Radical Fragment Ions in Collision-Induced Dissociation Mass Spectrometry

Shipei Xing[1], Tao Huan[1,*]

[1] Department of Chemistry, Faculty of Science, University of British Columbia, Vancouver Campus, 2036 Main Mall, Vancouver, V6T 1Z1, BC, Canada

[*] Author to whom correspondence should be addressed:

Dr. Tao Huan

Tel: (+1)-604-822-4891

E-mail: thuan@chem.ubc.ca

Website: https://huan.chem.ubc.ca/

Abstract

Collision-induced dissociation (CID) is a common fragmentation strategy in mass spectrometry (MS) analysis. A conventional understanding is that fragment ions generated in low-energy CID should follow the even-electron rule. As such, (de)protonated precursor ions should predominately generate (de)protonated fragment ions with very few radical fragment ions (RFIs). However, the extent to which RFIs present in MS^2 spectra has not been comprehensively investigated. This work uses the latest NIST 20 tandem mass spectral library to investigate of the occurrence of RFIs in CID MS^2 experiments. In particular, RFIs were recognized using their integer double bond equivalent (DBE) values calculated from their annotated molecular formulas. Our study shows unexpected results as 65.4% and 68.8% of MS^2 spectra contain at least 10% RFIs by ion-count (total number of ions) in positive and negative electrospray ionization (ESI) modes, respectively. Furthermore, we classified chemicals based on their compound classes and chemical substructures, and calculated the percentages of RFIs in each class. Results show that "Organic 1,3-dipolar compounds" and "Lignans, neolignans and related compounds" are the top 2 compound superclasses which tend to produce RFIs in their CID MS^2 spectra. Moreover, aromatic, arylbromide, heteroaromatic, alkylarylether, phenol, and conjugated double bond-containing chemicals are more likely to produce RFIs. We also found four possible patterns of change in RFI percentages as a function of CID collision energy. Finally, we demonstrate that the inadequate consideration of RFIs in most conventional bioinformatic tools might cause problems during in silico fragmentation and de novo annotation of MS^2 spectra. This work provides a further understanding of CID MS^2 mechanism, and the unexpectedly large percentage of RFIs suggests a need for consideration in the development of bioinformatic software for MS^2 interpretation.

Introduction

Collision-induced dissociation (CID) is a common ion activation technique used in mass spectrometry (MS) analysis to generate tandem MS (MS^2) spectra for chemical structure determination.[1-5] The CID process generates fragment ions to obtain a fragment ion spectrum. During the CID event, heterolytic fragmentation generates (de)protonated fragment ions and homolytic fragmentation generates radical fragment ions (RFIs). The CID collision energy is a laboratory frame collision energy, and the center of mass energy slightly varies for different precursor ions depending on their masses. In low-energy CID (energy less than 100 eV) used in MS-based chemical and biochemical analyses, it is commonly believed that CID predominantly generates fragmentation of protonated or deprotonated species. In comparison, RFIs are energetically not favorable and thus are rare. Another common belief is that RFIs are generated because there is a radical cation or anion precursor as the consequence of applying a high voltage during electrospray ionization (ESI). Besides several reports on RFIs in some targeted chemical classes,[6] the global investigation on the percentage of RFIs in CID has not been systematically studied.

With the development of high-resolution liquid chromatography-mass spectrometry (LC-MS) systems, it is now possible to achieve a comprehensive and untargeted coverage of chemical species in a biological or environmental sample. The application of CID then becomes critical to generate MS^2 spectra for chemical annotation.[7-9] In particular, due to the large volume of chemical signals detected in experiments and a limited number of chemical standards in MS^2 spectral libraries, de novo interpretation and in silico prediction of MS^2 spectra from chemical structures have become important.[10, 11] In the development of above-mentioned MS^2

interpretation programs, it is important to have a clear understanding of fragmentation mechanisms in order to develop powerful and robust bioinformatic tools. Conventionally, it is thought that since ESI produces even-electron species and the fragmentation method is of relatively low energy, CID should generate even-electron species almost exclusively as well—the chance of generating RFIs is exceedingly rare. However, to the best of our knowledge, there is no comprehensive study of the types of fragment ions generated in CID MS^2 at a global scale.

In this work, we studied the existence of RFIs in CID MS^2 spectra using the NIST 20 high-resolution MS^2 spectral library (https://www.nist.gov/srd/nist-special-database-20), hereafter referred to as NIST 20. The NIST 20 contains 1,026,717 MS^2 spectra for 27,613 unique chemical compounds (positive ion mode: 765,385 spectra for 26,600 chemicals; negative ion mode: 261,332 spectra for 11,675 chemicals). One important feature of NIST 20 is that fragment ions have been annotated with molecular formulas. Using the molecular formula information, we can calculate a double bond equivalent (DBE) value for each fragment ion. Since RFIs do not follow even-electron rules, their DBE values are integers. Using this information, we can determine whether an annotated fragment ion in NIST 20 is an RFI or not. The RFI information of all the chemicals in NIST 20 was then used for a comprehensive investigation, including (1) calculating the ion-count (total number of ions) and ion-intensity (total ion intensity) percentages of RFIs and plotting their distributions; (2) categorizing chemicals by their ontology classes and checking class-specific and substructure-specific RFI distributions; (3) investigating the relationship between RFIs and CID collision energy; and (4) summarizing the potential problems of not including RFIs in in silico MS^2 generation and de novo MS^2 interpretation. This work represents a systematic and holistic

study of RFIs in CID MS^2 spectra, providing guidance for the future development of bioinformatic tools for MS^2 interpretation.

Methods

Pretreatment of NIST 20 Tandem MS Spectral Library. NIST 20 was purchased from NIST through Isomass Scientific Inc. NIST 20 contains a total of 1,026,717 low-energy CID MS^2 spectra for 27,613 unique chemical compounds. It includes 765,385 spectra for 26,600 chemicals in positive ion mode and 261,332 spectra for 11,675 chemicals for negative ion mode. These high resolution MS^2 spectra were collected from Thermo Orbitrap mass spectrometers. More than 99.5% of the MS^2 spectra were obtained using nitrogen as the collision gas, while others used helium. Molecular formula annotation of all fragment ions was completed using MS Interpreter, a bioinformatic tool embedded in the NIST MS Search program. The detailed explanation of how NIST MS Search performs subformula annotation can be found in **Text S1**.

To prepare NIST 20 for the study, we first removed MS^2 spectra of uncommon precursor ions, such as the isotopic peak(s) of a precursor (e.g., M + 1, M + 2) and doubly and triply charged adducts (e.g., $[M + Na + H]^{2+}$). We also discarded MS^2 spectra with fewer than 5 annotated fragments. Furthermore, MS^2 spectra with radical precursor ions (**Figures S1 & S2**) were removed to ensure that all RFIs were generated from (de)protonated (or even-electron) precursor ions. When multiple MS^2 spectra were available for a given chemical compound, the MS^2 spectrum with the most fragment ions was used for further interpretation. It is important to note that not all fragment ions in NIST 20 have molecular formula annotations. Overall, 88.4% and 87.0% of the fragment ions are annotated in positive and negative ion modes, respectively. For a fragment with multiple annotations, only the smallest mass error one was kept.

Analysis of NIST 20. Data analysis was conducted using R language (version 4.0.3). The R package *CHNOSZ* (version 1.4.0) was used to parse and write molecular formulas. RFIs were

determined via the annotated subformula information. More specifically, double bond equivalent (DBE) values were calculated for given subformulas using the equation shown below. Letters represent the number of each chemical element in a molecular formula.

$$DBE = C + Si + 1 - \frac{H + F + Cl + Br + I + Na + K}{2} + \frac{N + P}{2}$$

Following the LEWIS rule that electrons in main group element-based molecules are shared such that s- and p-valence shells of all atoms are fully filled, fragment ions with non-integer DBE values are (de)protonated ions and fragment ions with integer DBE values are RFIs.

To study the relationship between RFIs and compound classes, chemical compounds were first systematically classified using ClassyFire[12] (**Tables S1** for positive ion mode results and **S2** for negative ion mode results). In brief, the InChIKey, a textual identifier for chemical substances, of each chemical in NIST 20 was used as an input for the function "get_classification" from the *classyfireR* package (version 0.3.6). The "get_classification" function assigned hierarchical classification results for each chemical, and the class levels of "superclass", "class", and "subclass" defined in ClassyFire[13] were used for further analysis. Moreover, only superclasses containing more than 0.1% of the total compounds were kept.

The relationship between chemical substructures and RFI percentages were investigated using the R package *rcdk* (version 3.5.0). The R package contains a total of 307 substructures from Chemistry Development Kit (CDK).[14] The entire CDK substructure list can be found in **Table S3**. To recognize chemical substructures, the InChIKey of each chemical compound was converted to a SMILES string using the PubChem Identifier Exchange Service platform

(https://pubchem.ncbi.nlm.nih.gov/idexchange). The SMILES string of a chemical is then used to get all possible fingerprint(s) in that structure using the function "get.fingerprint" from *rcdk*.

To understand the patterns of how RFI count and intensity percentages change as a function of collision energy, an algorithm was created. We first prepared an RFI percentage vector sorted by collision energy in ascending order. Then, we split the vector into two halves. For each half, Spearman correlation is calculated between the order of collision energy and the RFI percentages (X_i). After both Cor_1 (the first half) and Cor_2 (the second half) were calculated, the RFI pattern (e.g., pattern I, II, III, or IV) was determined using the following decision table:

Pattern	$Cor_1 \geq 0$	$Cor_1 < 0$
$Cor_2 \geq 0$	**I**	**II**
$Cor_2 < 0$	**III**	**IV**

Implications of RFIs in De Novo Annotation

To demonstrate the limited capacity of annotating RFIs in state-of-the-art bioinformatics tools, we tested NIST 20 MS^2 spectra using SIRIUS 4[15], one of the most commonly used MS^2 interpretation software. We randomly sampled 1000 RFI-containing MS^2 spectra from NIST 20 (500 per ionization mode) using their integer DBE values. These MS^2 spectra were then imported into SIRIUS 4 and subjected to molecular formula prediction and fragmentation tree calculation (see **Text S2** for the detailed SIRIUS 4 parameters). For all fragment ions interpreted by SIRIUS 4, their molecular formulas were used to determine whether they were (de)protonated ions or RFIs. These SIRIUS annotation results were then compared to the NIST annotated subformulas to calculate RFI annotation sensitivity (i.e., the fraction of RFIs correctly annotated by SIRIUS 4).

Results and Discussion

Radical Fragment Ions in NIST 20

A total of 765,385 spectra for 26,600 chemicals in positive ion mode and 261,332 spectra for 11,675 chemicals for negative ion mode were collected from NIST 20. After removing disqualified MS^2 spectra, including spectra with radical precursor ions, multiple-charged adducts, and fewer than 5 annotated fragments, a total of 470,841 MS^2 spectra for 24,140 chemicals in positive ion mode and 137,308 MS^2 spectra for 9,764 chemicals in negative ion mode were used for the following studies. It was interesting to find that 11.5 and 14.3% of the MS^2 spectra in positive and negative ion modes contained radical precursor ions, respectively (**Figure S2**). In addition, over 70% of the MS^2 spectra had at least 5 annotated fragments. The distributions of annotated MS^2 spectra fragments are presented in **Figure S3**.

Figure 1 illustrates the schematic workflow of investigating RFIs in NIST 20 MS^2 spectra. We first calculated the percentages of RFIs and (de)protonated ions in each NIST 20 MS^2 spectrum (**Tables S4 & S5**) and plotted their distributions. In particular, distributions of both ion-count and ion-intensity percentages were plotted throughout this work to gain a more comprehensive view of RFIs in MS^2 spectra. **Figures 2A** and **2C** show the results of NIST 20 MS^2 spectra in positive and negative ion modes, respectively. Here we consider MS^2 spectra with $\leqslant$ 10% RFIs as low-RFI and > 10% RFIs as high-RFI MS^2 spectra. In the positive ion mode MS^2 spectra, 34.6% (162,746 out of 470,841) are low-RFI and 65.4% (308,095 out of 470,841) are high-RFI MS^2 spectra. Similar results were also found in the negative ion mode MS^2 spectra, as 31.2% (42,798 out of 137,308) are low-RFI and 68.8% (94,510 out of 137,308) are high-RFI MS^2 spectra. The

results of these ion-count percentages were unanticipated, given the common belief that RFIs are very rare in low-energy CID MS^2 spectra.

Besides the ion-count percentages, we also studied the ion-intensity percentages of RFIs in both positive and negative ion modes (**Tables S6 & S7**). As shown in **Figures 2B and 2D**, for 74.2% (349,163 out of 470,841) of positive ion mode and 71.3% (97,930 out of 137,308) of negative ion mode MS^2 spectra, RFIs only account for less than 20% of the total ion intensities. A comparison to ion-count percentages clearly shows that although an unexpectedly high number of RFIs are found in MS^2 spectra, their ion intensities are relatively low. This might be related to their low chemical stability compared to (de)protonated ions.

Radical Fragment Ions and Their Precursor Compound Classes

To further understand which chemical compounds are more likely to generate RFIs in CID MS^2 experiments, we calculated both ion-count and ion-intensity percentages of RFIs and classified the corresponding chemical compounds using ClassyFire[13] on three class levels, including "superclass", "class", and "subclass". At the superclass level, for all 22,756 compounds in positive ion mode and 8,764 compounds in negative ion mode, 17 superclasses were assigned. **Figure 3A** shows the RFI count percentage distributions of the superclasses by descending median values (superclasses containing more than 0.1% of the total compounds were plotted here, 13 superclasses for each ion mode). As we can see from **Figure 3A**, the overall median RFI count percentage is 27.3% for positive ion mode and 21.2% for negative ion mode. Compound superclasses with RFI percentage medians larger than the overall median ("All" in the plot) were labelled in red and RFI percentage medians smaller than the overall median in blue. In both positive and negative ion

modes, “Organic 1,3-dipolar compounds” and “Lignans, neolignans and related compounds” are the top 2 compound superclasses and tend to produce RFIs in their CID MS^2 spectra. This can be attributed to their abundant conjugated π-bond systems, which help to stabilize RFIs with delocalized electrons. On the other side, RFIs are rarely found in MS^2 spectra of superclasses “Lipids and lipid-like molecules” and “Organic acids and derivatives”. This result agrees with our conventional understanding that compounds with long carbon chains are generally not preferable for RFIs compared to conjugated systems. Similar trends can be obtained using the distributions of RFI intensity percentages as shown in **Figure S4**.

Next, we generated sunburst plots of RFI percentage distributions in terms of the three levels of compound classes in both polarity modes. **Figure 3B** illustrates the sunburst plot of RFI count percentage in positive ion mode. The RFI count percentages of all compound classes at different class levels can be found detailed in **Table S8**. As we can see in **Figure 3B**, slices from the inner layer to the outer layer represent compound class levels of “superclass”, “class”, and “subclass”. The median RFI count percentage in each class was calculated, and their corresponding class blocks in **Figure 3A** were distinguished by color, where dark red denotes RFI count percentage higher than median and dark blue denotes lower than median. Interestingly, various compound classes that belong to the same superclass can behave substantially different from each other. For instance, both “Fatty acyls” and “Steroids and steroid derivatives” have the superclass “Lipids and lipid-like molecules”, but the median RFI count percentage of “Fatty acyls” is only 1.2% and much smaller compared to the 17.9% of “Steroids and steroid derivatives”. The fused ring system of steroid molecules render them more inclined to RFIs during the CID process. Similarly, the “Naphthacenes” class (37.7%) has higher RFI count percentage than “Benzene and substituted

derivatives" (31.2%), even though they are of the same superclass "Benzenoids". As "Naphthacenes" are four-ringed chemicals of polycyclic aromatic hydrocarbons, it is apparent that compounds with larger conjugated electron systems have higher RFI percentages. Similarly, the RFI count percentage in negative mode results are shown in **Figure S5** and **Table S9**. Moreover, we also generated sunburst plots and result tables using the ion-intensity percentages of RFIs. Relevant results can be found in **Figures S6-S7** and **Tables S10-S11**. These informative plots provide comprehensive knowledge of RFIs in the CID MS^2 spectra of various chemical classes.

Radical Fragment Ions and Chemical Substructures

Furthermore, we investigated which chemical substructure is more likely to lead to RFI generation in CID MS^2 events. In this study, a CDK substructure system containing 307 chemical substructures (**Table S3**) was selected. In total, 23,478 unique chemicals in positive ion mode (**Table S12**) and 9,411 unique chemicals in negative ion mode (**Table S13**) were successfully assigned with at least one CDK substructure. For each chemical substructure, we categorized all the chemical compounds into two groups based on the compound containing or not containing that specific chemical substructure. We then performed Mann–Whitney U test, a nonparametric test to determine statistical significance, between the RFI percentages (both ion-count and ion-intensity) of the two classes. Statistical results of positive and negative ion modes are tabulated in **Tables S14** and **S15**, respectively. Out of the 307 total substructures, 127 substructures have *P* values of less than 0.01 based on RFI count percentage in positive ion mode. Chemicals that contain any of these 127 substructures have significantly different RFI count percentages than those that do not. Of the 127 substructures, 65 have significantly higher RFI count percentages in the substructure-containing chemicals, suggesting that chemicals containing these substructures are more likely to

generate RFIs. In **Figure 4A**, we showcase four representative substructures that have the highest statistical significance (P < 1e-3). It can be clearly seen that all of these chemical substructures have conjugated π-bond systems, which contributed to their significantly higher RFI count percentages.

We also performed a similar analysis to all the compounds in negative ion mode. Negative ion mode analysis results show 94 substructures with P values of less than 0.01. Among them, 46 substructures lead to more RFI generation when a chemical contains it. Four of the top-ranked substructures, including arylfluoride, arylchloride, arylbromide and aryliodide, are shown in **Figure 4B**. The detailed results can be found in **Table S15**. Overall, the aromatic substructure consistently leads to more RFIs in both positive and negative ion modes.

Intensity of Radical Fragment Ions and CID Collision Energy

Furthermore, we tried to understand how the change of CID collision energy affects the production of RFIs. Our conventional understanding is that higher CID collision energy is more likely to generate RFIs. In this work, we investigated the correlation between RFI intensities and CID collision energies using the chemicals in NIST 20. An important feature of NIST 20 is that it provides MS^2 spectra collected from up to 24 different collision energies. We calculated RFI intensity percentages from MS^2 spectra at each collision energy and checked the change as a function of collision energy. After manually checking dozens of chemicals, we summarized four possible patterns as shown in **Figure 5A**. Type I, in which the percentage of RFI intensities keeps increasing with the increase of collision energy, is the most common. Interestingly, there are three

other types of RFI intensity percentage change; Type II, decreases and then increases; Type III, increases and then decreases; and Type IV, keeps decreasing.

We then automatically determined the type of RFI percentage for all 24,140 and 9,764 chemicals in positive and negative ionization modes, respectively, as MS^2 spectra at multiple collision energies were available. As shown in **Figure 5B**, most chemical compounds generate RFI percentages of Type 1, which account for 61.0% in positive ion mode and 40.5% in negative ion mode. An interpretation for the chemicals belonging to Type I is that most of their RFIs are of small structural pieces at the bottom leaves of fragmentation trees[16], and thus they are inclined to be produced under higher collision energies. As an example, we manually interpreted a fragmentation pathway for the MS^2 spectrum of lithocholic acid (**Figure S9)**. All the RFIs of lithocholic acid are the end products of the fragmentation pathway. Therefore, their intensities keep increasing with the increased collision energies. Conversely, Type IV RFI intensity percentages, those that decrease with collision energy, usually happens when the RFIs show up at the root branches of fragmentation trees. Although not very common, Type IV RFIs account for 9.8% in positive ion mode and 20.0% in negative ion mode. On the other side, Type II and Type III are more complicated. It is possible that in these two cases, RFIs show up at different positions in the fragmentation pathways.

Apart from RFI intensity percentages, we also looked into the distribution and patterns of RFIs as a function of collision energy using RFI count percentage in both polarity modes (see **Figure S8**). Likewise, Type I is the most common, accounting for 57.6% and 30.6% in positive and negative ion modes, respectively. The results above show that instead of being positively correlated with

collision energy, the pattern of RFIs varies and depends on the position of the RFI in the fragmentation pathway.

Potential Issues of Not Considering Radical Fragment Ions

A clear understanding of MS^2 spectra is critical to its interpretation in chemical annotation and unknown identification.[17] Currently, RFIs in MS^2 spectra are usually ignored during the process of untargeted metabolomics data. This leads to incomplete in silico predicted fragment ions in MS^2 spectra as well as missing or incorrect annotations of true RFIs in experimental MS^2 spectra. To understand this, we first summarized some well-established bioinformatic software that perform in silico fragmentation for unknown identification (**Table 1**). It can be clearly seen in the table that the majority of the software have not fully considered the existence of RFIs. To minimize the amount of false positive fragments as well as improve the computational speed, even-electron rules are usually applied while neglecting RFIs during the in silico prediction process. Given the considerable percentage of RFIs in our NIST 20 study, we believe that the incorporation of RFIs in the development of in silico MS^2 generation can significantly boost their performance.

Next, we demonstrated the limited RFI annotation of current bioinformatics tools using SIRIUS 4[15], which is one of the commonly used MS^2 interpretation software. By randomly sampling 1000 NIST 20 MS^2 spectra containing RFIs (500 per ionization mode) and comparing the annotated RFIs against NIST annotation, the distribution plots of RFI annotation sensitivity are shown in **Figures 6A** and **6C** for positive and negative ion modes, respectively. In general, 47.4% of the positively ionized MS^2 spectra and 57.8% of the negatively ionized MS^2 spectra have lower than 10% RFI annotation sensitivity. This low annotation sensitivity suggests that most RFIs

remain poorly annotated by SIRIUS. However, considering that the intrinsic design of SIRIUS 4 allows only a few common radical losses[18], this result can be expected. To further explore the relationship between RFI annotation sensitivity and MS^2 RFI percentage, we split the sampled MS^2 spectra into 5 groups according to their RFI count percentages. MS^2 spectra with RFI count percentages over 40% were merged together to ensure that there were enough MS^2 spectra for fair comparison. As seen in **Figures 6B** and **6D**, RFI annotation sensitivity does not show general preference for RFI percentage. No statistical significance ($P > 0.1$, one-way ANOVA) was observed among the annotation sensitivities of different groups. These results further demonstrate that RFIs in MS^2 spectra remain underestimated, and most RFIs in MS^2 spectra cannot be correctly identified.

Conclusion

This work provides a comprehensive study of RFIs using large-scale, high-quality, and well-annotated MS^2 spectra data from the NIST 20 MS spectral library. Our results of ion-count and ion-intensity percentages of RFIs suggest that RFIs are common in the CID MS^2 spectra of different classes of chemicals. The high occurrence of RFIs is well beyond our previous knowledge, which indicates a need for attention during the development of bioinformatic tools for in silico fragmentation as well as de novo MS^2 spectra interpretation. More importantly, the in-depth interpretation of RFIs extends our current understanding of the CID fragmentation mechanism and fragmentation pathway. It will also guide the development of more precise bioinformatic tools for the interpretation of MS^2 spectra, facilitating unknown chemical identification in MS-based chemical analysis.

Supporting Information

The Supporting Information is available free of charge.

Figure S1. Radical precursor ions ($M^{+\bullet}$ / $M^{-\bullet}$) in MS^2 spectra. **Figure S2.** Existence of radical precursor ions in positively and negatively ionized NIST 20 MS^2 spectra. **Figure S3.** Distribution of the number of annotated fragments in NIST 20. **Figure S4.** RFI intensity percentage distributions of different superclasses. **Figure S5.** The sunburst plot of RFI count percentage (medians) in negative ion mode. **Figure S6.** The sunburst plot of RFI intensity percentage (medians) in positive ion mode. **Figure S7.** The sunburst plot of RFI intensity percentage (medians) in negative ion mode. **Figure S8.** Distributions of four patterns of change in RFI count percentage with collision energy in both positive and negative ion modes. **Figure S9.** A fragmentation pathway example including RFIs. **Text S1.** Subformula annotation of NIST 20. **Text S2.** SIRIUS 4 parameter settings. **Table S1**. ClassyFire results of unique chemicals in NIST 20 (positive ion mode). **Table S2**. ClassyFire results of unique chemicals in NIST 20 (negative ion mode). **Table S3**. 307 CDK chemical substructure bits. **Table S4**. Ion-count percentage distribution of RFIs and protonated fragment ions in NIST 20 (positive ion mode). **Table S5**. Ion-count percentage distribution of RFIs and deprotonated fragment ions in NIST 20 (negative ion mode). **Table S6**. Ion-intensity percentage distribution of RFIs and protonated fragment ions in NIST 20 (positive ion mode). **Table S7**. Ion-intensity percentage distribution of RFIs and deprotonated fragment ions in NIST 20 (negative ion mode). **Table S8.** Ion-count percentage medians of RFIs in different compound classes in NIST 20 (positive ion mode). **Table S9.** Ion-count percentage medians of RFIs in different compound classes in NIST 20 (negative ion mode). **Table S10.** Ion-intensity percentage medians of RFIs in different compound classes in NIST 20 (positive ion mode). **Table S11.** Ion- intensity percentage medians of RFIs in different compound classes in NIST 20 (negative

ion mode). **Table S12**. CDK substructures of unique chemicals in NIST 20 (positive ion mode). **Table S13**. CDK substructures of unique chemicals in NIST 20 (negative ion mode). **Table S14**. Statistical analysis results of CDK substructures (positive ion mode). **Table S15**. Statistical analysis results of CDK substructures (negative ion mode).

Acknowledgments

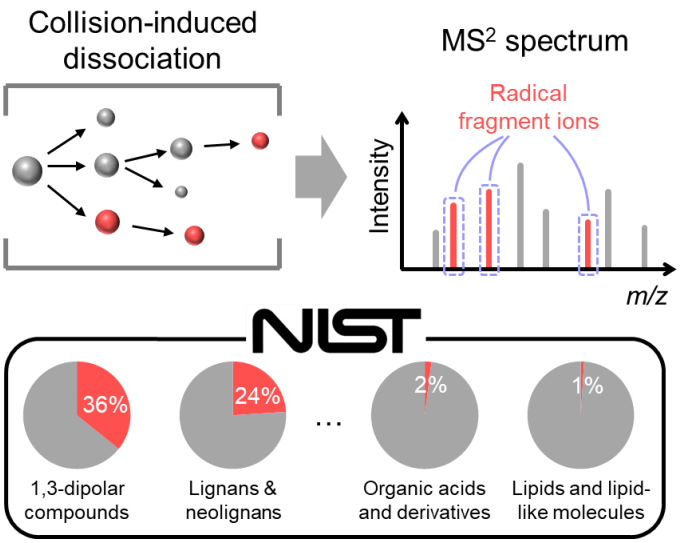
This study was funded by University of British Columbia Start-up Grant (F18-03001), Canada Foundation for Innovation (CFI 38159), New Frontiers in Research Fund/Exploration (NFRFE-2019-00789), and National Science and Engineering Research Council (NSERC) Discovery Grant (RGPIN-2020-04895). We also thank Alisa Hui for proofreading this manuscript.

References

1. Sleno, L., Volmer, D.A.: Ion activation methods for tandem mass spectrometry. Journal of mass spectrometry. **39**, 1091-1112 (2004)
2. Cooks, R.G.: Special feature: Historical. Collision‐induced dissociation: Readings and commentary. Journal of Mass spectrometry. **30**, 1215-1221 (1995)
3. Wells, J.M., McLuckey, S.A.: Collision‐induced dissociation (CID) of peptides and proteins. Methods in enzymology. **402**, 148-185 (2005)
4. Cody, R., Burnier, R., Freiser, B.: Collision-induced dissociation with Fourier transform mass spectrometry. Analytical chemistry. **54**, 96-101 (1982)
5. Rebick, C., Levine, R.: Collision induced dissociation: A statistical theory. The Journal of Chemical Physics. **58**, 3942-3952 (1973)
6. Chai, Y., Sun, H., Pan, Y., Sun, C.: N-centered odd-electron ions formation from collision-induced dissociation of electrospray ionization generated even-electron ions: single electron transfer via ion/neutral complex in the fragmentation of protonated N, N′-dibenzylpiperazines and protonated N-benzylpiperazines. Journal of the American Society for Mass Spectrometry. **22**, 1526-1533 (2011)
7. Rinschen, M.M., Ivanisevic, J., Giera, M., Siuzdak, G.: Identification of bioactive metabolites using activity metabolomics. Nature Reviews Molecular Cell Biology. **20**, 353-367 (2019)
8. Quinn, R.A., Melnik, A.V., Vrbanac, A., Fu, T., Patras, K.A., Christy, M.P., Bodai, Z., Belda-Ferre, P., Tripathi, A., Chung, L.K.: Global chemical effects of the microbiome include new bile-acid conjugations. Nature. **579**, 123-129 (2020)
9. Xing, S., Jiao, Y., Salehzadeh, M., Soma, K.K., Huan, T.: SteroidXtract: Deep Learning-Based Pattern Recognition Enables Comprehensive and Rapid Extraction of Steroid-Like Metabolic Features for Automated Biology-Driven Metabolomics. Analytical Chemistry. (2021)
10. Böcker, S., Letzel, M.C., Lipták, Z., Pervukhin, A.: SIRIUS: decomposing isotope patterns for metabolite identification. Bioinformatics. **25**, 218-224 (2009)
11. Wolf, S., Schmidt, S., Müller-Hannemann, M., Neumann, S.: In silico fragmentation for computer assisted identification of metabolite mass spectra. BMC bioinformatics. **11**, 1-12 (2010)
12. Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E., Greiner, R., Wishart, D.S.: ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. Journal of Cheminformatics. **8**, 61 (2016)
13. Feunang, Y.D., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., Fahy, E., Steinbeck, C., Subramanian, S., Bolton, E.: ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. Journal of cheminformatics. **8**, 1-20 (2016)
14. Willighagen, E.L., Mayfield, J.W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O.: The Chemistry Development Kit (CDK) v2. 0: atom typing, depiction, molecular formulas, and substructure searching. Journal of cheminformatics. **9**, 1-19 (2017)
15. Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., Dorrestein, P.C., Rousu, J., Böcker, S.: SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nature Methods. **16**, 299-302 (2019)
16. Rasche, F., Svatoš, A., Maddula, R.K., Böttcher, C., Böcker, S.: Computing Fragmentation Trees from Tandem Mass Spectrometry Data. Analytical Chemistry. **83**, 1243-1251 (2011)

17. Guo, J., Shen, S., Xing, S., Chen, Y., Chen, F., Porter, E.M., Yu, H., Huan, T.: EVA: Evaluation of Metabolic Feature Fidelity Using a Deep Learning Model Trained With Over 25000 Extracted Ion Chromatograms. Analytical Chemistry. (2021)
18. Rasche, F., Scheubert, K., Hufsky, F., Zichner, T., Kai, M., Svatoš, A., Böcker, S.: Identifying the Unknowns by Aligning Fragmentation Trees. Analytical Chemistry. **84**, 3417-3426 (2012)
19. Djoumbou-Feunang, Y., Pon, A., Karu, N., Zheng, J., Li, C., Arndt, D., Gautam, M., Allen, F., Wishart, D.S.: CFM-ID 3.0: Significantly improved ESI-MS/MS prediction and compound identification. Metabolites. **9**, 72 (2019)
20. Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S.: Searching molecular structure databases with tandem mass spectra using CSI: FingerID. Proceedings of the National Academy of Sciences. **112**, 12580-12585 (2015)
21. Ridder, L., van der Hooft, J.J., Verhoeven, S., de Vos, R.C., Bino, R.J., Vervoort, J.: Automatic chemical structure annotation of an LC–MS n based metabolic profile from green tea. Analytical Chemistry. **85**, 6033-6040 (2013)
22. Ridder, L., van der Hooft, J.J., Verhoeven, S., de Vos, R.C., van Schaik, R., Vervoort, J.: Substructure - based annotation of high - resolution multistage MSn spectral trees. Rapid Communications in Mass Spectrometry. **26**, 2461-2471 (2012)
23. Wang, Y., Kora, G., Bowen, B.P., Pan, C.: MIDAS: a database-searching algorithm for metabolite identification in metabolomics. Analytical chemistry. **86**, 9496-9503 (2014)
24. Tsugawa, H., Kind, T., Nakabayashi, R., Yukihira, D., Tanaka, W., Cajka, T., Saito, K., Fiehn, O., Arita, M.: Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. Analytical chemistry. **88**, 7946-7958 (2016)
25. Huan, T., Tang, C., Li, R., Shi, Y., Lin, G., Li, L.: MyCompoundID MS/MS search: metabolite identification using a library of predicted fragment-ion-spectra of 383,830 possible human metabolites. Analytical chemistry. **87**, 10619-10626 (2015)

TOC graphical abstract



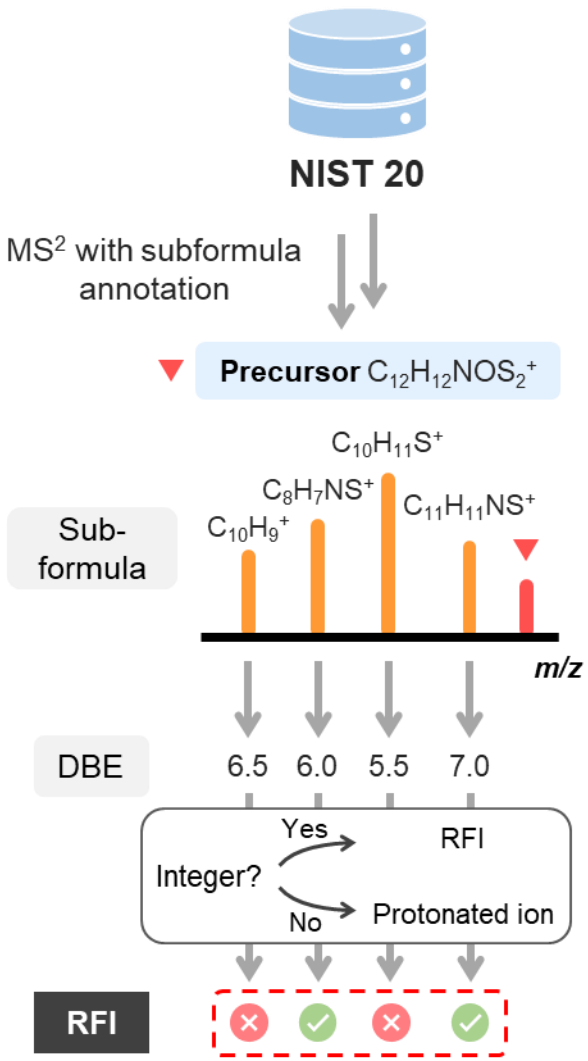


Figure 1. Schematic workflow of mining NIST 20 to automatically explore RFIs.

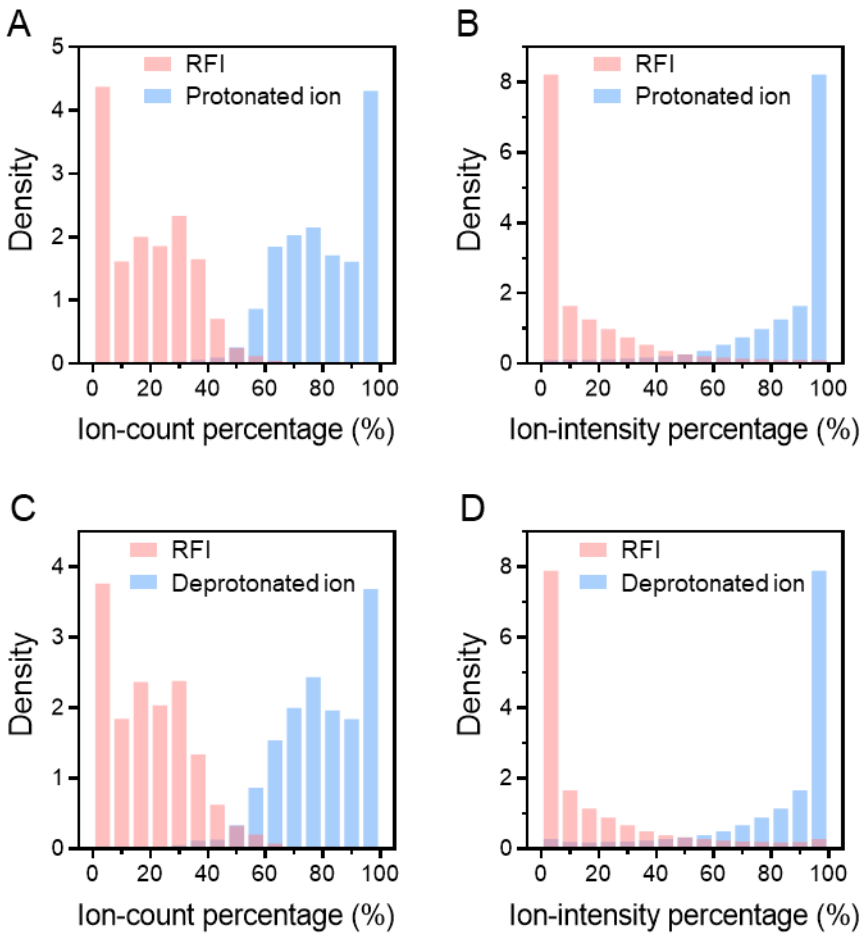
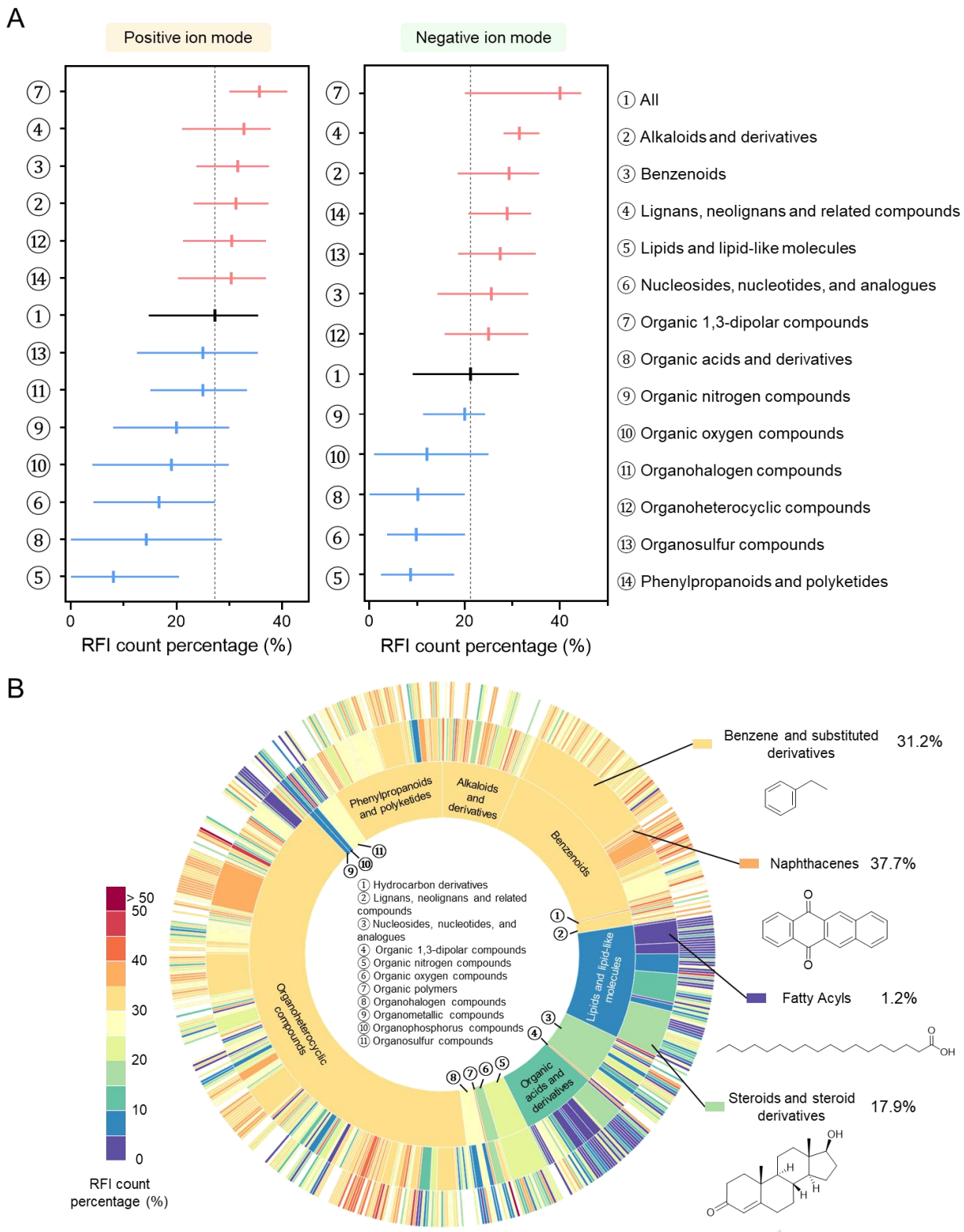


Figure 2. Distribution plots of RFIs & (de)protonated ions in NIST 20 library. (A) & (C) Ion-count distribution of RFIs and (de)protonated fragment ions. (B) & (D) Ion-intensity distribution of RFIs and (de)protonated fragment ions.


A
Positive ion mode
Negative ion mode
RFI count percentage (%)
RFI count percentage (%)
① All
② Alkaloids and derivatives
③ Benzenoids
④ Lignans, neolignans and related compounds
⑤ Lipids and lipid-like molecules
⑥ Nucleosides, nucleotides, and analogues
⑦ Organic 1,3-dipolar compounds
⑧ Organic acids and derivatives
⑨ Organic nitrogen compounds
⑩ Organic oxygen compounds
⑪ Organohalogen compounds
⑫ Organoheterocyclic compounds
⑬ Organosulfur compounds
⑭ Phenylpropanoids and polyketides
B
Phenylpropanoids and polyketides
Alkaloids and derivatives
Benzenoids
Lipids and lipid-like molecules
Organic acids and derivatives
Organoheterocyclic compounds
① Hydrocarbon derivatives
② Lignans, neolignans and related compounds
③ Nucleosides, nucleotides, and analogues
④ Organic 1,3-dipolar compounds
⑤ Organic nitrogen compounds
⑥ Organic oxygen compounds
⑦ Organic polymers
⑧ Organohalogen compounds
⑨ Organometallic compounds
⑩ Organophosphorus compounds
⑪ Organosulfur compounds
Benzene and substituted derivatives 31.2%
Naphthacenes 37.7%
Fatty Acyls 1.2%
Steroids and steroid derivatives 17.9%
RFI count percentage (%)

456

Figure 3. (A) RFI count percentages at the level of "Superclass" (median with interquartile range). The box plots were drew using median with interquartile. Compound superclasses containing more than 0.1% of the total compounds (13 superclasses in each ion mode) are shown. (B) The sunburst plot of RFI count percentage (medians) in positive ion mode. Slices from the inner layer to the outer layer represent class levels of "superclass", "class" and "subclass".

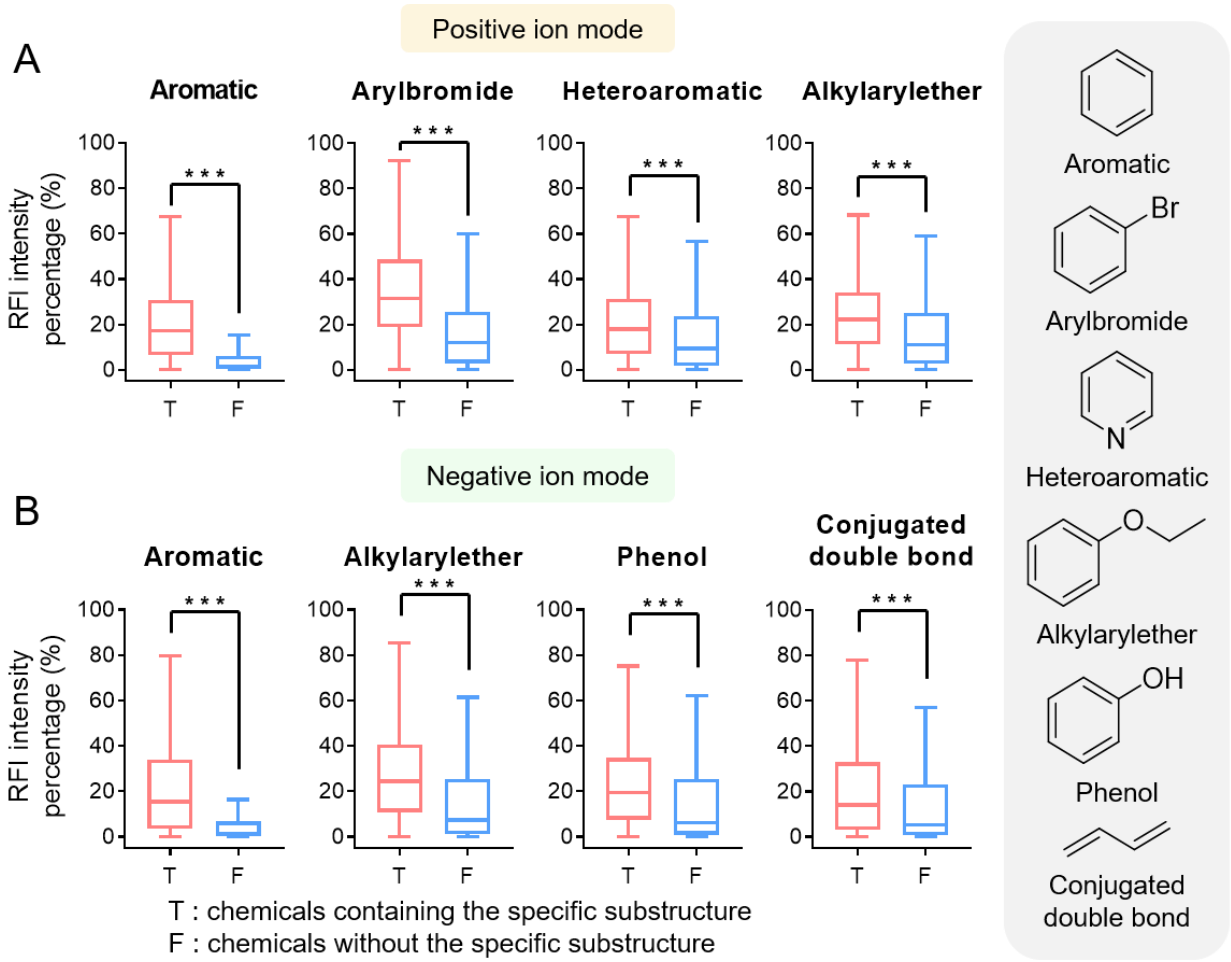


Figure 4. Representative chemical substructures that tend to produce RFIs when a chemical contains it in (A) positive ion mode and (B) negative ion mode. (*** on top of the box plot means $p < 0.001$, error bars indicate 95% confidence interval).

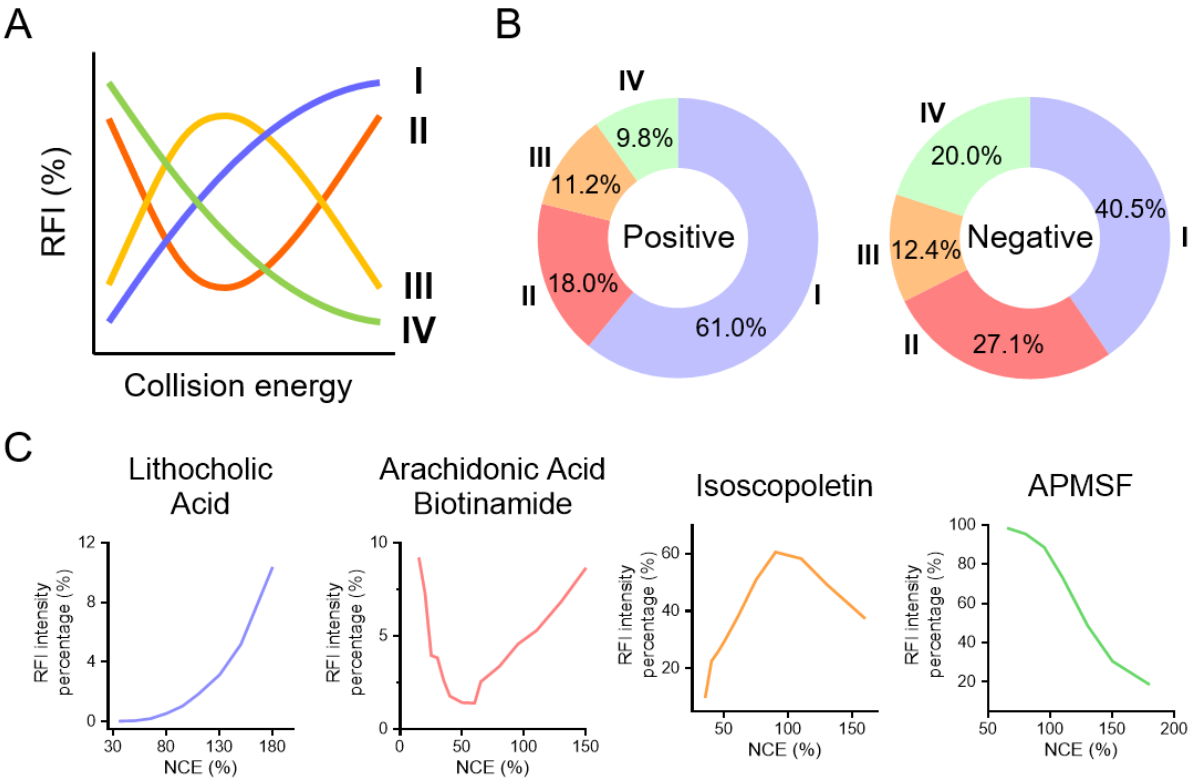


Figure 5. (A) Four patterns of how RFI percentage changes with collision energy. (B) Distributions of the four patterns in both positive and negative ion modes. (C) Representative examples of the four patterns. (NCE: normalized collision energy). APMSF: (4-Carbamimidoylphenyl) methanesulfonyl fluoride.

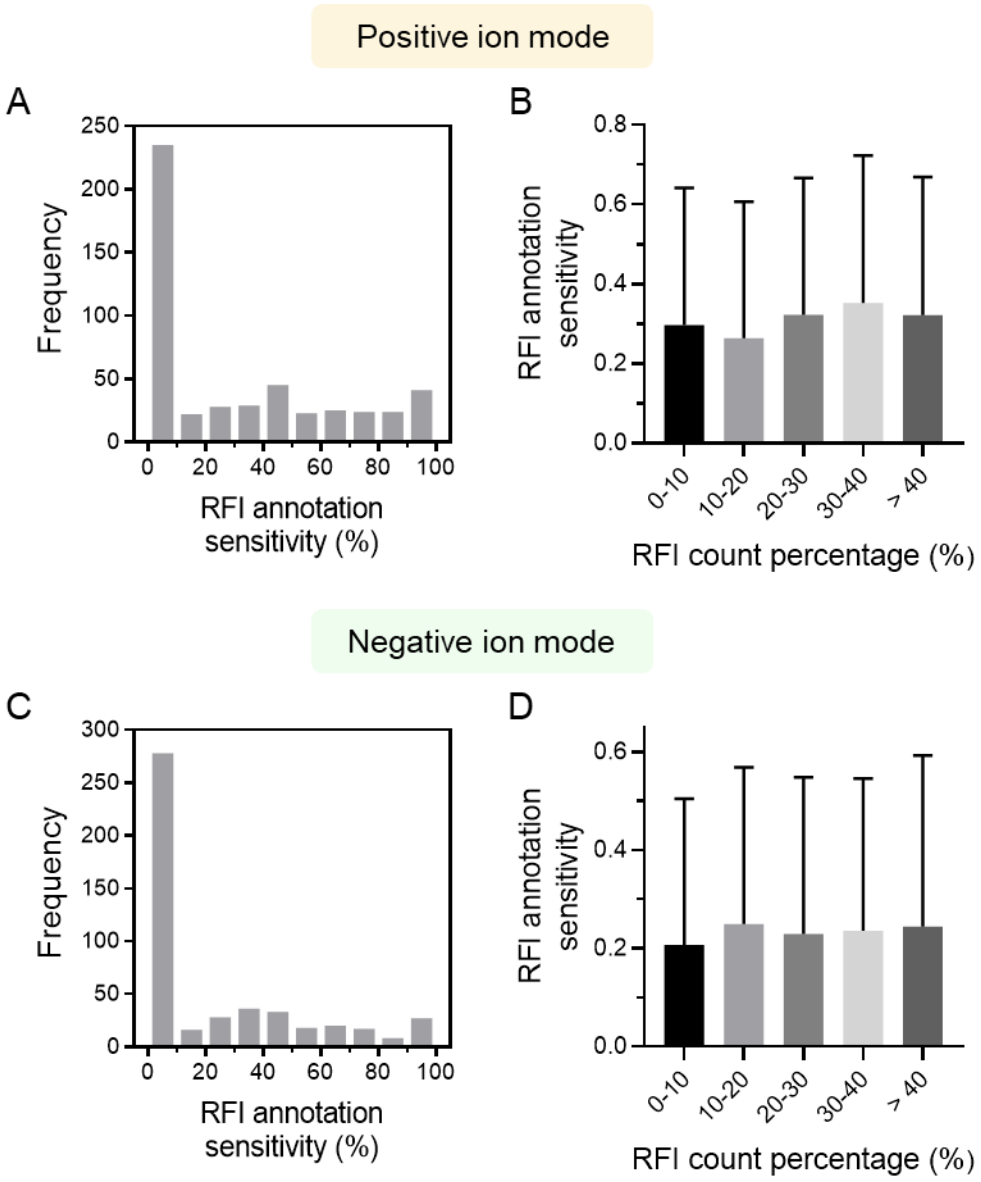


Figure 6. RFI annotation results using SIRIUS 4. (A) Distributions of RFI annotation sensitivity in positive ion mode. (B) RFI annotation sensitivity in terms of different RFI count percentages in positive mode (mean with SD shown). (C) Distributions of RFI annotation sensitivity in negative ion mode. (D) RFI annotation sensitivity in terms of different RFI count percentages in negative ion mode (mean with SD shown).

Table 1. Summary of representative in silico fragmentation tools (in alphabetical order) and their RFI implementations.

Tool name	Core algorithm for in silico fragmentation	RFI implementation
CFM-ID 3.0[19]	Models fragmentation as a stochastic, homogenous, Markov process involving state transitions between charged fragments.	**No.** Even-electron rule is applied, and no RFI is considered.
CSI:FingerID[20]	Computes fragmentation trees of MS^2 spectra. Predicts their fingerprints and compares them against the fingerprints of candidate compounds in the structure database.	**Partially.** A few common radical losses are taken into consideration.
MAGMa[21, 22]	Assigns pre-generated substructures to the fragment ions of high-resolution multistage MS^n data, and ranks the candidate molecules.	**No.** The maximum number of protons by which the mass is allowed to differ is set to the number of broken bonds plus one.
MetFrag[11]	A hybrid rule-based combinatorial approach. Simulates the fragmentation via breaking molecular bonds.	**Partially.** The matching function adds or removes a hydrogen to the fragment mass. A penalty is given in this case.
MIDAS[23]	A three-level fragmentation tree is constructed for each chemical structure. Three charged forms are considered.	**Yes.** Fragments in forms of $[F]^+$, $[F + H]^+$, and $[F + 2H]^+$ are considered.
MS-FINDER[24]	Hydrogen rearrangement during bond cleavage & even-electron rule for carbon and heteroatoms.	**Partially.** Up to two hydrogens can be added or removed to recognize RFIs, and RFIs are considered as irregular behaviors (semiresolved).
MycompoundID[25]	Heteroatom-initiated bond chopping & splitable-bond chopping	**No.** Only $[M + H]^+$ and $[M - H]^-$ are considered.