Prediction of a Large-Scale Database of Collision Cross-Section and Retention Time Using Machine Learning to Reduce False Positive Annotations in Untargeted Metabolomics.

Marie Lenski, Saïd Maallem, Gianni Zarcone, Guillaume Garçon, Jean-Marc Lo-Guidice, Sébastien Anthérieu, Delphine Allorge

metabolites

MDPI

Article

Prediction of a Large-Scale Database of Collision Cross-Section and Retention Time Using Machine Learning to Reduce False Positive Annotations in Untargeted Metabolomics

Marie Lenski [1,2,*], Saïd Maallem [1], Gianni Zarcone [1], Guillaume Garçon [1], Jean-Marc Lo-Guidice [1], Sébastien Anthérieu [1] and Delphine Allorge [1,2]

[1] ULR 4483, IMPECS—IMPact de l'Environnement Chimique sur la Santé humaine, CHU Lille, Institut Pasteur de Lille, Université de Lille, F-59000 Lille, France
[2] CHU Lille, Unité Fonctionnelle de Toxicologie, F-59037 Lille, France
* Correspondence: marie.lenski@univ-lille.fr

Abstract: Metabolite identification in untargeted metabolomics is complex, with the risk of false positive annotations. This work aims to use machine learning to successively predict the retention time (Rt) and the collision cross-section (CCS) of an open-access database to accelerate the interpretation of metabolomic results. Standards of metabolites were tested using liquid chromatography coupled with high-resolution mass spectrometry. In CCSBase and QSRR predictor machine learning models, experimental results were used to generate predicted CCS and Rt of the Human Metabolome Database. From 542 standards, 266 and 301 compounds were detected in positive and negative electrospray ionization mode, respectively, corresponding to 380 different metabolites. CCS and Rt were then predicted using machine learning tools for almost 114,000 metabolites. R^2 score of the linear regression between predicted and measured data achieved 0.938 and 0.898 for CCS and Rt, respectively, demonstrating the models' reliability. A CCS and Rt index filter of mean error ± 2 standard deviations could remove most misidentifications. Its application to data generated from a toxicology study on tobacco cigarettes reduced hits by 76%. Regarding the volume of data produced by metabolomics, the practical workflow provided allows for the implementation of valuable large-scale databases to improve the biological interpretation of metabolomics data.

Keywords: mass spectrometry-based metabolomics; ion mobility-mass spectrometry; metabolomics data analysis; machine learning; collision cross-section; retention time

Citation: Lenski, M.; Maallem, S.; Zarcone, G.; Garçon, G.; Lo-Guidice, J.-M.; Anthérieu, S.; Allorge, D. Prediction of a Large-Scale Database of Collision Cross-Section and Retention Time Using Machine Learning to Reduce False Positive Annotations in Untargeted Metabolomics. *Metabolites* **2023**, *13*, 282. https://doi.org/10.3390/metabo13020282

Academic Editors: Sarah Wille and Andrea E. Steuer

Received: 9 January 2023
Revised: 7 February 2023
Accepted: 12 February 2023
Published: 15 February 2023

1. Introduction

The metabolome of a biological system is influenced by physiological, pathological, or environmental conditions [1]. As it gathers the final products of the cellular process, the exhaustive measurement of metabolite changes by metabolomics provides dynamic and sensitive information closely linked to its phenotype. Investigating the metabolome allows the identification of metabolic fingerprints that can then be used as biomarkers and/or provides new mechanistic perspectives leading to a particular phenotype [2,3]. Recent developments in mass spectrometry (MS) technology, informatics, and analytical chemistry have made it possible to comprehensively analyze the metabolome [4,5] with a high level of sensitivity [6] compared to nuclear magnetic resonance-based metabolomics. Additionally, high throughput analyses like high-resolution MS in full scan mode allow the rapid achievement of large-scale studies of hundreds of compounds (untargeted metabolomics), which is an evident benefit over analyses that target a restricted number of metabolites or pathways (targeted metabolomics).

Untargeted metabolomics is a multistep process involving first sample collection, preparation, and analysis that generates data, then data processing and chemometrics that generate a candidate list of features, and finally, metabolite identification [7]. This final step

gives biological meaning to MS data [8]. A consensus by the Chemical Analysis Working Group of the Metabolomics Standards Initiative (MSI) reported different levels of confidence in the annotation depending on the method of identification used [9]. A confident and definitive identification (level 1) is hit when two or more orthogonal properties fit with data from authentic standard compounds in identical analytical settings. When the latter are unavailable, a comparison of experimental data with public libraries could lead to a putative compound annotation (level 2) or class annotation (level 3). Finally, unknown features discriminated with spectral data are classified at the lowest confidence level (level 4). Therefore, feature annotation is achieved by comparing experimental measurements to existing in-house or external databases of known metabolites to generate potential candidates [10]. Several commercial or open-source databases containing spectral data in libraries (Human Metabolome Database (HMDB) [11], Metlin [12] . . .) were developed and are continuously updated by the scientific community. However, confident and unequivocal structure identification could quickly be an issue when a candidate is not found (limited number of spectra) or when several candidates are proposed (false positives), increasing the probability of misidentification [13,14]. Therefore, it becomes important to use other readily obtained physicochemical properties for better metabolite identification.

Ion mobility-mass spectrometry (IMS-MS) is a fast two-dimensional separation of ions based on their mobility in a buffer gas. Importantly, this mobility is structure-dependent and is not affected by equipment or experimental factors (matrix effects, variations in mobile phase composition, and chromatography settings, ionization mode, acquisition settings . . .), unlike retention time (Rt) and mass spectrum [15], resulting in a high degree of repeatability and therefore facilitating database queries [16]. The physical property measured in IMS-MS is the collision cross-section (CCS). Thus, it provides the orthogonal separation to improve signal-to-noise, resolution, and isomeric metabolite separation [17], participating in the reduction of misidentification. However, the favorable contribution of CCS is currently limited by the poor availability of CCS reference values [18–22]. New experimental and computational approaches to predict those parameters for a large number of compounds is highly valuable. Several studies have developed, or applied machine learning-based prediction approaches [23]. Softwares like AllCCS [22], CCS Predictor [24], DeepCCS [25], MetCCS Predictor [26], or LipidCCS predictor [27] can efficiently generate a model when molecular descriptors are provided [15]. Molecular descriptors are numeric information generated by mathematical treatment of compound structures that characterize the physico–chemical properties of metabolites (ex: polarity, LogP . . .) [28]. In contrast, CSSBase is a web interface (https://CCSbase.net) (accessed on 6 May 2022) that provides access to a ready-to-use predictive model, allowing rapid prediction of CCS values directly from SMILES structures (Simplified Molecular Input Line Entry System representation), using a cluster-based prediction model [29]. This platform allows a broad coverage of chemical structure diversity and can thus be easily used in existing metabolomics workflows.

The Rt of a compound is defined by its chemical interactions with the chosen mobile phase and stationary phase. Metabolite retention can be improved by optimizing solvent gradient elution, nature, and dimensions of the chromatographic column or chromatographic settings [30]. Rt is often decisive in feature annotation but usually relies on the availability of authentic chemical standards that are applied to experimental conditions. In untargeted metabolomics, the transferability of Rt database between laboratories is not achievable because of the absence of standardized assays across different laboratories. Multiple machine learning models for the prediction of Rt have already been described, including quantitative structure–retention relationship (QSRR) models [31–35]. QSRR strategies have been used to accelerate the method development process by comparing predicted separation with different columns [36] or to enhance the confidence of identifications [37]. Software packages, such as the QSRR Automator [38], exist to automate Rt prediction model creation. Structure and chromatographic data from known metabolites, obtained from their SMILES and from chemical standards analyzed using a particular LC method, are used to generate a model. It identifies relations between chromatographic reten-

tion and the molecular descriptors, theoretically allowing to predict Rt for any metabolite whose molecular descriptors can be calculated [39].

In the present work, we aim to describe the workflow permitting the generation of a large-scale in-house database of Rt and CCS predicted with published machine learning models. Integration of these data with other sources of information, such as accurate mass, MSe fragmentation, and isotope pattern for facilitating the identification of compounds, is illustrated in an application to toxicology data.

2. Materials and Methods

2.1. Chemicals and Standards

Solutions used were: acetonitrile (UPLC-MS grade, Waters, Milford, MA, USA), methanol (UPLC-MS grade, Waters), Milli-Q purified water (Millipore, Burlington, MA, USA), formic acid (UPLC-MS grade, Honeywell, Charlotte, NC, USA), ammonium formate (Reagent-grade, Sigma-Aldrich, St. Louis, MO, USA), and chloroform (VWR Chemicals, Radnor, PA, USA). Chemical standards (MSMLS) were purchased from Sigma-Aldrich. This library was chosen for the broad chemical and functional diversity of metabolites included. It contains 634 standard metabolites sampled into seven 96-well plates at 5 µg per well, including 37 duplicates. An associated spreadsheet with information, such as metabolite identification, molecular formula, and SMILES was used to build our targeted database. The compounds were dissolved using two different solutions (5% methanol for plates 1–5 and chloroform:methanol:water 1:1:0.3 for plates 6–7) following the manufacturer's instructions to obtain a 20 µg mL^{-1} concentration. Stock solutions were pooled with a maximum of 12 compounds to obtain 56 solutions at 1.6 µg mL^{-1} to perform simple multiplex injections for LC-MS analysis.

2.2. LC-MS Conditions

Analyses were conducted on a liquid chromatograph system coupled to high-resolution mass spectrometry (LC-HRMS). Chromatographic separation was obtained with the following characteristics: Instrument: Acquity UPLC I-Class system (Waters); column: Acquity UPLC HSS T3 (1.8 µm, 150 × 2.1 mm; Waters); column temperature: 50 °C; flow rate: 0.4 mL min^{-1}; autosampler temperature: 10 °C; volume of injection: 15 µL. Separation was performed in a gradient elution mode. Mobile phases for the multistep gradient in the positive mode were solution A: aqueous solution of ammonium formate (3 mM) with 0.1% formic acid and solution B: acetonitrile with 0.1% formic acid (v/v). The elution gradient was: 100% A for 1 min, 0–1% B for 1 min, 1–3% B for 2 min, 3–99% B for 13 min, 99% B for 3 min, 99–0% B for 0.5 min, and 100% A for 2.5 min. Detection was performed on a Vion IMS-QToF mass spectrometer (Waters) with the following settings: ionization source: electrospray operating in positive (ESI+) and negative (ESI-) modes; source temperature: 120 °C; desolvation temperature: 600 °C; cone gas flow: 50 L h^{-1}; desolvation gas flow: 1000 L h^{-1}; capillary voltage: 0.5 kV in ESI+ or 2 kV in ESI-; m/z range: 50–1000; scan time: 0.25 s; lock mass reference: leucine enkephalin (m/z 556.2766) solution at 200 ng mL^{-1}; infusion intervals: 5 min; acquisition mode: high-definition MSe; low collision energy: 6 V; high collision energy ramp: 14–56 V; IMS drift gas and collision gas: nitrogen; ion mobility and mass calibrations solution: Major Mix IMS/ToF Calibration Kit (Waters). These parameters allow for achieving a mass resolving power of >20,000 FWHM.

Data analysis of the mixes was semi-automatically performed through the Unifi software (version 1.9.4.053 Waters MS Technologies, Manchester, UK) to obtain the Rt, response, and CCS of the standards after manual verification of the peak integration. Adducts considered were [M+H]+, [M+K]+, [M+Na]+, [M+Cl]−, [M+HCOO]−, and [M+CH3OO]−. Accurate mass, Rt, CCS, and fragmentation patterns were used to build a targeted database of LC and MS properties.

2.3. CCS Prediction

CCSBase is an electronic interface (https://CCSbase.net) (accessed on 6 May 2022) for accessing the CCS predictive model [29]. It calculates the predictive CCS values of adducts using SMILES. Original performances of CCSBase were described by Ross et al. [29] with an R^2 score, a mean absolute error, and a root mean squared error at 0.991, 3.83 $Å^2$, and 5.48 $Å^2$, respectively. In the study, all adducts considered by CCSBase, namely [M+H]+, [M+K]+, [M+Na]+, [M+Na-2H]−, [M+NH4]+, [M]+, [M-H]−, and [M]−, were taken in account. A batch prediction was performed for a dataset of metabolite structures freely available in HMDB v4.0 [11], which gathers up to 114,000 human metabolites, covering the majority of untargeted metabolomic data sets. A linear regression was performed comparing predicted and measured CCS of adducts from standard compounds. Measured CCS were included in this comparison if the standard were listed in HMDB, and if they presented common adducts with predictions. The coefficient of determination R^2 between the predicted and the experimental CCS data and mean absolute error permitted to evaluate the model. The best fit of linear regression was calculated, with an interval of +/− 2 standard deviations (SD). All statistical analyses and figure production of this manuscript were conducted under R language and environment [40].

2.4. Rt Prediction of Small Molecules

QSRR Automator [38] builds regression retention models. Based on their SMILES, chemical structures were converted into their numerical representation by expressing them through structural descriptors produced by informatic algorithms of QSRR Automator. First, using a defined training data set, the machine learning algorithm learns the "rule" between molecular descriptors and their experimental Rt values to establish prediction models and select the best model. QSRR algorithm identifies descriptors that positively impact model performance. Selection and optimization of regression algorithms were carried out by automated procedures and evaluated thanks to the R^2 score and mean absolute error. Then, the external validation data set is used to validate and evaluate the prediction error. Cross-validation (n = 10) provides an estimate of the accuracy of the Rt prediction for compounds that were not used in its development or optimization, evaluated thanks to the R^2 score, mean absolute error, and SD. Once a valid model was selected, Rt predictions were performed for metabolites from the HMDB v4.0. The best fit of linear regression was calculated with an interval of +/− 2 SD.

2.5. Reduction of the Occurrence of False Positive Annotations in Untargeted Metabolomics: Application to Toxicology Data

We analyzed LC-HRMS data from an ongoing study assessing the potential toxicity of tobacco cigarette fumes on human bronchial epithelial BEAS-2B cells to demonstrate the relevance of the predicted large-scale database of collision cross-section and retention time to metabolomics. The exposure protocol was adapted from Dusautoir et al. [41]. Briefly, BEAS-2B cells cultured at air–liquid interface was exposed to four puffs of tobacco cigarette emissions or to sterile air (negative control) in four replicates per exposure. Twenty-four hours after exposure, cell metabolism was quenched by the addition of ice-cold methanol:water (80:20, v/v) mixture. Cells were harvested using a cell scraper. Deproteinization was performed by adding the same methanolic mixture, vortexing, and centrifuging at 14,000× g at +4 °C for 15 min. Supernatants were concentrated to dryness with speedvac and reconstituted before injection in a water:methanol (90:10, v/v) mixture. After metabolomic analyses, LC-HRMS data were analyzed with Progenesis QI (Nonlinear Dynamics, UK) for feature extraction. Data normalization and statistical analyses were conducted under the R environment [40] on the features detected in ESI+ and ESI−. When searching against HMDB, two identification strategies were evaluated: (1) with an m/z (tolerance set at 5 mDa), isotope and fragmentation match only, and (2) with an m/z, isotope, fragmentation, CCS, and Rt match of the created predicted large-scale database.

3. Results

3.1. Analysis of Standard Compounds and Generation of an In-House Database

A total of 542 standards were originally used in the candidate database. Each set of data was manually examined for errors. Figure 1 describes the workflow used for the targeted database construction. The candidate sorting step has allowed the detection of 266 and 301 compounds in ESI+ or ESI−, respectively, corresponding to a total of 380 different metabolites. Accurate mass, Rt, CCS, and fragmentation were used to build a targeted database of LC and MS properties.

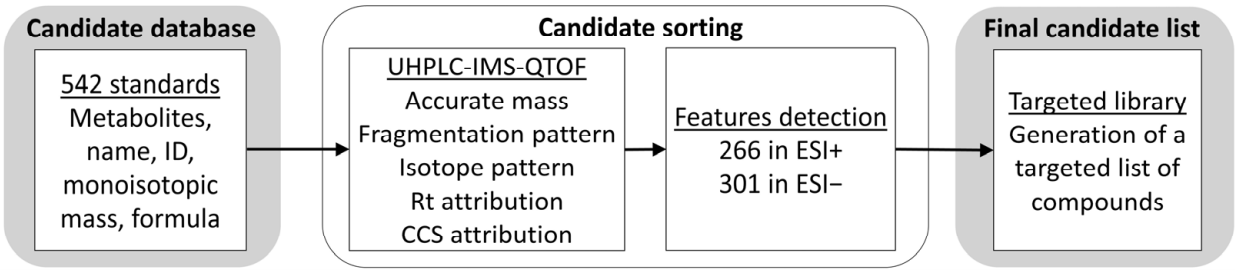

Figure 1. Workflow and results of the targeted-based metabolomic approach developed.

Metabolites belong to different chemical classes and are mainly carbohydrates, carboxylic acids, lipids, nucleotides, and organoheterocyclic compounds (Figure 2) involved in different pathways that reflect the metabolic status in biological matrices of interest.

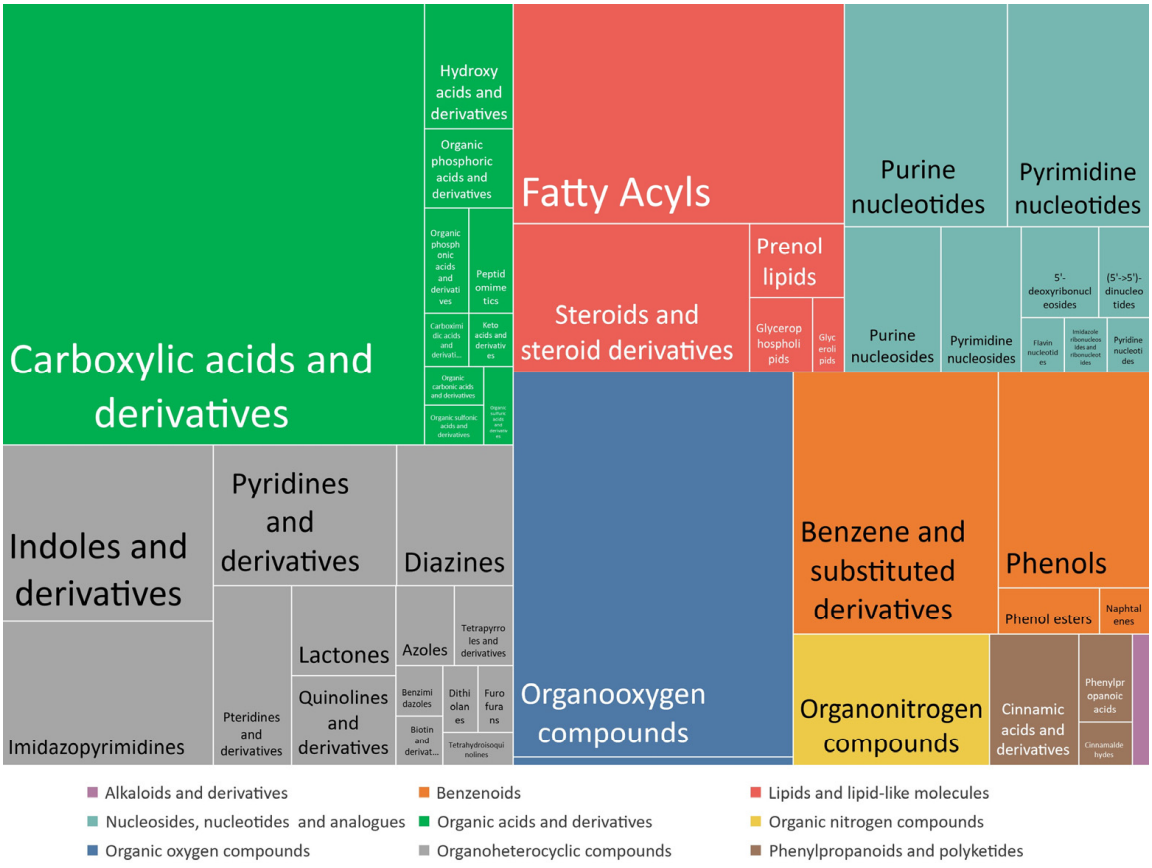

Figure 2. Chemical taxonomy of metabolites included in the targeted library and proportion of each superclass and class.

3.2. CCS Prediction and Creation of a CCS Database

The CCS database was generated according to the workflow described in Figure 3. Predictions were performed for almost 114,000 metabolites from the HMDB v4.0 database, generating 916,104 CCS adduct values. Results were validated with a validation set composed of 501 measured CCS adduct values from 297 standard compounds in both ionization modes.

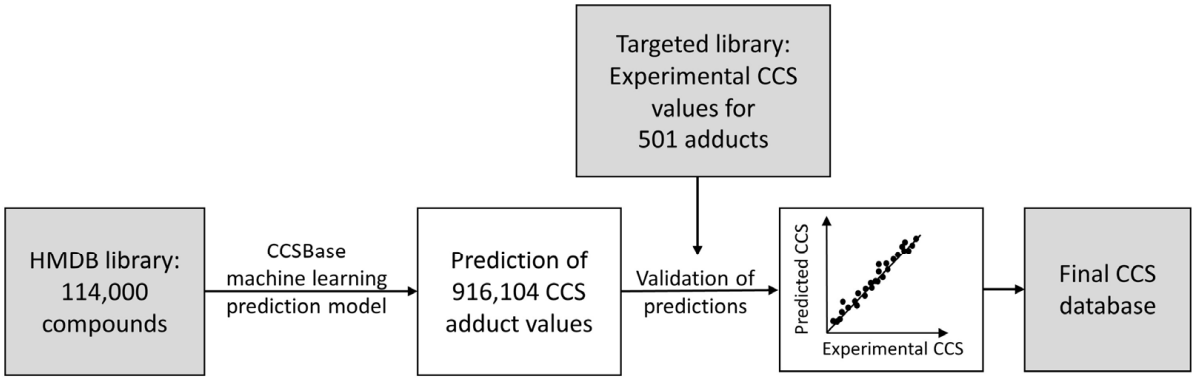

Figure 3. Workflow of CCS prediction.

Figure 4 illustrates the match between predicted and experimentally determined CCS. Outliers were kept as part of the data set in the absence of any evidence that they were the result of an error. The R^2 score of the linear regression achieves 0.938 and the mean absolute error was calculated at 3.94 Å^2, while the SD reaches 6.11 Å^2 or 3.36%. The predicted CCS = 0.95 × measured CCS + 7.92. The resulting output table of CCS allowed us to build our large-scale in-house reference database.

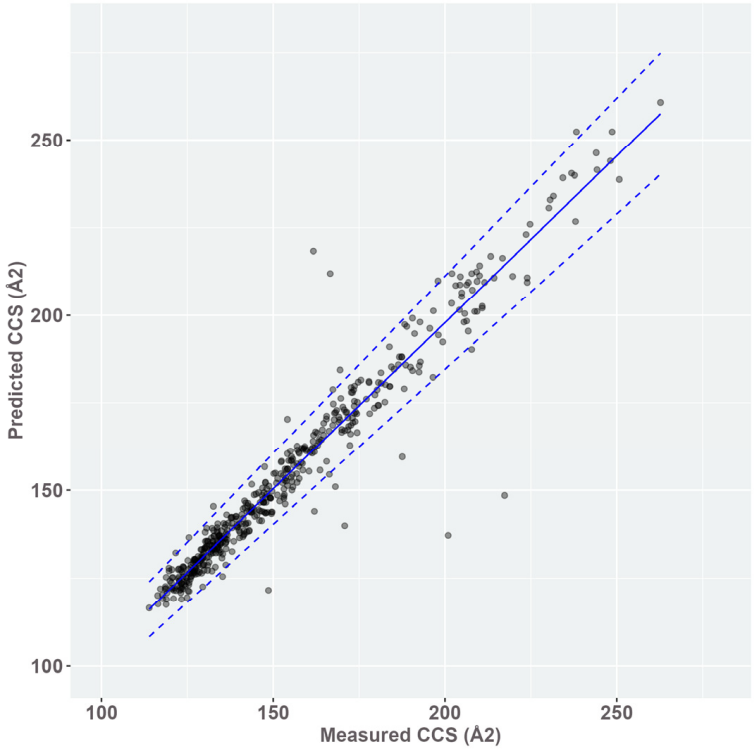

Figure 4. Predicted CCS (Å^2) using CCSBase algorithm [29] by measured CCS (Å^2). Validation was performed on the CCS of 501 adducts from 297 standard compounds measured in ESI− and ESI+. Linear regression: adjusted R^2 = 0.938. Blue line: best fit of linear regression. Dashed blue lines best fit +/− 6.72% (2 SD). Predicted CCS = 0.95 × measured CCS + 7.92.

3.3. Rt Prediction and Creation of an Rt Database

The Rt prediction workflow is described in Figure 5. In total, 204 compounds from the developed method were selected for the QSRR model; 114 were detected in both ESI+ and ESI−, while 90 were detected only in one ionization mode (45 for each ionization mode). Seven compounds were excluded due to incomplete data in molecular descriptors. Support vector regression (SVR) algorithm based on 113 molecular descriptors presented the best performances, with the R^2 score at 0.999 and the mean absolute error at 0.10 min for the training set. The validation set tested by cross-validation (n = 10) validated the model with the following performances: mean of cross-validation R^2 score 0.898, mean absolute error 0.81 min, and standard deviation of the mean absolute error 0.15 min. Detailed results are presented in Table S1 and Figure S1. Rt predictions were performed for almost 114,000 metabolites from HMDB v4.0. The resulting output table of Rt allowed us to build our large-scale in-house reference database.

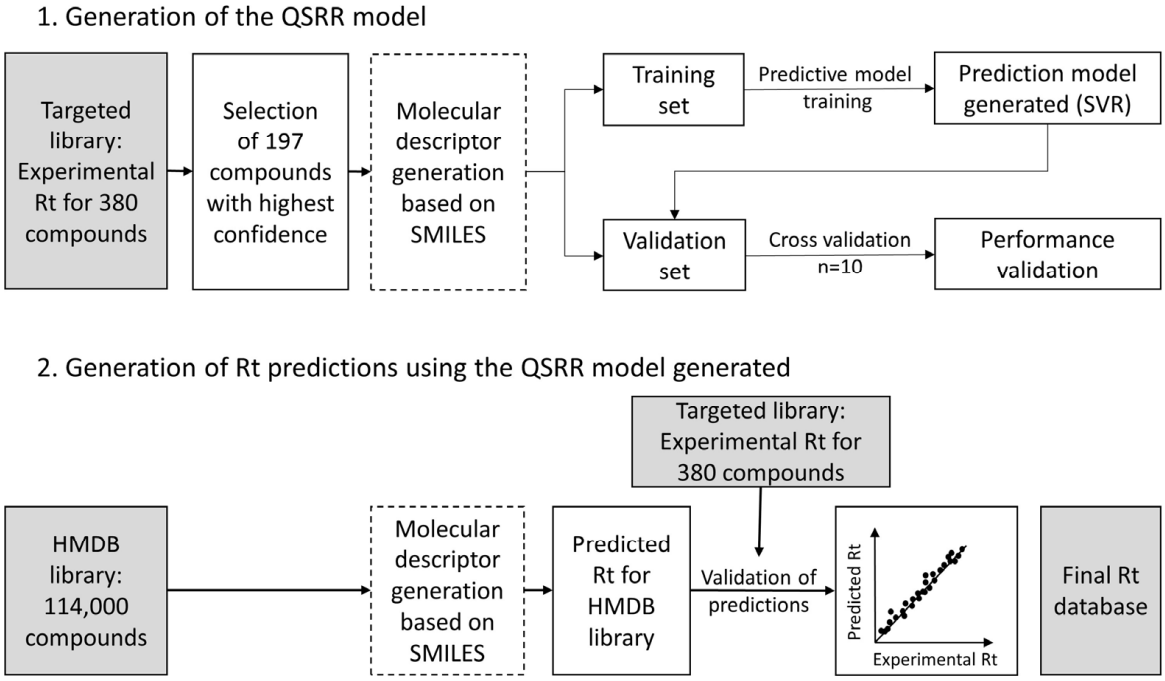

Figure 5. Workflow of Rt prediction.

3.4. Reduction of the Occurrence of False Positive Annotations in Untargeted Metabolomics: Application to Toxicology Data

We assessed the potential toxicity associated to tobacco cigarette fumes on the human bronchial epithelial BEAS-2B cells using metabolomics. Among the 3591 features detected in ESI+ and ESI−, 51 features were significantly deregulated by cigarette smoke compared to controls and needed to be identified. As illustrated in Figure 6a, 46 out of 51 features had one hit or more (90%). The number of hits exceeded 10 hits for the major part of the features. For the method combining m/z, CCS, and Rt match search, CCS and Rt match tolerances were set at 16 $Å^2$ and 1.1 minutes, respectively, according to the determined CCS and Rt index filter expressed as mean error +/− 2 SD. Only 37 out of 51 features (72%) had one or more metabolite hits. Seventy-six percent of hits were filtered using the predicted large-scale database (Figure 6b). The percentage of features with only one hit significantly increased with the additional CCS and Rt match (+53%), while the percentage of features with more than 10 hits decreased in the same conditions (−39%) (Figure 6c). For further identification, possible candidates for each compound are ranked by Progenesis QI on an overall score based on the m/z match, isotope similarity, fragmentation score, CCS, and Rt error (data not shown).

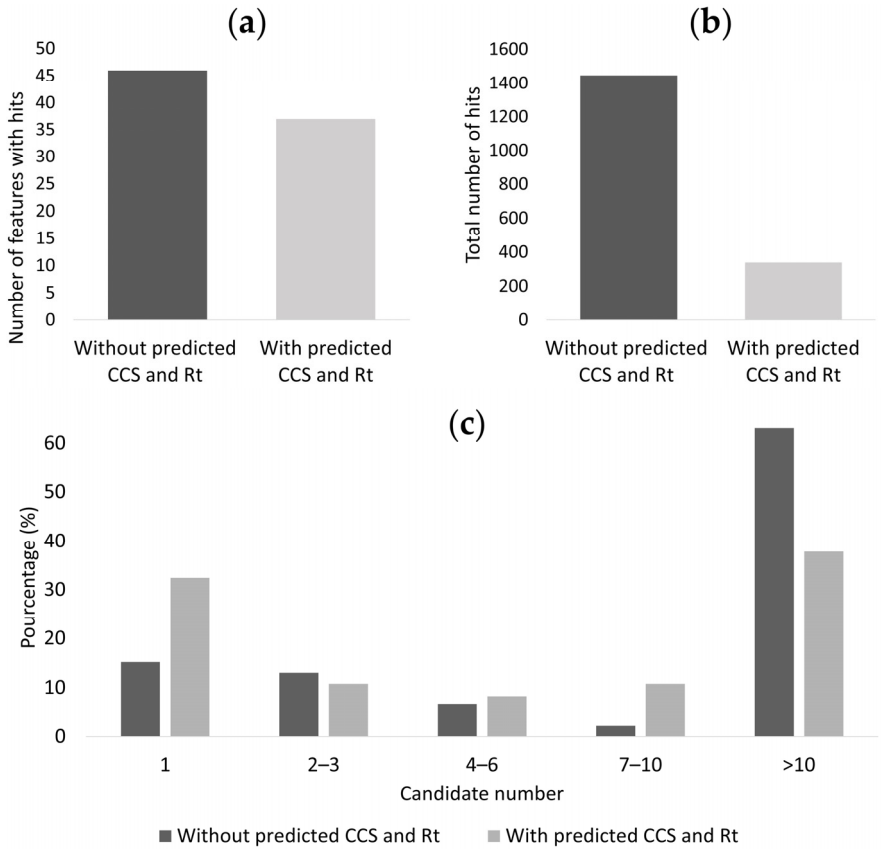

Figure 6. Contribution of the predicted CCS and Rt database to metabolite identification in untargeted metabolomics applied to toxicology. From the two match methods: (**a**) number of features with hits, (**b**) total number of hits (**c**), percentage distribution of features with different metabolite hits.

4. Discussion

LC-HRMS is an uncontestably powerful analytical approach employed in both targeted and untargeted metabolomics. We used an UPLC-IMS-QTOF to create an accurate in-house database using a commercial library of metabolite standards. Based on experimental results, we used two existing prediction models (CCSBase and QSRR Automator) to predict CCS and Rt values of a large-scale database to increase confidence in metabolite annotation. Associating Rt and CCS is relevant as it can provide complementary information coming from chromatographic and ion mobility separation or even replace other orthogonal properties (isotope similarity and fragmentation score) for putative compound annotation. Moreover, for all metabolites, predictions were performed for protonated and deprotonated ions as well as adducts, each having the same Rt but a different CCS [42]. The co-occurrence of adducts is common when analyzing heterogeneous biological samples [43]. Gathering predictions for multiple metabolic features of the same metabolite is valuable information, allowing for cross-validation of identification. The predicted database is presented in Table S2.

The comparison of mass spectra, Rt, CCS, and accurate mass of a feature with experimental data acquired from standard compounds measured under the same analytical settings permits achieving the highest level of identification confidence. An in-house database of 380 different metabolites was generated, allowing a confident and definitive identification (level 1 according to the MSI [9]). Metabolites were separated using an Acquity UPLC HSS T3 column that possesses superior polar-compound retention and aqueous mobile phase compatibility compared to more classical stationary phases. It could be used for the retention of mid-polar to apolar analytes. The Sigma library used contains several polar metabolites that cannot be retained in those analytical conditions, explaining

the lack of detection for some metabolites [44]. Moreover, with up to 114,000 chemicals deposited in HMDB, only a few percent of these compounds could be covered with authentic standards. Therefore, structure identification in untargeted metabolomics analyses remains a significant challenge. By predicting chromatographic Rt and CCS from experimentally acquired data, this targeted library represents a starting point to potentially give access to a detailed sample composition for future untargeted metabolomic studies. Using this methodology, we were able to drastically expand the number of metabolites at level 2 or level 3 annotations [9].

Predictions using machine learning are data-driven approaches providing predictions for metabolites with corresponding properties [45]. After the training of the model, predictions could be generated immediately for other compounds.

For CCS predictions, CCSBase is a machine learning-based prediction model built from a combined database, enabling to cover an important variety of structural compounds, participating in the transferability of this model. Indeed, large-scale CCS predictions were validated with our experimental data with a low bias and a high R^2 at 0.938. A CCS index filter defined as mean error +/− 2 SD, i.e., maximum 16.16 $Å^2$, could be used as the threshold for excluding false positives. This match tolerance, reflecting the deviation of analytes or family of analytes or type of adducts, is relatively large compared to other work demonstrating that median relative errors as low as 3 to 5% are reachable using other models [22,24–27]. However, excluding false positive identifications with a CCS match higher than the defined threshold remains of great importance when considering the number of possible matches when using m/z match, isotope similarity, and fragmentation score only. Moreover, this additional separation process participates in better detection of compounds presenting contaminant mass spectra due to the co-elution or a poor abundance.

For Rt predictions, the workflow was different as we trained and validated an accurate machine-learning model based on compounds with various physicochemical properties. The training set allowed the model to be trained, while the test set made of unknown data for the trained model allowed the model to be validated. With this strategy, the model was estimated with small error differences in favour of minimum overfitting. An Rt index filter defined as mean error +/− 2 SD, i.e., maximum 1.11 min, could be used as a threshold for eliminating the majority of misidentified compounds. Outliers could be due to software bias, random noise in the data used or errors in the attribution of standards. Naylor et al. described the QSRR Automator's original performances on various chromatographic columns. They showed errors in predictions within 1 min for the majority of predictions, and within 2 min for almost all predictions [38]. QSRR here performs comparably to previously published methods [34,36]. In relatively short run time methods, as in our method, many metabolites have very close Rt, including isomers with Rt that fall within 10 s of each other. Our database does not permit the distinction between those metabolites but is adequate to differentiate between clearly separated compounds of the same mass and reduce false positive identifications, leading to an advanced biological interpretation of results. Even if the generated model was based on compounds with various physicochemical properties separated and identified with an optimized method, particular attention should be dedicated to avoiding inaccurate results, including (i) compounds not retained in the column (ii) compounds retained after the observed Rt of the training data (iii) compounds with physicochemical properties that differ from the training set. For example, the in-house database presented here was generated from a large variety of chemical standards dedicated to metabolomics analyses but did not include complex high molecular weight compounds. Biased predictions for those metabolites should be excluded. The Rt database that we created is strictly related to our chromatographic conditions, so it can be directly useful only for those who decide to strictly adapt our choice of column, mobile phases, and flow rates. In reversed-phase chromatography, authors suggested that Rt from a defined method can be projected in other chromatographic settings as soon as the elution order of metabolites is preserved [46]. Most of the time, laboratories employ a distinct chromatographic setting depending on the separation required. We here presented

a practical workflow, with the objective of generating QSRR models and predictions for every set of LC conditions.

Some limitations of the present study must be mentioned. The number of compounds is relatively small, resulting in a limited number of experimental data that could influence the performance of models. However, this limitation is counterbalanced by the quality of data since we included data from authentic standards with the highest confidence possible. The resulting performances could have been further validated by performing a side-by-side comparison with other existing machine-learning tools. Such a comparison has already been described elsewhere [22,24–27]. Instead of that, the chosen strategy consisted of emphasizing the usefulness of our workflow with concrete application on biological data. Finally, by associating Rt predictions with CCS predictions, the generated large-scale database is strictly related to the instrumental configuration, but the workflow could be largely generated to other experimental conditions.

Most prediction models or workflows previously reported discuss one or the other predicted property (CCS or Rt), while only a few associate multi-dimensional information for metabolite annotation [47–50]. Interestingly, all of them are dedicated to lipids or exogenous compounds, while our workflow predicted a database including small molecules found in the human body, including water-soluble or lipid-soluble endogenous metabolites and exogenous compounds. Regarding the tremendous interest of the scientific community in metabolomics, providing a practical workflow is of large importance for analytical chemists or biologists who cannot develop machine learning models but who want to improve the biological interpretation of their metabolomics data. The usefulness of our combined large-scale predicted database was demonstrated with an application of biological data generated from a toxicology study on tobacco cigarettes. The results demonstrated that the introduction of CCS and Rt values for metabolite identification could significantly reduce false positive identifications, with the benefit of narrowing the search scope and improving the identification accuracy.

5. Conclusions

In this study, a workflow was introduced to remove false positive annotations in non-targeted metabolomics studies. The procedure includes the implementation of a combined CCS and Rt-restricted database starting from a commercial library of metabolite standards. This experimental database has been used to predict CCS and Rt of a large-scale dataset using existing machine learning tools. As illustrated by an application on a metabolomic study on tobacco cigarette toxicity, the presented workflow reduces the occurrence of false positive annotations in untargeted metabolomics and adds confidence to the identification of metabolites. This database has been integrated into the protocol used in our laboratory for untargeted metabolomics analyses and is freely downloadable. When making the assumption that the created database could be a representative subset of compounds present in the human metabolome, biological interpretation of metabolomics data is notably improved, giving new insights into biomarker research or mechanisms that generate a specific phenotype. We suggest using our data as a methodological starting point for the development of a large-scale in-house reference database based on artificial intelligence tools, providing a practical and effective workflow to improve the predictive confidence of metabolomic studies at a large-scale level.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/metabo13020282/s1, Figure S1: Performances of the SVR model using QSRR Automator algorithm; Table S1: Linear regression of Predicted Rt (min) using QSRR Automator algorithm by measured Rt (min) of compounds used from the developed method. This figure was generated by QSRR Automator from all the data available for model construction (including training data); Table S2: Large-scale in-house reference database.

Author Contributions: Conceptualization, M.L. and S.M.; methodology, S.M. and M.L.; investigation, M.L., S.M., G.Z., S.A. and D.A.; resources, D.A., G.G. and J.-M.L.-G.; writing—original draft preparation, M.L.; writing—review and editing, S.A., D.A., G.G., G.Z. and J.-M.L.-G.; visualization, S.M. and M.L.; supervision, D.A. and S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: Graphical abstract was modified from SMART (Servier Medical Art). http://smart.servier.com (accessed on 6 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

CCS	collision cross-section
ESI	electrospray ionization
HMDB	human metabolome database
HRMS	high-resolution mass spectrometry
IMS	ion mobility spectrometry
LC	liquid chromatography
m/z	mass to charge ratio
MS	mass spectrometry
MSI	metabolomics standards initiative
MSMLS	mass spectrometry metabolite library of standards (Sigma-Aldrich)
QSRR	quantitative structure retention relationships
QTOf	quadrupole time-of-flight
Rt	retention time
SD	standard deviation
SMILES	simplified molecular input line entry system
SVR	support vector regression
UHPLC	ultra high-performance liquid chromatography

References

1. Roessner, U.; Bowne, J. What Is Metabolomics All About? *BioTechniques* **2009**, *46*, 363–365. [CrossRef] [PubMed]
2. Beger, R.D.; Dunn, W.; Schmidt, M.A.; Gross, S.S.; Kirwan, J.A.; Cascante, M.; Brennan, L.; Wishart, D.S.; Oresic, M.; Hankemeier, T.; et al. Metabolomics Enables Precision Medicine: "A White Paper, Community Perspective". *Metabolomics Off. J. Metabolomic Soc.* **2016**, *12*, 149. [CrossRef] [PubMed]
3. Trifonova, O.P.; Maslov, D.L.; Balashova, E.E.; Lokhov, P.G. Current State and Future Perspectives on Personalized Metabolomics. *Metabolites* **2023**, *13*, 67. [CrossRef]
4. Ma, X. Recent Advances in Mass Spectrometry-Based Structural Elucidation Techniques. *Molecules* **2022**, *27*, 6466. [CrossRef] [PubMed]
5. Zarrouk, E.; Lenski, M.; Bruno, C.; Thibert, V.; Contreras, P.; Privat, K.; Ameline, A.; Fabresse, N. High-Resolution Mass Spectrometry: Theoretical and Technological Aspects. *Toxicol. Anal. Clin.* **2022**, *34*, 3–18. [CrossRef]
6. Patti, G.J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: The Apogee of the Omics Trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13*, 263–269. [CrossRef] [PubMed]
7. Barnes, S.; Benton, H.P.; Casazza, K.; Cooper, S.J.; Cui, X.; Du, X.; Engler, J.; Kabarowski, J.H.; Li, S.; Pathmasiri, W.; et al. Training in Metabolomics Research. II. Processing and Statistical Analysis of Metabolomics Data, Metabolite Identification, Pathway Analysis, Applications of Metabolomics and Its Future. *J. Mass Spectrom. JMS* **2016**, *51*, 535–548. [CrossRef]
8. Nash, W.J.; Dunn, W.B. From Mass to Metabolite in Human Untargeted Metabolomics: Recent Advances in Annotation of Metabolites Applying Liquid Chromatography-Mass Spectrometry Data. *TrAC Trends Anal. Chem.* **2019**, *120*, 115324. [CrossRef]
9. Sumner, L.W.; Amberg, A.; Barrett, D.; Beale, M.H.; Beger, R.; Daykin, C.A.; Fan, T.W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J.L.; et al. Proposed Minimum Reporting Standards for Chemical Analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics Off. J. Metabolomic Soc.* **2007**, *3*, 211–221. [CrossRef]

10. Yi, Z.; Zhu, Z.-J. Overview of Tandem Mass Spectral and Metabolite Databases for Metabolite Identification in Metabolomics. *Methods Mol. Biol. Clifton NJ* **2020**, *2104*, 139–148. [CrossRef]
11. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [CrossRef] [PubMed]
12. Guijas, C.; Montenegro-Burke, J.R.; Domingo-Almenara, X.; Palermo, A.; Warth, B.; Hermann, G.; Koellensperger, G.; Huan, T.; Uritboonthai, W.; Aisporna, A.E.; et al. METLIN: A Technology Platform for Identifying Knowns and Unknowns. *Anal. Chem.* **2018**, *90*, 3156–3164. [CrossRef] [PubMed]
13. Wen, Y.; Amos, R.I.J.; Talebi, M.; Szucs, R.; Dolan, J.W.; Pohl, C.A.; Haddad, P.R. Retention Index Prediction Using Quantitative Structure-Retention Relationships for Improving Structure Identification in Nontargeted Metabolomics. *Anal. Chem.* **2018**, *90*, 9434–9440. [CrossRef] [PubMed]
14. Zhou, Z.; Shen, X.; Tu, J.; Zhu, Z.-J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2016**, *88*, 11084–11091. [CrossRef]
15. Blaženović, I.; Shen, T.; Mehta, S.S.; Kind, T.; Ji, J.; Piparo, M.; Cacciola, F.; Mondello, L.; Fiehn, O. Increasing Compound Identification Rates in Untargeted Lipidomics Research with Liquid Chromatography Drift Time-Ion Mobility Mass Spectrometry. *Anal. Chem.* **2018**, *90*, 10758–10764. [CrossRef]
16. Hinnenkamp, V.; Klein, J.; Meckelmann, S.W.; Balsaa, P.; Schmidt, T.C.; Schmitz, O.J. Comparison of CCS Values Determined by Traveling Wave Ion Mobility Mass Spectrometry and Drift Tube Ion Mobility Mass Spectrometry. *Anal. Chem.* **2018**, *90*, 12042–12050. [CrossRef]
17. Zhang, X.; Kew, K.; Reisdorph, R.; Sartain, M.; Powell, R.; Armstrong, M.; Quinn, K.; Cruickshank-Quinn, C.; Walmsley, S.; Bokatzian, S.; et al. Performance of a High-Pressure Liquid Chromatography-Ion Mobility-Mass Spectrometry System for Metabolic Profiling. *Anal. Chem.* **2017**, *89*, 6384–6391. [CrossRef]
18. Zheng, X.; Aly, N.A.; Zhou, Y.; Dupuis, K.T.; Bilbao, A.; Paurus, V.L.; Orton, D.J.; Wilson, R.; Payne, S.H.; Smith, R.D.; et al. A Structural Examination and Collision Cross Section Database for over 500 Metabolites and Xenobiotics Using Drift Tube Ion Mobility Spectrometry. *Chem. Sci.* **2017**, *8*, 7724–7736. [CrossRef]
19. Righetti, L.; Bergmann, A.; Galaverna, G.; Rolfsson, O.; Paglia, G.; Dall'Asta, C. Ion Mobility-Derived Collision Cross Section Database: Application to Mycotoxin Analysis. *Anal. Chim. Acta* **2018**, *1014*, 50–57. [CrossRef]
20. Picache, J.A.; Rose, B.S.; Balinski, A.; Leaptrot, K.L.; Sherrod, S.D.; May, J.C.; McLean, J.A. Collision Cross Section Compendium to Annotate and Predict Multi-Omic Compound Identities. *Chem. Sci.* **2019**, *10*, 983–993. [CrossRef]
21. Hernández-Mesa, M.; Le Bizec, B.; Monteau, F.; García-Campaña, A.M.; Dervilly-Pinel, G. Collision Cross Section (CCS) Database: An Additional Measure to Characterize Steroids. *Anal. Chem.* **2018**, *90*, 4616–4625. [CrossRef] [PubMed]
22. Zhou, Z.; Luo, M.; Chen, X.; Yin, Y.; Xiong, X.; Wang, R.; Zhu, Z.-J. Ion Mobility Collision Cross-Section Atlas for Known and Unknown Metabolite Annotation in Untargeted Metabolomics. *Nat. Commun.* **2020**, *11*, 4334. [CrossRef] [PubMed]
23. Zhou, Z.; Tu, J.; Zhu, Z.-J. Advancing the Large-Scale CCS Database for Metabolomics and Lipidomics at the Machine-Learning Era. *Curr. Opin. Chem. Biol.* **2018**, *42*, 34–41. [CrossRef]
24. Rainey, M.A.; Watson, C.A.; Asef, C.K.; Foster, M.R.; Baker, E.S.; Fernández, F.M. CCS Predictor 2.0: An Open-Source Jupyter Notebook Tool for Filtering Out False Positives in Metabolomics. *Anal. Chem.* **2022**, *94*, 17456–17466. [CrossRef] [PubMed]
25. Plante, P.-L.; Francovic-Fontaine, É.; May, J.C.; McLean, J.A.; Baker, E.S.; Laviolette, F.; Marchand, M.; Corbeil, J. Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS. *Anal. Chem.* **2019**, *91*, 5191–5199. [CrossRef]
26. Zhou, Z.; Xiong, X.; Zhu, Z.-J. MetCCS Predictor: A Web Server for Predicting Collision Cross-Section Values of Metabolites in Ion Mobility-Mass Spectrometry Based Metabolomics. *Bioinforma. Oxf. Engl.* **2017**, *33*, 2235–2237. [CrossRef]
27. Zhou, Z.; Tu, J.; Xiong, X.; Shen, X.; Zhu, Z.-J. LipidCCS: Prediction of Collision Cross-Section Values for Lipids with High Precision To Support Ion Mobility-Mass Spectrometry-Based Lipidomics. *Anal. Chem.* **2017**, *89*, 9559–9566. [CrossRef]
28. Grisoni, F.; Ballabio, D.; Todeschini, R.; Consonni, V. Molecular Descriptors for Structure-Activity Applications: A Hands-On Approach. *Methods Mol. Biol. Clifton NJ* **2018**, *1800*, 3–53. [CrossRef]
29. Ross, D.H.; Cho, J.H.; Xu, L. Breaking Down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. *Anal. Chem.* **2020**, *92*, 4548–4557. [CrossRef]
30. Rainville, P.D.; Wilson, I.D.; Nicholson, J.K.; Isaac, G.; Mullin, L.; Langridge, J.I.; Plumb, R.S. Ion Mobility Spectrometry Combined with Ultra Performance Liquid Chromatography/Mass Spectrometry for Metabolic Phenotyping of Urine: Effects of Column Length, Gradient Duration and Ion Mobility Spectrometry on Metabolite Detection. *Anal. Chim. Acta* **2017**, *982*, 1–8. [CrossRef]
31. Stanstrup, J.; Neumann, S.; Vrhovšek, U. PredRet: Prediction of Retention Time by Direct Mapping between Multiple Chromatographic Systems. *Anal. Chem.* **2015**, *87*, 9421–9428. [CrossRef] [PubMed]
32. Falchi, F.; Bertozzi, S.M.; Ottonello, G.; Ruda, G.F.; Colombano, G.; Fiorelli, C.; Martucci, C.; Bertorelli, R.; Scarpelli, R.; Cavalli, A.; et al. Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification. *Anal. Chem.* **2016**, *88*, 9510–9517. [CrossRef] [PubMed]
33. Creek, D.J.; Jankevics, A.; Breitling, R.; Watson, D.G.; Barrett, M.P.; Burgess, K.E.V. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography-Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Anal. Chem.* **2011**, *83*, 8703–8710. [CrossRef] [PubMed]
34. Bonini, P.; Kind, T.; Tsugawa, H.; Barupal, D.K.; Fiehn, O. Retip: Retention Time Prediction for Compound Annotation in Untargeted Metabolomics. *Anal. Chem.* **2020**, *92*, 7515–7522. [CrossRef]

35. Liapikos, T.; Zisi, C.; Kodra, D.; Kademoglou, K.; Diamantidou, D.; Begou, O.; Pappa-Louisi, A.; Theodoridis, G. Quantitative Structure Retention Relationship (QSRR) Modelling for Analytes' Retention Prediction in LC-HRMS by Applying Different Machine Learning Algorithms and Evaluating Their Performance. *J. Chromatogr. B* **2022**, *1191*, 123132. [CrossRef]
36. Park, S.H.; De Pra, M.; Haddad, P.R.; Grosse, S.; Pohl, C.A.; Steiner, F. Localised Quantitative Structure-Retention Relationship Modelling for Rapid Method Development in Reversed-Phase High Performance Liquid Chromatography. *J. Chromatogr. A* **2020**, *1609*, 460508. [CrossRef]
37. Goryński, K.; Bojko, B.; Nowaczyk, A.; Buciński, A.; Pawliszyn, J.; Kaliszan, R. Quantitative Structure-Retention Relationships Models for Prediction of High Performance Liquid Chromatography Retention Time of Small Molecules: Endogenous Metabolites and Banned Compounds. *Anal. Chim. Acta* **2013**, *797*, 13–19. [CrossRef]
38. Naylor, B.C.; Catrow, J.L.; Maschek, J.A.; Cox, J.E. QSRR Automator: A Tool for Automating Retention Time Prediction in Lipidomics and Metabolomics. *Metabolites* **2020**, *10*, 237. [CrossRef]
39. Gritti, F. Perspective on the Future Approaches to Predict Retention in Liquid Chromatography. *Anal. Chem.* **2021**, *93*, 5653–5664. [CrossRef]
40. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2022.
41. Dusautoir, R.; Zarcone, G.; Verriele, M.; Garçon, G.; Fronval, I.; Beauval, N.; Allorge, D.; Riffault, V.; Locoge, N.; Lo-Guidice, J.-M.; et al. Comparison of the Chemical Composition of Aerosols from Heated Tobacco Products, Electronic Cigarettes and Tobacco Cigarettes and Their Toxic Impacts on the Human Bronchial Epithelial BEAS-2B Cells. *J. Hazard. Mater.* **2021**, *401*, 123417. [CrossRef]
42. Dunn, W.B.; Erban, A.; Weber, R.J.M.; Creek, D.J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; et al. Mass Appeal: Metabolite Identification in Mass Spectrometry-Focused Untargeted Metabolomics. *Metabolomics* **2013**, *9*, 44–66. [CrossRef]
43. Bittremieux, W.; Wang, M.; Dorrestein, P.C. The Critical Role That Spectral Libraries Play in Capturing the Metabolomics Community Knowledge. *Metabolomics Off. J. Metabolomic Soc.* **2022**, *18*, 94. [CrossRef] [PubMed]
44. Pezzatti, J.; González-Ruiz, V.; Codesido, S.; Gagnebin, Y.; Joshi, A.; Guillarme, D.; Schappler, J.; Picard, D.; Boccard, J.; Rudaz, S. A Scoring Approach for Multi-Platform Acquisition in Metabolomics. *J. Chromatogr. A* **2019**, *1592*, 47–54. [CrossRef] [PubMed]
45. Liebal, U.W.; Phan, A.N.T.; Sudhakar, M.; Raman, K.; Blank, L.M. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. *Metabolites* **2020**, *10*, 243. [CrossRef] [PubMed]
46. Domingo-Almenara, X.; Guijas, C.; Billings, E.; Montenegro-Burke, J.R.; Uritboonthai, W.; Aisporna, A.E.; Chen, E.; Benton, H.P.; Siuzdak, G. The METLIN Small Molecule Dataset for Machine Learning-Based Retention Time Prediction. *Nat. Commun.* **2019**, *10*, 5811. [CrossRef] [PubMed]
47. Tsugawa, H.; Ikeda, K.; Takahashi, M.; Satoh, A.; Mori, Y.; Uchino, H.; Okahashi, N.; Yamada, Y.; Tada, I.; Bonini, P.; et al. A Lipidome Atlas in MS-DIAL 4. *Nat. Biotechnol.* **2020**, *38*, 1159–1163. [CrossRef] [PubMed]
48. Mollerup, C.B.; Mardal, M.; Dalsgaard, P.W.; Linnet, K.; Barron, L.P. Prediction of Collision Cross Section and Retention Time for Broad Scope Screening in Gradient Reversed-Phase Liquid Chromatography-Ion Mobility-High Resolution Accurate Mass Spectrometry. *J. Chromatogr. A* **2018**, *1542*, 82–88. [CrossRef]
49. Celma, A.; Bade, R.; Sancho, J.V.; Hernandez, F.; Humphries, M.; Bijlsma, L. Prediction of Retention Time and Collision Cross Section (CCSH+, CCSH–, and CCSNa+) of Emerging Contaminants Using Multiple Adaptive Regression Splines. *J. Chem. Inf. Model.* **2022**, *62*, 5425–5434. [CrossRef]
50. Ross, D.H.; Cho, J.H.; Zhang, R.; Hines, K.M.; Xu, L. LiPydomics: A Python Package for Comprehensive Prediction of Lipid Collision Cross Sections and Retention Times and Analysis of Ion Mobility-Mass Spectrometry-Based Lipidomics Data. *Anal. Chem.* **2020**, *92*, 14967–14975. [CrossRef]