Analytica Chimica Acta 1070 (2019) 29–42

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

Targeting bioactive compounds in natural extracts - Development of a comprehensive workflow combining chemical and biological data

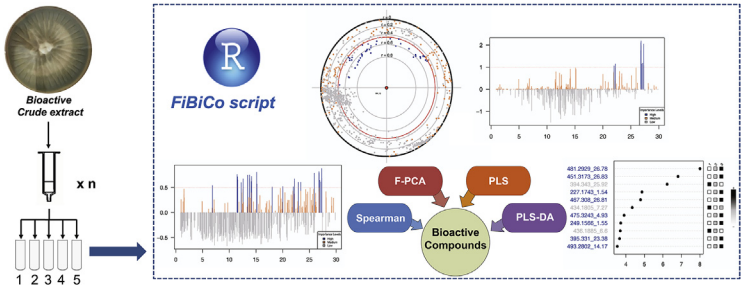
Lucie Ory [a], El-Hassane Nazih [a], Sahar Daoud [a], Julia Mocquard [a], Mélanie Bourjot [b], Laure Margueritte [c], Marc-André Delsuc [d], Jean-Marie Bard [a], Yves François Pouchus [a], Samuel Bertrand [a, e], Catherine Roullier [a, e, *]

[a] Université de Nantes, Mer Molécules Santé, MMS EA 2160, F-44000, Nantes, France
[b] UMR 7178 CNRS, Institut Pluridisciplinaire Hubert Curien, Faculté de Pharmacie, Université de Strasbourg, F-67401, Illkirch, France
[c] UMR 7200 CNRS, Laboratoire d'Innovation Thérapeutique, Faculté de Pharmacie, Université de Strasbourg, F-67401, Illkirch, France
[d] INSERM U596, UMR 7104 CNRS, Institut de Génétique et de Biologie Moléculaire et Cellulaire, Université de Strasbourg, F-67401, Illkirch, France
[e] Biogenouest, Université de Nantes, Corsaire-ThalassOMICS, F-44000, Nantes, France

HIGHLIGHTS

- Fractionation coupled to biochemometrics is an interesting strategy for drug discovery.
- The workflow proposed provides a new tool for the detection of bioactive compounds.
- R-FiBiCo script proved to be efficient in detecting bioactive compounds from a mixture.
- Performance evaluation of the script showed its ability to detect minor compounds.
- Mass spectrometry- and NMR-based biochemometrics approaches can be complementary.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 20 December 2018
Received in revised form
19 March 2019
Accepted 18 April 2019
Available online 23 April 2019

Keywords:
Metabolomics
Biochemometrics
Natural products
Liquid chromatography
Mass spectrometry
R script

ABSTRACT

In natural product drug discovery, several strategies have emerged to highlight specifically bioactive compound(s) within complex mixtures (fractions or crude extracts) using metabolomics tools. In this area, a great deal of interest has raised among the scientific community on strategies to link chemical profiles and associated biological data, leading to the new field called "biochemometrics". This article falls into this emerging research by proposing a complete workflow, which was divided into three major steps. The first one consists in the fractionation of the same extract using four different chromatographic stationary phases and appropriated elution conditions to obtain five fractions for each column. The second step corresponds to the acquisition of chemical profiles using HPLC-HRMS analysis, and the biological evaluation of each fraction. The last step evaluates the links between the relative abundances of molecules present in fractions (peak area) and the global bioactivity level observed for each fraction. To this purpose, an original bioinformatics script (encoded with R Studio software) using the combination of four statistical models (Spearman, F-PCA, PLS, PLS-DA) was here developed leading to the generation of a "Super list" of potential bioactive compounds together with a predictive score. This strategy was validated by its application on a marine-derived *Penicillium chrysogenum* extract exhibiting

* Corresponding author. Université de Nantes, Faculté de Pharmacie, MMS, BP 61112, F-44035, Nantes, France.
E-mail address: catherine.roullier@univ-nantes.fr (C. Roullier).

https://doi.org/10.1016/j.aca.2019.04.038

antiproliferative activity on breast cancer cells (MCF-7 cells). After the three steps of the workflow, one main compound was highlighted as responsible for the bioactivity and identified as ergosterol. Its antiproliferative activity was confirmed with an IC_{50} of 0.10 μM on MCF-7 cells. The script efficiency was further demonstrated by comparing the results obtained with a different recently described approach based on NMR profiling and by virtually modifying the data to evaluate the computational tool behaviour. This approach represents a new and efficient tool to tackle some of the bottlenecks in natural product drug discovery programs.

1. Introduction

In Natural Product (NP) research, bioassay-guided fractionation approach is mostly used to isolate bioactive metabolites from a crude extract [1–3]. Despite the fact that this method has proved to be efficient for the discovery of many active compounds including taxol, artemisinin or vinblastine [4], it is now often considered by industrials as a time-consuming, costly and risky investment [5]. In some cases, the activity originally observed on a mixture can be lost due to irreversible binding of the components to chromatographic resins, degradation, chemical modification, antagonistic or synergistic effects. Therefore, the need to improve productivity and efficiency in the discovery of new bioactive NPs to address limitations of bioassay-guided fractionation has resulted in the past few years in the emergence of several novel strategies.

Even if such strategies are considered as recent, initial strategies were started in the 80's. Pouchus et *al.* [6], Samuelsson et al. [7] and Cardellina et al. [8], were among the first people interested in improving the classical bioassay-guided fractionation method for active NP extracts. In 1989, Pouchus developed a mathematical script allowing a better purification of active compounds by calculating their relative quantities in each fraction. The program was also able to detect potential synergistic effects or the presence of several active compounds in an active mixture. In 1985 and 1993, respectively, Samuelsson et al. and Cardellina et al. studied the chemical nature of bioactive compounds in NP extracts. Based on a combination of several extraction solvents and several Solid Phase Extraction (SPE) columns, they managed to deduce the chemical properties of active compounds (size, polarity, stability, acido-basic properties …) and therefore develop appropriate and more focused purification strategies. Interestingly, the authors represented their results in an elution matrix with active fractions being highlighted, thus revealing the chemical profiles of the active compounds. Therefore, this approach also allowed them to better select their active extracts for subsequent investigation by avoiding probable synergistic effects and false positives (when activity was lost with fractionation) but also by performing preliminary dereplication (eliminating from further consideration) when activity of the fractions was observed with similar chemical profiles. This approach was further developed by Månsson et *al.* [9] as the Explorative Solid-Phase Extraction protocol (E-SPE), by implementing the approach with the acquisition of LC-UV-MS profiles for active fractions. These analyses provided additional chemical properties for the active compounds (MS and UV spectra), allowing better dereplication of the active constituents. However, the authors especially focused on recurrent peaks in active fractions to manually detect and highlight compounds responsible for the activity.

In the past few decades, metabolomics has appeared as a rapidly emerging and developing field in NP chemistry. Extracts analyses mainly by HPLC-$MS^{(n)}$ and NMR spectroscopy has been employed for different purposes such as chemical profiling and dereplication [10–13], biomarker characterization [14,15], quality control [16–18] and also for bioactive drug discovery [19–22]. Metabolomics studies generate a huge amount of data related to all detected compounds in a sample and need the use of bioinformatics tools to highlight the information of interest within the collected data. To this purpose many software have been developed such as MZmine [23], R Cran packages [24] or Python packages [25] and are now available for the community. Nevertheless, effective strategies for identification of compounds present in small amounts and their associated biological effect from a complex mixture are crucially lacking. While the most commonly used statistical analysis to compare the chemical composition of different mixtures remains principal component analysis (PCA), this approach is generally not sufficient in NPs drug discovery programs because it does not take into consideration bioactivity data. Consequently, a great deal of interest has risen among the scientific community to link chemical fingerprints to bioactivity data using statistical methods, leading to the new term “biochemometrics.” Such approaches would overcome some bioassay-guided fractionation limitations and provide a more comprehensive insight of compounds responsible for the activity. According to the literature, several statistical models were used for detection of active compounds such as Pearson correlation [26–29], partial least squares (PLS) [30–35], discriminant analysis (PCA-DA, PLS-DA, OPLS-DA) [21,28,36–41] and hierarchical cluster analysis (HCA) [42]. However, most of these studies use only one statistical model for interpretation, which may impair exhaustiveness and accuracy of highlighted features mainly because all of these models are not well adapted to delineate chemical fingerprints and bioactivity data relationships. A combination of them appears as an interesting solution to overcome the limitations of each model independently and increase performance of biochemometrics.

In this study, we propose a new workflow based on the association of both E-SPE and biochemometrics using the combination of multiple statistical models (PCA, Spearman, F-PCA, PLS, PLS-DA) to target bioactive compounds from extracts. This workflow was developed and applied in a real-case study, with an extract of a marine-derived fungal strain: *Penicillium chrysogenum* MMS5, presenting high antiproliferative activity on breast cancer cells (MCF-7 cell line), which was mainly due to the presence of high amounts of ergosterol [43]. The script, written using the open-source R Cran software, allowed to combine all data-mining strategies and was made accessible for the whole NP community (Supplementary information S1). This approach represents a new approach to tackle some of the bottlenecks currently existing in NP drug discovery programs.

2. Material and methods

2.1. Fungal strain, culture and extraction

Penicillium chrysogenum (MMS5) was collected in November

1994 from cockles in Le Croisic (France) and identified by ITS sequencing (Genbank accession number MK015724). This marine-derived fungal strain was 3-point inoculated on PDA medium (Potato Dextrose Agar) in 50 Erlenmeyer flasks (50 mL of media per flask) and incubated for 12 days at 27 °C. The crushed gelose and mycelium were extracted twice with 100 mL of a CH_2Cl_2/EtOAc (50:50, v/v) solvent mixture after 30 min ultrasound treatment for the first step extraction and after a night at room temperature incubation for the second. The supernatants were combined and filtered over Büchner and reextracted. The filtrates were dehydrated by Na_2SO_4 and filtered over filter paper and over a 0.45 μm regenerated cellulose membrane (Sartorius Stedim Biotech) to remove spores. Solvent was evaporated using a rotary evaporator to obtain a crude extract (710 mg).

2.2. Fractionation for E-SPE

Fractionations were performed on four columns of solid phase extraction (SPE) using Silica gel (Chromabond®, 6 mL, 1000 mg, pore size 60 Å, particle size 45 μm, from Macherey-Nagel), C_{18} (Chromabond®, 6 mL, 1000 mg, pore size 60 Å, particle size 45 μm from Macherey-Nagel), Sephadex™ LH20 (GE Healthcare Biosciences AB) and Strata™-X phase (6 mL, 100 mg, Phenomenex®). These phases are abbreviated in the following manuscript as SiOH, C18, LH20 and SX, respectively. For separation using Sephadex™ LH20, the dry powder was swelled in MeOH and manually packed (1.5 mL in 6 mL empty cartridges from Macherey-Nagel).

The different mobile phases used were composed of CH_2Cl_2/MeOH mixtures (from 1:0 to 1:1) for the normal phase Silica gel, MeOH/H_2O mixtures (from 0:1 to 1:0) for the reverse phase C_{18}, MeOH/H_2O mixtures (from 0:1 to 1:0 and additional MeOH + 1% formic acid) for the polymeric reverse phase Strata ™-X, and only MeOH for LH20. Fractionation was performed in multiple sub-fractions until no more measurable mass was recovered from the column. The sub-fractions were then pooled in 5 fractions according to mass amounts. In our study, after placing on top of each cartridge a frit with 50 mg of the dried crude extract obtained from MMS5 *P. chrysogenum*, nineteen fractions were obtained by the following successive mobile phases: SiOH-1 (6 mL CH_2Cl_2 and 3 mL CH_2Cl_2/MeOH (1:9)), SiOH-2 (3 mL CH_2Cl_2/MeOH (3:7)), SiOH-3 (9 mL CH_2Cl_2/MeOH (1:1)), SiOH-4 (9 mL CH_2Cl_2/MeOH (1:1)), SiOH-5 (9 mL CH_2Cl_2/MeOH (1:1)), C18-1 (3 mL H_2O, 3 mL MeOH/H_2O (1:2), 3 mL MeOH/H_2O (2:1) and 4 mL MeOH), C18-2 (4 mL MeOH), C18-3 (3 mL MeOH), C18-4 (6 mL MeOH), C18-5 (56 mL MeOH), LH20-1 (2.75 mL MeOH), LH20-2 (0.75 mL MeOH), LH20-3 (0.75 mL MeOH), LH20-4 (4 mL MeOH); LH20-5 (34 mL MeOH), SX-1 (4 mL H_2O, 3 mL MeOH/H_2O (1:2) and 3 mL MeOH/H_2O (2:1)), SX-2 (3 mL de MeOH), SX-3 (4 mL de MeOH), SX-4 (6 mL MeOH), SX-5 (21 mL MeOH and 59 mL MeOH + 1% formic acid). All fractions were dried under reduced pressure at room temperature.

2.3. Controls preparation

For all four columns, a blank sample was prepared by extracting the PDA medium used for cultivation in the same conditions as described above and eluting the extract with the same mobile phases on the four columns. All PDA medium sub-fractions were pooled to obtain a negative control (BM) for cytotoxic assay and chromatographic analyses. Similarly, a part of the crude extract from MMS5 previously spotted on a frit was eluted on a column without any phase (but equipped with a frit, top and bottom) with a mix of the different solvents to obtain a positive control (CE). Additionally, to make sure the activity was recovered in the fractions, a reconstituted crude extract (RCE) was obtained for all four columns by mixing the corresponding fractions in proportional amounts. A quality control (QC) for subsequent HPLC-MS analyses was also prepared by mixing all the samples (fractions and controls) in equal quantities.

2.4. HPLC-MS analyses

Analyses of fractions and controls were performed on a Shimadzu instrument consisting of an Ultra-Fast Liquid Chromatography coupled to UV detection and High Resolution Electrospray Ionization Mass Spectrometry combining Ion trap and Time of Flight analysers (UFLC-UV-ESI-IT-TOFMS). The unit consists of two LC-20ADxr pumps, a SIL-20ACxr autosampler, a CTO-20AC column oven, an SPD-M20A PDA detector and a MBC-20A system controller. High performance liquid chromatography analyses were performed on a Kinetex ™ C_{18} column (100 × 2.1 mm, 2.6 μm, Phenomenex) heated in an oven equilibrated at 40 °C. A mobile phase consisting of CH_3CN/H_2O (acidified with 0.1% formic acid) was used, starting with 15% CH_3CN during 2 min, then increasing linearly to 100% CH_3CN within 23 min, holding at 100% CH_3CN for another 5 min, then returning to the initial conditions within 1 min, and holding for 4 min, for a total run time of 35 min at a flow rate of 0.3 mL/min. The mass spectrometer was operated in full-scan mode. MS data were recorded in the ESI positive mode in the mass range of *m/z* 100–1000 with a mass accuracy of 7 ppm and a resolution of 10,000 at *m/z 500*, using the following parameters: heat block and curved desolvatation line temperatures at 200 °C; nebulizing nitrogen gas flow at 1.5 L min^{-1}; interface voltage at (+) 4.5 kV and detector voltage of the TOF analyser at 1.6 kV. UV-VIS spectra were detected and collected from 190 to 600 nm.

The samples were prepared in MeOH (UPLC/MS grade) at concentrations of 0.4 mg mL^{-1}, stored at 4 °C before injection of 5 μL for each. The analyses were performed randomly and included solvent blank samples (pure MeOH) and QCs injected regularly throughout the sequence.

2.5. Cytotoxicity assays

Human breast cancer MCF-7 cells were purchased from the European Collection of Animal Cell Cultures (ECACC, Salisbury, UK). 3-(4,5 Dimethyl-2-thiazolyl)-2,5-diphenyltetrazolium bromide (MTT) was purchased from Sigma Aldrich (Saint Quentin Fallavier, France). MCF-7 cells were cultured at 37 °C in a humidified incubator with 5% CO_2 in DMEM medium supplemented with 10% fetal bovine serum (FBS), 1% glutamine and 1% penicillin-streptomycin. Viability of MCF-7 was tested in 96-well plate at a density of 10 000 cells per well in 200 μL of culture medium and allowed to adhere overnight. Then the seeding medium was removed before cell treatment. Crude Extract (CE), fractions and Reconstituted Crude Extracts (RCE) were dissolved in DMSO and tested at a unique concentration of 50 μg mL^{-1}. Pure ergosterol and ergosterol-5,8-endoperoxyde (purchased from Sigma-Aldrich and from in-house library respectively) were dissolved in EtOH and then diluted in 0.1% BSA containing-medium in order to obtain concentrations of 0.001-0.01-0.1-1-12,5-25-50 μM. After 24 h of incubation, MTT assay was performed by removing 100 μL of the medium and adding 50 μL MTT (at 2.5 mg mL^{-1}) to each well. The mixture was further incubated for 4 h, and the liquid in the wells was removed thereafter. Dimethyl sulfoxide (DMSO 200 μL) was then added to each well to solubilize the formazan product and the absorbance was read at 570 nm. The relative inhibition was expressed as a percentage of the non-treated control, which corresponded to medium supplemented with the same final concentration of DMSO or EtOH.

2.6. LC-MS data treatment

LC-MS data obtained for studied mixtures were treated. Automatic feature detection between 0 and 30 min in positive mode in MZmine2.31 software [23] was achieved using the parameters selected according to the TOF-MS detector. Peak detection was performed with the "mass detection" algorithm with a noise level of "1.6E4" in centroid mode. Then, chromatograms were built for all detected ions with a minimum time span of 0.1 min and a minimum intensity of "8.0E4" counts in positive mode, allowing 80 ppm tolerance on *m/z* values. Peak deconvolution was applied to the generated chromatograms with the "baseline cut-off" algorithm using a min peak height of "8.0E4", a peak duration range of 0.1–5 min and a baseline level of 1.6E4. Deisotoping filter was applied using the "isotopic peaks grouper" module with tolerance parameters adjusted to 0.1 min and 0.01 on *m/z* values. Feature alignment was achieved with the "Join Aligner" module with a *m/z* tolerance of 0.01 and a retention time tolerance of 0.3 min followed by gap filling using the "Gapfiller" module, yielding a combined dataset. The features detected from blank MeOH and non-inoculated culture medium samples (BM) were removed from the generated matrix to focus on the features really corresponding to the fungus production.

2.7. Biochemometric analyses

To link chemical profiles and associated biological data of all fractions obtained, biochemometrics analyses were performed. The spectral data matrix obtained from Mzmine software [23] including *m/z*, retention time, and peak area for each detected ions was imported to Excel Microsoft office 2011 version 14.7.7 (2010 Microsoft Corporation) and merged with the bioactivity dataset (inhibition percentages at 50 $\mu g\,mL^{-1}$) to form the final matrix. The R FiBiCo script developed in the present work is presented in details in the "Results and Discussion" section. All statistical models used and included in the script such as Principal Component Analysis (PCA), Spearman, Focused- Principal Component Analysis (F-PCA), Partial Least Squares regression (PLS) and Partial Least Squares Discriminant Analysis (PLS-DA) were computed using the open source software Rstudio version 1.1.447 (2009–2018 RStudio, Inc) and the following packages: "MetaboAnalystR" [44], "psy" [45], "mdatools" [46], "readr" [47] and "psych" [48].

2.8. Virtual evaluation of the script efficiency

The initial data matrix was manually modified to evaluate the ability of the R FiBiCo script to detect potential bioactive compounds. A total of 65 matrices were designed by adding a virtual feature which peak area values in each fraction were defined according to different models of relationships between peak area values and biological activity (linear, exponential, logarithmic and two other relationships with monotonic and non-monotonic "S-shaped" curves). For all relationships peak area values in each fraction for the virtual feature were given in accordance to the antiproliferative activity on MCF-7 (in %) with lowest peak area value for the less active fraction C18-2 (except for the non-monotonic model) and the highest for the most active fraction SX-5. Based on the initial matrix, where peak area values ranged from "8.96E3" to "7.68E8", different matrices were then constructed with the additional virtual feature having peak area values either in the same range (from "8.96E3" to "7.68E8") or 1000 times less intense (from "8.96E3" to "7.68E5"). The different models were here obtained by using the following equations for peak area value range from "8.96E3" to "7.68E5": Eq. (1) for linear, Eq. (2) for exponential and Eq. (3) for logarithmic, with *A* corresponding to biological activity values (inhibition percentages as obtained on MTT assays) and P_{area} to calculated peak area values. For monotonic and non-monotonic "S-shaped" curves, peak area values were assigned manually to fit this type of curve.

$$A = 1.00\,E7 \times P_{area} - 18.94 \quad \text{Eq. (1)}$$

$$A = 0.95 \exp.(5.56\,E{-}6 \times P_{area}) \quad \text{Eq. (2)}$$

$$A = 16.44 \ln(P_{area}) - 168.53 \quad \text{Eq. (3)}$$

Moreover, a random variation of 10, 30 and 50% (using "randbetween" function in excel) on each peak area value defined for each model was applied. Additionally, three additional matrices were obtained with virtual peak area values defined according to activity groups. This time, random values were assigned (using "randbetween" function) to the virtual feature based on the following conditions: (1) in accordance with biological activity groups, with peak area values in group 1 (the less active) ranging from "1.00E4" to "5.00E5", in group 2 from "5.00E5" to "2.50E7" and in group 3 from "2.50E7" to "1.25 E9", (2) not in accordance with biological activities with values in group 3 < group 2 < group 1 or with values in group 2 < group 1 < group 3. Finally, a last assay was performed with completely random peak area values assigned to the virtual feature based on the initial matrix range (from "8.96E3" to "7.68E8") or on a smaller range (from "8.96E3" to "7.68E5") for the 19 samples.

These experiments were repeated three times to obtain three matrices for each condition (except for matrices where the peak area value of virtual feature range from "8.96E3" to "7.68E8" with a random variation of 10%), which were all analysed by the FiBiCo script. The results generated were compared, especially the position of the virtual feature in the "super list" and which models allowed its selection.

2.9. NMR pharmacophoric deconvolution approach

Fractions C18-1 to C18-5 were dissolved in deuterated methanol (10 × 0.75 mL) from Eurisotop, (Saint-Aubin, France) at a concentration of 5 $mg\,mL^{-1}$. Acquisitions were performed on a Bruker NMR spectrometer operating at 500 MHz. COSY spectra were acquired on 2 scans, 4134 points on F2 axis and 512 points on F1 axis. Spectral processing was performed using the Plasmodesma program, written in Python and based on the SPIKE library [49,50].

3. Results and Discussion

As identifying relevant active molecules from complex mixtures represents a major challenge in NP drug discovery, this study here proposes a complete workflow to target bioactive compounds from a crude extract, based on a combination of fractionation, metabolomics and chemometrics. It is divided into three steps, namely the Explorative-Solid Phase Extraction (E-SPE) followed by the acquisition of biological and chemical data and the biochemometrics analysis with the combination of four statistical models (Fig. 1). The last step required the development of the new FiBiCo script as described below and available in Supplementary information S1.

3.1. Development of the bioinformatics tool to FInd BIoactive COmpounds: "FiBiCo" script

3.1.1. Statistical models

So far, different models have been described in the field to link biological to chemical data, mostly linear Pearson correlation, PLS

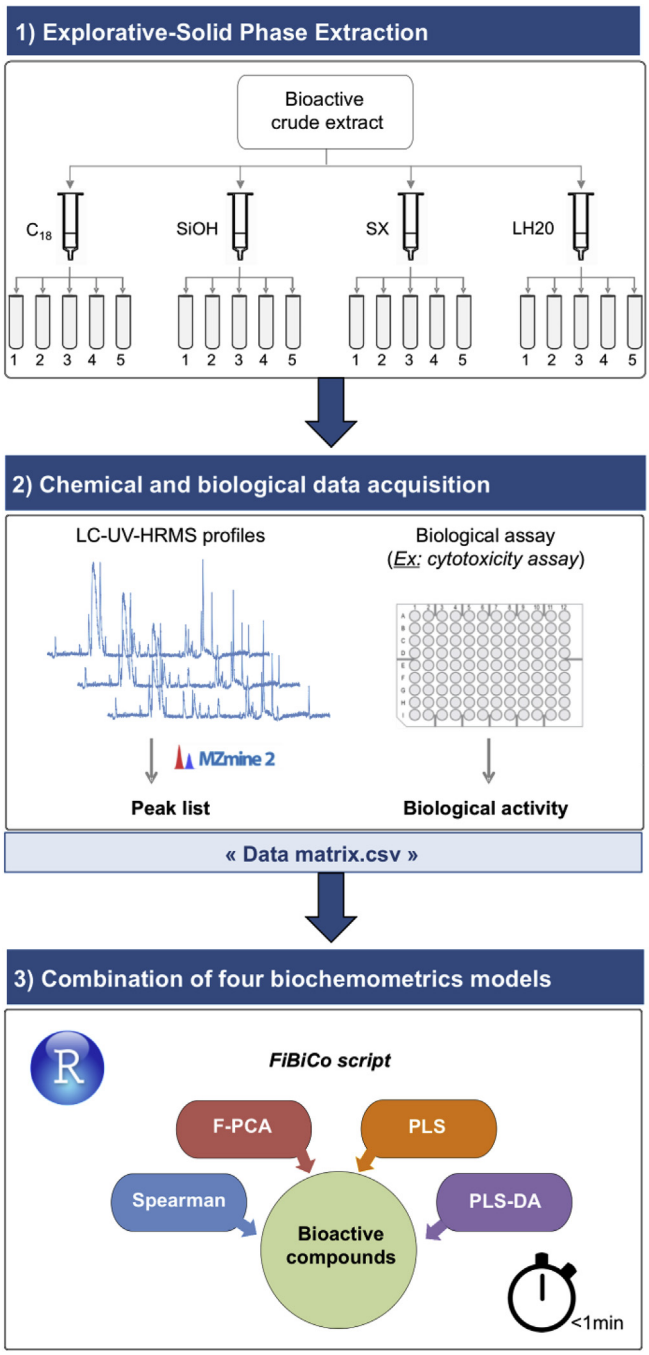


Fig. 1. Bioactive natural products discovery workflow based on a combination of Explorative-Solid Phase Extraction (E-SPE) and four biochemometrics approaches (Spearman, F-PCA, PLS, PLS-DA).

and PLS-DA. However, not many compare the results obtained from several analyses [37]. This is the reason why, in the present study, it was chosen to combine results from complementary univariate and multivariate models, namely Spearman correlation, focalised-PCA, PLS and PLS-DA. This should provide a more comprehensive view for the identification of bioactive compounds taking advantages of the differences between all those statistical approaches [51]. Spearman correlation is a univariate statistical approach, which measures the strength of correlation between two variables (*i.e.* peak area and biological activity) by evaluating their monotonic relationships (rank correlation coefficient). In this study, positive Spearman correlation scores (r) closer to 1 represented features, which peak area values seemed to increase along with biological activity. F-PCA, which was developed by Falissard et *al.* [52], while never used in the field of NP until now, revealed to be an interesting feature selection strategy. In fact, this model relying on Pearson correlation (linear correlation) allows to visualize simultaneously on a graphical representation, both correlations between a variable of interest Y (*i.e.* biological activity) and a set of variables X (*i.e.* feature peak areas) and correlations between X variables themselves. Like PCA, it corresponds to the projection of a correlation matrix in a two-dimensional plane, but unlike PCA, it accurately represents correlations of a given variable with the others (represented by the radius of the concentric circles) and even to test the statistical significance at the 5% level (represented by the red circle). The F-PCA plot center then corresponds to the variable of interest Y (biological activity) directing the analysis, whereas other variables (X) are represented by dots (*i.e* features). Their colour, localisation and organization around this central point allow to define the two types of correlations previously described. Consequently, the closer the feature is to the center, the closer to 1 (or-1) is the correlation coefficient. The red circle delimits statistical significance at the 5% level and allows to highlight dots with significant positive or negative correlation. Moreover, two dots close to one another indicate a strong positive correlation between these features themselves, while two diametrically opposed dots indicate a strong negative correlation between them. Partial Least Square (PLS) and Partial Least-Squares with Discriminant Analysis (PLS-DA) models are multivariate linear regression methods, commonly used in biochemometrics [53]. These supervised methods differ from PCA, because they allow to maximize the covariance of independent variables (peak area in our case) with a dependent variable (*i.e.* biological activity). PLS-DA allows to sharpen the separation between defined groups (*i.e.* nonactive/moderately or active fractions) of observations, by rotating PCA components such that a maximum separation among classes is obtained, and to understand which features carry the class separating information. Many values are employed in PLS for feature selection, including PLS loadings, weights, variable importance on projection (VIP), regression coefficients (RC), target projections, and selectivity ratio (SR) [54–56]. To date, VIP and selectivity ratio are the most popular ones in metabolomics [57].

The combination of all those models appeared as a new useful tool to comprehensively highlight features of interest. Metabolites appearing with a high score on several statistical models strengthen their importance to explain biological activity observed in the extract and fractions, and thus confirming their potential effect on the biological target studied.

3.1.2. Design of the script

This script was written using R Software (CRAN) to perform an automated processing of biochemometrics analyses on MacOS or Windows systems. It was designed to allow any user to perform the analyses on his/her own dataset. Therefore, it was very important that the operator could define and modify some parameters to improve methodology efficiency in accordance with his/her own data. In the same way, the file containing LC-MS and biological activity data to be read by the script was arranged as a simple matrix in a comma separated values (.csv) file. In this matrix, columns contained in the following order: sample names, groups, biological activity and peak areas or intensities of the different features "*m/z*_RT" detected. The script was divided into five steps (Fig. 2) and used "MetaboAnalystR" [44], "psy" [45], "mdatools" [46], "readr" [47],"psych" [48] packages.

The first step consists in filling information in the Rscript section "Define important information" with **(1)** general parameters

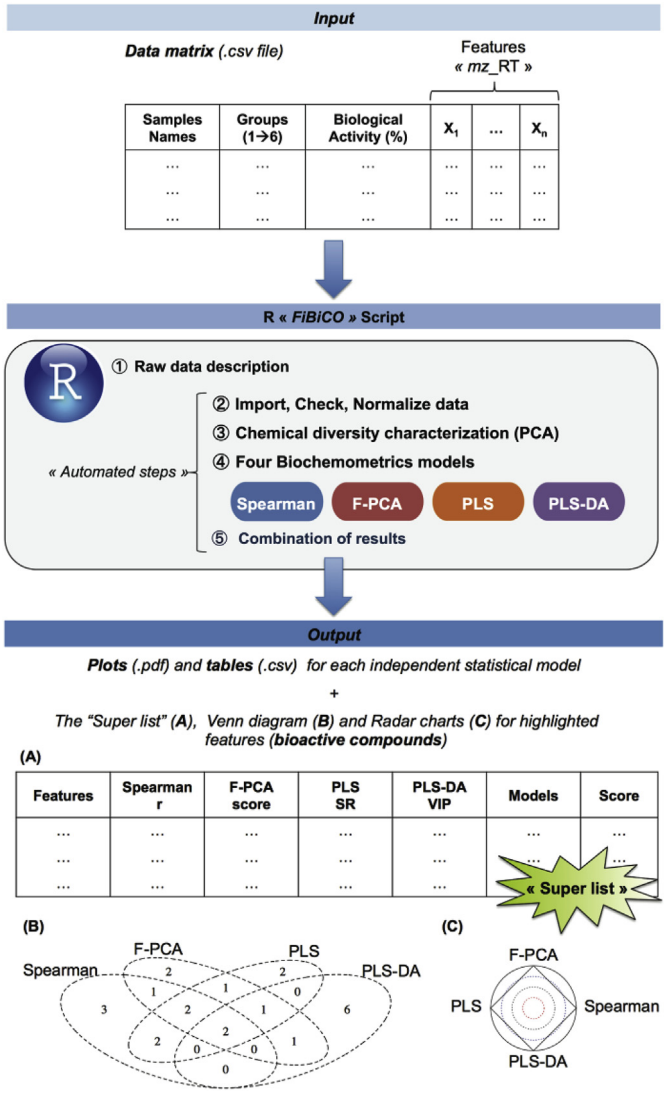


Fig. 2. Design of the "FiBiCo" script with its 5 major steps and the different results generated. At the end of the script, one bioactive compound can correspond to several highlighted features on (A), (B) and (C), as the same compound can generate several ions by electrospray ionization.

(operating system, file path to download the data matrix and working directory to export the results); **(2)** specific parameters according to the biological assay (biological activity, range and colours for groups ranked by ascending order); **(3)** parameters for normalization of raw data matrix (choosing options for sample normalization, data transformation and data scaling) [58] and **(4)** parameters for statistical analyses and graphical representations (as xy axes used for PCA and PLS with the associated titles and legends, etc..). After this first unique required contribution, the following 4 steps can be run successively without any more input from the operator. **The second step** of the script consists in importing, checking and normalizing the data matrix. **The third step** proceeds to the chemical diversity characterization of the fractions with PCA analysis. One output of this third step includes a graphical representation of the PCA score plot allowing the operator to assess if the best components to describe his/her data have been defined appropriately, and to highlight potential outliers. **The fourth step** is devised to perform biochemometrics analyses with the four chosen statistical models: Spearman, F-PCA, PLS and PLS-DA. Each model allows to determine the link between peak area (or intensities) and biological activity. Finally, **the last and most important step** in the biochemometrics tool designed in the present study, is the combination of all statistical results generating a "Super list" of features presenting a high score in the four models. It then highlights the metabolites for which the presence most probably explains the activity observed for the initial extract.

For the fourth step performing all biochemometrics analyses, a first output devised in the script was the generation and exportation in the working directory of one table (in a.csv file) per model with all the results obtained, together with the associated graphs (.pdf) as "Spearman score plot", "F-PCA score plot", "PLS score plot" and "PLS-DA score plot" (Supplementary information S2.1). To get a first overview of all detected peaks and all highlighted ones according to each model independently, each table was implemented with an additional classification column with importance levels as "High/Medium or Low". A combined graph representing the results of the four models was then created (with the colour code blue for "high", orange for "medium" and grey for "low"). To perform the classification, it was necessary to define conditions for each model (Table 1). While Spearman resulted in only one score of correlation ("r" coefficient) and was easier to classify, PLS and PLS-DA statistical models generated several important scores. Therefore, only few of them were selected for PLS and PLS-DA. So far, selectivity ratio for PLS and VIP score for PLS-DA have been reported to be the most important scores for interpretation of differences between groups or samples [33,35,59]. However, they did not reflect if the correlation was positive (activity was due to the presence of peak) or negative (activity was due to the absence of peak). Consequently, it was required to also take into account the sign value of regression coefficient for PLS model, and the peak area concentration of features in the groups for PLS-DA model. For F-PCA, the classification was carried out based on the graphical results, taking into account the three parameters defined as the correlation score (corresponding to the distance for each feature to the center of the graph), the sign of the correlation and the significance of the score (corresponding to features localised inside the red circle of the graph). This latter score was calculated as reported by Falissard et *al.* [52]. Some other parameters were calculated and reported in the table during the script (cf NA in Table 1). While they were not used for subsequent classification, they were kept available in the final table, as they could be useful for further in-depth interpretation.

For the final step of the script combining all the results from each independent model, it was chosen to allow the operator to define a maximal number of interesting ions for each model to obtain a suitable and easily interpretable list. It's important to note that the resulting final number of interesting features can be less if a smaller number of features (or even none) is in accordance with the defined limits in each model. For example, if the operator decided to choose a maximum of ten potential bioactive features per model, the "Super list" would contain a maximum of forty features (from 0 to 40). To focus on metabolites having a high score on several statistical results, the output table was designed to contain annotations identifying which model(s) allowed the selection of metabolites. Moreover, the distribution of this model-dependent selection was graphically represented on a Venn diagram. Additionally, after normalization of each model score (range 0–1), a global score for each feature was then calculated either by using the sum (range 0–4), the mean (range 0–1) or the Euclidean distance to 0 (range 0–2). Graphical representation of the "Super list" was finally generated as radar charts for each feature ordered by the calculated global score.

Table 1
Different values recovered from the statistical analyses performed by the FiBiCo script and conditions defined for correlation levels classification.

Models	Values generated and exported in the final tables	Classification of importance levels		
		High	Medium	Low
Spearman	Spearman coefficient score "r"	$r \geq 0.5$	$0.5 > r > 0$	$r \leq 0$
F-PCA	F-PCA score (combining distance and correlation sign)	F-PCA score > 0 & Significant[a]	F-PCA score > 0 & Non-significant	F-PCA score ≤ 0 Significant & non-significant
PLS	Selectivity Ratio (SR)	$SR \geq 1$	$SR < 1$	NA[b]
	Regression Coefficients (RC)	& RC > 0	& RC > 0	RC ≤ 0
	Variable of Importance in the projection (VIP)	NA	NA	NA
	Distances 1 and 2[c]	NA	NA	NA
PLS-DA	Variable of Importance in the projection (VIP)	$VIP \geq 1$	$1 > VIP \geq 0.5$	$VIP < 0.5$
	Mean peak areas or intensities for each group C_1, C_2, …, C_x[d]	$C_x > C_1$ & $C_x > C_2$ & … & $C_x > C_{(x-1)}$	$C_x > C_1$ & $C_x > C_2$ & … & $C_x > C_{(x-1)}$	NA
	Regression Coefficients (RC)	NA	NA	NA

[a] F-PCA score (dot on the score plot) localised in the red significant circle (meaning F-PCA score > limit value of significance at 5% level calculated by the "psy" package and depending on the data, especially the number of features).
[b] NA (not applicable) is noted when these values calculated by the script were not used for the classification of importance levels.
[c] On the PLS loadings plot, distances 1 and 2 correspond to the distance to the point of biological activity, and to the regression line, respectively.
[d] The number of defined groups being "x" (maximum x value = 6) and the xth group being the most active; NA (not applicable) is noted when these values calculated by the script were not used for the classification of importance levels.

3.2. Application to the marine-derived P. chrysogenum MMS5 extract

Following an initial screening of fungal extracts from our in-house library on MCF-7 breast cancer cell proliferation, extracts from the marine-derived *Penicillium chrysogenum* MMS5 isolated from cockles displayed strong inhibition whatever the growth medium used with 57–66% inhibition on CYA, YES and PDA media at 10 μg mL^{-1} (data not shown). It was found that the activity of this extract was related to high amounts of ergosterol, which was recently described as cytotoxic on MCF-7 cells [43] (Supplementary information S2.2). The proposed workflow was then applied to this MMS5 extract to evaluate its ability to highlight this bioactive compound.

3.2.1. Explorative-solid phase extraction: bioactivity elution matrix

Explorative-solid phase extraction on the crude extract was carried out with three stationary phases classically used in the NP field (SiOH, C_{18}, LH20) and a polymeric phase (StrataX). Those phases allow the fractionation of chemical constituents of the extract according to their physicochemical properties, such as polarity and hydrophobicity. As one fraction from SiOH (SiOH-5) did not present any measurable mass, a total of 19 fractions (4, 5, 5 and 5 for SiOH, C_{18}, LH20 and StrataX, respectively) were submitted to biological assay (MCF-7 proliferation inhibition), along with the positive control corresponding to the crude extract eluted without any stationary phase (CE) and a reconstructed crude extract (RCE). These latter RCE were obtained by mixing the fractions recovered from each column in the same proportions. This allowed evaluation of potential degradation of compounds present in the CE. For an easier visual representation, data results were sorted in an elution matrix (Fig. 3), where fractions were coloured in three groups according to their proliferation inhibition activity: red (0–20%), orange (20–40%) and green (40–60%). The activities ranged from 0 to 54% and the amount of fraction collected suggested a homogeneous separation of NPs present in the crude extract among them. As an example, all fractions from C_{18} column presented similar amounts (around 7 mg), while their corresponding antiproliferative activities were different, from no inhibition (C18–1, C18–2 and C18-3) to higher inhibition (43% and 50% for C18–4 and C18-5, respectively). SiOH column was the less efficient separation strategy as most of the weight was recovered in fraction SiOH-2 along with most of the activity, contrarily to SiOH-5 which was not tested due to the absence of mass.

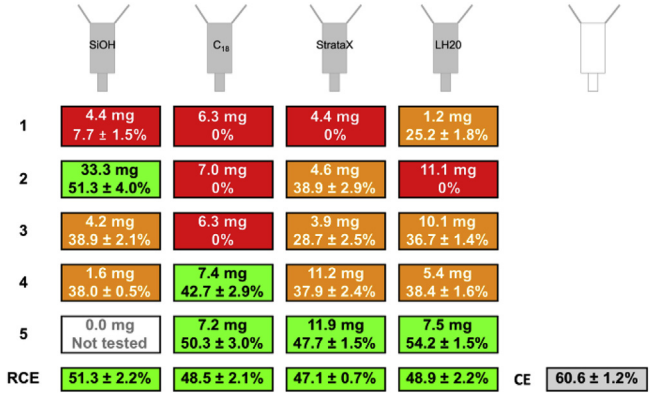


Fig. 3. Bioactivity elution matrix of the active crude extract indicating for each fraction the corresponding dry weight and proliferation inhibition percentage obtained at 50 μg mL^{-1} on MCF-7 cells (mean of triplicates ± SD), with the 3-group colour code defined as red (0–20%) for inactive, orange (20–40%) for moderately active and green (40–60%) for highly active samples. RCE and CE correspond to recombined crude extracts with fractions mixed in the same proportions and crude extract, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Interestingly, from this elution matrix, a first insight into the chemical nature of bioactive compounds could be hypothesized as being rather non-polar. Indeed, highest activity levels were observed in the last fractions of reversed phase and in the second fractions of normal phase columns. From this perspective, LH20 and StrataX columns also seemed to release the most active compounds in the last fractions meaning hydrophobic interactions may play an important role in these columns [60]. Interestingly, sufficiently distinct activity profiles were observed for C_{18}, LH20 and StrataX columns, allowing further comparison of their respective fraction compositions. It is important to note that several compounds should be responsible for the activity as a slight inhibition was observed for LH20-1 while this activity was lost for the following fraction and recovered afterwards. Finally, the maximum of proliferation inhibition obtained for individual fractions appeared at similar levels as RCE, meaning there was no loss of activity during fractionation and that activity was not depending on synergistic effect between compounds from different fractions. However, activities of the four RCE (48–51%) while close to the initial CE activity (61%), were slightly lower, probably in relation to the mass loss on

the column.

3.2.2. Chemical profiling and samples characterization

All fractions obtained from E-SPE were analysed with liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS) to compare their chemical profiles. Peaks only appearing in active fractions and not in inactive fractions were potential candidates for the observed bioactivity. However, no differences were revealed by manual comparison of the chromatograms, as their noticeable differences (corresponding to major peaks) were not linked to activity modulation (Supplementary information S2.3). Moreover, this manual inspection tends to be a very time-consuming task, in particular when differences are not obvious and if the active compounds correspond to low intensity peaks. Therefore, in this configuration, a metabolomics strategy appears more appropriate for the analysis of LC-MS profiles in regards to measured cell proliferation inhibition. This allows a more comprehensive description of the chemical composition of each fraction. Thus, automatic peak detection was here performed on raw chemical profiles by using Mzmine2 software [20], which allowed to obtain a list of 795 individual features among all samples, corresponding to peaks characterised by their *m/z* values and associated retention times (referred as "*m/z*_RT"), with their respective peak areas. The aligned data matrix was amended with biological data (inhibition percentages and groups as defined previously) and applied to the newly developed R FiBiCo script for statistical analyses.

After the first two steps consisting of filling needed information, defining appropriate parameters, uploading and normalizing the data, the script proceeded to the third step. The chemical diversity of fractions obtained from the E-SPE was then investigated using principal component analysis (PCA), an unsupervised multivariate statistical approach. PCA is based on a linear transformation that preserves as much of the variance in the original data as possible in a lower dimensional space (usually two- or three-dimensions). The first three components of the PCA explained 78.6% of the total variability of the model (with 41.7%, 28.3% and 8.3% for components 1, 2 and 3 respectively). PCA score plot (Fig. 4A) obtained for the components 1 versus 2, revealed that the chemical composition of the fractions from fractionation using SiOH column were very different from the other fractions. One exception was SiOH-2, which corresponded to the most active fraction of this specific fractionation. This confirms that separation on this phase was orthogonal to the other E-SPE phases. Interestingly, the PCA seemed to show that bioactive samples stood out from the others according to component 1. However, no feature on the loadings plot (Fig. 4B) related to the observed biological activity could be selected from PCA. Therefore, additional specific statistical models provided by the R FibiCo script were needed to provide accurate insight into the biologically active chemical constituents.

3.2.3. Biochemometrics analysis

After running the fourth step of the script corresponding to the biochemometrics analysis, the four statistical models returned very interesting features "highly" related to the biological activity (Table 2). This observation was consistent with the possibility to find bioactive metabolites produced by *P. chrysogenum*. Differences in features importance levels were observed. For example, among the 795 features detected by LC-MS, the script allowed to highlight 50, 42, 38 and 51 features with "high" importance level for Spearman, F-PCA, PLS and PLS-DA, respectively.

These features of interest were highlighted in blue ("high" importance level) on the combined graph representing the results of the four models (Fig. 5). Spearman score plot (Fig. 5A) displayed two main groups of chromatographic peaks indicating potential

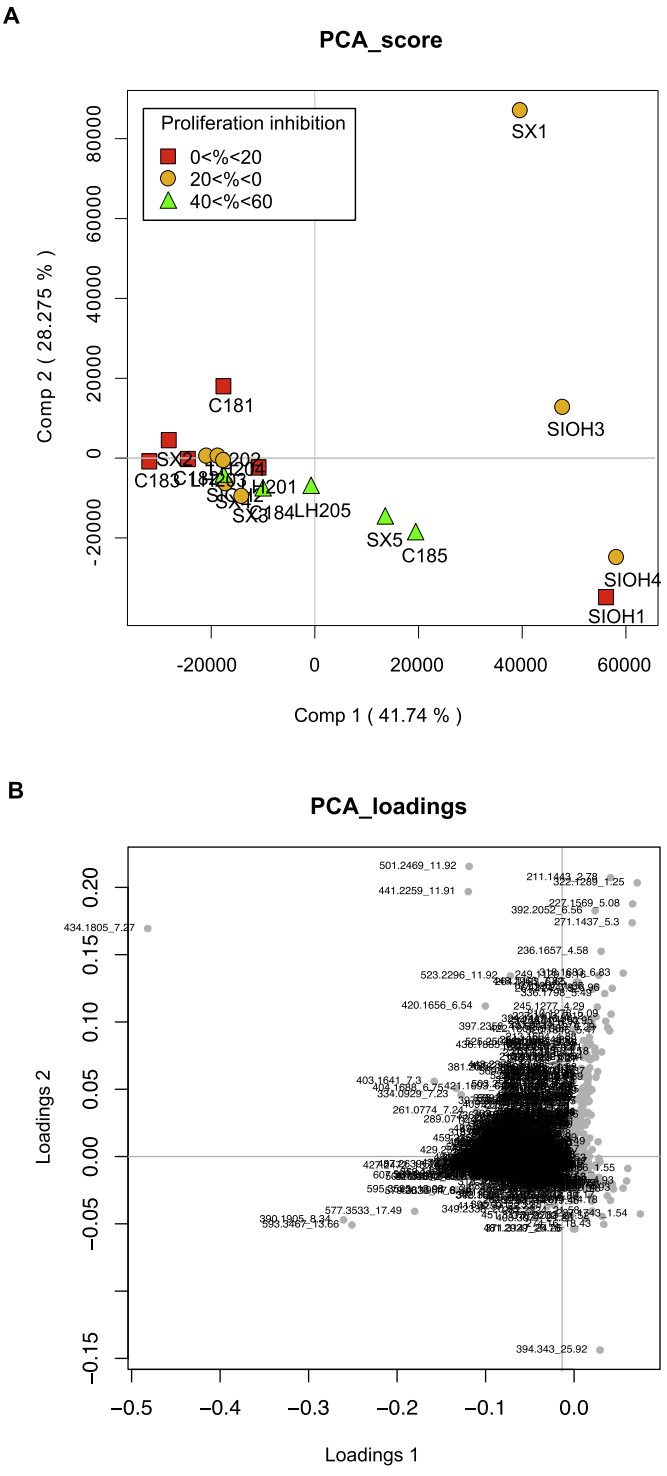


Fig. 4. Principal Component Analysis (PCA) score plot (A) of fractions based on the 795 features (LC-MS peaks), with the 3-group colour code according to proliferation inhibition (MCF-7) defined as red (0–20%), orange (20–40%) and green (40–60%) and the corresponding PCA loadings plot (B). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

bioactive compounds were either moderately polar (RT = 10–15 min) or apolar (RT = 20–28 min). Similarly, F-PCA (Fig. 5B) also revealed two groups of "high" importance features: one in the upper right corner and another in the upper left corner presenting many dots being more significantly correlated to the

Table 2
Number of LC-HRMS features related to proliferation inhibition of MCF-7 cell lines with high, medium or low importance levels in the case of *Penicillium chrysogenum* MMS5 originating data matrix.

Importance Levels	Spearman	F-PCA	PLS	PLS-DA
High	50	42	38	51
Medium	140	164	221	72
Low	605	589	536	672
Total	795	795	795	795

activity (because closer to the central red point representing the biological activity variable) with r coefficients between 0.6 and 0.8. In accordance with the previously defined parameters for the classification, these features (coloured in blue) were highlighted because they had both an absolute F-PCA score value significantly different from 0 at the 5% level (inside the red circle) and a positive correlation sign. Among these highlighted features, F-PCA allowed to obtain additional information of the relationship between features themselves. In fact, many of these dots were closely located on the graph because they also correlated between them, meaning they possessed similar presence pattern across fractions. However, no strong negative correlation was observed between them (absence of diametrically opposed blue dots). The fact that two groups of features were observed in contiguous quadrants could mean that not only one compound (but at least two) has an influence on the activity observed. Interestingly, PLS model results (Fig. 5C) highlighted similar peaks as Spearman in the RT range 20–28 min. Additionally, the two first VIP from PLS-DA model (Fig. 5D) seemed to highlight features at 26.8 min presenting high area value in the most active group of samples (group 3). The results obtained from PLS and PLS-DA gave satisfying R^2 values (0.82 and 0.96 respectively), meaning the models were correctly fitting the data, allowing to trust their results [61,62]. From these first observations, it appeared interesting to compare highlighted features and especially bring out those which were common in all models. This is where the last step of the FiBiCo script was carried out to provide a "Super list" of most interesting features and their corresponding radar charts (Supplementary information S2.4 and S2.5), based on the combination of the top ten of each model. In the case of *P. chrysogenum* MMS5 a total of 23 highly interesting features was present in the "Super list". The corresponding Venn diagram (Fig. 6) displayed the importance of combining the results from the four different statistical analyses as they highlight complementarily the features of importance. In fact, among the 23 highlighted in the "Super list", only 2 were highlighted by all the models, while 13 were selected by only one of them, *i.e.* 3, 2, 2 and 6 features by Spearman, F-PCA, PLS and PLS-DA respectively (Fig. 6A). It is important to note that complementarity between the models remained even when increasing the number of features to select in each model to the top-20, as in this case, only 6 (out of 41) were

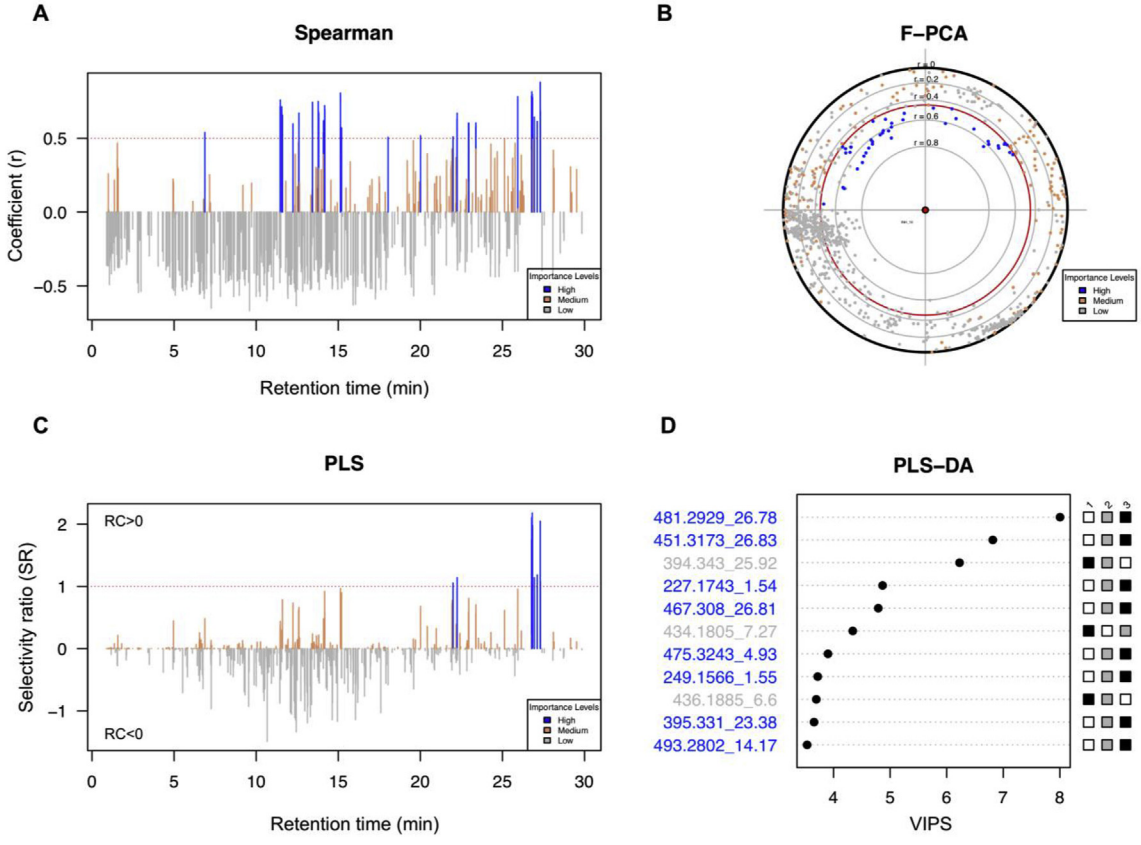


Fig. 5. Graphical representation of results obtained from the four statistical analyses: (A) Spearman, (B) F-PCA, (C) PLS and (D) PLS-DA. A colour-code "blue/orange/grey" was defined according to the feature importance level, with blue corresponding to features with high importance levels, most probably related to bioactive compounds. Orange refers to lower but positive relationships, while grey corresponds to features with low importance levels. The limit value used to select high important features was represented as a red line for Spearman (correlation coefficient (r) $\geq$ 0.5) and PLS (Selectivity Ratio (SR) $\geq$ 1 and Regression Coefficient (RC) $\geq$ 0). For F-PCA model, the red circle corresponded to the limit value of significance for the F-PCA score at the 5% level (features inside this circle have positive or negative correlations significantly different from 0). For PLS-DA, the greyscale on the right represents the mean peak area values in each group for the top-11 features, which are presented by decreasing VIP value ($\geq$1) from top to bottom. Highly important features (in blue) correspond to those where the mean peak area value in the most active group is higher than in the others. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

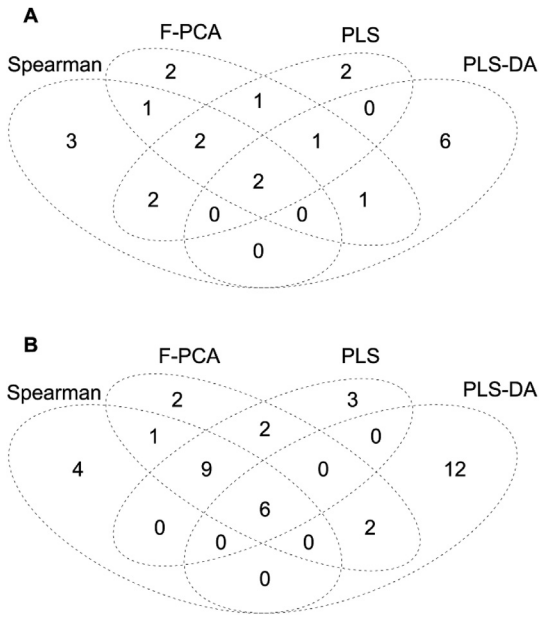


Fig. 6. Venn diagrams representing overlapping interesting features returned by the 4 models of the script (Spearman, F-PCA, PLS and PLS-DA) with the combination of (A) the top-10 features from each model giving a total of 23 features and (B) the top-20 giving 41 features.

highlighted by all the models (Fig. 6B). The user should then be careful on the definition of this number of features (10, 20 or more) to select in each model for the construction of the "Super list", according to his/her own data, as if too restrictive, it may exclude features of interests in some models.

3.2.4. Identification of highlighted hits and confirmation of biological activity

The "Super list" of 23 features was then thoroughly investigated to annotate and identify the potential bioactive compounds. Overlay of their extracted ion chromatograms (XICs) revealed a perfect overlap for 12 out of the 23 features. Consequently, this overlap suggested that the 12 features arose from the same molecule. It was identified as ergosterol, as expected, by comparison to HPLC-UV-HRMS profiles of a standard (Supplementary information S2.6). The workflow proposed and developed in the present article, combining E-SPE and biochemometrics, was then successful in identifying the main bioactive compound responsible for the initial activity observed from a complex mixture (MMS5 *P. chrysogenum* extract). It is important to note that, while present in large amounts in the extract, ergosterol was not detected at high levels by mass ESI-spectrometry, because of a poor ionization. This means that the script developed allowed to highlight minor peaks as corresponding to potential bioactive compounds.

The other 11 features highlighted by the "Super list" were further investigated to understand if they could also have an influence on the activity observed or be considered as false positives. By grouping them according to peak shapes and retention times, eight other putative molecules could be revealed with the following retention times: 1.5 min, 4.2 min, 14.2 min, 15.1 min, 23.4 min, 25.9 min, 26.9 min and 27.3 min. Among these, annotation was tentatively carried out based on the observed adducts and the predicted molecular formula (Supplementary information S2.7). One of them matched with ergosterol-5,8 endoperoxyde, which was further confirmed by the injection of a standard. Its biological activity on MCF-7 breast cancer cell was then confirmed with a proliferation inhibition percentage of 48.7% at 50 μM, which is much lower than ergosterol but in accordance with our results (lower score: 1.34 compared to 3.51 for ergosterol). Moreover, annotation of chaetoglobosin derivatives could be proposed for the compound at 15.1 min (global score of 1.91). Interestingly, cytotoxicities were reported on breast cancer cells lines for chaetoglobosins A, C and G with IC_{50} of 37.56, 19.97 and 38.77 μM on MDA-MB-231 cells, respectively [63]. The last compound with a good score at 27.3 min had a hit for an ergosterol derivative, which could make sense regarding the activities of both ergosterol and ergosterol endoperoxyde in this series. Other annotations included putative diketopiperazine derivatives for compounds at 1.5 and 4.2 min, which biological activities on breast cancer cells were not reported. It is important to note here that these corresponded to signals that were not detected in the crude extract. It could be due to a concentration phenomenon inherent to the fractionation step. For other features, due to the difficulty to predict the molecular formula or a number of matches in databases too important, a good annotation could not be obtained. Further purification work is then currently in progress towards these other features that were not identified. In fact, even if we cannot ascertain that all the features highlighted by the script have an activity on MCF-7 cells, these first results leading to the annotation of two other ergosterol derivatives, with one of them presenting proliferation inhibition properties, together with chaetoglobosin derivatives already reported to inhibit breast cancer cells proliferation, tends to prove our method is quite powerful.

3.3. Performance evaluation of the R FiBiCo script

3.3.1. Estimation of the robustness of the approach

As calculating false positive and false negative rates revealed impossible here on a real case study, because it would require the isolation and testing of all the molecules from the natural crude extract, alternatively, an assessment of the robustness of the approach based on the literature was performed. So far, among the metabolites reported from *P. chrysogenum* species (around 200), only 15 metabolites have been reported to inhibit breast cancer cells proliferation (including MCF-7) (Supplementary information S2.8). Four compounds in this list matched (MS and UV spectra) with features from the peak list (obtained from MMS5 *P. chrysogenum*), namely ergosterol, ergosterol-5,8-endoperoxide, roquefortine C and meleagrin. Additionally, the molecular formula $C_{32}H_{36}N_2O_5$ (corresponding to chaetoglobosins A, C and G) also appeared to be a good match with one feature at 15.1 min. Its UV spectrum could not be recovered and compared due to low intensity. For the other compounds reported to have antiproliferative activity on breast cancer cells from *P. chrysogenum* species, no matches within the peak list of 795 features were found. This is easily understandable as the conditions used in the present study may have not been appropriate to obtain the biosynthesis of these other metabolites. Among these matches, ergosterol, ergosterol-5,8-endoperoxide and chaetoglobosins corresponding features were effectively highlighted by the R FiBiCo script, while meleagrin and roquefortine C were not. However, the activity of these two latter compounds originating from the same biosynthetic pathway [64], may be questionable as meleagrin has been tested twice on MCF-7 cells with completely different results [65,66]. One reports an IC_{50} of 1.9 μM while the other reports no activity at all. It seems like the MCF-7 cell line used in the present study is not sensitive to this type of molecules. Preliminary results on fractions enriched with meleagrin and roquefortine C (data not shown) tend to prove this absence of inhibition on this cell line.

Given to these different investigations (dereplication on the "Super list" in Table S2.7 and the list of molecules from *P. chrysogenum* reported for their activity on breast cancer cells in

Table S2.8), we can assume the false positive and false negative rates of this method are quite low. Nevertheless, the main purpose of the method described in the present article is to highlight molecules of interest in an extract and orientate further purification work in a more rational way. Even if, there may be false positives, it still helps drastically reducing the number of molecules to be further investigated for their biological activity. Additionally, false positives could correspond in many cases, to molecules presenting the same chromatographic behaviour than the active ones, and then probably to chemically related derivatives, which would also be worth isolating to study structure-activity relationships. However, if no false positive is preferred, one easy option is to focus only on hits returned by the 4 models. In the present case, this would only return two features only corresponding to ergosterol (See Supplementary information Table S2.7).

3.3.2. Virtual training

While the application of the workflow to a real case proved to be successful in the detection of the main active compound, further evaluation of the FiBiCo script was performed with virtual modifications of the data matrix to verify its ability to detect interesting features and investigate its limits. A virtual feature was then added on the initial MMS5 *P. chrysogenum* data matrix with different peak area values according to proliferation inhibition for each sample. A total of 16 matrices were constructed with one additional feature, which peak area values in the 19 samples, were attributed

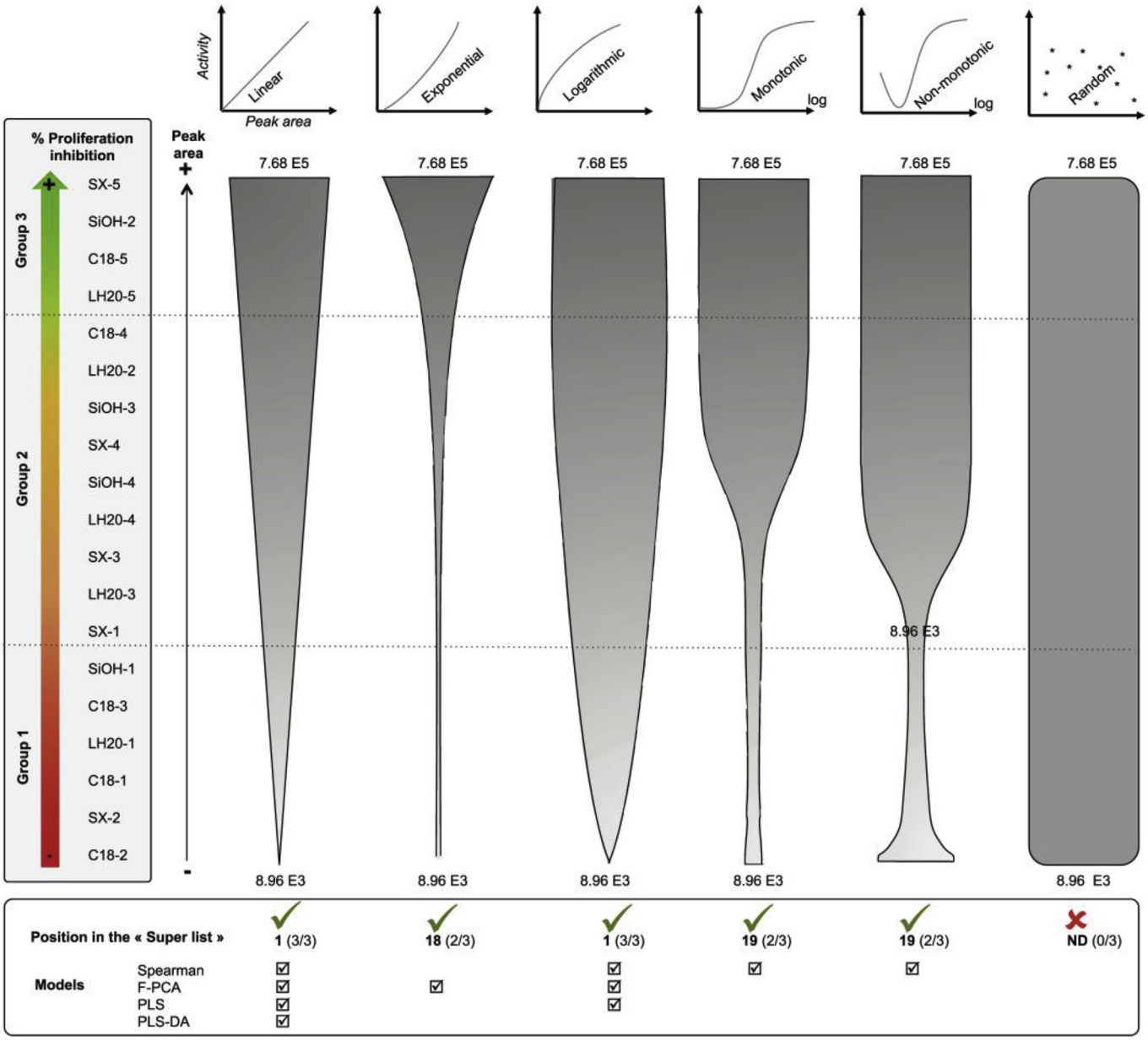


Fig. 7. Virtual evaluation of the FiBiCo script performance with the addition of a virtual feature in the MMS5 *P. chrysogenum* data matrix. The figure presents the conditions tested for the additional feature designing the different relationships between peak area value and biological activity. In accordance with biological activity obtained from the 19 samples, a peak area value was assigned for the additional feature to represent six different relationships: linear, exponential, logarithmic, monotonic, non-monotonic, and random. To be closer to real conditions, a random variation of 30% was applied on the peak area values of the additional feature for each model to obtain a final range from "8.96E3" to "7.68E5". The R FiBiCo script was applied on the modified initial matrix with this additional feature generating its position in the "Super list" and revealing which models allowed its selection. These experiments were performed in triplicate. Additional conditions tested are presented in Supplementary information S2.9.

according to linear, exponential, logarithmic, monotonic, non-monotonic or random relationships to biological activity. To avoid perfect fit and be closer to real conditions, random variations of 30 and 50% were applied on each peak area value defined for this additional feature in the different matrices. After uploading the matrices on the R FiBiCo script, the results obtained were analysed and reported on a graphical representation (Fig. 7 and Supplementary information S2.9). For each assay, the position of the additional virtual feature in the "Super list" was picked together with the information on which of the biochemometrics models allowed the selection of this feature. Data results obtained with a random variation of 30% performed in triplicate for all relationships (Fig. 7) revealed that R FiBiCo script was able to highlight the additional feature for all meaningful relationships except for the random model, for which no hit was expected. The selection of the additional feature was very good when the relationship was linear or logarithmic, i.e. detected by three or four statistical models (Spearman, F-PCA, PLS and PLS-DA) and with the best global score in the super list (ranking 1st). However, for other activity-to-peak area relationships, some statistical models appeared to be more adapted. For example, for monotonic and non-monotonic relationships, Spearman correlation was the only model able to catch the additional feature. Moreover, while its global score of ranked it in the 19th position, it actually ranked in the two first positions in the Spearman model alone, proving the relevance of Spearman for this type of relationship. Another example is given by the exponential relationship, which was only detected by F-PCA model (in positions 3–5). These results demonstrate the importance to use the combination of the four biochemometrics models to highlight interesting features, because in dose-response curves, different relationships can be observed between the amount of a compound and its biological activity. It is important to note that the previously described experiments were performed on values, which were quite low and in a small range ("8.96E3" – "7.68E5"), with a 30% variation on each. These were chosen to test the model and approach its limits. Other experiments on broader ranges of values (Supplementary information S2.9) showed that increasing the maximal peak area allowed a better detection of the additional feature presenting an interesting relationship with activity. This means that for highly detected compounds, more statistical models may respond. In addition, the smaller the variation of the values (10%), the better the detection as well. Further assays were also performed using groups of biological activity (not the activity for each of the 19 samples). They allowed to confirm the efficiency of the script in the detection, because the additional feature was highlighted only if its peak area values in samples of a group were in accordance with biological activity (i.e. group3>group2>group1 or group2<group3<group1).

3.3.3. Comparison with the NMR pharmacophoric deconvolution approach

The recrudescence of "omics" and the development of new strategies in bioactive NP research allowed to compare our strategy using LC-MS analyses with a recently biochemometrics approach developed by Margueritte et al. [49] using NMR pharmacophoric deconvolution approach. This strategy was based on the use of an NMR fingerprint obtained by automatic differential analysis of 2D NMR data to target and identify bioactive natural product(s) in a complex crude extract. The NMR pharmacophoric fingerprint is obtained through the common cross-peaks in the $^{1}H–^{1}H$ COSY NMR spectra of the active fractions. Common cross-peaks give partial structural information about the activity-bearing compound because successive fraction obtained by the fractionation should contain shared molecules. Therefore, $^{1}H–^{1}H$ COSY spectra were recorded for the five C18 fractions obtained from *P. chrysogenum* crude extract. NMR data was processed by Plasmodesma program. This program written in Python was developed to process automatically NMR data set and provide the best analysis differential of NMR data. To achieve this, Plasmodesma divides NMR data in buckets and measures different variables. Then, the generated bucketlist were used by a Python script on the Jupiter notebook environment. Areas values of buckets were subjected to a cleaning and symmetrization steps to remove a large part of the noise and artefacts. The correlation analysis between spectral features and activity levels was performed with the set of regression tools from the scikit-learn library. A linear regression was applied between areas values and the bioactivity of fractions to produce the pharmacophoric fingerprint (Fig. 8B). Its chemical shifts of correlation peaks suggested a terpenoid skeleton. The overlay of ergosterol spectrum (Fig. 8A) and the pharmacophoric fingerprint (Fig. 8B) allowed to highlight 14 common cross-peaks (Fig. 8C), suggesting ergosterol was the main molecule responsible for the biological activity. This result was in accordance with the previous FiBiCo analysis further confirming its efficiency.

The use of NMR data was of great interest in this presented case, as ergosterol presents an unusual ionization pattern with low signal in ESI-MS, even though it is a major component of the extract. Therefore an analytical method having a more linear link between content and signal should provide an alternate accurate approach.

In fact, while NMR is able to detect all hydrogenated molecules, it suffers from a lack of sensitivity, making minor compounds much

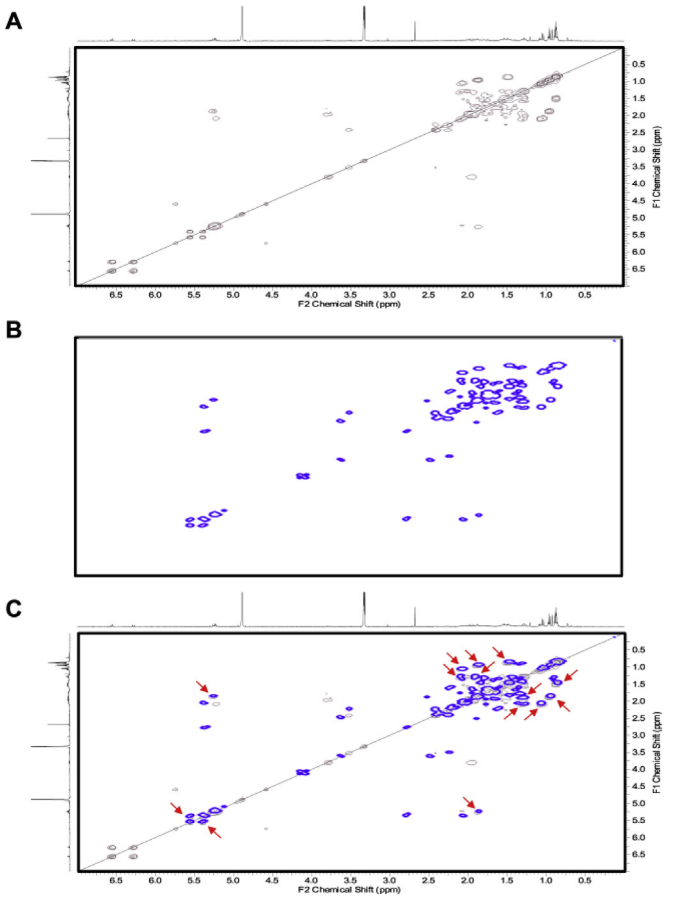


Fig. 8. NMR pharmacophoric approach applied to the five C_{18} fractions from *P. chrysogenum* MMS5 crude extract. A) $^{1}H–^{1}H$ COSY NMR spectrum of ergosterol, B) automatic pharmacophoric fingerprint, C) overlay of both with red arrows corresponding to common features. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

more difficult to detect. On the contrary, MS is highly sensitive but is limited by the ability of molecules to ionize depending on the ionization source of the instruments. Therefore, it is important to remember that all strategies reported in the literature have their limits and can be complementary.

4. Conclusion

Currently, the central challenge in NP research is the discovery of new bioactive compounds. With the recrudescence of "omics" studies in this domain, several strategies based on NMR and MS analyses, are currently emerging to highlight and identify bioactive compound from natural complex mixtures at a very early stage. In this study, a complete workflow is proposed, which combines explorative SPE and chemometrics tools using LC-MS data with the integration of four statistical models (Spearman, F-PCA, PLS and PLS-DA) to highlight potential bioactive compounds. Additionally, a script encoded with R software (named FiBiCo) was here developed to automatically highlight the most interesting features by including the complementary results of all statistical analyses, and made available for the scientific community (Supplementary information S1).

This complete workflow applied to a bioactive *P. chrysogenum* MMS5 crude extract, where ergosterol was the main active component, successfully highlighted this compound as responsible for the proliferation inhibition of MCF-7 breast cancer cells. This allowed to prove the efficiency of the developed R FiBiCo script. This is even more interesting as ergosterol was very poorly ionised with electrospray ionization showing that minor compounds can be highlighted, which is essential in bioactive natural product research. The performance of the FiBiCo script was evaluated with virtual modifications of the data matrix, proving its ability to detect different types of relationships between activities and chemical compounds amounts. The complementarity of the statistical models proved to be very important, especially if a low false positive rate is preferred. Moreover, the performance of the script in picking only one feature (among 796), with low values in a small range and with a high variation (30–50%), proved its efficiency to "find the needle in the haystack". The workflow was also evaluated by comparing the results obtained to those given by another recently described automatic deconvolution strategy using NMR data. The pharmacophoric fingerprint revealed also ergosterol as the bioactive compound, strengthening the validity of the workflow described and developed in the present study.

Perspective of this work will be to apply the methodology to other bioactive extracts, in the context of drug discovery screening programs. It effectively responds to one of the many issues encountered in this field, by rationalizing subsequent purification work to directly focus on compounds responsible for the activity.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Cancéropôle Grand Ouest (CGO), La Ligue Contre le Cancer and the French Ministry of Higher Education and Research (PhD grant). The authors also acknowledge J.M. Huvelin, A. Burghelea and T. Robiou du Pont for their technical participation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2019.04.038.

References

[1] F.E. Koehn, G.T. Carter, The evolving role of natural products in drug discovery, Nat. Rev. Drug Discov. 4 (2005) 206–220.
[2] M.G. Weller, A unifying review of bioassay-guided fractionation, effect-directed analysis and related techniques, Sensors 12 (2012) 9181–9209.
[3] F. Bucar, A. Wube, M. Schmid, Natural product isolation – how to get from biological material to pure compounds, Nat. Prod. Rep. 30 (2013) 525–545.
[4] D.J. Newman, G.M. Cragg, Natural products as sources of new drugs from 1981 to 2014, J. Nat. Prod. 79 (2016) 629–661.
[5] J.W.H. Li, J.C. Vederas, Drug discovery and natural products: end of an era or an endless frontier? Science 325 (2009) 161–165.
[6] Y.F. Pouchus, A.F. Benslimane, J.F. Verbist, SESAME: an expert system for bioassay-directed isolation of active compounds, Tetrahedron Comput. Methodol. 2 (1989) 55–64.
[7] G. Samuelsson, G. Kyerematen, M.H. Farah, Preliminary chemical characterization of pharmacologically active compounds in aqueous plant extracts, J. Ethnopharmacol. 14 (1985) 193–201.
[8] J.H. Cardellina, M.H.G. Munro, R.W. Fuller, K.P. Manfredi, T.C. McKee, M. Tischler, H.R. Bokesch, K.R. Gustafson, J.A. Beutler, M.R. Boyd, A chemical screening strategy for the dereplication and prioritization of HIV-inhibitory aqueous natural products extracts, J. Nat. Prod. 56 (1993) 1123–1129.
[9] M. Månsson, R.K. Phipps, L. Gram, M.H.G. Munro, T.O. Larsen, K.F. Nielsen, Explorative solid-phase extraction (E-SPE) for accelerated microbial natural product discovery, dereplication, and purification, J. Nat. Prod. 73 (2010) 1126–1132.
[10] J.-L. Wolfender, G. Marti, A. Thomas, S. Bertrand, Current approaches and challenges for the metabolite profiling of complex natural extracts, J. Chromatogr. A 1382 (2015) 136–164.
[11] P.-M. Allard, G. Genta-Jouve, J.-L. Wolfender, Deep metabolome annotation in natural products research: towards a virtuous cycle in metabolite identification, Curr. Opin. Chem. Biol. 36 (2017) 40–49.
[12] J.Y. Yang, L.M. Sanchez, C.M. Rath, X. Liu, P.D. Boudreau, N. Bruns, E. Glukhov, A. Wodtke, R. De Felicio, A. Fenner, W.R. Wong, R.G. Linington, L. Zhang, H.M. Debonsi, W.H. Gerwick, P.C. Dorrestein, Molecular networking as a dereplication strategy, J. Nat. Prod. 76 (2013) 1686–1699.
[13] J.-L. Wolfender, J.-M. Nuzillard, J.J.J. van der Hooft, J.-H. Renault, S. Bertrand, Accelerating metabolite identification in natural product research: toward an ideal combination of LC-HRMS/MS and NMR profiling, *in silico* databases and chemometrics, Anal. Chem. 91 (2019) 704–742.
[14] D.G. Cox, J. Oh, A. Keasling, K.L. Colson, M.T. Hamann, The utility of metabolomics in natural product and biomarker characterization, Biochim. Biophys. Acta Gen. Subj. 1840 (2014) 3460–3474.
[15] M.D. Luque de Castro, F. Priego-Capote, The analytical process to search for metabolomics biomarkers, J. Pharm. Biomed. Anal. 147 (2018) 341–349.
[16] F. Perrotti, C. Rosa, I. Cicalini, P. Sacchetta, P. Del Boccio, D. Genovesi, D. Pieragostino, Advances in lipidomics for cancer biomarkers discovery, Int. J. Mol. Sci. 17 (2016).
[17] K.M. Lee, J.Y. Jeon, B.J. Lee, H. Lee, H.K. Choi, Application of metabolomics to quality control of natural product derived medicines, Biomol. Ther. 25 (2017) 559–568.
[18] E.O. Olawode, R. Tandlich, C. Garth, 1H-NMR Profiling and chemometric analysis of selected honeys from South Africa, Zambia, and Slovakia, Molecules 23 (2018) 2–19.
[19] S.L. Robinette, R. Brüschweiler, F.C. Schroeder, A.S. Edison, NMR in metabolomics and natural products research: two sides of the same coin, Acc. Chem. Res. 45 (2012) 288–297.
[20] M. Fillet, M. Frédérich, The emergence of metabolomics as a key discipline in the drug discovery process, Drug Discov. Today Technol. 13 (2015) 19–24.
[21] M. Mandrone, A. Coqueiro, F. Poli, F. Antognoni, Y.H. Choi, Identification of a collagenase-inhibiting flavonoid from *Alchemilla vulgaris* using NMR-based metabolomics, Planta Med. (2018) 941–946.
[22] F. Olivon, P.-M. Allard, A. Koval, D. Righi, G. Genta-Jouve, J. Neyts, C. Apel, C. Pannecouque, L.F. Nothias, X. Cachet, L. Marcourt, F. Roussi, V.L. Katanaev, D. Touboul, J.L. Wolfender, M. Litaudon, Bioactive natural products prioritization using massive multi-informational molecular networks, ACS Chem. Biol. 12 (2017) 2644–2651.
[23] T. Pluskal, S. Castillo, A. Villar-Briones, M. Orešič, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinf. 11 (2010) 2–11.
[24] R. R Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2018. https://www.R-project.org/.
[25] P.C. Team, Python: a dynamic, open source programming language. https://www.python.org/, 2015.
[26] T. Inui, Y. Wang, S.M. Pro, S.G. Franzblau, G.F. Pauli, Unbiased evaluation of bioactive secondary metabolites in complex matrices, Fitoterapia 83 (2012) 1218–1225.

[27] S. Bertrand, A. Azzollini, A. Nievergelt, J. Boccard, S. Rudaz, M. Cuendet, J.L. Wolfender, Statistical correlations between HPLC activity-based profiling results and NMR/MS microfraction data to deconvolute bioactive compounds in mixtures, Molecules 21 (2016) 2–13.
[28] L.A. Richards, C. Oliveira, L.A. Dyer, A. Rumbaugh, F. Urbano-Muñoz, I.S. Wallace, C.D. Dodson, C.S. Jeffrey, Shedding light on chemically mediated tri-trophic interactions: a 1H-NMR network approach to identify compound structural features and associated biological activity, Front. Plant Sci. 9 (2018) 1–12.
[29] L.-F. Nothias, M. Nothias-Esposito, R. da Silva, M. Wang, I. Protsyuk, Z. Zhang, A. Sarvepalli, P. Leyssen, D. Touboul, J. Costa, J. Paolini, T. Alexandrov, M. Litaudon, P.C. Dorrestein, Bioactivity-based molecular networking for the discovery of drug leads in natural product bioassay-guided fractionation, J. Nat. Prod. 81 (2018) 758–767.
[30] G. D'Urso, C. Pizza, S. Piacente, P. Montoro, Combination of LC–MS based metabolomics and antioxidant activity for evaluation of bioactive compounds in *Fragaria vesca* leaves from Italy, J. Pharm. Biomed. Anal. 150 (2018) 233–240.
[31] O.M. Kvalheim, H. Yan Chan, I.F.F. Benzie, Y. tong Szeto, A.H. Chung Tzang, D.K. wah Mok, F. Tim Chau, Chromatographic profiling and multivariate analysis for screening and quantifying the contributions from individual components to the bioactive signature in natural products, Chemometr. Intell. Lab. Syst. 107 (2011) 98–105.
[32] J. Xu, Q.S. Xu, C.O. Chan, D.K.W. Mok, L.Z. Yi, F.T. Chau, Identifying bioactive components in natural products through chromatographic fingerprint, Anal. Chim. Acta 870 (2015) 45–55.
[33] J.J. Kellogg, D.A. Todd, J.M. Egan, H.A. Raja, N.H. Oberlies, O.M. Kvalheim, N.B. Cech, Biochemometrics for natural products research: comparison of data analysis approaches and application to identification of bioactive compounds, J. Nat. Prod. 79 (2016) 376–386.
[34] L.K. Caesar, J.J. Kellogg, O.M. Kvalheim, R.A. Cech, N.B. Cech, Integration of biochemometrics and molecular networking to identify antimicrobials in *Angelica keiskei*, Planta Med. 84 (2018) 721–728.
[35] E.R. Britton, J.J. Kellogg, O.M. Kvalheim, N.B. Cech, Biochemometrics to identify synergists and additives from botanical medicines: a case study with *Hydrastis canadensis* (goldenseal), J. Nat. Prod. 81 (2018) 484–493.
[36] G. Xie, M. Ye, Y. Wang, Y. Ni, M. Su, H. Huang, M. Qiu, A. Zhao, X. Zheng, T. Chen, W. Jia, Characterization of pu-erh tea using chemical and metabolic profiling approaches, J. Agric. Food Chem. 57 (2009) 3046–3054.
[37] Y. Fujimura, K. Kurihara, M. Ida, R. Kosaka, D. Miura, H. Wariishi, M. Maeda-Yamamoto, A. Nesumi, T. Saito, T. Kanda, K. Yamada, H. Tachibana, Metabolomics-driven nutraceutical evaluation of diverse green tea cultivars, PLoS One 6 (2011) 3–16.
[38] K.M. Chan, G.G.L. Yue, P. Li, E.C.W. Wong, J.K.M. Lee, E.J. Kennelly, C.B.S. Lau, Screening and analysis of potential anti-tumor components from the stipe of *Ganoderma sinense* using high-performance liquid chromatography/time-of-flight mass spectrometry with multivariate statistical tool, J. Chromatogr. A 1487 (2017) 162–167.
[39] D.M. Kulakowski, S.B. Wu, M.J. Balick, E.J. Kennelly, Merging bioactivity with liquid chromatography-mass spectrometry-based chemometrics to identify minor immunomodulatory compounds from a Micronesian adaptogen, *Phaleria nisidai*, J. Chromatogr. A 1364 (2014) 74–82.
[40] D.A. Chagas-Paula, T. Zhang, F.B. da Costa, R.A. Edrada-Ebel, A metabolomic approach to target compounds from the *Asteraceae* family for dual COX and LOX inhibition, Metabolites 5 (2015) 404–430.
[41] Y. Chen, J. Luo, Q. Zhang, L. Kong, Identification of active substances for dually modulating the renin-angiotensin system in *Bidens pilosa* by liquid chromatography-mass spectrometry-based chemometrics, J. Funct. Foods 21 (2016) 201–211.
[42] A. Patras, N.P. Brunton, G. Downey, A. Rawson, K. Warriner, G. Gernigon, Application of principal component and hierarchical cluster analysis to classify fruits and vegetables commonly consumed in Ireland based on in vitro antioxidant activity, J. Food Compos. Anal. 24 (2011) 250–256.
[43] X. Li, Q. Wu, Y. Xie, Y. Ding, W.W. Du, M. Sdiri, B.B. Yang, Ergosterol purified from medicinal mushroom *Amauroderma rude* inhibits cancer growth *in vitro* and *in vivo* by up-regulating multiple tumor suppressors, Oncotarget 6 (2015) 17832–17846.
[44] J. Chong, J. Xia, MetaboAnalystR: an R package for comprehensive analysis of metabolomics data, Bioinformatics 34 (2018) 4313–4314.
[45] B. Falissard, psy: various procedures used in psychometry. R package version 1.1. https://CRAN.R-project.org/package=psy, 2012.
[46] S. Kucheryavskiy, mdatools: multivariate data analysis for chemometrics. R package version 0.9.1. https://CRAN.R-project.org/package=mdatools, 2018.
[47] H. Wickham, J. Hester, R. Francois, readr: read rectangular text data. R package version 1.1.1. https://CRAN.R-project.org/package=readr, 2017.
[48] W. Revelle, Psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, 2018. R package version 1.8.10, https://CRAN.R-project.org/package=psych.
[49] L. Margueritte, P. Markov, L. Chiron, J.-P. Starck, C. Vonthron-Sénécheau, M. Bourjot, M.-A. Delsuc, Automatic differential analysis of NMR experiments in complex samples, Magn. Reson. Chem. 56 (2018) 469–479.
[50] L. Chiron, M.-A. Coutouly, J.-P. Starck, C. Rolando, M.-A. Delsuc, SPIKE a Processing Software Dedicated to Fourier Spectroscopies, ArXiv., 2016, p. 1608, 06777.
[51] E. Saccenti, H.C.J. Hoefsloot, A.K. Smilde, J.A. Westerhuis, M.M.W.B. Hendriks, Reflections on univariate and multivariate analysis of metabolomics data, Metabolomics 10 (2014) 361–374.
[52] B. Falissard, Focused principal component analysis: looking at a correlation matrix with a particular interest in a given variable, J. Comput. Graph. Stat. 8 (1999) 906–912.
[53] P.S. Gromski, H. Muhamadali, D.I. Ellis, Y. Xu, E. Correa, M.L. Turner, R. Goodacre, A tutorial review: metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding, Anal. Chim. Acta 879 (2015) 10–23.
[54] O.M. Kvalheim, T. V Karstang, Interpretation of latent-variable regression models, Chemometr. Intell. Lab. Syst. 7 (1989) 39–51.
[55] B. Worley, R. Powers, Multivariate analysis in metabolomics, Curr. Metabolomics. 1 (2015) 92–107.
[56] Y.H. Yun, B.C. Deng, D.S. Cao, W.T. Wang, Y.Z. Liang, Variable importance analysis based on rank aggregation with applications in metabolomics for biomarker discovery, Anal. Chim. Acta 911 (2016) 27–34.
[57] M. Farrés, S. Platikanov, S. Tsakovski, R. Tauler, Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation, J. Chemom. 29 (2015) 528–536.
[58] R.A. Van den Berg, H.C. Hoefsloot, J.A. Westerhuis, A. k. Smilde, M.J. van der Van derWerf1, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genomics 7 (2006) 1–15.
[59] H.-W. Cho, S.B. Kim, M.K. Jeong, Y. Park, N.G. Miller, T.R. Ziegler, D.P. Jones, Discovery of metabolite features for the modelling and analysis of high-resolution NMR spectra, Int. J. Data Min. Bioinform. 2 (2008) 176–192.
[60] V. Prakash, P.K. Nandi, Interaction of amino acids, N-acetyl amino acid esters, thymine and adenine with Sephadex LH-20 gel, J. Chromatogr. 106 (1975) 23–31.
[61] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Assessment of PLSDA cross validation, Metabolomics 4 (2008) 81–89.
[62] H.Y. Kim, M.Y. Lee, H.M. Park, Y.K. Park, J.C. Shon, K.-H. Liu, C.H. Lee, Urine and serum metabolite profiling of rats fed a high-fat diet and the anti-obesity effects of caffeine consumption, Molecules 20 (2015) 3107–3128.
[63] S. Huang, H. Chen, W. Li, X. Zhu, W. Ding, C. Li, Bioactive chaetoglobosins from the mangrove endophytic fungus *Penicillium chrysogenum*, Mar. Drugs 14 (2016) 1–12.
[64] H. Ali, M. Ries, J. Nijland, P. Lankhorst, T. Hankemeier, R. Bovenberg, R. Vreeken, A. Driessen, A branched biosynthetic pathway is involved in production of roquefortine and related compounds in *Penicillium chrysogenum*, PLoS One 8 (2013) 1–12.
[65] M.S. Mady, M.M. Mohyeldin, H.Y. Ebrahim, H.E. Elsayed, W.E. Houssen, E.G. Haggag, R.F. Soliman, K.A. El Sayed, The indole alkaloid meleagrin, from the olive tree endophytic fungus *Penicillium chrysogenum*, as a novel lead for the control of c-Met-dependent breast cancer proliferation, migration and invasion, Bioorg. Med. Chem. 24 (2016) 113–122.
[66] Z. Shang, X. Li, L. Meng, C. Li, S. Gao, C. Huang, B. Wang, Chemical profile of the secondary metabolites produced by a deep-sea sediment-derived fungus *Penicillium commune* SD-118, Chin. J. Oceanol. Limnol. 30 (2012) 305–314.