Reconstruction of HMBC Correlation Networks: A Novel NMR-based Contribution to Metabolite Mixture Analysis

Ali Bakiri, Jane Hubert, Romain Reynaud, Carole Lambert, Agathe Martinez, Jean- Hugues Renault, Jean-Marc Nuzillard

1 **Reconstruction of HMBC Correlation Networks: A Novel NMR-based Contribution to**

2 **Metabolite Mixture Analysis**

3

4 Ali Bakiri,[1,2] Jane Hubert,[1,*] Romain Reynaud,[2] Carole Lambert, [2] Agathe Martinez,[1] Jean-

5 Hugues Renault,[1] and Jean-Marc Nuzillard[1]

6

7 [1] Institut de Chimie Moléculaire de Reims, UMR CNRS 7312, SFR CAP'SANTE, Université de

8 Reims Champagne-Ardenne, Reims, France

9 [2] Givaudan France, Active Beauty Department, Pomacle, France

10

11 *e-mail: jane.hubert@univ-reims.fr

1   **Abstract**

2   Here is introduced a new *in silico* method for the dereplication of natural metabolite mixtures

3   based on HMBC and HSQC spectra that inform about short-range and long-range H-C

4   correlations occurring in the carbon skeleton of individual chemical entities. Starting from the

5   HMBC spectrum of a metabolite mixture, an algorithm was developed in order to recover

6   individualized HMBC footprints of the mixture constituents. The collected H-C correlations are

7   represented by a network of NMR peaks connected each other when sharing either a ^{1}H or

8   ^{13}C chemical shift value. The obtained network is then divided into clusters using a community

9   detection algorithm, and finally each cluster is tentatively assigned to a molecular structure

10   by mean of a NMR chemical shift database containing the theoretical HMBC and HSQC

11   correlation data of a range of natural metabolites. The proof of principle of this method is

12   demonstrated on a model mixture of 3 known natural compounds and then on a real-life bark

13   extract obtained from the common spruce (*Picea abies* L.).

1 **Graphical abstract**

2

3
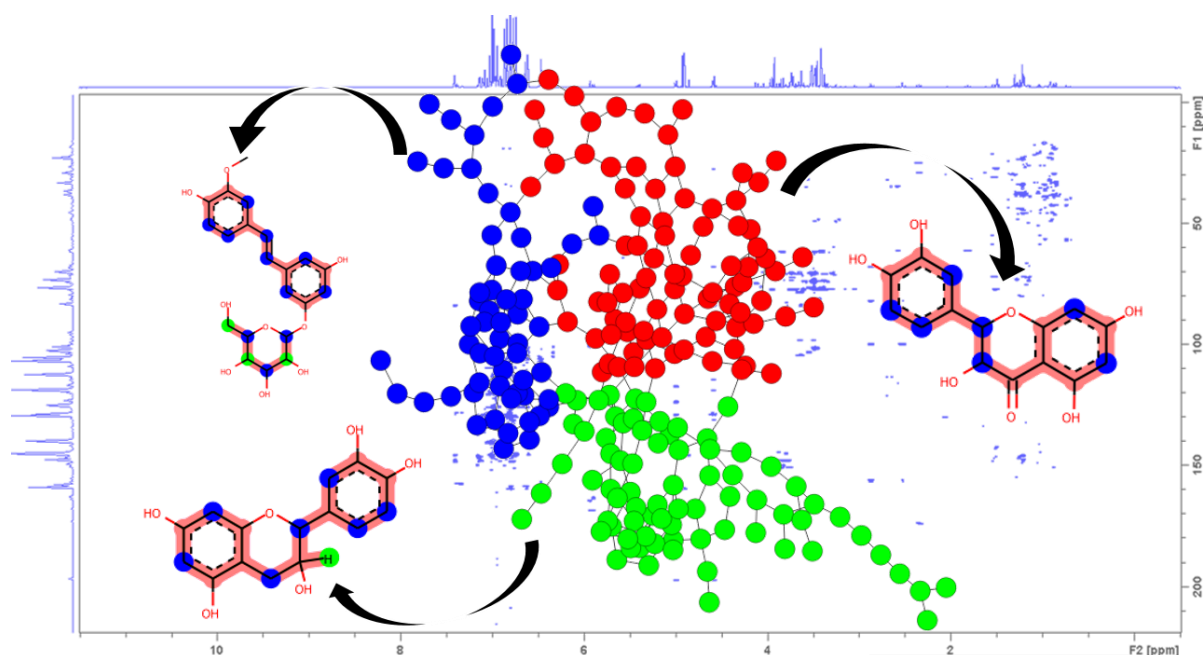
4 **INTRODUCTION**

5 Identification and quantification of chemical entities directly within complex mixtures is

6 gaining a great attention in the field of natural product research and more generally in all

7 metabolomics studies dealing with the analysis of low molecular weight metabolites in

8 biological systems (Wist, 2016)(Gaudêncio & Pereira, 2015)(Bingol et al., 2016) (Kim, Choi, &

9 Verpoorte, 2011). Over the last decade, a range of new analytical strategies have emerged to

10 accelerate metabolite identification processes, mainly based on the so-called dereplication

11 concept (Beutler, Alvarado, Schaufelberger, Andrews, & McCloud, 1990)ref revue wolfender

12 (Gaudêncio & Pereira, 2015; Hubert, Nuzillard, & Renault, 2015). Among the panel of

13 analytical techniques available, Nuclear Magnetic Resonance (NMR) remains by far the most

14 efficient method to unambiguously identify small organic molecules (Smolinska, Blanchet,

15 Buydens, & Wijmenga, 2012)(Schripsema, 2010). NMR has the unique ability to probe

1   structural information at the atom level for NMR-active nuclei such as ^{1}H, ^{13}C, ^{15}N and ^{31}P.

2   Proton (^{1}H nucleus) NMR spectroscopy is mostly used due to its high sensitivity, but suffers

3   from frequent signal overlaps which confuse data interpretation when dealing with samples

4   of complex composition. Conversely, ^{13}C NMR is less sensitive than ^{1}H NMR, nonetheless it

5   provides a larger chemical shift dispersion (240 ppm) that reduces signal overlaps and leads

6   to well-resolved spectra containing individualized narrow peaks. Along with the ability to

7   obtain information about the carbon skeleton of all detected organic molecules, these

8   advantages have progressively prompted ^{13}C NMR in the core of recent dereplication

9   procedures (Clendinen et al, 2014; Laude et al, 1986; Bighelli et al, 1994; Hubert et al, 2014).

10   In addition to ^{1}H and ^{13}C NMR, 2D NMR spectra such as HMBC (Heteronuclear Multiple Bond

11   Correlation spectroscopy) and HSQC (Heteronuclear Single Quantum Correlation

12   spectroscopy) are extremely informative. Both provide important information about atom

13   connectivity. Such spectra have also attracted an increasing interest in recent dereplication

14   and metabolomics studies   (Bingol & Bruschweiler, 2013 + ref pauli hsqc mask actae

15   triterpenes)

16   The second breakthrough in mixture analysis strategies is the application of mathematical

17   treatments to facilitate the interpretation of complex analytical data (Wiklund, n.d.). Along

18   with the multivariate statistical analyses classically used in metabolomics, new mathematical

19   methods such as data fusion (Acar, Bro, & Smilde, 2015) and graph theory are more and more

20   applied to rise up the relevance of complex mixture studies (Garg et al., 2015). For instance,

21   molecular networking has recently emerged in the context of mass spectrometry (MS) as an

22   efficient way to handle the complexity of MS/MS data. It relies on the observation that

23   structurally similar metabolites share similar MS/MS fragmentation patterns. On the basis of

24   this property, the correlations between MS/MS spectral data of a natural sample are

1 measured and used to construct a molecular network in which each cluster of similar

2 fragments represents a specific group of structurally close metabolites. (Winnikoff, Glukhov,

3 Watrous, Dorrestein, & Gerwick, 2014) (Allard et al., 2016; Yang et al., n.d.)

4 Here is presented a new HMBC-based dereplication method that uses a networking approach

5 for the deconvolution of complex NMR spectra of metabolite mixtures. This method exploits

6 the ability of HMBC experiments to provide connectivity information between ^{1}H and ^{13}C

7 atoms located at their vicinity. Starting from the HMBC spectrum of a metabolite mixture, the

8 strategy aims to build the network of ^{1}H-^{13}C HMBC correlations and to highlight individualized

9 patterns representing specific molecular fragments or even whole chemical structures. The

10 isolated patterns are then identified by database search.

11 The proof of concept of this HMBC-based atom networking approach is demonstrated here

12 on a synthetic model mixture of 3 known natural compounds: (+)-catechin, taxifolin and *E*-

13 isorhapontin, and then on a genuine extract obtained from the bark extract of *Picea abies.*

14

15 **METHODS**

16 **Sample preparation.** The first analyzed sample was a model mixture composed of 3.9 mg of

17 taxifolin, 0.4 mg of (+)-catechin, and 0.7 mg of *E*-isorhapontin (all purchased from Sigma-

18 Aldrich, Steinheim, Germany). The three compounds were directly dissolved in 550 μL of

19 deuterated dimethyl sulfoxide (DMSO-*d6*, Eurisotop, Saint-Aubin, France). The second sample

20 was a crude methanol extract obtained from the barks of *P. abies*, prepared as described

21 previously (Angelis et al., 2016). Briefly, 2 kg of barks were collected in the French Champagne-

22 Ardenne territory in 2014 within the context of forestry activities (Voucher specimen JH-2014-

23 9, Faculty of Pharmacy, University of Reims Champagne-Ardenne, France). After drying at

1 30 °C for 72 h, the barks were powdered in a hammer mill and extracted with methanol under

2 magnetic stirring at ambient temperature for 20 h. After filtration, methanol was evaporated

3 to dryness and an aliquot of the resulting extract (20 mg) was dissolved in 550 μL of DMSO-*d6*

4 for NMR analyses.

5 **NMR analysis and peak picking.** NMR analyses were performed at 298 K on an Avance III-600

6 spectrometer (Bruker, Karlsruhe, Germany) equipped with a cryoprobe optimized for [1]H

7 detection and with cooled ^{1}H, ^{13}C and ^{2}H coils preamplifiers. HSQC and HMBC spectra were

8 recorded using standard Bruker pulse programs. For the acquisition of HSQC spectra (pulse

9 sequence hsqcedetgpsisp2.2), the size of FID was set at 1024 in the F2 dimension and 512 in

10 the F1 dimension, and the number of scans was set at 4. For the acquisition of HMBC spectra

11 (pulse sequence hmbcetgpl3nd), the size of FID was set at 2048 in the F2 dimension and 512

12 in the F1 dimension, and the number of scans was 4. The spectra were manually phased and

13 baseline corrected using the TOPSPIN 3.2 software (Bruker). Once the HMBC spectra recorded

14 and processed, a semi-automated peak picking was performed to obtain the list of

15 correlations. The peak picking was performed on the different regions of the spectrum first

16 automatically, then manually in order to remove possibly meaningless peaks caused by

17 thermal noise, t_1 noise, or residual ^1J peak pairs. The automatic peak picking process was based

18 on the minimum intensity set to the lowest contour level (Bruker processing window in

19 Topspin 3.2).

20 **Prediction of HMBC and HSQC data**. A database of theoretical HMBC connectivity data of a

21 range of natural metabolites was created as follows: the ^{1}H and ^{13}C NMR chemical shift values

22 of ≈ 2700 natural metabolites (October 2017) were calculated using the NMR Workbook Suite

23 ACD/Labs software (Ontario, Canada). From these data, HMBC correlations (pairs of ^{1}H and

1   ^{13}C NMR chemical shifts) were generated using a locally written Python program. For each

2   molecule of the database, this program creates a list of all H-C atom pairs that might be visible

3   in an HMBC spectrum, *i.e.* all H-C pairs in which atoms are placed 2 or 3 bonds away. The

4   correlation between chemical shifts of atoms separated by 4 bonds are not observed in all

5   molecular structures and were therefore not taken into account. Once this list is established,

6   the corresponding predicted chemical shifts values of each atom pair are included in the

7   database (**Fig. 1**). In the same way, the HSQC correlations of all metabolites of the database

8   were determined for the ^{1}H-^{13}C atom pairs located at a distance of 1 bond. In addition to the

9   ^{1}H and ^{13}C chemical shifts values, the parity of the number of H atoms attached to each C atom

10   was also considered. Thus, an additional column was added to the HSQC correlation table in

11   order to store the parity data.

12

Figure 1| Construction of the theoretical 2D NMR database from the predicted ¹H and ¹³C NMR chemical shifts of natural metabolites. 1. Listing of all H-C atom pairs separated by 2 or 3 bonds. 2. Determination of the theoretical HMBC correlations list. 3. Addition of the created list to the database.
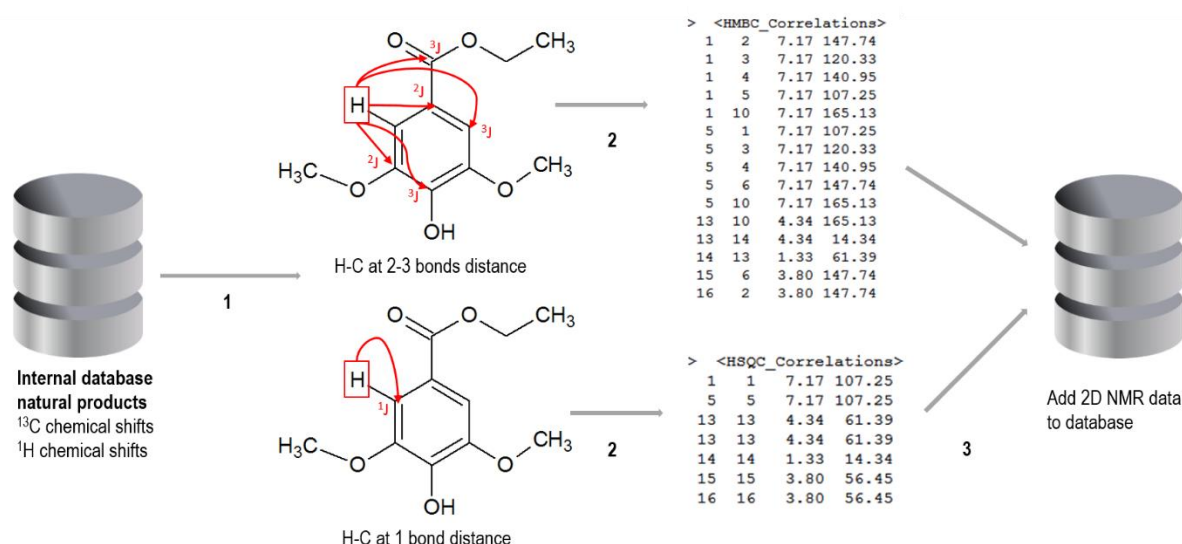
1 **From experimental HMBC to ¹H-¹³C chemical shift network.** The ¹H-¹³C chemical shift

2 networks were generated from the peak list of the experimental HMBC spectra using an

3 algorithm written in Python. The principle of this algorithm is to extract the HMBC correlation

4 data from the experimental HMBC peak list and then to create a network edge list by

5 connecting each peak (vertex) to its adjacent neighbors on the same row of the HMBC

6 spectrum in the F1 dimension (having the same ¹³C NMR chemical shift) or on the same

7 column in the F2 dimension (having the same ¹H NMR chemical shift). The obtained edge list

8 is then converted to a network graph using the igraph 0.7.1 package for Python 2.7. The

9 clustering of network vertices was performed using the Louvain community detection

10 algorithm (Blondel et al.). The Louvain method is a modularity optimization-based algorithm.

11 Modularity is a quality function that measures the strength of connectivity between the nodes

of the same community. It measures the density of edges inside the community compared to

edges outside the community. The modularity is defined as followed (J. Newman):

$$Q = \sum (e_{ii} - a_i^2)$$

e_{ii} : The fraction of edges that connects vertices from the same group.

e_{ij} : The fraction of edges that connects vertices from group i to vertices from group j and $a_i = \sum e_{ij}$

The Louvain algorithm is a bottom-up algorithm that starts with isolated vertices as

communities that will be merged progressively into meta-communities according to the gain

of modularity until an optimal graph partition is reached (Blondel et al.). The "RBER" method

is a variant of the Louvain algorithm that uses Erdös-Rényi null-model (Newman, M et al 2004),

in which each edge has the same probability of appearing. It includes a resolution parameter

that allows the control of the community's size, in general, a higher resolution parameter

leads to smaller communities and vice versa. The RBER method was applied for the detection

of the HMBC peak clusters in HMBC networks, with a resolution parameter set at 0.2. This

value was selected by trial and error and found to give the best results.

Metabolite identification. Metabolite identification was performed by comparing the

theoretical HMBC correlations of the database metabolites to the experimental HMBC

correlations of the clusters extracted by network analysis. For this purpose, an algorithm was

written in Python language to obtain the lists of experimental correlations corresponding to

each cluster of the HMBC correlation network. Then, these lists of correlations are compared

to the theoretical HMBC correlations of database metabolites. For each theoretical HMBC

correlation of the database, the algorithm attempts to find an experimental correlation that

1 matches both ^{1}H and ^{13}C chemical shift values within a defined tolerance radius, typically set

2 at 1.5 ppm for ^{13}C NMR chemical shifts and 0.15 ppm ^{1}H NMR chemical shifts. The percentage

3 of theoretical correlations that match experimental correlations for each metabolite of the

4 database is then calculated to evaluate its likeliness to be present within the analyzed mixture.

5 Finally, the algorithm returns a list of metabolites for each cluster, ranked in the decreasing

6 order of matching percentage (**Fig. 2**). As signals belonging to a putatively identified

7 metabolite may be divided between several clusters, a second step is then performed to

8 compare the theoretical HMBC correlations of the top ranked 20 molecules of each cluster to

9 the complete experimental list of HBMC correlations. By this way, an improved list of

10 molecules ranked according to their decreasing percentage of matching to the whole HMBC

11 spectrum is obtained. The resulting list is submitted to an additional validation step which

12 consists in the comparison of the theoretical HSQC correlations of the top ranked molecules

13 to the HSQC correlations detected experimentally. This step is performed using another locally

14 developed algorithm which compares the ^{1}H and ^{13}C chemical shifts values of the HSQC

15 correlations together with their parity, therefore theoretical correlations are matched to the

16 experimental one only if both their chemical shifts and parity match. From this step, an

17 additional score (percentage of theoretical HSQC correlations that match experimental HSQC

18 correlations) is obtained for each proposed metabolite. This score reinforces the identification

19 accuracy.

20

Figure 2 | HMBC atom networking workflow for metabolite mixture identification.

2

RESULTS

3
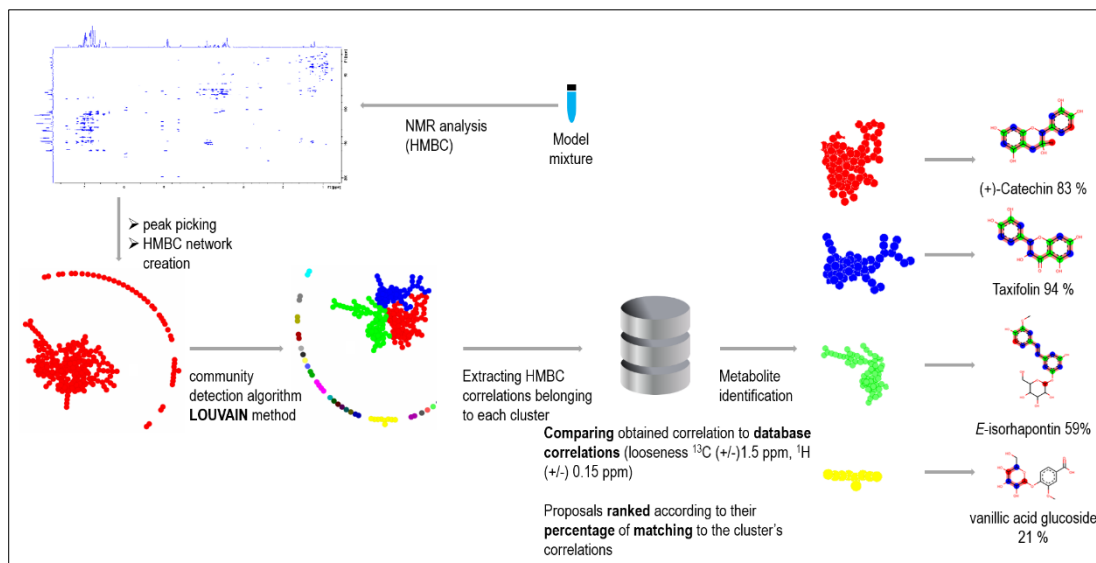
Proof of concept on a model mixture. A mixture of 3 commercially available compounds (2

flavonoids and 1 stilbene), taxifolin, (+)-catechin and *E*-isorhapontin, was used as a model in order to

establish a first proof of concept of the method. The proportions of the mixture were chosen so that a

molar ratio of about 10 was set between the most and the least abundant mixture components, thus

giving an indication of the concentration dynamics that can possibly be handled. After NMR analysis of

the mixture, a semi-automated peak picking was performed to collect all ^{1}H-^{13}C HMBC connectivity

data from the experimental spectrum. The resulting list of ^{1}H and ^{13}C chemical shift pairs was exported

as a text file and used as input to create the HMBC-derived correlation network. A network of 267

nodes representing ^{1}H-^{13}C HMBC correlations was obtained (**Fig. 3**). The nodes sharing either a ^{1}H or a

^{13}C chemical shift value were connected by an edge. As the analyzed model mixture was composed of

3 molecules, the HMBC correlation network was expected to form 3 non-connected sub-networks

representing each of the three molecules. However, only a big sub-network of 222 nodes was obtained,

along with a small sub-network of 9 nodes and further sub-networks of 1 or 2 nodes. Non-expected

interconnections were in fact due to the presence of wide, overlapping ^{1}H multiplet signals and of some identical chemical shift values between the three molecules. The major sub-network of 222 nodes was supposed to correspond to the three molecules of the model mixture. In an attempt to more efficiently isolate chemical shift clusters corresponding to individual molecule, a community detection algorithm was used. After the testing of several community detection approaches (Porter, Mason A et al), the Louvain method using the RBER method, which includes a resolution parameter, was found to be the most appropriate: i) the network is unweighted and undirected and the modularity based algorithms are the most appropriate to detect communities in this type of networks and ii) the resolution parameter allows the control of the communities size. By contrast, the default modularity algorithm tends to form very small communities that do not contain enough information for the identification of the metabolites. Applying this method with a resolution parameter set at 0.2, four main clusters (C1-C4) were obtained from the initial HMBC network (**Fig. 3**). The experimental ^{1}H-^{13}C HMBC correlations of each cluster were then compared to the theoretical ^{1}H-^{13}C HMBC correlations of the database metabolites. As a result, four lists of molecules having the highest likelihood to correspond to each cluster were obtained (**Table 1**). The proposed molecules were sorted according to their decreasing score value which represent the percentage of matching between experimental and theoretical HMBC correlations of each molecule. For cluster C1, the list of proposed molecules only contained flavan-3-ols (**Table 1**). The highest score was obtained for (+)-catechin (83 %) which was indeed one of the three constituents of the model mixture. For cluster C2, a list of dihydroflavonols was also obtained (**Table 1**), the highest score being attributed to taxifolin (94 %). For cluster C3, a list of 3 stilbenes and one tri-galloyl glucose was obtained (**Table 1**) with the highest score attributed to *E*-isorhapontin (54 %). This score was low as compared to the metabolites proposed for the two previous clusters, because the chemical structure of *E*-isohrapontin contains a glucose moiety and a methoxy group that do not share so many correlation with the aglycone moiety, so that the community detection algorithm affects the signals of the two moiety in separate communities. In addition, some correlations that would have allowed the connection of the communities were absent from the peak

1   list. Cluster C4 was a small cluster containing just a few correlation values characteristic of a sugar

2   moiety (**Table 1**). This cluster corresponded to the glucose part of *E*-isorhapontin. A second comparison

3   was performed for the 20 top ranked molecules in each of the clusters C1-C4. In this case, all theoretical

4   HMBC correlation values of the proposed molecules were compared to the whole experimental HMBC

5   peak list. This step allowed to confirm the presence of taxifolin, (+)-catechin and *E*-isorhapontin in the

6   model mixture with score values increased up to 97 % for taxifolin, 93 % for (+)-catechin and 70 % for

7   the *E*-isorhapontin.

8   A final search was then performed to check the matching between the theoretical HSQC

9   correlations of the top ranked molecules to the HSQC chemical shift values collected in the

10   experimental spectrum. HSQC data inform about one-bond H-C correlations and therefore

11   provide information highly complementary to HMBC data to improve the identification

12   accuracy. Taxifolin, (+)-catechin and *E*-isorhapontin achieved all an HSQC score value of 100 %,

13   thus providing an additional proof of their presence in the model mixture.

14

Figure 3 | HMBC-network obtained after NMR analysis of the model mixture and
clustering of the vertices. Colors represent the different clusters (Red C1, Blue C2, Green C3,
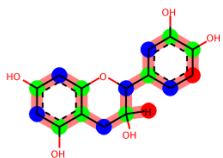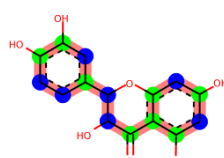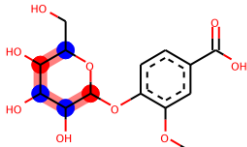Yellow C4).
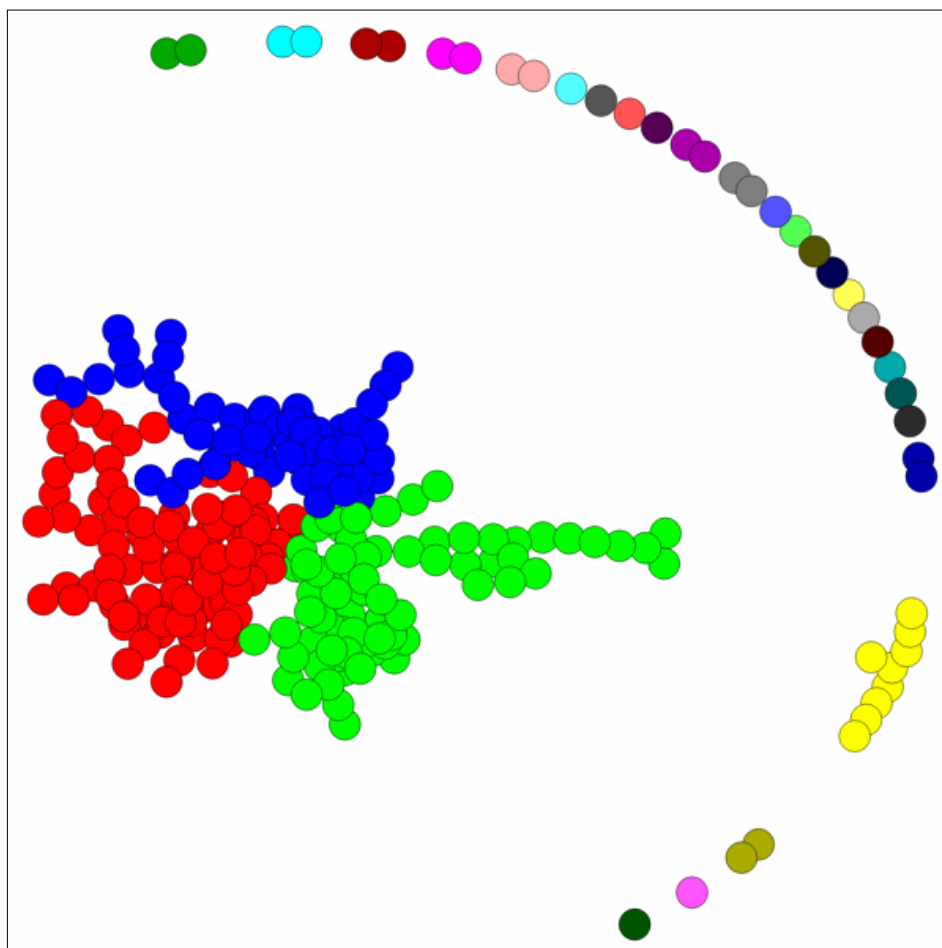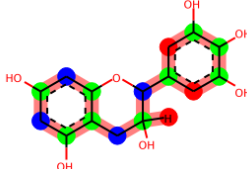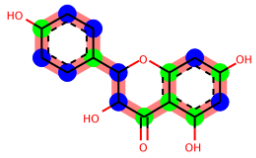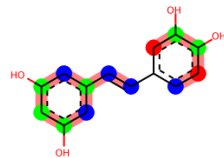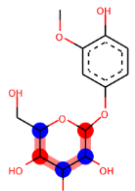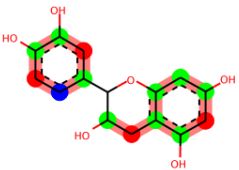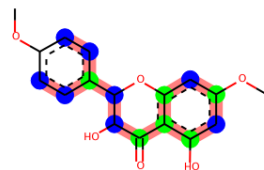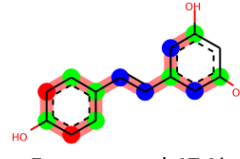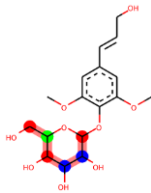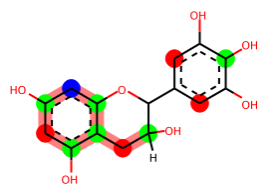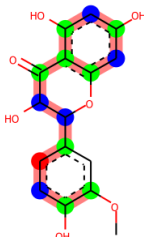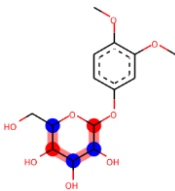
1

2

3

4

Table 1 | Proposed list of metabolites obtained for the model mixture. Colors highlight atoms for which the chemical shifts values of the experimental HMBC peak list matched the HMBC predicted data; Pink: matching molecular features, Blue: matching both ^{1}H and ^{13}C chemical shifts, Green: matching only ^{13}C chemical shifts, Red: Matching only ^{1}H chemical shifts.

Cluster 1 (C1)	Cluster 2 (C2)	Cluster 3 (C3)	Cluster 4 (C4)
(+)-catechin 83 %	taxifolin 94 %	*E*-isorhapontin 54 %	vanillic acid glucoside 21 %
gallocatechin 58 %	dihydrokaempferol 81 %	astringinin 52 %	tachioside 20 %
(-)-epicatechin 49 %	aromadendrin7'4'dimethyl_ether 71 %	*E*-resveratrol 47 %	syringin 20 %
(-)-epigallocatechin 46 %	3'-*O*-methyltaxifolin 64 %	1,2,6-Tri-*O*-galloylglucose 44 %	3,4-dimethoxyphenyl-D-glucopyranoside 19 %

1

2

Applicability to an authentic natural extract. The pertinence of the dereplication workflow based on HMBC correlation networking was then evaluated on a crude methanol bark extract obtained from *P. abies*, a common tree growing in temperate climate regions . Fractionation by centrifugal partition chromatography as well as NMR-based chemical profiling of this extract were previously described, revealing the presence of compounds belonging to various chemical classes including stilbenes, flavonoids or phenolic acids (Angelis et al., 2016). In view of this chemical diversity, this extract was considered to be a suitable model for the testing of the real life applicability of the method. A semi-automated peak picking was performed to collect all $^1$H-^{13}C HMBC connectivity data from the experimental HMBC spectrum of the extract. The resulting peak list was used to create a HMBC correlation network in the same way as described for the model mixture. A network of 642 nodes was obtained, containing one large sub-network of 556 nodes and two small sub-networks of 11 and 8 nodes, respectively. The other sub-networks contained less than 5 nodes. Applying the Louvain community detection algorithm to this network with the resolution parameter set at 0.2, six clusters were obtained (C1-C6) (**Fig. 4**). The experimental HMBC data of these clusters were automatically compared to the theoretical HMBC data of the database metabolites. For Cluster C1, a list of highly similar stilbenes was obtained, the top ranked molecules of the list were *E*-resveratrol, *E*-piceid, *E*-isorhapontin, and *E*-astringinin with high score values of 76 %, 76 %, 73 % and 71 %, respectively. All these metabolites shared a *E*-resveratrol skeleton (5-[(1E)-2-(4-Hydroxyphenyl) ethenyl]-1,3-benzenediol), either linked or not to a glycosidic moiety, suggesting a highly likely presence of these molecules in the extract. For cluster C2, a list of sugars was proposed by the database (**Table 2**), with the highest score obtained for saccharose (63 %). A list of terpenoids and tocopherols was proposed for cluster C3, but the obtained score values were relatively poor, all below 50 % at this stage of the procedure. For cluster C4,

16

1 a list of flavonoids was obtained, the top ranked molecules were (+)-catechin and taxifolin

2 derivatives with (+)-catechin having the highest score (52 %). Cluster C5 was a small cluster of

3 only 11 HMBC correlations for which a list of various sugars with low score values was

4 obtained. This cluster contained in fact the HMBC correlation values of either a simple sugar

5 or the sugar moiety of a glycosylated molecule. Cluster C6 contained only 8 correlations that

6 did not allow prioritizing any relevant result.

7

Figure 4| HMBC-network obtained after NMR analysis of the bark extract of *P. abies*. Colors

represent the different clusters: red C1, green C2, blue C3, yellow C4, pink C5 and light blue

C6.

8

9
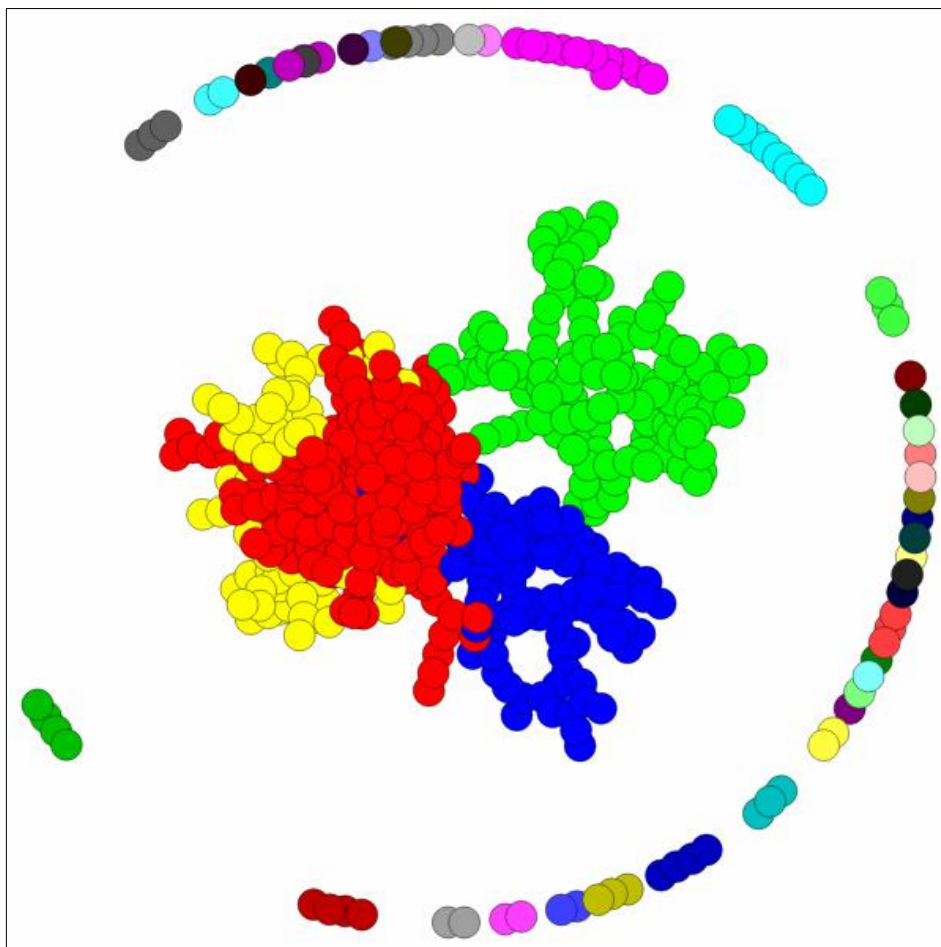
1   The search for clusters of HMBC correlations resulted in the identification of specific groups

2   of molecules. However, as the HMBC fingerprint of the molecules contained in the extract

3   were in some cases divided between several clusters, a second database search was

4   performed for the 20 top ranked molecules of each cluster to compare their theoretical HMBC

5   correlations to the total experimental HMBC correlations. As a result, a new list of molecules

6   was obtained with improved scores, mainly containing stilbenes, flavonoids and sugars. The

7   new scores of the top ranked molecules of cluster C1 increased, putting the glycosylated

8   molecules in the top of the list with score values of 95 % for *E*-isorhapontin, 94 % for *E*-piceid,

9   91 % for *E*-astringin, and 92 % for *E*-rhaponticin. The scores of stilbene aglycones increased up

10 to 81 % for *E*-resveratrol, 82 % for *E*-astringinin and 79 % *E*-rhapontigenin. The score of

11 saccharose identified from cluster C2 also increased up to 74 %, and the score of 3,4-

12 dimethoxyphenyl-β-*D*-glucopyranoside, that was initially ranked at 12th position in the cluster

13 C2, arrived after the second step with a score value of 84 %. For cluster C3, the scores of the

14 proposed molecules did not increased significantly after the second step of the database

15 search, all remaining below 60 %. The score value of (+)-catechin identified in cluster C4 was

16 increased up to 81 %, and other flavonoids including taxifolin, dihydrokaempferol, and

17 taxifolin 3'-*O*- β-D-glucopyranoside arrived also in the top ranked molecules with score values

18 of 87 %, 86 % and 85 %, respectively.

19 A final search was then performed to check the matching between the theoretical HSQC

20 correlations of the top ranked molecules to the HSQC chemical shift values collected in the

21 experimental spectrum. The final list of proposed molecules is given in **Table 2**, with fourteen

22 molecules achieving a score value of 100 %.

1   In summary, among the top ranked molecules obtained over the three step of our

2   identification procedure, fourteen molecules including *E*-isorhapontin, *E*-astringin, *E*-piceid,

3   saccharose, (+)-catechin, taxifolin, taxifolin 3'-*O*- β-D-glucopyranoside, 3,4-dimethoxyphenyl-

4   β-D-glucopyranoside, *E*-resveratrol, *E*-piceatannol, *E*-rhapontigenin, *E*-rhaponticin, maltose,

5   and dihydrokaempferol were putatively identified in the methanol bark extract of *P. abies*

6   (**Table 2**). In order to unambiguously validate the presence of these molecules in the extract,

7   their 1D and 2D NMR data were manually checked by going back to the experimental 1D and

8   2D NMR spectra. The presence of eight compounds was fully confirmed (**Fig. 6**), namely *E*-

9   isorhapontin, *E*-astringin, *E*-piceid, saccharose, (+)-catechin, taxifolin, taxifolin 3'-*O*-β-D-

10   glucopyranoside, and 3,4-dimethoxyphenyl-β-D-glucopyranoside. *E*-resveratrol was easily

11   eliminated from the list of hits because no HMBC correlation corresponding to the highly

12   characteristic symmetrical part of the 1,3-benzenediol moiety of the resveratrol carbon

13   skeleton was detected, a correlation that is always present in all the spectra of compounds

14   bearing this structural feature and that is due to the existence of a strong 3J coupling (*J. Am.*

15   *Chem. Soc.*, 1969, 91 (17), pp 4940–4941, pas mieux…). Similarly, *E*-piceatannol and *E*-

16   rhapontigenin were also eliminated from the list of hits because no HMBC correlation

17   corresponding to the symmetrical part of the 1,3-benzenediol moiety was detected by NMR.

18   Concerning *E*-rhaponticin, which is a positional isomer of *E*-isorhapontin, the manual checking

19   of experimental NMR data for the validation of *E*-isorhapontin enabled to discard *E*-

20   rhaponticin from the list of hits thanks to three main indices (**Fig. 5**): First, we confirmed *via*

21   the interpretation of the HMBC spectrum of the crude extract that the methoxy group of the

22   putatively identified *E*-rhaponticin was directly linked to C-1 at 147.7 ppm. Secondly, the

23   proton at 6.99 ppm linked to C-4 at 119.9 ppm exhibited HMBC correlations with C-2 at

24   146.4 ppm, with C-3 at 114.8 ppm, with C-6 at 109.1 ppm, but not with C-1 at 147.7 ppm. An

1   HMBC correlation was also observed between the proton at 7.14 linked to C-6 and C-1 at 147.7

2   ppm, indicating that C1 belongs to the same carbon skeleton as C-2, C-3, C-4, and C-6.

3   Therefore C-1 and C-4 were placed in a *para* position from each other. ∑Thirdly, the

4   experimental chemical shift value of C-3 was 114.8 ppm, a value which is very close to the

5   predicted value of the same carbon position for *E*-isorhapontin at 115.0 ppm, while the same

6   carbon position of the *E*-rhaponticin is predicted at 110.6 ppm. Therefore the validation of the

7   structure of *E*-isorhapontin lead to discard the possible presence of its very similar isomer *E*-

8   rhaponticin in the methanol bark extract of *P. abies*. For dihydrokaempferol and maltose, the

9   experimental chemical shift values and 2D NMR connectivity data were not all confirmed after

10   manual checking, therefore both compounds were also discarded from the list of hits.

Figure 5 | Experimental HMBC correlations observed for E-isorhapontin. Chemical shift

values: H4→C2 (6.99 ppm→146.4 ppm); H4→C3 (6.99 ppm→114.8 ppm), H4→C6 (6.99 ppm,

^{13}C 109.1 ppm), H6→C4 (7.14 ppm→119.9 ppm), H6→C1 (7.14 ppm→147.7 ppm), H6→C2

(7.14 ppm→146.4 ppm), H8→C1 (3.9 ppm→ 147.7ppm).
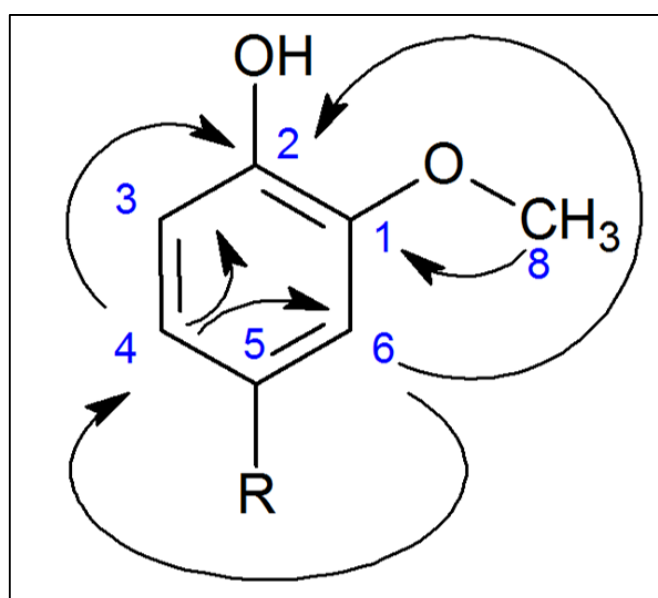
Table 2. Proposal list for the *P. abies* bark extract.

	Putatively identified metabolites	Score 1 (HMBC cluster only)	Score 2 (HMBC full spectrum)	Score HSQC	Validation
C1	*E*-resveratrol	76	81	100	eliminated*
	E-piceid	76	94	100	confirmed
	E-isorhapontin	73	95	100	confirmed
	E-piceatannol (astringinin)	71	82	100	eliminated*
	E-rhapontigenin	69	79	100	eliminated*
	E-pterostilbene	68	73	91	eliminated
	E-astringin	66	91	100	confirmed
	7,3'-dihydroxy-4'-methoxy-isoflavone	66	66	75	eliminated
	rhaponticin	65	92	100	eliminated*
	E-resveratrol-3-*O*-D-glucuronide	64	79	93	eliminated

C2	**saccharose**	**63**	**74**	**100**	**confirmed**
	maltose	**59**	**72**	**100**	**eliminated***
	sophorose	59	65	79	eliminated
	myricetin 3-*O*-rutinoside	57	70	82	eliminated
	kojibiose	57	65	86	eliminated
	kaempferol 3,4'-diglucoside	54	61	75	eliminated
	rutinose	53	62	92	eliminated
	vanillic acid glucoside	53	63	91	eliminated
	laricitrin 3-*O*-glucoside	52	62	67	eliminated
	Laricitrin 3-*O*-rutinoside	52	65	67	eliminated
C3	13-epimanool	47	58	44	eliminated
	cycloisosativene	41	49	69	eliminated
	neointermedeol	39	59	58	eliminated
	veticadinol	39	55	50	eliminated
	phylloquinone	39	55	44	eliminated
	docosadienoic acid	39	54	89	eliminated
	gamma-tocopherol	38	58	48	eliminated
	beta-tocopherol	37	60	45	eliminated
	delta-tocopherol	36	57	55	eliminated
	alpha-tocopherol	35	52	42	eliminated
C4	**(+)-catechin**	**52**	**81**	100.0	**confirmed**
	taxifolin	**41**	**88**	100.0	**confirmed**
	(-)-epicatechin	40	62	88.9	eliminated
	syringetin	40	55	50.0	eliminated
	(+)-catechin gallate	37	81	72.7	eliminated
	dihydrokaempferol	36	86	100.0	**eliminated***
	(-)-epigallocatechin	33	51	75.0	eliminated
	taxifolin 3'-*O*-glucoside	**31**	**85**	**100.0**	**confirmed**
	epi-afzelechin	30	48	70.0	eliminated
	flavogallonic acid	30	40	50.0	eliminated
C5	monogalloylglucose	21	76	78	eliminated
	lactose	20	60	86	eliminated
	maltose	**20**	**72**	**100**	**eliminated***
	laricitrin 3-*O*-glucoside	19	62	67	eliminated
	3,3'-di-*O*-methylellagic acid 4'-*O*-beta-D-xylopyranoside	19	53	80	eliminated
	myricetin 3-*O*-glucoside	18	70	82	eliminated
	3,3',4-tri-*O*-methylellagic acid-4'-*O*-beta-D-glucopyranoside	17	47	67	eliminated
	quercetin-3-*O*-alpha-D-arabinopyranoside	17	48	55	eliminated
	cibarian	16	44	27	eliminated
	3,4-dimethoxyphenyl-β-D-glucopyranoside	**16**	**84**	**100**	**confirmed**

1   *High final score but eliminated after manual checking

2

3

4

5

1 **Figure 6 | Chemical structures of the eight confirmed metabolites in the *P. abies* bark extract.**

2 Colors highlighted atoms with chemical shifts that matched chemical shift values from the

3 experimental HMBC peak list; Pink: matching molecular features, Blue: matching both ^{1}H and [13]C

4 chemical shifts, Green: matching only ^{13}C chemical shifts, Red: Matching only ^{1}H chemical shifts.

5

E- isorhapontin   *E-*astringin   *E-*piceid   saccharose

(+)-catechin   taxifolin   taxifolin 3'-*O*-β-*D*-glucopyranoside   3,4-dimethoxyphenyl-β-*D*-glucopyranoside

6
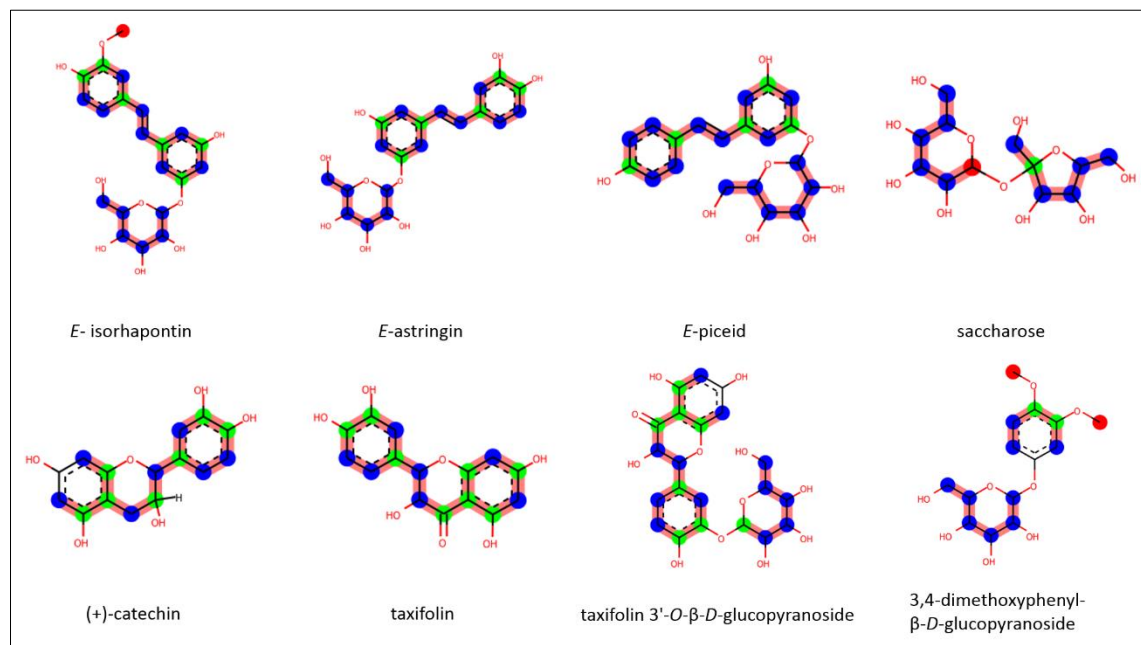
7 **Conclusion**

8 A new method exploiting HMBC and HSQC correlation data was developed in order to

9 assist and accelerate the identification of natural metabolites directly within mixtures,

10 while increasing the identification accuracy thanks to the richness of the information

11 brought by HMBC and HSQC spectra. To the best of our knowledge, this is the first

12 application of the network theory for the exploitation of heteronuclear 2D NMR data in

13 the context of natural mixture analysis. In this paper, the proof of concept of the method

14 was illustrated by two samples, a simplified model mixture of three commercial

15 compounds and a real life natural extract. In the crude methanol extract of the bark of

16 *Picea abies*, the method allowed the accurate identification of 8 molecules, namely *E-*

1   isorhapontin, *E*-astringin, *E*-piceid, saccharose, (+)-catechin, taxifolin, taxifolin 3'-*O*-β-*D*-

2   glucopyranoside and 3,4-dimethoxyphenyl-O-β-D-glucopyranoside. This method achieves

3   a better performance when applied to the analysis of simplified fractions with well

4   resolved HMBC spectra. The performance of this method is indeed highly dependent on

5   the HMBC spectrum quality because the latter determines the reliability of the initial peak

6   picking step. The method allows an early stage chemical profiling of the natural extracts

7   and it could be very complementary to MS-based dereplication by combining the different

8   advantages of each method in order to cross validate the results and have more accurate

9   insight into the chemical composition of the extracts.

10