Revelation of the metabolic pathway of hederacoside C using an innovative data analysis strategy for dynamic multiclass biotransformation experiments

Revelation of the metabolic pathway of Hederacoside C using an innovative data analysis strategy for dynamic multiclass biotransformation experiments

Laura Peeters[1], Charlie Beirnaert[2], Anastasia Van der Auwera[1], Sebastiaan Bijttebier[1], Tess De Bruyne[1], Kris Laukens[2], Luc Pieters[1], Nina Hermans[1], Kenn Foubert[1]

Affiliation

[1] Natural Products & Food Research and Analysis (NatuRA), Department of Pharmaceutical Sciences, University of Antwerp, Universiteitsplein 1, 2610 Antwerp, Belgium

[2] Adrem Data Lab, Department of Mathematics - Computer Sciences, University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium

Correspondence

Laura Peeters, Natural Products & Food Research and Analysis (NatuRA), Department of Pharmaceutical Sciences, University of Antwerp, Universiteitsplein 1, 2610 Antwerp, Belgium, E-mail: laura.peeters@uantwerpen.be, Tel.: +32 3 265 27 20

Highlights

- A workflow was established for unbiased screening of metabolites.
- *In vitro* biotransformation was performed, using GI enzymes and fecal microflora.
- An automated workflow was developed to analyze longitudinal LC-MS data.
- The unbiased screening revealed the stepwise elimination of sugar moieties.
- The data analysis workflow can replace laborious human data revision processing.

Abstract

Although some herbal remedies have been used for decades, little is known about the active compounds and the mechanism of action. Many natural products, such as glycosides, can be considered as prodrugs, which become active after biotransformation.

To optimize the workflow of *in vitro* biotransformation followed by automated data analysis, hederacoside C was used as a model compound for saponins. Hederacoside C was subjected to gastrointestinal enzymes and fecal microflora. Samples were analyzed with UHPLC-PDA-HRMS before, during and after *in vitro* biotransformation, which allowed the monitoring of the relative abundances of the compound and its metabolites.

The data-analysis workflow was optimized to render as much information as possible from the longitudinal LC-MS data. XCMS was used to convert the raw data into features via peak-picking, followed by grouping, and EDGE was used for the extraction of significant differential profiles. To evaluate if the workflow was suitable for dynamic multiclass metabolomics data, an interactive Shiny web app was developed in R to rate the quality of the resulting features. These ratings were used to train a random forest model for predicting experts response. A performance analysis revealed that the random forest model was capable of correctly predicting the reviewers response in most cases (AUC 0.926 with 10 fold cross validation).

The automated data analysis workflow was used for unbiased screening for metabolites and revealed the biotransformation of hederacoside C. As expected, a decrease in relative abundance of hederacoside C was observed over time. Additionally, the relative abundance of metabolites increased, illustrating the biotransformation of hederacoside C, especially in the colon phase, where microbial fermentation takes place. Stepwise progressive elimination of sugar moieties was the major metabolic pathway.

Keywords: hederacoside C, biotransformation, metabolomics, data-analysis

1 INTRODUCTION

The use of plants for therapeutic purposes is as ancient as human history. The earliest recorded evidence of their use in Indian, Chinese, Egyptian, Greek, Roman and Syrian texts dates back about 5000 years. Up to today, herbal medicines are still popularly consumed. Although many herbal remedies have been used for centuries or even longer, in many cases little is known about their active constituents and mechanism of action [1, 2]. Identifying these active compounds can provide new insights in the continuing search for new therapeutic agents.

It is well known that many natural products are so-called prodrugs, which become active after biotransformation [3, 4]. Nevertheless, this aspect is usually overlooked when searching for new therapeutic agents using classical approaches. In nature, many phytochemicals occur in their glycosylated form. After oral administration of a herbal extract, it is inevitably brought into contact with gastrointestinal enzymes and intestinal microflora, which might lead to transformation of the compounds. Before absorption from the gastrointestinal tract into the circulation, glycosylated compounds are presumably hydrolyzed with loss of their sugar moieties. The absorbed aglycones (or metabolites thereof) may further be biotransformed to the final active components [5, 6].

However, revealing metabolic pathways is complex since herbal extracts comprise a mixture of compounds covering a wide range of bioactive and inactive constituents, including prodrugs [7-9]. Therefore, *in vitro* models have been designed to simulate biotransformation processes [7, 10, 11]. To facilitate optimization of the experimental setup and the data analysis workflow for biotransformation studies of extracts, preliminary experiments were conducted with a single compound. Saponins comprise a wide range of secondary metabolites present in food and medicinal plants. They consist of a triterpene aglycone and one, two or sometimes three glycosyl moieties, i.e. the mono-, bi- and tridesmodic saponins respectively [12, 13]. The saponin hederacoside C was selected as a model compound for bidesmosidic saponins to investigate its *in vitro* biotransformation (Figure 1).

Concentrations of phytochemicals and their metabolites display continuous and complex modulations in function of time. Their concentration can increase, decrease or show any combination of these patterns during biotransformation. The number of methods available to analyze such dynamic data is limited and most of them are not specifically designed for metabolomics [14]. These methods are powerful analytical tools for case vs control experiments, but they are difficult to apply in multi-class experiments. In this study samples are compared with a blank and a negative control, generating a longitudinal three class experiment. The negative control sample gives additional information on how the compound behaves in time in absence of intestinal microflora. Therefore, a new approach is

required to compare dynamic data of three groups in order to elucidate the metabolic pathway of hederacoside C.

2 MATERIALS AND METHODS

2.1 CHEMICALS

Ultrapure water with a resistivity of 18.2 MΩ.cm at 25 °C was generated with a Millipore™-purification system. UHPLC-grade methanol and acetonitrile were purchased from Biosolve (Dieuze, France). Formic acid was delivered by Sigma-Aldrich (St. Louis, USA) and hederacoside C (>99%) was acquired from Bio-Connect (Huissen, The Netherlands). All other chemicals and biochemicals were purchased from Sigma-Aldrich (USA).

2.2 BIOTRANSFORMATION

A previously developed and validated *in vitro* gastrointestinal biotransformation model was used to mimic human biotransformation processes [10]. This model allows the study of biotranformation processes in the stomach, small intestine and colon using a culture of pooled human feces, avoiding extensive *in vivo* studies.

2.2.1.1 Preparation of digestive juices and fecal suspension

In the *in vitro* gastrointestinal simulation model, digestive conditions and enzymes were chosen to mimic human *in vivo* conditions. The pepsin solution was prepared by dissolving 5.4% (w/v) of pepsin powder from porcine gastric mucosa in 0.1 M HCl (2000 FIP-U mL^{-1}). To obtain a pancreatin-bile mixture, 0.4% (w/v) pancreatin from poricine pancreas and 0.8% (w/v) bile extract (porcine) were dissolved in 0.1 M $NaHCO_3$ (32000 FIP-U lipase, 143600 FIP-U amylase, 16400 FIP-U protease and 58.4 mmol bile L^{-1}).

Human fecal donors (n=3) were selected, aged 25 to 58 years, non-smokers, non-vegetarians, without any history of gastrointestinal disease, not treated with antibiotics in the last three months and with normal defecation. A suspension of 10% (w/v) feces was prepared by homogenizing the pooled fecal sample with a sterile phosphate buffer solution (0.1 M, pH 7.0) in a stomacher (Lab-blender 400, Seward Medical, London, UK) during 3 min. The phosphate buffer solution consists of $NaH_2PO_4.2H_2O$ (1.03% w/v), sodium thioglycolate broth (3.45% v/v) and glycerol (17%). Before use, the buffer solution was sonicated and autoclaved (1 bar, 121 °C). Particulate food material was removed from the slurry by using a sterile filter bag consisting of a full-page filter (Bagpage® R/25 400 mL, VWR International, Haasrode, Belgium). The mixed fecal pool was stored at -80 °C.

Prior to use, the fecal suspension was cultivated. To 10% (v/v) pooled fecal suspension in phosphate buffer 90% (v/v) sterile adapted WCB (Wilkins Chalgren Broth, Oxoid, Hampshire, UK) was added in an anaerobic glove box (5% H_2, 5% CO_2 and 90% N_2) (Jacomex Globe (Box) T_3, TCPS, Belgium). The composition of the adapted growth medium was: WCB (33 g L^{-1}), L-cysteine (0.5 g L^{-1}), Tween 80 (2.8 mL L^{-1}) and resazurine sodium salt (1.4 mg L^{-1}). The medium was autoclaved at 121 °C for 15 minutes. The bacteria were cultivated in an anaerobic environment for 17 h at 37 °C under continuously stirring. Afterwards, 80% sterile adapted WCB was added and the suspension was incubated for another hour in order to obtain a bacteria suspension of 10^8 CFU mL^{-1}. The viable cell concentrations were determined by means of decimal dilution series of the bacteria samples plated onto WCA (Wilkins Chalgren Agar, Oxoid, Hampshire, UK).

2.2.1.2 Simulation of the stomach, small intestine and colon

An amount of 10 mg hederacoside C was accurately weighed in triplicate. Additionally, a negative control sample (NC) also containing 10 mg hederacoside C and a method blank sample (MB) without test compound were included.

The samples were mixed with 47 mL ultrapure water and 3 mL pepsin solution. To simulate the gastric stage, the pH was adjusted to 2.0 using 6 M HCl. Afterwards, the samples were incubated for 60 min in a shaking water bath at 37 °C (120 strokes min^{-1}).

The small intestinal stage was simulated by adding 50 mL of ultrapure water to the gastric digest. An amount of 1 M $NaHCO_3$ was added to the negative control sample to increase the pH to 7.5. The same amount was added to the three samples using a small dialysis bag in order to obtain a gradual pH change. The samples were continuously stirred at 37 °C for 1.5 h. After 30 min, 15 mL pancreatin-bile mixture was added to the digest and stirring was continued.

Afterwards, 50 mL WCB lacking fecal suspension was added to the negative control sample and the pH was adjusted to 5.8 using 1 M HCl. The same volume of 1 M HCl was added to the three samples and the blank. Thereafter all samples were transferred to an anaerobic glove box. A volume of 50 mL bacteria in WCB (10^8 CFU mL^{-1}) was added to the samples and the blank and all mixtures were stirred for 48 h. An aliquot of each sample was taken at several time points: before biotransformation (t0), after the gastric stage (S), after the small intestinal phase (SI) and at several time points during the colon phase (C2, C4, C6, C10, C14, C18, C22, C24, C30, C39 and C48) (Figure 2). Every sample was diluted with methanol (1:2) and centrifuged for 5 min at 3500 rpm (approximately 1450 g). The supernatant was collected and samples were 10 times diluted with methanol:water (60:40) and analyzed.

2.3 INSTRUMENTAL ANALYSIS

For the qualitative UPLC-DAD-QTOF analyses of the samples, an aliquot of 5 µL was injected on a BEH-Shield-RP18 column (100 mm x 2.1 mm, 1.7 µm, Waters, Milford, MA, USA). The temperature of the column was kept at 40 °C. The mobile phase solvents consisted of water + 0.1% formic acid (A) and acetonitrile + 0.1% formic acid (B) and the gradient was set as follows (min/B%): 0/2, 1/2, 14/26, 24/65, 26/100, 29/100, 31/2, 41/2. The flow rate was set at 0.4 mL min^{-1}. For detection, accurate mass measurements were done using a Xevo G2-XS QTof spectrometer (Waters, Milford, MA, USA) coupled with an ACQUITY LC system equipped with MassLynx version 4.1 software.

During the first analysis, full scan data were recorded in ESI (+) and ESI (-) mode from m/z 50 to 1500 and the analyzer was used in sensitivity mode (approximate resolution: 22000 FWHM). The spray voltage was set at either +1.5 kV and -1.0 kV; cone gas flow and desolvation gas flow at 50.0 L/h and 1000.0 L/h respectively; and source temperature and desolvation temperature at 120 °C and 550 °C respectively. Data were also recorded using MS^E in positive and negative ionization modes (two analyses per mode), and a ramp collision energy from 20 till 30 V was applied to obtain additional structural information. Leucine encephalin was used as lock mass.

2.4 DATA ANALYSIS

Due to the longitudinal aspect (measurement in function of time) of the experiment, data are complex and difficult to interpret. Moreover, the gastrointestinal enzymes and fecal microflora present in the samples cause a lot of matrix interference, which makes analysis challenging. To select the most interesting time profiles, a workflow was optimized to render as much information as possible from the longitudinal LC-MS data. Data were converted to the open source mzXML format to allow XCMS processing [15]. The XCMS CentWave algorithm was used to convert the raw data into features via peak-picking followed by grouping. Although XCMS has the ability to discover features that are different between groups, it is not able to take the longitudinal aspect into account. Therefore EDGE was used to extract significant differential profiles (16). EDGE fits two models to every feature. The null model assumes there is no difference between the groups and fits a single cubic spline to all the data of a feature. The alternative model fits a cubic spline to each group and calculates the goodness of fit. The difference between these groups, expressed as p-values, reflects the improvement in goodness of fit.

However, this approach generates two p-values for every feature, one for sample vs blank and one for sample vs negative control, which is non-trivial to interpret or combine. To circumvent this problem, a machine learning model was constructed to obtain a single score for each feature. Additionally, false positive and false negative hits can be reduced by training the random forest model, which in turn can

be used to filter them out of the remaining data. The training dataset was composed of 217 MS time profiles in negative ion mode, selected on their low p-values from the two EDGE analyses for sample vs method blank and sample vs negative control. Each training sample consisted of 70 features: 14 time points per sample, 3 replicates per sample, 1 negative control and 1 blank sample. To review the quality of the significant features and to train the machine learning model, a Shiny web application was constructed, named tinderesting [17]. The app queries the user for the quality of features and uses these results to build the model in the background. Figure 3 shows a time profile for an *m/z* value (894.37) in negative ion mode. There is a change in intensity of the signal for the sample, which is not observed in negative control or blank. This profile can be classified as interesting by swiping right. To validate whether this model can be used to adequately predict whether a time profile is a false discovery or not, a 10-fold cross validation was performed on the training dataset.

3 RESULTS AND DISCUSSION

3.1 Hederacoside C metabolites identification

Hederacoside C consists of hederagenin linked to sugar moieties via the OH-groups at C-3 and C-28 of the aglycon. Therefore, extensive biotransformation after oral intake is expected. Gastrointestinal biotransformation was simulated *in vitro* to identify the metabolites and to monitor their levels using high resolution MS data.

The total ion chromatogram (TIC) of hederacoside C before and after biotransformation (C48, colon phase after 48 h) showed significant differences (Figure 4). Hederacoside C showed a molecular ion peak at *m/z* 1219.6063 $[M-H]^-$ and a retention time of 16.06 min. After the colon phase (C48) the extracted ion chromatogram (XIC) displayed no signal for the pseudomolecular ion *m/z* 1219.6063 $[M-H]^-$. The complete disappearance of the hederacoside C signal clearly suggests the formation of metabolites.

In agreement with literature, it was observed that biotransformation was mainly observed in the colon phase [10, 11, 12]. Breynaert et al. reported that biotransformation of chlorogenic acid was complete after 6 h of colon phase. However, they suggested that biotransformation may require more time for more complex compounds. To ensure complete biotransformation of hederacoside C, a complex saponin, samples were exposed to fecal microflora for up to 48 h.

Indeed several metabolites could be detected during the colon phase and their structures were tentatively assigned when both the mass/charge (*m/z)* ratios and the molecular formulae of the precursor ion and product ions were in agreement (Table 1).

The tentative metabolic pathway of hederacoside C is outlined in Figure 5. Consecutive removal of sugar moieties decreases polarity, which results in an increasing retention time. Deglycosylation reactions by stepwise cleavage of sugar moieties were first observed after 24 h of colon phase, leading to the formation of compound 2, 3, 4 and 5. After 39 h, compound 6 was formed. After 48 h the aglycon (compound 7) could be detected. However, all other metabolites, except compound 2, were still present in the extract after biotransformation. The data support the theory that gastrointestinal biotransformation of saponins leads to a loss of sugar moieties [12, 13, 18]. These modifications enhance absorption of the metabolites through the colonic epithelium.

These results were obtained by laborious expert scientist revision processing. The TIC of the samples were studied on every time point, but method blank and negative control were not taken into account. It is not straightforward to analyze a longitudinal multiclass experiment by comparing the chromatograms of the different samples. Although it is usually overlooked, the negative control sample provides information on the stability of the compound over time.

3.2 Data analysis

An automated data analysis workflow was used for unbiased screening for metabolites. To that end, expensive commercially available software and free open source software were compared with the developed workflow [17]. Critical aspects that contribute into the choice of a data analysis program include file compatibility, computational requirements, user friendliness and data visualization. Commercially available software has the advantage that it is compatible with the obtained data files. However, these software packages are not always able to handle the complexity of the data. They can handle comparison between groups or time levels but are not capable to compare longitudinal multiclass data. Data analysis of these different results is usually done manually by the user, which is time-consuming and can be biased. Over the last years, free and open access workflows and pipelines have gained popularity [19]. However, choosing a suitable workflow is an art without predefined rules. The choices for measurement and data analysis are driven by the biological question, the data generating process, the experimental design and the assumptions of the data analysis methods. An advantage of these open source workflows is that they can be shared, improving verifiability and reproducibility. However, hardly any methods incorporating dynamics exist so far [14].

In our case, statistical analysis to find features that display different expression over time is performed via the EDGE method [20]. This method is robust and performant for longitudinal two class experiments. However, this approach generates two p-values for every feature, one for sample vs blank and one for sample vs negative control, which is non-trivial to interpret or combine. A machine learning model was constructed to obtain a single score for each feature. Expert knowledge is

incorporated into the analysis pipeline via the in-house developed tinderesting app as a valuable addition. This includes the rating of images of data entities on their importance and the subsequent training of the machine learning model that can be used to analyze the data. Therefore, the advantage of the developed automated data analysis workflow is to be found in the interaction of the researchers with their data, with respect to the longitudinal multi class setup of the experiment [17].

The automated data analysis workflow ranked all features based on the difference in signal between sample, negative control and method blank. To validate the performance of the workflow, the presence of the tentatively identified metabolites, indicated by expert scientists data analysis, was verified using these features.

The automated data analysis workflow shows a time profile for hederacoside C during biotransformation. A decrease in relative abundance in the samples (HED) is observed over time, starting from C6 (colon phase after 6 h), which is not observed in the negative control (NC) (Figure 6A).

In addition, the signal was absent in the blank (MB), as expected. These observations are in agreement with the results of the expert scientist revision process, discussed in 3.1. The feature is rated as interesting by tinderesting with a ranking of 44 out of 10307 features. Over time, the intensity decreases in the samples and the signal is absent at the end of the gastrointestinal simulation, while it is still present in the negative control.

On the other hand, the relative abundance of the metabolites is increasing over time, which implies elimination of sugar moieties during gastrointestinal biotransformation.

Figure *6*: Time profile of hederacoside C (A) and hederagenin 3-O-α-L-rhamnopyranosyl-(1->2)-O-α-L-arabinopyranosyl ester (B) in negative ion mode.

B shows the time profile of compound 5, illustrating the increase in signal over time.

The plots in Figure 7 show all features, in negative and positive mode respectively. The p-values obtained with EDGE for sample versus blank (MBp) are plotted against the p-values for sample versus negative control (NCp). The color of the data points refers to the score obtained with tinderesting. The plots illustrate that a low p-value for sample vs method blank corresponds with a high tinderesting score. However, the correlation between the high tinderesting score and the p-value for sample vs negative control is not straightforward. This shows that it is not trivial to combine the two outcomes of the EDGE analysis. The machine learning model circumvents this problem as the outcome is a single score for every feature, independent of the number of classes in the experiment. For the metabolites found by expert scientist revision data processing p-values close to zero are observed as well as scores

close to one for the machine learning model (Table 2). All previously identified metabolites are ranked in the top 100 out of 10307 features. This implies that the machine learning model is capable of predicting interesting features.

Compounds **3** and **5** are present in the top 10 as ranked by tinderesting. Other compounds in the top 10 were tentatively identified as adducts of compound **2** (retention time of 16.2 min) and compound **3** (retention time of 21.0 min), based on the retention time and fragmentation pattern. For example, the compound ranked as number 2 according to tinderesting, with an *m/z* value of 787.4244, could be identified as the ^{37}Cl-adduct of compound **3**.

To validate if the model can adequately predict whether a time profile is interesting or not, a 10 fold cross validation was performed on the training dataset, resulting in an AUC of 0.926. Therefore, this automated data analysis workflow can replace laborious expert scientist revision processing in an unbiased manner, allowing rapid scoring of large amounts of data. Moreover, additional information is obtained since the automated workflow compares time profiles of sample with blank and negative control instead of studying single sample points.

4 CONCLUSION

In vitro biotransformation via a gastrointestinal simulation model followed by metabolomics profiling is an innovative concept to disclose metabolic pathways. An automated data analysis workflow was used for unbiased screening for metabolites and revealed the biotransformation of hederacoside C. The results of the automated data analysis workflow match the results of the expert scientist revision processing. Hederacoside C showed a decrease in relative abundance over time, especially in the colon phase, while the relative abundance in the method blank and negative control remained unchanged. Additionally, the relative abundance of metabolites increased, illustrating the biotransformation of hederacoside C. The major metabolic pathway is the stepwise elimination of sugar moieties. Since the automated data analysis workflow is able to predict interesting features, laborious expert scientist revision processing can be avoided. This insight smooths the path for future experiments with complex herbal extracts.

References

[1] Pal SK, Shukla Y. Herbal medicine: Current status and the future. Asian Pacific J Cancer Prev. 2003;4:281-8.
[2] Pal S, Shukla Y. Herbal medicine: Current status and the future2002. 281-8 p.
[3] Rautio J, Kumpulainen H, Heimbach T, Oliyai R, Oh D, Jarvinen T, et al. Prodrugs: design and clinical applications. Nat Rev Drug Discov. 2008;7(3):255-70.
[4] Kumar SV, Saravanan D, Kumar B, Jayakumar A. An update on prodrugs from natural products. Asian Pacific Journal of Tropical Medicine. 2014;7:S54-S9.
[5] Osbourn AE. Preformed Antimicrobial Compounds and Plant Defense against Fungal Attack. The Plant cell. 1996;8(10):1821-31.
[6] Mahdi JG. Biosynthesis and metabolism of β-d-salicin: A novel molecule that exerts biological function in humans and plants. Biotechnology reports (Amsterdam, Netherlands). 2014;4:73-9.
[7] Aura A-M. Microbial metabolism of dietary phenolic compounds in the colon. Phytochemistry Reviews. 2008;7(3):407-29.
[8] Stevens JF, Maier CS. The Chemistry of Gut Microbial Metabolism of Polyphenols. Phytochem Rev. 2016;15(3):425-44.
[9] Crozier A, Jaganath IB, Clifford MN. Dietary phenolics: chemistry, bioavailability and effects on health. Nat Prod Rep. 2009;26(8):1001-43.
[10] Breynaert A, Bosscher D, Kahnt A, Claeys M, Cos P, Pieters L, et al. Development and Validation of an in vitro Experimental GastroIntestinal Dialysis Model with Colon Phase to Study the Availability and Colonic Metabolisation of Polyphenolic Compounds. Planta Med. 2015;81(12-13):1075-83.
[11] Jaganath IB, Mullen W, Lean ME, Edwards CA, Crozier A. In vitro catabolism of rutin by human fecal bacteria and the antioxidant capacity of its catabolites. Free Radic Biol Med. 2009;47(8):1180-9.
[12] Wang HY, Hua HY, Liu XY, Liu JH, Yu BY. In vitro biotransformation of red ginseng extract by human intestinal microflora: metabolites identification and metabolic profile elucidation using LC-Q-TOF/MS. J Pharm Biomed Anal. 2014;98:296-306.
[13] Kong H, Wang M, Venema K, Maathuis A, van der Heijden R, van der Greef J, et al. Bioconversion of red ginseng saponins in the gastro-intestinal tract in vitro model studied by high-performance liquid chromatography-high resolution Fourier transform ion cyclotron resonance mass spectrometry. J Chromatogr A. 2009;1216(11):2195-203.
[14] Smilde AK, Westerhuis JA, Hoefsloot HC, Bijlsma S, Rubingh CM, Vis DJ, et al. Dynamic metabolomic data analysis: a tutorial review. Metabolomics. 2010;6(1):3-17.
[15] Stanstrup J, Gerlich M, Dragsted LO, Neumann S. Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. Analytical and Bioanalytical Chemistry. 2013;405(15):5037-48.
[16] Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. PNAS. 2005;102(36):5.
[17] Beirnaert C, Peeters L, Meysman P, Bittremieux W, Foubert K, Custers D, et al. Using expert driven machine learning to enhance dynamic metabolomics data analysis. bioRxiv. 2018.
[18] Wan J-Y, Liu P, Wang H-Y, Qi L-W, Wang C-Z, Li P, et al. Biotransformation and metabolic profile of American ginseng saponins with human intestinal microflora by liquid chromatography quadrupole time-of-flight mass spectrometry. Journal of Chromatography A. 2013;1286:83-92.
[19] Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic acids research. 2018;46(W1):W486-W94.
[20] Leek JT, Monsen E, Dabney AR, Storey JD. EDGE: extraction and analysis of differential gene expression. Bioinformatics. 2006;22(4):507-8.

Tables and figures

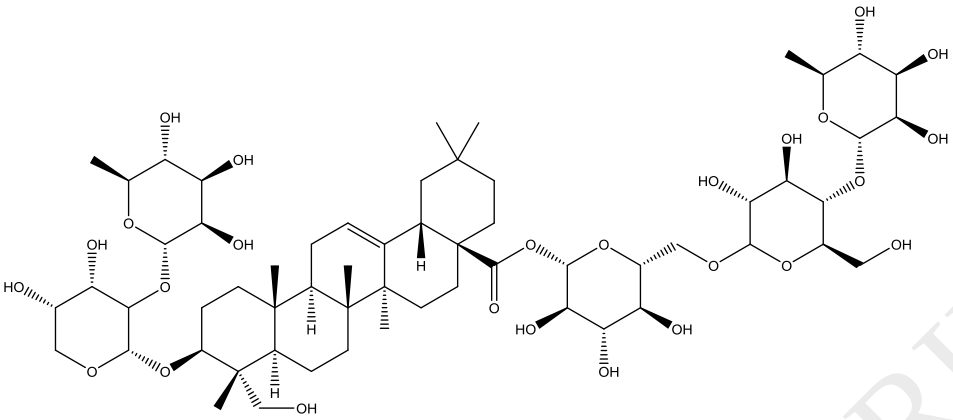

Figure 1: The chemical structure of hederacoside C.

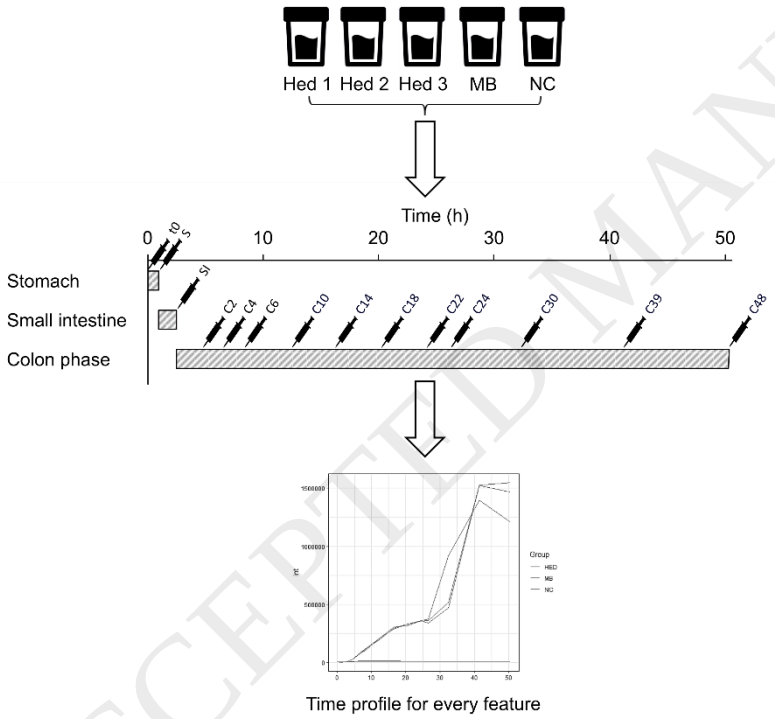

Figure 2: Overview of sample points: samples were taken at several time points: before biotransformation (t0), after the gastric stage (S), after the small intestinal phase (SI) and at several time points during the colon phase (C): 2 h, 4 h, 6 h etc.

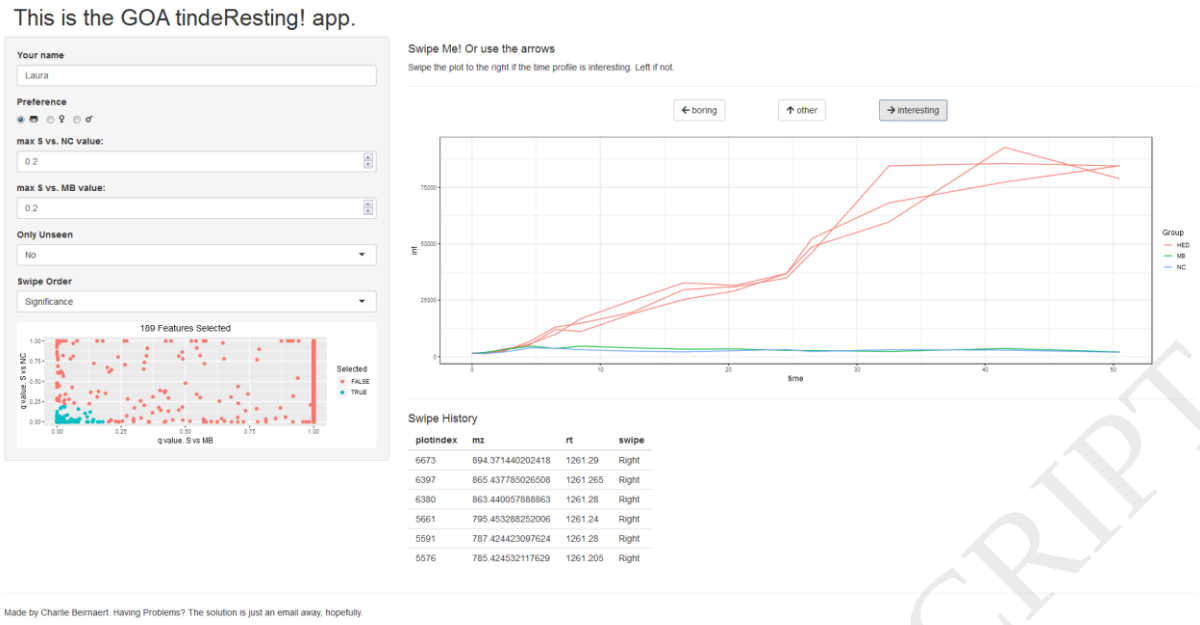

Figure 3: Screenshot of the tinderesting app for an *m/z* value (894.37).

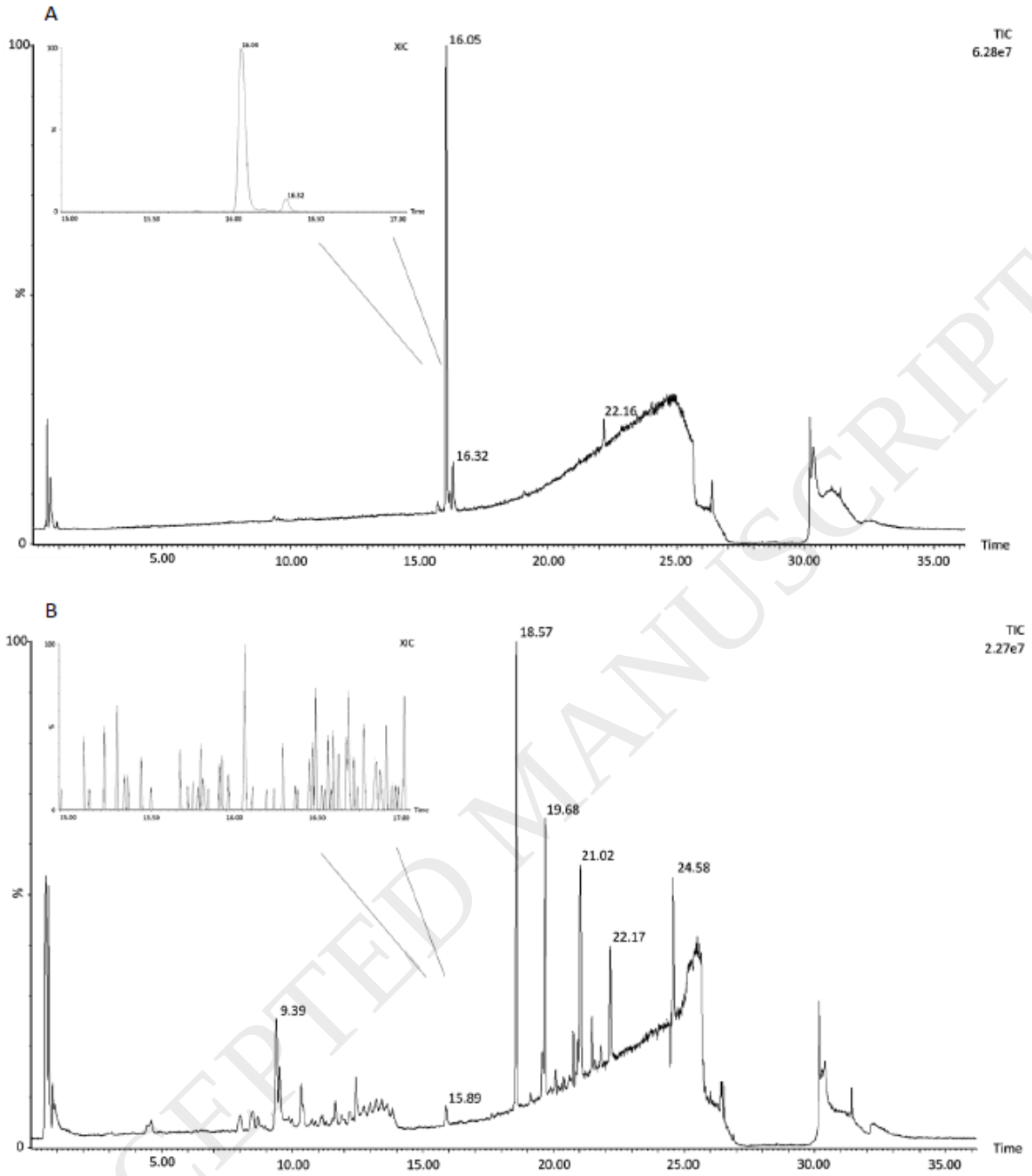

Figure 4: The total ion chromatograms (TIC) in negative ion mode of hederacoside C before (A) and after (B) biotransformation. The extracted ion chromatogram (XIC) of *m/z* 1219.6063 $[M-H]^-$ is shown in detail.

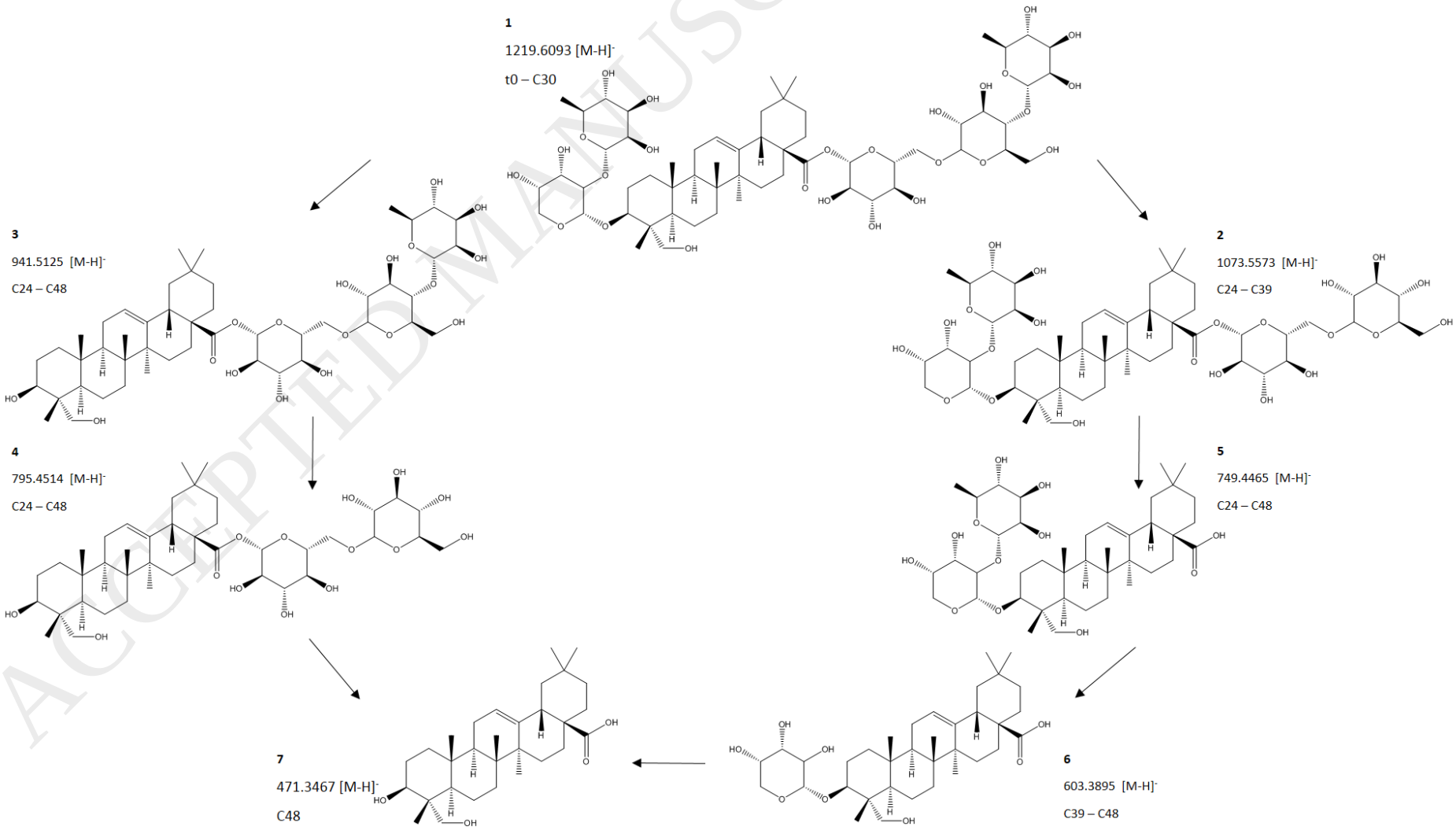

Figure 5: Metabolic pathway of hederacoside C by *in vitro* biotransformation.

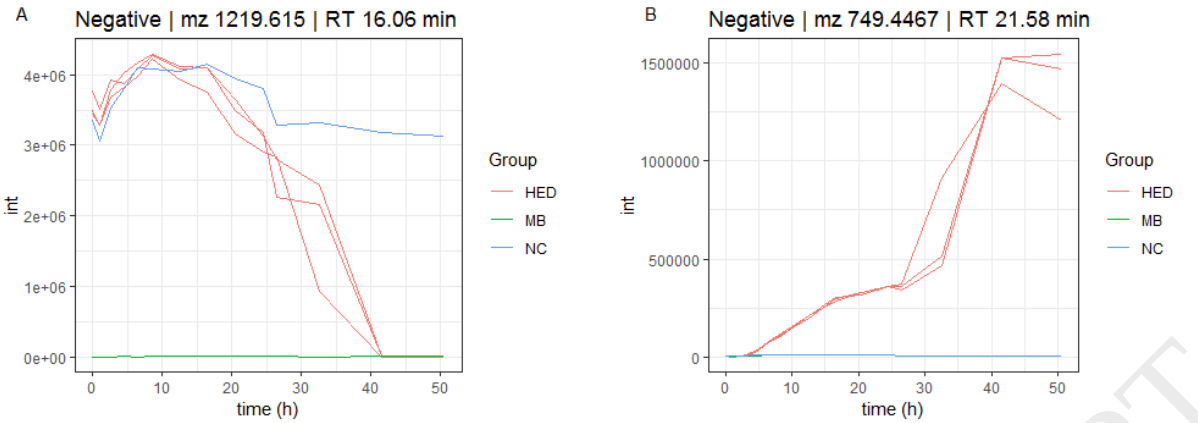

Figure 6: Time profile of hederacoside C (A) and hederagenin 3-*O*-α-L-rhamnopyranosyl-(1->2)-*O*-α-L-arabinopyranosyl ester (B) in negative ion mode.

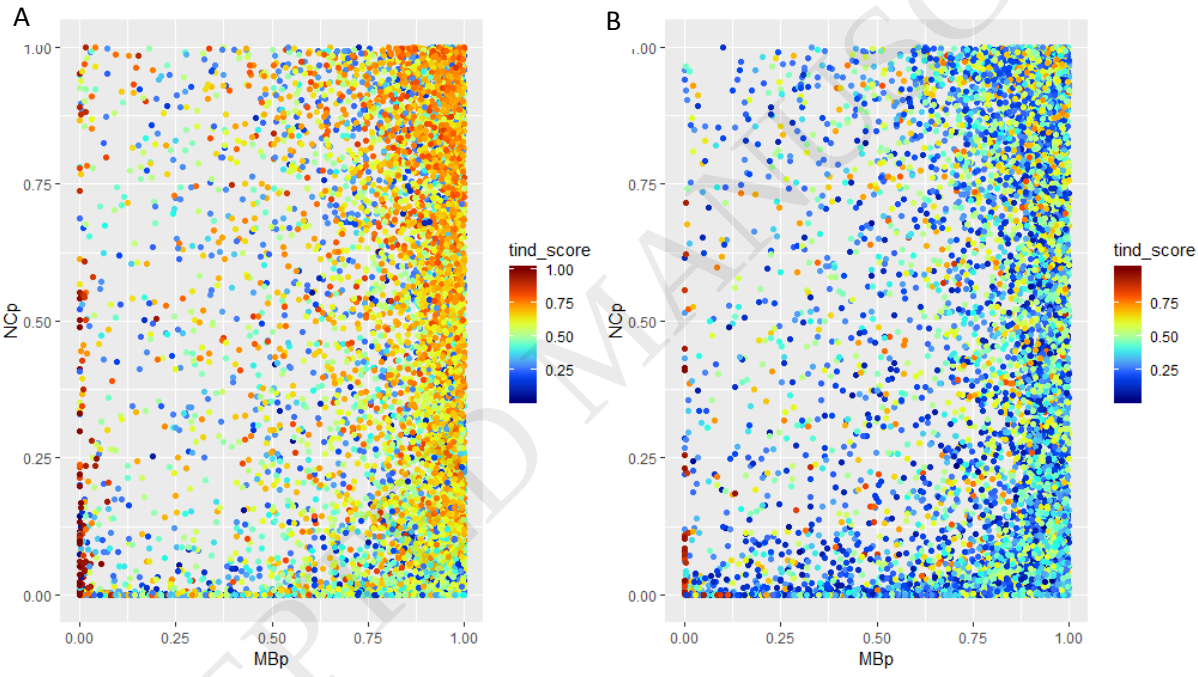

Figure 7: p-values of sample vs method blank (MBp) and sample vs negative control (NCp) are plotted against the tinderesting score for negative ion mode (A) and positive ion mode (B). A low p-value for sample vs method blank corresponds with a high tinderesting score (red dots), implying that the machine learning model is capable of predicting interesting features.

Table 1: Tentatively identified metabolites of hederacoside C.

	Compound	Molecular formula	Retention time (min)		HESI full MS[c]	Δ ppm
1	hederagenin 3-*O*-α-L-rhamnopyranosyl-(1->2)-*O*-α-L-arabinopyranoside 28-*O*-α-L-rhamnopyranosyl-(1->4)-*O*-β-D-glucopyranosyl-(1->6)-*O*-β-D-glucopyranosyl ester	$C_{59}H_{96}O_{26}$	16.06	neg	1219.6093 $[M-H]^-$	-1.6
				pos	1221.6246 $[M+H]^+$; 1243.6063 $[M+Na]^+$	-1.8
2	hederagenin 3-*O*-α-L-rhamnopyranosyl-(1->2)-*O*-α-L-arabinopyranoside 28-β-D-glucopyranosyl-(1->6)-*O*-β-D-glucopyranosyl ester	$C_{53}H_{86}O_{22}$	16.23	neg	1073.5573 $[M-H]^-$; 1119.5609 $[M-H+CH_2O_2]^-$	3.8
				pos	1075.5682 $[M+H]^+$; 1097.5554 $[M+Na]^+$	-0.7
3	hederagenin 28-*O*-α-L-rhamnopyranosyl-(1->4)-*O*-β-D-glucopyranosyl-(1->6)-*O*-β-D-glucopyranosyl ester	$C_{48}H_{78}O_{18}$	17.54	neg	941.5125 $[M-H]^-$; 987.5165 $[M-H+CH_2O_2]^-$	1.6
				pos	-	-
4	hederagenin 28-*O*-β-D-glucopyranosyl-(1->6)-*O*-β-D-glucopyranosyl ester	$C_{42}H_{68}O_{14}$	21.02	neg	795.4514 $[M-H]^-$	-2.1
				pos	-	-
5	hederagenin 3-*O*-α-L-rhamnopyranosyl-(1->2)-*O*-α-L-arabinopyranosyl ester	$C_{41}H_{66}O_{12}$	21.58	neg	749.4465 $[M-H]^-$; 795.4465 $[M-H+CH_2O_2]^-$	-1.5
				pos	751.4639 $[M+H]^+$; 773.4453 $[M+Na]^+$	0.8
6	hederagenin 3-*O*-α-L-arabinopyranosyl ester	$C_{35}H_{56}O_8$	21.47	neg	603.3895 $[M-H]^-$; 649.3955 $[M-H+CH_2O_2]^-$	-0.3
				pos	-	-
7	hederagenin	$C_{30}H_{48}O_4$	24.00	neg	471.3467 $[M-H]^-$	-1.5
				pos	-	-

Table 2: Ranking of the previously identified compounds via tinderesting.

Compound	*[m/z]*	Rt (min)	p-value blank vs sample	p-value NC vs sample	Tinderesting score	Ranking tinderesting
1	1219.6136	16.3	4.66E-15	2.30E-08	0.996	44
1	1219.615	16.1	7.44E-15	3.21E-11	0.992	65
2	1073.5542	16.2	4.91E-09	2.45E-09	0.998	27
3	749.4467	21.6	4.88E-15	4.88E-15	0.994	53
3	749.448	21.0	5.00E-15	9.44E-15	1.000	1
3	749.4474	16.3	6.90E-12	2.58E-05	0.988	78
4	603.3896	21.5	4.51E-10	4.49E-10	0.988	74
4	603.3896	22.3	7.95E-09	8.11E-09	0.996	36
5	941.5114	17.5	1.51E-10	1.48E-10	1.000	5
6	795.4527	17.7	5.63E-05	5.66E-05	0.994	54
6	795.4533	21.0	0.00E+00	0.00E+00	0.996	40
6	795.4542	16.1	3.42E-13	7.83E-11	0.996	41
7	471.3478	25.2	2.22E-16	2.22E-16	0.990	68