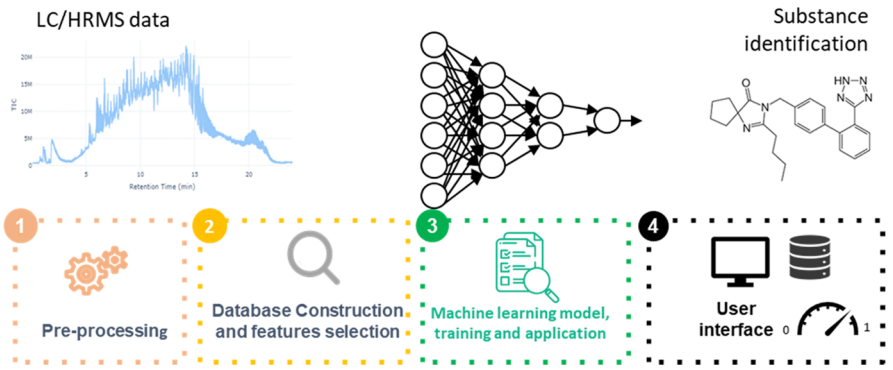
Development and Implementation of Automated Qualification Processes for the Identification of Pollutants in an Aquatic Environment from High-Resolution Mass Spectrometric Nontarget Screening Data

Francois Lestremau, Alexandre Levesque, Abdelmoughit Lahssini, Tanguy Magnan de Bornier, Romain Laurans, Azziz Assoumani, Hugues Biaudet

Development and Implementation of Automated Qualification Processes for the Identification of Pollutants in an Aquatic Environment from High-Resolution Mass Spectrometric Nontarget Screening Data

Francois Lestremau,* Alexandre Levesque, Abdelmoughit Lahssini, Tanguy Magnan de Bornier, Romain Laurans, Azziz Assoumani, and Hugues Biaudet



ABSTRACT: Environmental pollution monitoring represents a major challenge due to the growing presence of a large and diverse number of potential contaminants. In complement to target analysis, nontarget analysis, via liquid chromatography (LC) coupled to high-resolution mass spectrometry (HRMS), is increasingly used to provide a more comprehensive characterization of pollution. The challenge associated with this type of analysis is particularly related to the data treatment for substance identification. One of the main limitations is that all data must be manually reviewed, which is tedious and time-consuming. Machine learning algorithms aim to reproduce human behavior, and their capabilities were therefore evaluated to automatically identify substances in suspect screening approaches. After selecting the relevant features produced from LC/HRMS, seven different machine learning models were evaluated for each of the three different databases, which resulted in the selection of logistic regression (LR) and random forest (RF)-based algorithms. An interface was built to rank the identified substances and to assess the performance of the developed models. The LR model provided the best results when retention times were available. The developed LR and RF models were determined complementarily, particularly when no retention times were available. However, limitations were noticed when using a database containing different HRMS technologies.

KEYWORDS: *suspect screening, LC/HRMS, machine learning, identification, contaminants, water*

INTRODUCTION

With the improvement of capabilities of the high-resolution mass spectrometer, nontarget screening, which aims to provide a more representative view of the presence of organic substances in samples, has been a growing field in the past few decades in the characterization of environmental pollution.[1,2] Coupled to gas or liquid chromatography, nontarget screening strategies were first originally applied to metabolomics studies.[3−6] It has been extended to many fields of environmental studies to characterize pollutants, for instance, in atmospheric, soil, or water matrices.[7,8]

Different categories can be distinguished for nontarget screening, being used for fingerprint pattern comparison of data samples from different days/sites or for aiming at the identification of detected substances.[9,10] Identification of substances detected by nontarget screening approaches can therefore be performed using mass spectra databases in a so-called "suspect screening".[11] The constitution of databases is critical to ensure that a large number of substances can be

accurately identified. Laboratories that developed their own databases can generally inject only a limited amount of analytical standards, between a hundred up to a thousand, essentially due to the cost limitation of purchasing them. To overcome this limitation, databases from external sources can be employed. Manufacturers therefore provide spectral databases, which can contain more than thousands of substances. Collaborative databases have also been developed by the scientific community. MassBank,[12,13] for instance, contains overall up to 70,000 spectra from various types of instrumentations and acquisition modes provided by diverse voluntary contributor organisms. However, differences in fragmentation patterns, in the number of fragments, and in their relative intensities can occur between these different types of instruments/vendors according to the technologies set up even if several papers have reported that comparability between QTOF spectra and Orbitrap spectra was providing a good correlation.[14,15]

For data treatment, dedicated software are used to automate the substance identification (either from vendors or from independent groups, for instance, EnviMass,[16] MS DIAL,[17] or patRoon[18]). Different strategies have been developed depending on the experimental data (molecular ion, retention time, isotopic profile, fragmentation data) to provide assistance to the user for substance identification. However, as there is a huge amount of data to take into account and most of them can suffer from more or less pronounced interferences, it is common practice that experts manually validate all substance identifications since the risk of false-positives can be large. All proposed identifications have therefore to be reviewed, which is time-consuming and represents a tedious task. This task is also likely to become even more time-consuming as databases are expected to grow and the number of substances to be identified will therefore increase. This will lead to a completely unacceptable amount of time dedicated to data analysis. An additional downside is that the validation depends on the experience and view of the expert who carries on this process. Therefore, the validation procedure must be carefully defined to obtain a reproducible process among different experts.

Artificial intelligence/machine learning approaches are increasingly used to improve mass spectrometry-based data treatment due to major improvements in recent years in computing capabilities. Therefore, for example, metabolomic communities have explored capabilities of using machine learning to highlight biomarkers of diseases.[19,20] Deep learning methods have also been investigated for example to improve peak picking[21] or mass spectra comparison.[22]

This study was therefore dedicated to evaluating the use of artificial intelligence to develop a scoring system that can provide improved assistance to users for the verification of substance identification in the suspect screening approach. It ideally aims to provide a reproducible and automated validation process without the need of a manual review. The study focused particularly on water sample contaminants. Various parameters were considered for the development of the scoring system with a focus on processing data-independent analysis results. Two main categories of databases were considered to develop a dedicated data treatment algorithm: a first one containing data generated with the same type of instruments and a second one constituted from different types of instruments. The obtained algorithms were applied on a test data set to establish the degree of confidence of the scoring system and the possibility of automating substance identification.

EXPERIMENTAL SECTION

Data Sources. *Data Samples.* Data were obtained from the analysis of surface and wastewater samples from different sites across France as part of a national French monitoring campaign.[23] All details about the samples and their extraction procedure are described in this report. The samples extracted onto the solid phase (SPE) were analyzed using an LC/QTOF 6550 system (Agilent technologies). Chromatographic analysis used a Zorbax Aq column (150 mm × 2.1 mm, 1.8 μm) (Agilent technologies) with 1 mM ammonium acetate and methanol as mobile phases. Detailed chromatographic conditions are presented in SI 1 in Table SI 1. All samples' data were produced in the Agilent format (.d) and converted in the mzML format using MS convert software (Proteowizard[24]).

Data Acquisition Mode. Analysis in the data-independent mode was preferred to provide an exhaustive sample characterization since it has the advantage that all ionized substances that reach the collision cell are fragmented. Two collision energies, 20 and 40 eV, were systematically acquired in positive and negative modes for all data and included in the process of the scoring model.

Databases. Three databases were used in this work.

- A database created by our laboratory from the injection of 207 standards in the positive mode (POS) and 80 in the negative mode (NEG) in the same analytical conditions as used for the sample analyses.
- A database provided by Agilent in the sdf format, which included 919 substances in POS and 295 in NEG.
- An extract of the MassBank Europe database—Release version 2021.03.[13] Only spectra containing high-resolution MS/MS information were considered for this study representing a total of 1717 substances in POS and 738 in NEG. As MassBank data contained spectra from many different conditions and instruments, the data were curated and homogenized to fit with the intended study. All details about preprocessing of MassBank data are provided in SI 2.

The laboratory and vendor databases included for all entries mass spectra fragmentation at 10, 20, and 40 eV. Some entries could include, if relevant, adduct spectra. Details on those data are provided in Table SI 2. All databases were gathered, formatted, and cleaned using homemade scripts. After this preprocessing step, they were uploaded and stored in a PostgreSQL database.

Data Treatment. All data were treated using a developed in-house interface used to extract the relevant features and display the results generated by the machine learning approach (classification scores). Details on the data treatment are provided in SI 3. General information about the treatment of nontarget screening data and classification aspects have been provided by Fisher et al.[25] and can help facilitate the comprehension of the results of this study.

Classification Problem and Machine Learning Approach. Suspect screening relies on looking for a large number of substances from a reference database into an experimental sample. For each individual substance investigated, the challenge is to certify if it is present or not in the sample, which practically is being able to identify the substance or not.

Table 1. Summary of All Features Considered for the Development of the Scoring System

type	features	description	study
fragmentation at 20 eV	cos_sim_20	cosine similarity between the experimental fragmentation spectrum and the theoretical one at 20 eV	A,B,C
	MassBank_sim_20	MassBank similarity between the experimental fragmentation spectrum and the theoretical one at 20 eV	A,B,C
	nb_matched_peaks_20	number of matched peaks between the experimental fragmentation spectrum and the theoretical one at 20 eV	only C
	ratio_peaks_found_20	ratio of peaks matched at 20 eV (= nb_matched_peaks_20/total_nb_peaks_20)	A,B,C
	nb_high_peaks_20	number of experimental fragments at 20 eV with a mass >100	only C
	error_sum_20	cumulated sum of errors in mass peak by peak between the experimental fragmentation spectrum and the theoretical one at 20 eV	A,B,C
fragmentation at 40 eV	cos_sim_40	cosine similarity between the experimental fragmentation spectrum and the theoretical one at 40 eV	A,B,C
	MassBank_sim_40	MassBank similarity between the experimental fragmentation spectrum and the theoretical one at 40 eV	A,B,C
	nb_matched_peaks_40	number of matched peaks between the experimental fragmentation spectrum and the theoretical one at 40 eV	only C
	ratio_peaks_found_20	ratio of peaks matched at 40 eV (= nb_matched_peaks_40/total_nb_peaks_40)	A,B,C
	nb_high_peaks_20	number of experimental fragments at 40 eV with a mass >100	only C
	error_sum_40	cumulated sum of errors in mass peak by peak between the experimental fragmentation spectrum and the theoretical one at 40 eV	A,B,C
isotopic profile	cos_sim_isotopic	cosine similarity between the experimental isotopic spectrum and the theoretical one	A,B,C
	nb_matched_peaks_isotopic	number of matched peaks between the experimental isotopic spectrum and the theoretical one at 20 eV	A,B,C
other considered features	tic_flag	1 if apex intensity is higher than an intensity defined by expert; else 0	A,B,C

This constitutes a binary classification problem, which can be solved using different approaches. One particular method is to use machine learning based on common properties, so-called features, to provide an answer to this challenge with a confidence value associated. Many machine learning algorithms offer binary classification. One of the critical aspects of this study was to select the most relevant features and model to achieve an optimum classification.

Features Selected for Scoring System Evaluation. Feature Overview. To develop the algorithm and rank the probability of substance identification, a scoring system was developed. Liquid chromatography/high-resolution mass spectrometry (LC/HRMS) data contain different types of information generated from chromatography, the MS TIC scan, or the MS fragmentation pattern. Therefore, the following features, considered the most significant, were selected for developing the scoring system (Table 1).

As more variability was expected when comparing with a database including multiple vendor spectra (MassBank data), additional features were specifically included for this study (Study C).

Details about the choice and the description of each feature are provided in SI 3.

Construction of the Database Used for the Training Phase. A training phase was carried out to develop the scoring model. Representative experimental samples were manually characterized and labeled. For this set of substances, the selected features were built in comparison with the considered reference databases, and various algorithms were trained. The training data were built so that the two classes, substance is present/is not present in the experimental sample, were equally populated to avoid any bias.

Substances Considered for the Training. Twenty SPE extracts of surface water samples from a national French monitoring campaign were considered as case studies to train the model (results for this study have been previously published in a report[23]). The data of these SPE extracts were treated with vendor software, and the same two databases (laboratory and vendor) were used in this study. Using the vendor database in both cases allowed one to also consider a larger number of substances and to include situations where no retention times were available. For the classification of the common approach, an adaptation of the Schymanski scale[26] was used. Substances for which retention times were matching and at least or more than one fragment were determined were classified as level 1. Substances without retention times but for which at least three fragments were determined were classified as level 2. To minimize the rate of false-positives, only these two levels of classification, and therefore the corresponding substances, were used to train the model. All results were manually inspected with Agilent vendor software (Masshunter) and validated using the criteria described previously. Although this approach is not fully based on matching with retention times of analytical standards (when they were not used) and can therefore represent a larger risk to include some false-positives, this choice of using real sample data over using injected analytical standards was preferred since it could provide more representative samples and a different level of intensities, particularly in the case of data-independent analysis where a larger noise is produced from the various sample matrices for instance. This mode of analysis was selected particularly since all ionized substances from the analyzed samples produce fragmentation data and thereby a higher level of interference, leading to a more challenging data interpretation.

Overall and across all 20 samples in positive and negative ionization modes and in the data-independent analysis mode, approximately 1200 substances (600 for each category) were selected for the training data set (including many identical substances found in the different samples with different abundances and impacted with different matrix effects).

Substances Not Present Considered for the Training. To be able to train the algorithm, a category where molecules are not present in the samples had to be defined. It had to

Table 2. Performances Obtained with Different Classifiers for Study A

	model	fitting time	scoring time	accuracy	precision	recall	F1_score
5	random forests	0.324726	0.030378	0.965347	0.966728	0.965054	0.965296
4	quadratic discriminant analyses	0.008564	0.008156	0.961425	0.963795	0.960994	0.961312
0	logistic regression	0.018897	0.011027	0.953808	0.959158	0.953066	0.953484
2	support vector machines	0.019733	0.009340	0.953808	0.960903	0.952846	0.953399
1	decision tree	0.010815	0.008656	0.953771	0.955632	0.953510	0.953699
6	K-nearest neighbors	0.005651	0.013044	0.946116	0.954928	0.945000	0.945545
7	Bayes	0.007140	0.007359	0.934351	0.937782	0.935157	0.934185
3	linear discriminant analyses	0.010990	0.009642	0.924811	0.928917	0.924048	0.924512

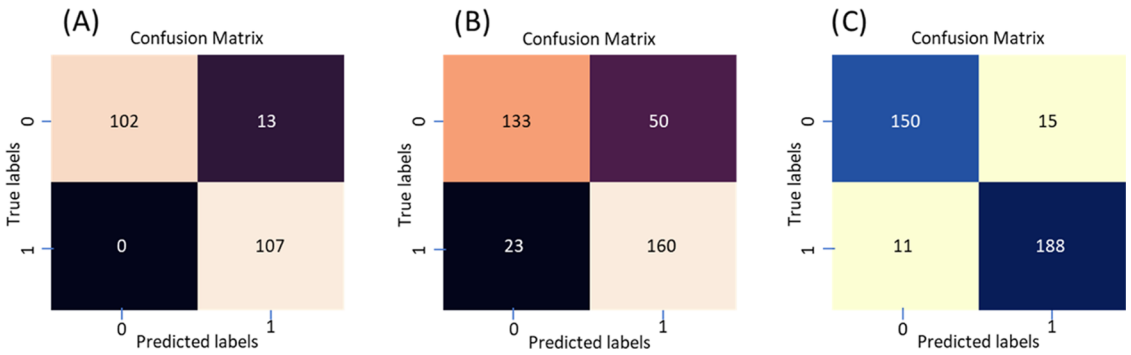


Figure 1. Confusion matrix on the selected model using (A) laboratory database (Study A), (B) vendor database (Study B), and (C) MassBank database (Study C).

represent roughly the same size in number as the category considered for the molecules present in the samples. One approach could have been to consider substances in the databases that had not been determined in the samples. However, this approach can potentially highlight substances that were not identified using the common laboratory approach (with vendor software) but that were in fact present in the sample (false-negative). Therefore, instead of considering all of the molecules of the databases not identified, another approach was preferred. To generate a set of molecules classified as not present in the sample, it was chosen to label the peaks with an identical molecular ion but with retention times different from those of the molecules previously identified in the samples.

Development and Test of the Different Algorithms. Machine Learning Models Considered. The Python Package scikit-learn for machine learning models was used to tackle the data classification.[27] The following eight classification models implemented within this package were tested: decision tree/random forest (RF), KNN classifier, MLP classifier, logistic regression (LR), Bayes classifier, and SVM classifier (some comparison elements can be found elsewhere[28]).

Evaluation of the Models. For the evaluation phase, the same approach was used for all evaluated algorithms to provide a fair comparison between them. Thus, the labeled data set was divided into 80% dedicated to training (i.e., the training set) and 20% dedicated to testing (i.e., the test set). The different trainings were performed with a cross-validation (k-fold = 10) for which the training data were shuffled and split into 10 random subsets to perform small independent trainings. This method usually results in less biased or less optimistic estimates of the classifier than a simple train/test split. The same standard measures were used to evaluate the models. It was based on the confusion matrices of the models evaluated on the test set. The numbers of correct and incorrect predictions are summarized and broken down into each class with count values.

Evaluation of Performance of the Developed Models with the Test Set. The models were evaluated on a test set, which comprised data from four different surface water samples generated in the same analytical conditions, and that was independent of the training set used for the devolvement of the algorithm. All data were reviewed manually and classified according to the Schymanski scale[26] adapted in level 1 and level 2 categories. Correlation between manual classification and scoring obtained for each model was then performed.

For the three studies performed and the models selected (based with and without retention times on QTOF data and on MassBank data—see the results part), the scoring obtained for each model whatever the data processing was calculated. For example, even if the data treatment was performed with the laboratory database, the "confidence" scores were also calculated with the model developed without considering the retention times (based also on logistic regression) and with the model developed from the MassBank Database (based on RF and using more features). That provided additional information that could be used to compare the different models. That was particularly relevant for the model without using the retention times, which has been selected based on logistic regression although during the development, RF was providing better results.

RESULTS AND DISCUSSION

Development of the Scoring Model with the Same HRMS Technology. Mass spectrometry data and particularly fragmentation data can be influenced by the technology (and experimental conditions) used. Therefore, to minimize the number of factors having an impact on the variability of the data produced, only data produced by an Agilent QTOF system were initially considered.

The first part of this work was therefore dedicated to evaluating an automated scoring system using a unique instrumentation type. Two cases were considered: when the retention times in the database were available (Study A), usually using a laboratory database; and when a vendor database is used therefore without using the retention times feature (Study B). Eight machine learning models were evaluated in each case.

Study Using Retention Times (Study A). The best performance could be observed with the "Random forest" (RF) algorithm with an accuracy of 96% (Table 2). Even if the performances obtained were slightly lower (95% accuracy), it was preferred to select the "logistic regression" (LR) algorithm. Indeed, when a visual inspection of the results with high scores determined with the RF model was performed, some inconsistencies were observed on certain substances, and moreover, the logistic regression model was more understandable to rank the substance scores based on their confidence level. The distribution of the different parameters according to their classification is displayed in Figure SI 1, and the coefficient obtained per considered parameter is shown in Table SI 4.

To visualize the separation of the two classes with the chosen algorithm, Figure SI 3 displays a PCA separating the zones of the two classes 0 (in red) and 1 (in blue) created by the algorithm. It can be noticed that both sets of data are clearly separated. The confusion matrix performed on the test set (Figure 1A) also demonstrated the good performances of the developed model. No false-negatives were noticed on the training set with an acceptable limited number of false-positives. This result was determined acceptable since no suspected substances would be missed when processing the data and a manual double-check of the remaining doubtful cases can be manageable. A robust probability score, obtained with the selected LR model, enabled a ranking of the substances based on the confidence of their presence (1 being the highest).

Study Not Using Retention Times (Study B). As for the study using retention times, the best performance could be observed with the RF model but with a lower accuracy at 92% (Table SI 5). To be in accordance and for the same reason as for the choice of study A, the LR model, for which a lower accuracy of 0.81 was determined, was finally preferred.

To visualize the separation of the two classes with the LR model, Figure SI 6 displays a PCA separating the zones of the two classes 0 (in red) and 1 (in blue) defined by the algorithm. While both sets of data are correctly separated, it is not as clear as to what was obtained for the study using the retention times. This is also reflected by the confusion matrix performed on the test set (Figure 1B), which also displayed lower performances than for the other developed model. The retention times are indeed strong indicators in the identification process of a substance and bring a superior level of confidence than when not used or available in the database. There is therefore more uncertainty in the identification process when a model without considering this parameter is used.

Evaluation with a Database Using Multiple HRMS Technologies (Study C). The second part of the work was dedicated to evaluating if the same approach could be potentially used with a database including data produced from a large panel of technologies/vendors. The MassBank database was therefore used. Data were curated and homogenized, and then, a ranking was performed to select only one entry per substance considered. Additional features were also included to reflect the expected increase in variabilities in the reference spectra from multiple vendors (see Experimental Section, Table 1).

Development, Training, and Comparison of Different Algorithms. The eight classification models were tested on MassBank data. The same set of sample data as the one used for the first part was also selected. Therefore, only the substances that were common between MassBank and the substances determined in the samples using the standard laboratory procedure were considered, which represented overall 1220 substances. Since more variability was expected, the pool of samples was extended with seven wastewater samples also previously characterized by the laboratory. Standard solutions of pesticides and pharmaceuticals injected at three different concentrations, low, middle, and high, were also used to have known substances at different levels of intensities. Overall, 2155 different substances were used instead of 1200 in the first phase of the project.

The same approach as for the first part was set up, with two categories defined, which corresponded to the presence of the substance in the sample (label 1) and when the substance was not present (label 0).

Model Developed (Model C). Average performances for each of the eight models are presented in Table SI 6, with the best performance observed for the RF algorithm reaching an accuracy of 94%. This model was therefore selected. The confusion matrix performed on the test set (Figure 1C) indicated a robust classification with a very low rate of false-positives and false-negatives.

An overall summary of the three investigated studies (Studies A, B, and C) is presented in Table SI 7.

Evaluation of the Performances of the Developed Models. Four data sets from river sample analyses were processed with the three different models to reproduce the approach of assessing a confidence level for suspect analysis (LR with retention times—Study A; LR without retention times—Study B; and RF—Study C)/database (laboratory, vendor, and MassBank). This process provided a score for every substance matching with a database entry. All data were also manually reviewed and classified according to the Schymanski scale[32]. The scoring obtained with the machine learning process was then compared with the manual review.

It can also be pointed out that while for each database used, a specific model was selected (for instance, an LR data treatment performed for Study A with the laboratory database, for instance), the scores based on LR and RF models were anyway systematically calculated. That feature was particularly useful considering Study C based on RF and including more parameters than for Studies A and B when data treatments were performed on the LR model.

Comparison of Performances Depending on Databases. With the threshold of scoring set at LR $\geq$ 0.8 (out of a maximum scoring of 1), comparable results could be obtained with Studies A and B regarding the percentage of substances that could be assessed with the highest level of confidence (Figure SI 9). For the positive ionization mode (POS), 86% of the substances were classified as level 1 using the laboratory database, while 84% were set at level 2 with the vendor database. For the negative mode (NEG), 74% of the highest confidence level was determined in both cases. The negative mode produces generally less fragments than the positive

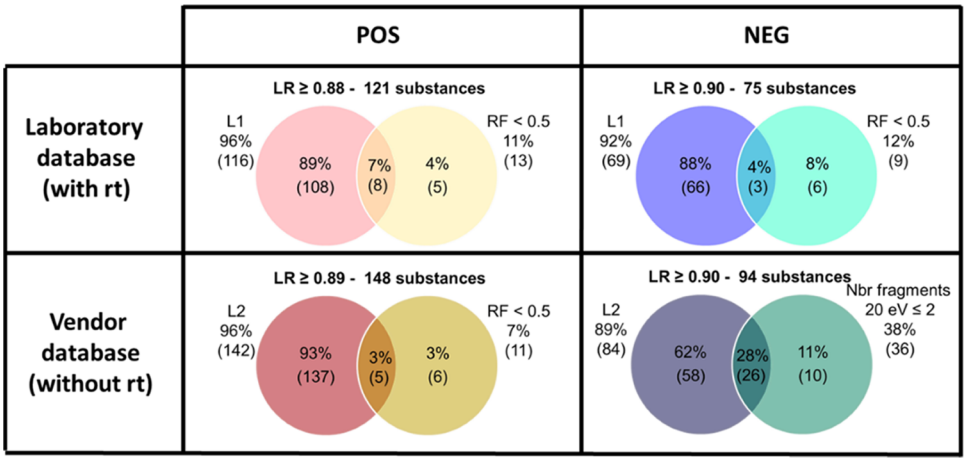


Figure 2. Performances with the highest possible confidence levels obtained depending on the model/databases used and the ionization modes (LR, logistic regression; RF, random forest; rt, retention time; L1/L2, level 1/2; in bracket, number of substances for each category).

mode, therefore leading to less confidence for substances with a low number of fragments.

Through the evaluation of the test results, it was determined that the RF model could be complementary to the use of the LR model. When using the LR model with retention times (Study A), RF did not bring any additional value since retention time provides a sufficiently strong feature. For the LR model developed without retention times (Study B), it was noticed that some relevant substances (level 2) had a score <0.8. In this case (Study B), the RF model (which includes additional features) could provide 20 more substances in addition to the LR model (note that using only the RF model was not as efficient as using only the LR model).

The RF model (with the score set at ≥0.85) presented however some limitations when used with the MassBank database (Study C) with the percentage of substances set at level 2 at 57% in POS and 44% in NEG. The MassBank database used presented many substances in POS and particularly in NEG with fragments of low mass (and therefore poorly selective), particularly at 20 eV.

Evaluation of an Automated Validation of Substance Identification. An automated validation process for substance identification was evaluated by defining a scoring threshold depending on the ionization mode and models/databases used (Figure 2). As highlighted previously, the RF model seemed complementary to the LR model. In the positive mode, for the match obtained for both the laboratory and vendor databases, setting the LR thresholds at, respectively, ≥0.88 and 0.89 while ruling out substances with RF < 0.5 allowed one to automatically classify at their highest level of confidence (level 1 for the laboratory database, example of acetaminophen in Figure S10; and level 2 for the vendor database, example of clarithromycin and diltiazem in Figures S11 and S12) all of the substances falling in these settings (representing about 88% of the substances above the set RF threshold). The remaining substances would then have to be evaluated by an expert review.

For the negative mode, the same features (with thresholds LR ≥ 0.90 and RF < 0.5) could be used with the laboratory database with the validation of all substances with these settings (representing 88% of the substances with LR ≥ 0.90).

The same NEG threshold of the LR score at 0.9 was determined for the vendor database. However, using the RF score <0.5 would not work out. It was however noticed that if the feature "number of fragments at 20 eV" was used and the number of fragments >2 was selected, all substances falling in this category were level 2, representing 62% of the 94 substances with the LR score ≥ 0.90. This parameter was not as selective as the RF factor and led therefore to a lower percentage of substances that could be automatically qualified. Substances analyzed in the negative mode do not fragment as much as those analyzed in the positive mode, particularly at 20 eV, which can be considered the most informative fragmentation level. However, even if fragmentation at 40 eV particularly in data-independent analysis produced more interferences, using a secondary collision proved to be complementary to only using one level of fragmentation.

For the RF model developed with the MassBank database, no rules could be determined for automated identification with either POS or NEG modes.

CONCLUSIONS

A machine-learning-based approach has been developed to provide better guidance and aiming at an automated validation process for the identification of substances in the suspect LC HRMS mode. For the first part of the study, using the same HRMS technology, eight classification models were evaluated, and the logistic regression model was selected, which provided robust scoring ranking of identified substances with or without using the retention time feature. In a second part, the study was then continued with the use of an open-source reference database including data from different HRMS technologies. Additional features were included to consider for the larger MS spectra variability, and the random forest model was considered the most efficient. However, this approach was determined to present limitations for a database that comprises data from different HRMS technologies, particularly for substances presenting a low fragmentation MS pattern, which prevented a possible automated validation process.

Overall, this study demonstrates the potential of using machine learning approaches to facilitate the data treatment process and constitutes a first step for an automated process for the identification of environmental substances in LC/HRMS data.

The best approach was determined using both logistic regression and random forest, so further work could include a model combining both algorithms. Other improvements could

also be considered with additional features such as intensity, a better integration of substances with low MS fragmentation, or using a predictive model for retention times or MS fragmentation data. Moreover, application on larger data sets, including spiked samples with analytical standards, will provide better refinement on the boundary and robustness of the developed approaches. Machine learning approaches can also be used to implement a simplified scoring model.[29]

AUTHOR INFORMATION

Corresponding Author

Francois Lestremau − *Hydrosciences Montpellier, Univ Montpellier, IMT Mines Ales, IRD, CNRS, Ales 30100, France;* orcid.org/0000-0002-0959-3477; Email: Francois.lestremau@mines-ales.fr

Authors

Alexandre Levesque − *SIA Partners, Paris 75000, France*

Abdelmoughit Lahssini − *SIA Partners, Paris 75000, France*

Tanguy Magnan de Bornier − *SIA Partners, Paris 75000, France*

Romain Laurans − *SIA Partners, Paris 75000, France*

Azziz Assoumani − *Institut National de l'Environnement Industriel et des Risques (INERIS), Verneuil en Halatte 60550, France;* orcid.org/0000-0002-3774-659X

Hugues Biaudet − *Institut National de l'Environnement Industriel et des Risques (INERIS), Verneuil en Halatte 60550, France*

Author Contributions

F.L.: writing original draft preparation, formal analysis, data contributor, method validation, review and editing. A.L.: writing original draft preparation, formal analysis, machine learning development, software development, review and editing. A.L.: formal analysis, machine learning development, software development. T.M.d.B.: formal analysis, machine learning development, software development, review and editing. R.L.: machine learning development, software development, review and editing. A.A.: writing original draft preparation, formal analysis, review and editing. H.B.: formal analysis, review and editing

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was funded as part of "AMI Intelligence Artificielle" by DITP and DINSIC and from the "France Relance" project by DINUM. The authors acknowledge Agilent for proving the vendor database in the sdf format and authorizing its use for the scope of this project.

REFERENCES

(1) Hollender, J.; van Bavel, B.; Dulio, V.; Farmen, E.; Furtmann, K.; Koschorreck, J.; Kunkel, U.; Krauss, M.; Munthe, J.; Schlabach, M.; Slobodnik, J.; Stroomberg, G.; Ternes, T.; Thomaidis, N. S.; Togola, A.; Tornero, V. High Resolution Mass Spectrometry-Based Non-Target Screening Can Support Regulatory Environmental Monitoring and Chemicals Management. *Environ. Sci. Eur.* **2019**, *31*, No. 42.

(2) Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. Nontarget Screening with High Resolution Mass Spectrometry in the Environment: Ready to Go? *Environ. Sci. Technol.* **2017**, *51*, 11505−11512.

(3) Hollywood, K.; Brison, D. R.; Goodacre, R. Metabolomics: Current Technologies and Future Trends. *Proteomics* **2006**, *6*, 4716−4723.

(4) Antignac, J. P.; Courant, F.; Pinel, G.; Bichon, E.; Monteau, F.; Elliott, C.; le Bizec, B. Mass Spectrometry-Based Metabolomics Applied to the Chemical Safety of Food. *Trends Anal. Chem.* **2011**, *30*, 292−301.

(5) Sangster, T.; Major, H.; Plumb, R.; Wilson, A. J.; Wilson, I. D. A Pragmatic and Readily Implemented Quality Control Strategy for HPLC-MS and GC-MS-Based Metabonomic Analysis. *Analyst* **2006**, *131*, 1075−1078.

(6) Theodoridis, G.; Gika, H. G.; Wilson, I. D. LC-MS-Based Methodology for Global Metabolite Profiling in Metabonomics/Metabolomics. *Trends Anal. Chem.* **2008**, *27*, 251−260.

(7) Ibáñez, M.; Sancho, J. V.; Hernández, F.; McMillan, D.; Rao, R. Rapid Non-Target Screening of Organic Pollutants in Water by Ultraperformance Liquid Chromatography Coupled to Time-of-Light Mass Spectrometry. *Trends Anal. Chem.* **2008**, *27*, 481−489.

(8) Díaz, R.; Ibáñez, M.; Sancho, J. V.; Hernández, F. Target and Non-Target Screening Strategies for Organic Contaminants, Residues and Illicit Substances in Food, Environmental and Human Biological Samples by UHPLC-QTOF-MS. *Anal. Methods* **2012**, *4*, 196−209.

(9) Krauss, M.; Singer, H.; Hollender, J. LC−High Resolution MS in Environmental Analysis: From Target Screening to the Identification of Unknowns. *Anal. Bioanal. Chem.* **2010**, *397*, 943−951.

(10) Bletsou, A. A.; Jeon, J.; Hollender, J.; Archontaki, E.; Thomaidis, N. S. Targeted and Non-Targeted Liquid Chromatography-Mass Spectrometric Workflows for Identification of Transformation Products of Emerging Pollutants in the Aquatic Environment. *Trends Anal. Chem.* **2015**, *66*, 32−44.

(11) Gago-Ferrero, P.; Krettek, A.; Fischer, S.; Wiberg, K.; Ahrens, L. Suspect Screening and Regulatory Databases: A Powerful Combination to Identify Emerging Micropollutants. *Environ. Sci. Technol.* **2018**, *52*, 6881−6894.

(12) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences. *J. Mass Spectrom.* **2010**, *45*, 703−714.

(13) MassBank | MassBank Europe Mass Spectral DataBase. https://massbank.eu/MassBank/ (accessed March 04, 2022).

(14) Oberacher, H.; Sasse, M.; Antignac, J. P.; Guitton, Y.; Debrauwer, L.; Jamin, E. L.; Schulze, T.; Krauss, M.; Covaci, A.; Caballero-Casero, N.; Rousseau, K.; Damont, A.; Fenaille, F.; Lamoree, M.; Schymanski, E. L. A European Proposal for Quality Control and Quality Assurance of Tandem Mass Spectral Libraries. *Environ. Sci. Eur.* **2020**, *32*, No. 43.

(15) Oberacher, H.; Reinstadler, V.; Kreidl, M.; Stravs, M. A.; Hollender, J.; Schymanski, E. L. Annotating Nontargeted LC-HRMS/MS Data with Two Complementary Tandem Mass Spectral Libraries. *Metabolites* **2019**, *9*, No. 3.

(16) enviMass. https://www.envibee.ch/eng/enviMass/overview.htm (accessed March 19, 2022).

(17) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; Vandergheynst, J.; Fiehn, O.; Arita, M. MS-DIAL:

Data-Independent MS/MS Deconvolution for Comprehensive Metabolome Analysis. *Nat. Methods* **2015**, *12*, 523−526.
(18) Helmus, R.; ter Laak, T. L.; van Wezel, A. P.; de Voogt, P.; Schymanski, E. L. PatRoon: Open Source Software Platform for Environmental Mass Spectrometry Based Non-Target Screening. *J. Cheminform.* **2021**, *13*, No. 1.
(19) Giordano, S.; Takeda, S.; Donadon, M.; Saiki, H.; Brunelli, L.; Pastorelli, R.; Cimino, M.; Soldani, C.; Franceschini, B.; di Tommaso, L.; Lleo, A.; Yoshimura, K.; Nakajima, H.; Torzilli, G.; Davoli, E. Rapid Automated Diagnosis of Primary Hepatic Tumour by Mass Spectrometry and Artificial Intelligence. *Liver Int.* **2020**, *40*, 3117−3124.
(20) Ledesma, D.; Symes, S.; Richards, S. Advancements within Modern Machine Learning Methodology: Impacts and Prospects in Biomarker Discovery. *Curr. Med. Chem.* **2021**, *28*, 6512−6531.
(21) Chetnik, K.; Petrick, L.; Pandey, G. MetaClean: A Machine Learning-Based Classifier for Reduced False Positive Peak Detection in Untargeted LC−MS Metabolomics Data. *Metabolomics* **2020**, *16*, No. 117.
(22) Huber, F.; van der Burg, S.; van der Hooft, J. J. J.; Ridder, L. MS2DeepScore: A Novel Deep Learning Similarity Measure to Compare Tandem Mass Spectra. *J. Cheminform.* **2021**, *13*, No. 84.
(23) Togola, A.; Guillemain, C.; Lestremau, F.; Coureau, C.; Margoum, C.; Soulier, C. Applicabilité de La Technique de ≪ screening Non Ciblé ≫ Pour La Surveillance Prospective. *Réseau de Surveillance prospective-AQUAREF ; Rapport BRGM/RP-70108-FR*, 2020.
(24) Adusumilli, R.; Mallick, P.Data Conversion with ProteoWizard MsConvert In *Methods in Molecular Biology*; Springer, 2017; Vol. *1550*, pp 339−368.
(25) Fisher, C. M.; Peter, K. T.; Newton, S. R.; Schaub, A. J.; Sobus, J. R. Approaches for assessing performance of high-resolution mass spectrometry-based non-targeted analysis methods. *Anal. Bioanal. Chem.* **2022**, *414*, 6455−6471.
(26) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48*, 2097−2098.
(27) Pedregosa, F.; Michel, V.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Thirion, B.; Grisel, O.; Dubourg, V.; Passos, A.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(28) Bahel, V.; Pillai, S.; Malhotra, M. In *A Comparative Study on Various Binary Classification Algorithms and Their Improved Variant for Optimal Performance*, 2020 IEEE Region 10 Symposium, TENSYMP; IEEE: Dhaka, Bangladesh, 2020; pp 495−498.
(29) Alygizakis, N.; Lestremau, F.; Gago-Ferrero, P.; Gil-Solsona, R.; Arturi, K.; Hollender, J.; Schymanski, E. L.; Dulio, V.; Slobodnik, J.; Thomaidis, N. Towards a Harmonized Identification Scoring System in LC-HRMS/MS Based Non-Target Screening (NTS) of Emerging Contaminants. *TrAC, Trends Anal. Chem.* **2023**, *159*, No. 116944.