cells

MDPI

Article

Computational Workflow to Study the Diversity of Secondary Metabolites in Fourteen Different *Isatis* Species

Doudou Huang [1,†], Chen Zhang [1,†], Junfeng Chen [1], Ying Xiao [1], Mingming Li [2], Lianna Sun [1], Shi Qiu [1,*] and Wansheng Chen [1,2,*]

[1] Research and Development Center of Chinese Medicine Resources and Biotechnology, Institute of Chinese Materia Medica, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China; hdd890920@163.com (D.H.); chen870826@126.com (C.Z.); cjf12347831@foxmail.com (J.C.); xiaoyingtcm@shutcm.edu.cn (Y.X.); sssnmr@163.com (L.S.)

[2] Department of Pharmacy, Changzheng Hospital, Second Military Medical University, Shanghai 200433, China; limingming.email@foxmail.com

* Correspondence: davidhugh@msn.cn (S.Q.); chenwansheng@shutcm.edu.cn (W.C.)

† These authors contributed equally to this work.

Abstract: The screening of real features among thousands of ions remains a great challenge in the study of metabolomics. In this research, a workflow designed based on the MetaboFR tool and "feature-rating" rule was developed to screen the real features in large-scale data analyses. Seventy-four reference standards were used to test the feasibility, with 83.21% of real features being obtained after MetaboFR processing. Moreover, the full workflow was applied for systematic characterization of 14 species of the genus *Isatis*, with the result that 87.72% of real features were retained and 69.19% of the in-source fragments were removed. To gain insights into metabolite diversity within this plant family, 1697 real features were tentatively identified, including lipids, phenylpropanoids, organic acids, indole derivatives, etc. Indole derivatives were demonstrated to be the best chemical markers with which to differentiate different species. The rare existence of indole derivatives in *Isatis cappadocica* (*cap*) and *Isatis cappadocica* subsp. Steveniana (*cap*S) indicates that the biosynthesis of indole derivatives could play a key role in driving the chemical diversity and evolution of genus *Isatis*. Our workflow provides the foundations for the exploration of real features in metabolomics, and has the potential to reveal the chemical composition and marker metabolites of secondary metabolites in plant fields.

Keywords: untargeted metabolomics; high-resolution mass spectrometry; real features screening; chemical characterization; genus *Isatis*

Citation: Huang, D.; Zhang, C.; Chen, J.; Xiao, Y.; Li, M.; Sun, L.; Qiu, S.; Chen, W. Computational Workflow to Study the Diversity of Secondary Metabolites in Fourteen Different *Isatis* Species. *Cells* **2022**, *11*, 907. https://doi.org/10.3390/cells11050907

Academic Editors: Franz Hadacek and Petr Karlovsky

Received: 16 January 2022
Accepted: 28 February 2022
Published: 6 March 2022

1. Introduction

Metabolomics has been widely applied by a broad spectrum of researchers interested in defining the biological roles of biomarkers, novel metabolites, and new drug candidates, as well as for disease diagnosis [1–3]. Untargeted and targeted metabolomics are the two most common methodologies for comprehensive and targeted analyses to provide global or simple metabolic overviews [4,5]. For sample sets of complex mixtures for which little information is available, untargeted metabolomics can provide an unbiased discriminatory analysis for global metabolite detection, and may also yield insights into biochemical functions [6]. Advances in liquid chromatography coupled with mass spectrometry (LC-MS) have become integral metabolomics platforms, with each resulting signal commonly being referred to as a "feature". On average, the representative untargeted approach is usually performed with a tremendous amount of MS features across tens to hundreds of samples, which makes the exploration of metabolite features even harder [7]. Until recently, numerous open-source tools have been developed to define features for the generation of peak tables, with the most common ones being XCMS, MZmine2, and MS-DIAL [8–10].

Although an enormous and diverse collection of MS data is critical for metabolomics studies, the interpretation of results remains a challenge owing to the ever-increasing complexity of feature attributions, which complicates the determination of real features in peak tables.

In untargeted metabolomics, the peak table obtained with a feature-detection tool contains all the data acquired from a metabolomics study [11], implying that the quality of the peak table may influence data interpretation. Variations in experimental conditions and multiple sample sets may lead to erroneous assignments of feature attributions in MS data in a peak table, including parental features, adducts, fragments, isotopes, dimers and contaminants [12–14]. Real features cover, e.g., parental features, while adducts represent features according to real metabolites and are derived from real sample sources. Parental features, including quasi-molecular ions, are presented as a proton adduct of $[M + H]^+$ in positive-ion mode or $[M - H]^-$ in the negative-ion mode [13]. Meanwhile, adducts are formed by the addition of a molecular ion to a metabolite within the ion source. Notably, each adduct may coexist and correlate with the parental feature or other adducts. Such correlations inform the discovery of parental features accurately and rapidly, in particular for metabolites that have no parental features in MS. In contrast, the existence of enormous amounts of in-source fragments, which are produced during a series of dissociation events as a consequence of weak bonds that are broken in metabolites in the ion source, complicates feature annotation owing to the inherent diversity of metabolites [12]. Among the technical advances of the last decade, certain novel tools have attempted to address this challenge by focusing on improving the grouping of fragments with parental features; these tools include CAMERA, RAMClust and MS-CleanR [15–17]. MS-CleanR, as the most recently developed tool, utilizes full-featured generic filters and feature-clustering functions, and produces a user-friendly peak table [17]. Unfortunately, any inaccuracies in the selection of parental features will yield a peak table of poor quality. Given these potential issues, it is vital that a rule be implemented for accurate screening and selecting parental features among the inherent abundance of potential features with the purpose of "one feature to one metabolite".

Chemical diversity is generated through the evolution of biosynthesis pathways in plants [18]. Although mass spectrometry can yield more structural information, the recognition of real features among thousands of features derived from diverse metabolites is still a significant obstacle in plant metabolomics. Here, we have developed a workflow which contains a self-developed tool named MetaboFR, aiming to obtain high quality peak tables in metabolomics. We demonstrate the efficiency of our workflow on the 14 different species from genus *Isatis*. Genus *Isatis* is a widely distributed plant, comprising 80 species from the Middle East to the Mediterranean region [19]. In China, *Isatis indigotica* Fort is important in traditional Chinese medicine (TCM), and is used for the treatment of fever, flu, and inflammation [20]. Its derivatives played a key role as an antiviral medicine in the SARS outbreak in 2002. Previous studies revealed the main compounds in *Isatis indigotica*, including indole alkaloids, organic acids, flavonoids, lignans, nucleosides, etc. Indigo, as a principal indole alkaloid in *I. indigotica*, has long been used as a natural dye around the world [21]. In addition, *Isatis tinctoria*, better known as woad, is widely used as a source of indigo and has been used as a medicinal plant in Europe for centuries. Modern studies show that it has anti-inflammatory, antitumor, antimicrobial and antioxidant activities [19]. However, only *Isatis tinctoria* and *Isatis indigotica* have attracted much attention, and little is known about the chemistry of other species in this genus. By employing our workflow, we are able to provide comprehensive chemical annotations based on real features screening (Table S1), which will give us metabolic insights into the chemical diversity and plant evolution of secondary metabolites in *Isatis* L. genus.

2. Materials and Methods

2.1. Materials and Chemicals

In June 2018, the following *Isatis* species were planted on our behalf by the private company (Zealquest, Shanghai, China): *Isatis indigotica* Fort. (*ind*), *Isatis buschiana* (*bus*), *Isatis cappadocica* (*cap*), *Isatis cappadocica* subsp. Steveniana (*cap*S), *Isatis costata* C. A. Mey. (*cos*C), *Isatis tinctoria* L.(*tin*), *Isatis tinctoria* L. var. *tinctoria* (*tin*V), *Isatis japonica* Miq. (*jap*), *Isatis lusitanica*L. (*lus*), *Isatis glauca* Aucherex Boiss. (*gla*A), Tetraploid *Isatis indigotica* Fort. (4n = 28), (4*ind*), *Isatis oblongata* DC. (*obl*), *Isatis violascens* Bunge. (*vio*), and *Isatis minima* Bunge. (*min*). Information regarding the source of the seeds (14 species) is shown in Table S2. Leaves of these 14 species were harvested four months after planting. All voucher specimens were deposited in the Institute of Chinese Materia Medica, Shanghai University of Traditional Chinese Medicine (SHUTCM, Shanghai, China). The phenotypes of 14 species are presented in Figure S1.

Pure distilled water used for LC-MS analysis was purchased from Watsons Water (Watsons, Hong Kong, China). HPLC-grade methanol, acetonitrile, and formic acid were purchased from Fisher Scientific (Thermo, Waltham, MA, USA). Warfarin and analytical-grade ammonium acetate were obtained from Sigma-Aldrich (Sigma-Aldrich, Darmstadt, Germany).

2.2. Standard Solution Preparation

A test mixture containing 74 reference standards was selected for validation of our workflow, and was prepared at three levels of concentration, i.e., 10, 50, and 100 ug mL^{-1} in MeOH. Seventy-four reference standards were purchased from Sigma-Aldrich (Darmstadt, Germany, purity $\geq$ 98%); detailed information is shown in Table S3.

The following 12 reference standards for the characterization of genus *Isatis* were purchased from Dalian Meilun Biotechnology (Meilun, Dalian, China) and prepared at 100 ug mL^{-1} in MeOH: isovitexin, vicenin-2, guanosine, indigo, indicant, indirubin, indoxyl, isatin, pheophorbide a, pinoresinol, pinoresinol-4-O-D-glucose, and matairesinol-4-O-D-glucose (purity $\geq$ 98%).

2.3. Sample Preparation

Leaves of *Isatis* species were freeze-dried and passed through a 40-mesh sieve. The extract method from our previous study [22] was applied with some modifications. Briefly, the ground leaves (20.0 mg) were extracted in 8.0 mL methanol containing warfarin (200 ng mL^{-1}) as an internal reference. The mixture was vortexed for 30 s and sonicated for 30 min (40 kHz, 250 W). The extract was then centrifuged at 4 °C for 10 min at 20,000× *g*. A 3-µL aliquot was subjected to analysis with UHPLC-QTOF-MS/MS (QTOF, quadrupole-time-of-flight). Finally, a blank sample (methanol) and a quality-control (QC) sample (aliquot of all samples) were also prepared for the LC-MS analysis.

2.4. LC-MS Analysis

For metabolic data acquisition for 14 species in genus *Isatis*, 1290 UHPLC combined with 6530 QTOF-MS (Agilent, Santa Clara, CA, USA) system was used. The samples were separated through ACQUITY BEH C18 (2.1 × 100 mm, 1.7 µm) (Waters Technologies, Milford, MA, USA) with a column temperature of 35 °C. In this research, gradient elution was applied using 2 mM ammonium acetate in water (A) and acetonitrile (B) as the elution phase. The elution procedure was conducted as follows: 95–90% A from 0 to 2 min; 90–48% A from 2 min to 10 min; 48–25% A from 10 min to 15 min; 25–5% A from 15 min to 25 min; and 5% A sustained from 25 min to 35 min. Another 4 min 95% A was used for re-equilibration, and the flow rate was set as 0.3 mL min^{-1}. In addition, QC samples were employed every 10 injections to monitor the system stability. The mass spectrometer parameters were set according to our previous study with positive-ion as the acquisition mode [23].

2.5. Data Processing

The data processing workflow, shown in Figure 1, was divided into three steps. In step 1, the AnalysisBaseFileConverter (https://www.reifycs.com/AbfConverter/, accessed on 16 January 2022) software was used for raw mass spectrum data extraction, and MS-DIAL version 4.24 (https://prime.psc.riken.jp/compms/msdial/main.html, accessed on 16 January 2022) was used for MS peak detection and alignment with the following parameters [10]: retention time range, 1–30 min (1–18 min for reference standard samples); retention time tolerance, 0.15 min; MS^1 mass range, 100–1200 (100–1000 min for reference standard samples); MS^1 tolerance, 0.01 Da; MS^2 mass range, 0–1200 (0–1000 min for reference standard samples); MS^2 tolerance, 0.02 Da; minimum peak height, 5000 amplitude (15,000 for reference standard samples). Other parameters, including mass slice width and MS/MS abundance cut off, were set according to our previous study [23]. An internal standard was used for peak heights normalization.

In step 2, MS-CleanR (https://github.com/eMetaboHUB/MS-CleanR, accessed on 16 January 2022) was used to process the peak tables extracted from MS-DIAL [17]. In the MS-CleanR processing, several parameters, including minimum blank ratio and maximum retention time tolerance, were set according to our previous study [24], in which the minimum Pearson correlation coefficient was 0.8, and $\alpha = 0.05$ indicated statistical difference. The peak table was generated as "MS_peaks-clusters_final" in MS-CleanR and was used for further MetaboFR processing.

In step 3, the peak table "MS_peaks-clusters_final" was processed by MetaboFR based on the "feature-rating" (FR) rule. A tutorial of MetaboFR is included in the Supplementary Information. The mass tolerance of adduct flagging and fragment removal was set to 0.01 and 0.05 Da, respectively. For adduct flagging, six adduct groups were defined and imported for characterization of genus *Isatis*: $[M + H]^+$ to $[M + NH^4]^+$, $[M + H]^+$ to $[M + Na]^+$, $[M + H]^+$ to $[M + K]^+$, $[M + NH^4]^+$ to $[M + Na]^+$, $[M + NH^4]^+$ to $[M + K]^+$ and $[M + Na]^+$ to $[M + K]^+$.

2.6. Statistical Analysis

The normalized peak tables processed by different processing approaches were imported into SIMCA-P 14.1 (Umetrics AB, Umea, Sweden). All data were scaled by unit variance scaling, and all variables were pareto-scaled before autofitting. The principal component analysis (PCA) model was applied to determine unsupervised pattern recognition by importing different peak table compositions.

2.7. Metabolite Identification

An in-house library, containing 269 compounds reported from the entire genus *Isatis*, was established in order to obtain precise chemical interpretation results. Information about the 269 compounds, including chemical names and formulas, is listed in Table S4, and the structures are displayed in Figures S2–S6. To identify unknown metabolites, we also drew upon major natural product databases, including KNApSAcK, PlantCyc, LipidMaps, NANPDB and UNPD. Conveniently, an in silico software, MS-FINDER (https://prime.psc.riken.jp/compms/msfinder/main.html, accessed on 16 January 2022), integrates these databases and provides each compound with Simplified Molecular Input Line Entry Specification (SMILES) for chemical classification with the Classyfire function (https://classyfire.wishartlab.com, accessed on 16 January 2022) [25,26].

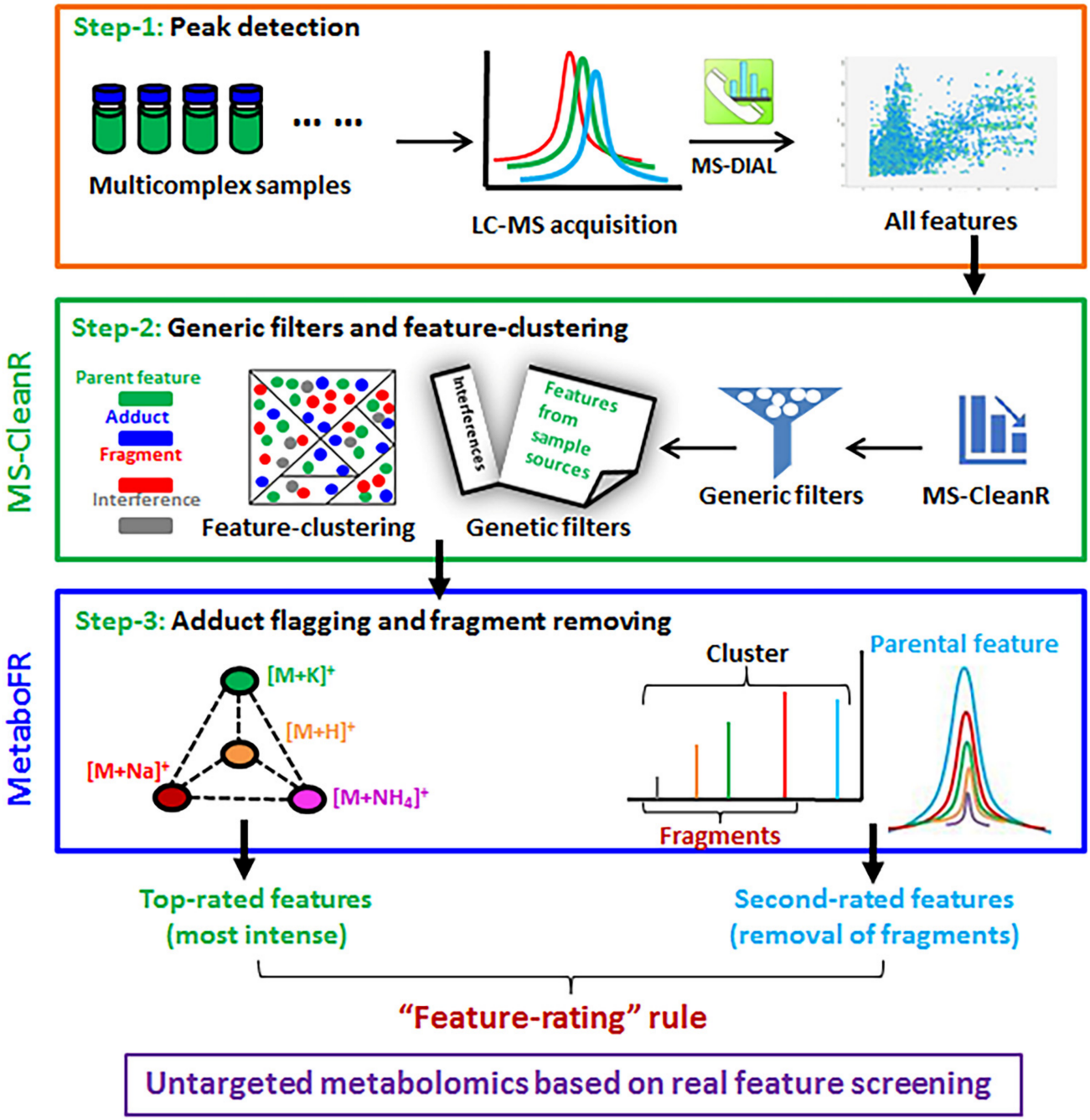

Figure 1. Data analysis workflow for comprehensive characterization of complex mixtures of metabolites based on the location of real features. Step 1: MS-DIAL was applied for the peak detection. Step 2: MS-CleanR was conducted utilizing the generic filters and feature clustering. Step 3: The developed MetaboFR tool combined with the "feature-rating" rule was applied for adduct flagging and in-source fragment removal.

3. Results and Discussion

3.1. The Workflow of "Feature-Rating" Rule and MetaboFR

The purpose of our workflow is to remove both the false features from nonsample sources and interfering metabolites in order to improve the quality of peak tables and identification results [27]. First, as shown in Figure 1, MS-DIAL was used as a tool for peak extraction, alignment and normalization to generate a peak table in an untargeted metabolomics analysis in Step 1 [10]. In Step 2, generic filters and feature-clustering

from the MS-CleanR tool were used to remove the interference signals and implement feature-clustering based on an MS-DIAL peak character estimation algorithm and Pearson's correlation [28]. Any aberrant features resulting from signals in the "blank" sample, as well as any other unusual mass defects, were removed by the generic filters, which were also applied to establish a relative standard deviation threshold among the sample classes for the purpose of eliminating metabolites in the same sample class that were unstable [17]. Step 3 involves a FR rule based on the screening of real features. The real features are defined as "top-rated features" (TRFs) and "second-rated features" (SRFs), based on differences in their MS attributions. TRFs are parental features that either correlated with corresponding adduct ions or do not have a corresponding quasi-molecular ion, but rather, only corresponding adduct ions. In contrast, SRFs are defined as parental features that only have isolated quasi-molecular ions. For TRF screening, orienting the correlations between adducts and quasi-molecular ions can increase the efficiency of parental-feature identification. Notably, some molecules have no adduct form, e.g., $[M + H]^+$ or $[M - H]^-$ in MS; in such cases, it is imperative to take advantage of adduct correlations to identify parental features. However, for SRF screening, in-source fragments pose the greatest challenge when distinguishing with SRFs because both have isolated quasi-molecular ions in MS, which confounds the ability to distinguish between the two. In a summary, locating adduct correlations and reducing the number of fragments are two core concepts in the FR rule when exploring real features among thousands of MS signals.

Here, an R package called MetaboFR was developed to simultaneously capture adduct correlations and reduce in-source fragments by embedding the results after MS-CleanR in step 3 in our workflow. Desirable adduct correlations can be imported into the R package. The flagging of adduct types is based on the mass difference among each feature cluster, and the mass tolerance (typically 0.01 Da) is also tunable by user (as detailed in the tutorial on MetaboFR). All the features flagged with adduct types are regarded as TRFs and are labeled in the last blank column in the peak table. In addition, the reduction of in-source fragments to explore SRFs is another function of MetaboFR. Among the peak table generated from MS-CleanR, if the feature shown in MS^1 (Average.Mz after MS-CleanR processing) also exists in MS^2 (MS.MS. spectrum after MS-CleanR processing), based on other features in each feature cluster, it will be regarded as an in-source fragment and will be removed from the peak table. The mass tolerance for fragment screening is tunable for users (typically 0.05 Da). After removal of the fragments, the retained features (except retained TRFs) are regarded as SRFs. Finally, only the most intense TRFs (TRFs most) in each adduct correlation are manually selected to the final peak table, which are applied, together with SRFs, for metabolomic analyses.

3.2. Validation of Workflow by Applying Reference Standards

To validate our process, we applied a mixture of 74 reference standards to benchmark our workflow in negative-ion mode (Figure 2). Detailed information on the reference standards is shown in Table S3. Three levels of concentrations along with QC and blank samples were acquired and imported into MS-DIAL for peak detection. The obtained peak table, containing 248 featuress was processed by MS-CleanR to generate the clustered peak table, with the result of 96 feature clusters covering 203 features (see Figure 2C and Table S4). MetaboFR was applied to directly process Table S4; 131 features were retained after adduct flagging and fragment removal (Figure 2C and Table S5).

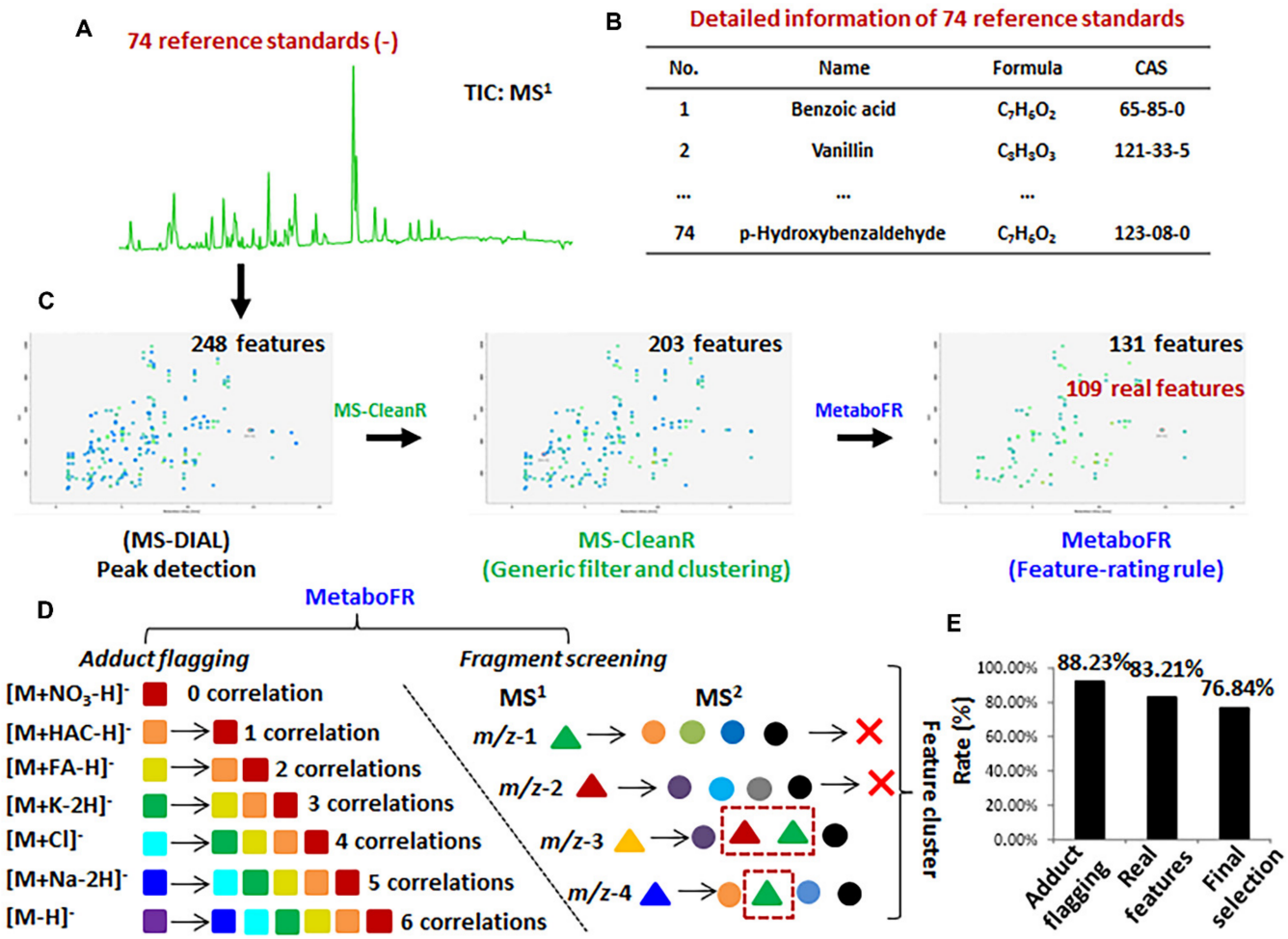

Figure 2. Seventy-four reference standards were used to validate our workflow. (**A**) A total ion current (TIC) chromatogram of 74 reference standards in negative−ion mode. (**B**) Detailed information about the 74 reference standards (Table S3). (**C**) The evolution of feature amounts corresponding to three individual steps in our workflow. (**D**) The selection of adduct correlations and the rule of in−source fragment recognition. (**E**) The effect of our workflow on the final feature table of 74 reference standards.

To locate the TRFs by adduct flagging, 21 adduct correlations derived from seven adduct types were imported into MetaboFR with a mass tolerance setting of 0.01 Da (see Figure 2D). Unlike conventional adduct flagging, flagging based on the adduct correlations is able to label the adduct types even for features without quasi-molecular ions in MS. In the fragment screening step, 72 features were removed based on the fragment screening rule shown in Figure 2D, and 40 dehydrogenation features without adducts were explored as SRFs. It should be noted that the flagged adduct features were defined as TRFs and were not be screened as fragment candidates. As a result of MetaboFR tool processing, 68 features were flagged as adduct types; the accuracy of flagging was 88.23%, whereas only five adducts were omitted due to unreasonable feature clustering observed by manual inspection (Figure 2E and Table S6). Among the total of 131 features, the existence of 49.62% adduct features revealed that adducts exhibit great values for the screening of parental features, especially for the MS behavior of five features with no visible quasi-molecular ions (Table S6). In order to explain the tool performance in more detail, Figure 3A shows a series of MS features from a retention time 7.41 min after processing by MS-CleanR [17]. Three compounds, i.e., epmedin C, etoposide and prednisolone, were identified among twelve features, comprising ten adducts and two fragments, by manually annotating.

Notably, all three compounds exhibited multiple adduct types along with an absence of quasi-molecular ions, indicated that flagging the adduct types through adduct correlations is essential and effective. After the application of MetaboFR, nine adduct types were labeled and were regarded as TRFs, while two fragments derived from epmedin C and prednisolone were removed due to the existence in the MS^2 spectrum of corresponding parental features. Finally, the most intense of TRFs were selected to represent the parental features of three compounds for further analysis. As shown in Figure 3B, one compound named p-hydroxybenzaldehyde (m/z 121.0295) was not detectable, as it was recognized as a fragment from an interfering feature, i.e., m/z 179.0355. After processing using the MetaboFR tool, 109 real features according to 73 reference standards within the total of 131 features were retained, including 69 TRFs and 40 SRFs (Table S7), which significantly increased the ratio of real features in the peak table compared with processed with only MS-DIAL or MS-CleanR (Figure 2C). After the final selection of the most intense TRFs, 73 parental features and 22 interferences were included in the final peak table with an accuracy rate of parental features of 76.84% (Figure 2E). A manual inspection of final result is described in Table S7.

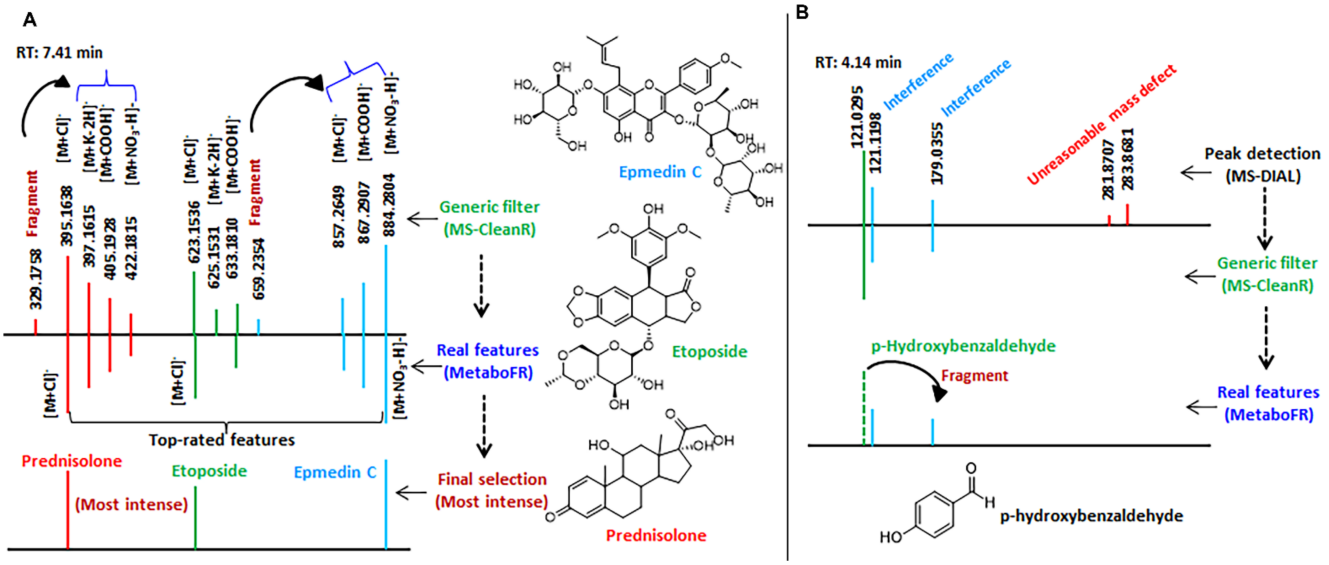

Figure 3. Two instances are illustrated to explain the performance of the tool in more detail. (**A**) The effect of MetaboFR and FR rule on the MS features at a retention time of around 7.41 min after processing by MS−CleanR. (**B**) One compound named p−hydroxybenzaldehyde (m/z 121.0295) could not be detected as it was recognized as a fragment from an interfering feature, i.e., m/z 179.0355.

3.3. Application of Workflow on Multicomplex Samples

To evaluate the application of our workflow on multicomplex samples, we set up an experiment for the metabolic profiling of 14 different species in plants of genus *Isatis*. Key chromatographic conditions were optimized to improve both the sensitivity and resolution. The ACQUITY BEH C_{18} column in positive-ion mode was selected because it has greater column capacity and yields better resolution and peak distribution (see Figure S7). The use of 2 mM ammonium acetate as the aqueous phase increased the number of detected ions and suppressed signal noise (Figure S8). According to published information on metabolites, the accepted standard mixture for metabolomics studies contains two flavonoids, five indole-related derivatives, one nucleoside, one chlorophyll and three lignans [29–31]. Such a mixture was injected into LC-MS to determine the m/z values and adduct types (Figure S9). Isovitexin, pheophorbide a, isatin and guanosine have common adduct types in positive-ion mode, in terms of the isolated quasi-molecular ions or after correlating with the corresponding adduct ions. In particular, pinoresinol 4-O-D-glucose only had adduct ions but no quasi-molecular ions. Based on the adducts identified in the different

metabolites, six adduct correlations derived from $[M + NH_4]^+$, $[M + Na]^+$, and $[M + K]^+$ were selected to enhance the accuracy of the annotating adducts.

A total of 80 pieces of raw data (i.e., 71 leaf samples from 14 *Isatis* species, 1 blank sample and 8 QC samples) were acquired and imported into the MS-DIAL to generate a peak table (4895 features). The raw data were uploaded to Metabolights (MTBL4254). After processing in MS-CleanR, a total of 823 features were defined as interferences, and were removed by generic filters, yielding a new peak table named "MS_peaks-clusters_final" that contained 4072 features, along with 1935 feature clusters (Table S8). Subsequently, Table S8 was treated using MetaboFR, together with six imported adduct correlations. A feature table containing 3426 features covering 1085 adduct flags was obtained. Meanwhile, 646 features were screened as the in-source fragments and removed directly by MetaboFR, whereas the remaining 2341 features were explored and treated as SRFs. After the final selection, the most intense TRFs were retained and 2825 features were included in the final peak table (see Table S9). In order to evaluate the performance of MetaboFR, a detailed manual annotation of each feature was carried out with the peak table after the implementation of the generic filters with the rules shown in Figure S10 (Table S9).

3.4. Detailed Annotation of Peak Tables after MetaboFR and FR Analysis of Genus Isatis

Among the 4072 features listed in Table S8, 1105 TRFs, 2020 SRFs, 860 fragments and 87 interferences (unknown adducts, isotopes, etc., referred to as unrecognized interferences) were annotated following a manual inspection. It should be noted that the feature attributions of fragments were labeled according to the Pearson's Correlation ($\geq$0.8) to the corresponding quasi-molecular ions within a time range of 0.03 min (Figure S5). Figure 4 presents the entire evolution of the peak table; the corresponding features with different attributions are listed in bar charts by comparing with detailed manual annotations. It is clear that a large quantity of fragments persisted even after the generic filters had been applied. After implementing our workflow and comparing the quality of the peak table obtained from MS-DIAL/MS-CleanR, the accuracy rate of real features was improved to 87.72%, whereas the number of fragments decreased to 9.38% (Figure 5A,B). As shown in Figure 5C,D, the false annotations compared with the manual inspection were caused from the false recognition of adducts and fragments by metaboFR processing. In Figure 5E,F, the occurrences of false recognized and unrecognized adducts or fragments were due to unreasonable feature clustering and generic filtration. As shown in Figure 5E, feature m/z 436.0651 (RT: 4.24 min) and m/z 420.0889 (4.30 min) were incorrectly recognized as $[M + K]^+$ and $[M + Na]^+$ within the same feature cluster. For the manual inspection, feature m/z 420.0889 exhibited no correlation and a separable retention time (>0.03 min) with m/z 436.0651, which was correlated to another feature, i.e., m/z 691.1317, at the same retention time in this cluster, indicating that m/z 420.0889 was an in-source fragment of m/z 691.1317. Among the unrecognized adducts, Figure 5F shows that two features, i.e., m/z 381.1161 (RT: 4.12 min, sodium adduct) and m/z 397.0894 (4.11 min, potassium adduct), exhibited adduct correlations after peak detection. However, m/z 397.0894 was removed by the generic filter due to its unstable MS behavior among the sample classes, which was unable to flag both features by MetaboFR.

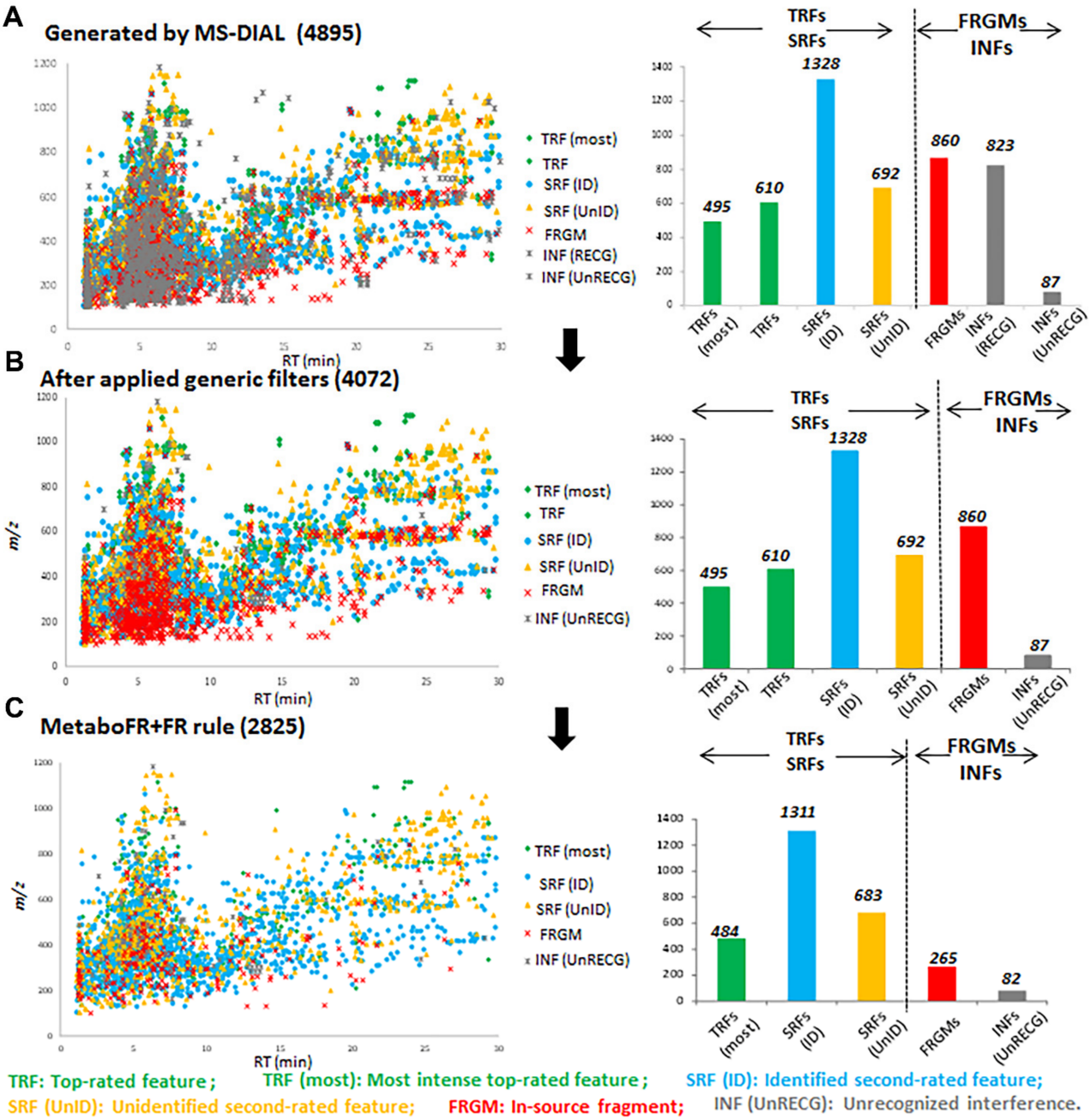

Figure 4. Evolution of the peak table by implementing the FR rule and quantifying the corresponding features with different attributions following a manual inspection, presented in bar charts. (**A**) The peak table generated by peak detection from MS−DIAL. (**B**) The application of generic filters eliminated the recognized interference. (**C**) MetaboFR combined with the FR rule preserved the most prominent TRFs and SRFs to yield a peak table that included most of the real features.

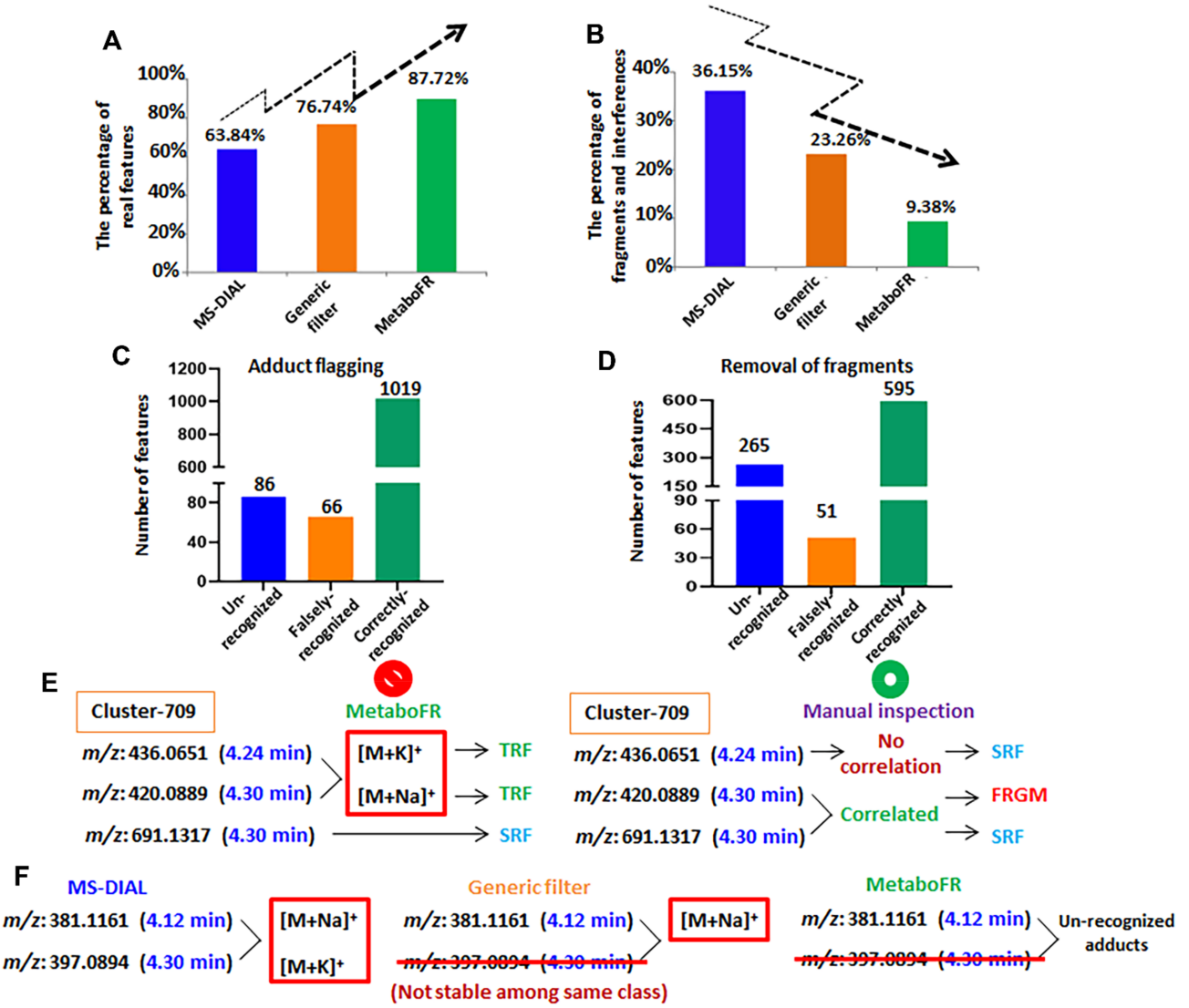

Figure 5. The effect of implementing our workflow to screen real features in metabolomics studies. (**A**) The percentage of real features identified by our workflow exceeds that of the co-implementation of MS−DIAL and MS−CleanR. (**B**) The existence of in-source fragments decreases substantially upon implementation of our workflow. (**C**) The effect of adduct flagging by MetaboFR tool. (**D**) The effect of fragment removal by MetaboFR. (**E**) False adduct recognition led to the false annotation of corresponding features by MetaboFR. (**F**) Some unstable features were removed by generic filters to cause the unrecognition of adduct correlations.

It was also noted that the quality of the imported peak table can greatly influence the results of a metabolomics study. The two- (2D) and three-dimensional (3D) PCA results by importing only validated real parental features are shown in Figure 6C, indicating that 14 *Isatis* species were unsupervised clustered into five groups. Similar clustering results were observed when the MetaboFR and FR rules were implemented, revealing nearly the same distribution when validated real features were imported (Figure 6B,C). Interestingly, only the most intense imported validated TRFs (495 features, Table S10) yielded five clusters, revealing that TRFs with a relatively high degree of confidence can be applied to output the most relevant results for complicated datasets (Figure 6D). Figure 6A presents a totally different distribution of different species in our PCA analysis compared with the other three PCA results, indicating that the lower ratio of real features may give rise to incorrect interpretations in metabolomics studies.

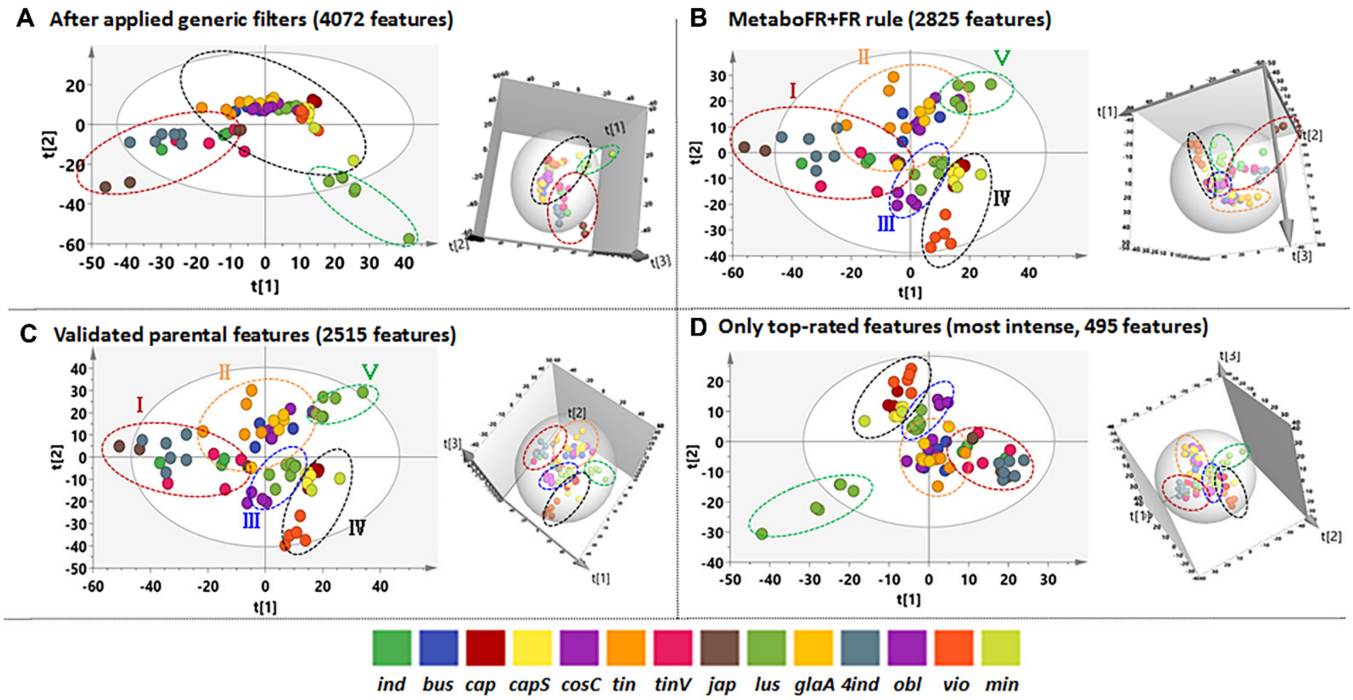

Figure 6. Different peak tables yield differences in the statistical results for metabolomics studies. (**A**) The 2D and 3D PCA results after the application of MS−CleanR (4072 features). (**B**) The 2D and 3D PCA results obtained by importing the features after the application of the MetaboFR and FR rules (2825 features) (**C**) The 2D and 3D PCA results obtained by importing only the validated parental features (2515 features) (**D**) The 2D and 3D PCA results obtained by importing only the most intense TRFs (495 features).

3.5. Chemical Analysis of Genus Isatis

Of the 2515 parental features, 109 metabolites were identified in the in-house database. The identification rate for the most intense TRFs (495 features; see Table S10) was 74.54%, and that for the SRFs was 65.74%. Consequently, 1697 metabolites were tentatively identified to give a comprehensive chemical interpretation of 14 *Isatis* species (see Table S11). The major metabolites were lipids, phenylpropanoids, organic acids and indole derivatives (Figure 7A). Five groups were clustered based on our PCA analysis; the relative contents of major metabolites among the 14 species is presented in Figure 7A. Four species in group I had a greater proportion of indoles and their derivatives. In particular the *4ind* compared with natural diploid progenitor (*ind*) had a higher proportion of indoles, alkaloids and phenylpropanoids. Group II contained four species derived from Europe and exhibited abundant organic acids and carbohydrates. Compared with groups I and II, groups III and V had a higher proportion of phenylpropanoids and fewer indole derivatives. The chemical compositions of *cap* and *cap*S were similar to those of the two desert species of *vio* and *min*, which were dominated by lipids yet lacked indole derivatives. Based on our characterization, it was found that the indole-related metabolites, which contain at least one indole moiety, exhibit tremendous chemical diversity and interspecies differences. As the representative metabolites in *Brassicaceae*, indole-related metabolites exhibit multiple bio-activities and are widely applied as pharmacological molecules [32]. Furthermore, 120 diversified indole-related metabolites were selected (Table S12); their relative contents among 14 species are shown as a heat-map in Figure 7B. The result clearly shows that few indole-related metabolites existed in group IV, especially for the bare accumulation in *cap* and *cap*S. Figure 7C indicates the relative content of nine metabolites involved in the indole-related biosynthesis pathways. Accordance semiquantitative results were obtained, indicating that only small amounts of these metabolites were present in group IV, and only rarely were any of these metabolites detected in *cap* and *cap*S, demonstrating that the

differences in accumulation were derived from the biosynthetic capacities of indole-related metabolites in plants of genus *Isatis*. In order to measure the diversity of indole-related compounds, we evaluated the contents of indole in plants of genus *Isatis* (Figure S11), proving that the existence of interspecific variability mainly occurred via downstream indole-related biosynthesis pathways among different species. Therefore, based on metabolomics data mining from a large dataset, the indole-related metabolites were found to be important chemical markers for distinguishing individual species of the genus *Isatis*. In addition, *cap* and *capS* may be used as natural mutants to reveal indole related biosynthesis in plants of genus *Isatis*. To summarize, we expect that our approach will facilitate metabolomics studies involving massive datasets, and as such, will provide more abundant chemical information in various scientific fields. Also, the proposed method will help to find suitable plants for the exploration of key metabolites during natural product biosynthesis.

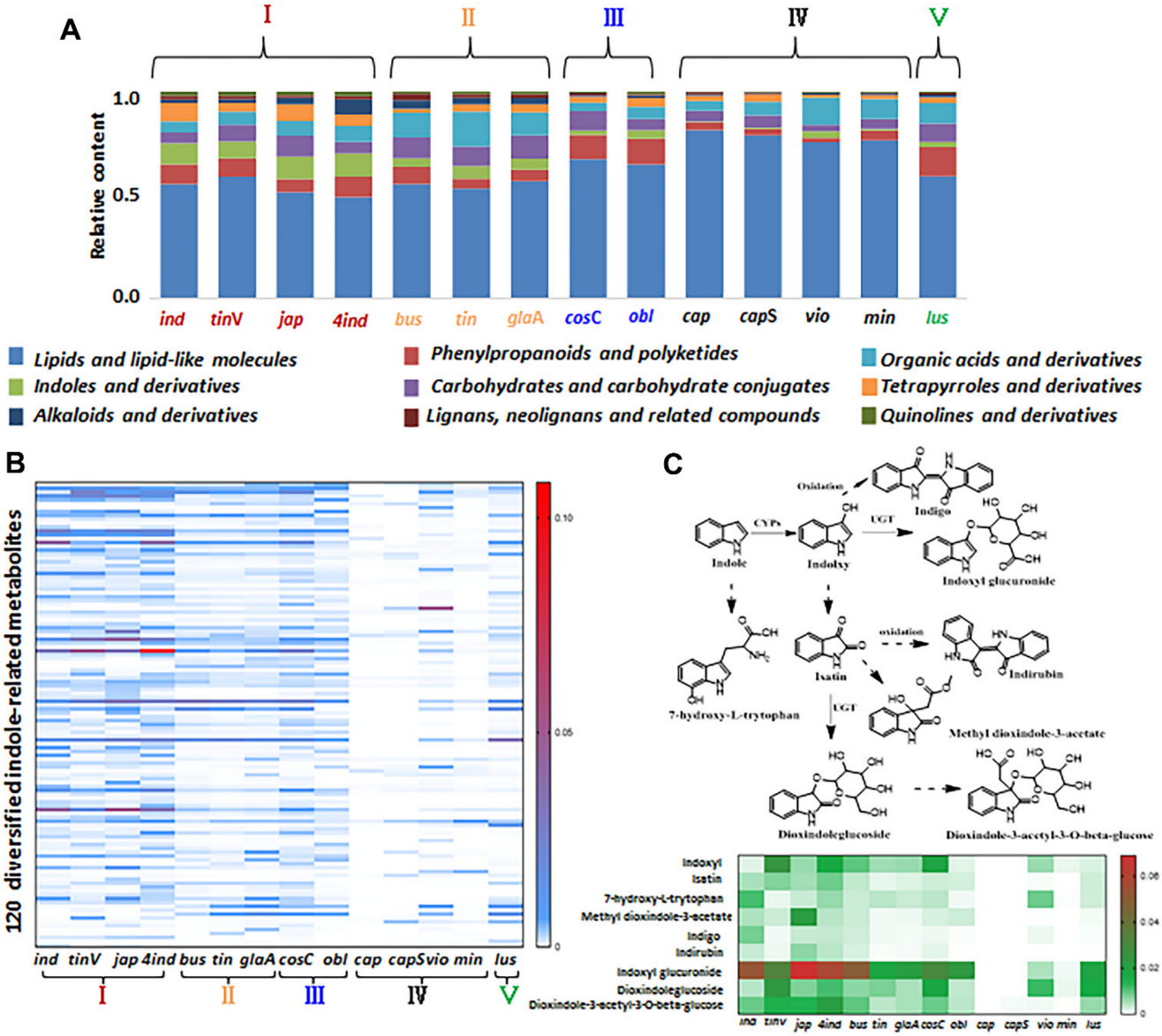

Figure 7. Comprehensive chemical interpretation and distribution of major metabolites among the 14 species of genus *Isatis*. (**A**) Bar chart showing the relative content of different classes of metabolites among the 14 species. (**B**) A heat-map showing the distribution differences of 120 indole-related metabolites among the 14 species within the five groups. (**C**) A heat−map showing the distribution differences of nine indole−related metabolites belonging to indole−related biosynthetic pathways among 14 species.

4. Conclusions

The systematic characterization of complex mixtures of metabolites remains a substantial challenge, especially for real feature screening. Therefore, combined with the workflow of real features screening and unbiased identification, we obtained as much structural information regarding detectable metabolites as possible. In our research, MetaboFR (Supplementary file S13), based on the MS-DIAL/MS-CleanR suite, was developed to screen real features and a comprehensive workflow, from raw data to final annotated peak list, was provided. The utility of this workflow is demonstrated by the fact that by analyzing secondary metabolites in 14 species of genus *Isatis*, 87.72% of real features were retained and 69.19% of the in-source fragments were removed. After careful manual checking, 1697 MS features were tentatively identified. Moreover, indole derivatives, which have been shown to be the medicinal basis of the use of *Isatis indigotica* Fort. and *Isatis tinctoria* L. [22,33], were explored as chemical markers by comparing differences in metabolites among 14 different species. More importantly, indole derivatives were rarely found in *cap* and *cap*S, which may provide natural mutant plants for indole derivative biosynthesis. Significant differences in indole derivative biosynthesis also indicated the chemical evolution of indole derivatives in genus *Isatis*. To summarize, we expect that our approach will facilitate metabolomics studies with massive datasets and provide more abundant chemical information for natural product research and other scientific fields.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cells11050907/s1, Figure S1. The phenotypes of 14 *Isatis* species. Figure S2. Structural information of metabolites for li-1~66. Figure S3. Structural information of metabolites for li-67~121. Figure S4. Structural information of metabolites for li-122~170. Figure S5. Structural information of metabolites for li-171~218. Figure S6. Structural information of metabolites for li-219~269. Figure S7. Investigation of different chromatographic columns for achieving better separation and peak capacity (*ind* as an example). Figure S8. Investigation of different mobile phase additives for achieving better ionization (*ind* as an example). Figure S9. A standard mixture contains 12 purity compounds with different chemical classifications for the demonstration of mass features and adducts types before the large-scale acquisition. Figure S10. The operation of detailed manual annotation of each feature in each cluster. Figure S11 The relative content of indole compound by applying GC-MS among 12 species. Supplementary Table S1 The tentative metabolites identification with SMILES. Table S2. The information of the source of the seeds (14 species). Table S3. The detailed information of 74 reference standards. Table S4. An in-house library covers 269 metabolites from phytochemical research (up to 2020) in the whole *Isatis* L. genus. Table S5. The feature table processed by MSCleanR (reference standards). Table S6. The feature table processed by MetaboFR (reference standards). Table S7. Validation results of detection of 73 reference standards by application of feature-rating rule. Table S8. The feature table processed by MSCleanR (*Isatis* samples). Table S9. The feature table processed by MetaboFR (*Isatis* samples). Table S10 Metabolite annotation and documentation for LC-MS data (495 features). Table S11 The tentative metabolites identification in 14 *Isatis*. Table S12 Comparation of 120 indole-related metabolites among 14 species. Supplementary file S13. Tutorial of MetaboFR.

Author Contributions: Conceptualization, D.H., S.Q. and W.C.; methodology, S.Q.; software, D.H.; validation, C.Z.; formal analysis, S.Q.; investigation, Y.X. and M.L.; resources, J.C. and L.S.; data curation, D.H.; writing—original draft preparation, D.H. and S.Q.; writing—review and editing, C.Z., S.Q. and W.C.; visualization, Y.X.; supervision, W.C.; project administration, S.Q.; funding acquisition, S.Q. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (grant no. 81903745), the Research Project of Science and Technology Commission of Shanghai Municipalit (grant no. 21DZ2202300). This research was also supported by the National Key R&D Program of China (2019YFC1711000).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Raw data were uploaded to Metabolights (MTBL4254).

Acknowledgments: The authors thank Zealquest company (Shanghai, China) for providing planting place of genus *Isatis*.

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

References

1. Li, C.; Zhang, J.M.; Wu, R.J.; Liu, Y.; Hu, X.; Yan, Y.Q.; Ling, X.M. A novel strategy for rapidly and accurately screening biomarkers based on ultraperformance liquid chromatography-mass spectrometry metabolomics data. *Anal. Chim. Acta* **2019**, *31*, 47–56. [CrossRef] [PubMed]
2. Treutler, H.; Tsugawa, H.; Porzel, A.; Gorzolka, K.; Tissier, A.; Neumann, S.; Balcke, G.U. Discovering regulated metabolite families in untargeted metabolomics studies. *Anal. Chem.* **2016**, *88*, 8082–8090. [CrossRef] [PubMed]
3. Wishart, D.S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* **2016**, *15*, 473–484. [CrossRef]
4. Gorrochategui, E.; Jaumot, J.; Lacorte, S.; Tauler, R. Data analysis strategies for targeted and untargeted LC-MS metabolomics studies: Overview and workflow. *Trends Anal. Chem.* **2016**, *82*, 425–442. [CrossRef]
5. Wolfender, J.J.; Nuzillard, J.M.; Van der Hooft, J.J.J.; Renault, J.H.; Bertrand, S. Accelerating metabolite identification in natural product research: Toward an ideal combination of liquid chromatography-high-resolution tandem mass spectrometry and NMR profiling, in silico databases, and chemometrics. *Anal. Chem.* **2019**, *91*, 704–742. [CrossRef]
6. Lai, Z.J.; Tsugawa, H.; Wohlgemuth, G.; Mehta, S.; Mueller, M.; Zheng, Y.X.; Ogiwara, A.; Meissen, J.; Showalter, M.; Takeuchi, K.; et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Methods* **2018**, *15*, 53–56. [CrossRef]
7. Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.J.; Mehta, S.S.; Wohlgemuth, G.; Barupal, D.K.; Showalter, M.R.; Arita, M.; et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spec. Rev.* **2017**, *37*, 513–532. [CrossRef]
8. Gowda, H.; Ivanisevic, J.; Johnson, C.H.; Kurczy, M.E.; Benton, H.P.; Rinehart, D.; Nguyen, T.; Ray, J.; Kuehl, J.; Arevalo, B.; et al. Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses. *Anal. Chem.* **2014**, *86*, 6931–6939. [CrossRef]
9. Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinform.* **2010**, *11*, 395–405. [CrossRef]
10. Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **2015**, *12*, 523–526. [CrossRef] [PubMed]
11. Plyushchenko, I.; Shakhmatov, D.; Bolotnik, T.; Baygildiev, T.; Nesterenko, P.N.; Rodin, I. An approach for feature selection with data modelling in LC-MS metabolomics. *Anal. Methods* **2020**, *12*, 3582–3591. [CrossRef] [PubMed]
12. Ran, J.; Liu, X.Y.; Zheng, F.J.; Zhao, X.J.; Lu, X.; Zeng, Z.D.; Lin, X.H.; Xu, G.W. Removal of false positive features to generate authentic peak table for high-resolution mass spectrometry-based metabolomics study. *Anal. Chim. Acta* **2019**, *1067*, 79–87.
13. Tian, L.; Lin, L.; Bayen, S. Optimization of the post-acquisition data processing for the non-targeted screening of trace leachable residues from reusable plastic bottles by high performance liquid chromatography coupled to hybrid quadrupole time of flight mass spectrometry. *Talanta* **2019**, *193*, 70–76. [CrossRef] [PubMed]
14. Marney, L.C.; Siegler, W.C.; Parsons, B.A.; Hoggard, J.C.; Wright, B.W.; Synovec, R.E. Tile-based Fisher-ratio software for improved feature selection analysis of comprehensive two-dimensional gas chromatography-time-of-flight mass spectrometry data. *Talanta* **2013**, *115*, 887–895. [CrossRef]
15. Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T.R.; Neumann, S. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **2012**, *84*, 283–289. [CrossRef]
16. Broeckling, C.D.; Afsar, F.A.; Neumann, S.; Ben-Hur, A.; Prenni, J.E. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **2014**, *86*, 6812–6817. [CrossRef]
17. Fraisier-Vannier, O.; Chervin, J.; Cabanac, G.; Puech, V.; Fournier, S.; Durand, V.; Amiel, A.; André, O.; Benamar, O.A.; Dumas, B.; et al. MS-CleanR: A feature-filtering workflow for untargeted LC-MS based metabolomics. *Anal. Chem.* **2020**, *92*, 9971–9981. [CrossRef]
18. Kang, K.B.; Ernst, M.; van der Hooft, J.J.; da Silva, R.R.; Park, J.; Medema, M.H.; Sung, S.H.; Dorrestein, P.C. Comprehensive mass spectrometry-guided phenotyping of plant specialized metabolites reveals metabolic diversity in the cosmopolitan plant family Rhamnaceae. *Plant J.* **2019**, *98*, 1134–1144. [CrossRef]
19. Speranza, J.; Miceli, N.; Taviano, M.F.; Ragusa, S.; Kwiecien, I.; Szopa, A.; Ekiert, H. *Isatis tinctoria* L. (Woad): A review of its botany, ethnobotanical uses, phytochemistry, biological activities, and biotechnological studies. *Plants* **2020**, *9*, 298–337.
20. Qin, G.W.; Xu, R.S. Recent advances on bioactive natural products from Chinese medicinal plants. *Med. Res. Rev.* **1998**, *6*, 375–382. [CrossRef]
21. Ghosh, R.; Chakraborty, A.; Biswas, A.; Chowdhuri, S. Depicting the inhibitory potential of polyphenols from *Isatis indigotica* root against the main protease of SARS CoV-2 using computational approaches. *J. Biomol. Struct. Dyn.* **2020**, *9*, 1–12. [CrossRef] [PubMed]

22. Xiao, Y.; Feng, J.X.; Li, Q.; Zhou, Y.Y.; Bu, Q.T.; Zhou, J.H.; Tan, H.X.; Yang, Y.B.; Zhang, L.; Chen, W.S. IiWRKY34 positively regulates yield, lignan biosynthesis and stress tolerance in *Isatis indigotica*. *Acta Pharm. Sin. B* **2020**, *10*, 2417–2432. [CrossRef] [PubMed]
23. Deng, J.L.; Ma, Y.; He, Y.Q.; Yang, H.; Chen, Y.H.; Wang, L.; Huang, D.D.; Qiu, S.; Tao, X.; Chen, W.S. A Network Pharmacology-Based Investigation to the Pharmacodynamic Material Basis and Mechanisms of the Anti-Inflammatory and Anti-Viral Effect of *Isatis indigotica*. *Drug. Des. Dev. Ther.* **2021**, *15*, 3193–3206. [CrossRef] [PubMed]
24. Tong, Q.; Zhang, C.; Tu, Y.; Chen, J.F.; Li, Q.; Zeng, Z.; Wang, F.Y.; Sun, L.N.; Huang, D.D.; Li, M.M.; et al. Biosynthesis-based spatial metabolome of *Salvia miltiorrhiza* Bunge by combining metabolomics approaches with mass spectrometry-imaging. *Talanta* **2022**, *238*, 123045. [CrossRef]
25. 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. *Anal. Chem.* **2016**, *88*, 7946–7958. [CrossRef] [PubMed]
26. Feunang, Y.D.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; et al. ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminform.* **2016**, *8*, 61–80. [CrossRef]
27. Broadhurst, D.; Kell, B. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* **2006**, *2*, 171–196. [CrossRef]
28. Tsugawa, H.; Nakabayashi, R.; Mori, T.; Yamada, Y.; Takahashi, M.; Rai, A.; Sugiyama, R.; Yamamoto, H.; Nakaya, T.; Yamazaki, M.; et al. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nat. Methods* **2019**, *16*, 295–298. [CrossRef]
29. Nguyen, T.; Marcelo, P.; Gontier, E.; Dauwe, R. Metabolic markers for the yield of lipophilic indole alkaloids in dried woad leaves (*Isatis tinctoria* L.). *Phytochemistry* **2019**, *163*, 89–98. [CrossRef]
30. Nguyen, T.K.O.; Jamali, A.; Grand, E.; Morreel, K.; Marcelo, P.; Gontier, E.; Dauwe, R. Phenylpropanoid profiling reveals a class of hydroxycinnamoyl glucaric acid conjugates in *Isatis tinctoria* leaves. *Phytochemistry* **2017**, *144*, 127–140. [CrossRef]
31. Mohn, T.; Plitzko, I.; Hamburger, M. A comprehensive metabolite profiling of *Isatis tinctoria* leaf extracts. *Phytochemistry* **2009**, *70*, 924–934. [CrossRef] [PubMed]
32. Katz, E.; Nisani, S.; Chamovitz, D.A. Indole-3-carbinol: A plant hormone combatting cancer. *F1000Research* **2018**, *7*, 689–697. [CrossRef] [PubMed]
33. Chen, Q.; Lan, H.Y.; Peng, W.; Rahman, K.; Liu, Q.C.; Luan, X.; Zhang, H. *Isatis indigotica*: A review of phytochemistry, pharmacological activities and clinical applications. *J. Pharm. Pharmacol.* **2021**, *73*, 1137–1150. [CrossRef] [PubMed]