[Supporting Information]

PlantMAT: A Metabolomics Tool for Predicting the Specialized Metabolic Potential of a System and for Large-scale Metabolite Identifications

Feng Qiu,[1,2] Dennis D. Fine,[1] Daniel J. Wherritt,[1,3] Zhentian Lei,[1,2] and Lloyd W. Sumner[1,2,*]

[1]Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401

[2]Department of Biochemistry, University of Missouri, Schweitzer Hall, Columbia, MO 65211

[3]Department of Chemistry, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249

[*]Corresponding Author's Email: sumnerlw@missouri.edu

Contents **Page**

Experimental Sections

UHPLC-MS/MS

The column used is a Waters Acquity UPLC BEH C18 column (2.1 mm i.d., 150 mm length, 1.7 μm particles). A column flow of 560 μL/min was used. Solvents were water with 0.05% formic acid (A) and acetonitrile (B). The solvent gradient started at 95% A (aqueous) at time 0 and decreased to 30% A in 30 min. Then the gradient decreased further to 5% A in three min followed by 5% A for 3 min. The column is pre-condition column at 95% A for 4 min at the end of each run. The total run time is 40 min. The column is maintained at 60 °C by a column heater. The sample compartment temperature is maintained at 10 °C.

The mass spectrometer was a Bruker maXis impact quadrupole time of flight (QToF). An ion-source collision-induced dissociation (CID) energy of 45 eV is always present in the ESI. For MS/MS CID fragmentation high purity nitrogen gas was used. The electrospray conditions were end plate offset: 500 V; capillary transfer tube: 4000 V; nebulizer (nitrogen): 15 psi; drying gas (nitrogen): flow 6.0 L/min and temperature: 180 °C. A mass range of *m/z* 100–1500 was scanned at 6.0 Hz.

UHPLC-MS-SPE

For UHPLC-MS-SPE-NMR, the column eluent is split between the automated solid phase extractor (ASPE) unit and Bruker maXis QToF MS with a split of ~20/1. A Bruker NMR-MS Bridge is used for splitting the column eluent and for adding make-up flow to the split eluent before it enters the ESI source. The higher split flow goes through the flow cell of the PDA before it enters the ASPE unit. A make-up flow of 1.75 mL/min of water containing 0.05% formic acid is added to this flow prior to trapping. The smaller flow is directed towards the Bruker maXis QToF MS. A make-up flow of 50 μL/min acetonitrile is provided by the Bruker NMR-MS-Bridge.

Solid phase extraction was done using 1 mm i.d. × 10 mm length cartridges containing ~2.6 mg of Waters Oasis HLB resin (hydrophilic lipophilic balance). The cartridges are conditioned with 1 mL of acetonitrile and 1 mL of water containing 0.05% formic acid at 1 mL/min.

Peak trapping is accomplished with a Bruker/Spark-Holland Prospekt II controlled by the Bruker HyStar software. Fifteen to twenty injections of extract were performed. After trapping,

the cartridges are dried with nitrogen for 5 minutes. The extracted compounds are eluted from the cartridges using 30 μL of methanol-d_4 into a 1.7 mm NMR tube with a Gilson automatic liquid handler.

NMR

All NMR data were recorded and processed using Bruker's TopSpin 3.2 software. Proton spectra were acquired with WET solvent suppression (pulse program: wetdc). A total of 64–512 scans were collected into 16k data points. COSY (pulse sequence: cosygpmfppqf) spectra were run with 16–64 scans and 128 increments. HSQC (pulse sequence: hsqcedetgpsp.3) spectra were run with 32–512 scans and 128 increments. HMBC (pulse sequence: hmbcetgpl3nd) were run with 128 scans and 128 increments. 1D gradient selected TOCSY (pulse sequence: selmlgp) were recorded with mixing times of 0.02, 0.04, 0.08, and 0.16 seconds, respectively. 1D gradient selected ROESY (pulse sequence: selrogp) were recorded with a spinlock time of 400,000 μs. All spectra were run at temperature 300 K.

Table S1. Putative Identifications of Glycosides in *M. truncatula* Aerial Extracts[a]

Peak #	Rt /min	Exact mass	n_C	Aglycone	H	A	D	P	C	F	M	n_G	n_G'
15	3.1	449.1094	1	Eriodictyol	1	0	0	0	0	0	0	1	1
18	3.7	637.1404	4	Chrysoeriol	1	1	0	0	0	0	0	3	2
				Kaempferid	1	1	0	0	0	0	0	3	2
26	4.7	637.1045	3	Luteolin	0	2	0	0	0	0	0	2	2
				Fisetin	0	2	0	0	0	0	0	2	2
				Kaempferol	0	2	0	0	0	0	0	2	2
29	5	621.1102	3	Apigenin	0	2	0	0	0	0	0	2	2
				Galangin	0	2	0	0	0	0	0	2	2
				Genistein	0	2	0	0	0	0	0	2	2
34	5.5	431.0980	10	Apigenin	1	0	0	0	0	0	0	1	1
				Galangin	1	0	0	0	0	0	0	1	1
				Genistein	1	0	0	0	0	0	0	1	1
36	6.3	491.1199	1	Tricin	1	0	0	0	0	0	0	1	1
41	7.1	505.0986	2	Tricin	0	1	0	0	0	0	0	1	1
42	7.3	253.0507	3	Hispidol	0	0	0	0	0	0	0	1	1
				7,4'-Dihydroxyflavone	0	0	0	0	0	0	0	1	1
				Daidzein	0	0	0	0	0	0	0	1	1
46	7.6	767.1474	3	Apigenin	0	2	0	0	1	0	0	6	3
				Galangin	0	2	0	0	1	0	0	6	3
				Genistein	0	2	0	0	1	0	0	6	3
47	7.8	797.1551	5	Apigenin	0	2	0	0	0	1	0	6	6
				Galangin	0	2	0	0	0	1	0	6	6
				Genistein	0	2	0	0	0	1	0	6	6
48	8.6	767.1458	3	Apigenin	0	2	0	0	1	0	0	6	6
				Galangin	0	2	0	0	1	0	0	6	6
				Genistein	0	2	0	0	1	0	0	6	6
66	10.9	1383.6064	6	Zanhic acid	2	0	1	3	0	0	0	210	1
70	11.1	1265.5425	4	Zanhic acid	1	1	1	2	0	0	0	180	2
71	11.1	1235.5312	7	Zanhic acid	0	1	1	3	0	0	0	60	20
72	11.2	1265.5479	4	Zanhic acid	1	1	1	2	0	0	0	180	6

[a]The table gives the following results for each PlantMAT-identified peak: the number of all the possible aglycone/glycosyl/acyl combinations (n_C) from combinatorial enumeration; the most probable combination suggested by in silico MS/MS annotation (H–hexose, A–uronic acid, D–6-deoxyhexose, P–pentose, C–coumaric acid, F–ferulic acid, M–malonic acid); the number of all the possible glycosyl sequences (n_G); and the number of glycosyl sequences (n_G') which have the highest matching scores.

Table S1 (Continued)

Peak #	Rt /min	Exact mass	n_C	Aglycone	H	A	D	P	C	F	M	n_G	n_G'
73	11.3	1235.5360	7	Zanhic acid	3	0	1	0	0	0	1	60	60
				Zanhic acid	0	1	1	3	0	0	0	60	60
76	11.4	1103.4891	3	Zanhic acid	0	1	1	2	0	0	0	30	2
78	11.5	1383.6046	6	Zanhic acid	2	0	1	3	0	0	0	210	2
79	11.6	1383.6083	6	Zanhic acid	2	0	1	3	0	0	0	210	6
84	11.7	1251.5642	4	Zanhic acid	2	0	1	2	0	0	0	90	2
85	11.8	1089.5107	4	Zanhic acid	1	0	1	2	0	0	0	30	30
87	11.9	1089.5115	4	Zanhic acid	1	0	1	2	0	0	0	30	4
94	12.7	971.4836	9	Bayogenin	1	1	1	0	0	0	0	12	6
98	12.9	1381.5905	17	Medicagenic acid	1	1	1	3	0	0	0	420	2
102	13.1	1249.5496	8	Medicagenic acid	1	1	1	2	0	0	0	180	6
104	13.2	1219.5375	14	Medicagenic acid	0	1	1	3	0	0	0	60	1
105	13.3	1367.6154	14	Medicagenic acid	2	0	1	3	0	0	0	210	2
107	13.4	1087.4933	5	Medicagenic acid	0	1	1	2	0	0	0	30	2
108	13.4	1235.5684	9	Medicagenic acid	2	0	1	2	0	0	0	90	4
109	13.5	1205.5584	13	Medicagenic acid	1	0	1	3	0	0	0	60	2
119	14.3	1071.4987	6	Gypsogenic acid	0	1	1	2	0	0	0	30	6
				Polygalagenin	0	1	1	2	0	0	0	30	6
129	15.5	911.4655	5	Medicagenic acid	0	0	1	2	0	0	0	6	6
135	16.8	1013.5320	12	Soyasapogenol B	2	0	1	0	0	0	1	30	3
136	16.9	1043.5452	9	Soyasapogenol B	0	1	1	2	0	0	0	30	4
137	17.3	1011.5165	11	Soyasapogenol B	0	1	2	0	0	0	1	30	30
138	17.3	911.5017	9	Soyasapogenol B	0	1	1	1	0	0	0	12	2
139	17.4	941.5095	10	Soyasapogenol B	1	1	1	0	0	0	0	12	1
140	17.4	1129.5410	10	Soyasapogenol B	0	1	1	2	0	0	1	180	2
144	18.0	997.4997	10	Soyasapogenol B	0	1	1	1	0	0	1	60	4
145	18.1	1027.5080	15	Soyasapogenol B	1	1	1	0	0	0	1	60	4

Table S2. Putative Identifications of Glycosides in *M. truncatula* Root Extracts

Peak #	Rt /min	Exact mass	n_C	Aglycone	H	A	D	P	C	F	M	n_G	n_G'
12	3.8	433.1140	2	Naringenin	1	0	0	0	0	0	0	1	1
34	9.7	515.1156	5	Formononetin	1	0	0	0	0	0	1	2	1
38	11.0	517.1335	5	Medicarpin	1	0	0	0	0	0	1	2	1
43	12.1	987.4831	4	Bayogenin	2	1	0	0	0	0	0	6	3
44	12.1	1059.4980	5	Bayogenin	3	0	0	0	0	0	1	10	2
45	12.4	973.5019	2	Bayogenin	3	0	0	0	0	0	0	2	2
46	12.5	825.4289	4	Bayogenin	1	1	0	0	0	0	0	3	2
47	12.7	971.4842	9	Hederagenin	2	1	0	0	0	0	0	6	3
48	12.7	987.4811	4	Medicagenic acid	3	0	0	0	0	0	0	2	2
51	13.0	971.4854	9	Hederagenin	2	1	0	0	0	0	0	6	1
52	13.1	825.4260	4	Medicagenic acid	2	0	0	0	0	0	0	2	2
53	13.2	867.4378	2	Bayogenin	1	0	0	1	0	0	1	12	2
56	13.4	809.4324	8	Hederagenin	1	1	0	0	0	0	0	3	2
57	13.5	823.4137	5	Gypsogenic acid	1	1	0	0	0	0	0	3	2
				Polygalagenin	1	1	0	0	0	0	0	3	2
58	13.5	911.4294	4	Medicagenic acid	2	0	0	0	0	0	1	6	3
61	13.8	971.4850	9	Hederagenin	2	1	0	0	0	0	0	6	3
64	14.2	851.4438	4	Hederagenin	1	0	0	1	0	0	1	12	6
66	14.3	865.4245	5	Gypsogenic acid	1	0	0	1	0	0	1	12	2
				Polygalagenin	1	0	0	1	0	0	1	12	2
68	14.5	955.4908	15	Soyasapogenol E	2	1	0	0	0	0	0	6	3
70	14.7	1043.5070	10	Hederagenin	3	0	0	0	0	0	1	10	6
71	14.7	927.4935	5	Hederagenin	2	0	0	1	0	0	0	6	3
72	14.8	955.4903	15	Soyasapogenol E	2	1	0	0	0	0	0	6	3
75	15.2	809.4345	8	Hederagenin	1	1	0	0	0	0	0	3	2
82	16.0	851.4441	4	Hederagenin	1	0	0	1	0	0	1	12	2
88	16.5	825.4278	4	Medicagenic acid	2	0	0	0	0	0	0	2	2
91	16.8	663.3745	3	Medicagenic acid	1	0	0	0	0	0	0	1	1
94	17.0	663.3763	3	Medicagenic acid	1	0	0	0	0	0	0	1	1
96	17.4	941.5122	10	Soyasapogenol B	1	1	1	0	0	0	0	12	12

[a]The table gives the following results for each PlantMAT-identified peak: the number of all the possible aglycone/glycosyl/acyl combinations (n_C) from combinatorial enumeration; the most probable combination suggested by in silico MS/MS annotation (H–hexose, A–uronic acid, D–6-deoxyhexose, P–pentose, C–coumaric acid, F–ferulic acid, M–malonic acid); the number of all the possible glycosyl sequences (n_G); and the number of glycosyl sequences (n_G') which have relative matching score of 1.00.

Table S2 (Continued)

Peak #	Rt /min	Exact mass	n_C	Aglycone	H	A	D	P	C	F	M	n_G	n_G'
97	17.8	647.3797	5	Hederagenin	0	1	0	0	0	0	0	1	1
99	18.1	1027.5130	15	Soyasapogenol B	1	1	1	0	0	0	1	60	12
100	18.2	661.3604	2	Gypsogenic acid	0	1	0	0	0	0	0	1	1
				Polygalagenin	0	1	0	0	0	0	0	1	1
103	18.8	939.4963	13	Soyasapogenol E	1	1	1	0	0	0	0	12	4
104	19.5	1025.4980	13	Soyasapogenol E	1	1	1	0	0	0	1	60	8

Table S3. MS/MS Spectra of Ten Glycosides Used for the Validation of PlantMAT

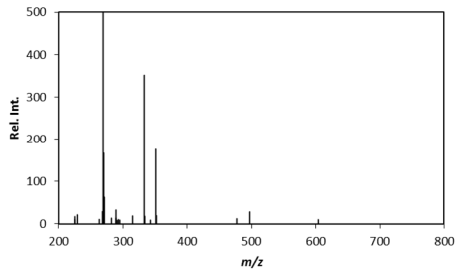
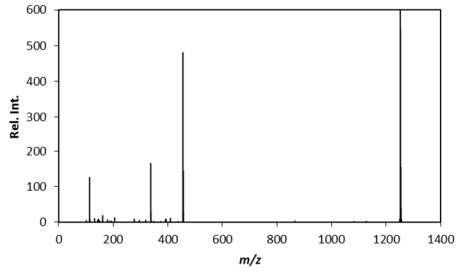
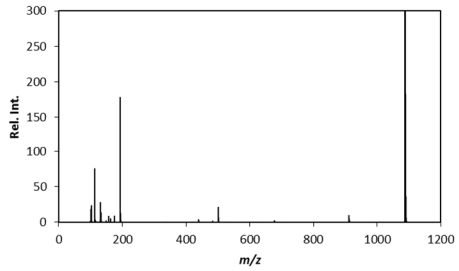
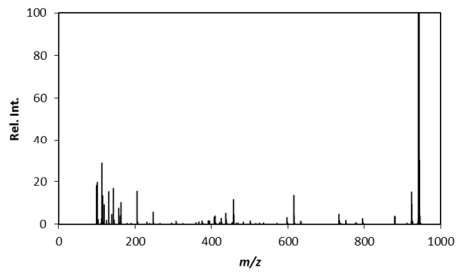
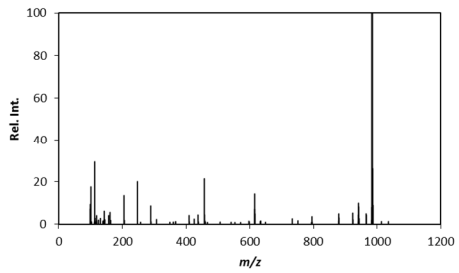
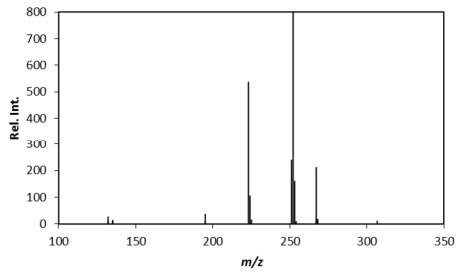
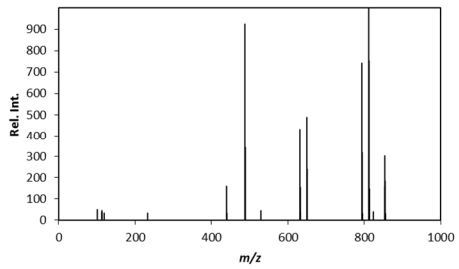
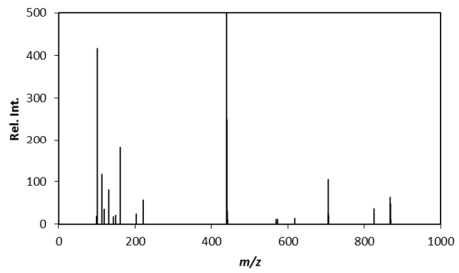
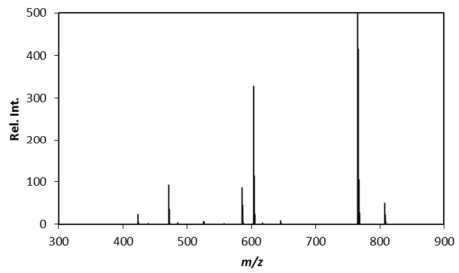
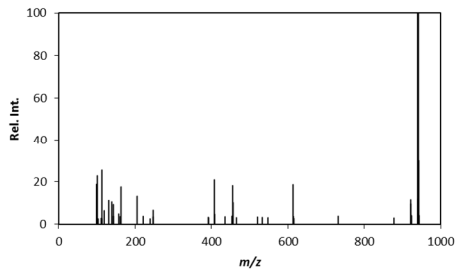
Peak #	MS/MS Spectra	Annotated Ions (m/z, % base peak, annotation)		
A-46		269.0459	30.17	$[Agly-H]^-$
		497.1043	0.86	$[M-Cou-2H_2O-2CO_2-H]^-$
A-84		455.3159	16.37	$[Agly-H_2O-CO_2-H]^-$
		865.4574	0.10	$[M-2Hex-H_2O-CO_2-H]^-$
		1251.5626	34.06	$[M-H]^-$
A-107		439.3210	0.16	$[Agly-H_2O-CO_2-H]^-$
		465.3003	0.00	$[Agly-H_2O-H_2O-H]^-$
		483.3127	0.07	$[Agly-H_2O-H]^-$
		501.3196	0.92	$[Agly-H]^-$
		677.3464	0.08	$[M-dHex-2Pen-H]^-$
		911.4588	0.41	$[M-HexA-H]^-$
		1087.4943	44.07	$[M-H]^-$
A-139		457.3653	0.56	$[Agly-H]^-$
		615.3871	0.66	$[M-Hex-dHex-H_2O-H]^-$
		733.4494	0.23	$[M-dHex-H_2O-CO_2-H]^-$
		751.4688	0.09	$[M-dHex-CO_2-H]^-$
		795.4623	0.09	$[M-dHex-H]^-$
		923.5022	0.73	$[M-H_2O-H]^-$
		941.5100	48.46	[M−H]−
A-145		457.3682	1.04	$[Agly-H]^-$
		615.3904	0.70	$[M-Hex-dHex-Mal-H_2O-H]^-$
		879.5013	0.24	$[M-Mal-H_2O-CO_2-H]^-$
		923.5032	0.26	$[M-Mal-H_2O-H]^-$
		941.5175	0.21	$[M-Mal-H]^-$
		965.5074	0.24	$[M-H_2O-CO_2-H]^-$
		983.5215	48.57	$[M-CO_2-H]^-$

Table S3 (Continued)

Peak #		Annotated Ions (m/z, % base peak, annotation)		
R-34		267.0664	70.00	$[Agly-H]^-$
		335.0532	2.22	$[M-Hex-H_2O-H]^-$
R-44		487.3393	13.68	$[Agly-H]^-$
		529.3522	0.66	$[M-3Hex-CO_2-H]^-$
		631.3828	6.39	$[M-2Hex-Mal-H_2O-H]^-$
		649.3908	7.23	$[M-2Hex-Mal-H]^-$
		793.4280	10.98	$[M-Hex-Mal-H_2O-H]^-$
		811.4445	14.80	$[M-Hex-Mal-H]^-$
		853.4591	4.48	$[M-Hex-CO_2-H]^-$
R-58		439.3224	38.02	$[Agly-H_2O-CO_2-H]^-$
		705.3799	4.00	$[M-Hex-CO_2-H]^-$
		825.4133	1.38	$[M-Mal-H]^-$
		867.4434	2.41	$[M-CO_2-H]^-$
R-71		471.3475	3.70	$[Agly-H]^-$
		585.3777	3.44	$[M-2Hex-H_2O-H]^-$
		603.3881	13.15	$[M-2Hex-H]^-$
		765.4421	40.14	$[M-Hex-H]^-$
R-103		455.3537	0.91	$[Agly-H]^-$
		613.3712	0.94	$[M-Hex-dHex-H_2O-H]^-$
		731.4391	0.19	$[M-dHex-H_2O-CO_2-H]^-$
		877.4881	0.15	$[M-H_2O-CO_2-H]^-$
		921.4814	0.58	$[M-H_2O-H]^-$
		939.4936	50.23	$[M-H]^-$

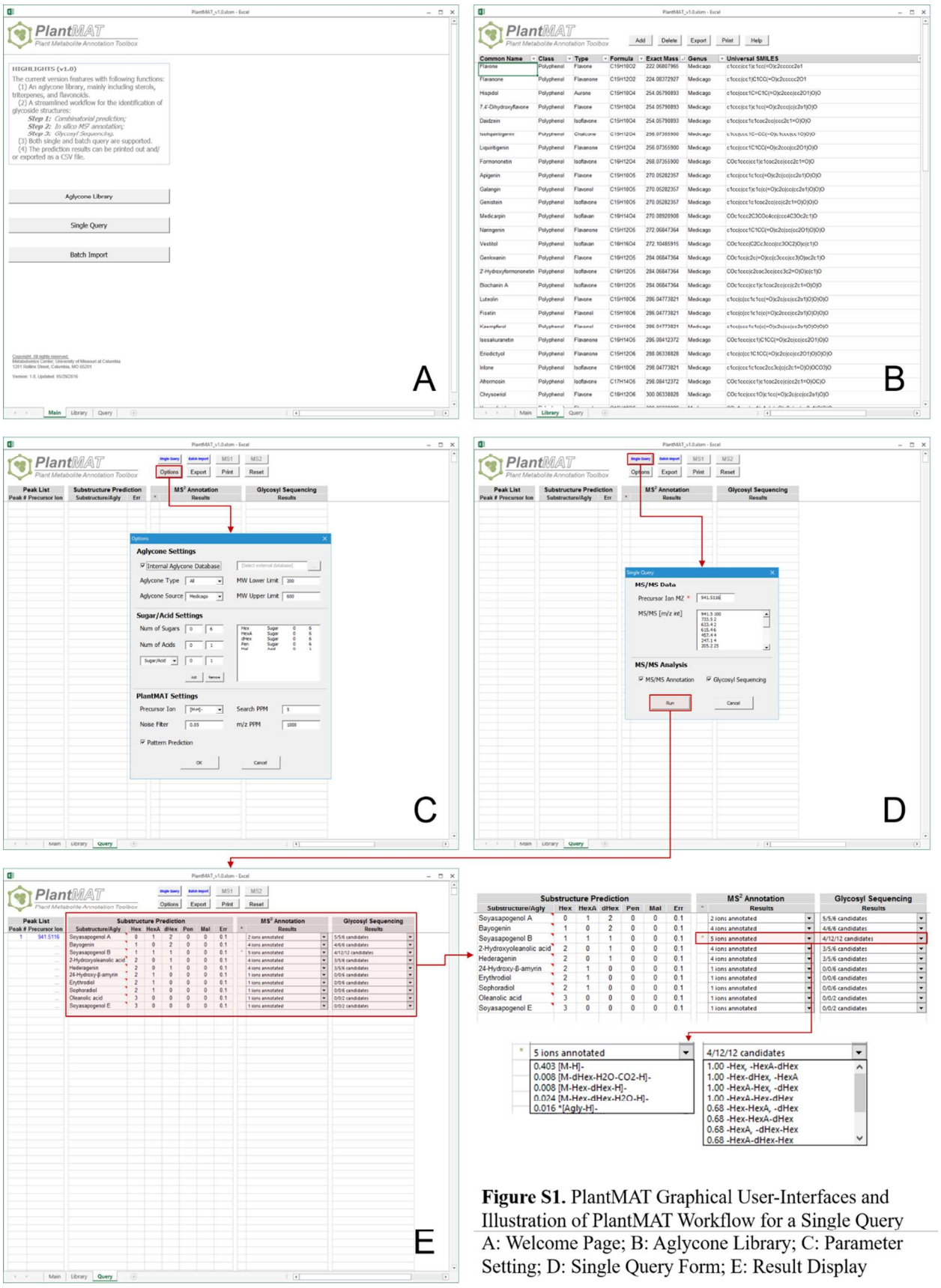


Figure S1. PlantMAT Graphical User-Interfaces and Illustration of PlantMAT Workflow for a Single Query
A: Welcome Page; B: Aglycone Library; C: Parameter Setting; D: Single Query Form; E: Result Display

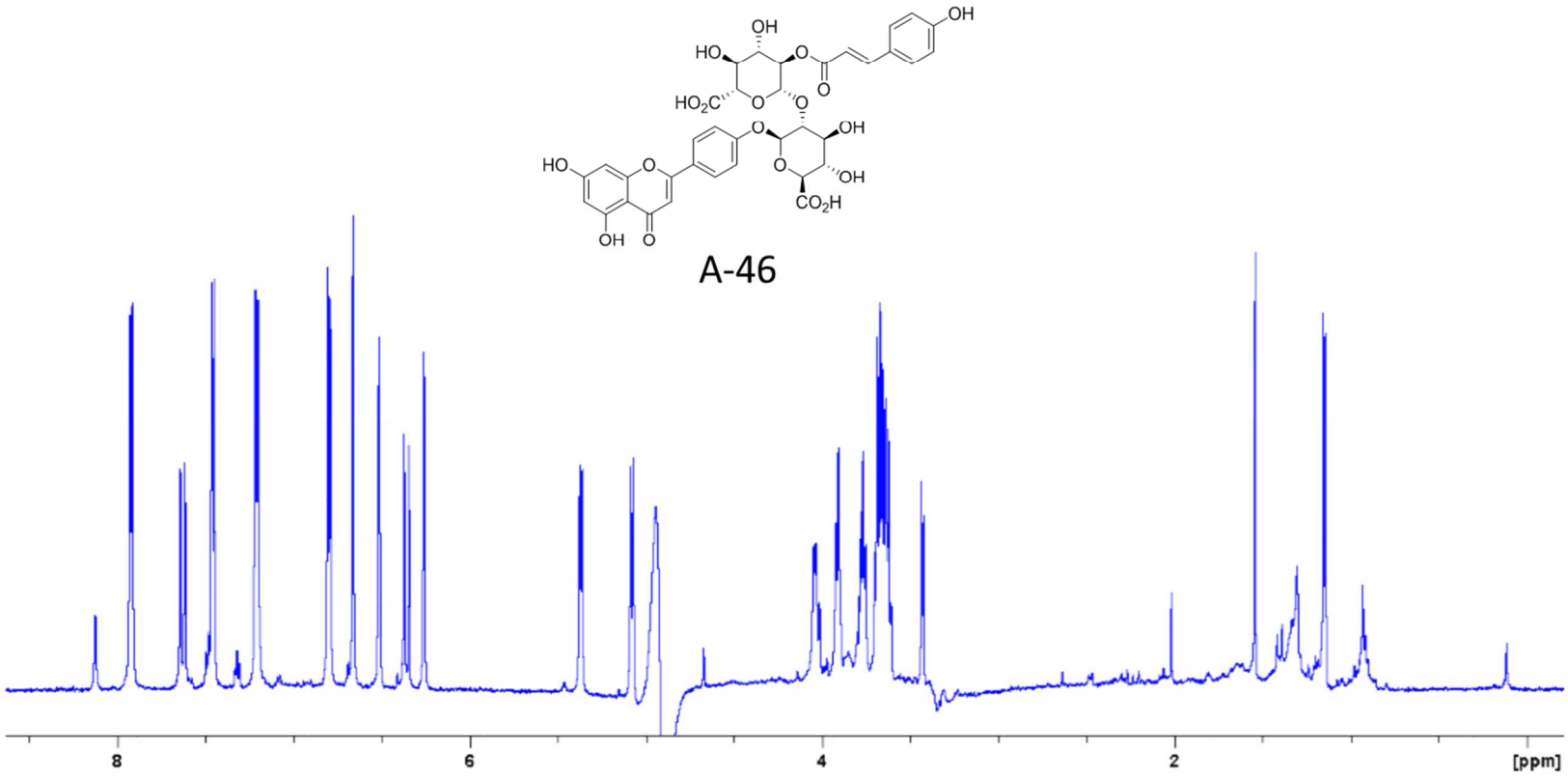


Figure S2. ^{1}H NMR Spectrum of Peak A-46.

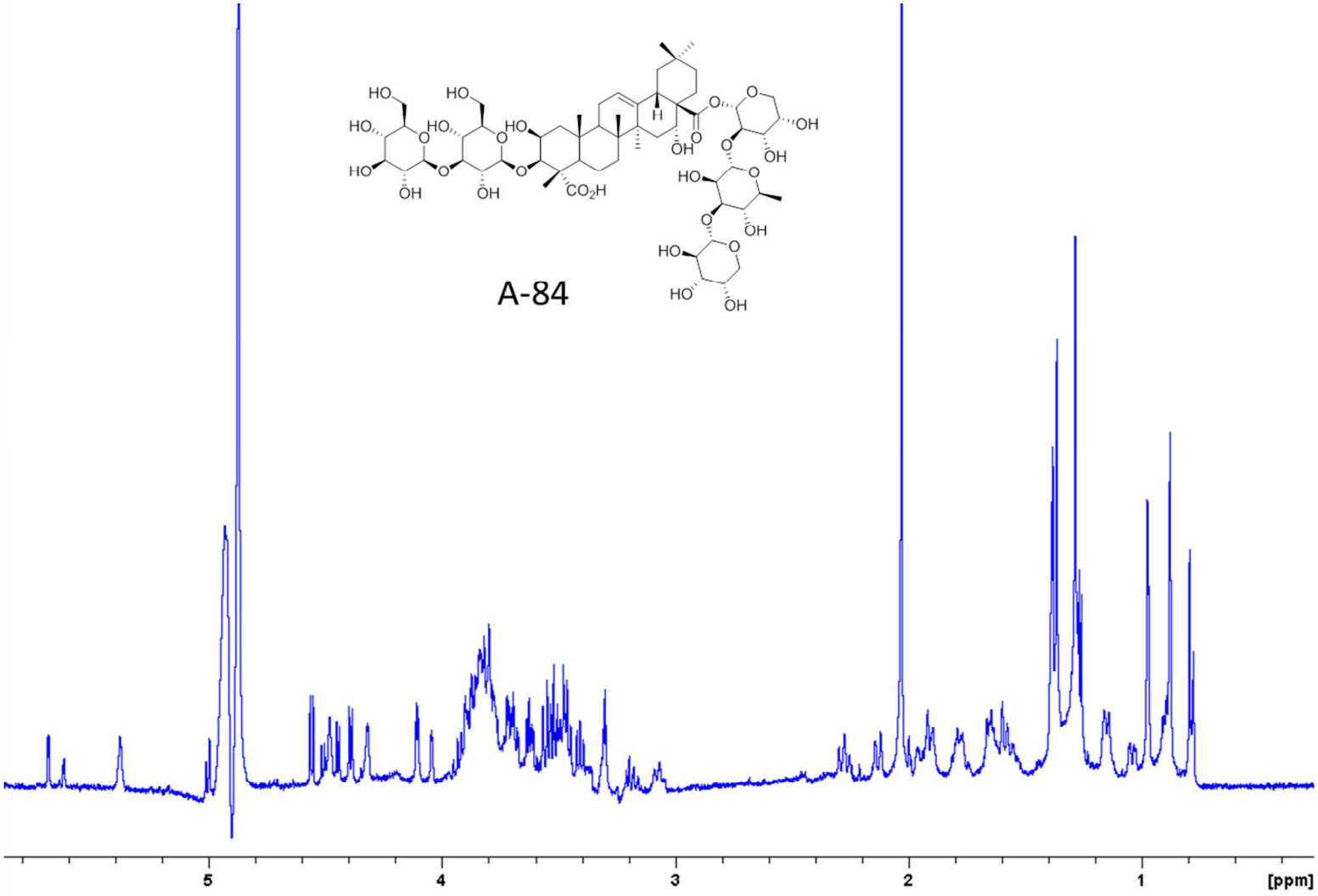


Figure S3. ^{1}H NMR Spectrum of Peak A-84.

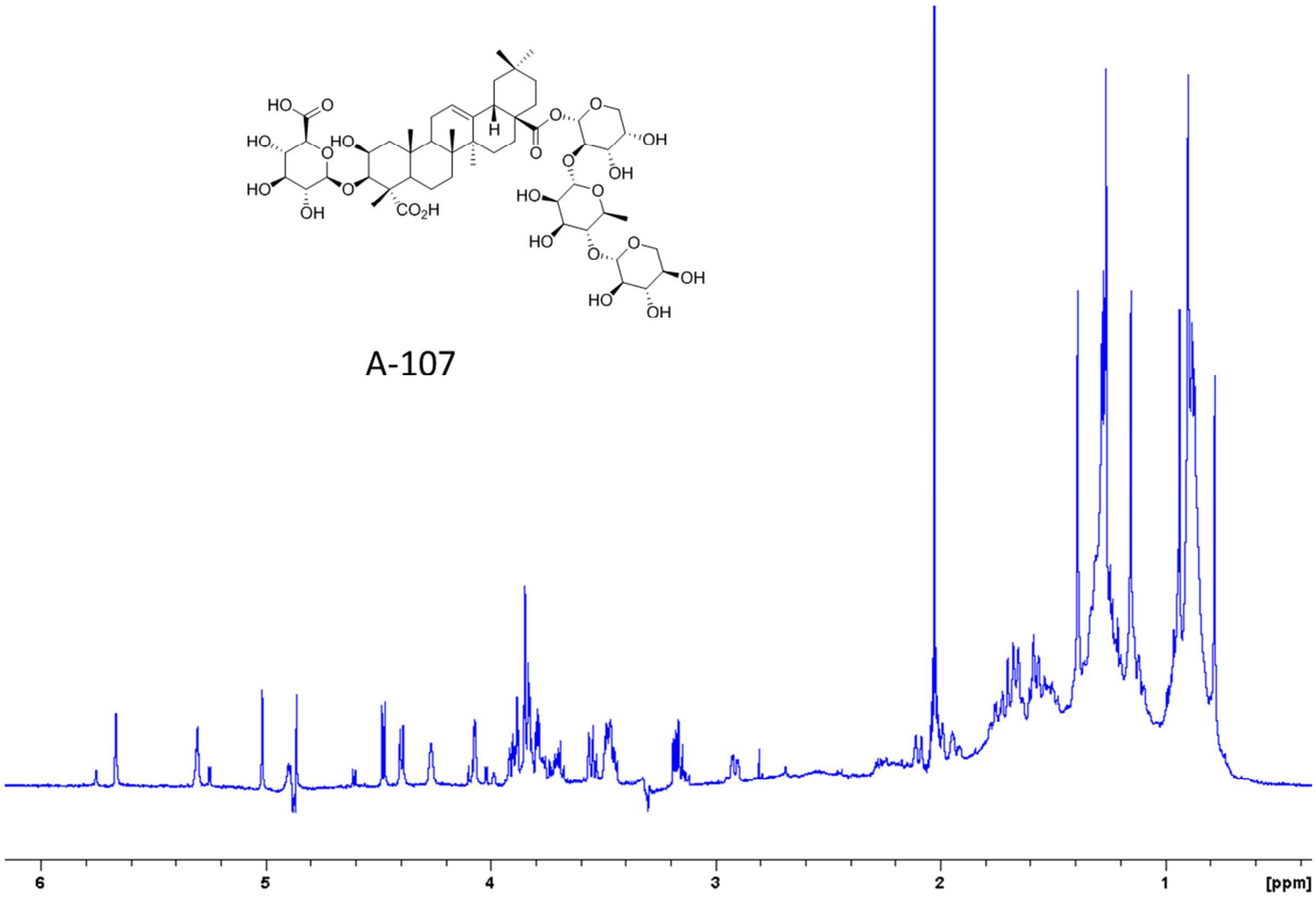


Figure S4. ^{1}H NMR Spectrum of Peak A-107.

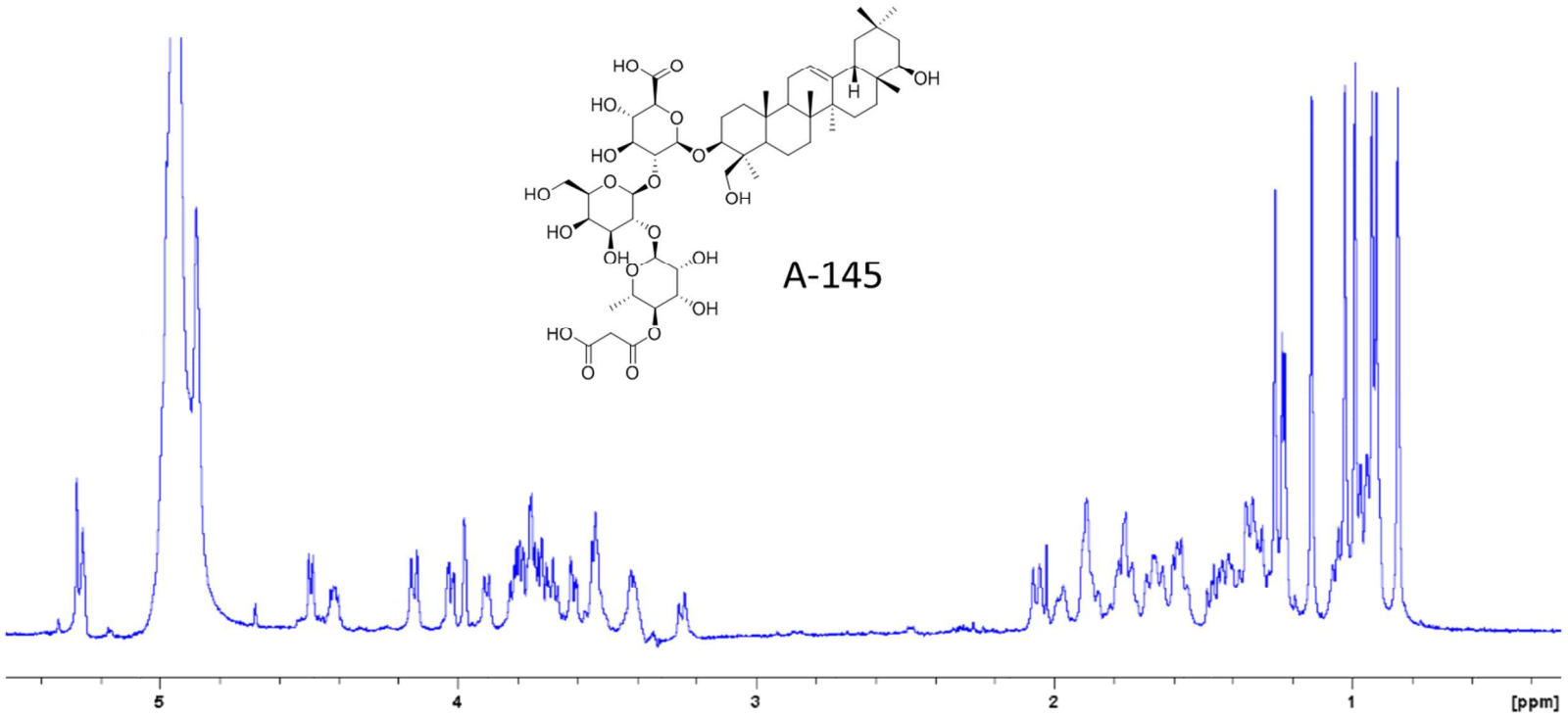


Figure S5. ^{1}H NMR Spectrum of Peak A-145.

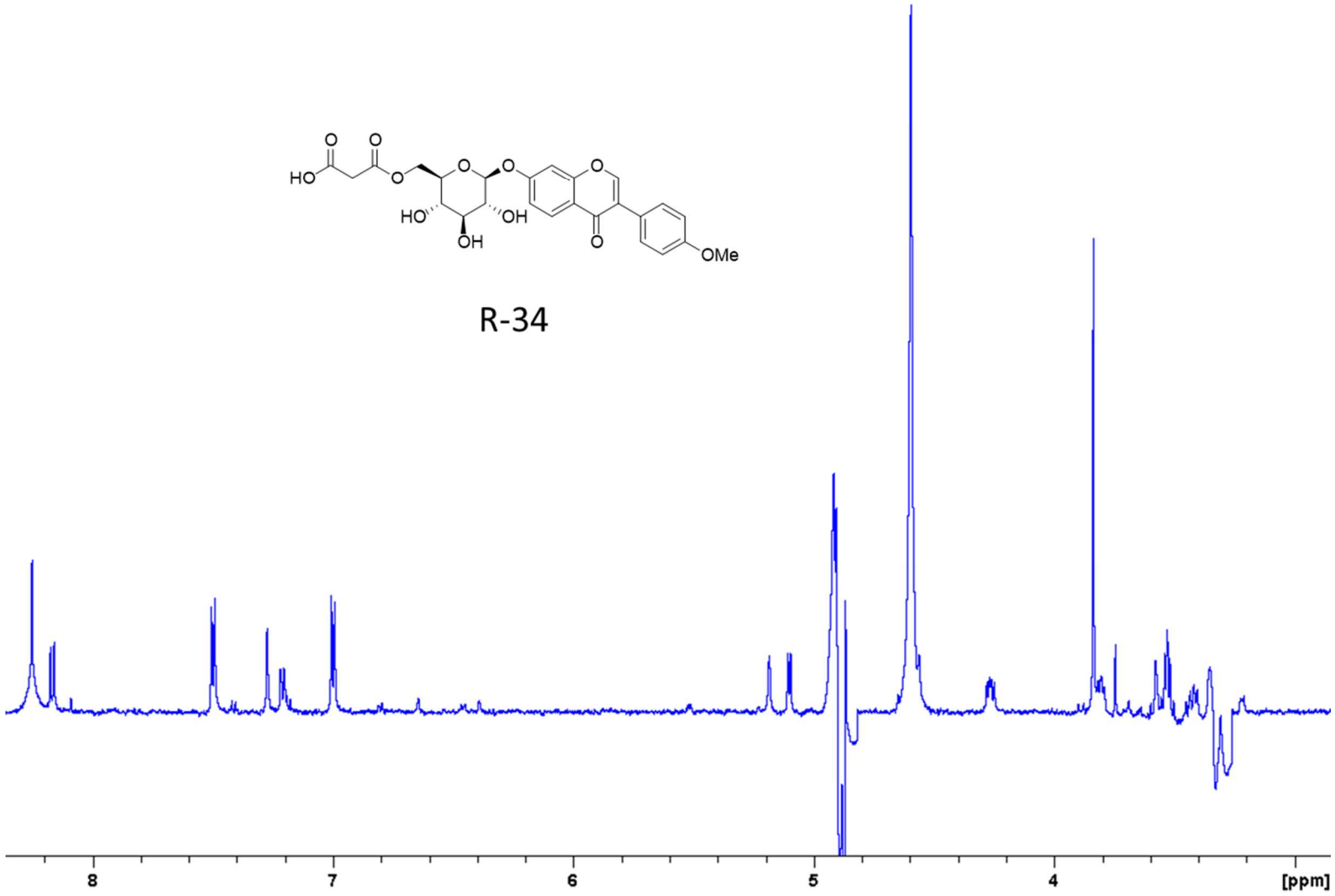


Figure S6. ^{1}H NMR Spectrum of Peak R-34.

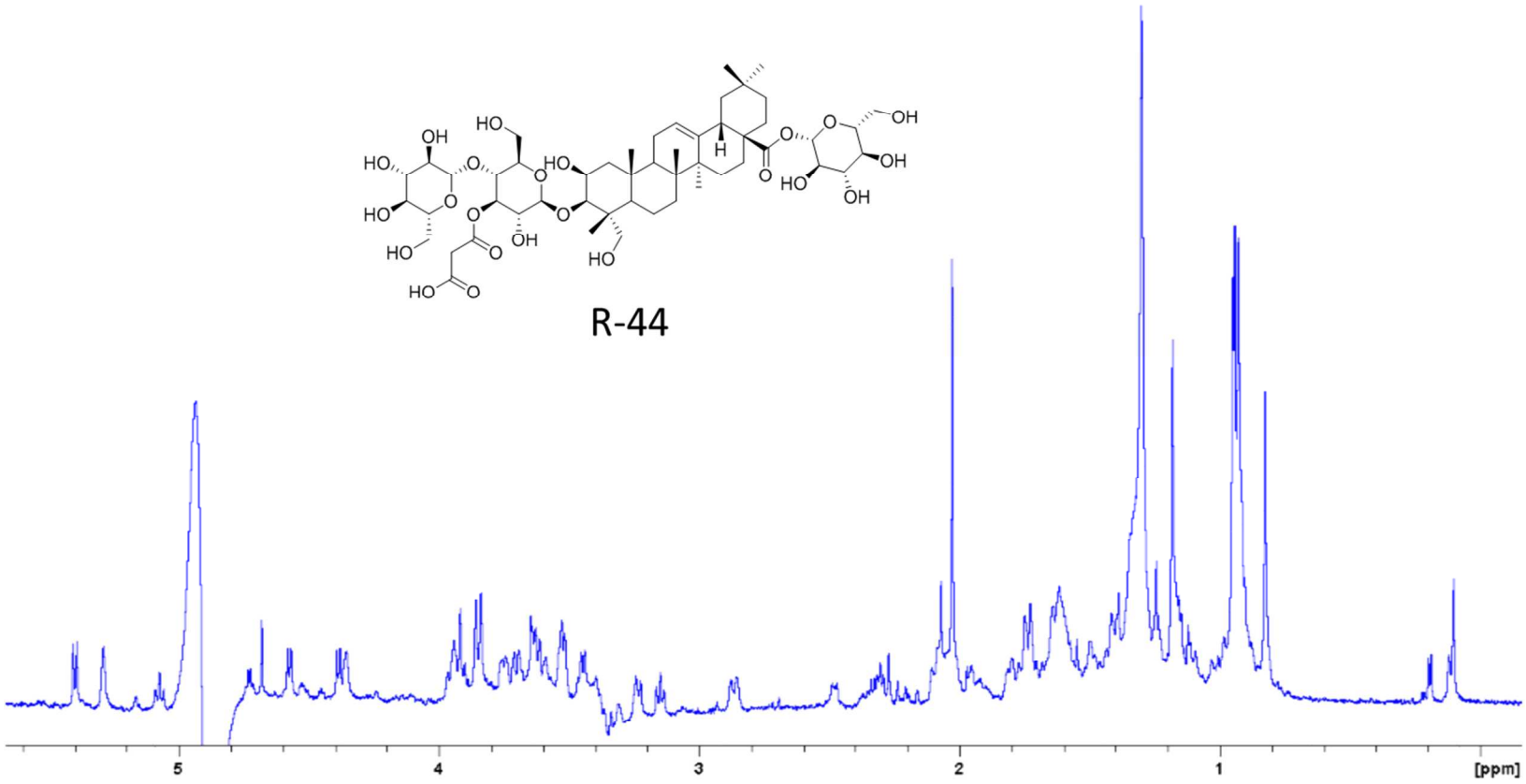


Figure S7. ^{1}H NMR Spectrum of Peak R-44.

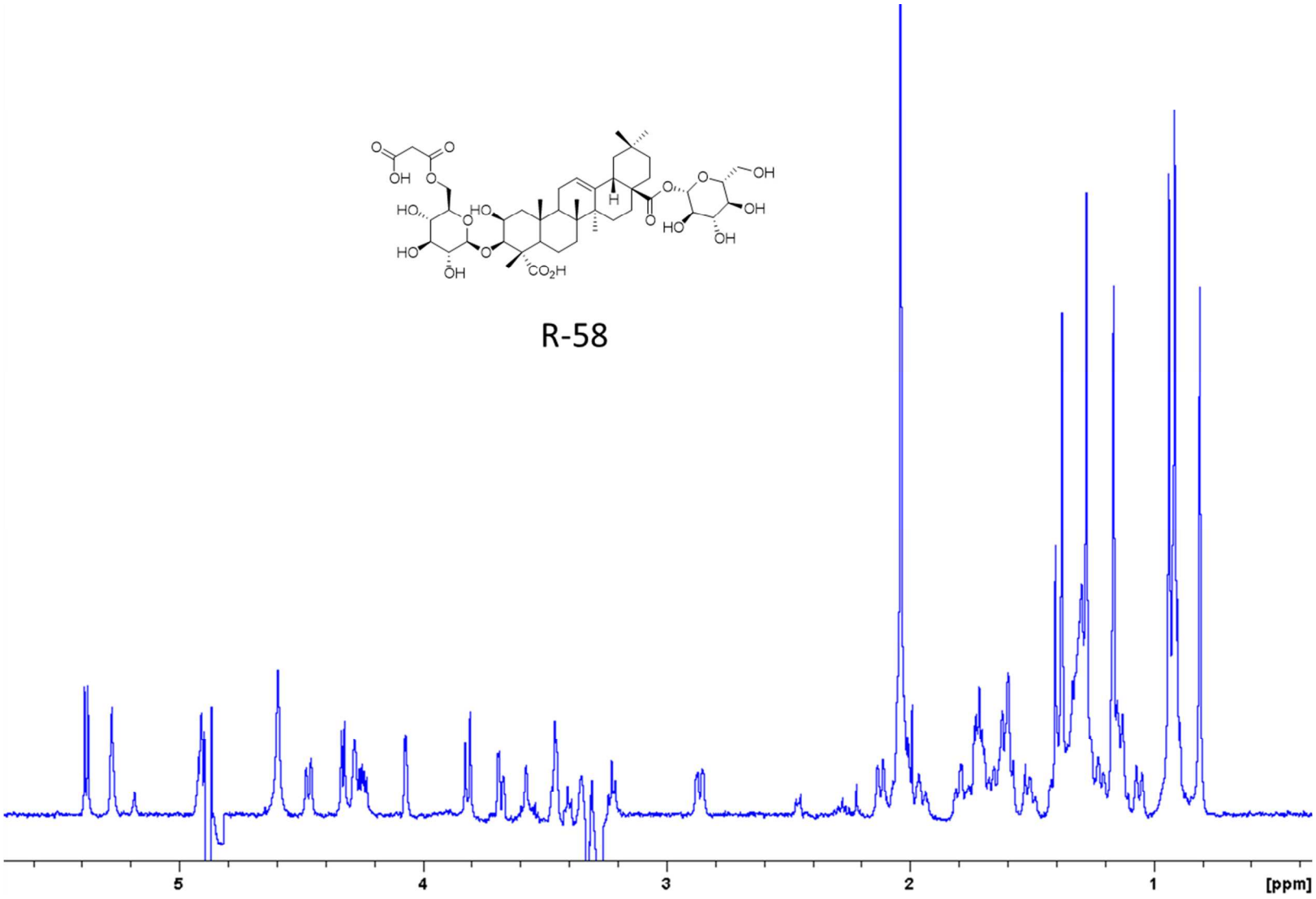


Figure S8. ^{1}H NMR Spectrum of Peak R-58.

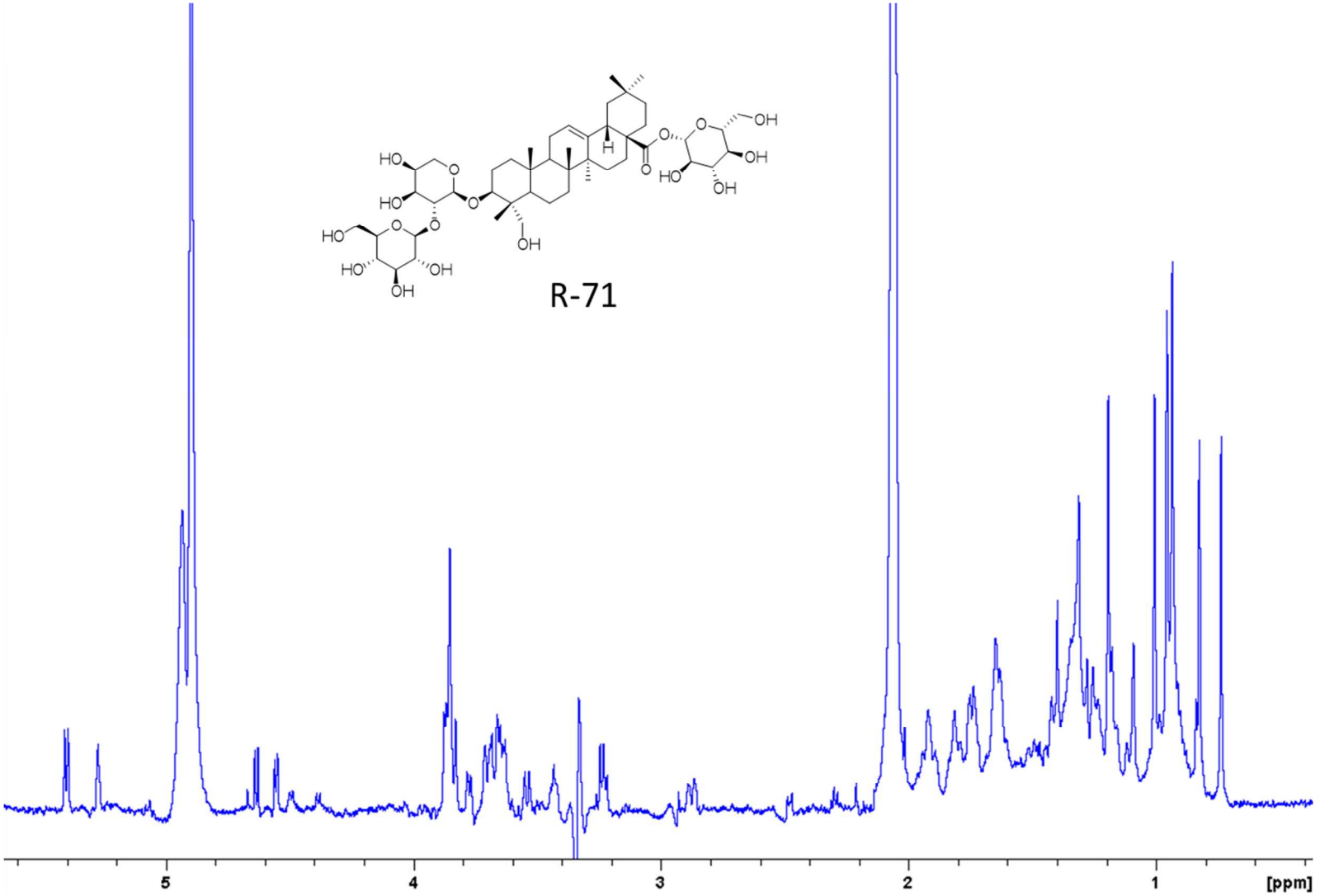


Figure S9. ^{1}H NMR Spectrum of Peak R-71.

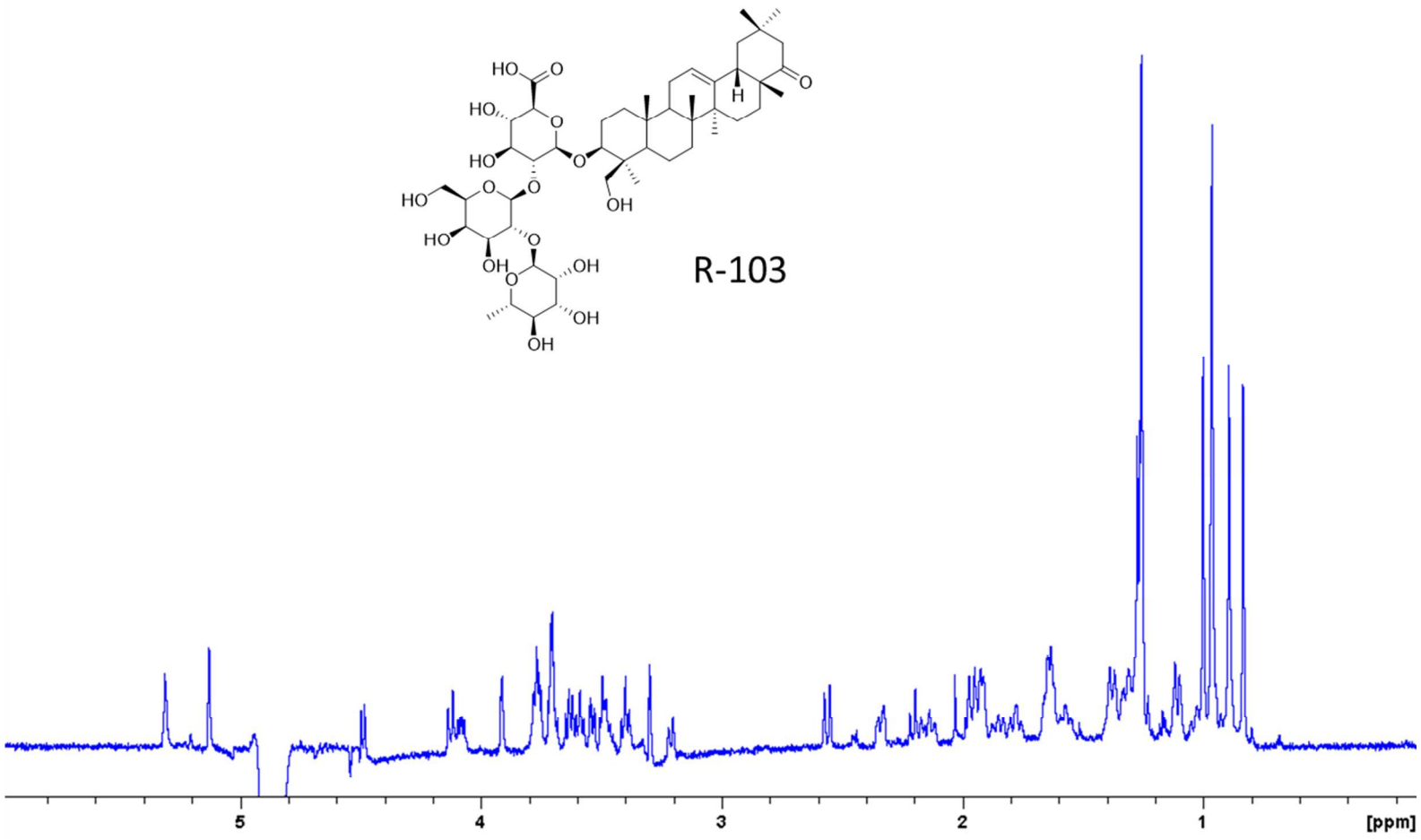


Figure S10. ^{1}H NMR Spectrum of Peak R-103.