Targeted Dereplication of Microbial Natural Products by High-Resolution MS and Predicted LC-Retention Time

Justine Chervin,[†,#] Marc Stierhof,[†,#] Ming Him Tong,[†,#] Doe Peace,[†] Kine Østnes Hansen, [‡] Dagmar Solveig Urgast,[†] Jeanette Hammer Andersen,[‡] Yi Yu,[§] Rainer Ebel,[†] Kwaku Kyeremeh,[⊥] Veronica Paget,[▽] Gabriela Cimpan,[▽] Albert Van Wyk,[▽] Hai Deng,[†] Marcel Jaspars,[†] Jioji N. Tabudravu[†,]*

[†]*The Marine Biodiscovery Centre, Department of Chemistry, University of Aberdeen, AB24 3UE, Scotland, UK.*

[‡]*Marbio, UiT The Arctic University of Norway, Breivika, N-9037, Tromsø , Norway.*

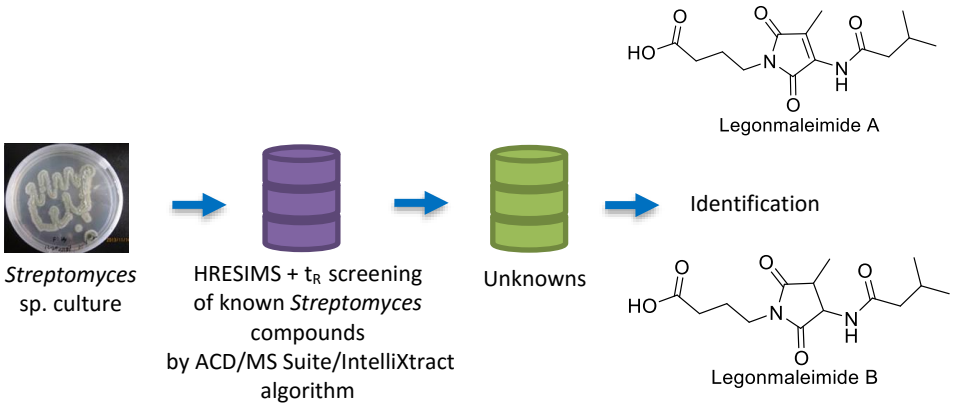
[§]*Key Laboratory of Combinatory Biosynthesis and Drug Discovery (Ministry of Education), School of Pharmaceutical Sciences, Wuhan University, 185 East Lake Road, Wuhan 430071, China.*

[⊥]*Marine and Plant Research Laboratory of Ghana, Department of Chemistry, University of Ghana, Accra, P.O. Box LG 56, Ghana.*

[▽]*Advanced Chemistry Development, UK Ltd., Venture House, Arlington Square, Downshire Way, Bracknell, Berks. RG12 1WA, UK.*

ABSTRACT

A new strategy for the identification of known compounds in *Streptomyces* extracts that can be applied in the discovery of natural products is presented. The strategy incorporates screening a database of 5,553 natural products including 5,102 structures from *Streptomyces* sp. alone, using a high throughput LCMS data processing algorithm that utilises HRMS data and predicted LC retention times (t_R) as filters for rapid identification of compounds in the natural product extract. The database named StrepDB contains for each compound, the structure, molecular formula, molecular mass, and LC predicted retention time. All identified compounds are annotated and color coded for easier visualization. It is an indirect approach to quickly assess masses (which are not annotated) that may potentially lead to the discovery of new or novel structures. In addition, a spectral database named MbcDB was generated using ACD/Spectrus DB Platform. MbcDB contains 665 natural products, each with structure, experimental HRESIMS, MS/MS, UV, and NMR spectra. StrepDB was used to screen a mutant *Streptomyces albus* extract that led to the identification and isolation of two new compounds: legonmaleimides A and B, the structures of which were elucidated with the aid of MbcDB and spectroscopic techniques. The structures were confirmed by computer assisted structure elucidation (CASE) methods using ACD/Structure Elucidator Suite. The developed methodology suggests a pipeline approach to the dereplication of extracts and discovery of novel natural products.


Legonmaleimide A
Identification
Streptomyces sp. culture
HRESIMS + t_R screening of known *Streptomyces* compounds by ACD/MS Suite/IntelliXtract algorithm
Unknowns
Legonmaleimide B

The need for increasing productivity in the discovery of new or novel drug-like compounds has resulted in the search for new techniques and strategies to streamline the drug discovery pipeline. Central to this is the rapid identification of known compounds and analogues, a process known as chemical dereplication[1–3] enabling resources to be focused on extracts that can yield new or novel compounds. The most commonly employed methods involve the use of liquid chromatography coupled to a photodiode array detector and a high resolution mass spectrometer (LC-DAD-MS).[4,5] The use of HRESIMS offers advantages over other methods, for example UV and NMR due to its high sensitivity and versatility, enabling it to produce more than one type of data in a single experiment such as positive, negative and fragmentation. However, determining molecular structures is more challenging and despite advances in MS technology, the process still needs NMR techniques to provide full structural information. Chemical dereplication methods available today use either one or both of these methods and generally fall into two main categories: targeted or non-targeted dereplication. Non-targeted chemical dereplication methods are less focused and often result in wasted time and resources due to reisolation of known compounds. Targeted chemical dereplication offers advantages as it allows resources to be utilized on isolation of new compounds or targeted isolation of known compounds for example with interesting biological activity.[4] For example LC-DAD-TOFMS was used to successfully identify known fungal metabolites by comparison of UV and MS data with reference standards.[6] The use of a database containing HRMS, MS/MS and UV data was used to rapidly identify the presence of aflatoxins which are nuisance mycotoxins in crude fungal extracts.[7] The use of MS/MS data to build molecular networks can identify compounds that are structurally related,[8,9] but requires MS/MS data in integrated compound databases for their identification. Current MS/MS libraries containing experimental MS/MS data are known to cover only a limited

number of natural products, estimated to be around 10%.[9] The combination of LCMS and phylogenetic finger printing was used in the targeted discovery of cyanobacterial compounds.[10] NMR spectroscopy, despite the problem of low sensitivity is increasingly being used for dereplication purposes.[11–13] The use of experimental LC retention time by comparing the retention time of an unknown with the retention time of a known standard compound is a commonly employed method of identifying compounds when used orthogonally with another data source like HRMS or MS/MS.[7,14] However, it is not practical to purchase every known compound in the natural products database for measurement of experimental retention times considering the fact that there are about 250,000 of natural products[15] known to date. Use of predicted retention times, however is a possible alternative and has been studied for metabolite identification.[16] Prediction of retention times follows the concept of quantitative structure retention relationship (QSRR) which is based on the physicochemical nature of the analyte-column interactions that determine retention.[17] It has been successfully applied to HPLC for identification of specific classes of compounds such as peptides,[18] steroids,[19] and lipids.[20] Application in chemically diverse group such as organic synthetic compounds has been carried out successfully,[21] but has yet to be applied in the identification of natural products. The huge diversity of structures in natural products is believed to be a challenge for any retention time prediction model,[16] but, it has the potential to work successfully as an extra filter when used with HRMS in a high throughput screening strategy to rapidly improve the identification of known natural products which is often one of the major bottle-necks[22] in drug discovery.

The goal of this study was to identify new natural products easily from *Streptomyces* sp. through rapid screening of known compounds held in a *Streptomyces* natural products database, StreDB.

To facilitate this requires first of all the construction of the database, StrepDB to contain the necessary information such as structures, molecular formulas, molecular masses, and predicted LC-retention times for all compounds. And, secondly to determine and optimize the accuracy of the method in its ability to identify known standard compounds in an extract. Complimentary to this was the construction of a spectral natural product database, MbcDB which contains NMR, LCMS, MS/MS, and UV data of 665 natural products. Spectral databases are crucial for identifying known compounds or providing spectral information for the elucidation of new structures. To test this new approach we build two databases (StrepDB and MbcDB), and looked at the screening, isolation and characterization of new pyrrolidine alkaloids related to the legonmycins and the legonindolizidines[23] in a mutant *Streptomyces* extract.

RESULTS AND DISCUSSION

StrepDB. The Antimarin database[24] (version 2011) holds 15,831 unique masses and 35,850 molecular structures. The bacterial genus *Streptomyces* alone contributes about 6,845 structures to this database. These *Streptomyces* derived structures were imported to ACD/ChemSketch[25] and checked for structural irregularities using the 'clean structure' option available in ChemSketch. After removal of duplicated structures the remaining MDL MOL files (5,098 or 74.5% of the original structures) were imported to IntelliXtract (an add-in software of ACD/MS Workbook Suite)[25] to generate StrepDB. StrepDB holds 5,100 compounds from *Streptomyces* and 453 compounds from other sources giving a total of 5,553 compounds with structures, exact masses, molecular formulae, and predicted LC retention times (Table 1, Figures S3-S8). Two known compounds (jasplakinolide and antibiotic A-23187) were analysed by LCMS and the HRMS data used to optimize the data preprocessing settings (S43). The green colour-coded output files

shown in Figures S1 and S2 indicated that the accuracy target of 5 ppm has been observed for both compounds. Optimization of preprocessing settings is important as it affects the peak-picking, alignment and data accuracy of the applied algorithm,[26,27] particularly in complex biological samples. To determine the impact of retention time, and retention time window on the number of identified peaks, a crude sample of *Streptomyces* sp. MA37[23] was analysed by LCMS, and the data processed by IntelliXtract and screened by StrepDB using two different settings for retention time. The first was keeping both the retention time (t_R), and retention time window (t_R window) unfilled (S46). The second was using the predicted values for t_R, but used seven different settings for t_R window: no value, 0.0, 1.5, 2.0, 2.5, 3.0, and 4.0 minutes (S46 – S51, S3). The results (even though the number of sample is limited) showed an inverse correlation (R^2 = 0.97) between the retention time window used and the number of annotated compounds removed (S52-S59), considered as false hits. For example, a retention time window of 1.5 minutes removed up to 90% of hits when HRMS is used alone indicating the importance of retention time (and retention time window) as an additional filter in compound identification. To demonstrate the accuracy of the strategy a known component of *Streptomyces* sp. MA37, legonmycin A[23] was processed where the predicted retention time, structure, and other information were inputted into StrepDB and the LCMS data analysed as before. This time the mass 253.1555 that had previously been 'unlabeled' was annotated correctly as legonmycin A (Figure S60). The difference between experimental and predicted retention times of legonmycin A was 2.6 minutes, well within the 4.0 minute retention time window set in StrepDB.

No.	Reference M...	Formula	Structure	Label	tR (Min)	tR Window
823	245.08	C12H11N3O3		Caerulomycin-B	6.4	4.0
824	259.0957	C13H13N3O3		Caerulomycin-C	11.3	4.0
825	403.138	C19H21N3O7		Caerulomycin-D	5.4	4.0
826	214.0954	C9H14N2O4		Cairomycin A	1.7	4.0
827	225.1113	C10H15N3O3		Cairomycin B	1.9	4.0

Table 1 StrepDB contains 5,553 compounds including 5,102 structures from *Streptomyces* with names, molecular formulae, high resolution monoisotopic masses, and calculated LC-retention times. More compounds can be added either directly to the table or via XML files.

MbcDB. This database holds spectral data for 665 natural products from marine and terrestrial sources representing several classes of natural products such as peptides (30), alkaloids (120), terpenes (300) and others (215). A library of 502 of compounds was obtained from Enzo Life Sciences UK[28] while the rest (163) came from our in-house compound collections. The database holds information such as compound structure, references, text files such as SMILES, and InChIKey, HRESIMS, LRMS/MS, ^{1}H NMR and UV spectra for all compounds. HRESIMS and MS/MS data are in positive mode with some in the negative mode (<10%). For a selection of compounds (<10%) there are ^{13}C and 2D NMR data such as COSY, HSQC and HMBC (Figure

S9). Since all the spectral data are stored as raw data files these can be reprocessed if required (Figures 1, S10-S13).

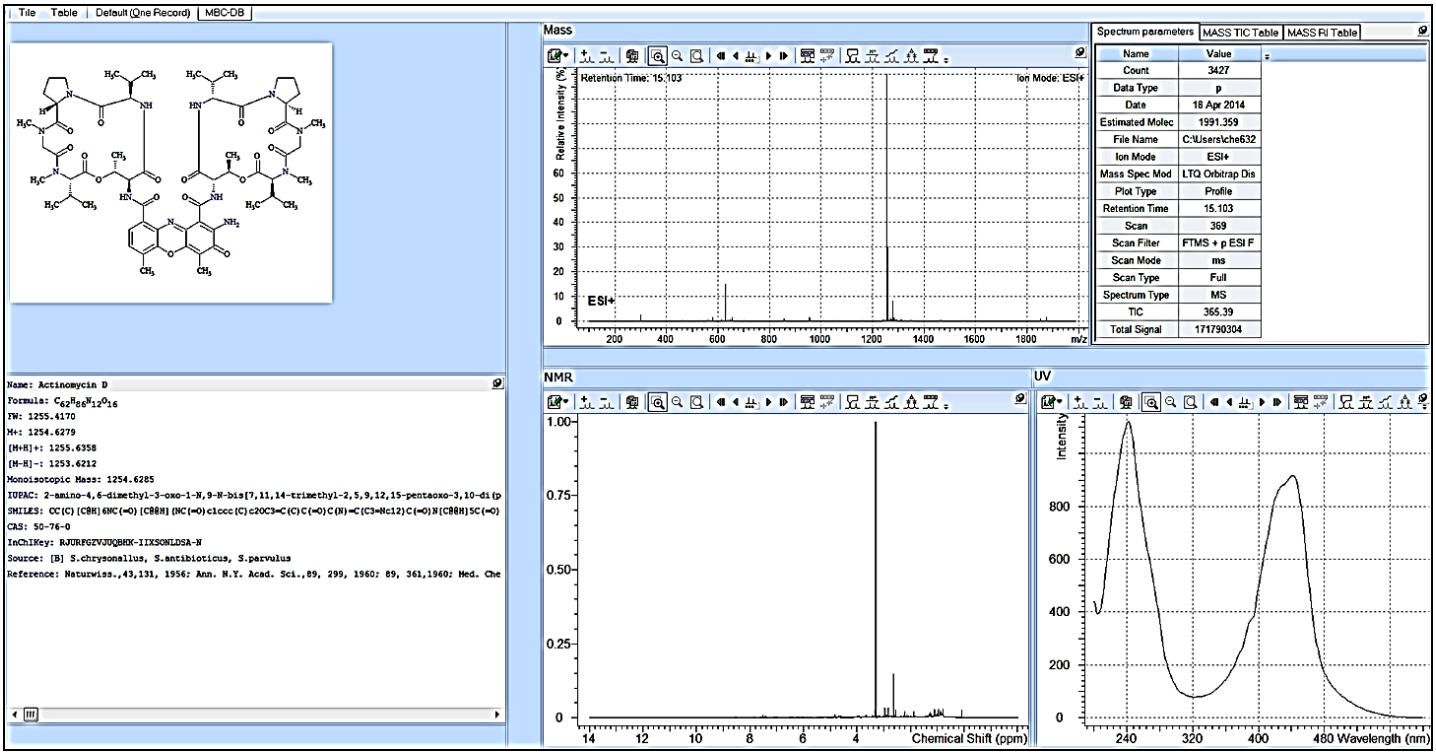


Figure 1. MbcDB contains 665 compounds with spectral data such HRMS, LRMS/MS, UV and ^{1}H NMR. The database is searchable based on compound name, molecular formula, exact mass, structure, and substructure. Spectra of unknown compounds such as UV, 1D & 2D NMR, MS and MS/MS can be searched for similarities with spectra held within MbcDB (e.g. Figures S19, S28, S44, S45).

Calculation of LC Retention Times. The LC-retention times of compounds in StrepDB were calculated using ACD/ChromGenius[25] The software uses physicochemical parameters like log*P*, log*D*, molecular weight (MW), molecular volume (MV), molar refractivity (MR), polar surface area (PSA), H-acceptors (NA), and H-donors (ND) to create a knowledge base using a set of 'training' compounds that had been measured previously.[25] To create the knowledge base, 417

natural products from MbcDB with structures, monoisotopic masses, and LCMS data (Figures S14, S15) were downloaded to ChromGenius. Calculations were performed for the method used (solvents, pH, temperature, column type, column size, and experimental LC-retention time) to build the prediction model. Retention times were predicted for each compound using the 10% most similar compounds in the knowledge base obtained by a Dice coefficient similarity search.[29] An overall correlation value (R^2= 0.75) between the experimental and predicted retention times of the 417 standard compounds was obtained (Figure 2). Analysis of the data showed that 76% of the standard (training) compounds had retention time deviation (experimental minus predicted) between 0.0-2.0 minutes, 93% between 0.0-4.0, and the rest (7%) between 4.2-10.3 minutes (Table S61). The model was then used to calculate the retention times of the 5,553 compounds held in StrepDB (Table 1).

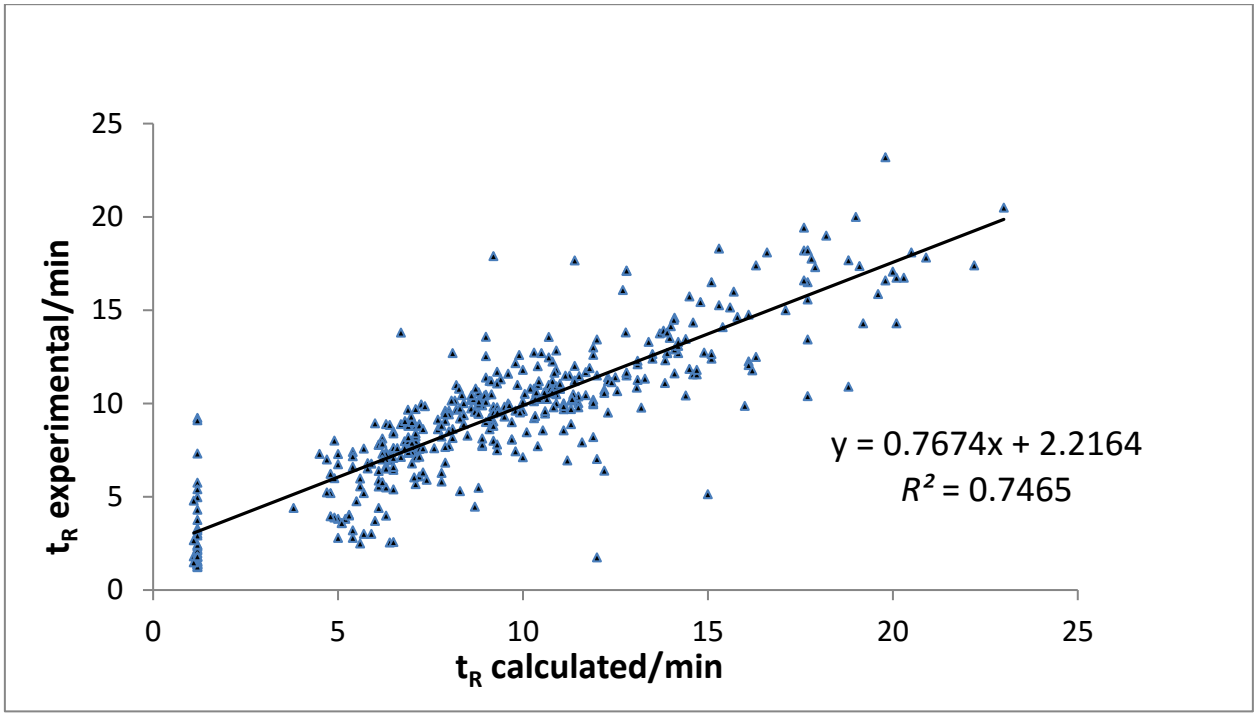


Figure 2. Correlation between predicted and experimental LC retention times, n = 417.

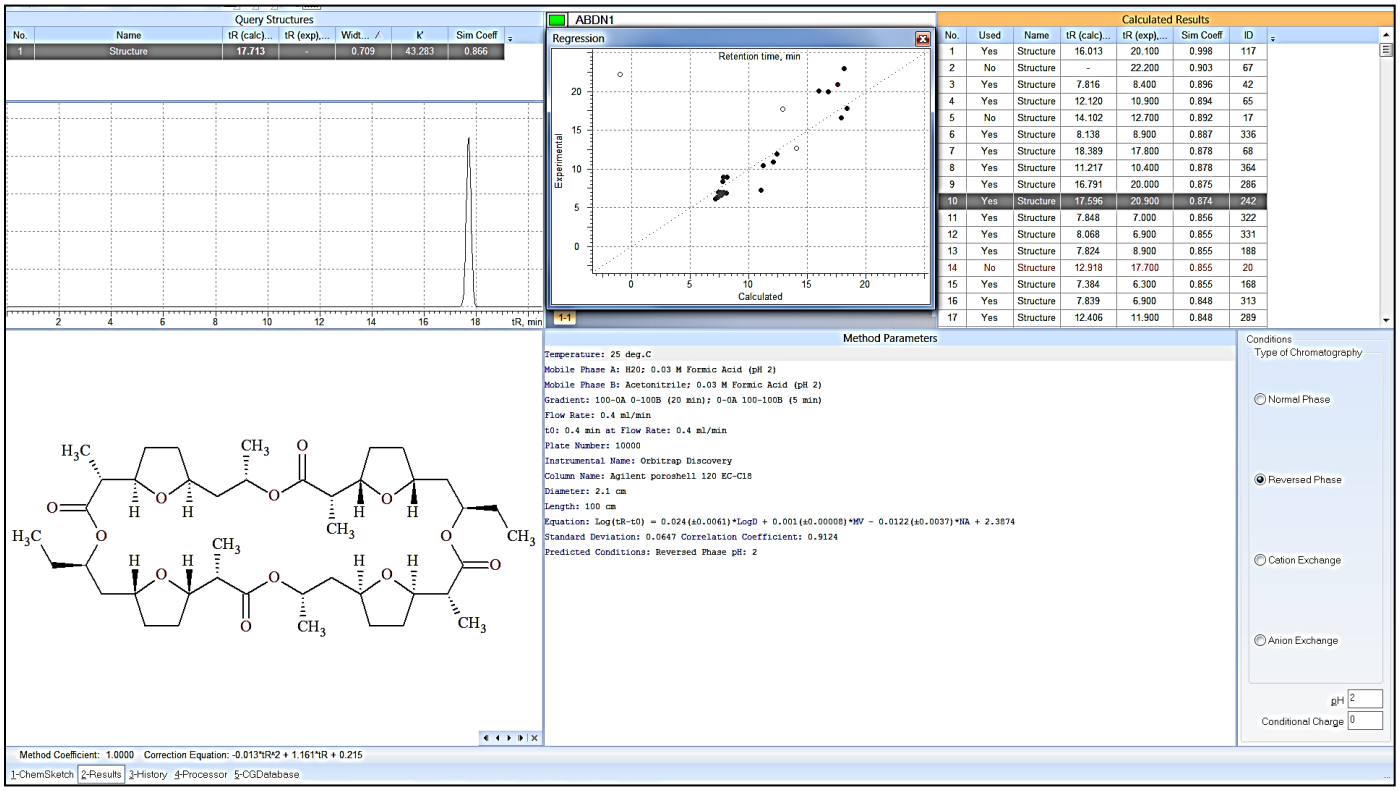


Figure 3. The calculated LC elution profile of dynactin (structure shown), correlation between predicted and experimental LC retention times of the 26 compounds used in the calculation are shown. A similarity coefficient (Sim Coeff) of 0.87 was obtained. A similarity coefficient of 1.0 indicates a perfect match between calculated and experimental retention times.

Dereplication using StrepDB. The *Streptomyces* strain used in this study was obtained from a soil sample from the University of Ghana, Africa and the subsequent construction of the legonmycin gene cluster, in-frame deletion and expression of the mutant strain have been described previously.[23] The crude sample of *Streptomyces albus, ΔlgnC* was fractionated into four fractions: water-butanol, water-methanol, dichloromethane, and hexane based on polarity using a modified Kupchan method.[30,31] The four fractions were subjected to LCMS analysis, but

only the water-butanol fraction was prioritized over the other three for compound isolation workup based on the amount (dried weight) of fraction and relative intensity of the ions to the quality control sample (reserpine). Data processing of this fraction and subsequent screening of StrepDB by IntelliXtract using a retention time window of 4.0 (± 2.0) minutes resulted in 71 masses between 200-500 Da, of which four masses were annotated as matching 10 compounds in StrepDB (Table S41). The structures of these compounds are easily displayed by switching to 'Table of Components' view (Figure S42). The data suggested that the remaining 67 masses (94% of the total masses) were not in StrepDB, and were potentially new compounds. A UV profile filter was then applied manually to target only the compounds related to the legonindolizidines,[23] resulting in eleven masses with the expected UV profile (Table S16, Figure S17). Inspection of the data indicated that one of the compounds of interest had been identified as legonindolizidine A (Figure S42). Legonindolizidines[23] A (**3**) and B (**4**) have previously been isolated from this strain, and their structures characterised based on ^{1}H NMR, MS and MS/MS data.[23] Legonindolizidine B was also identified (results not shown), after applying a retention time window of 6.0 minutes. Out of the nine unidentified masses of interest (based on UV profile), two were isolated, purified, and the structures elucidated as new legonmaleimides. The yield of the remaining seven compounds after HPLC purification were far too low for measurement of 1D and 2D NMR data, and hence their structures remained to be determined.

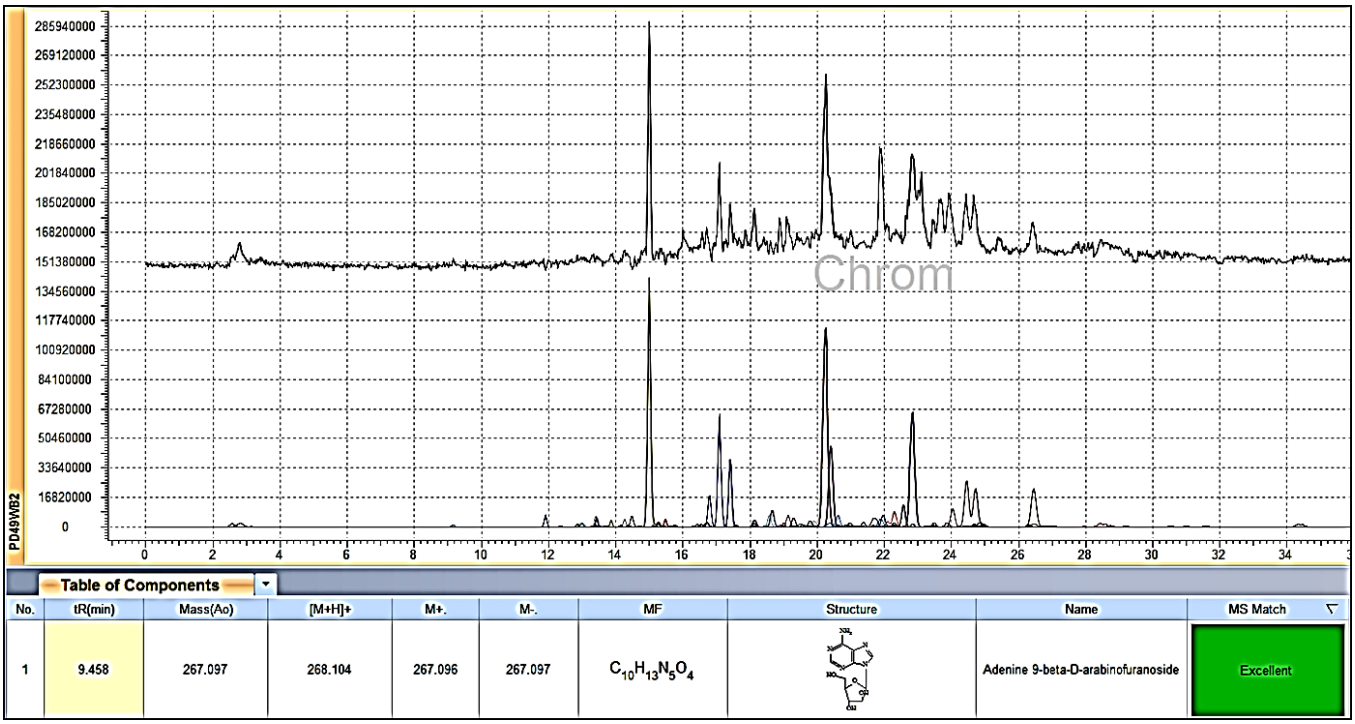
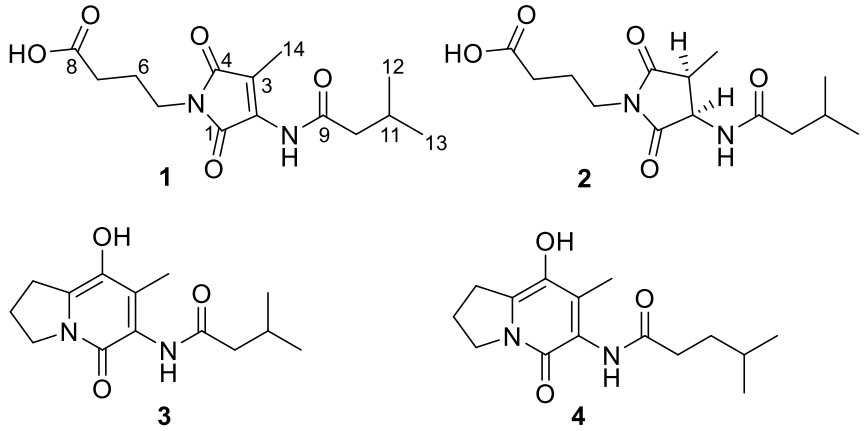


Figure 4. IntelliXtract output file of the water-butanol fraction showing the presence of adenine-9-beta-D-arabinofuranoside. The total ion chromatogram (top) is overlaid with the extracted ion chromatogram (bottom) for visualization of peak retention times. All extracted ions can be visualized as either a mass list (Table S41) or as a table of component containing structures (Table S42). The dark-green in the MS/Match column has indicated an excellent match with the compound adenine 9-beta-D-arabinofuranoside in StrepDB.

Isolation and Structure Determination. The water-butanol fraction was purified by a C_{18} HPLC column using a water-methanol gradient to afford the new compounds legonmaleimides A (**1**) and B (**2**), and the known compounds legonindolizidines[23]A (**3**) and B (**4**).



HRESIMS of legonmaleimide A (**1**) gave a molecular formula of $C_{14}H_{21}O_5N_2$ requiring 6 degrees of unsaturation (Figure S18). A ^{1}H NMR spectrum search of MbcDB suggested proton signal similarities to the known legonindolizidines (Figure S19).[23] Interpretation of ^{1}H, edited HSQC, and HMBC NMR data (Figures S20, S21, S23) indicated the presence of four methylenes (δ_C 44.7, 37.6, 34.9, 25.4 ppm), one methine (δ_C 25.9 ppm), two sp^2 (δ_C 132.9, 122.2 ppm), three amides (δ_C 172.2, 172.1, 168.0 ppm), and one carboxylic acid group (δ_C 180.0 ppm) suggesting the presence of one ring in the structure of **1**. The full structure was assigned by the interpretation of 1D and 2D NMR data in particular COSY (Figure S22), HMBC correlations (Table 2, Figure 5), and by comparison of NMR data with those of farinomalien, a maleimide-bearing compound from the entomopathogenic fungus *Paecilomyces farinosus*.[32] The COSY spectrum indicated two fragments: H5-H6-H7 and H10-H11-H12/H13. HMBC correlations from H14 to C2, C3, C4 and from H5 to C1 and C2 established the maleimide core and the butanoic acid group suggesting the bonding of C5 to the maleimide nitrogen. ^{13}C chemical shifts suggest that the 3-methylbutanamide unit was linked to the maleimide core at C2. Examples of compounds with the 3-methylbutanamide have previously been isolated from this bacterial strain.[23]

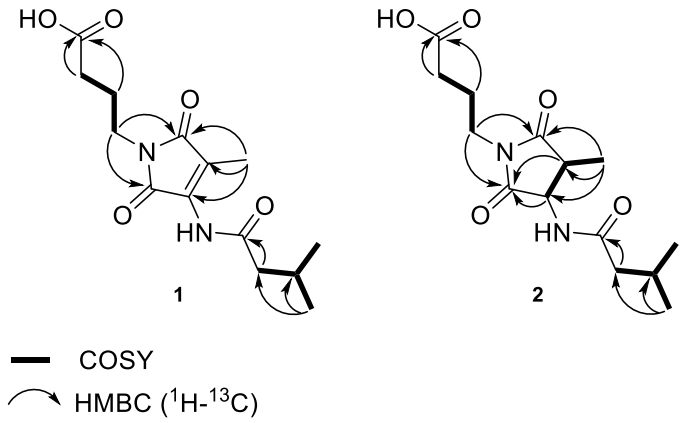


Figure 5. COSY and HMBC correlations of **1** and **2.**

Legonmaleimide B (**2**) showed a formula of $C_{14}H_{23}O_5N_2$ based on HRESIMS and requiring 5 degrees of unsaturation (Figure S27). A ^{1}H NMR spectrum search of MbcDB suggested similarities in proton profile to the known legonindolizidines (Figure S28).[23] Interpretation of 1D and 2D NMR data (Figures S29-S31) suggested that compound **2** had lost the double bond between C2 and C3. 2D COSY and HMBC correlations (Figures S32, S33) support the proposed structure shown in Figure 5. The absolute stereochemistries of legonmaleimide (**2**) at positions C2 and C3 have not been determined. The relative stereochemistry, however, was determined by a 2D ROESY NMR experiment that showed a strong correlation between H2 to H3 indicating a *cis*-configuration (Figures S34, S35).

Additional evidence for the two structures were provided by MS fragmentation data (Figures S26, S36), and by Computer Assisted Structure Elucidation (CASE) using ACD/Structure Elucidator Suite (version 2015.2.5).[25] 1D NMR (^{1}H), and 2D NMR data including HSQC, COSY, ROESY and HMBC plus the molecular formula were entered into ACD/Structure Elucidator and all possible structures calculated (Figures S37, S40) in Fuzzy Structure Generation (FSG) mode to detect and resolved non-standard correlations in COSY and HMBC

NMR data.[15,33,34] Calculations for compound **1** produced 77 possible structures

Table 2. ^{1}H, ^{13}C, ^{1}H-^{1}H COSY and ^{1}H-^{13}C HMBC NMR data of compounds **1** and **2** at 600/150 MHz in CD_3OD

	1					2			
pos.	δ_C , type	δ_H (*J* in Hz)	COSY ^{1}H-^{1}H	[a]HMBC	pos.	δ_C , type	δ_H (J in Hz)	COSY ^{1}H-^{1}H	[a]HMBC
1	168.0, C				1	176.1, C			
2	132.9, C				2	51.5, CH	4.68, d, 8.8	3	
3	122.2, C				3	38.3, CH	3.08, m	2, 14	4
4	172.1, C				4	179.8, C			
5	37.6, CH_2	3.50, t, 7.0	6	1, 4	5	37.9, CH_2	3.60, t, 6.8	6	1, 4
6	25.4, CH_2	1.83, m	5, 7	8	6	22.7, CH_2	1.90, m	5, 7	8
7	34.9, CH_2	2.12, overlap	6	6	7	30.8, CH_2	2.08, t, 7.2	6	6
8	180.0, C				8	175.2, C			
9	172.2, C				9	174.9, C			
10	44.7, CH_2	2.30, dd, 7.2, 7.2	11	11, 12, 13	10	44.4, CH_2	2.14, dd, 6.7, 7.1	11	11
11	25.9, CH	2.10, m	10, 12, 13	10, 12, 13	11	26.1, CH	2.08, m	10, 13, 14	10, 12, 13
12	21.3, CH_3	0.98, d, 6.7	11	11	12	21.5, CH_3	0.99, d, 6.5	11	11
13	21.3, CH_3	0.98, d, 6.7	11	11	13	21.5, CH_3	0.98, d, 6.5	11	11
14	9.0, CH_3	1.97, s		4	14	9.8, CH_3	1.14, d, 7.8	3	4

[a]HMBC correlations optimized for 8.0 Hz are from proton(s) stated to the indicated carbon.

which were then ranked according to the differences between the experimental and calculated ^{13}C data.[35] The low chemical shift deviations of the HOSE-code (d_A),[36] incremental Method (d_I),[37] and Artificial Neural Networks (d_N)[37] suggested that the proposed structure is correct. The top 25 candidates are shown in Figure S38 where the proposed candidate was placed at the number 1 position. A similar calculation performed for compound **2** yielded 92 possible structures. The top 25 candidates are shown in Figure S39 where the proposed structure for **2** was placed at the number 1 position.

Plausible biogenetic pathway. Compound **1** is likely to be derived from the biosynthetic intermediate legonindolizidine A[23] (**3**) through hydroxylation at C5, followed by ring rearrangement to generate a reactive aldehyde species **5**, which is then oxidized to the corresponding carboxylic acid **1**. Compound **1** is then reduced into **2** (Figure 10).

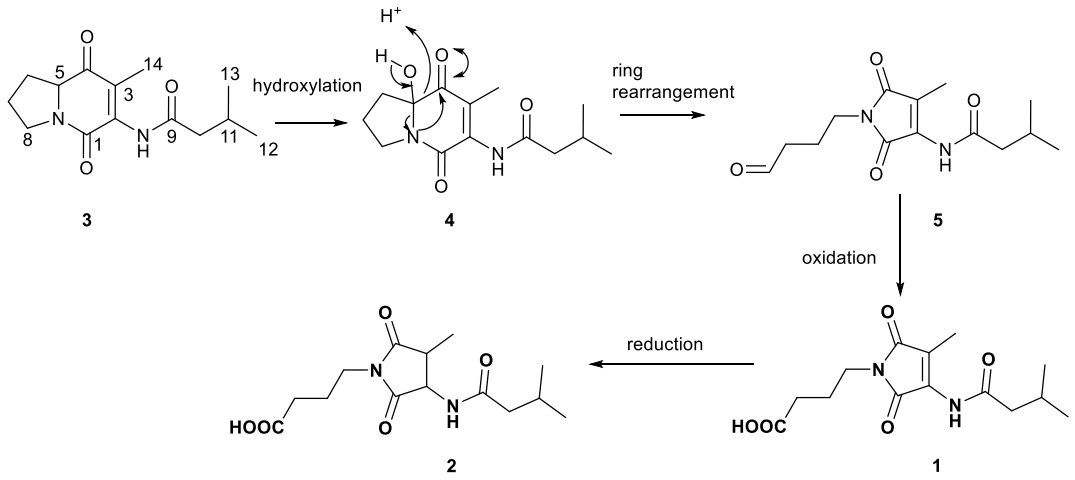


Figure 10. Proposed biosynthetic pathway of compounds **1** and **2**.

Compound **1** was tested against the human melanoma cell line A2058 and was found to be inactive (100% cell survival rate at 250 μM). Compound **2** has not been tested in any biological assay.

EXPERIMENTAL SECTION

General Experimental Procedures. NMR data, both 1D and 2D were recorded on a Bruker AVANCE III HD Prodigy TCI Cryoprobe at 600 and 150 MHz for ^{1}H and ^{13}C respectively. This instrument was optimized for ^{1}H observation with pulsing/decoupling of ^{13}C and ^{15}N with 2H

lock channels equipped with shielded z-gradients and cooled preamplifiers for ^{1}H and ^{13}C. The ^{1}H and ^{13}C chemical shifts were referenced to the solvent signals (δ_H 3.31 and δ_C 49.00 in CD_3OD). High resolution mass spectrometry data were measured using a ThermoScientific LTQXL-Discovery Orbitrap coupled to an Accela UPLC-DAD system. The following conditions were used for mass spectrometric analysis: capillary voltage 45 V, capillary temperature 320 °C, auxiliary gas flow rate 10 -20 arbitrary units, sheath gas flow rate 40-50 arbitrary units, spray voltage 4.5 kV, mass range 100-2000 amu (maximum resolution 30,000). Optical rotation measurements were recorded using a Bellingham & Stanley, Model ADP410 Polarimeter at 589 nm. Semi-preparative HPLC purifications were performed using an ACE 5 C_{18}, 250 x 100 mm column connected to an Agilent 1100 HPLC system consisting of a binary pump, degasser, photodiode array (DAD) and a preparative fraction collector. This system was also used to record the UV profile for measurement of the molar extinction coefficient (ε). IR was measured using a Perkin Elmer FT-IR (UATR Two) spectrometer. All solvents were of HPLC grade.

General Procedure for HPLC Analysis. Analysis by LCMS was performed on a C_{18} (Agilent Poroshell 120, EC C_{18}, 2.7 µm, 2.1 x 100 mm) column. The gradient was from 100% water (0.1% formic acid) to 100% acetonitrile (0.1% Formic acid) in 25 minutes and kept at this solvent for another 5 minutes before equilibration with the starting solvent for another 5 minutes. The flow rate was 0.4 mL/min, and column temperature 25 °C. For the determination of the molar absorptivity an ACE 5 C_4 (250 x 10 mm) column was used with an HPLC gradient starting from 5% MeOH (in water) to 100% MeOH in 10 minutes and kept at that solvent for another 10 minutes, with a flow rate of 2.0 mL/min. For compound **1**, 30 µL of 0.1 mg/mL was injected to the column. For compound **2**, 10 µL of 1.4 mg/mL was used.

StrepDB. Compound structures were copied to ACD/ChemSketch (version 2015.2.5) from AntiMarin[24] and then converted to MDL MOL files. These were then imported into the Ion Presence List in IntelliXtract within ACD/MS Workbook Suite (version 2015.2.5) and saved as an XML file. Other information like compound names, LC retention time, retention time window and MS delta were either added directly to the table or via Microsoft Excel (version 2010) which were then reconverted to XML format before loading to MS Workbook Suite. An LC retention time window of 4 (± 2) minutes was added to accommodate variations in experimental retention time.

LCMS data (RAW file format) were loaded to ACD/MS Work Book Suite via ACD/Spectrus and processed using optimized preprocessing settings using the IntelliXtract algorithm (S43). The *m/z* range was set between 200-500 Da, and LC-retention time from 2-35 minutes. The mass accuracy value was set as 0.5 Da; the mass value for HRMS data processing approach was set as ± 0.005 Da; mass accuracy for peak labelling 0.005 Da, and mass accuracy for structure search set as 0.0005 Da. Similar settings were used for the IntelliTarget algorithm (within ACD/MS Workbook Suite); Additional settings for IntelliTarget: Target analysis mass accuracy for MC generation 0.0003 Da; type of mass: peak top; mass accuracy for assignment: 0.0003 Da; FFQ threshold set as zero.

MbcDB. NMR data (FID) from Varian/Agilent and Bruker were processed by ACD/NMR Workbook Suite, solvent referenced, extracted as single spectrum and loaded to Spectrus DB (in update mode) via the 'DB update key' within ACD/Spectrus Processor. Compound structures and associated information were loaded via ChemSketch. MS data were loaded via MS Workbook Suite, UV profile (derived from HPLC or LCMS) were loaded via the ACD/Optic

Workbook or ACD/MS Workbook Suite. MbcDB was saved as CryptoForge Document (CFD) file format.

LC-Retention Time Calculation. Compounds (n = 417) with experimental LC-retention times were downloaded from the MbcDB and converted to a SDF file. These values were imported into ACD/ChromGenius (version 2015)[25] to create a knowledge base for the given chromatographic method. ACD/ChromGenius created a prediction equation for each compound by relating retention time to key predicted physicochemical parameters. The prediction accuracy of the knowledge base was estimated using a leave one out approach for each compound.[38] This knowledge base was then used to calculate the retention times of the compounds in StrepDB. Retention times were predicted using the 10% most similar compounds in the knowledge base obtained by a Dice coefficient similarity search.[29]

Microbiology. The *Streptomyces albus*::minimal cassette Δ*lgnC* has been previous studied in which the key gene *lgnC* in the minimal cassette of the *lgn* gene cluster, responsible for the biosynthesis of the bacterial pyrrolizidine alkaloids, legonmycins A and B was genetically inactivated.[23] The *Streptomyces albus* mutant was cultured on MS medium (mannitol 20g/L, soybean meal 20g/L, agar 20g/L, pH 7.2) for spore formation. For fermentation, *Streptomyces albus* was cultured on ISP2 medium. Modified ISP4 medium [10 g/L soluble starch, 2 g/L $(NH_4)_2SO_4$, 1 g/L K_2HPO_4, 1 g/L $MgSO_4 \cdot 7H_2O$, 1 g/L NaCl, 1 g/L tryptone, 0.5 g/L yeast extract, 1 g/L peptone, trace element solution (1 mL/L), pH 7.2 before sterilization] containing the final concentration of 30 mM Mg^{2+} was used for conjugation of *Streptomyces albus*.

Isolation of the Legonmaleimides. The methanol extract of *Streptomyces albus*::minimal cassette Δ*lgnC* culture was dried and partitioned into four fractions: water-butanol, water-methanol, dichloromethane and hexane according to polarity.[30,31] The water-butanol fraction was prioritised over the others and dereplicated using StrepDB based on sample dried weight, and compounds of interest based on HRMS and UV profile (Table S16, Figure S17). Final purification was performed by reversed phase HPLC using water and methanol as solvents. The gradient started from 95% Water to 50% MeOH in 20 minutes and then reached 100% MeOH in another 10 minutes before equilibration for a further 5 minutes in the starting solvents to yield legonmaleimide A (2.0 mg) and legonmaleimide B (1.5 mg).

*Legonmaleimide A (**1**)* ^{1}H and ^{13}C NMR data (CD_3OD, 600 and 150 MHz, respectively), see Table 2; HRESIMS 297.1447 $[M+H]^+$ Δ 0.1 ppm calculated for $C_{14}H_{21}O_5N_2$; UV (MeOH) λ_{max} (log ε) 244 (3.96), 340 (3.04) nm; IR (MeOH, cm^{-1}) : 3329, 2957, 1702, 157, 1076, 626.

*Legonmaleimide B (**2**)* ^{1}H and ^{13}C NMR data (CD_3OD, 600 and 150 MHz, respectively), see Table 2; HRESIMS *m/z* 299.1603 $[M+H]^+$ Δ 0.7 ppm from calculated for $C_{14}H_{23}O_5N_2$. UV (MeOH) λ_{max} (log ε) 210 (2.94) nm; $[\alpha]^{17.5}_{D}$ +90.9 (c 3.7 MeOH).

ASSOCIATED CONTENT

Supporting Information

^{1}H NMR, COSY, ROESY, HSQC, HMBC, HRESIMS, ES^+ fragmentation, CASE data for compounds **1 and 2**. StrepDB, and MbcDB components. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Telephone: +44 1224 273105. Fax: +44 1224 272921. E-mail: j.tabudravu@abdn.ac.uk

Author Contributions

[#]Justine Chervin, Marc Stierhof, and Ming Him Tong contributed equally to this work.

Notes

The authors declare the following competing financial interest(s): G. C., V. P., and A. V. W. are full-time employees at ACD/Labs. The other authors declare no competing financial interest.

DEDICATION

Dedicated to Professor Phil Crews, of University of California, Santa Cruz, for his pioneering work on bioactive natural products.

ACKNOWLEDGEMENTS.

This work was supported in part by the University of Aberdeen Knowledge Transfer Fund grant 032 UZZ0101 (to J.T.). The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013 under grant agreement no 312184 'PharmaSea' to M.J., R.E., J.T., H.D., J. H. A., K.Ø.H. R. E., K.K., H.D., and M.J. acknowledge the financial support of the Leverhulme Trust-Royal Society Africa Award (AA090088). J.T., J. C., D S.U., and M.S. acknowledge Russell Gray of the Spectroscopy Lab., Marine Biodiscovery Centre, University of Aberdeen for NMR training and help.

REFERENCES

1. Bradshaw, J. *et al.* A Rapid and Facile Method for the Dereplication of Purified Natural Products. *J. Nat. Prod.* **64,** 1541–1544 (2001).
2. Corley, D. G. & Durley, R. C. Strategies for Database Dereplication of Natural Products. *J. Nat. Prod.* **57,** 1484–1490 (1994).
3. Williamson, R. T. *et al.* New Diffusion-Edited NMR Experiments To Expedite the Dereplication of Known Compounds from Natural Product Mixtures. *Org. Lett.* **2,** 289–292 (2000).
4. Hubert, J., Nuzillard, J.-M. & Renault, J.-H. Dereplication strategies in natural product research: How many tools and methodologies behind the same concept? *Phytochem. Rev.* (2015). doi:10.1007/s11101-015-9448-7
5. Gaudêncio, S. P. & Pereira, F. Dereplication: racing to speed up the natural products discovery process. *Nat Prod Rep* **32,** 779–810 (2015).
6. Klitgaard, A. *et al.* Aggressive dereplication using UHPLC–DAD–QTOF: screening extracts for up to 3000 fungal secondary metabolites. *Anal. Bioanal. Chem.* **406,** 1933–1943 (2014).
7. El-Elimat, T. *et al.* High-Resolution MS, MS/MS, and UV Database of Fungal Secondary Metabolites as a Dereplication Protocol for Bioactive Natural Products. *J. Nat. Prod.* **76,** 1709–1716 (2013).
8. Yang, J. Y. *et al.* Molecular Networking as a Dereplication Strategy. *J. Nat. Prod.* **76,** 1686–1699 (2013).
9. Allard, P.-M. *et al.* Integration of Molecular Networking and *In-Silico* MS/MS Fragmentation for Natural Products Dereplication. *Anal. Chem.* **88,** 3317–3323 (2016).

10. Salvador-Reyes, L. A., Engene, N., Paul, V. J. & Luesch, H. Targeted Natural Products Discovery from Marine Cyanobacteria Using Combined Phylogenetic and Mass Spectrometric Evaluation. *J. Nat. Prod.* **78,** 486–492 (2015).
11. Pierens, G. K., Mobli, M. & Vegh, V. Effective Protocol for Database Similarity Searching of Heteronuclear Single Quantum Coherence Spectra. *Anal. Chem.* **81,** 9329–9335 (2009).
12. Williams, R. B. *et al.* Dereplication of natural products using minimal NMR data inputs. *Org Biomol Chem* **13,** 9957–9962 (2015).
13. Pauli, G. F. *et al.* Essential Parameters for Structural Analysis and Dereplication by ^{1}H NMR Spectroscopy. *J. Nat. Prod.* **77,** 1473–1487 (2014).
14. Peironcely, J. E. *et al.* Automated Pipeline for De Novo Metabolite Identification Using Mass-Spectrometry-Based Metabolomics. *Anal. Chem.* **85,** 3576–3583 (2013).
15. Williams, A., Martin, G. & Rovnyak, D. *Modern NMR Approaches to the Structure Elucidation of Natural Products*. (The Royal Society of Chemistry, 2016). doi:10.1039/9781849735186
16. Creek, D. J. *et al.* Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Anal. Chem.* **83,** 8703–8710 (2011).
17. Kaliszan, R. QSRR: Quantitative Structure-(Chromatographic) Retention Relationships. *Chem. Rev.* **107,** 3212–3246 (2007).
18. Kaliszan, R. *et al.* Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships. *PROTEOMICS* **5,** 409–415 (2005).

19. Salo, M., Sirén, H., Volin, P., Wiedmer, S. & Vuorela, H. Structure-retention relationships of steroid hormones in reversed-phase liquid chromatography and micellar electrokinetic capillary chromatography. *J. Chromatogr. A* **728,** 83–88 (1996).

20. Aicheler, F. *et al.* Retention Time Prediction Improves Identification in Nontargeted Lipidomics Approaches. *Anal. Chem.* **87,** 7698–7704 (2015).

21. Falchi, F. *et al.* Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification. *Anal. Chem.* **88,** 9510–9517 (2016).

22. Hou, Y. *et al.* Microbial Strain Prioritization Using Metabolomics Tools for the Discovery of Natural Products. *Anal. Chem.* **84,** 4277–4283 (2012).

23. Huang, S. *et al.* Discovery of a Single Monooxygenase that Catalyzes Carbamate Formation and Ring Contraction in the Biosynthesis of the Legonmycins. *Angew. Chem. Int. Ed.* **54,** 12697–12701 (2015).

24. Blunt, J. W., Munro, M. H. G. & Laatsch, H. AntiMarin Database. (2006).

25. ACD/Labs.com :: Your Partner in Chemistry Software for Analytical and Chemical Knowledge Management, Chemical Nomenclature, and In-Silico PhysChem and ADME-Tox. Available at: http://www.acdlabs.com/. (Accessed: 15th November 2016)

26. Brodsky, L., Moussaieff, A., Shahaf, N., Aharoni, A. & Rogachev, I. Evaluation of Peak Picking Quality in LC−MS Metabolomics Data. *Anal. Chem.* **82,** 9177–9187 (2010).

27. Gürdeniz, G., Kristensen, M., Skov, T. & Dragsted, L. O. The Effect of LC-MS Data Preprocessing Methods on the Selection of Plasma Biomarkers in Fed vs. Fasted Rats. *Metabolites* **2,** 77–99 (2012).

28. Enzo Life Sciences. Available at: http://www.enzolifesciences.com/. (Accessed: 9th March 2017)

29. Zou, K. H. *et al.* Statistical validation of image segmentation quality based on a spatial overlap index1. *Acad. Radiol.* **11,** 178–189 (2004).

30. Tabudravu, J. N. & Jaspars, M. Stelliferin Riboside, a Triterpene Monosaccharide Isolated from the Fijian Sponge *Geodia g lobostellifera*. *J. Nat. Prod.* **64,** 813–815 (2001).

31. Kupchan, S. M. *et al.* Tumor inhibitors. 126. New cytotoxic neolignans from Aniba megaphylla Mez. *J. Org. Chem.* **43,** 586–590 (1978).

32. Putri, S. P., Kinoshita, H., Ihara, F., Igarashi, Y. & Nihira, T. Farinomalein, a Maleimide-Bearing Compound from the Entomopathogenic Fungus *Paecilomyces farinosus*. *J. Nat. Prod.* **72,** 1544–1546 (2009).

33. Elyashberg, M., Williams, A. & Blinov, K. in *New Developments in NMR* P001–P004 (Royal Society of Chemistry, 2011). doi:10.1039/9781849734578-FP001

34. Rateb, M. E., Tabudravu, J. & Ebel, R. in *Nuclear Magnetic Resonance* (ed. Ramesh, V.) **45,** 240–268 (Royal Society of Chemistry, 2016).

35. Elyashberg, M., Williams, A. J. & Blinov, K. Structural revisions of natural products by Computer-Assisted Structure Elucidation (CASE) systems. *Nat. Prod. Rep.* **27,** 1296 (2010).

36. Bremser, W. Hose — a novel substructure code. *Anal. Chim. Acta* **103,** 355–365 (1978).

37. Smurnyy, Y. D., Blinov, K. A., Churanova, T. S., Elyashberg, M. E. & Williams, A. J. Toward More Reliable ^{13}C and ^{1}H Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches. *J. Chem. Inf. Model.* **48,** 128–134 (2008).

38. Steinberger, L. & Leeb, H. Leave-one-out prediction intervals in linear regression models with many variables. (2016).