomaidis:** Investigation, Review and editing

References

[1] J. Hollender, B. van Bavel, V. Dulio, E. Farmen, K. Furtmann, J. Koschorreck, U. Kunkel, M. Krauss, J. Munthe, M. Schlabach, J. Slobodnik, G. Stroomberg, T. Ternes, N.S. Thomaidis, A. Togola, V. Tornero, Environmental Sciences Europe 31 (2019) DOI: 10.1021/acs.est.7b02184.

[2] Z. Wang, G.W. Walker, D.C.G. Muir, K. Nagatani-Yoshida, Environmental Science & Technology 54 (2020) 2575 DOI: 10.1021/acs.est.9b06379.

[3] S. Finckh, L.-M. Beckers, W. Busch, E. Carmona, V. Dulio, L. Kramer, M. Krauss, L. Posthuma, T. Schulze, J. Slootweg, P.C. Von der Ohe, W. Brack, Environment International 164 (2022) 107234 DOI: 10.1016/j.envint.2022.107234.

[4] S. Petromelidou, D. Margaritis, C. Nannou, C. Keramydas, D.A. Lambropoulou, Science of The Total Environment 848 (2022) 157696 DOI: 10.1016/j.scitotenv.2022.157696.

[5] W. Yang, Y. Tang, L. Jiang, P. Luo, Y. Wu, Y. Cao, X. Wu, J. Xiong, Science of The Total Environment 809 (2022) 151117 DOI: 10.1016/j.scitotenv.2021.151117.

[6] Y. Han, L.-X. Hu, T. Liu, J. Liu, Y.-Q. Wang, J.-H. Zhao, Y.-S. Liu, J.-L. Zhao, G.-G. Ying, Science of The Total Environment 837 (2022) 155705 DOI: 10.1016/j.scitotenv.2022.155705.

[7] F. Menger, P. Gago-Ferrero, K. Wiberg, L. Ahrens, Trends in Environmental Analytical Chemistry 28 (2020) e00102 DOI: 10.1016/j.teac.2020.e00102.

[8] W.-L. Chen, S.-C. Lin, C.-H. Huang, S.-Y. Peng, Y.S. Ling, Science of The Total Environment 750 (2021) 141519 DOI: 10.1016/j.scitotenv.2020.141519.

[9] F. Freeling, N.A. Alygizakis, P.C. von der Ohe, J. Slobodnik, P. Oswald, R. Aalizadeh, L. Cirka, N.S. Thomaidis, M. Scheurer, Science of The Total Environment 681 (2019) 475 DOI: 10.1016/j.scitotenv.2019.04.445.

[10] E.L. Schymanski, J. Jeon, R. Gulde, K. Fenner, M. Ruff, H.P. Singer, J. Hollender, Environmental Science and Technology 48 (2014) 2097 DOI: 10.1021/es5002105.

[11] J.A. Charbonnet, C.A. McDonough, F. Xiao, T. Schwichtenberg, D. Cao, S. Kaserzon, K.V. Thomas, P. Dewapriya, B.J. Place, E.L. Schymanski, J.A. Field, D.E. Helbling, C.P. Higgins, Environmental Science & Technology Letters 9 (2022) 473 DOI: 10.1021/acs.estlett.2c00206.
[12] M. Pourchet, L. Debrauwer, J. Klanova, E.J. Price, A. Covaci, N. Caballero-Casero, H. Oberacher, M. Lamoree, A. Damont, F. Fenaille, J. Vlaanderen, J. Meijer, M. Krauss, D. Sarigiannis, R. Barouki, B. Le Bizec, J.P. Antignac, Environment International 139 (2020) 105545 DOI: 10.1016/j.envint.2020.105545.
[13] B. González-Gaya, N. Lopez-Herguedas, A. Santamaria, F. Mijangos, N. Etxebarria, M. Olivares, A. Prieto, O. Zuloaga, Chemosphere 274 (2021) 129964 DOI: 10.1016/j.chemosphere.2021.129964.
[14] P. Gago-Ferrero, A. Krettek, S. Fischer, K. Wiberg, L. Ahrens, Environmental Science and Technology 52 (2018) 6881 DOI: 10.1021/acs.est.7b06598.
[15] S. Samanipour, J.A. Baz-Lomba, N.A. Alygizakis, M.J. Reid, N.S. Thomaidis, K.V. Thomas, Journal of Chromatography A 1501 (2017) 68 DOI: 10.1016/j.chroma.2017.04.040.
[16] E.L. Schymanski, H.P. Singer, J. Slobodnik, I.M. Ipolyi, P. Oswald, M. Krauss, T. Schulze, P. Haglund, T. Letzel, S. Grosse, N.S. Thomaidis, A. Bletsou, C. Zwiener, M. Ibanez, T. Portoles, R. de Boer, M.J. Reid, M. Onghena, U. Kunkel, W. Schulz, A. Guillon, N. Noyon, G. Leroy, P. Bados, S. Bogialli, D. Stipanicev, P. Rostkowski, J. Hollender, Analytical Bioanalytical Chemistry 407 (2015) 6237 DOI: 10.1007/s00216-015-8681-7.
[17] P. Gago-Ferrero, A.A. Bletsou, D.E. Damalas, R. Aalizadeh, N.A. Alygizakis, H.P. Singer, J. Hollender, N.S. Thomaidis, Journal of Hazardous Materials 387 (2020) 121712 DOI: 10.1016/j.jhazmat.2019.121712.
[18] ISO Standard 21253, Water quality - Multi-compound class methods Criteria for the identification of target compounds by gas and liquid chromatography and mass spectrometry (2019).
[19] R. Helmus, T.L. ter Laak, A.P. van Wezel, P. de Voogt, E.L. Schymanski, Journal of Cheminformatics 13 (2021) 1 DOI: 10.1186/s13321-020-00477-w.
[20] N.A. Alygizakis, P. Oswald, N.S. Thomaidis, E.L. Schymanski, R. Aalizadeh, T. Schulze, M. Oswaldova, J. Slobodnik, TrAC Trends in Analytical Chemistry 115 (2019) 129 DOI: 10.1016/j.trac.2019.04.008.
[21] N. Caballero-Casero, L. Belova, P. Vervliet, J.-P. Antignac, A. Castaño, L. Debrauwer, M.E. López, C. Huber, J. Klanova, M. Krauss, A. Lommen, H.G.J. Mol, H. Oberacher, O. Pardo, E.J. Price, V. Reinstadler, C.M. Vitale, A.L.N. van Nuijs, A. Covaci, TrAC Trends in Analytical Chemistry 136 (2021) 116201 DOI: 10.1016/j.trac.2021.116201.
[22] P. Rostkowski, P. Haglund, R. Aalizadeh, N. Alygizakis, N. Thomaidis, J.B. Arandes, P.B. Nizzetto, P. Booij, H. Budzinski, P. Brunswick, A. Covaci, C. Gallampois, S. Grosse, R. Hindle, I. Ipolyi, K. Jobst, S.L. Kaserzon, P. Leonards, F. Lestremau, T. Letzel, J. Magnér, H. Matsukami, C. Moschet, P. Oswald, M. Plassmann, J. Slobodnik, C. Yang, Analytical and Bioanalytical Chemistry 411 (2019) 1957 DOI: 10.1007/s00216-019-01615-6.
[23] R. Bade, L. Bijlsma, T.H. Miller, L.P. Barron, J.V. Sancho, F. Hernández, Science of The Total Environment 538 (2015) 934 DOI: 10.1016/j.scitotenv.2015.08.078.
[24] C. Feng, Q. Xu, X. Qiu, Y.e. Jin, J. Ji, Y. Lin, S. Le, J. She, D. Lu, G. Wang, Chemosphere 271 (2021) 129447 DOI: 10.1016/j.chemosphere.2020.129447.
[25] D. Pasin, C.B. Mollerup, B.S. Rasmussen, K. Linnet, P.W. Dalsgaard, Analytica Chimica Acta 1184 (2021) 339035 DOI: 10.1016/j.aca.2021.339035.
[26] E. Kováts, Helvetica Chimica Acta 41 (1958) 1915 DOI: 10.1002/hlca.19580410703.
[27] R. Aalizadeh, N. Alygizakis, E.L. Schymanski, M. Krauss, T. Schulze, M. Ibáñez, A.D. McEachran, A. Chao, A.J. Williams, P. Gago-Ferrero, A. Covaci, C. Moschet, T. Young, J. Hollender, J. Slobodnik, N. Thomaidis, Analytical Chemistry 93 (2021) 11601 DOI: 10.1021/acs.analchem.1c02348
[28] MassBank, Available at https://massbank.eu/MassBank/, Last accessed 21 September 2022. 2022.
[29] MoNA, Available at https://mona.fiehnlab.ucdavis.edu/, Last accessed 21 September 2022. 2022.

[30] mzCloud, Available at https://www.mzcloud.org/, Last accessed 21 September 2022. 2022.
[31] K. Dührkop, M. Fleischauer, M. Ludwig, A.A. Aksenov, A.V. Melnik, M. Meusel, P.C. Dorrestein, J. Rousu, S. Böcker, Nature Methods 16 (2019) 299 DOI: 10.1038/s41592-019-0344-8.
[32] F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner, D.S. Wishart, Analytical Chemistry 93 (2021) 11692 DOI: 10.1021/acs.analchem.1c01465.
[33] C. Ruttkies, E.L. Schymanski, S. Wolf, J. Hollender, S. Neumann, Journal of Cheminformatics 8 (2016) 3 DOI: 10.1186/s13321-016-0115-9.
[34] L. Ridder, J.J.J. van der Hooft, S. Verhoeven, Mass Spectrometry 3 (2014) S0033 DOI: 10.5702/massspectrometry.S0033.
[35] M. Loos, nontarget: Detecting Isotope, Adduct and Homologue Relations in LC–MS Data https://CRAN.R-project.org/package=nontarget. 2016.
[36] C.D. Broeckling, F.A. Afsar, S. Neumann, A. Ben-Hur, J.E. Prenni, Analytical Chemistry 86 (2014) 6812 DOI: 10.1021/ac501530d.
[37] A. Celma, L. Ahrens, P. Gago-Ferrero, F. Hernández, F. López, J. Lundqvist, E. Pitarch, J.V. Sancho, K. Wiberg, L. Bijlsma, Chemosphere 280 (2021) 130799 DOI: 10.1016/j.chemosphere.2021.130799.
[38] D.H. Ross, J.H. Cho, L. Xu, Analytical Chemistry 92 (2020) 4548 DOI: 10.1021/acs.analchem.9b05772.
[39] M. Meringer, E.L. Schymanski, Metabolites 3 (2013) 440 DOI: 10.3390/metabo3020440.
[40] M. Meringer, S. Reinker, J. Zhang, A. Muller, Communications in Mathematical and in Computer Chemistry 65 (2011) ISSN: 0340-6253.
[41] N. Dodder, K. Mullen, OrgMassSpecR: Organic Mass Spectrometry (Available at https://cran.r-project.org/web/packages/OrgMassSpecR/index.html, Last accessed 21 January 2022). 2017.
[42] A. Togola, C. Guillemain, F. Lestremau, C. Coureau, C. Margoum, C. Soulier, APPLICABILITE DE LA TECHNIQUE DE « SCREENING NON CIBLE » POUR LA SURVEILLANCE PROSPECTIVE. Réseau de Surveillance prospective-AQUAREF ; Rapport BRGM/RP-70108-FR (Accessible at https://professionnels.ofb.fr/sites/default/files/pdf/documentation/Pollution/LotC_4.pdf, Last accessed 21 January 2022). 2020.
[43] R. Gil-Solsona, M.-C. Nika, M. Bustamante, C.M. Villanueva, M. Foraster, M. Cosin-Tomás, N. Alygizakis, M.D. Gómez-Roig, E. Llurba-Olive, J. Sunyer, N.S. Thomaidis, P. Dadvand, P. Gago-Ferrero, Environmental Science & Technology Letters 8 (2021) 1077 DOI: 10.1021/acs.estlett.1c00848.
[44] R. Gil-Solsona, S. Rodriguez-Mozaz, M.S. Diaz-Cruz, A. Sunyer-Caldu, T. Luarte, J. Hofer, C. Galban-Malagon, P. Gago-Ferrero, MethodsX 8 (2021) 101193 DOI: 10.1016/j.mex.2020.101193.
[45] K. Kiefer, A. Muller, H. Singer, J. Hollender, Water Research 165 (2019) 114972 DOI: 10.1016/j.watres.2019.114972.
[46] R. Tautenhahn, C. Böttcher, S. Neumann, BMC Bioinformatics 9 (2008) 504 DOI: 10.1186/1471-2105-9-504.
[47] G. Libiseller, M. Dvorzak, U. Kleb, E. Gander, T. Eisenberg, F. Madeo, S. Neumann, G. Trausinger, F. Sinner, T. Pieber, C. Magnes, BMC Bioinformatics 16 (2015) 118 DOI: 10.1186/s12859-015-0562-8.
[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Journal of Machine Learning Research 12 (2011) 2825.
[49] European Commission, Commission decision of 12 August 2002 implementing Council Directive 96/23/EC concernign the performance of analytical methods and the interpratation of results Official Journal of the European Commission 221 (2002) 8.
[50] A. Assoumani, F. Lestremau, M. Salomon, C. Ferret, B. Lepot, Campagne Emergents Nationaux 2018 (EMNAT 2018) - Résultats de la recherche de contaminants émergents dans les eaux de surface et les rejets de STEU (2020).