# *MetIDfyR*, an open-source R package to decipher small-molecule drugs metabolism through high resolution mass spectrometry

Vivian Delcourt,[*,†,‡] Agnès Barnabé,[†,‡] Benoit Loup,[†] Patrice Garcia,[†] François André,[†] Benjamin Chabot,[†] Stéphane Trévisiol,[†] Yves Moulard,[†] Marie-Agnès Popot,[†] and Ludovic Bailly-Chouriberry[†]

†*GIE-LCH, Laboratoire des Courses Hippiques, Verrières-le-Buisson, France*

‡*Equal contribution*

E-mail: v.delcourt@lchfrance.fr

**Abstract**

After administration to humans or animals, small-molecule drugs most frequently undergo several biochemical transformations by the endogenous enzymatic machinery, called phase I and phase II metabolism. These molecular processes allow organisms to eliminate xenobiotics through modification of their chemical properties and generate metabolites. With recent advances in analytical chemistry, LC-HRMS/MS has become an essential tool for metabolite discovery and detection. Even if most common drug transformations have already been extensively described, manual search of drug metabolites in LC-HRMS/MS datasets is still a common practice in toxicology laboratories, disabling efficient metabolite discovery. Furthermore, the availability of free open-source software for metabolite discovery is still limited. In this article, we present *MetIDfyR*, an open-source and cross-platform R package for *in-silico*

drug phase I/II biotransformations prediction and mass-spectrometric data mining. *MetIDfyR* has proven efficacy for advanced metabolite identification in semi-complex and complex mixtures in *in-vitro* or *in-vivo* drug studies and is freely available at github.com/agnesblch/MetIDfyR.

# Introduction

Metabolism of small-molecule drugs is the scientific field focused on determining the fate of compounds absorbed by an organism which most frequently leads to the compound's elimination and ensures detoxification. This is a crucial step involved in multiple stages of a compound's life such as drug discovery and development,[1–3] preliminary *in-vitro* and *in-vivo* experiments, pharmaco-kinetics/dynamics studies[4,5] and forensics,[6–10] including drug-testing of human and animal athletes.[11–17]

Over the past few decades, multiple key techniques and findings conducted the scientific community to expand knowledge on these complex molecular pathways. This includes the use of radioisotopes,[18] immunoassays, the identification of enzymes responsible of various and possibly complex biotransformations,[19] leading to the definition of most prevalent biochemical reactions for detoxification[20] and key actors of these mechanisms such as the P450 cytochrome enzyme family.[21,22] In parallel, analytical chemistry applied to drug metabolism underwent several breakthroughs in terms of techniques, sensitivity and precision by means of advanced nuclear magnetic resonance[23] and mass spectrometry.[24] Because of its versatility and sensitivity, liquid-chromatography hyphenated to mass spectrometry (LC-MS) has become an essential tool for drug metabolite identification, structure elucidation and detection.[1,7,25,26] In addition, the precise mass measurement enabled by high-resolution tandem mass-spectrometry (HRMS/MS) allows the elementary composition (EC) determination of the detected compounds and fragments, increasing confidence in compound identification and complex mixture characterization *via* non-targeted strategies such as data-dependent

(DDA) and data-independent (DIA / SWATH) acquisition strategies[27].

In analytical chemistry laboratories studying small-molecule drugs and their associated metabolites, the determination of relevant metabolites was associated with the administration of compounds to animals, which sometimes had to be repeated to allow metabolites identifications. Because of ethical and welfare concerns, the administration of drugs to animal or human subjects are declining, promoting *in-vitro* preliminary approaches which allow rapid, cost-effective and mostly accurate metabolite definition. *In-vitro* techniques rely on mimicking the endogenous enzymatic machinery employing recombinant cytochrome P450,[28] cultured hepatocytes,[29] homogenized liver extracts,[15] liver and kidney microsomes[30] or electrochemical devices.[31,32] Drug of interest (DoI) is mixed with reaction media and incubated prior to extraction followed by LC-HRMS analysis. LC-HRMS data can then either be evaluated employing dedicated software or manually, which is still a common practice.

Based on the DoI EC and various chemical transformations mass-shifts, the analyst can calculate theoretical masses of putative metabolites which can be searched in experimental data. However, manual processing is time-consuming and can be error-prone, especially in case of successive modifications, possibly leading to incomplete metabolite identification and suboptimal detection. In addition, cross-platform and database-free solution for metabolite prediction and evaluation through LC-HRMS data are still limited. In this article, we present the development of MetIDfyR, a R[33,34] package to predict and detect metabolites out of EC(s) and LC-HRMS/MS data. MetIDfyR evaluates each predicted biotransformation *via* a multi-layer strategy. Each detected *m/z* is evaluated at MS level, considering its intensity and quality of isotope pattern and also at MS/MS level, through MS/MS fragment matching and MS/MS spectra similarity assessment with DoI. This package has shown rapid and efficient metabolite identification capabilities in semi-complex and highly complex matrices. Indeed, MetIDfyR could highlight multiple referenced metabolites of LGD-4033 (also named

"Ligandrol or VK5211"), an investigational selective androgen receptor modulator (SARM) after horse *in-vitro* phase I metabolism experiments but also Cocaine and its related metabolites in a positive case of drug abuse in human urine. MetIDfyR uses computer processor parallelization allowing high-throughput analysis and all steps may be realized *via* a code-free Shiny[35] graphical user interface and is freely available at github.com/agnesblch/MetIDfyR.

# Materials and methods

## Chemicals and Reagents

Ultrapure water was generated using a Milli-Q 7010 (Merck-Millipore, Darmstadt, Germany). LC-MS grade acetonitrile (ACN), Isopropanol, ammonium hydroxide ($NH_4OH$), dichloromethane and methanol (MeOH) used for metabolite processing were purchased from Carlo-Erba (Val-de-Reuil, France). ACN and formic acid (FA) used for LC-HRMS/MS analysis were ULC-MS grades purchased from Biosolve (Dieuze, France). Male and female S9 horse liver fractions, were purchased from XenoTech LLC (Kansas City, Missouri, USA). Dihydronicotinamide adenine dinucleotide phosphate (NADPH, N5130) and magnesium chloride (MgCl2, M1028) were purchased from Merck (Darmstadt, Germany). Certified material LGD-4033 was purchased from Cayman Chemicals (Ann Arbor, Michigan, USA). Desalting C18 pipette tips were purchased from Pierce (Thermo Scientific, Waltham, Massachusetts, USA). BCX2 (C8-Benzenesulfonic acid) solid phase extraction 96-well plates were purchased from UCT (Bristol, Pennsylvania, USA). *Aspergillus melleus* Proteinase was purchased from Sigma (Saint-Louis, Missouri, USA) and *E. coli K12* $\beta$-glucuronidase was purchased from Roche Diagnostics (Basel, Switzerland).

## Sample preparation

### *In-vitro* metabolism experiment

*In-vitro* metabolism was performed using S9 fractions from horse liver microsomes. Briefly, horse male and female S9 fractions were mixed and suspended at a final concentration of 0.5 mg/mL in phosphate buffered saline solution (50 mM, pH 7.4) containing $MgCl_2$ (5 mM). NADPH was used as a cofactor for phase I reactions and was added to reaction mixture to reach 5 mM. Finally, 1 µL of 1 mg/mL MeOH stock solution of LGD-4033 was added to the reaction mixture (100 µL). Reactions were incubated overnight using a thermomixer with pulsed rotation (800 rpm, 10 s pulse with 10 s pause) at 37 °C.

Reactions were quenched by adding one volume of ice-cold ACN to the reaction mixture followed by protein precipitation by centrifugation (10 000 rpm, 5 min). Supernatant was collected and stored at 4 °C until analysis. Prior to LC-HRMS/MS analysis, half of the reaction volume was desalted on a C18 pipette tip and analyzed by LC-HRMS/MS.

### Extraction of cocaine and associated metabolites in a positive case human urine sample

Human urine sample previously tested positive for Cocaine, Benzoylecgonine and Ecgonine methyl ester by means of an accredited method and certified reference materials was subjected to mixed-mode solid-phase extraction (SPE). Briefly, human urine (1.5 mL) was buffered with 0.3 mL of PBS (1 M, pH 5.8) to reach final pH range within 6.0-6.2 to which 25 µL of proteinase and 25 µL of *E. coli K12* $\beta$-glucuronidase were added and incubated for 1 hour at 55 °C.

Hydrolysed urine was extracted on 96-well BCX2 SPE plates. SPE media was conditioned with 500 µL MeOH, 500 µL $H_2O$ and 500 µL of PBS and 1.6 mL of hydrolysed urine were

allowed to flow through SPE media. The media was washed by 1.5 mL of acetic acid in $H_2O$ (1 mol/L) and 1 mL of MeOH, and compounds were eluted by 1.8 mL of isopropanol/32 % $NH_4OH$/dichloromethane (0.18/0.02/0.8, v/v/v). The extract was dried and resuspended in 100 µL of $H_2O$/MeOH (0.95/0.05, v/v) and analyzed by LC-HRMS/MS.

## Liquid chromatography

Reverse phase separation was performed on a Nexera X2R UHPLC (Shimadzu, Nakagyō-ku, Japan) coupled to a hybrid quadrupole-Orbitrap high resolution mass spectrometer (Q-Exactive HF, Thermo Scientific, Bremen, Germany). Compounds were separated on a Poroshell 120 EC-C18 (Agilent Technologies, California, USA) column (150 × 2.1 mm) packed with 2.7 µm particles at a constant 300 µL/min flow rate during a 20 min binary gradient. 0.2 % FA in $H_2O$ and 0.2 % FA in ACN were used as mobile phases A and B, respectively. Gradient was set as follows : 0-1 min isocratic 2 % B, 1-17 min linear increase to 90 % B, 17-18 isocratic 90 % B, 18-18.5 min linear decrease to 2 % B, 18.5-20 min isocratic 2 % B.

## Mass spectrometry

Liquid chromatography was connected to the mass spectrometer *via* a heated-electrospray (H-ESI II, Thermo Scientific) probe. Voltage was set to 3 kV for positive and negative ionization modes. Capillary and probe heater temperatures were set to 320 and 350 °C, respectively. Sheath and auxiliary gases were set to 40 and 20, respectively. S-lens RF level was set to 50.

**LGD-4033 datasets**

For data-independent acquisition (DIA) experiment, the method was split into four scan events (two positive and two negative) composed of a full scan MS acquisition within 250-650 $m/z$ range, automatic gain control (AGC Target) set to 3e6 and maximum accumulation time (MaxIT) set to 20 ms and a DIA loop. The DIA scans were defined to select the com-

plete 250-647 *m/z* range and fragment it by 9 successive isolations of all precursor ions within a 45 *m/z* window, windows being overlapped by 1 *m/z* (1st window: 250-295, 2nd: 294-339 and so on until 9th: 602-647). Selected ions within the *m/z*-ranges were fragmented in the HCD cell with 20, 35, and 50 as stepped normalized collision energy values (NCE), AGC Target was set to 1e6 and MaxIT to 23 ms. Data were recorded in profile mode with mass resolutions set to 60 000 and 15 000 at *m/z* 200 for MS and MS/MS acquisition, respectively.

For data-dependent acquisition (DDA) experiment, method consisted into two Full MS-ddMS$^2$ scan events (one in positive and one in negative mode). Full MS parameters were identical to the DIA method. DDA-MS/MS scans were set to select the ten most abundant precursor ions of the full MS scan within a quadrupole isolation of 1.2 *m/z*, AGC Target was set to 1e5, NCE and MaxIT were identical to the DIA method. Dynamic exclusion was set to 5 s. Mass resolution were set to 60 000 and 15 000 at *m/z* 200 for MS and MS/MS acquisitions, respectively.

**Cocaine datasets**

Data were recorded in DIA positive mode. Method consisted in a Full scan MS within 100-600 *m/z* range, AGC Target was set to 3e6 and MaxIT to 20 ms and a DIA loop. The DIA scans were defined to select the complete 100-581 *m/z* range and fragment it by 20 successive isolations of all precursors within a 25 *m/z* window, windows being overlapped by 1 *m/z* (1st window: 100-125, 2nd window: 124-149 and so on until 20th: 556-581). Selected *m/z*-ranges were fragmented in the HCD cell with 20, 40, and 60 as stepped NCE value, AGC Target was set to 1e6 and MaxIT to 23 ms. Data were recorded in profile mode with mass resolutions set to 60 000 and 15 000 at *m/z* 200 for MS and MS/MS acquisition, respectively.

Data were converted to the mzML universal format[36] file using ProteoWizard MSconvert tool (version 3.0.20090)[37] using profile mode default configuration, except in case of peak-

picked data analysis, where peak picking of MS and MS/MS data was enabled.

## Chemoinformatics

MetIDfyR allows the prediction of metabolites in LC-HRMS/MS samples based on the Drug of interest (DoI) elemental composition (EC) and few parameters. The software is divided in few steps (Figure 1) and involves multiple R packages, notably Rdisop,[38] MSnBase[39] and mzR[40] to compute spectral properties and mzML file mining (see supplementary information for complete list and versions). To optimize throughput, MetIDfyR can parallelize *in-silico* transformations and MS searches, thanks to the doParallel package.

Based on the DoI EC and expected adducts in both ESI polarities, a list of possible metabolites is generated (Figure 1 A). The number of successive transformations ($N$ successive loops of transformations) can be set to consider simple ($N = 1$) or complex ($N \geq 2$) transformations, listed in Table 1. Transformation list can be edited to fit user's needs. The predicted metabolites' singly charged $m/z$ are calculated and theoretical isotope patterns determined using Rdisop.

Once the complete list of unique ECs is generated, chromatograms of each predicted $m/z$ are automatically extracted from the mzML file (Figure 1 B). If signal above a defined intensity cut-off value is detected, peak detection is performed by searching local maxima, generating $m/z$-retention-time (mz-RT) pair(s). Secondly, spectral properties of each mz-RT are assessed to evaluate metabolite detection. At MS level, an intensity conditional score is calculated ($R_{Intensity}$, Supplementary Equation 1) and the isotopic distribution of the selected mz-RT is compared to theoretical distribution through the calculation of the absolute isotopic deviation ($iAScore$).

At MS/MS level, DIA or DDA MS/MS spectra of each mz-RT are compared with parent

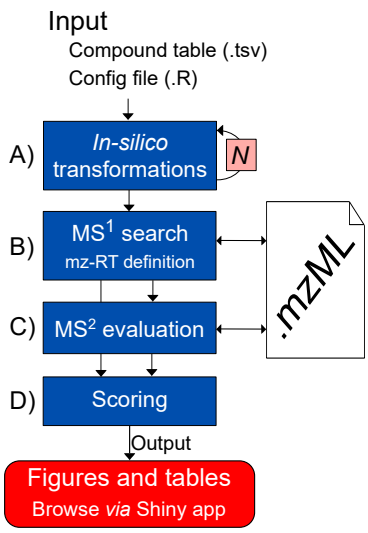


Figure 1: Schematic representation of the MetIDfyR analysis pipeline. $N$ refers to the number of successive transformation loops.

drug spectra (provided externally or automatically picked in the chromatogram) considering common and/or shifted $m/z$ ($MS2p$) (Figure 1 C). $MS2p$ is calculated as to a $MS2p$ of 100 % means all DoI fragments are found as common and/or shifted in the mz-RT MS/MS spectra. On the other hand, a $MS2p$ of 0 % means none of the DoI fragments were found as common or shifted in the mz-RT spectra. Additionally, a composite $MS2cos\theta$ [41] is computed comparing DoI spectra with the metabolite spectra where shifted $m/z$ are unshifted (see Result section for case example). To enable MS/MS evaluation when the DoI is not detected (*i.e.* complete metabolization), it is possible to incorporate an external DoI MS/MS spectra as a tabulation separated table with positive and/or negative fragments $m/z$ and their relative intensities, which is similar to identification software using a spectral database. Multiple drug/metabolite searches in a single file is also possible incorporating multiple compounds, ECs, and external spectra (optional) linked to a single mzML file into a compound table (Figure 1).

Finally, a score to evaluate each mz-RT is calculated (Figure 1 D), based on MS ($R_{Intensity}$ and $iAScore$) and MS/MS ($MS2p$ and $MS2cos\theta$) components (Equation 1). However, the

Table 1: Summary of modifications tested by MetIDfyR.
* Loss of halogens is only tested if the molecule to be modified contain a halogen atom.

| **Name of modification** | **Raw formula modification** | **Mass change (a.m.u.)** | **Phase** |
|---|---|---|---|
| Hydroxylation | +O | 15.99491 | I |
| Deoxydation | -O | -15.99491 | I |
| Carbonyl loss | -CO | -27.99491 | I |
| Oxidative deamination | -$NH_2$ +OH | -0.98402 | I |
| Defluorination* | -HF | -20.00623 | I |
| Dechlorination* | -HCl | -35.97668 | I |
| Debromination* | -HBr | -79.92616 | I |
| Dephenylation | -$C_6H_4$ | -77.03913 | I |
| Demethylation | -$CH_2$ | -14.01565 | I |
| Methylation | +$CH_2$ | 14.01565 | I |
| Acetylation | +$C_2OH_2$ | 42.01056 | I |
| Reduction | +$H_2$ | 2.01565 | I |
| Hydration | +$H_2O$ | 18.01056 | I |
| Dehydration | -$H_2O$ | -18.01056 | I |
| Decarboxylation | -$CO_2$ | -43.98983 | I |
| Oxydation of C-C bond | -$H_2$ | -2.01565 | I |
| Glucuronylation | +$C_6H_8O_6$ | 172.0008 | II |
| Sulfation | +$SO_3$ | 96.95955 | II |

search space exponentially expands with increasing values of $N$, yielding in a higher probability of false positive hit. To counter this effect and adjust each candidate metabolite score according to the search space it has been predicted from, MS and MS/MS score components are adjusted by $1/\sqrt{N}$.

$$Score = \frac{1}{\sqrt{N}}\left(\frac{1}{2}\cdot R_{Intensity}\cdot iAScore \; + \; \frac{1}{2}\cdot MS^2 cos\theta \cdot \frac{MS2p}{100}\right) \tag{1}$$

As an output, MetIDfyR generates paneled figures using ggplot2,[42] for each mz-RT pair above an intensity threshold defined prior to analysis. The figure includes the extracted ion chromatogram, the mz-RT mass spectrum with experimental-theoretical isotopic pattern comparison and a composite MS/MS spectra where top spectra and bottom spectra refer to DoI and mz-RT MS/MS spectra, respectively. Finally, results are compiled within an analysis summary table, support of the web interface visualisation tool powered by Shiny[35] where candidate browsing is possible and enables report generations.

## Computer configurations

The aim of MetIDfyR is to enable cross-platform identification of drug metabolites on any computer configuration running any operating system compatible with the R environment. In this context, our strategy was to evaluate MetIDfyR on a calculation server but also a conventional computer with different hardwares and running with different operating systems.

### Calculation server

MetIDfyR ran on a Debian-based (bullseye/sid) HPE ProLiant DL380 Gen10 (Hewlett Packard Enterprise) server equipped with a 12-core Intel Xeon-Silver 4214 processor (2.2 GHz), 48 GB memory and with R version 3.6.2 installed. MetIDfyR analyses were launched as bash Rscript commands.

### Regular laptop

To evaluate MetIDfyR capabilities on a broadly available configuration and for speed comparisons, MetIDfyR was tested on a Windows 10 HP 250 G6 Probook (Hewlett Packard) regular laptop equipped with a 4-core Intel core i5-8265 processor (1.8 GHz), 8 GB memory with R version 3.6.3 and RStudio[43] installed. MetIDfyR searches were launched *via* RStudio Rscripts commands.

# Results and discussion

Untargeted metabolite discovery relies on the detection of a drug of interest (DoI) which potentially underwent one or successive biochemical modifications, most of the time increasing its hydrophilic properties. While these modifications are most frequently predictable, manual hypothesis testing on the detection of each of these modifications in a LC-HRMS file is a repetitive, time-consuming and an error-prone process. On the other hand, the pre-

diction and automatic determination of mass spectrometric properties of one compound's putative metabolites is an informatically automatable process which already exists in some dedicated proprietary or free softwares, with few specific examples in the open-source R mass-spectrometry tools and packages. In this project, we intended to develop an efficient and fast tool to simplify drug metabolism studies, allowing scientists to evaluate drug-metabolite candidates easily, without the need of extensive informatics resources, mass-spectrometric expertise or proprietary software.

The aim of MetIDfyR is to provide an automated approach performing each individual step which an analyst would realize to detect biotransformations of a DoI. This includes : (1) calculation of transformations (and their combinations) related mass-shifts ; (2) evaluation of isotopic profile in case of signal detection ; (3) evaluation of MS/MS spectra and comparison with DoI ; (4) putative metabolite ranking based on previous steps results. Additionally, the tool had to provide confidence into analysis through the high number of transformations tested, the absence of manual intervention, observation of known metabolites in model experiments and be significantly faster than manual processing. Finally, the tool would include a "code-free" user interface to allow most analytical scientists to perform straightforward drug metabolism investigations.

## Calculation performances

As shown on Figure 2 obtained after analysis of the *in-vitro* phase I metabolism LGD-4033 datasets with various configurations, MetIDfyR displays rapid molecule-testing properties. Indeed, analysis speed varies between 15 molecules/min ($N = 1$ with single core search in DDA) to 400 molecules/min ($N = 5$ with ten cores search in DIA) on a calculation-dedicated computer in peak-picked mode (Figure 2, left), where $N$ is the number of successive transformation loops. Because profile data structure is considerably more complex, MetIDfyR offers slower testing capabilities, with speeds varying between 10 molecules/min

to 150 molecules/min (Figure 2, right). As expected, speed is increasing with higher core number for transformations combinations $N \geq 2$, whereas it remains limited for $N = 1$. This later observation may be a consequence of the limited number of predicted molecules (86 molecules tested for $N = 1$), probably not high enough to reach maximum analysis speed. Thus, both peak-picked and profile data analyses are considered as high-throughput when employing multiple core strategies.

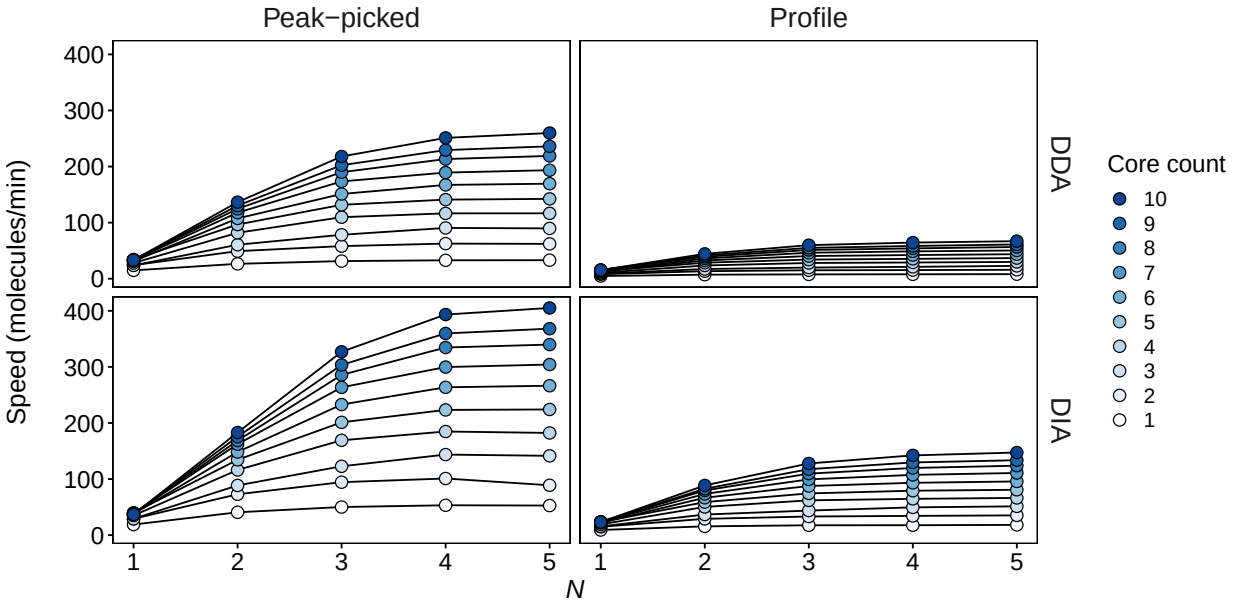


Figure 2: MetIDfyR calculation speed on the LGD-4033 DIA and DDA datasets. Line and point shapes refer to number of modifications ($N$).

Interestingly, even if the DIA dataset displays a greater spectrum count than its DDA equivalent in the LGD-4033 datasets ($> 30$ %), analysis of a DIA dataset is significantly faster than DDA (Figure 2). This observation may be a consequence of the complex precursor selection process occurring in DDA mode, yielding in an extensive high number of different MS/MS scan headers in such data. In addition, precursor selection in DIA approaches is not driven by the relative abundance of precursor ions, guaranteeing complete fragmentation in a specified $m/z$ range. As a consequence, DIA-MS/MS spectra contain fragment ions of co-eluted and co-selected precursors, leading to the observation of fragment

ions unrelated to DoI or metabolites. These aspects may have an impact on MetIDfyR scoring capabilities. However, this phenomenon did not greatly affect later results in tested conditions. Similarly, DIA approaches are more subject to MS/MS ion-suppression effects, in case of co-elution and co-fragmentation of a highly abundant unrelated compound. These properties should be kept in mind of analysts when reviewing the results.

To demonstrate compatibility and assess calculation speeds on a common computer, we analyzed the profile-mode LGD-4033 DIA dataset on a laptop with core count of 3 and 4 and compared speeds with the calculation server in similar conditions. First, analogous speed behaviour was observed as to increases with $N$ and core numbers with both configurations. Laptop speed reaches maximal values for $N = 3$ & 4, which may be linked to laptop hardware (Figure 3) and be explained by memory and/or core overload when using 100 % of available cores (Figure 3). More generally, analysis speed was significantly higher when analysis ran on a calculation server (Wilcoxon rank sum test p-value = 0.0104), as expected. However, considering the analysis speed reached by the Laptop setup (Figure 3, dashed line) with maximum speed being close to 20 molecules tested/min, it is convincing that the molecule testing speed of MetIDfyR surpass manual processing. Thus, even if a calculation-dedicated computer provides higher testing speed, MetIDfyR does not require high-end calculation capabilities and may be run on a regular computer.

## Metabolites identification

### *In-vitro* generated metabolites of LGD-4033

To assess the package ability to identify relevant metabolites, we performed the metabolite identification of compounds undergoing multiple modifications during metabolism such as LGD-4033. In addition, the metabolism of LGD-4033 has already been extensively studied and determined by multiple research groups in human and horse,[17,44,45] which makes it an adequate model compound (Supplementary Figure 1).

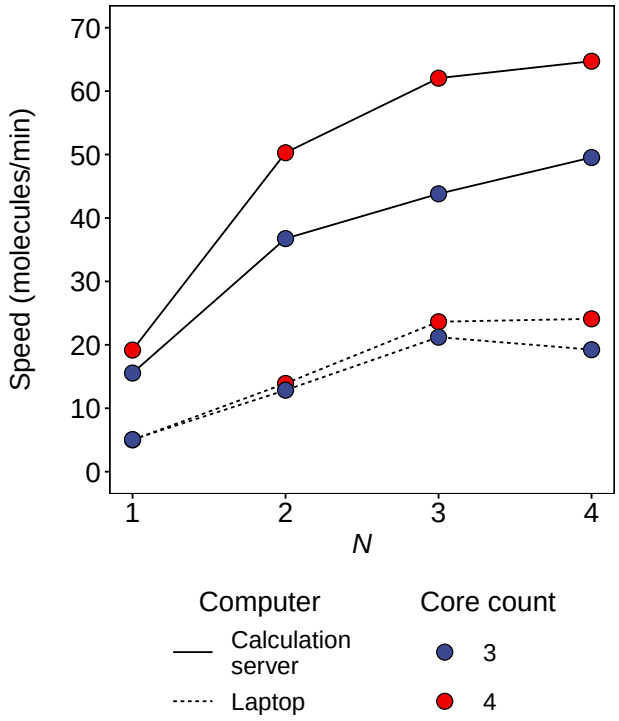


Figure 3: MetIDfyR analysis speed comparison between a calculation dedicated server and a common laptop. Analyses were run on the LGD-DIA profile mode dataset.

After Phase I *in-vitro* experiments with horse liver S9 fractions followed by LC-HRMS/MS experiments, DIA data files were subjected to MetIDfyR for metabolite identification. As shown in Figure 4, MetIDfyR could identify all six known Phase I metabolites referenced in the literature and their epimers or position isomers when the number of modifications ($N$) was equal or greater than three. Additionally, MetIDFyR could identify the [LGD-4033 +HCOO]$^-$ adduct as well as in-source fragments [LGD-4033 -HF +H]$^+$ and [LGD-4033 -HF +$H_2O$]$^+$ as these ions were predicted by the software during *in-silico* transformation, enabling their detection. Even though these findings are not directly related to metabolite detection, this property may be useful to refine related ions for later analyses as it is generally observed in usual LC-MS detection strategies. Additionally, these findings exhibited a possible epimerization of LGD-4033 during Phase I metabolism as two additional negatively charged [LGD-4033 -H]$^-$ and [LGD-4033 +HCOO]$^-$ species were detected (Figure 4, LGD-

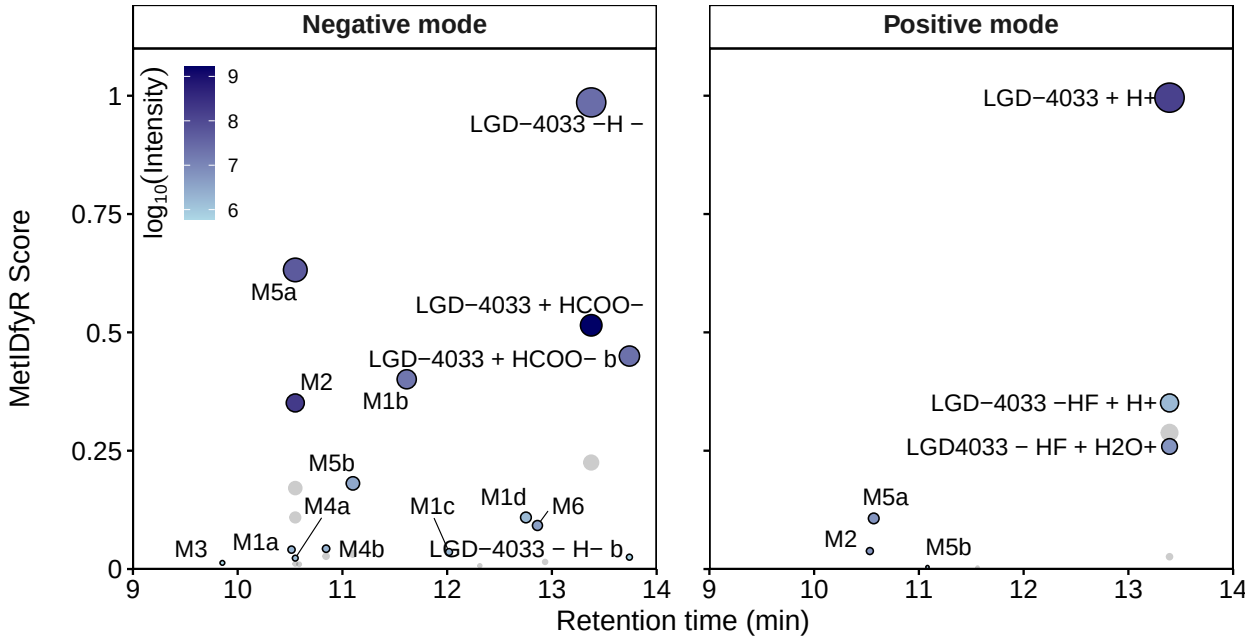


Figure 4: Scatter MetIDfyR score plot of LGD-4033 and its *in-vitro* generated metabolites detected by MetIDfyR, where each point is related to a candidate mz-RT metabolite and point size is related to its MetIDfyR score (see Supplementary Figure 1 for structures according to Culter *et al.* 2019[17]). Colour scale of identified metabolites is relative to MS intensity.

4033b and LGD-4033 +HCOO-b, respectively).

The scoring approach implemented in MetIDfyR successfully ranked best mz-RT found in the dataset as top candidate metabolites as shown on Figure 4 where larger points indicate higher scores determined by the software. However, even if the software succeeded in finding *in-vitro* generated metabolites in the LGD-4033 datasets, the data is considered as moderately complex.

### Drug of abuse metabolite identification in real-case human urine extract

To fully validate that MetIDfyR would successfully identify most prevalent metabolites in a complex matrix, metabolite search was performed in a human urine extract that had been previously tested positive for Cocaine, Benzoylecgonine and Ecgonine methyl ester using

an accredited method and certified reference materials. After analysis, MetIDfyR succefully identified Cocaine, Benzoylecgnonine (demethylation of Cocaine) and Ecgonine methyl ester (dephenylation and carbonyl loss of Cocaine) in the urine extract (Figure 5 A), in accordance with prior analysis. Remarkably, MetIDfyR also identified Cocaethylene (transesterification of Cocaine by Ethanol) a well-documented co-metabolite marker of alcohol consumption associated with Cocaine (Figure 5 A & B) and a Cocaine isomer which could be Pseudococaine/Isococaine. This finding is of major importance since it demonstrates that MetIDfyR can highlight unforeseen metabolites. Indeed, MetIDfyR revealed that Cocaine was convincingly associated with ethanol which is highly relevant in a forensic context. In addition, these five compounds were ranked as top 5 candidate metabolites, confirming the relevance of the comprehensive scoring approach of MetIDfyR (Eq. 1). Additionally, deeper inspection of the candidates allowed the identification of two supplementary metabolites: Ecgonine and para or meta-Hydroxybenzoylecgonine ranked 52 and 136 out of 962 candidates (Supplementary Table). However, their low intensities explain their respective low scores. As a conclusion, cocaine, one cocaine isomer and five cocaine metabolites were detected by MetIDfyR, reaching high completeness of metabolite characterization, considering the most prevalent cocaine metabolites (Supplementary Figure 2) and the high complexity of the extract.

Overall, the experiments presented here describe effectively the software capability. Indeed, MetIDfyR is intended to highlight top candidates out of all possible candidates present within LC-HRMS/MS data. Analyses of *in-vitro* generated LGD-4033 and *in-vivo* human urine cocaine metabolites successfully highlighted referenced metabolites of both compounds.

## Result browsing and user interface

Thanks to the Shiny user interface implemented in MetIDfyR, candidate manual inspection is straightforward and user-friendly. Indeed, based on the summary table output (Figure 1),

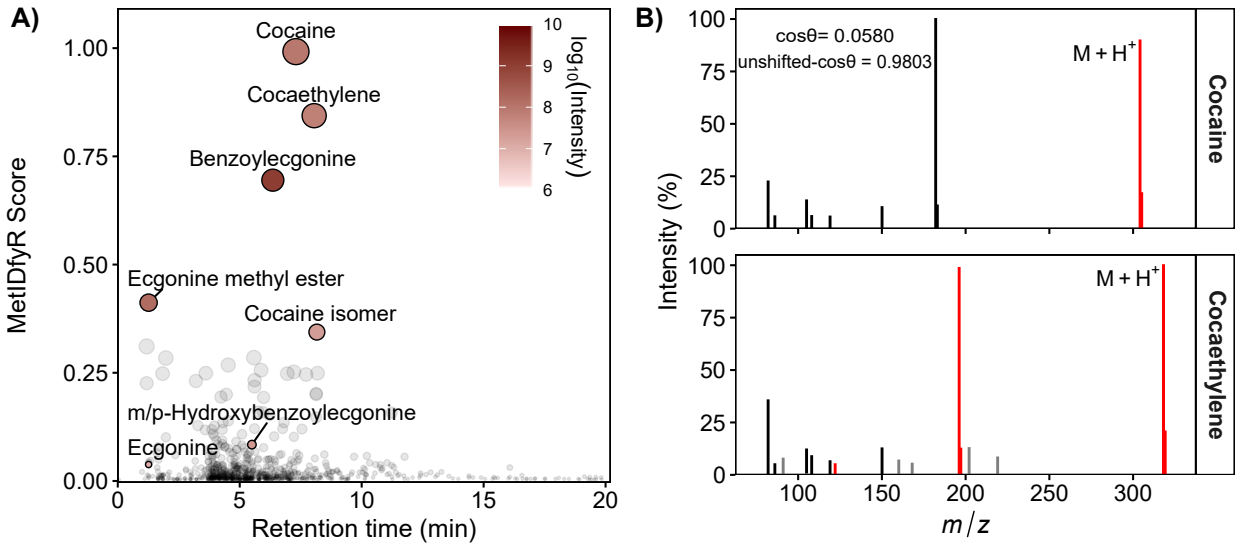


Figure 5: A. Scatter score plot of Cocaine and its metabolites identified in a positive case urine extract by MetIDfyR. Each point is related to a candidate mz-RT metabolite and point size is related to its MetIDfyR score (see Supplementary Figure 2 for structures adapted from Huestis *et al.*, 2007[46]). Colour scale of identified metabolites is relative to MS intensity. B. MS/MS spectra of Cocaine (top) and Cocaethylene (bottom) identified by MetIDfyR where $+CH_2$ shifted peaks are marked red.

analysts can review detected compounds based on their scores, but also each compound's figure and fragment table. Once reviewing the data, the analyst may export the results as a metabolite PDF report.

# Conclusion and outlook

In this article, we presented MetIDfyR a free, cross-platform and open-source R-based analysis pipeline to assist analysts during *in-vitro* and/or *in-vivo* drug metabolism studies. The tool is designed to analyze positive and/or negative LC-HRMS/MS data from standard file format mzML, enabling cross-manufacturer analysis. MetIDfyR has proven metabolite identification capabilities of two model compounds in semi-complex matrix for horse *in-vitro* S9 fraction metabolism, but also complex matrix in a positive case of drug abuse in human urine. These results demonstrate its efficiency towards comprehensive LC-HRMS/MS analysis of drug biotransformations. The candidate ranking process is based on intensity,

isotopic distribution and MS/MS spectral properties *via* spectral comparison with drug of interest and has shown efficiency to identify top candidate metabolites in DDA/DIA-based LC-HRMS/MS datasets. MetIDfyR may be run on calculation servers or conventional computers running operating systems compatible with R, with testing speeds reaching hundreds of transformations assessed per minute in LC-HRMS/MS data. MetIDfyR comes with a built-in Shiny app, allowing the user to perform each steps of the analysis pipeline (*i.e.* configuration, launching and candidate metabolites review).

While the tool allows straightforward use with possible high-throughput analysis, future improvements for metabolite predictions will be focused on adducts and charge detection to ensure maximal metabolite detection and on structural analysis, with the implementation of simplified molecular-input line-entry system (SMILES[47]) to enable structure-specific modification prediction such as acyl chain hydrolysis.

# Supporting Information Available

The following files are available free of charge.

- Supplementary information.pdf: Additional information relative to MetIDfyR package dependencies and $R_{Intensity}$ calculation description.
- Supplementary data: MetIDfyR output table obtained after analysis of LGD-DIA and Cocaine profile datasets.
- Code and example datasets: Code and peak-picked LGD-4033 and Cocaine example datasets are freely available at github.com/agnesblch/MetIDfyR

# References


(1) Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting drug metabolism: experiment and/or computation? *Nature reviews Drug discovery* **2015**, *14*, 387–404.

(2) T Issa, N.; Wathieu, H.; Ojo, A.; W Byers, S.; Dakshanamurthy, S. Drug metabolism in preclinical drug development: a survey of the discovery process, toxicology, and computational tools. *Current drug metabolism* **2017**, *18*, 556–565.

(3) Zhang, Z.; Tang, W. Drug metabolism in drug discovery and development. *Acta Pharmaceutica Sinica B* **2018**, *8*, 721–732.

(4) Toutain, P.-L.; Reymond, N.; Laroute, V.; Garcia, P.; Popot, M.-A.; Bonnaire, Y.; Hirsch, A.; Narbe, R. Pharmacokinetics of meloxicam in plasma and urine of horses. *American journal of veterinary research* **2004**, *65*, 1542–1547.

(5) Popot, M.-A.; Jacobs, M.; Garcia, P.; Loup, B.; Guyonnet, J.; Toutain, P.; Bailly-Chouriberry, L.; Bonnaire, Y. Pharmacokinetics of tiludronate in horses: A field population study. *Equine veterinary journal* **2018**, *50*, 488–492.

(6) Hess, C.; Brockmann, C.; Doberentz, E.; Madea, B.; Musshoff, F. Unintentional lethal overdose with metildigoxin in a 36-week-old infant–post mortem tissue distribution of metildigoxin and its metabolites by liquid chromatography tandem mass spectrometry. *Forensic science international* **2014**, *241*, e23–e27.

(7) Sundström, M.; Pelander, A.; Ojanperä, I. Comparison between drug screening by immunoassay and ultra-high performance liquid chromatography/high-resolution time-of-flight mass spectrometry in post-mortem urine. *Drug testing and analysis* **2015**, *7*, 420–427.

(8) Tynon, M.; Homan, J.; Kacinko, S.; Ervin, A.; McMullin, M.; Logan, B. K. Rapid and sensitive screening and confirmation of thirty-four aminocarbonyl/carboxamide (NACA) and arylindole synthetic cannabinoid drugs in human whole blood. *Drug testing and analysis* **2017**, *9*, 924–934.

(9) Partridge, E.; Teoh, E.; Nash, C.; Scott, T.; Charlwood, C.; Kostakis, C. The Increasing Use and Abuse of Tapentadol and Its Incorporation Into a Validated Quantitative Method. *Journal of analytical toxicology* **2018**, *42*, 485–490.

(10) Gaunitz, F.; Lehmann, S.; Thomas, A.; Thevis, M.; Rothschild, M. A.; Mercer-Chalmers-Bender, K. Post-mortem distribution of the synthetic cannabinoid MDMB-CHMICA and its metabolites in a case of combined drug intoxication. *International journal of legal medicine* **2018**, *132*, 1645–1657.

(11) Moulard, Y.; Bailly-Chouriberry, L.; Boyer, S.; Garcia, P.; Popot, M.-A.; Bonnaire, Y. Use of benchtop exactive high resolution and high mass accuracy orbitrap mass spectrometer for screening in horse doping control. *Analytica chimica acta* **2011**, *700*, 126–136.

(12) Scarth, J. P.; Teale, P.; Kuuranne, T. Drug metabolism in the horse: a review. *Drug Testing and Analysis* **2011**, *3*, 19–53.

(13) Hansson, A.; Knych, H.; Stanley, S.; Thevis, M.; Bondesson, U.; Hedeland, M. Characterization of equine urinary metabolites of selective androgen receptor modulators (SARMs) S1, S4 and S22 for doping control purposes. *Drug testing and analysis* **2015**, *7*, 673–683.

(14) Decloedt, A.; Bailly-Chouriberry, L.; Vanden Bussche, J.; Garcia, P.; Popot, M.-A.; Bonnaire, Y.; Vanhaecke, L. Mouldy feed: A possible explanation for the excretion of anabolic-androgenic steroids in horses. *Drug testing and analysis* **2016**, *8*, 525–534.

(15) Wong, J. K.; Chan, G. H.; Leung, D. K.; Tang, F. P.; Wan, T. S. Generation of phase II in vitro metabolites using homogenized horse liver. *Drug testing and analysis* **2016**, *8*, 241–247.

(16) Wong, J.; Choi, T.; Kwok, K.; Lei, E.; Wan, T. Doping control analysis of 121 prohibited substances in equine hair by liquid chromatography–tandem mass spectrometry. *Journal of pharmaceutical and biomedical analysis* **2018**, *158*, 189–203.

(17) Cutler, C.; Viljanto, M.; Hincks, P.; Habershon-Butcher, J.; Muir, T.; Biddle, S. Investigation of the metabolism of the selective androgen receptor modulator LGD-4033 in equine urine, plasma and hair following oral administration. *Drug testing and analysis* **2019**,

(18) Bachmann, C.; Bickel, M. History of drug metabolism: the first half of the 20th century. *Drug metabolism reviews* **1985**, *16*, 185–253.

(19) Brodie, B. B.; Gillette, J. R.; La Du, B. N. Enzymatic metabolism of drugs and other foreign compounds. *Annual review of biochemistry* **1958**, *27*, 427–454.

(20) Tucker, G. Drug metabolism. *British journal of anaesthesia* **1979**, *51*, 603–618.

(21) Guengerich, F. P. Cytochrome P450s and other enzymes in drug metabolism and toxicity. *The AAPS journal* **2006**, *8*, E101–E111.

(22) Guengerich, F. P. Cytochrome p450 and chemical toxicology. *Chemical research in toxicology* **2007**, *21*, 70–83.

(23) Lindon, J. C.; Nicholson, J. K.; Sidelmann, U. G.; Wilson, I. D. Directly coupled HPLC-NMR and its application to drug metabolism. *Drug metabolism reviews* **1997**, *29*, 705–746.

(24) Wen, B.; Zhu, M. Applications of mass spectrometry in drug metabolism: 50 years of progress. *Drug metabolism reviews* **2015**, *47*, 71–87.

(25) Youdim, K. A.; Saunders, K. C. A review of LC–MS techniques and high-throughput approaches used to investigate drug metabolism by cytochrome P450s. *Journal of Chromatography B* **2010**, *878*, 1326–1336.

(26) Jacobs, P. L.; Ridder, L.; Ruijken, M.; Rosing, H.; Jager, N. G.; Beijnen, J. H.; Bas, R. R.; van Dongen, W. D. Identification of drug metabolites in human plasma or serum integrating metabolite prediction, LC–HRMS and untargeted data processing. *Bioanalysis* **2013**, *5*, 2115–2128.

(27) Maurer, H.; Meyer, M. R. High-resolution mass spectrometry in toxicology: current status and future perspectives. *Archives of toxicology* **2016**, *90*, 2161–2172.

(28) Parikh, A.; Gillam, E. M.; Guengerich, F. P. Drug metabolism by Escherichia coli expressing human cytochromes P450. *Nature biotechnology* **1997**, *15*, 784.

(29) Flasch, M.; Bueschl, C.; Woelflingseder, L.; Schwartz-Zimmermann, H. E.; Adam, G.; Schuhmacher, R.; Marko, D.; Warth, B. Stable isotope-assisted metabolomics for deciphering xenobiotic metabolism in mammalian cell culture. *ACS Chemical Biology* **2020**,

(30) Peters, F. T.; Meyer, M. R. In vitro approaches to studying the metabolism of new psychoactive compounds. *Drug testing and analysis* **2011**, *3*, 483–495.

(31) Lohmann, W.; Karst, U. Generation and identification of reactive metabolites by electrochemistry and immobilized enzymes coupled on-line to liquid chromatography/mass spectrometry. *Analytical chemistry* **2007**, *79*, 6831–6839.

(32) Fangmeyer, J.; Scheeren, S.; Schmid, R.; Karst, U. Fast Online Separation and Identification of Electrochemically Generated Isomeric Oxidation Products by Trapped Ion Mobility-Mass Spectrometry. *Analytical Chemistry* **2019**,

(33) Ihaka, R.; Gentleman, R. R: a language for data analysis and graphics. *Journal of computational and graphical statistics* **1996**, *5*, 299–314.

(34) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2019.

(35) RStudio Inc, Easy web applications in R. 2013; URL: `http://www.rstudio.com/shiny/`.

(36) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shofstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D., et al. mzML—a community standard for mass spectrometry data. *Molecular & Cellular Proteomics* **2011**, *10*.

(37) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. ProteoWizard: open source software for rapid proteomics tools development. *Bioinformatics* **2008**, *24*, 2534–2536.

(38) Böcker, S.; Letzel, M. C.; Lipták, Z.; Pervukhin, A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* **2008**, *25*, 218–224.

(39) Gatto, L.; Lilley, K. S. MSnbase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics* **2011**, *28*, 288–289.

(40) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J., et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature biotechnology* **2012**, *30*, 918.

(41) Wan, K. X.; Vidavsky, I.; Gross, M. L. Comparing similar spectra: from similarity index to spectral contrast angle. *Journal of the American Society for Mass Spectrometry* **2002**, *13*, 85–88.

(42) Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer-Verlag New York, 2016.

(43) RStudio Team, RStudio: Integrated Development Environment for R. RStudio, Inc.: Boston, MA, 2015.

(44) Thevis, M.; Lagojda, A.; Kuehne, D.; Thomas, A.; Dib, J.; Hansson, A.; Hedeland, M.; Bondesson, U.; Wigger, T.; Karst, U., et al. Characterization of a non-approved selective androgen receptor modulator drug candidate sold via the Internet and identification of in vitro generated phase-I metabolites for human sports drug testing. *Rapid Communications in Mass Spectrometry* **2015**, *29*, 991–999.

(45) Hansson, A.; Knych, H.; Stanley, S.; Berndtson, E.; Jackson, L.; Bondesson, U.; Thevis, M.; Hedeland, M. Equine in vivo-derived metabolites of the SARM LGD-4033 and comparison with human and fungal metabolites. *Journal of Chromatography B* **2018**, *1074*, 91–98.

(46) Huestis, M. A.; Darwin, W. D.; Shimomura, E.; Lalani, S. A.; Trinidad, D. V.; Jenkins, A. J.; Cone, E. J.; Jacobs, A. J.; Smith, M. L.; Paul, B. D. Cocaine and metabolites urinary excretion after controlled smoked administration. *Journal of analytical toxicology* **2007**, *31*, 462–468.

(47) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.