Machine Learning for Enhanced Identification in RPLC/HRMS Non-Targeted Workflows

Hiu-Lok Ngan,[†] Viktoriia Turkina,[‡] Denice van Herwerden,[‡] Hong Yan,[†] Zongwei Cai,[*,†] and Saer Samanipour[*,‡,§]

[†]State Key Laboratory of Environmental and Biological Analysis, Department of Chemistry, Hong Kong Baptist University, Hong Kong, P. R. China

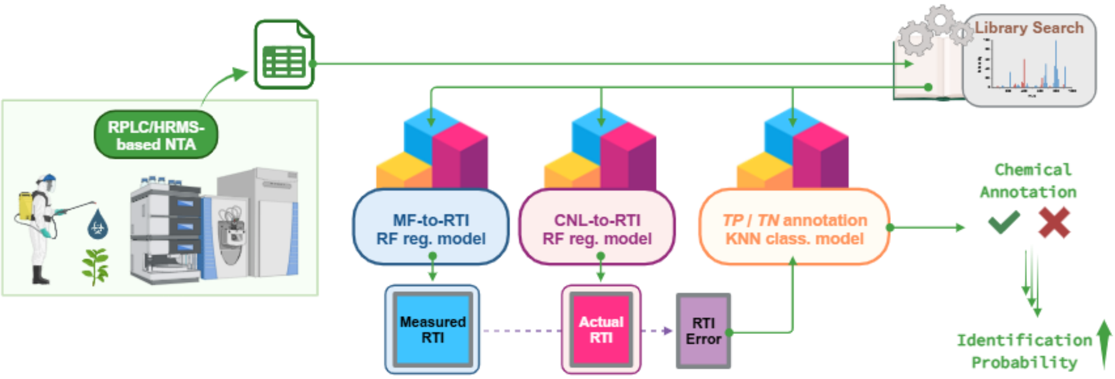
[‡]Van ’t Hoff Institute for Molecular Sciences (HIMS), University of Amsterdam, Amsterdam, 1098 XH, the Netherlands

[§]UvA Data Science Center, University of Amsterdam, Amsterdam, 1012 WP, the Netherlands

***Corresponding authors: Drs. Zongwei Cai and Saer Samanipour**

E-mails: zwcai@hkbu.edu.hk; s.samanipour@uva.nl

Abstract Graphic



Abstract

In HRMS-based non-targeted analysis (NTA) without access to any retention information of unknown compounds, spectral matching is one of the most employed approaches for the assessment of chemical identification probability (IP). Recently, within the metabolomics community, the use of true positive (*TP*) probability has been proposed as an alternative to the conventional confidence assessment approaches. In this study, a combination of information extracted from the MS/MS spectra and calibrant-free predicted retention time indices (RTIs) yielded the probability of *TP* for each chemical annotation by integrating 3 machine learning (ML) models. Firstly, they include a molecular fingerprint (MF)-to-RTI model trained by 4,713 calibrants. Then a cumulative neutral loss (CNL)-to-RTI model was trained by 485,577 experimental spectra. Finally, a binary classification model was trained by 1,686,319 *TP* and true negative (*TN*) annotations. Our results demonstrated a high correlation (training: $R^2 = 0.96$; testing: $R^2 = 0.88$) between MF-derived and CNL-derived RTI values, suggesting reduced RTI error for *TP* annotations. By incorporating the output parameters from a previously developed library search algorithm, monoisotopic mass, and RTI error for *TP* determination, the k-nearest neighbors algorithm achieved a weighted $F1$ score of 0.65 and a Matthews correlation coefficient of 0.30 for the annotations with their spectral matching scores ≥50% of total score. The attained ML models were applied to RPLC/HRMS NTA of pesticide mixtures that were spiked in solvent blank and 100× and 10× diluted black tea matrix. The chemical IPs of *TP* candidates were increased by 54.5%, 52.1%, and 46.7%, respectively. This work demonstrates the application of ML at large-scale model training to enhance chemical IP of unknown compounds.

Keywords

Non-targeted screening, Identification confidence, Quantitative structure-retention relationship (QSRR), MS/MS spectral reference library, Supervised learning, Model transferability

Scientific Contribution

Three supervised machine learning models were developed to improve the identification probabilities of unknown compounds in non-targeted RPLC/HRMS. Our work provided a higher confidence *in silico* analysis solution for the early-stage data analytics, aiding to shortlist chemicals-of-interest for further chemical identification by reference standards.

1. Introduction

Mass spectrometry (MS)-based non-targeted analysis (NTA) is a high-throughput technique for profiling sample analytes, distinguishing itself from targeted analysis through its discovery-oriented approach, which does not require prior knowledge of compound structures. The combination of reversed-phase liquid chromatography (RPLC) with high-resolution MS (HRMS) has become a prominent methodology for conducting NTA of the chemical exposome [1, 2], including contaminants such as pesticides [3], and per- and polyfluoroalkyl substances (PFAS) [4]. Researchers investigating human exposure to environmental chemicals not only focus on the exposome but also examine pollutant-induced alterations in the metabolome. Consequently, RPLC/HRMS-based NTA is increasingly employed for uncovering biological insights in environmental toxicology [5].

Accurate molecular annotation is crucial for compound identification. In LC/MS-based NTA, each deconvoluted two-dimensional (2D) LC/MS tensor is characterized by the information from the retention time (RT) and the mass-to-charge ratio (*m/z*) domains. To communicate confidence of compound identification, the Chemical Analysis Working Group at the Metabolomics Standards Initiative proposed a minimum metadata requirement for non-novel metabolite identification in 2007 [6]. For each annotated LC/MS tensor, its identification can only be putative if its measured *m/z* value and tandem MS (MS/MS) spectrum are just matched with its reference spectra from external libraries while the verification by chemical standard data acquired on the same platform is missed. Since 2014, similar criteria have been introduced in environmental analysis [7–9], emphasizing the necessity of further experimental effort [10], such as at least matching the retention behavior of reference standard on the same instrumental setting to achieve higher confidence match [7].

Although applying an in-house database of standard RT and MS/MS spectrum is the only efficient way to reach the highest identification confidence (IC) in RPLC/HRMS-based NTA, expanding the database requires significant resources. Accurate *in silico* analysis at the early stage can aid to screen out a panel of chemicals whose annotated identities require further validation by reference standard. Practically, it involves spectral matching against the *m/z* values of molecular ions and their associated fragment ions [11, 12]. Several approaches have been introduced for determining matching tolerance, including false discovery rate (FDR) estimation [13], modified cosine, and neutral loss-based spectral alignment [14]. Additionally, novel software solutions, such as DeepMASS [15], MAW [16], and ULSA [12], have been developed to rank candidate compounds based on spectral similarity or spectral entropy similarity [17]. However, these approaches only handle information in the *m/z* domain. Considering multiple hits resulted for each MS/MS spectrum during reference library search, incorporating retention behavior into a computational process is necessary. Recently, the concept of "identification probability" (IP) was proposed by Metz et al [18], advocating for the measurement of ambiguity rather than confidence to improve transferable annotation results across analytical platforms. In our study, we aim to investigate whether incorporating chemical retention informatics with the information extracted from mass spectrum through machine learning (ML) modeling techniques enhance the compound IP in RPLC chemical space.

Model transferability is the ability of a model to accurately predict beyond the data it was originally trained on [19]. It requires diverse data sources and substantial volume during statistical modeling. Previous studies have reported on ML models for RT prediction in LC-based analyses of chemicals, such as the chemical exposome [20–24]. However, the transferability of these models is often limited by the specific chromatographic conditions [25–29], as RT can vary significantly with individual chromatographic methods. A harmonized RT index (RTI) scale is therefore necessary to offer a chromatographical method-independent alternative, although there are different schemes of RT index (RTI). It has its potentials of RTI model transferability and of integration with spectral information to enhance IC [30–33].

Herein, we present a computational approach to achieve calibrant-free transferable prediction of RTI value on a harmonized scale and integrate it with reference spectral match for compound IP enhancement in RPLC/HRMS-based NTA through statistical modeling. Compounds-of-interest with high predicted IP thus are suggested to have further experimental validation. Among the 3 ML models involved in our identification workflow, there are 2 models built on the quantitative structure-retention relationship (QSRR) to horizontally compute expected RTI values separately for the chemicals beyond the calibrants' data. Upon MS/MS information being extracted through reference library search, a binary classification model that incorporates information from both retention and *m/z* domains is used to determine true positive (*TP*) annotations. The integrative use of our models is demonstrated by RPLC/HRMS data on various pesticides-spiked blanks and black tea samples. Incorporation of the developed ML models increased the IP of the chemical exposome by half on average, showing this computational approach benefits from non-targeted screening in scoping chemical compounds that require further identification.

2. Materials and Methods

2.1. Overall Roadmap

In this work, 3 ML models and the MS/MS spectral matching algorithm, ULSA [12], were employed. As the roadmap illustrated in **Figs. 1A and S01**, the first model (Model 1) is a random forest (RF) regression model that correlates molecular fingerprint (MF) to the true RTI values of calibrants from 3 comparable scales. The second model (Model 2) is another RF regression model that utilizes experimental MS/MS spectra from various external spectral databases to predict the expected RTI values on a harmonized scale without knowledge about the extract structure. During modeling Models 1 and 2, train/test split were performed based on MF leverage rather than RTI to ensure the model was trained and tested on structurally diverse compounds. The third model (Model 3) is a k-nearest neighbors (KNN) binary classification model designed to predict the acceptance of matched compound annotations from ULSA or any other spectral matching algorithms with the spectral matching scores ≥50% of total score. This model integrates feature deviations from both the RTI and *m/z* domains. The computational method was validated using both internal train/test splits and external independent datasets, including RPLC/HRMS data from pesticide analyses in blank spike and tea extracts.

2.2. Model 1: Molecular Fingerprint-Based Regression Model

A previously reported RF model for predicting molecular RTI was retrained for this study with some modifications [33]. In brief, this model correlates the pre-selected MF features to calibrant RTI for transferable prediction of expected RTI value. For each canonical simplified molecular-input line-entry system (SMILES)-identified molecule, 780 AtomPairs2DFingerprintCount (APC2D) and 881 PubChem Fingerprinter (PubChem) MF features were computed using PaDEL [34]. From the features in the PubChem system, only 148 bits (i.e., from the 115th to the 262nd)—specifically those describing cyclic structure—were selected and condensed as 10 features to complement with the APC2D features used to describe non-cyclic structure as previously described [33]. Co-correlation among features was examined to consider the portion of features used for each tree's training (discussed in **Supplementary Information**). As a result, each compound was represented by a 790-bit MF vector.

Different from the previous study, stratification based on the calibrants' MFs was involved in work. Transferable prediction of RTI values was achieved by training a model on 4,713 calibrants across 3 comparable scales of RTI in RPLC: the C3–14 n-alkylamide system (Amide) [35], the RTI system developed by Aalizadeh et al. from the University of Athens (UoA) [31], and the C0–23 cocamide diethanolamine homologous series (Cocamide) [36]. Details on model training and validation are discussed in **Supplementary Information**. To assess model transferability, accuracy of the predicted RTI value on the harmonized scale was tested by the true values of 1,237 untrained calibrants in terms of mean absolute error. The list of compounds used to train and test this MF-to-RTI RF regression model was available online (see file 1 on 10.5281/zenodo.14752891).

2.3. Model 2: Cumulative Neutral Loss-Based Regression Model

In addition to predicting expected RTI values on a harmonized scale from 2D chemical descriptors, a second RF regression model was established to correlate empirical positive electrospray ionization (ESI+) MS/MS spectral data to the RTI values predicted from Model 1, following a strategy similar to a prior study [32]. Feature for this Model 2 was based on cumulative neutral loss (CNL), a method recently applied for matching analog molecules [14, 37, 38]. This approach has proven advantageous for chemical componentization [14, 39]. Based on the previous results [32, 40], a list of 15,961 high probability masses-of-interest (MOIs) was pre-selected as features. The list is available online (see file 2 on 10.5281/zenodo.14752891). Moreover, monoisotopic mass was included as an additional feature, as it was previously reported to be one of the most discriminative to CNL-based model due to its intrinsic relationship with molecular weight [32].

Model 2's training dataset was compiled from various MS/MS reference spectral data from known compounds that have chemical identifiers, specifically InChIKeys and SMILES IDs. Databases including MassBank EU, MoNA, and NIST LC-MS/MS provided multiple spectra for specific compounds. In total, there were 27,211 distinct molecules, consisting of 693,685 MS/MS spectra used for training (the list is available online, see file 3 on 10.5281/zenodo.14752891). Each spectrum contains at least 2 MOIs. To address the occurrence of multiple compounds, Hao Guo et al. implemented a random train/test split strategy, reserving only 1 spectrum for training and testing while using others for validation [41]. This approach effectively manages sample balance, ensuring each compound carries equal weight in the model, which is particularly important for regression modeling. However, the output labels of

Model 2, the predicted expected RTI values generated by Model 1, were primarily localized in the center rather than scattered in the margin regions (around RTI values of 0 and 1,500), making edge case overloading neglectable. In our study, we aimed to include as many molecules as currently available in the relevant spectral databases. This CNL-to-RTI RF regression model was trained on 485,577 query MS/MS spectra. Details on model training and validation are discussed in **Supplementary Information**. To evaluate the exchangeability between Model 1 and Model 2, exploratory data analysis (EDA) and quartile analysis were performed to study the similarity in distribution and the closeness of the predicted RTI values with the true RTI values of the compound calibrants used to train Model 1. Predictive power of Model 2 was tested on 208,104 query spectra by using R^2 value and root mean square error (RMSE, calculated according to **eq1**) as the performance metrics.

2.4. Spectral Reference Library and Universal Library Search Algorithm

ULSA, a previously developed MS/MS spectral matching algorithm [12], was employed in this study to annotate compounds by matching MS/MS spectra from various public spectral databases, including MassBank EU, MoNA, and GNPS. During the execution of ULSA, 7 different parameters were summed to derive a final performance score using the formula outlined in **eq2**. The final scores ranged from 0 to 7, with higher values indicating increased reliability of the assigned annotation based on spectral matching. “FinalScoreRatio” (S_8) was then calculated by dividing the final score by 7. In **eq2**, MS_{tol} represents the user-defined *m/z* tolerance in Daltons for an ion to be accepted as a positive match. The 7 features (S_{1-7}) that were extracted from the spectra for feature selection in Model 3 included: S_{1-2}, the fractions of matched fragments to the reference spectra (denoted as $f_{matched}/f_{reference}$ in **eq2**) and to the total fragments in the user spectrum ($f_{matched}/f_{user}$); S_3, the observed mass error in the precursor ion ($MS_{err,pre}$); S_4, the absolute average mass error of all matched fragment ions ($\left|MS_{err,frag}\right|$); S_5, the standard deviation of mass error for matched fragments ($STD_{err,frag}$);

and S_{6-7}, the forward ($MF_{forward}$) and reverse ($MF_{reverse}$) match factors calculated via dot product. Importantly, these output parameters are from statistical calculations that can be computed directly from the spectra, making it rather independent from ULSA implementation. Also, the publicly available spectral libraries are separated to ULSA. For any reference MS/MS spectral data that is newly acquired, uploading it to these databases allowing them to be matched. That is particularly valuable for annotating the chemicals of emerging concern.

2.5. Model 3: Predicting Acceptance of Molecular Annotation

Model 3 is a binary classification model aimed at determining whether a molecular annotation suggested by ULSA was *TP* and should be accepted (the outputs were labeled as "1"), or it was indicates true negative (*TN*) and should be rejected (labeled as "0"). The features for this model include the RTI difference between the RTI values derived from Models 1 (RTI_{MF}) and 2 (RTI_{CNL}), along with monoisotopic mass and 8 candidate features obtained from ULSA. We hypothesize that a larger RTI error indicates *TN* annotations, while a smaller error correlates with *TP* annotations. Some features as discussed in **Supplementary Information** exhibit right-skewed distribution were log-transformed, yet all data were scaled by mean normalization for model selection and optimization. Several ML algorithms, including logistic regression (LR), decision tree (DT), RF, and KNN, were evaluated. Pearson correlation was applied to eliminate redundant features that have less feature importance and *r* values >0.80 for fair comparison among models. Further feature selection focuses on the model's sensitivity to excluded features, prioritizing those with the most negative contribution to predictions. Details information on model selection and models' hyper-parameter tuning are discussed in **Supplementary Information**. A set of semi-synthetic MS/MS spectra was prepared following methodologies from prior research [12]. In short, HRMS spectra with a mass resolution >5,000, acquired using an ESI+ source, were randomly selected from databases such as MassBank EU, MoNA, GNPS, and NIST. To mimic false negative (*FN*) ions (i.e., loss ions) caused by ion suppression or acquisition defects, 10% fragments were removed randomly. Data augmentation was applied by adding noise to simulate false positive (*FP*) ions. For instance, if the mass tolerance for MS/MS spectral matching was set to ±20 mDa, *FP* ions with random mass errors in a range of ±30 mDa were incorporated to the accurate masses of each fragment ion. These semi-synthetic sample spectra were then analyzed by ULSA, yielding 4,368,902 annotated MS/MS spectra. Since all annotated spectra with a "FinalScoreRatio" (S_8) <0.50 are in 100% of *TNs*, ML was deemed unnecessary for distinguishing *TP* from *TN*. Ultimately, 1,686,319 spectra were used for training and 421,381 for testing. Among the training samples, 1,535,009 spectra were *TN* (label "0"), while 151,310 instances represented *TP* (label "1"). Nine additional replicates of *TP* spectra were used to balance the training dataset. The optimal binary classification model was tested on an external independent dataset, which comprised of experimental data of pesticides spiked into blank samples at varying concentrations (1, 2.5, 5, 10, 25, 50, 100, and 1,000 ppb). Because both *TP* and *TN* are known in the blank spikes, the optimal model was primarily assessed based on the Matthews correlation coefficient (MCC) score (**eq3**) and secondarily by weighted *F*1 score (**eq4**).

2.6. Application of the Models in the Identification Workflow: NTA of Pesticides-Spiked Black Tea Samples

After developing all the predictive models, they were integrated into the NTA pipeline for aiding pesticide annotation as a real-world application. Their overall effectiveness in enhancing compound IPs was evaluated using a set of RPLC-MS/MS data from pesticides-spiked black tea extract samples, as previously reported in a CNL model development study [40]. The pesticide standards were part of the LC/MS Pesticide Comprehensive Mix Kit (Part Number 5190-0551) from Agilent Technologies. Pesticide mixtures were spiked in solvent blank and 10× and 100× diluted black tea matrix respectively, resulting 1 ppm final solutions. The details of data acquisition were outlined in the prior CNL model study [40].

- Data Pre-Processing

To extract HRMS information required as the model input, a user interface, namely "jHRMS ToolBox", was previously developed for no-code application [42]. Detailed descriptions of its component modules and their modules' input/output were summarized in **Table S1** for facilitating MS community to reassemble the pipeline. To introduce, firstly, the upstream processing of LC/HRMS files requiring the open MS format, ".mzXML". Raw data files conversion was accomplished using the "MS_Import" module, built on MSConvertGUI (64-bit, ProteoWizard [43]). Secondly, this ecosystem also incorporates the self-adjusting feature detection (SAFD) algorithm for extracting LC-MS/MS features [44]. Thirdly, componentization was achieved by executing the "CompCreate" module, which correlates the extracted features with the probabilistic CNL model [40]. Finally, the resultant MS/MS spectra were then analyzed using ULSA for compound annotation [12]. It returned a spreadsheet that contains necessary information for the identification workflow, including the suggested compound, *m/z* value of the molecular ion, a list of *m/z* values of the correlated fragments, and the scores S_{1-8}.

- Models Application

In **Figs. 1B** and **S01**, the resultant spreadsheet was fed into Models 1 and 2 to calculate RTI error (as defined by **eq5**). Subsequently, Model 3 was deployed for each annotated RPLC/HRMS tensor, utilizing the necessary ULSA-returning scores, monoisotopic mass, and RTI error to predict whether the suggested chemical annotation was acceptable.

Model 1 (the MF-to-RTI model) was deployed for obtaining the annotated compounds' RTI values (considered as measured values). Leverage (h_{ii}) was used to assess whether an annotated compound was in the applicability domain (AD, the 95% leverage threshold) of Model 1 (0.275) based on the training dataset. It is calculated by the equation shown in **eq6**, where X represents the matrix of training data, and x_i is the 790-bit vector for an individual data query. If the leverage value of a suggested annotation is within AD, the obtained RTI value has low uncertainty and can be treated as a ULSA-measured value. Similar, Model 2 (the CNL-to-RTI model) was deployed for obtaining the actual RTI values if and only if a RPLC/HRMS tensor's spectral data was within the AD of Model 2 (h_{ii} <0.146). Upon computing the RTI values from both Models 1 and 2, RTI error, the difference between measured and actual RTI values, was then calculated and proceeded to Model 3 as one of the features, namely "DeltaRi".

Model 3 was deployed for *TP* annotation determination. A script for its application is available (see Section **Computations and Code Availability**). In short, the RPLC/HRMS tensors with the values of S_8 ≥0.50 and within the ADs of Models 1 and 2 were considered. Independent variables included logarithm of S_1, S_3, S_5, logarithm of S_8, logarithm of a value of monoisotopic mass/1000, and "DeltaRi". They were normalized by mean before Model 3 implementation. For real sample analyses, recall (as defined by **eq7**) was chosen as the primary performance metric since *TN* and *FP* are always unknown in real sample.

- Identification Probability Calculation

In ambiguity measurement of a candidate RPLC/HRMS tensor, a collection of reference spectra can be matched. For each compound hit, multiple spectra of the same compound may correspond. Hence, a weighted probability of *TP* (*P*(*TP*)) was calculated for each hit. A decision threshold of 0.50 was then applied to determine whether one should retain or exclude that hit. The shortlisted hits for each RPLC/HRMS tensor were then used to calculate the IP according to **eq8**. In other words, if ULSA implementation yielded 5 hits, the probability of *TP* from ULSA determination (*P*(ULSA)) was calculated as 1/5. Since different numbers of reference spectra (1, 5, 10, 20, and 50) were matched for the respective hits, the weighted *P*(*TP*) for each hit was calculated as the summation of *P*(*TP*) values for each spectrum-specific annotation derived from Model 3, followed by a division of the number of spectra. Thus, the calculations would be Σ(*P*(*TP of the 1st hit*))/1, Σ(*P*(*TP of the 2nd hit*))/5, Σ(*P*(*TP of the 3rd hit*))/10, Σ(*P*(*TP* *of the 4th hit*))/20, and Σ(*P*(*TP of the 5th hit*))/50. If only 2 hits remained from the 5, the IP would then be 1/2.

2.7. Computations and Code Availability

All scripts for ML modeling and data visualization were written in Julia v.1.6 using Visual Studio Code (Microsoft), while some data visualization was conducted in Python v.3.10 on Jupyter Notebook (Anaconda3). The scripts (with step-by-step instructions) for models application and IP determination and also the developed models are available in the following repositories respectively: git@github.com:TommyNHL/exposomeIDproba.git (scripts); and git@bitbucket.com:hiulokngan/modelsExposomeProjUvA.git (models). Details regarding computational power and packages used can be found in **Table S2**. Any updates will be published on both repositories with a citation of the final DOI for this work.

3. Results and Discussion

3.1. Performance of Model 1 (the Molecular Fingerprint-to-Retention Time Index Model)

The first model developed in this study was utilized to predict RTI_{MF} (considered as measured RTI for each suggested annotation) on a harmonized scale. It was trained by 4,713 compound calibrants with 5,048 true TRI values across 3 similar RTI scale systems. The mean absolute error of the predicted values of the trained calibrant was 89.52. Contrasting to the mean absolute difference (78.56) among true RTI values from different scales of individual calibrants with publicly available MS/MS spectra (the list is available online, see file 4 on 10.5281/zenodo.14752891), similar uncertainty might exist for the predicted values from Model 1. The uncertainty (with a mean absolute error of 111.15) was slightly increased when the model tested by 1,263 true values of 1,237 calibrants. In the histograms depicted in **Figs. 2A and 2B**, the true RTI values (colored in purple-blue) and CNL_{MF} (in sky-blue) exhibited similar distributions among the data used for training and beyond modeling respectively. Hence, the retrained MF-to-RTI model has been validated to be transferable in this study.

3.2. Model 2 (the Cumulative Neutral Loss-to-Retention Time Index Model)

The second model established in this study was employed to predict RTI_{CNL} (considered as actual RTI for each suggested annotation) on a harmonized scale that was defined in Model 1. Model 2 was trained to correlate the pre-selected CNL masses with RTI_{MF} by 485,577 query MS/MS spectra and tested by 208,104 spectra. Both Models 1 and 2 were used to compute RTI error of each chemical annotation upon universal library search. It was calculated by the difference between RTI_{MF} and RTI_{CNL}. If a chemical annotation is *TP*, both predicted RTI values should be similar and thus the resultant error is close to 0. Wrong annotation gives a large different in RTI_{MF} (measured value) from RTI_{CNL} (actual value), resulting larger RTI error in a *TN* annotation.

- Model Performance

Due to QSRR, distribution of RTI_{CNL} (green bars) were similar to RTI_{MF} (sky-blue bars) and the true RTI values (purple-blue bars) of chemical calibrants used to train (**Fig. 2A**) and test (**Fig. 2B**) Model 1. Variances of RTI error from Model 1 (sky-blue box) was smaller than that from Model 2 (green box, **Fig. 2C**) for the data it was originally trained on Model 1. However, both models showed similar RTI error variance for the extended data (data not used to train Model 1, see **Fig. 2D**). These similar deviation patterns demonstrated the exchangeability of RTI values that were computed from Models 1 and 2.

For the model trained with a 7:3 train/test split, the RMSEs were 51.8 and 92.8, with R^2 values of 0.96 and 0.88 for the training and testing datasets (**Table 1** and **Figs. 2E and 2F**). These results were comparable to the predictive power of the CatBoost model that was trained on NORMAN dataset by different descriptors by Boelrijk et al. [32]. Our retrained RF model demonstrated the predicted values had better correlation with MF-derived RTIs (train: R^2 = 0.94; test: 0.85 in Boelrijk et al.'s work). Larger RMSEs were observed in our study (train: 44.0; test: 67.0 in Boelrijk et al.'s work) might be due to the RTI scale adopted in this study was across 3 comparable RTI scaling systems. The comparable predictive performance with the previous work confirming that this newly constructed Model 2 was transferable and could effectively predict expected RTI values based on QSRR.

- Feature Interpretation

In this study, the spectral data collected showed mass distribution of the CNL masses during ESI+ MS/MS analysis lying into range between 0–500 Da (see **Fig. S02**), consistent with the findings from van Herwerden et al. [40]. CNL masses with lower weight showed higher frequency than the higher masses (the list of counts for each CNL mass is available online, see file 2 on 10.5281/zenodo.14752891). Similar distribution also can be seemed in the semi-synthetic data (see **Fig. S03**).

3.3. Model 3: Binary Classification Model

- Model Performance

Model 3 was a KNN model that incorporated 6 features, namely “RefMatchFragRatio”, “MS1Error”, “MS2ErrorStd”, “FinalScoreRatio”, monoisotopic mass, and RTI error. This model only applied to the individual spectral matching results with the scores of ≥50% for *TP*/*TN* annotation determination. Our optimized model was a RTI error inclusive model. The area under the receiver operating characteristics curve (AUROC) and the area under the precision-recall curve (AUPRC) were 0.95 and 0.93 in training and testing datasets, respectively (**Fig. 3A**). RTI error inclusive model exhibited slightly greater AUROC and AUPRC values than the values of RTI error exclusive model (train: AUROC = 0.94, AUPRC = 0.92; test: AUROC = 0.94, AUPRC = 0.91). If only considered positive results, the recalls were 0.99 for training and testing datasets (see **Table 1**). Our KNN model achieved a weighted *F*1 score of 0.65 and a MCC score of 0.30 for pesticides-spiked blank samples, with a recall of 0.66 (**Table 1**). Notably, the predictive power of RTI error inclusive model resulted in better recall (0.60) for determining acceptance of the chemical exposome annotation when the samples presented with 10× and 100× diluted black tea matrix than the RTI error exclusive model (recall = 0.54). We demonstrated that ML modeling by adopting semi-synthetic data was feasible to achieve acceptable model transferability, annotating RPLC/HRMS tensors in higher IC.

By deploying Model 3, the default probability threshold for accepting an annotation as *TP* was set at 0.50, resulting in *TP* and *FP* rates of 98.8% and 23.4%, respectively (**Fig. 3C**), with a FDR of 19.1%. FDR-controlled cut-offs for the probability of accepting an annotation as *TP* were 0.89 and 0.78, respectively (**Fig. 3D**). These decision thresholds could be applied to calculate recalls for individual samples in different levels of IC. For 1 ppm samples, their recalls increased as the matrix effect decreased (**Fig. 3E**). When the matrix effect was absent, as solution concentrations increased, so did the number of deconvoluted RPLC/HRMS tensor and the number of positive candidates to be identified by our identification workflow. However, Model 3 exhibited no predictive power for the analyses of 1 ppb solution, while it functioned stably for the analyses of chemical contaminants with concentrations of 2.5 ppb or above.

- Feature Interpretation

Since the developed binary classification model indicated discriminative power to *TP*/*TN* chemical annotations, we conducted permutation analysis to evaluate feature importance. All KNN models with MCC scores >0.30 on validation dataset (the experimental pesticides-spiked blanks) during model selection and hyperparameter tuning were examined. Our results confirmed that the excluded features, logarithm of S_2 ("UsrMatchFragRatio") and the difference between S_6 ("DirectMatch") and S_7 ("ReverseMatch"), were one of the factors contributing to the models' predictive power on experimental data. The most significant feature was S_3, "MS1Error" (**Fig. 3F**), while the remaining features indicating importances occasionally with positive and negative contributions that were close to 0.00. These findings were only specific to the binary classification models we attained, while they did not reflect the intrinsic predictive values of features. Hence, Model 3 was significantly sensitive to the differences between the measured molecular ion mass and its actual *m/z* values reported from public reference spectra. Incorporation of other least discriminative features was still critical for Model 3 to exhibit its predictive power on experimental data.

3.4. Performance Comparison with Solely MS/MS Spectral Library Search

To examine the extent of improvement achieved by deploying our models (Models 1, 2, and 3) accompanying with MS/MS spectral reference library search, we calculated the IP of molecular features based on spectral matching and ML-assisted determination of annotation acceptance or rejection. Performance comparison was based on 2 different definitions of "hit".

The first definition of hit refers to a collection of reference spectra matched for measuring the ambiguity of a candidate compound. We averaged the IPs of all positive results to facilitate overall performance comparison. As shown in **Fig. 3G**, for 1 ppm sample solutions, integration of our 6-component KNN model (Model 3) enhanced average IPs from 44% to 68% (24% difference) for the blank spike and from 48% to 73% (25% difference) in samples presented with 100× diluted tea matrix, representing increments of 54.5% and 52.1%, respectively. The presence of 10× diluted tea matrix slightly affected the improvement from ML incorporation, which revealed a 21% higher IP than singly ULSA alone (from 45% to 66%), equivalent to 46.7% increase.

Molecular annotation via spectral library search typically ranks based on score of matched reference spectra to measure annotation confidence [12, 16]. We also compared the performance by defining a hit as a single reference spectrum match upon ranking by matching score. Ranking was performed for library search-suggested annotations and ML-assisted annotations using "FinalScoreRatio" and by "*P*(*TP*)" without a defined acceptance threshold. An alternative IP, as defined in **eq8** for ranking analysis, represented the occurrence frequency of *TP* annotations in the top-ranked hits. As shown in **Fig. 4**, IPs were improved slightly (1–

3%) on average for samples with tea matrices due to ML incorporation, while remaining comparable for spiked blanks. This finding emphasized the importance of considering identification ambiguity of a candidate compound from a collection of reference spectral and multiple platforms rather than a single reference spectrum hit [18]. *TP* hit rates were calculated as the ratio of RPLC/HRMS tensors with any *TP* hit at least once in the top ranks to the total identified spectra. Library search achieved *TP* hit rates of 87–88%, 83–87%, and 76–81% for top-5, top-3, and top-1 hits, respectively. Our assessments found that ML incorporation improved these rates to 89–95%, 87–90%, and 78–83%.

Collectively, chemical IP could be enhanced by applying our developed models. The improvement would be greater if the number of reference spectra increased. This observation was supported by our results of top-1, top-3, top-5 analyses and the ambiguity analyses that included a collection of matched reference spectra (i.e., top-many analyses).

3.5. Potentials and Limitations

In this work, we demonstrated a novel approach of integrative ML models to enhance chemical IP in RPLC/HRMS-based NTA. Our results demonstrated the effectiveness of our developed models, particularly Model 3, which integrates RTI error computed by the RTI values on a harmonized scale from 2 QSRR-based models to enhance molecular annotation IPs. Model 1 (the MF-to-RTI model) exhibited transferable RTI prediction and its exchangeability with Model 2 (the CNL-to-RTI model). Model 2 showed strong predictive power for the expected RTIs, capable to predict RTI values of chemical compounds beyond the data used to modeling. The integration of ML incorporation with reference spectral library searches in our identification workflow for the measurement of ambiguity not only improved recall in real samples with and without tea matrices but also increased the confidence of annotation acceptance compared to single library searches. Application of Model 3 resulted in acceptable weighted *F*1 and MCC scores for the blank spike samples, alongside notable increases in IPs with diluted tea matrix. Greater IP improvements were observed as the number of reference spectra increased. In our work, for both tea matrix-containing and matrix-free samples, computing a chemical IP from its corresponding collection of all available reference MS/MS spectra improved half in annotation confidence. Our work provided a higher confidence *in silico* analysis solution for the early-stage data analytics, aiding to shortlist chemicals-of-interest for further chemical identification by reference standards.

However, a primary limitation inherent in our approach is the reliance on the quality and diversity of the spectral reference libraries. Incomplete or biased libraries may hinder the identification of highly structurally diverse compounds or those with limited accessible reference spectral data. To address this challenge, it is essential to continuously update and expand the spectral libraries, ensuring they encompass a broad spectrum of compounds encountered in real-world samples. Furthermore, ongoing validation with various real-world samples of *TP* and *TN* MS/MS spectra will be essential to fine-tune the models and ensure their reliability across various analytical conditions.

Abbreviations

2D	Two-dimensional
AD	Applicability domain
Amide	C3–14 n-Alkylamide system
APC2D	AtomPairs2DFingerprintCount
AUPRC	Area under the precision-recall curve
AUROC	Area under the receiver operating characteristics curve
CNL	Cumulative neutral loss
Cocamide	C0–23 Cocamide diethanolamine homologous series
CV	Cross-validation
DT	Decision tree
EDA	Exploratory data analysis
ESI+	Positive electrospray ionization
FDR	False discovery rate
FN	False negative
FP	False positive
HRMS	High-resolution mass spectrometry

IC	Identification confidence
IP	Identification probability
KNN	k-Nearest neighbors
LR	Logistic regression
m/z	Mass-to-charge ratio
MCC	Matthews correlation coefficient
MF	Molecular fingerprint
ML	Machine learning
MOIs	Masses-of-interest
MoNA	MassBank of North America
MS/MS	Tandem mass spectrometry
NTA	Non-targeted analysis
P(*TP*)	Probability of true positive
P(ULSA)	Probability of true positive from ULSA determination
PFAS	Per- and polyfluoroalkyl substances
PubChem	PubChem Fingerprinter
QSRR	Quantitative structure-retention relationship
R^2	Coefficient of correlation
RF	Random forest
RI	Retention index
RMSE	Root mean square error
RPLC	Reversed-phase lipid chromatography
RT	Retention time
RTI	Retention time index
SAFD	Self-adjusting feature detection
SMILES	Simplified molecular-input line-entry system
TN	True negative
TP	True positive
ULSA	Universal Library Search Algorithm
UoA	the RI system developed at the University of Athens

Supplementary Information

A word file with details regarding modeling, identification workflow, computational power, libraries used, and the additional figures and tables for the supporting text.

Author Information

- **Disclosure**

The authors declare that they have no conflicts of interest or personal relationships that could influence the work reported in this paper.

▪ Author Contributions

Conceptualization, H.-L.N., V.T., and S.S.; resources, H.-L.N., V.T., D.v.H., Z.C., and S.S.; data curation, H.-L.N., V.T., and D.v.H.; writing-original draft preparation, H.-L.N.; writing-review and editing, H.-L.N., V.T., Y.H., and S.S.; visualization, H.-L.N. and Y.H.; supervision, Y.H., Z.C., and S.S.; project administration, Z.C. and S.S.; funding acquisition, Z.C. and S.S. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

H.-L. Ngan acknowledges the University Grant (UG) Studentship (R20PSCI000679), the sponsorship from the Graduate School of Hong Kong Baptist University (HKBU), the Environmental Monitoring and Computational Mass Spectrometry (EMCMS, www.emcms.info) group for their insights and feedback, and Dr. S. Samanipour for his host of overseas research experience program. S. Samanipour and V. Turkina thank the ChemistryNL for financial support. H. Yan thanks the Start-up Grant for New Academics – YAN Hong (165520). The authors from HKBU thank the financial support from the Kwok Chung Bo Fun Charitable Fund for the establishment of the Kwok Yat Wai Endowed Chair of Environmental and Biological Analysis.

References

1. Manz KE, Feerick A, Braun JM, et al (2023) Non-targeted analysis (NTA) and suspect screening analysis (SSA): A review of examining the chemical exposome. J Expo Sci Environ Epidemiol 33:524–536.
2. Samanipour S, Barron LP, van Herwerden D, et al (2024) Exploring the chemical space of the exposome: How far have we gone? JACS Au 4:2412–2425.
3. Guo Z, Zhu Z, Huang S, Wang J (2020) Non-targeted screening of pesticides for food analysis using liquid chromatography high-resolution mass spectrometry-a review. Food Additives and Contaminants - Part A 37:1180–1201.
4. Zweigle J, Bugsel B, Zwiener C (2022) FindPFΔS: Non-target screening for PFAS-comprehensive data mining for MS2 fragment mass differences. Anal Chem 94:10788–10796.
5. Chen Y, Ngan HL, Song Y, et al (2024) Chronic real-ambient PM2.5 exposure exacerbates cardiovascular risk via amplifying liver injury in mice fed with a high-fat and high-cholesterol diet.

Environment & Health 2:221–232.

6. Sumner LW, Amberg A, Barrett D, et al (2007) Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). Metabolomics 3:211–221.

7. Schymanski EL, Jeon J, Gulde R, et al (2014) Identifying small molecules via high resolution mass spectrometry: Communicating confidence. Environ Sci Technol 48:2097–2098.

8. Hollender J, Schymanski EL, Ahrens L, et al (2023) NORMAN guidance on suspect and non-target screening in environmental monitoring. Environ Sci Eur 35:75.

9. Alygizakis N, Lestremau F, Gago-Ferrero P, et al (2023) Towards a harmonized identification scoring system in LC-HRMS/MS based non-target screening (NTS) of emerging contaminants. TrAC - Trends in Analytical Chemistry 159:116944.

10. Ciccarelli D, Samanipour S, Rapp-Wright H, et al (2024) Bridging knowledge gaps in human chemical exposure via drinking water with non-target screening. Crit Rev Environ Sci Technol 55:190.

11. Hulleman T, Turkina V, O'Brien JW, et al (2023) Critical assessment of the chemical space covered by LC-HRMS non-targeted analysis. Environ Sci Technol 57:14101–14112.

12. Samanipour S, Reid MJ, Bæk K, Thomas K V. (2018) Combining a deconvolution and a universal library search algorithm for the nontarget analysis of data-independent acquisition mode liquid chromatography-high-resolution mass spectrometry results. Environ Sci Technol 52:4694–4701.

13. Scheubert K, Hufsky F, Petras D, et al (2017) Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun 8:1494.

14. Bittremieux W, Schmid R, Huber F, et al (2022) Comparison of cosine, modified cosine, and neutral loss based spectrum alignment for discovery of structurally related molecules. J Am Soc Mass Spectrom 33:1733–1744.

15. Ji H, Xu Y, Lu H, Zhang Z (2019) Deep MS/MS-aided structural-similarity scoring for unknown metabolite identification. Anal Chem 91:5629–5637.

16. Zulfiqar M, Gadelha L, Steinbeck C, et al (2023) MAW: The reproducible Metabolome Annotation Workflow for untargeted tandem mass spectrometry. J Cheminform 15:32.

17. Li Y, Kind T, Folz J, et al (2021) Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. Nat Methods 18:1524–1531.

18. Metz TO, Chang CH, Gautam V, et al (2024) Introducing "identification probability" for automated and transferable assessment of metabolite identification confidence in metabolomics and related studies. Anal Chem 97:1–11.

19. Assis J, Serrão EA, Fragkopoulou E, et al (2024) Misconception of model transferability precludes estimates of seagrass community reorganization in a changing climate. Nat Plants 10:1071–1074.

20. Munro K, Miller TH, Martins CPB, et al (2015) Artificial neural network modelling of pharmaceutical residue retention times in wastewater extracts using gradient liquid chromatography-high resolution mass spectrometry data. J Chromatogr A 1396:34–44.

21. McEachran AD, Mansouri K, Newton SR, et al (2018) A comparison of three liquid chromatography (LC) retention time prediction models. Talanta 182:371–379.

22. Aalizadeh R, Nika MC, Thomaidis NS (2019) Development and application of retention time prediction models in the suspect and non-target screening of emerging contaminants. J Hazard Mater 363:277–285.

23. Feng C, Xu Q, Qiu X, et al (2021) Evaluation and application of machine learning-based retention time prediction for suspect screening of pesticides and pesticide transformation products in LC-HRMS. Chemosphere 271:129447.

24. Song D, Tang T, Wang R, et al (2024) Enhancing compound confidence in suspect and non-target screening through machine learning-based retention time prediction. Environmental Pollution 347:123763.

25. Bonini P, Kind T, Tsugawa H, et al (2020) Retip: Retention time prediction for compound annotation in untargeted metabolomics. Anal Chem 92:7515–7522.

26. Stanstrup J, Neumann S, Vrhovšek U (2015) PredRet: Prediction of retention time by direct mapping between multiple chromatographic systems. Anal Chem 87:9421–9428.

27. Ruttkies C, Schymanski EL, Wolf S, et al (2016) MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. J Cheminform 8:3.

28. Bach E, Szedmak S, Brouard C, et al (2018) Liquid-chromatography retention order prediction for metabolite identification. Bioinformatics 34:i875–i883.

29. Kretschmer F, Harrieder EM, Hoffmann MA, et al (2024) RepoRT: A comprehensive repository for small molecule retention times. Nat Methods 21:153–155.

30. Rigano F, Arigò A, Oteri M, et al (2021) The retention index approach in liquid chromatography: An historical review and recent advances. J Chromatogr A 1640:461963.

31. Aalizadeh R, Alygizakis NA, Schymanski EL, et al (2021) Development and application of liquid chromatographic retention time indices in HRMS-based suspect and nontarget screening. Anal Chem 93:11601–11611.

32. Boelrijk J, van Herwerden D, Ensing B, et al (2023) Predicting RP-LC retention indices of structurally unknown chemicals from mass spectrometry data. J Cheminform 15:28.

33. van Herwerden D, Nikolopoulos A, Barron LP, et al (2024) Exploring the chemical subspace of RPLC: A data driven approach. Anal Chim Acta 1317:342869.

34. Yap CW (2011) PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474.

35. Hall LM, Hill DW, Menikarachchi LC, et al (2015) Optimizing artificial neural network models for metabolomics and systems biology: An example using HPLC retention index data. Bioanalysis 7:939–955.

36. Aalizadeh R, Nikolopoulou V, Thomaidis NS (2022) Development of liquid chromatographic retention index based on cocamide diethanolamine homologous series (C(n)-DEA). Anal Chem 94:15987–15996.

37. Xue J, Guijas C, Benton HP, et al (2020) METLIN MS2 molecular standards database: A broad chemical and biological resource. Nat Methods 17:953–954.

38. Aisporna A, Benton HP, Chen A, et al (2022) Neutral loss mass spectral data enhances molecular similarity analysis in METLIN. J Am Soc Mass Spectrom 33:530–534.

39. Mohanty I, Mannochio-Russo H, Schweer J V., et al (2024) The underappreciated diversity of bile acid modifications. Cell 187:1801–1818.

40. van Herwerden D, O'Brien JW, Lege S, et al (2023) Cumulative neutral loss model for fragment deconvolution in electrospray ionization high-resolution mass spectrometry data. Anal Chem 95:12247–12255.

41. Guo H, Xue K, Sun H, et al (2023) Contrastive learning-based embedder for the representation of tandem mass spectra. Anal Chem 95:7888–7896.

42. Herwerden D van, Kant E, Jackson M, et al (2024) Modular open-access and open-source julia language toolbox for processing of HRMS data: jHRMSToolBox

43. Chambers MC, MacLean B, Burke R, et al (2012) A cross-platform toolkit for mass spectrometry and proteomics. Nat Biotechnol 30:918–920.

44. Samanipour S, O'Brien JW, Reid MJ, Thomas K V. (2019) Self adjusting algorithm for the nontargeted feature detection of high resolution mass spectrometry coupled with liquid chromatography profile data. Anal Chem 91:10800–10807.

Formulae and Equations

Eq1	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
Eq2	$S_8 = \frac{f_{matched}}{f_{user}} + \frac{f_{matched}}{f_{reference}} + \frac{MS_{tol} - MS_{err,pre}}{MS_{tol}} + \frac{f_{matched}}{f_{reference}} \cdot \frac{MS_{tol} - \lvert MS_{err,frag} \rvert}{MS_{tol}} + \frac{f_{matched}}{f_{reference}} \cdot \frac{2 \cdot MS_{tol} - STD_{err,frag}}{2 \cdot MS_{tol}} + MF_{forward} + MF_{reverse} = S_1 + S_2 + \frac{MS_{tol} - S_3}{MS_{tol}} + S_2 \cdot \frac{MS_{tol} - S_4}{MS_{tol}} + S_2 \cdot \frac{2 \cdot MS_{tol} - S_5}{2 \cdot MS_{tol}} + S_6 + S_7$
Eq3	$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
Eq4	$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$
Eq5	$RTI\ error = RTI_{MF} - RTI_{CNL}$
Eq6	$h_{ii} = x_i^T (X^T X)^{-1} x_i$
Eq7	$Recall = \frac{TP}{TP + FN}$
Eq8	$P(Exposome\ Identification\ for\ each\ RPLC/HRMS\ tensor) = \frac{1}{N}$
Eq9	$P(Exposome\ Identification\ for\ each\ RPLC/HRMS\ tensor\ in\ Top - X\ of\ Hits\ of\ reference\ spectra) = \frac{N}{X}$

Figure Legends:

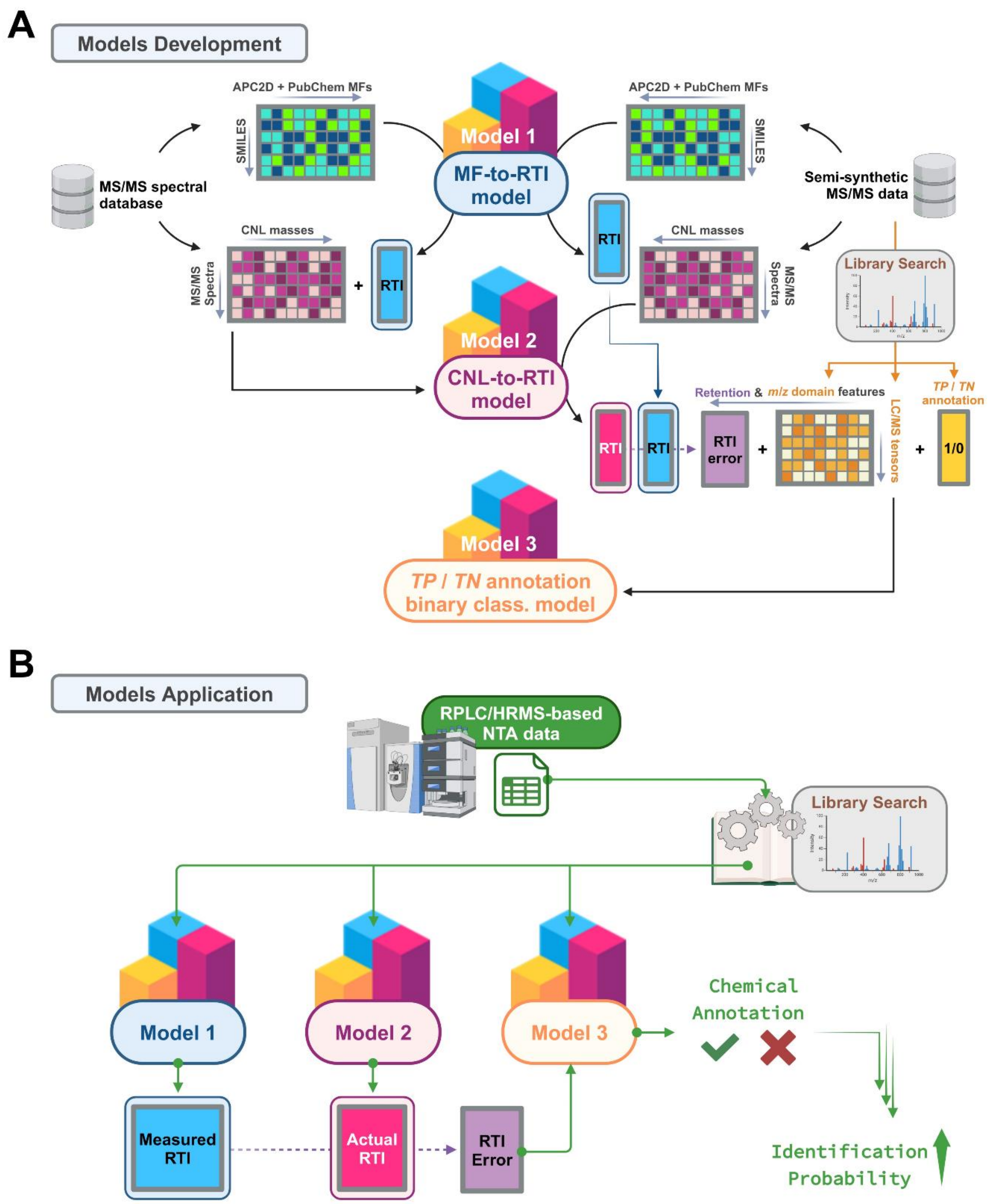
Figure 1: A diagram of the roadmap in this study. **(A)** Three models were developed in this study. Model 1 was a random forest (RF) regression (reg.) model for predicting molecular fingerprint (MF)-derived retention time index (RTI). Model 2 was a RF reg. model for predicting cumulative neutral loss (CNL)-derived RTI. A set of semi-synthetic MS/MS data was prepared. Models 1 and 2 were deployed to predict the expected RTI values for the semi-synthetic data. These data were processed by library search, outputting spectral matching scores for Model 3 training. Model 3 was a k-Nearest neighbors (KNN) binary classification (class.) model was trained by incorporating features from retention and *m/z* domains to predict true positive (*TP*) true negative (*TN*) chemical annotation. **(B)** The 3 developed models were applied to analyze reversed-phase liquid chromatography/high-resolution mass spectrometry (RPLC/HRMS)-based non-targeted analysis (NTA) data. By incorporating the models, the chemical identification probability could be enhanced. *Abbreviations: APC2D, AtomPairs2DFingerprintCount; PubChem, PubChem Fingerprinter.*

Figure 2: Performance assessment of the quantitative structure-retention relationship (QSRR)-based models. Histograms illustrating the distributions of true retention time indices (RTIs) of calibrants used to **(A)** train and **(B)** test Model 1 and the values predicted from Models 1 and 2. Boxplots presenting the variances of the predicted values to the true values of calibrants used to **(C)** train and **(D)** test Model 1. Correlation plots showing correspondence between molecular fingerprint (MF)-derived and cumulative neutral loss (CNL)-derived RTIs for **(E)** training and **(F)** testing datasets of Model 2. Green dots representing compounds used in Model 1's training, yet sky-blue dots indicating new compounds within the RPLC chemical space after applicability domain filtering.

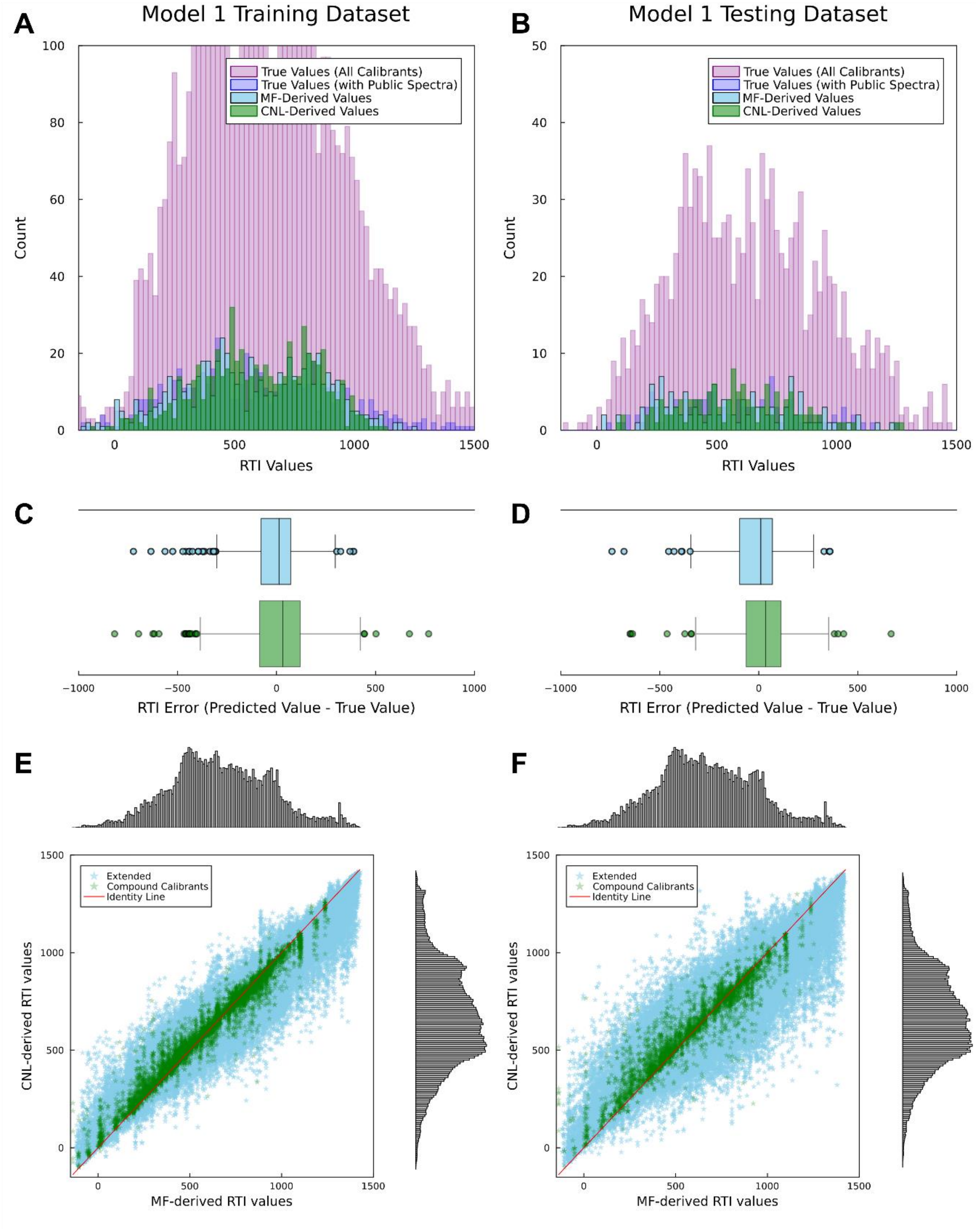
Figure 3: Performance assessment of the true positive/true negative annotation binary classification model. (A) Receiver operating characteristics curves and **(B)** precision-recall curves for the retention time error (RTI) inclusive and exclusive models. **(C)** True positive (*TP*) and false negative rates and **(D)** false discovery rates (FDR) across various probability of true positive (*P*(1) or *P*(*TP*)) thresholds. **(E)** Recalls of individual LC-MS/MS samples at different *P*(*TP*) cut-off thresholds. Pesticide standards were part of the LC/MS Pesticide Comprehensive Mix Kit (Part Number 5190-0551) from Agilent Technologies (PestMix). Pesticide mixtures that were spiked in solvent blank and 10× and 100× diluted black tea matrix respectively to result 1 ppm final solutions were analyzed. For blank spike (No Tea) samples, various final solution concentrations (1, 2.5, 5, 10, 25, 50, 100, and 1,000 ppb) were analyzed. **(F)** Feature permutation importance analysis for all k-nearest neighbors models with Matthews correlation coefficient scores >0.30 on blank spike dataset during model selection and hyperparameter tuning. **(G)** Integration of machine learning (ML) with Universal Library Search Algorithm (ULSA) enhanced the average identification probabilities by 24% and 25% for the blank spike and the samples with the presence of 100× diluted tea matrix, equivalent to 54.5% and 52.1% increments than solely ULSA implementation, yet the presence of 10× tea matrix reduced this improvement to and 21%, equivalent to 46.7% increment. *Abbreviations: nd, not detected.*

Figure 4: Performance assessment by ranking. Integration of machine learning (ML) with Universal Library Search Algorithm (ULSA) slightly enhanced the average identification probability by 1–3% in top-5, top-3, and top-1 hits. True positive (*TP*) hit rates were the ratio of molecular annotations with any *TP* hit at least once in the top ranks to the total identified spectra. *Abbreviations: nd, not detected.*

Figure 1

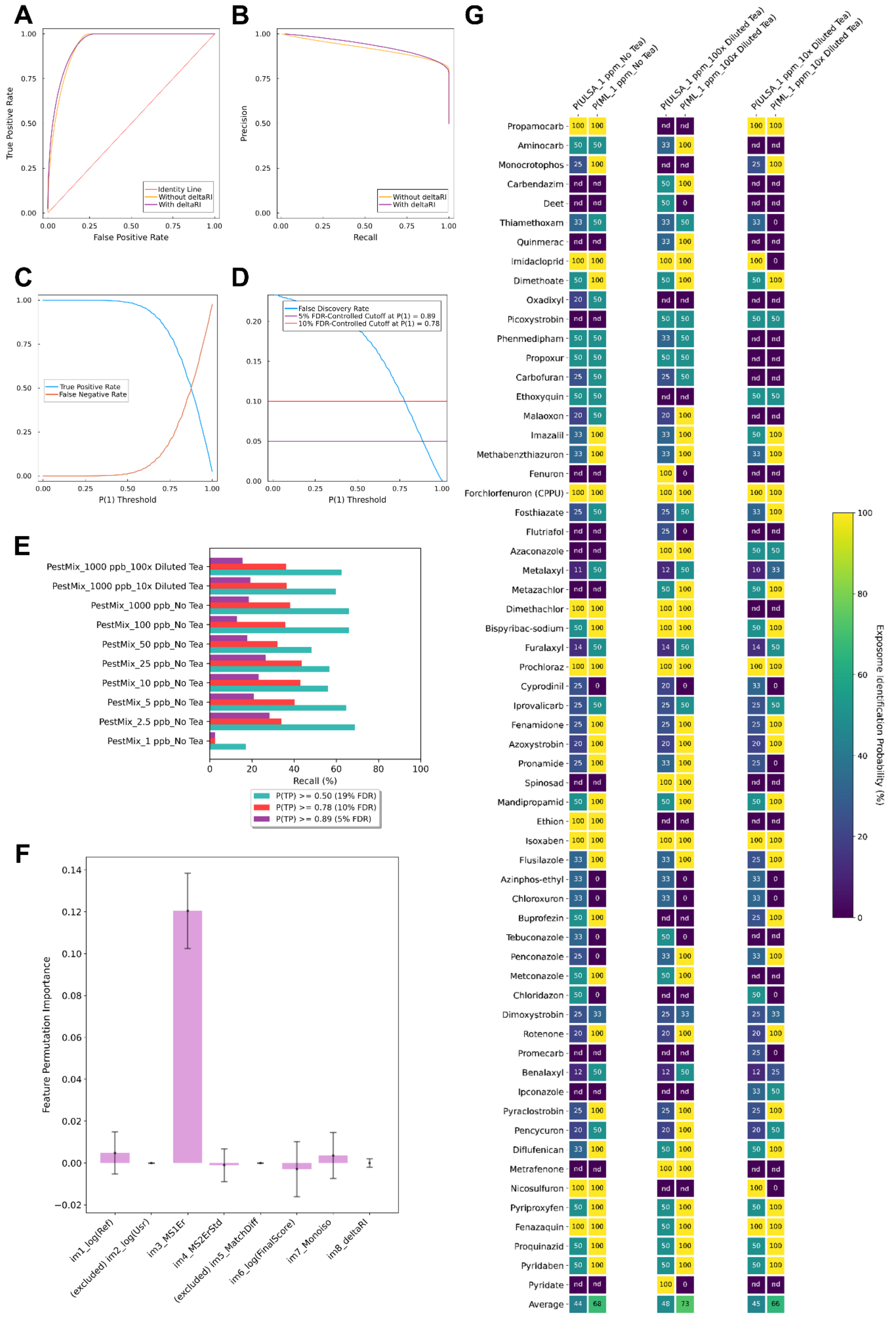

A
Models Development
APC2D + PubChem MFs
SMILES
MS/MS spectral database
CNL masses
MS/MS Spectra
+
RTI
Model 1
MF-to-RTI model
APC2D + PubChem MFs
SMILES
Semi-synthetic MS/MS data
RTI
CNL masses
MS/MS Spectra
Library Search
Model 2
CNL-to-RTI model
Retention & m/z domain features
TP / TN annotation
LC/MS tensors
RTI
RTI
RTI error
+
+
1/0
Model 3
TP / TN annotation binary class. model
B
Models Application
RPLC/HRMS-based NTA data
Library Search
Model 1
Model 2
Model 3
Chemical Annotation
✓ ✗
Measured RTI
Actual RTI
RTI Error
Identification Probability

647 **Figure 2**



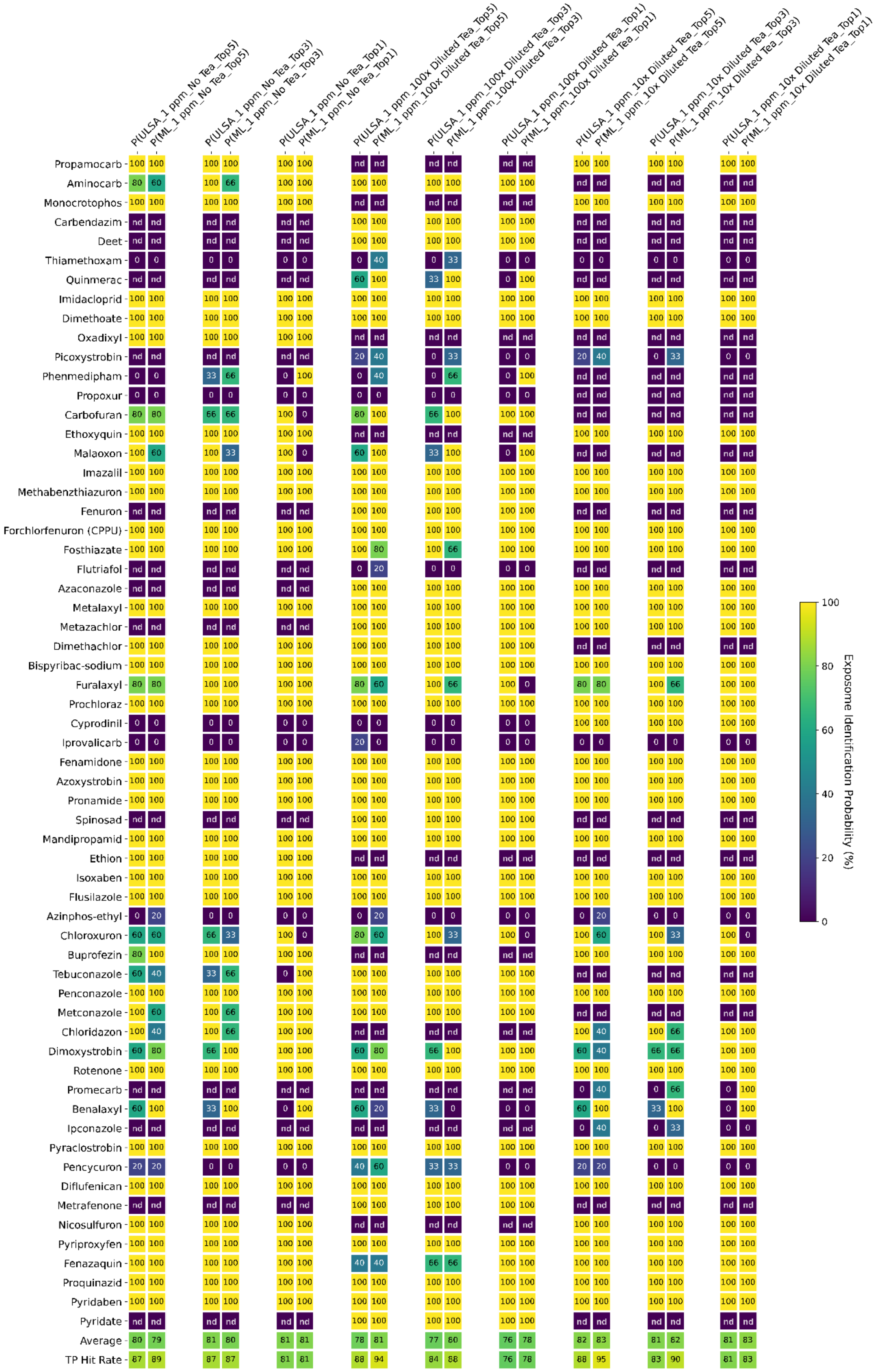
648

Figure 3



	P(ULSA_1 ppm_No Tea)	P(ML_1 ppm_No Tea)	P(ULSA_1 ppm_100x Diluted Tea)	P(ML_1 ppm_100x Diluted Tea)	P(ULSA_1 ppm_10x Diluted Tea)	P(ML_1 ppm_10x Diluted Tea)
Propamocarb	100	100	nd	nd	100	100
Aminocarb	50	50	33	100	nd	nd
Monocrotophos	25	100	nd	nd	25	100
Carbendazim	nd	nd	50	100	nd	nd
Deet	nd	nd	50	0	nd	nd
Thiamethoxam	33	50	33	50	33	0
Quinmerac	nd	nd	33	100	nd	nd
Imidacloprid	100	100	100	100	100	0
Dimethoate	50	100	50	100	50	100
Oxadixyl	20	50	nd	nd	nd	nd
Picoxystrobin	nd	nd	50	50	50	50
Phenmedipham	50	50	33	50	nd	nd
Propoxur	50	50	50	50	nd	nd
Carbofuran	25	50	25	50	nd	nd
Ethoxyquin	50	50	nd	nd	50	50
Malaoxon	20	50	20	100	nd	nd
Imazalil	33	100	33	100	50	100
Methabenzthiazuron	33	100	33	100	33	100
Fenuron	nd	nd	100	0	nd	nd
Forchlorfenuron (CPPU)	100	100	100	100	100	100
Fosthiazate	25	50	25	50	33	100
Flutriafol	nd	nd	25	0	nd	nd
Azaconazole	nd	nd	100	100	50	50
Metalaxyl	11	50	12	50	10	33
Metazachlor	nd	nd	50	100	50	100
Dimethachlor	100	100	100	100	nd	nd
Bispyribac-sodium	50	100	100	100	50	100
Furalaxyl	14	50	14	50	14	50
Prochloraz	100	100	100	100	100	100
Cyprodinil	25	0	20	0	33	0
Iprovalicarb	25	50	25	50	25	50
Fenamidone	25	100	25	100	25	100
Azoxystrobin	20	100	20	100	20	100
Pronamide	25	100	33	100	25	0
Spinosad	nd	nd	100	100	nd	nd
Mandipropamid	50	100	50	100	50	100
Ethion	100	100	nd	nd	nd	nd
Isoxaben	100	100	100	100	100	100
Flusilazole	33	100	33	100	25	100
Azinphos-ethyl	33	0	33	0	33	0
Chloroxuron	33	0	33	0	33	0
Buprofezin	50	100	nd	nd	25	100
Tebuconazole	33	0	50	0	nd	nd
Penconazole	25	0	33	100	33	100
Metconazole	50	100	50	100	nd	nd
Chloridazon	50	0	nd	nd	50	0
Dimoxystrobin	25	33	25	33	25	33
Rotenone	20	100	20	100	20	100
Promecarb	nd	nd	nd	nd	25	0
Benalaxyl	12	50	12	50	12	25
Ipconazole	nd	nd	nd	nd	33	50
Pyraclostrobin	25	100	25	100	25	100
Pencycuron	20	50	20	100	20	50
Diflufenican	33	100	50	100	50	100
Metrafenone	nd	nd	100	100	nd	nd
Nicosulfuron	100	100	nd	nd	100	0
Pyriproxyfen	50	100	50	100	50	100
Fenazaquin	100	100	50	100	100	100
Proquinazid	50	100	50	100	50	100
Pyridaben	50	100	50	100	50	100
Pyridate	nd	nd	100	0	nd	nd
Average	44	68	48	73	45	66

Figure 4



	P(ULSA_1 ppm_No Tea_Top5)	P(ML_1 ppm_No Tea_Top5)	P(ULSA_1 ppm_No Tea_Top3)	P(ML_1 ppm_No Tea_Top3)	P(ULSA_1 ppm_No Tea_Top1)	P(ML_1 ppm_No Tea_Top1)	P(ULSA_1 ppm_100x Diluted Tea_Top5)	P(ML_1 ppm_100x Diluted Tea_Top5)	P(ULSA_1 ppm_100x Diluted Tea_Top3)	P(ML_1 ppm_100x Diluted Tea_Top3)	P(ULSA_1 ppm_100x Diluted Tea_Top1)	P(ML_1 ppm_100x Diluted Tea_Top1)	P(ULSA_1 ppm_10x Diluted Tea_Top5)	P(ML_1 ppm_10x Diluted Tea_Top5)	P(ULSA_1 ppm_10x Diluted Tea_Top3)	P(ML_1 ppm_10x Diluted Tea_Top3)	P(ULSA_1 ppm_10x Diluted Tea_Top1)	P(ML_1 ppm_10x Diluted Tea_Top1)
Propamocarb	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100
Aminocarb	80	60	100	66	100	100	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Monocrotophos	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100
Carbendazim	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Deet	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Thiamethoxam	0	0	0	0	0	0	0	40	0	33	0	0	0	0	0	0	0	0
Quinmerac	nd	nd	nd	nd	nd	nd	60	100	33	100	0	100	nd	nd	nd	nd	nd	nd
Imidacloprid	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Dimethoate	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Oxadixyl	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
Picoxystrobin	nd	nd	nd	nd	nd	nd	20	40	0	33	0	0	20	40	0	33	0	0
Phenmedipham	0	0	33	66	0	100	0	40	0	66	0	100	nd	nd	nd	nd	nd	nd
Propoxur	0	0	0	0	0	0	0	0	0	0	0	0	nd	nd	nd	nd	nd	nd
Carbofuran	80	80	66	66	100	0	80	100	66	100	100	100	nd	nd	nd	nd	nd	nd
Ethoxyquin	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100
Malaoxon	100	60	100	33	100	0	60	100	33	100	0	100	nd	nd	nd	nd	nd	nd
Imazalil	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Methabenzthiazuron	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Fenuron	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Forchlorfenuron (CPPU)	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Fosthiazate	100	100	100	100	100	100	100	80	100	66	100	100	100	100	100	100	100	100
Flutriafol	nd	nd	nd	nd	nd	nd	0	20	0	0	0	0	nd	nd	nd	nd	nd	nd
Azaconazole	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	100	100	100	100	100	100
Metalaxyl	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Metazachlor	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	100	100	100	100	100	100
Dimethachlor	100	100	100	100	100	100	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Bispyribac-sodium	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Furalaxyl	80	80	100	100	100	100	80	60	100	66	100	0	80	80	100	66	100	100
Prochloraz	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Cyprodinil	0	0	0	0	0	0	0	0	0	0	0	0	100	100	100	100	100	100
Iprovalicarb	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
Fenamidone	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Azoxystrobin	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Pronamide	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Spinosad	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Mandipropamid	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Ethion	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd
Isoxaben	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Flusilazole	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Azinphos-ethyl	0	20	0	0	0	0	0	20	0	0	0	0	0	20	0	0	0	0
Chloroxuron	60	60	66	33	100	0	80	60	100	33	100	0	100	60	100	33	100	0
Buprofezin	80	100	100	100	100	100	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100
Tebuconazole	60	40	33	66	0	100	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Penconazole	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Metconazole	100	60	100	66	100	100	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Chloridazon	100	40	100	66	100	100	nd	nd	nd	nd	nd	nd	100	40	100	66	100	100
Dimoxystrobin	60	80	66	100	100	100	60	80	66	100	100	100	60	40	66	66	100	100
Rotenone	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Promecarb	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	0	40	0	66	0	100
Benalaxyl	60	100	33	100	0	100	60	20	33	0	0	0	60	100	33	100	0	100
Ipconazole	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	nd	0	40	0	33	0	0
Pyraclostrobin	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Pencycuron	20	20	0	0	0	0	40	60	33	33	0	0	20	20	0	0	0	0
Diflufenican	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Metrafenone	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Nicosulfuron	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100
Pyriproxyfen	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Fenazaquin	100	100	100	100	100	100	40	40	66	66	100	100	100	100	100	100	100	100
Proquinazid	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Pyridaben	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
Pyridate	nd	nd	nd	nd	nd	nd	100	100	100	100	100	100	nd	nd	nd	nd	nd	nd
Average	80	79	81	80	81	81	78	81	77	80	76	78	82	83	81	82	81	83
TP Hit Rate	87	89	87	87	81	81	88	94	84	88	76	78	88	95	83	90	81	83

Table 1: Summary of the Machine Learning Models' Predictive Power.

ML[a] model	Model 1	Model 2		Model 3 (with feature "DeltaRi"; without "DeltaRi")		
Function	to predict RTI[b] from MF[c]	to predict RTI[b] from CNL[d]		to predict acceptance/rejection decision of a ULSA[e]-suggested molecular annotation from a MS/MS spectrum		
Performance metric		R^2	RMSE[f]	Weighted $F1$ score	MCC[g]	Recall
Training		0.9634	51.8	0.89; 0.90	0.77; 0.79	0.99; 1.00
Testing		0.8824	92.8	0.89; 0.89	0.77; 0.79	0.99; 0.99
Validation				0.65; 0.65	0.30; 0.30	0.66; 0.66
Real Sample						0.60; 0.54

[a] Machine learning.

[b] Retention time index.

[c] Molecular fingerprint.

[d] Cumulative neutral loss.

[e] Universal Library Search Algorithm.

[f] Root mean square error.

[g] Matthews correlation coefficient.