Chemical Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: A. G. Beck, R. Anyaeche, P. Wijewardhane, S. Iyer, Y. Fu, J. K. Liu, J. Zhang, K. Alzarieni, E. Feng, R. T. Hilger, C. J. Welch, H. I. Kenttamaa and G. Chopra, *Chem. Sci.*, 2026, DOI: 10.1039/D5SC07324C.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

View Article Online
DOI: 10.1039/D5SC07324C

Interpretable Machine Learning-based Automated HPLC/MS^2 Platform using Ion-Molecule Reactions for the Identification of Functionalities in Analytes

Armen G. Beck,[a,+] Ruth O. Anyaeche,[a,+] Prageeth Wijewardhane,[a,+] Sanjay Iyer,[a] Yue Fu,[a] Judy Kuan-Yu Liu,[a] Jifa Zhang,[a] Kawthar Z. Alzarieni,[b] Erlu Feng,[a] Ryan T. Hilger,[a] Christopher Welch,[d] Hilkka I. Kenttämaa,[a*] Gaurav Chopra[a,c*]

[a] Department of Chemistry, Purdue University, 560 Oval Drive, West Lafayette, IN, USA.

[b] Department of Medicinal Chemistry and Pharmacognosy, Faculty of Pharmacy, Jordan University of Science and Technology, P.O. Box 3030, Ar-Ramtha Street, Irbid 22110, Jordan

[c] Department of Computer Science (*by courtesy*), Purdue Institute for Drug Discovery, Regenstrief Center for Healthcare Engineering, Purdue Center for Cancer Research, Purdue Institute for Inflammation, Immunology and Infectious Disease, Purdue Institute for Integrative Neuroscience, West Lafayette, IN 47909 USA

[d] Indiana Consortium for Analytical Science & Engineering (ICASE), Indianapolis, Indiana 46202, USA

+These authors contributed equally to this work

* E-mail: hilkka@purdue.edu; gchopra@purdue.edu

Abstract

Identification of unknown compounds in complex mixtures is a time-consuming and challenging problem in several areas of chemistry. High-performance liquid chromatography (HPLC) coupled to tandem mass spectrometry (MS^2) based on collision-activated dissociation (CAD) is a standard approach used to identify unknown compounds in complex mixtures. However, CAD often produces similar fragmentation patterns for isomeric or related ionized analytes, which makes it difficult to differentiate between similar ions. MS/MS methods based on diagnostic gas-phase ion-molecule reactions provide a powerful, predictable, and reliable alternative for the differentiation of isomeric or similar ions via the identification of their specific functional groups. However, the interpretation of the experimental results, the selection of appropriate neutral reagents for new analytes, and the optimization of the conditions for reagent introduction is a manual, time-consuming and challenging process. We have developed a chemical graph-based interpretable machine learning approach that enables automated identification of functionalities in

View Article Online
DOI: 10.1039/D5SC07324C

previously unknown protonated analytes, which facilitates the differentiation of isomeric or otherwise similar compounds. Furthermore, this approach significantly advances prior methods used to study ion–molecule reactions by enabling, for the first time, automated selection of neutral reagents for previously unstudied analytes and algorithmic optimization of reagent-dependent pulsing-in and pumping-out times for reagents introduced into the mass spectrometer via pulsed valves. This study establishes a foundation for fully automated HPLC/MS^2 platforms, enabling the differentiation of similar unknown compounds in complex mixtures with broad applications across chemical sciences.

1. Introduction

High-performance liquid chromatography (HPLC) coupled with tandem mass spectrometry (MS^2) based on collision-activated dissociation (CAD) is a widely used technique to obtain structural information for unknown compounds in complex mixtures.[1–4] However, this technique often produces uninformative data,[5] usually cannot be used to identify specific functional groups in ionized analytes,[6,7] and rarely can be used to differentiate isomeric ions.[5,8–10] Further, examination of several authentic ions is usually required in order to identify an analyte ion. Even then, the identification may fail as many isomeric ions fragment in the same manner and because CAD may cause isomerization. On the other hand, MS^2 methods based on diagnostic and predictable gas-phase ion-molecule reactions have been utilized successfully to differentiate many isomeric ions and to identify specific functional groups in unknown ionized analytes.[11–19] Furthermore, this can be done without the aid of standards because ion-molecule reactions are highly predictable after their mechanisms have been delineated. Ion-molecule reactions have been utilized, for example, to differentiate between isomers including isomeric glucuronides[10,20,21] and to identify *N*-nitrosamines,[19] carboxylic acids,[12,22] amides,[13] and amines.[23] Most of these experiments involve protonation of the analytes by using ionization techniques such as atmospheric pressure chemical ionization[23] (APCI) or electrospray ionization (ESI).[6,12] The protonated analyte molecules are transferred from the ion source into a reaction region, isolated, and allowed to react with neutral reagents. This is usually performed by using ion trap instruments, such as Fourier-transform ion cyclotron resonance (FT-ICR) or quadrupole ion trap mass spectrometers,[17,24]

View Article Online
DOI: 10.1039/D5SC07324C

or triple quadrupole instruments.[25] Only minor modifications need to be performed on such commercially available instruments.[26]

Gas-phase ion-molecule reactions of protonated compounds with reagents such as tris(dimethylamino)borane (TDMAB), trimethyl borate (TMB), and 2-methoxypropene (MOP) have been studied extensively and can be used for the identification of specific functional groups (sometimes followed by CAD), such as sulfoxide, sulfone, urea, and N-oxide.[17,18,27,28] To achieve this, neutral reagents used in gas-phase ion-molecule reactions can be introduced into a mass spectrometer via a continuous flow.[10,18,28] This approach, however, is limited to the use of one reagent at any given time, preventing high-throughput screening. To address this issue and enable high-throughput screening of complex mixtures, a home-built nine pulsed-valve inlet system has been developed for rapid introduction of several reagents into a linear quadrupole ion trap (LQIT) for ion-molecule reaction studies while the analytes are eluting from an HPLC.[29] The mass spectrometry experiments can be automated and are fast enough to be performed during HPLC separation.

However, the design of the pulsing sequence for the reagents (the time period during which the reagents are pulsed into the ion trap (pulsing-in time) and the time period during which they are pumped out (pumping-out time)) requires prior knowledge of the appropriate pulsing-in and pumping-out times for each reagent. Furthermore, interpretation of the results and prediction of a neutral reagent for the identification of a new, previously unstudied functionality is challenging. While the method used thus far for the screening of suitable reagents has been effective, it is time-consuming and resource-intensive, requiring consumption of chemicals and solvents. Therefore, implementing an automated system offers a significant advantage by accelerating reagent selection and minimizing experimental waste, while still being informed by the accumulated expert knowledge. To achieve high-throughput screening of analytes and to avoid human bias in the design of the experiments and in the interpretation of the results, we introduce an automated machine-learning (ML) guided HPLC/MS^2 system to facilitate the identification of functionalities in unknown

View Article Online
DOI: 10.1039/D5SC07324C

analytes in complex mixtures. This ML guided HPLC/MS^2 platform introduces, for the first time, ML automated capabilities for identifying suitable neutral reagents for analytes when no direct structural information is available, optimal selection of reagents, and algorithmic optimization of pulsing-in and pumping-out times, thereby enabling a robust, scalable and automated approach for structural elucidation of compounds in complex mixtures with applications for chemical, biological, and pharmaceutical discovery. Additionally, interpretable ML models, and the learned relations between nontraditional functionalities and reactivity, are introduced for TDMAB and TMB. Overall, this platform allows for the identification of functionalities present in analytes that may not be possible by traditional CAD mediated MS^2 experiments.

2. Experimental Section

2.1. Materials

Pyridine N-oxide (purity 95%), methyl phenyl sulfone (purity ≥98%), phenyl sulfoxide (purity 96%), tris(dimethylamino)borane (purity 99%), and 2-methoxypropene (purity 97%) were obtained from Sigma-Aldrich (Saint Louis, MO, USA). Trimethyl borate (98%) was obtained from Fluka (Buchs, Switzerland). Liquid chromatography/mass spectrometry (LC/MS) grade methanol and water were purchased from Fisher Scientific (Pittsburgh, PA, USA). All chemicals were used as received unless otherwise indicated.

2.2. Mass Spectrometry Instrumentation

All experiments were performed using a slightly modified Thermo Scientific linear quadrupole ion trap (LQIT) mass spectrometer (LTQ XL, Thermo Scientific, San Jose, CA, USA) coupled with an atmospheric-pressure chemical ionization (APCI) source and operated in the positive-ion detection mode. The APCI source conditions were as follows: 300 °C vaporization temperature, 30 (arbitrary units) sheath gas (N_2) flow rate, 10 (arbitrary units) auxiliary gas (N_2) flow rate, and 4.0 kV discharge voltage. The capillary voltage was set to 10 V, and the tube lens voltage was set to 40 V. The voltages for the ion optics were optimized using the automated tuning feature of the instrument, LTQ Tune Plus, for the normal mass range

View Article Online
DOI: 10.1039/D5SC07324C

from *m/z* 50 up to *m/z* 500. The instrument was operated using the LTQ Tune Plus interface and Xcalibur 2.2 software. The mass spectrometer was coupled to a Surveyor Plus high-performance liquid chromatograph (HPLC) equipped with a quaternary pump, autosampler, and photodiode array (PDA) detector. The HPLC/MS^2 experiments were automated using a data-dependent analysis method described in detail below.

Three reagents, 2-methoxypropene, trimethyl borate and tris(dimethylamino)borane (MOP, TMB and TDMAB, respectively), were used in this study. A pulsed valve inlet system was used to introduce each neutral reagent, one after another, as described previously.[29] The pulsed valve inlet included nine pulsed-valve stems containing two-way tee connectors (Parker Hannifin, Cleveland, OH, USA), a plunger used to isolate each pulsed-valve stem from the instrument, and a variable leak valve (purchased from MKS Instruments, Andover, MA, USA) to introduce a continuous flow of helium through the manifold.[29] About 5 μL of MOP, TMB, or TDMAB was injected into the different pulsed-valve injection ports by using a 10 μL syringe. A LabVIEW program used to create the pulsing sequence was linked to a high-voltage pulsed-valve driver (built by the Jonathan Amy Facility for Chemical Instrumentation at Purdue), and connected to the nine pulsed valves (Series 9, VAC-1250 PSIG, Parker Hannifin, Cleveland, OH, USA).[29] The pulsed-valve driver supplied a voltage of up to 300 V to the pulsed valves. Each pulsed valve was either triggered manually or automatically (upon ion isolation) by using the LabVIEW program. The LabVIEW program was used to generate different pulse sequences, *i.e.*, different times for opening the pulsed valves with varying timed delays between the pulses.

2.3. **High-Performance Liquid Chromatography**

The above mass spectrometer was coupled to a Surveyor Plus high-performance liquid chromatograph (HPLC) equipped with a quaternary pump, autosampler, and photodiode array (PDA) detector. For HPLC/MS^2 analysis, all samples were introduced using an autosampler with a partial loop of 10 µL injection volume. An equimolar mixture containing 100 ppm of each analyte was prepared in methanol. A

View Article Online
DOI: 10.1039/D5SC07324C

Zorbax SB-C18 column (4.6 × 250 mm, 5 µm particle size, Agilent Technologies) was used. The mobile phase solvents were pure LC/MS grade water (A) and LC/MS grade methanol (B). The gradient was as follows: 0-5 min isocratic elution at 20% B; 5 – 15 min linear increase to 95% B, 15 – 25 min isocratic elution at 95% B, 25 – 27 linear decreases to 20% B, and 27 – 30 min isocratic elution at 20% B. Gradient conditions are expressed in terms of the mobile phase B, assuming that the remainder of the composition is mobile phase A. The flow rate was 500 µL/min, and the column temperature was 30 °C. The eluates were subsequently ionized by APCI and transferred into the linear quadrupole ion trap for analysis. The HPLC/MS^2 experiments were automated using a data-dependent analysis method described in detail below.

2.4. **Data-dependent Analysis**

HPLC/MS^2 experiments were carried out by using the data-dependent scanning feature of Xcalibur 2.2. This feature was used to isolate, one after another, the three most abundant ions detected in the APCI mass spectrum (isolation width 2.0 m/z units; q value 0.25) measured for each eluting compound. This ensured that the protonated analyte was included among the isolated ions, particularly in cases where signals from protonated solvent and solvent related compounds were also present, as solvent-related ions could dominate the spectra. This enabled automated experiments without prior knowledge on the HPLC elution times or MW of the analytes. Additionally, the analytes must produce a signal great enough to overcome noise latently present due to the use of HPLC caused by solvent and other contaminants. A consistent signal-to-noise ratio of at least three was used to identify real signals based on previous work that reports excellent detection limits for this approach (from 50 pM to 250 nM, with the average being 50-100 nM).[30] The isolated ions were allowed to react for 30 ms with the reagents that were introduced, one after another, into the ion trap via the external pulsed-valve inlet system, as described previously.[18] Given the established behavior of these systems and their known fast reaction kinetics, a reaction time of 30 ms was deemed appropriate and adequate for the current study.[17,18,27,28] The first ion isolation event triggered the pulsed-valve system. Reagents were pulsed into the ion trap several times until the experiment had completed (see discussion below). The reagents were introduced sequentially starting with MOP, followed

View Article Online
DOI: 10.1039/D5SC07324C

by TMB and finally TDMAB. The pulsing-in time set in the LabVIEW program for all reagents was 150 μs unless otherwise specified. Since the reagents were introduced into the ion trap several times, a 1 s pumping-out time (unless otherwise specified) was used between the closing of a pulsed valve that introduced MOP and opening of another. The pumping-out times for TMB and TDMAB were set at 1.2 s and 2 s, respectively, unless otherwise noted. In the program used to operate the instrument, the charge state of ions for the detection was set to +2. Though ions were singly charged, this setting allowed for the detection of product ions larger than the protonated analyte.[12]

2.5. **Development and Validation of the Decision Tree Models**

Decision tree models were generated to identify functional groups in protonated analytes based on whether a diagnostic ion-molecule reaction product was formed or not. For each reagent and diagnostic product type (for MOP, models had been previously published), the available reactions (listed in Tables S6–S7) were split into a training set and a small external test set (TDMAB: 30 training + 5 test; TMB: 21 training + 5 test). A diverse training set, comprising aliphatic and aromatic mono- and polyfunctional analytes, as well as compounds bearing functionalities structurally related to the target group, was systematically evaluated to probe selectivity and potential interferences. As summarized in Tables S6 and S7, this breadth of chemical space reflects the established approach routinely taken during the development of ion-molecule reaction-based methods to confirm that a reaction can be considered diagnostic. The training set was used for model fitting and internal validation, while the held-out test set was used to assess generalization.

To document the structural diversity of the external test analytes, we computed the average pairwise Tanimoto similarity for each diagnostic product test set (**Table S12**). The very low similarities observed for the TDMAB models (~0.065) and low diversity within the TMB models (0.30-0.36) confirm that the held-out analytes span within the studied reaction space, thereby providing a meaningful assessment of model generalization.

View Article Online
DOI: 10.1039/D5SC07324C

All reactions were stoichiometrically balanced. RDKit [30] was used to convert input SMILES strings to Morgan fingerprints with a bit length of 2048 and radii of one, two, or three. All input SMILES were canonicalized by RDKit prior to conversion to Morgan fingerprints, ensuring consistent molecular representations across all models. For each reagent and diagnostic product ion, we systematically varied the fingerprint radius and the branching ratio cutoff used to define a "hit" (for example, yield > 0.1), as summarized in Tables S1-S5. Note: the product ion branching ratios correspond to their individual abundances divided by the sum abundance of all product ions. The radius 1 yielded the best performance for the TMB adduct-Me_2O model, whereas radius 2 was optimal for all remaining diagnostic product ion models, and these radii were used for the final decision trees.

Decision tree models were generated to identify functional groups in protonated analytes based on whether a diagnostic ion-molecule reaction product was formed or not. For each reagent and diagnostic product ion type (for MOP, models have been previously published), the available reactions (listed in Tables S6–S7) were split into a training set and a small external test set (five reactions) as mentioned before. The training set was used for model fitting and internal validation, while the held-out test set was used to assess generalization.

To quantify the reliability of the model on the training data, we applied leave-one-out cross validation (LOOCV) for each combination of fingerprint radius and branching ratio cutoff. For every left-out reaction, a decision tree was trained on all remaining reactions and used to predict the left-out reaction label. A confusion matrix comparing predicted and true labels was then computed, and the corresponding kappa statistic was used as a measure of agreement beyond chance. This LOOCV procedure was repeated 100 times for each cutoff to account for internal randomness in tree construction, and the maximum kappa value across these repeats was recorded as the primary evaluation metric with F1 scores and false discovery rates reported as well (Tables S1–S5).

View Article Online
DOI: 10.1039/D5SC07324C

In addition, for each cutoff, we trained an ensemble of 10,000 decision trees on the training set and used this ensemble to generate predictions on the held-out test set. For every test reaction, the fraction of trees predicting the correct outcome ("formed" or "not formed") was computed, providing an estimate of predictive consistency on unseen reactions. The entries labelled "Compounds 1–5" in Tables S1–S5 report these fractions for each test reaction across branching ratio cutoffs and fingerprint radii. Together, the LOOCV kappa values and the test-set consistency guided the choice of optimal fingerprint radius and branching ratio cutoff for each reagent and diagnostic product.

The distribution of "formed" versus "not formed" labels varies across diagnostic product ions, and because this distribution is further influenced by the branching-ratio cutoff used to define a hit (as shown in Tables S1–S5), the kappa statistic provides a more appropriate measure of model performance than raw accuracy, as it accounts for the agreement expected from these changing class proportions.

Once the hyperparameters were selected, a final interpretable decision tree was trained on the full training dataset for each reagent by using the chosen cutoff and fingerprint radius. This step was implemented in the script decision_tree.jl, which also reports 6-fold cross validation accuracy as an additional sanity check. The resulting trees were then analyzed to identify fingerprint bits and corresponding structural motifs that are most predictive of diagnostic reactivity. These motifs were classified as traditional or nontraditional functional groups and used to construct the functional-group identification module and the neutral reagent selection module. Functional groups are classified as traditional or nontraditional based on whether a synthetic organic chemist would typically recognize the corresponding molecular fragment as a known functional group. All machine learning-based decision tree models (**Figure 6b-d**) developed for TMB and TDMAB reagents and their identified functional groups are described in the discussion section. For each reagent, we trained separate decision tree models for each diagnostic product ion channel (for example, for TMB: adduct, adduct - MeOH, and adduct - Me_2O; for TDMAB: adduct, adduct - DMA, and adduct - 2 DMA). Each of these diagnostic product ion trees was trained independently on a relevant set of reactions but with a different binary label indicating whether that specific diagnostic product ion was "formed" or

View Article Online
DOI: 10.1039/D5SC07324C

"not formed". In **Figure 6b and 6d**, each dotted box corresponds to one such diagnostic-product–ion specific tree. For visualization, the trees are displayed side-by-side and conceptually combined into a "composite" reagent-level scheme, but in the implementation each diagnostic product tree is used separately when extracting functional group-dependent patterns and making predictions. These extracted groups form the basis for the functional-group identification and reagent selection modules described later in the manuscript.

A single decision tree model was selected rather than ensemble approaches (e.g., Random Forest) because interpretability was a primary objective of this work. A standalone tree provides an explicit, human-readable set of decision rules that can be directly mapped to functional groups and chemical reactivity patterns, enabling mechanistic insight into diagnostic ion–molecule reactions. In contrast, ensemble models average over many trees, making reconstruction of clear functional-group determinants substantially more difficult. Additionally, the decision tree framework maintains continuity with our previously published MOP model and is well matched to the modest dataset size while remaining highly interpretable. [28]

2.6. **Functional Group Identification Module**

High-resolution accurate-mass measurements were combined with an ensemble of bootstrapped decision tree models and expert-based reactivity heuristics for the prediction of the most probable functionalities in unknown analytes. The decision-tree models predict whether a diagnostic ion–molecule reaction product is formed. Functional-group assignments are subsequently obtained by interpreting these predicted reaction outcomes using the extracted structure–reactivity motifs and established ion–molecule reaction rules.

The module used a mass spectrum measured after an ion-molecule reaction, the identity of the reagent used in the experiment, and the m/z-value of the protonated analyte and its measured elemental composition (performed on a different instrument equipped with a high-resolution orbitrap mass analyzer) as inputs and predicted plausible functional groups as the output. The mass differences between the m/z value of the protonated analyte and any detected product ions were determined. A dataset of the determined mass differences and functional groups that could form the detected diagnostic product ion was prepared based

View Article Online
DOI: 10.1039/D5SC07324C

on machine learning and experimental data. If a certain mass difference value was not found in the dataset, that value was ignored. Finally, further filtering was carried out using the elemental compositions measured for the protonated analytes (using a different instrument) to filter out functional groups that the analyte could not contain (**Figure 5b**). Thus, the machine learning component predicts reaction behavior, while functional-group identification arises from the interpretation of these predictions within the framework of known ion–molecule reactivity.

2.7. Reagent Selection Module

The first step in the design of an experiment that enables successful and rapid identification of functional groups in unknown protonated analytes based on diagnostic ion-molecule reactions is to identify a neutral reagent that reacts with the protonated analyte to form diagnostic products. Therefore, a module was developed for this purpose based on the measured elemental composition of the protonated analyte and its ring and double bond equivalent (RDBE) value. The module contained two methods. One used known and published expert-based data (empirical data) on ion-molecule reactions.[18] The other method was developed using functionalities identified by machine learning-based decision trees. Both methods in the module used elemental compositions and RDBE values of the protonated analyte as the input to provide a prioritized reagent list as the output (**Figure 6a**). If both methods in the script predict the same neutral reagent, then this is ranked to be the first choice for the analyte of interest (**Figure 6b**).

2.8. Optimization of Reagent Pulsing-in and Pumping-out Times

Pulsing-in time is the length of time during which the pulsed valve is open to allow a reagent into the ion trap whereas the pumping-out time is the total time difference between the opening of two pulsed valves. Optimized pulsing-in and pumping-out times, though analyte independent, are reagent dependent and sensitive to day-to-day instrumental/environmental irregularities. To ensure that a reagent is introduced and evacuated effectively, an automated approach for the selection of appropriate pulsing-in and pumping-out times was developed using an in-house metaheuristic package (Paddy).[34] Paddy was integrated into a graphical user interface (GUI) programmed in Python 3 to facilitate easy accessibility to the full software

View Article Online
DOI: 10.1039/D5SC07324C

implementation named Paddy-PUMP. Paddy-PUMP was written with a tkinter front-end user interface and used Paddy for optimization in its back-end code. The pulsing-in and pumping-out times were randomly initiated and allowed to propagate between 70-180 μs and 1-4 s, respectively. Each parameter pair (pulsing-in and pumping-out time combination) was tested in triplicate by introducing the neutral reagent into a mass spectrometer four times and monitoring the length of time needed to pump it out based on the detection of the protonated reagent generated upon reactions with protonated methanol dimers. The resulting data were transferred into the respective extracted ion profiles showing the abundance of the protonated reagent as a function of time.. Pulsing-in and pumping-out time parameters were optimized by considering separation quality of extracted ion profile peaks. A peak resolution (R) (eq. 1) of 2.5 was used as the target quantity for sufficient separation between peaks. The first iteration of the optimization was initiated by randomly generating a set of five pairs of pulsing-in and pumping-out times, with subsequent times being dependent on the evaluated performance of the previous iteration.

$$R = \frac{t_2 - t_1}{w_2 + w_1}$$, t = time point of the peak maximum (min), w = width of extracted ion profile peak at half height,R = resolution.…….(1)

Paddy was used to minimize the difference between the average resolution of sets of paired peaks and 2.5, making R = 2.5 defined as the maximum of the resulting fitness function (eq. 2).

$$\text{Fitness} = -\left|\frac{\sum_{n=1}^{3} R_n}{3} - 2.5\right| \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

A solution to the optimization problem was defined as a parameter pair generating an average resolution value within ±0.05 of 2.5. Determination of the peak width at half height was performed using the 'find peaks' and 'peak widths' functions from the 'signal' module in the SciPy library (Figure S1).[35]

With the intent of excluding noise that would otherwise be selected as peaks, a data-driven threshold for peak selection was developed. Gaussian mixture models (GMMs) were employed using the 'mixture' module in the Scikit Learn library[36] to facilitate the data-dependent aspect when calculating the

View Article Online
DOI: 10.1039/D5SC07324C

threshold for peak selection. This was accomplished by fitting a GMM with six components to a 2-D list comprised of sorted intensity values and their index, with no regularization of covariance matrices. Once mixture models were fit, they were used to assign extracted ion profile data points to one of the six GMM components (Figure S2). Of the six subpopulations, the maximal value of the subpopulation with the lowest average abundance value in the extracted ion profile was used to calculate the peak height threshold (eq. 3).

$$\text{Threshold} = 1500 + 2\left(\max(K_1)\right),\ K_1 = \text{subpopulation with lowest average value} \quad \ldots\ldots\ldots\ldots (3)$$

The GUI, Paddy-PUMP, facilitated optimization of reagent introduction by writing 'recipe' files for the LabVIEW pulsed-valve driver software in an iterative manner. Recipe files were written such that pulse sequences mirrored the parameters of the pulsing-in and pumping-out times generated by Paddy. Each recipe file was written with five preprogrammed times for a standardized set of reagent injections to initialize the recipe file. This served by providing a signature to denote the start of an experiment in the resulting extracted ion profiles, while also regulating the pulsed-valve system after a period of inactivity (Figure S1). The five preprogrammed introductions of a reagent were then followed by the corresponding times of the pulsing-in and pumping-out times generated by Paddy for the given iteration. Each parameter pair was written into the recipe file as a quadruplet with a three-second delay between different parameters. Extracted ion profiles were exported from Thermo Xcalibur Qual Browser and subsequently processed by Paddy-PUMP via the methods mentioned prior in this section. If no solution was generated by the sequences in the recipe file, fitness values were assigned to the respective parameter pairs and were used to propagate a subsequent iteration. If the peak selection pipeline identified a number of peaks other than a quadruplet per pair of parameters, the fitness value was set to -9999999, effectively eliminating it from further propagation. Upon completion of the optimization of pulsing-in and pumping-out times, a textbox notification informed the experimenter that a 'solutions' file containing suitable parameters was generated.

View Article Online
DOI: 10.1039/D5SC07324C

3. Results and Discussion

The goal of this study was to automate a HPLC/MS^2 system based on diagnostic gas-phase ion-molecule reactions for the rapid identification of functionalities in unknown compounds in mixtures without human interaction. The overview of the machine-learning automated HPLC/MS^2 system is discussed first, followed by description of the data dependent experiment used to automate the isolation of ionized analytes eluting from HPLC without knowing their retention times or m/z-values ahead of time. After this, functional-group identification using machine-learning based decision tree model is discussed, followed by the prediction of suitable neutral reagents for new types of analytes. Finally, the optimization of the pulsing-in and pumping-out times of different reagents is discussed.

Ion-molecule reaction experiments began with the introduction of the selected reagents into the nine pulsed-valve inlet system. After this, analytes were injected into the HPLC for separation. A post-column flow splitter was employed to divert the majority of the flow to waste, while approximately 20 µL/min was directed into the mass spectrometer. The analytes were protonated by APCI as they eluted from the HPLC and then transferred into the ion trap. Data dependent scan methods were used to isolate the three most abundant ions detected in the measured APCI mass spectrum, one after each other, and allowed them to react with reagents introduced, one after each other, from the pulsed-valve system. The generated data were used in the machine learning modules to identify the functionalities in the analytes (**Figure 1**). Further, in experiments on previously unknown protonated analytes, the machine learning modules were used to predict suitable neutral reagents (**Figure 1**).

View Article Online
DOI: 10.1039/D5SC07324C

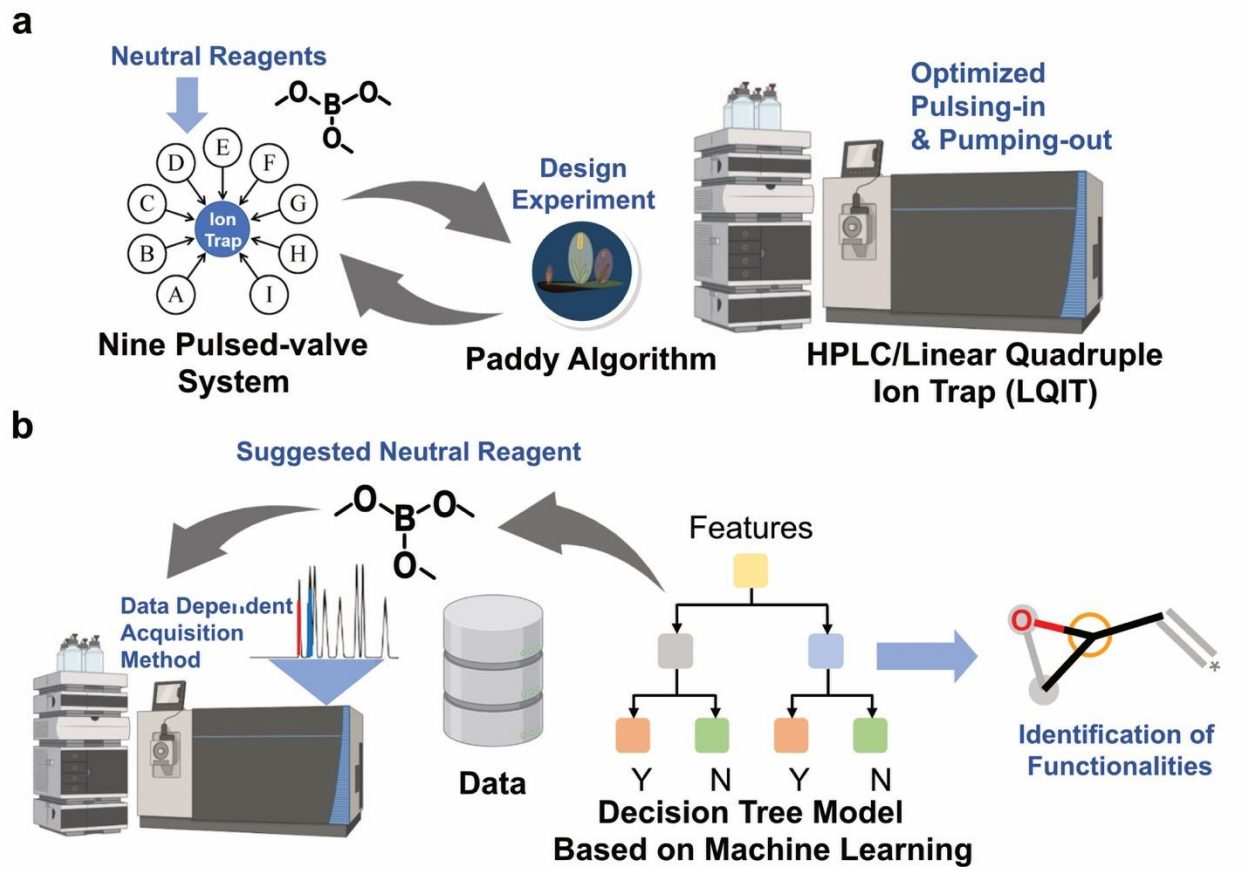

Figure 1. Schematic diagrams showing the overview of the machine-learning automations introduced for HPLC/ MS^2 system based on diagnostic gas-phase ion-molecule reactions for the structural characterization of unknown compounds in mixtures. a) An evolutionary algorithm, paddy, is used to optimize the pulsing-in and pumping-out times of neutral reagents. b) workflows utilizing decision tree models are used for both the identification of functionalities in analytes and the suggestion of neutral reagent for subsequent experimentation.

Data-dependent experiments

Data-dependent scanning features available on the LQIT mass spectrometer can aid the automation of functional-group identification in ion-molecule reaction experiments and improve rapid screening of unknown analytes. In these experiments, the Xcalibur software algorithms signal the mass spectrometer to perform MS^2 experiments for the three most abundant ions detected in the initial MS scan (APCI mass spectrum). An overview of the data dependent analysis method is provided first, followed by an example that demonstrates this approach.

View Article Online
DOI: 10.1039/D5SC07324C

When performing a HPLC/MS^2 experiments manually, the unknown mixture first has to be examined by HPLC/MS to determine the elution time and MW of each analyte before manually selecting a MS^2 program to analyze each analyte. In sharp contrast, an automated experiment based on data dependent scans requires little to no prior knowledge of the analytes, including their retention times or their ions *m/z* values. These LC/MS^2 experiments involved an initial MS scan (**Figure 2**) of the eluted analyte to obtain an APCI mass spectrum. The three most abundant ions detected in the APCI mass spectrum are automatically isolated (isolation width of 2 m/z-units, q value 0.25), one after another, by using separate MS^2 scans (**Figure 2**) and allowed react (for 30 ms) with the neutral reagents introduced, one after another, by the pulsed valves (**Figure 3a, Figures S4** and **S5**).

The LABVIEW program triggers the pulsed-valve system upon the very first ion isolation, which allows the reagents to begin to enter the ion trap from the pulsed valve system, one after another, followed by pumping them out before the next reagent is introduced. The order of the introduced reagents as well as the open time of each pulsed valve (pulsing-in time) and the time delay between two pulses (pumping-out time) determine the *pulsing sequence* for the neutral reagents. One complete pulsed-valve cycle took approximately five seconds (**Figure 2**). This cycle was repeated until the first analyte had eluted from the HPLC. The system automatically paused reagent pulsing until the second analyte began to elute, at which point it automatically re-initiated the sequence, starting with acquisition of an APCI mass spectrum.

View Article Online
DOI: 10.1039/D5SC07324C

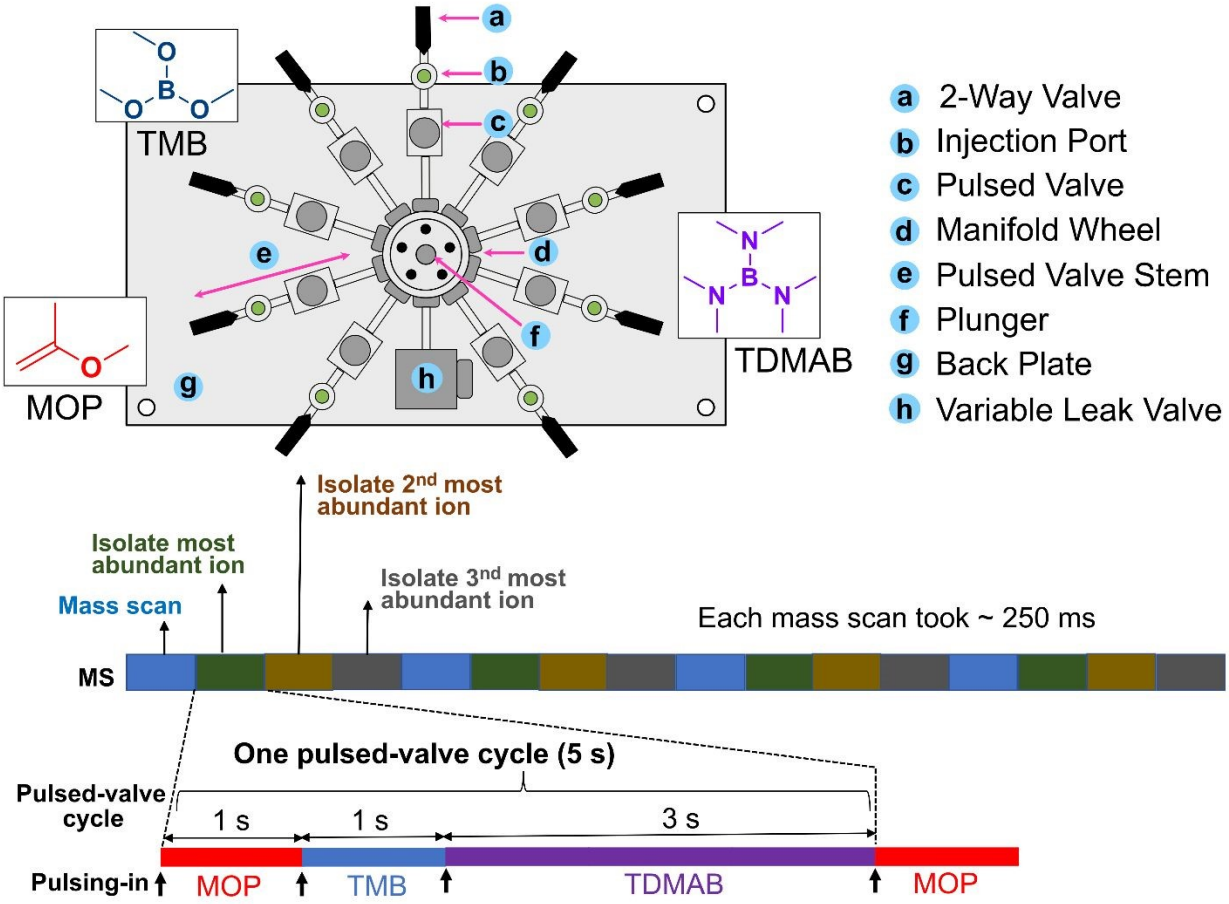

Figure 2. Timeline of one pulsed-valve cycle. An APCI mass spectrum is measured for each eluate. The first ion isolation event triggers the pulsed valve cycle. The three most abundant ions in each APCI mass scan are isolated one after each other, allowed to react with the reagents introduced one after another by the pulsed valve system via a variable leak valve, and the products are detected.

An example of the above process is illustrated below for a mixture of pyridine N-oxide, methyl phenyl sulfone and diphenyl sulfoxide. The analytes eluted from the HPLC one after the other (**Figure S3**), were subjected to APCI, and all ions generated for each analyte were transferred into the ion trap for the measurement of an APCI mass spectrum (**Figure 3b**). The three most abundant ions were isolated, one after another, and allowed to react with MOP, TMB and TDMAB introduced from the pulsed valves, one after another (**Figure 3b**). The most abundant ions (*m/z* 96) generated from the first eluting compound, pyridine

View Article Online
DOI: 10.1039/D5SC07324C

N-oxide, reacted with TDMAB to form a proton transfer product (*m/z* 144) and an adduct that had lost a dimethylamine molecule (*m/z* 194). The last product is diagnostic for the N-oxide functionality. The other two ions detected in the APCI mass spectrum, those of *m/z* 191 and 141, did not react with TDMAB (**Figure 3b**). This process was repeated as methyl phenyl sulfone and diphenyl sulfoxide eluted from the HPLC (**Figures S4** and **S5**). In the case of methyl phenyl sulfone (**Figure S4**), the most abundant ions (*m/z* 157) reacted with TMB to form a methanol adduct (*m/z* 189), an adduct that had lost a dimethyl ether molecule (*m/z* 215) and an adduct that had lost a methanol molecule (*m/z* 229). The last product is diagnostic for the sulfone functionality. The other two ions detected in the APCI mass spectrum (*m/z* 189 and 174) did not react with TMB. For diphenyl sulfoxide (**Figure S5**), the most abundant ions (*m/z* 203) reacted with MOP to form an adduct (*m/z* 275). This product is diagnostic for the sulfoxide functionality. The other two ions detected in the APCI mass spectrum (m/z 405 and 205) did not react with MOP. Therefore, this experiment enabled the identification of the N-oxide, sulfone, and sulfoxide functionalities in the three analytes in one HPLC run.

View Article Online
DOI: 10.1039/D5SC07324C

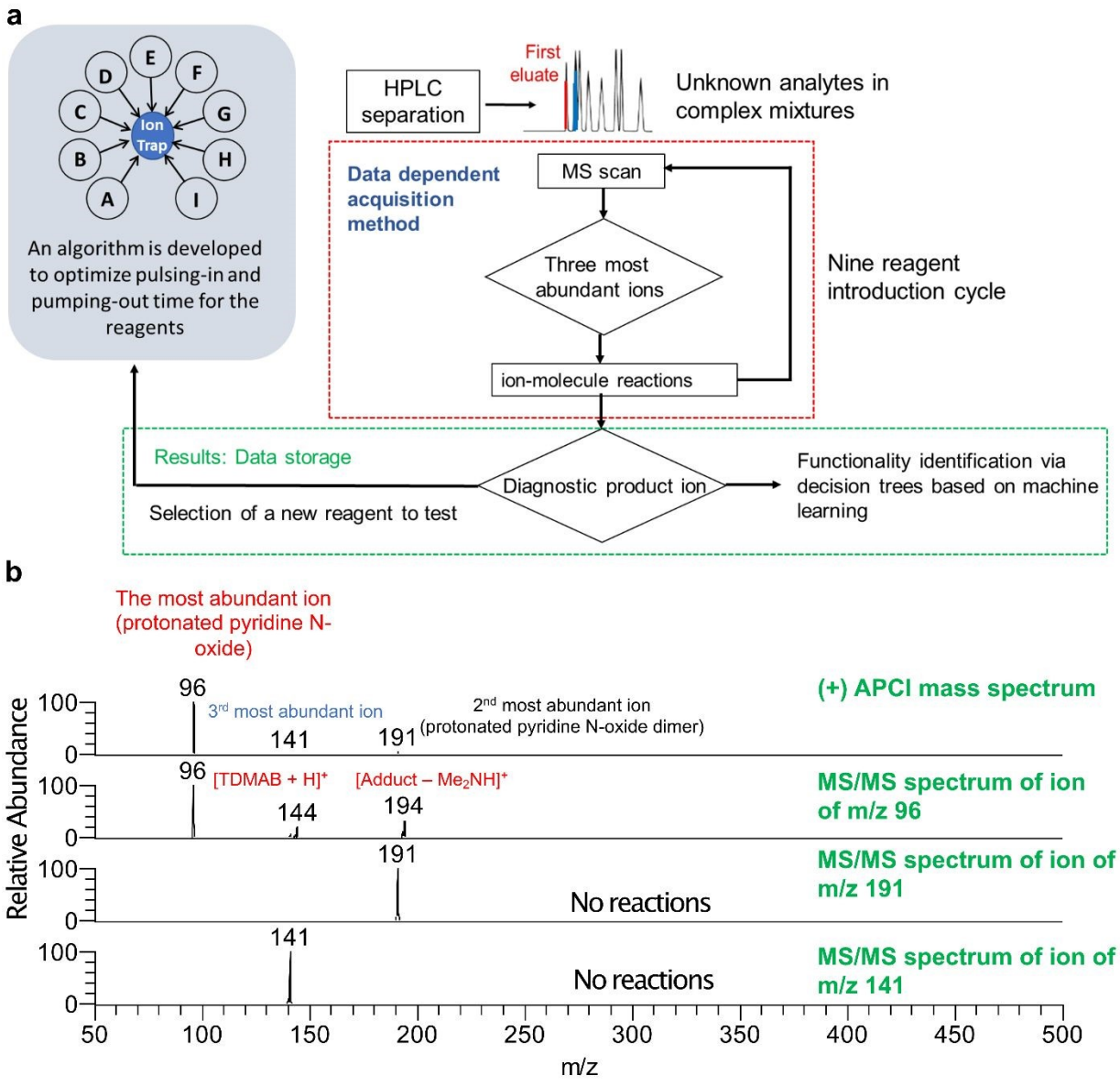

Figure 3. a) Overview of the automation of the data dependent acquisition method. Unknown analytes in complex mixtures are separated by HPLC and APCI mass spectra are measured for them as they elute from the HPLC. The three most abundant ions in each APCI mass spectrum are isolated, one after another, and allowed to undergo ion-molecule reactions with the neutral reagents introduced by the pulsed valve system one after another. **b)** Top: The APCI mass spectrum measured for pyridine N-oxide. Below: the MS^2 mass spectra measured for the three most abundant ions after isolation and reactions with TDMAB for 30 ms.

View Article Online
DOI: 10.1039/D5SC07324C

Functional-group identification module

The functional-group identification module was developed using the decision tree models developed for the three reagents as well as by using experimental data obtained from literature.[18] The mass spectrum measured after an ion-molecule reaction, the reagent used for the experiment, the measured m/z-value of the protonated analyte, and the measured elemental composition of the protonated analyte corresponded to the input for the module. The module searches for all the ions above a pre-defined relative abundance cutoff in the ion-molecule reaction mass spectrum and creates a list of m/z values of the detected ions. Then, the mass differences between the protonated analyte and the detected ion-molecule reaction products are calculated (detailed in the methods section). The module contains a dataset that associates functional groups with specific mass difference values (as explained in the methods section). Finally, the functional groups that are found to correlate with any of the specific mass difference values are selected as the predicted functional groups. Further filtering is performed based on the elemental composition determined for each analyte. However, if the elemental composition of a predicted functional group does not match the elemental composition determined for the unknown analyte, that functional group will be rejected. The functional group prediction module gives two predictions, one based on functionalities identified by the machine learning decision trees, and another based on traditional functional groups (rule-based method) identified using experimental data. Therefore, the rule-based method effectively serves as a built-in baseline. This allows users to directly compare mass-difference rules with the decision tree predictions where the decision tree-identified structural motifs provide additional discriminatory power beyond simple mass-difference matching.

The main objective for the development of the functional-group identification module was to be able to automatically identify functional groups in an unknown protonated analyte given its elemental composition, its gas-phase ion-molecule reaction mass spectrum and other inputs. **Figure 4a** summarizes how the ion-molecule reaction data were extracted and used in the module for functional group identification. The reaction outcome (**Figure 4a**) denotes whether diagnostic product ions are formed upon reactions with a

View Article Online
DOI: 10.1039/D5SC07324C

specific reagent or not. The workflow to identify the functional groups in an unknown analyte is illustrated using the reaction of protonated methyl phenyl sulfone with TMB (**Figure 4b**). The MS^2 mass spectrum measured after the reactions shows three product ions. The mass spectrum is given as an input to the module along with the determined elemental composition, RDBE, and the measured m/z value of the ionized analyte. The module tests all detected product ions with abundances above a predefined threshold value for the mass differences. Then the module matches all identified mass difference values with the values that should be obtained for the diagnostic product ions formed with the TMB reagent. By jointly considering the elemental composition, the RDBE value, and the mass difference associated with each diagnostic product ion, the model effectively constrains the set of chemically plausible assignments and suppresses overlap between functionalities, thereby minimizing the risk of misassignment.

After such ions (diagnostic ions) have been identified, the module matches them with functional groups that can produce such mass differences by using machine learning and expert-based functional groups.

View Article Online
DOI: 10.1039/D5SC07324C

After performing the above steps, the sulfone functional group was predicted as the most probable functional group in the analyte (**Figure 4b**).

To test the reliability of the module, the reactions of protonated diphenyl sulfoxide, methyl phenyl sulfone, and pyridine N-oxide were tested with MOP, TMB and TDMAB reagents. These reactions were not included in the training data set. The module predicted the functional groups shown in **Figure 5a** for these analytes. The results are largely in agreement with the actual functional groups of these analytes. A brief summary of these examples is as follows: MOP selectively reacted with the protonated sulfoxide

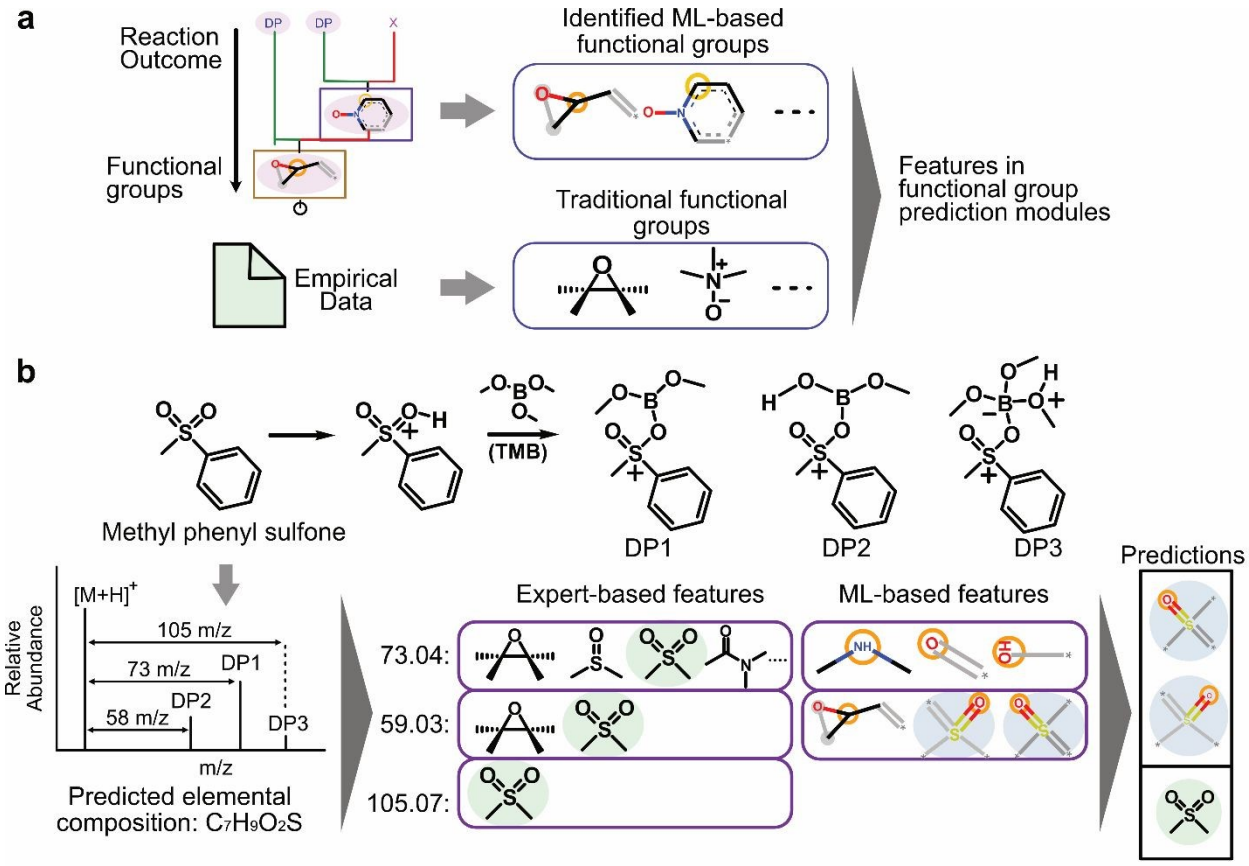

Figure 4. Functional group prediction module. **a)** Functional group identification using machine-learning based decision tree models and experimental data. Decision tree models trained with ion-molecule reaction data obtained for each neutral reagent were used to identify functional groups that can form diagnostic products during gas-phase ion-molecule reactions with specific neutral reagents. These identified functional groups were incorporated as features to the functional-group identification module. **b)** The module uses an MS^2 spectrum of the ion-molecule reactions and the measured elemental composition of the unknown analyte as inputs. The output of the module provides functional groups that are likely to exist in the unknown analyte. Diagnostic product ions formed during the reaction are denoted as P1, P2 and P3.

View Article Online
DOI: 10.1039/D5SC07324C

analyte, TMB selectively reacted with the protonated sulfone analyte, and TDMAB selectively reacted with the protonated N-oxide analyte. These examples illustrate how the diagnostic product decision trees, the machine-learning identified structural motifs, the expert-defined rules, and the functional-group identification module converge to produce chemically intuitive functional-group assignments. Full, step-by-step decision-path case studies are provided in Section S1.

However, in some cases, such as the reaction of protonated pyridine *N*-oxide with MOP, the module returned multiple plausible functional groups, although the correct assignment remained among the options. The mass-difference values for each analyte are shown in **Figure 5b** together with the predicted functional groups. Finally, it should be noted that the module does not return a functional-group prediction when no diagnostic product is formed.

When no diagnostic mass-difference match survives the elemental-composition filter, the platform refrains from suggesting a functional group and flags the analyte as lacking a supported assignment under the current rules.

View Article Online
DOI: 10.1039/D5SC07324C

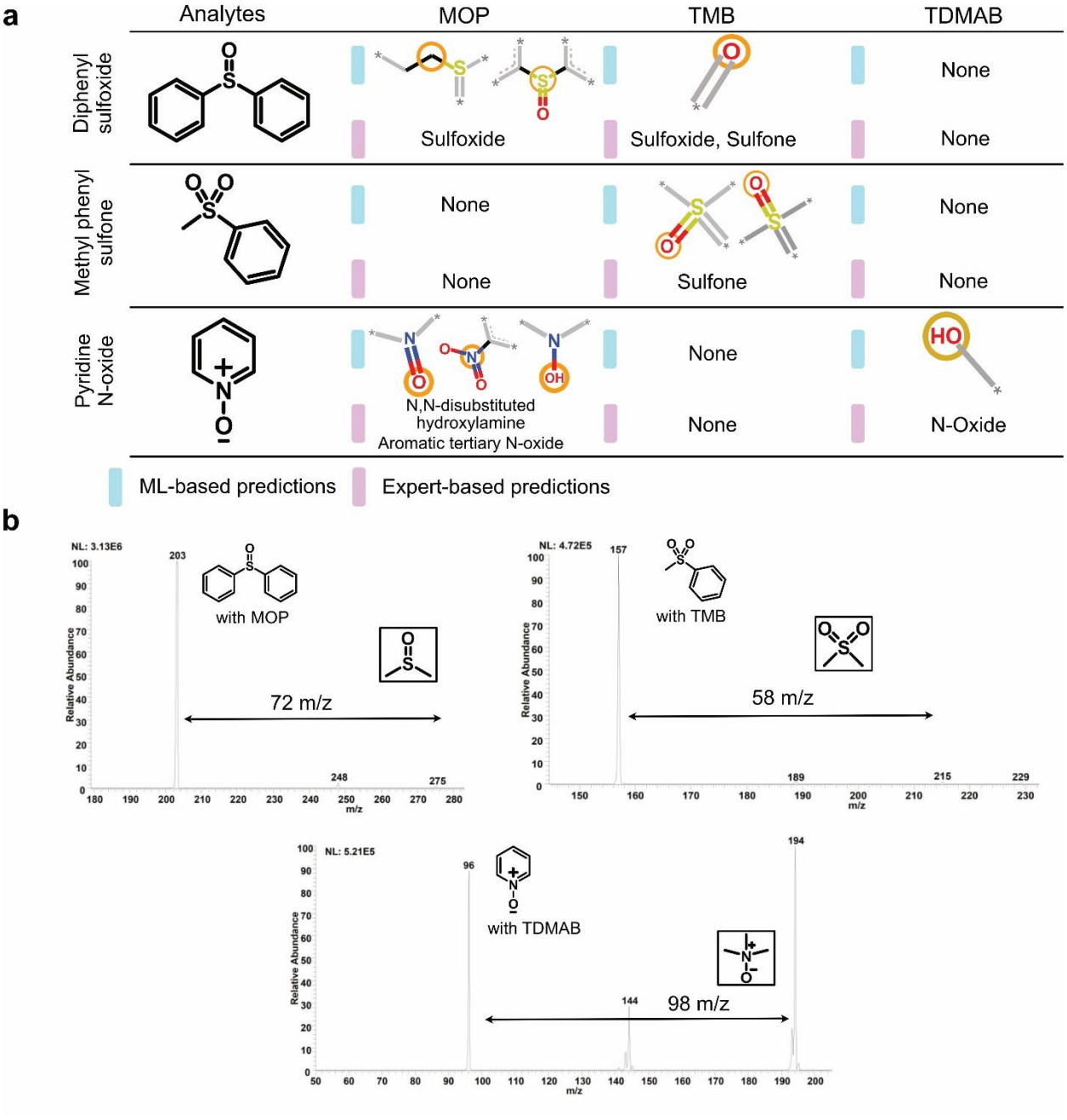

Figure 5. a) Results obtained for several examples by using the developed functional-group identification module. Each row shows results for a different example and the columns show the neutral reagents and the possible functional groups identified based on the reactions with each neutral reagent. b) Mass spectra measured after ion-molecule reactions of the protonated analytes with selected reagents and the mass differences that contributed to the selection of the plausible functional groups.

Performance of decision tree models

View Article Online
DOI: 10.1039/D5SC07324C

As discussed above, machine-learning based decision tree models were developed to identify functional groups based on whether a diagnostic product ion was "formed" or "not formed" during gas-phase ion-molecule reactions. Full decision trees were developed for TMB and TDMAB by using annotated datasets as these decision trees have not been published previously.

To quantify model interpretability, we report model complexity metrics including tree depth (3-5), number of internal nodes (4-12), number of leaves (5-13), and average training-sample support per leaf (1-3). These compact decision trees enable direct mapping of splits to specific Morgan fingerprint bits. Using the decoding procedure described in Section 2.5, we identified the atom environments associated with each major split and grouped them into traditional or nontraditional functional groups. Three representative examples, reactivity of protonated sulfoxides toward TMB, protonated *N*-oxides toward MOP, and protonated tertiary-amines toward TDMAB are presented in Section S1 in the Supporting Information. In all cases, the decision paths reflect intuitive chemical reactivity principles.

The ionic reaction products detected for TMB upon reactions with various protonated analytes were a stable TMB adduct, TMB adduct – MeOH and TMB adduct – Me_2O. Hence, three decision tree models were developed, one for each of these three reaction pathways (**Figure 6b**). Performance of each decision tree was measured using the Cohen's kappa value of the test set. The Cohen's kappa values were 0.74, 0.77, and 0.90 for TMB adduct – MeOH, TMB adduct – Me_2O, and TMB adduct products, respectively (Cohen's kappa values describe the strength of agreement based on inter-model reliability; values > 0.6 indicate a very good or substantial agreement between models and > 0.8 indicates an almost perfect agreement). These kappa values correspond to the best-performing combinations of branching-ratio cutoff and Morgan fingerprint radius of the test reactions.

Morgan fingerprint radii were utilized as hyperparameters to fine-tune the developed model. They were varied to optimize the amount of information that should be represented in a chemical structure such as the analyte structure representation as input. On the other hand, measured product ion branching ratios provide the relative abundances (in %) of primary product ions. These values were used to more accurately

View Article Online
DOI: 10.1039/D5SC07324C

decide whether a product ion was diagnostic or not. The selected branching ratios and Morgan fingerprint radii are shown in bold in **Tables S1**, **S2** and **S3** for the three products: TMB adduct – MeOH, TMB adduct – Me_2O, and TMB adduct. All the data used for preparing the decision trees are shown in **Table S6** and the final developed TMB decision tree is a combination of all three developed decision trees for TMB and is shown in **Figure 6b**. All test set kappa values obtained indicate an extraordinary ability of the model to classify these diagnostic products, and hence to identify functionalities using diagnostic products.

Similarly, another decision tree model was developed for TDMAB reagent. TDMAB generates two diagnostic product ions upon reactions with certain protonated analytes, which have been previously reported in the literature.[18] The product ions are an adduct that had eliminated a dimethylamine (DMA) molecule (adduct-DMA), and an adduct that had eliminated two dimethylamine molecules (adduct-2DMA). Two separate decision trees were developed for these two reaction pathways and combined to get the full tree for TDMAB neutral reagent. The Cohen's kappa values obtained were 0.73 and 1.00 for Adduct-DMA and Adduct-2DMA decision trees respectively. These well-trained decision trees were combined to create the main TDMAB decision tree (**Figure 6a**), which was used to identify functionality features that were used to develop the functional-group prediction module.

Although the resulting decision paths are simple and interpretable, they reflect reagent-specific structure–reactivity relationships learned directly from the experimental data, and these patterns extend beyond what could be captured by a manually curated list of traditional functional groups. All reactions used for model training and testing are experimentally determined ion-molecule reactions, drawn from both previously published our own studies and the measurements reported in this work, as listed in Tables S6-S7. While the external test sets demonstrate generalization across structurally diverse analytes, the current dataset remains limited in size. Expansion to additional analytes and reagents will further strengthen the generalizability of the approach and broaden its applicability.

View Article Online
DOI: 10.1039/D5SC07324C

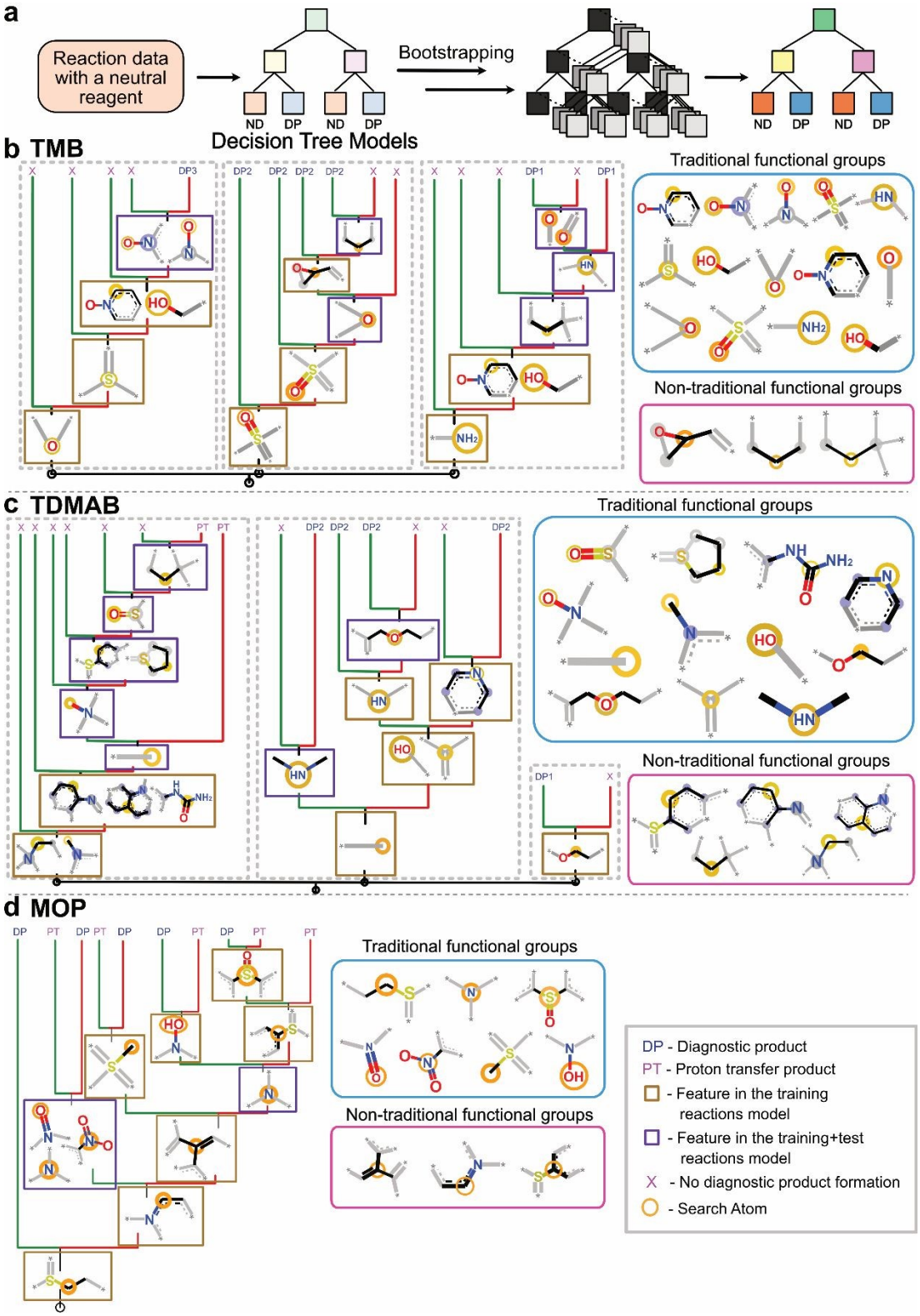

Figure 6. Each dotted box represents a separate decision tree trained for a single diagnostic product (e.g., a specific TMB or TDMAB adduct); the trees are shown together to illustrate the combined reagent-level reactivity, but they are used independently in the functional-group prediction module. **a)** Machine-learning decision tree models were trained using ion-molecule reaction data. Bootstrapping was used to check the robustness of the models for different neutral reagents: **b)** TMB, **c)** TDMAB, and **d)** MOP.

View Article Online
DOI: 10.1039/D5SC07324C

Reagent prediction module

A module was developed to predict reagents that are likely to react in an informative manner with an unknown protonated analyte (module workflow shown in **Figure 7a).** The functional groups that were identified based on machine learning-based decision tree models were used for the reagent prediction module. The module uses a dataset that contains the measured elemental compositions and RDBE values of known compounds that can form diagnostic products with the reagents considered. These data are matched with the measured elemental compositions and the RDBE values of unknown protonated analytes to make the best prediction of functional groups. If the elemental compositions and RDBE values of the predicted functional groups match the measured elemental composition of the unknown protonated analyte and predicted RDBE values of the functional groups of the unknown analyte, then the module will look up the trained decision tree models to propose reagents that may react with those predicted functional groups in a diagnostic manner. If both machine learning and expert-based approaches suggest the same reagent, it will be prioritized as the top prediction (see **Figure 7b**). Finally, all the predictions obtained for reactions of protonated diphenyl sulfoxide, methyl phenyl sulfone, and pyridine *N*-oxide with MOP, TMB and TDMAB reagents (**Figure 7c**) are shown as the predicted reagents based on machine learning as well as by expert-based approach. Interestingly, the predicted reagents matched the actual reagents used to carry out diagnostic gas-phase ion-molecule reactions for these testing reactions (**Table S11**).

View Article Online
DOI: 10.1039/D5SC07324C

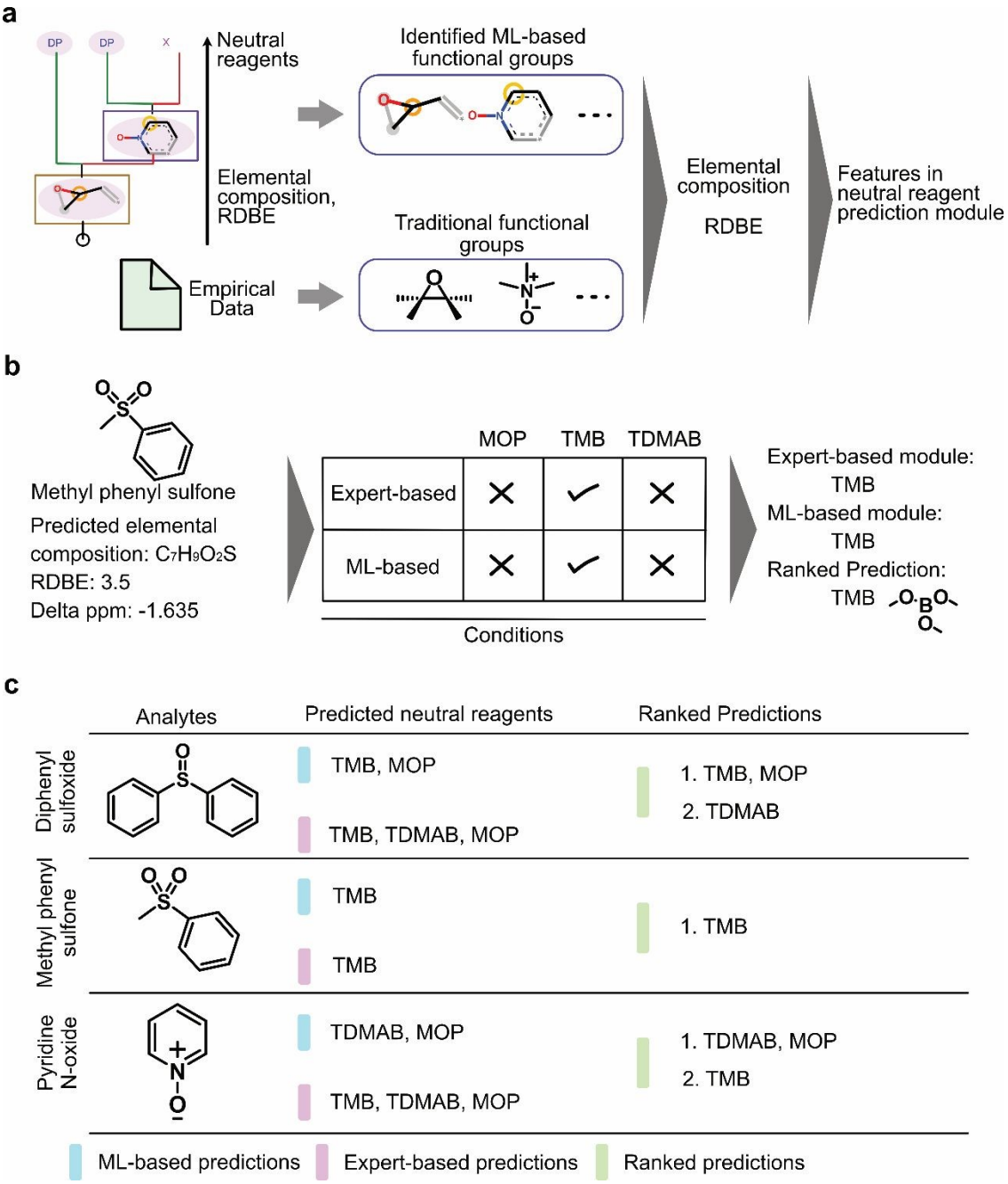

Figure 7. a) Neutral reagent prediction module: Workflow of identifying functional groups by using decision tree models and experimental data. **b)** Neutral reagent prediction process for the reaction of protonated methyl phenyl sulfone with the neutral reagent TMB. **c)** Neutral reagents predicted for testing the structures of the selected analytes.

Optimization of reagent pulsing-in and pumping-out times

The reagent pulsing-in time is the time (in μs) that the pulsed valve is opened to allow the reagent to enter

View Article Online
DOI: 10.1039/D5SC07324C

the ion trap, while reagent pumping-out time is the time (in seconds) used to pump a reagent out of the ion trap before the next pulsed valve is opened (**Figure 8a**). Inadequate pumping-out times of reagents may result in different reagents residing simultaneously in the ion trap, which may complicate the interpretation of the measured data. Previously, the pulsing-in and pumping-out times of the reagents were determined manually. Manual determination is tedious, time-consuming, and inefficient. Paddy-PUMP was utilized to develop automated optimization of the pulsing-in and pumping-out times for MOP, TMB, and TDMAB.

Since the above reagents, MOP, TMB and TDMAB, have been extensively studied, their pulsing-in and pumping-out times that were used previously were first tested by pulsing each reagent into the instrument four times. For MOP, the pulsing-in time used previously[26] was 150μs and pumping-out time 1.0 s. The respective extracted ion profile (**Figure 8b**) for the product ions (*m/z* 73) formed upon reactions of protonated methanol dimer with MOP sightly overlapped near the base of each peak (**Figure 8b**, left). On the other hand, TMB was introduced into the ion trap with a pulsing-in time of μs and pumping-out time of 1.2 s, as described in a previous study[29] The extracted ion profiles (**Figure 8b)** for the ion of *m/z* 105 (protonated TMB) formed upon reactions of TMB with protonated methanol dimer also indicated slightly overlapping peaks (**Figure 8b**, middle). Lastly, TDMAB was introduced into the ion trap μs and a pumping-out time of 1 s, again according to prior work[29] The extracted ion profile (**Figure 8b)** for the ion of m/z 144 (protonated TDMAB) formed upon reactions of TDMAB with protonated methanol dimer had clear and more distinct overlap in its peaks (**Figure 8b**, right). The extracted ion profiles (**Figure 8b**) show that the 1.0 s pumping-out time for MOP and TDMAB and 1.2 s pumping-out time for TMB were insufficient to fully pump away these reagents, suggesting that these reagents remained in the ion trap and did not completely evacuate from the ion trap before the next pulse.

To ensure that each reagent introduced into the ion trap has been fully removed before the introduction of another reagent, the pulsing-in and the pumping-out times of the reagents were optimized. The selection of peak maxima in extracted ion profiles was facilitated via GMM-based heuristics as described above in the methods section (**Figure S2**). This approach resulted in the successful identification

View Article Online
DOI: 10.1039/D5SC07324C

of an optimal pulsing-in schedule for the reagents while excluding noise. Peak apexes were selected to assign a time point (in minutes) to the peak, whereas noise and product ions in the extracted ion profiles with signals below the data dependent thresholds were ignored. An illustrative example of the GMM-based peak method can be found in the supporting information (**Figure S2a** and **b**). Paddy-PUMP was used to optimize pulsing-in and pumping-out times by utilizing peak pair resolutions in the extracted ion profiles, as mentioned above in the methods section (**Figure 8c**). The pulsing-in and pumping-out times for MOP were optimized over the course of six iterations (**Table S8**), with two solutions being generated in the last iteration. The two optimized sets of pulsing-in and pumping-out times were 80 μs and 1.5 s as well as 70 μs and 1.4 s, respectively. The pulsing-in and pumping-out times of TMB were optimized over six iterations (**Table S9**), with the optimized pulsing-in and pumping-out times being 110 μs and 1.9 s, respectively. Lastly, the pulsing-in and pumping-out times of TDMAB were optimized over three iterations (**Table S10**), with the optimized pulsing-in and pumping-out times being 170 μs and 3 s, respectively. These optimized values corresponded to shorter pulsing-in times and slightly longer pumping-out times for the selected reagents compared to the previous values.

Optimized pulsing-in and pumping-out times, produced using Paddy-PUMP for all three reagents, display satisfactory resolutions between the peaks in the extracted ion profiles (**Figure 8d**). Evolutionary pathways for solutions produced using Paddy-PUMP across the parameter space for each reagent are depicted as contour maps in **Figure 8e**. Contour maps displaying the optimization of pulsing-in and pumping-out times with Paddy-PUMP for each individual iteration can be found in the supporting information (**Figure S6-8**). Lastly, a GUI for Paddy-PUMP was developed to facilitate human-in-the-loop experimentation, with successful identification of sufficient pulsing-in and pumping-out times resulting in a pop-up window being displayed to the experimenter (**Figure 8f**). A demo video showcasing the workflow with Paddy-PUMP is provided in the supporting information (https://doi.org/10.5281/zenodo.17173211).

View Article Online
DOI: 10.1039/D5SC07324C

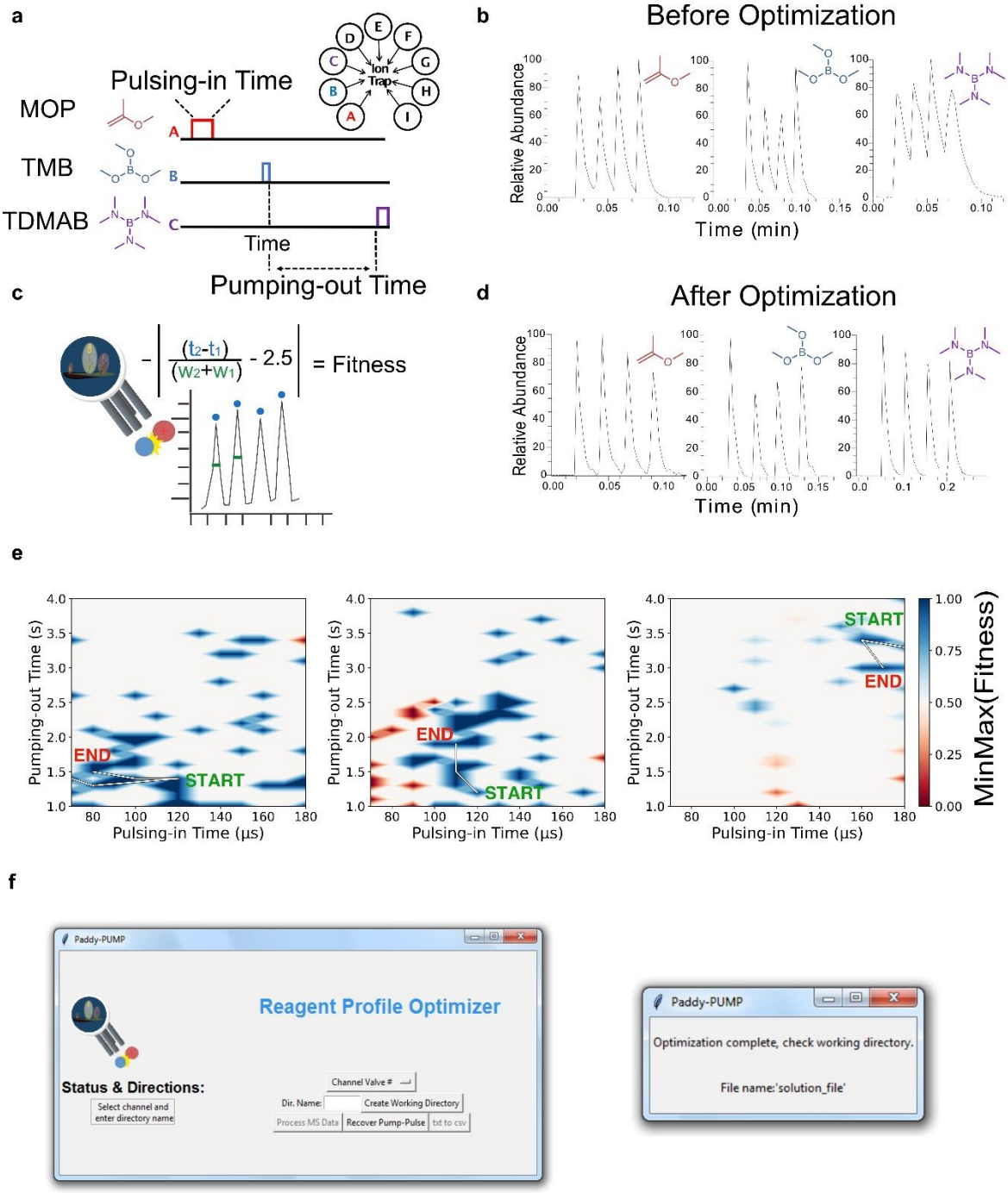

Figure 8. a) Schematic introducing the pulsing-in and pumping-out times of a pulsing sequence when MOP, TMB, and TDMAB are pulsed into the ion trap. **b**) Extracted ion profiles as a function of time are indicative of unoptimized pulsing-in and pumping-out times for MOP, TMB and TDMAB. **c**) Depiction of Paddy-PUMP and a schematic displaying the use of resolution within the fitness function being maximized during optimization. Peak width and peak apex are denoted in green and blue text, respectively, in both the equation and extracted ion profile. **d**) Extracted ion profiles of the product ions of reactions between protonated methanol dimer and neutral reagent of the optimized pulsing-in and pumping-out time 'solutions' produced using Paddy-PUMP. **e**) Contour plots displaying the normalized fitness (via min-max normalization) of the sampled pulsing-in and pumping-out times, and paths taken in generating solutions. **f**) GUI display of Paddy-Pump and pop-up window that informs the experimentor that an optimization experiment is completed.

View Article Online
DOI: 10.1039/D5SC07324C

4. Conclusions

An HPLC/APCI MS^2 LQIT experiment based on diagnostic gas-phase ion-molecule reactions and introduction of reagents via a set of nine pulsed valves were coupled with bootstrapped human interpretable machine learning models to automate the identification of functionalities in protonated analytes in mixtures. Furthermore, an automated system was developed to optimize the reagent pulsing-in and pumping-out times for the sequential introduction of several reagents by use of an in-house evolutionary algorithm.

A decision tree developed[31] previously for MOP was utilized in this research, together with new decision trees developed for TDMAB and TMB. Bootstrapping techniques were used to select reagent-specific decision trees from 10,000 models and the performance was assessed using Kappa statistics to select the best performing model (kappa > 0.7 indicates good inter-model reliability compared to the model prediction by chance). The identification of probable functionalities in the unknown protonated analytes is based on diagnostic product ions formed upon ion-molecule reactions as well as the ring and double-bond equivalents and elemental compositions obtained from high-resolution accurate mass measurements for the protonated analyte. The diagnostic ion-molecule reaction products were identified based on the m/z differences between the protonated analytes and the detected product ions. The machine learning models trained on known diagnostic ion-molecule reactions identified traditional (expert-based) and nontraditional (graph-based) functional group features. Development of a method for the selection of suitable reagents for previously unstudied analytes was based on a machine learning model that utilized the measured elemental composition of the protonated analyte and its ring and double bond equivalent as well as the decision trees developed for the neutral reagents.

The work presented here demonstrates that the combination of machine learning and tandem mass spectrometry based on diagnostic gas-phase ion-molecule reactions can be used to rapidly identify functional groups in unknown protonated compounds directly in mixtures. Incorporating machine learning with ion-molecule reactions will facilitate the interpretation of the data and selection of reagents for previously unstudied analytes. Because the machine learning workflow is reagent-agnostic, the same

View Article Online
DOI: 10.1039/D5SC07324C

training, validation, and functional-group extraction procedures can be applied to historic and newly studied reagents, allowing this platform to be systematically expanded as additional ion–molecule reaction data become available. The Paddy-PUMP method was used to optimize the reagent pulsing-in and pumping-out times. The optimal reagent conditions for MOP (pulsing-in time: 80 μs, pumping-out time 1.5 sec; or pulsing-in time: 70 μs, pumping-out time: 1.4 s), TMB (pulsing-in time: 110 μs, pumping-out time 1.9 s) and TDMAB (pulsing-in time: 170 μs, pumping-out time: 3.0 s) were determined. These optimized values provided shorter pulsing-in times and slightly longer pumping-out times compared to the previously used values. Computer codes, scripts and additional information can be found in GitHub (https://github.com/chopralab/cbm_ml_automation) and in the supporting information (SI).

Demo (SI Supporting Movie: PADDY_PUMP_DEMO.mp4, https://doi.org/10.5281/zenodo.17173211. The demo shows the process from designing to performing experiments for the optimization of pulsing-in and pumping-out times of reagents by use of the Paddy-PUMP app.

Data Availability

All data and computer code related to the manuscript is available at GitHub at https://github.com/chopralab/cbm_ml_automation.

Acknowledgements

This work was supported, in part, by the NSF I/UCRC Center for Bioanalytical Metrology (Award 1916991), the United States Department of Defense USAMRAA award W81XWH2010665 through the Peer Reviewed Alzheimer's Research Program, the National Institutes of Health (NIH) award, R01MH128866 by National Institute of Mental Health, and NIH National Center for Advancing Translational Sciences U18TR004146 and ASPIRE Challenge and Reduction-to-Practice awards to G.C. The Purdue University Center for Cancer Research funded by NIH grant P30 CA023168 is also acknowledged. The authors thank all members of Center for Bioanalytical Metrology for their guidance

View Article Online
DOI: 10.1039/D5SC07324C

and support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing Interests

The authors declare the following competing financial interest(s): G.C. is the Director of the Merck-Purdue Center funded by Merck Sharp & Dohme, a subsidiary of Merck and the co-founder of Meditati Inc., BrainGnosis Inc. and LIPOS BIO Inc. The remaining authors declare no competing interests.

Author Contribution

Armen G. Beck: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization

Ruth O. Anyaeche: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization

Prageeth Wijewardhane: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization

Sanjay Iyer: Software, Methodology, Investigation

Yue Fu: Visualization, Methodology

Judy Kuan-Yu Liu: Visualization, Methodology, Conceptualization

Jifa Zhang: Methodology, Conceptualization

Kawthar Z. Alzarieni: Visualization, Methodology, Conceptualization

Erlu Feng: Methodology

Ryan T. Hilger: Software, Methodology

Christopher Welch: Resources

Hilkka I. Kenttämaa: Writing – review & editing, Supervision, Methodology, Resources, Funding acquisition, Conceptualization

Gaurav Chopra: Writing – review & editing, Supervision, Methodology, Resources, Funding acquisition, Conceptualization, Project administration

View Article Online
DOI: 10.1039/D5SC07324C

References

1 M. S. Lee and E. H. Kerns, LC/MS applications in drug development, *Mass Spectrom Rev*, 1999, 18, 187–279.

2 J. Jakopic, M. M. Petkovsek, A. Likozar, A. Solar, F. Stampar and R. Veberic, HPLC–MS identification of phenols in hazelnut (Corylus avellana L.) kernels, *Food Chem*, 2011, 124, 1100–1106.

3 H. Kanazawa, A. Okada, Y. Matsushima, H. Yokota, S. Okubo, F. Mashige and K. Nakahara, Determination of omeprazole and its metabolites in human plasma by liquid chromatography-mass spectrometry, *J Chromatogr A*, 2002, 949, 1–9.

4 R. G. Cooks, K. L. Busch and G. L. Glish, Mass Spectrometry: Analytical Capabilities and Potentials, *Science (1979)*, 1983, 222, 273–291.

5 A. Shimizu, T. Ohe and M. Chiba, A Novel Method for the Determination of the Site of Glucuronidation by Ion Mobility Spectrometry-Mass Spectrometry, *Drug Metabolism and Disposition*, 2012, 40, 1456–1459.

6 T. M. Jarrell, C. L. Marcum, H. Sheng, B. C. Owen, C. J. O'Lenick, H. Maraun, J. J. Bozell and H. I. Kenttämaa, Characterization of organosolv switchgrass lignin by using high performance liquid chromatography/high resolution tandem mass spectrometry using hydroxide-doped negative-ion mode electrospray ionization, *Green Chem.*, 2014, 16, 2713–2727.

7 S. S. Uppal, S. E. Beasley, M. Scian and M. Guttman, Gas-Phase Hydrogen/Deuterium Exchange for Distinguishing Isomeric Carbohydrate Ions, *Anal Chem*, 2017, 89, 4737–4742.

8 P. C. Schmid, J. Greenberg, T. L. Nguyen, J. H. Thorpe, K. J. Catani, O. A. Krohn, M. I. Miller, J. F. Stanton and H. J. Lewandowski, Isomer-selected ion–molecule reactions of acetylene cations with propyne and allene, *Physical Chemistry Chemical Physics*, 2020, 22, 20303–20310.

9 J. Y. Kong, Z. Yu, M. W. Easton, E. Niyonsaba, X. Ma, R. Yerabolu, H. Sheng, T. M. Jarrell, Z. Zhang, A. K. Ghosh and H. I. Kenttämaa, Differentiating Isomeric Deprotonated Glucuronide Drug Metabolites via Ion/Molecule Reactions in Tandem Mass Spectrometry, *Anal Chem*, 2018, 90, 9426–9433.

10 E. Niyonsaba, M. W. Easton, E. Feng, Z. Yu, Z. Zhang, H. Sheng, J. Kong, L. F. Easterling, J. Milton, H. R. Chobanian, N. R. Deprez, M. T. Cancilla, G. Kilaz and H. I. Kenttämaa, Differentiation of Deprotonated Acyl-, N -, and O -Glucuronide Drug Metabolites by Using Tandem Mass Spectrometry Based on Gas-Phase Ion–Molecule Reactions Followed by Collision-Activated Dissociation, *Anal Chem*, 2019, 91, 11388–11396.

11 J. Somuramasami, B. E. Winger, T. A. Gillespie and H. I. Kenttämaa, Identification and counting of carbonyl and hydroxyl functionalities in protonated bifunctional analytes by using solution derivatization prior to mass spectrometric analysis via ion-molecule reactions, *J Am Soc Mass Spectrom*, 2010, 21, 773–784.

12 S. C. Habicht, N. R. Vinueza, E. F. Archibold, P. Duan and H. I. Kenttämaa, Identification of the Carboxylic Acid Functionality by Using Electrospray Ionization and Ion−Molecule Reactions in a Modified Linear Quadrupole Ion Trap Mass Spectrometer, *Anal Chem*, 2008, 80, 3416–3421.

13 K. M. Campbell, M. A. Watkins, S. Li, M. N. Fiddler, B. Winger and H. I. Kenttämaa, Functional Group Selective Ion/Molecule Reactions: Mass Spectrometric Identification of the Amido Functionality in Protonated Monofunctional Compounds, *J Org Chem*, 2007, 72, 3159–3165.

14 M. A. Watkins, J. M. Price, B. E. Winger and H. I. Kenttämaa, Ion−Molecule Reactions for Mass Spectrometric Identification of Functional Groups in Protonated Oxygen-Containing Monofunctional Compounds, *Anal Chem*, 2004, 76, 964–976.

15 P. Schorr and D. A. Volmer, Using differential ion mobility spectrometry to perform class-specific ion-molecule reactions of 4-quinolones with selected chemical reagents, *Anal Bioanal Chem*, 2019, 411, 6247–6253.

16 Y. Song and R. G. Cooks, Atmospheric pressure ion/molecule reactions for the selective detection

View Article Online
DOI: 10.1039/D5SC07324C

of nitroaromatic explosives using acetonitrile and air as reagents, *Rapid Communications in Mass Spectrometry*, 2006, 20, 3130–3138.

17 H. Sheng, P. E. Williams, W. Tang, J. S. Riedeman, M. Zhang and H. I. Kenttämaa, Identification of the Sulfone Functionality in Protonated Analytes via Ion/Molecule Reactions in a Linear Quadrupole Ion Trap Mass Spectrometer, *J Org Chem*, 2014, 79, 2883–2889.

18 J. K. Liu, E. Niyonsaba, K. Z. Alzarieni, V. M. Boulos, R. Yerabolu and H. I. Kenttämaa, Determination of the compound class and functional groups in protonated analytes via diagnostic gas-phase ion-molecule reactions, *Mass Spectrom Rev*, 2023, 42, 1508–1534.

19 J. K.-Y. Liu, E. Feng, Y. Fu, W. Li, X. Ma, H. Sheng, J. Kong, Y. Liu, M. Hicks, B. Xiang, Z. Liu, J. Pennington and H. I. Kenttämaa, A Diagnostic Nitrosamine Detection Approach for Pharmaceuticals by Using Tandem Mass Spectrometry Based on Diagnostic Gas-Phase Ion-Molecule Reactions, *Anal Chem*, 2022, 94, 13795–13803.

20 L. J. Chyall and H. I. Kenttämaa, Gas-phase reactions of the 4-dehydroanilinium ion and its isomers, *Journal of Mass Spectrometry*, 1995, 30, 81–87.

21 R. O. Anyaeche, K. Z. Alzarieni, R. Kumar, J. R. Milton, J. Kaur, H. Sheng and H. I. Kenttämaa, Differentiation of the Aziridine Functionality from Related Functional Groups in Protonated Analytes by Using Selective Ion-Molecule Reactions Followed by Collision-Activated Dissociation in a Linear Quadrupole Ion Trap Mass Spectrometer, *J Org Chem*, 2023, 88, 8865–8873.

22 G. E. Reid, K. D. Roberts, E. A. Kapp and R. J. Simpson, Statistical and Mechanistic Approaches to Understanding the Gas-Phase Fragmentation Behavior of Methionine Sulfoxide Containing Peptides, *J Proteome Res*, 2004, 3, 751–759.

23 M. Fu, R. J. Eismin, P. Duan, S. Li and H. I. Kenttämaa, Ion–molecule reactions facilitate the identification and differentiation of primary, secondary and tertiary amino functionalities in protonated monofunctional analytes in mass spectrometry, *Int J Mass Spectrom*, 2009, 282, 77–84.

24 J. M. J. Nuutinen, A. Irico, M. Vincenti, E. Dalcanale, J. M. H. Pakarinen and P. Vainiotalo, Gas-Phase Ion−Molecule Reactions between a Series of Protonated Diastereomeric Cavitands and Neutral Amines Studied by ESI-FTICRMS: Gas-Phase Inclusion Complex Formation, *J Am Chem Soc*, 2000, 122, 10090–10100.

25 W. J. Meyerhoffer and M. M. Bursey, Differentiation of the isomeric 1,2-cyclopentanediols by ion-molecule reactions in a triple-quadrupole mass spectrometer, *Organic Mass Spectrometry*, 1989, 24, 169–175.

26 Y. Fu, C. J. Brown, J. T. Johnson, B. M. Marsh, J. R. Gilbert, E. Feng and H. I. Kenttämaa, Modification of a Quadrupole/Orbitrap/Linear Quadrupole Ion Trap Tribrid Mass Spectrometer for Diagnostic Gas-Phase Ion–Molecule Reactions, *J Am Soc Mass Spectrom*, 2023, 34, 426–434.

27 H. Sheng, P. E. Williams, W. Tang, M. Zhang and H. I. Kenttämaa, Identification of the sulfoxide functionality in protonated analytes via ion/molecule reactions in linear quadrupole ion trap mass spectrometry, *Analyst*, 2014, 139, 4296–4302.

28 E. Feng, X. Ma and H. I. Kenttämaa, Characterization of Protonated Substituted Ureas by Using Diagnostic Gas-Phase Ion-Molecule Reactions Followed by Collision-Activated Dissociation in Tandem Mass Spectrometry Experiments, *Anal Chem*, 2021, 93, 7851–7859.

29 J. Y. Kong, R. T. Hilger, C. Jin, R. Yerabolu, J. R. Zimmerman, R. W. Replogle, T. M. Jarrell, L. Easterling, R. Kumar and H. I. Kenttämaa, Integration of a Multichannel Pulsed-Valve Inlet System to a Linear Quadrupole Ion Trap Mass Spectrometer for the Rapid Consecutive Introduction of Nine Reagents for Diagnostic Ion/Molecule Reactions, *Anal Chem*, 2019, 91, 15652–15660.

30 L. F. Easterling, R. Yerabolu, R. Kumar, K. Z. Alzarieni and H. I. Kenttämaa, Factors Affecting the Limit of Detection for HPLC/Tandem Mass Spectrometry Experiments Based on Gas-Phase Ion–Molecule Reactions, *Anal Chem*, 2020, 92, 7471–7477.

31 J. Fine, J. Kuan-Yu Liu, A. Beck, K. Z. Alzarieni, X. Ma, V. M. Boulos, H. I. Kenttämaa and G.

View Article Online
DOI: 10.1039/D5SC07324C

Chopra, Graph-based machine learning interprets and predicts diagnostic isomer-selective ion–molecule reactions in tandem mass spectrometry, *Chem Sci*, 2020, 11, 11849–11858.

32 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J Chem Inf Comput Sci*, 1988, 28, 31–36.

33 G. Landrum, RDKit: A Software Suite for Cheminformatics, Computational Chemistry, and Predictive Modeling, 2013, preprint.

34 A. G. Beck, S. Iyer, J. Fine and G. Chopra, Paddy: an evolutionary optimization algorithm for chemical systems and spaces, *Digital Discovery*, 2025, 4, 1352–1371.

35 P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. Pietro Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C. N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D. A. Nicholson, D. R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G. A. Price, G.-L. Ingold, G. E. Allen, G. R. Lee, H. Audren, I. Probst, J. P. Dietrich, J. Silterra, J. T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J. L. Schönberger, J. V. de Miranda Cardoso, J. Reimer, J. Harrington, J. L. C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N. J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P. A. Brodtkorb, P. Lee, R. T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T. J. Pingel, T. P. Robitaille, T. Spura, T. R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y. O. Halchenko and Y. Vázquez-Baeza, SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat Methods*, 2020, 17, 261–272.

36 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine Learning in Python, *The Journal of Machine Learning Research*, 2011, 12, 2825–2830.

View Article Online
DOI: 10.1039/D5SC07324C

Data Availability

All data and computer code related to the manuscript is available at GitHub at https://github.com/chopralab/cbm_ml_automation.