Hybrid Transcription Factor Engineering Activates the Silent Secondary Metabolite Gene Cluster for (+)-Asperlin in *Aspergillus nidulans*

Michelle F. Grau[†], Ruth Entwistle[‡], Yi-Ming Chiang[†§], Manmeet Ahuja[‡∇], C. Elizabeth Oakley[‡], Tomohiro Akashi[‖], Clay C. C. Wang*[†⊥], Richard B. Todd*[#], Berl R. Oakley*[‡]

[†] Department of Pharmacology and Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, California 90089, United States

[‡] Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66045, United States

[§] Department of Pharmacy, Chia Nan University of Pharmacy and Science, Tainan City 71710, Taiwan

[‖] Division of OMICS analysis, Center for Neurological Diseases and Cancer, Nagoya University Graduate School of Medicine, Nagoya, Japan

[⊥] Department of Chemistry, University of Southern California, Dornsife College of Letters, Arts, and Sciences, Los Angeles, California 90089, United States

[#] Department of Plant Pathology, Kansas State University, 4024 Throckmorton Plant Sciences Center, Manhattan, KS 66506, United States

[∇] Industrial Biotechnology Division, Reliance Technology Group, Reliance Industries Limited, Reliance Corporate Park, Thane Belapur Road, Ghansoli, Navi, Mumbai 400701, India.

*Corresponding authors: clayw@usc.edu; rbtodd@ksu.edu; boakley@ku.edu

Table of Contents

1. SUPPLEMENTAL METHODS

1.1 Isolation of Secondary Metabolites

For scaling up to isolate compound **1**, 1 L of LMM (20 125-mL flasks were used containing 50 mL of medium each) inoculated with 1.0 x 10^9 spores L^{-1} of *A. nidulans* strain LO4909 was incubated at 37°C with shaking at 180 rpm. For *alcA*(p) induction, 50mM of MEK was added to the culture(s) 42 h after inoculation. Culture medium were collected 72 h after induction by vacuum filtration. The culture medium partitioned with ethyl acetate (EtOAc; 1 L) after acidification by 1N HCl to pH = 3. The EtOAc layer was collected and evaporated *in vacuo* to yield compound **1** without further purification.

For scaling up to isolate compound **2**, 2 L of LMM (2 2-Liter flasks were used containing 1 L of medium each) inoculated with 1.0 x 10^9 spores L^{-1} of *A. nidulans* strain LO9721 was incubated at 37°C with shaking at 180 rpm. For *alcA*(p) induction, 50 mM of MEK was added to the culture(s) 42 h after inoculation. Culture medium and hyphae were collected 72 h after induction by vacuum filtration. The culture medium partitioned with ethyl acetate (EtOAc; 2 L), and the EtOAc layer was evaporated *in vacuo* (crude extract 184.7 mg). Thin Layer Chromatography was carried out (Merck TLC Silica Gel 60 RP-C_{18} F_{254S} glass plates 20 x 20 cm) on the crude extract, with the correct compound identified by UV visualization. A razor blade was used to scrape the silica containing the product off the plate. The silica was placed in a fritted funnel and flushed with EtOAc. The filtrate was collected and the solvent was removed *in vacuo* resulting in the isolation of **(2)** (114.4 mg).

1.2 Detailed Structural Characterization

Compound **1** was isolated as a white amorphous powder. The molecular formula was found to be $C_8H_{10}O_2$ by its ^1H NMR, ^{13}C NMR (Figures S7-S9) and HRESIMS spectral data (Figure S6), representing four indices of hydrogen deficiency (IHD). The ^1H and ^{13}C NMR in $CDCl_3$ exhibited signals for three disubstituted olefins [δ_H 5.67, 5.97, 6.37, 6.62, 7.15, and 7.18 (each 1H); δ_C 116.0, 122.0, 126.0, 136.1, 138.7, 141.0], one carboxylic acid [δ_H 10.87 (1H, br s); δ_C 172.3], and one methyl group [1.85 (3H, br d, J = 6.8 Hz)]. This together with the molecular formula of **1** indicated that **1** is a linear trienoic acid. 2D NMR correlations (^1H-^1H COSY, gHMQC and gHMBC) also support the structure (data not shown). The double bond configurations were determined to be 2*Z*, 4*Z*, and 6*E* based on the coupling constants of H-2 and H-3 (J = 11.3 Hz), H-3 and H-4 (J = 11.1 Hz), and H-4 and H-5 (J = 14.6 Hz). Therefore, compound 1 was assigned as (2*Z*,4*Z*,6*E*)-octa-2,4,6-trienoic acid.

Compound **2** was isolated as a colorless oil. The molecular formula was found to be $C_{10}H_{14}O_6$ from HRESIMS (Figure S6) and both ^1H and ^{13}C NMR data (Figures S7, and S10-S11), indicating **2** has four IHD. ^{13}C NMR spectrum of **2** exhibited one olefin (δ_C 124.8 and 140.5) and two ester or carboxylic acid carbonyl carbons (δ_C 161.46 and 169.70) in the down-field region. ^1H, ^{13}C, gHSQC NMR spectra indicated compound **2** contains two methyl [δ_H 1.36 (3H, d, J = 5.2 Hz), δ_C 17.0 (q) and δ_H 2.11 (3H, s), δ_C 20.5 (q)] and four oxymethine groups [δ_H 3.03 (1H, dd, J = 7.0, 2.1 Hz), δ_C 54.9 (d); δ_H 3.06 (1H, dq, J = 5.2, 2.1 Hz), δ_C 54.5 (d); δ_H 4.09 (1H, dd, J =7.2, 2.8 Hz), δ_C 78.8 (d); and δ_H 5.29 (1H, dd, J = 5.7, 2.8 Hz), δ_C 62.1 (d)]. Because compound **2** has four IHD but only contains one olefin and two ester or carboxylic acid carbonyl carbons, **2** must contain a cyclic ether or an epoxide moiety. This, together with the fact that two oxymethine groups (δ_H 3.03, δ_C 54.8 and δ_H 3.06, δ_C 54.5) coupling to each other has a relatively high field chemical shift, indicated that **2** has an epoxide functional group. HSQC and long-range HMBC correlations allowed full assignment of the structure (Figure S7). Comparison of ^{1}H and ^{13}C NMR data of **2** with (+)-Asperlin in the literature[1] confirmed the identity of compound **2**.

References:

1. Argoudelis, A. D.; Zieserl, J. F. The Structure of U-13,933, a New Antibiotic. *Tetrahedron Lett.* **1966**, *7* (18), 1969–1973.

2. FIGURES

Figure S1. Correct coding sequence for *afoA* and corresponding amino acid sequence of its protein product (AfoA). Intron sequences are shown in red.

```
    M  A  C  P   T  R  R   G  R  Q   Q  P  G  F   A  C  E
    ATGGCGTGTC CCACCAGACG AGGACGACAG CAGCCCGGCT TTGCATGCGA
  1 ---------+ ---------+ ---------+ ---------+ ---------+ 50

     E  C  R   R  R  K  A   R  C  D   R  V  R   P  K  C  G
    GGAGTGTCGC CGCCGCAAAG CGCGCTGTGA TCGCGTGCGT CCGAAATGCG
 51 ---------+ ---------+ ---------+ ---------+ ---------+ 100

      F  C  T   E  N  E   L  Q  C  V   F  V  D   K  R  Q
    GGTTCTGCAC TGAGAATGAG CTGCAGTGTG TGTTCGTTGA CAAGAGGCAG
101 ---------+ ---------+ ---------+ ---------+ ---------+ 150

    Q  R  G  P   I  K  G   Q  I  T   S  M  Q  S   Q  L  A
    CAGAGGGGTC CGATCAAAGG GCAGATCACC TCGATGCAGT CGCAGCTGGG
151 ---------+ ---------+ ---------+ ---------+ ---------+ 200

    TAGGTGTTTG TCTTGTCTCA TTGTATCTCG TCTCGTCTGC GCTTTTGTGA
201 ---------+ ---------+ ---------+ ---------+ ---------+ 250

    TTATGGGGCT GCCATGTTTC CGGTCCGGAC ACAGGCATCT GCAAGGCCCG
251 ---------+ ---------+ ---------+ ---------+ ---------+ 300

    CCGCTGTGCT CCCCCGATCT GCAGGGACCA ATGCAGCTGG TTCTGGAGCT
301 ---------+ ---------+ ---------+ ---------+ ---------+ 350

    TGTGCTGTGC TGCTTCCCTG TCTTTCCACA TGGTCGAGTC GAGCGAGCTA
351 ---------+ ---------+ ---------+ ---------+ ---------+ 400

                                     T  L  R  W   Q  L  D
    GCTAACATGG GATGCCTCAT GCTTTCAGCA ACGCTTCGAT GGCAGCTTGA
401 ---------+ ---------+ ---------+ ---------+ ---------+ 450

     R  Y  L   R  H  R  P   P  P  S   I  T  M   A  G  E  L
    TCGATACCTG CGACATCGAC CTCCCCCGTC CATAACCATG GCCGGCGAGC
451 ---------+ ---------+ ---------+ ---------+ ---------+ 500

      D  E  P   P  A  D   I  Q  T  M   L  D  D   F  D  V
    TCGATGAGCC ACCAGCGGAT ATCCAGACGA TGCTGGATGA CTTTGATGTA
501 ---------+ ---------+ ---------+ ---------+ ---------+ 550

    Q  V  A  A   L  K  Q   D  A  T   A  T  T  T   M  S  T
    CAGGTCGCCG CGCTGAAGCA GGATGCCACG GCAACCACCA CAATGTCGAC
551 ---------+ ---------+ ---------+ ---------+ ---------+ 600

     S  T  A   L  M  P  A   P  A  I   S  S  K   D  A  A  P
    GTCGACAGCT CTCATGCCTG CCCCAGCCAT CTCATCTAAA GATGCTGCTC
601 ---------+ ---------+ ---------+ ---------+ ---------+ 650

      A  G  A   G  L  S   W  P  D  P   T  W  L   D  R  Q
    CTGCTGGTGC TGGTTTATCG TGGCCTGACC CAACCTGGCT GGATCGCCAG
651 ---------+ ---------+ ---------+ ---------+ ---------+ 700
```

```
     W  Q  D  V   S  S  T   S  L  V   P  P  S  D   L  T  V
     TGGCAGGATG TCAGCAGTAC CAGCCTCGTC CCTCCATCAG ACCTGACAGT
 701 ---------+ ---------+ ---------+ ---------+ ---------+ 750

      S  S  A   T  T  L  T   D  P  L   S  F  D   L  L  N   E
     CTCGTCGGCC ACTACCCTAA CCGACCCTCT CAGCTTCGAC CTTTTGAACG
 751 ---------+ ---------+ ---------+ ---------+ ---------+ 800

       T  P  P   P  P  S   T  T  T  T   T  S  T   T  R  R
     AGACTCCTCC TCCTCCTTCT ACGACGACAA CAACGTCGAC GACGAGGCGA
 801 ---------+ ---------+ ---------+ ---------+ ---------+ 850

     D  S  C  T   K  V  M   L  T  D   L  I  R  A   E  L
     GACTCATGTA CTAAGGTCAT GTTAACTGAC CTCATCCGGG CTGAATTGTA
 851 ---------+ ---------+ ---------+ ---------+ ---------+ 900

     CACTACCTAA CTGATTTGTC TACCATGACA CCTGACTGAC AATGTGCAGA
 901 ---------+ ---------+ ---------+ ---------+ ---------+ 950

     D  Q  L  Y   F  D  R   V  H  A   F  C  P  I   I  H  R
     GACCAACTCT ACTTCGACCG GGTCCACGCC TTCTGCCCCA TCATCCACCG
 951 ---------+ ---------+ ---------+ ---------+ ---------+ 1000

      R  R  Y   F  A  R  V   A  R  D   S  H  T   P  A  Q   A
     GCGACGGTAC TTTGCGCGGG TCGCCCGAGA TAGCCATACC CCAGCACAGG
1001 ---------+ ---------+ ---------+ ---------+ ---------+ 1050

       C  L  Q   F  A  M   R  T  L  A   A  A  M   S  A  H
     CATGTCTGCA GTTCGCCATG CGAACGCTCG CAGCGGCAAT GTCTGCTCAC
1051 ---------+ ---------+ ---------+ ---------+ ---------+ 1100

     C  H  L  S   E  H  L   Y  A  E   T  K  A  L   L  E  T
     TGCCATCTTA GCGAGCATCT CTATGCCGAG ACCAAGGCCC TCTTGGAGAC
1101 ---------+ ---------+ ---------+ ---------+ ---------+ 1150

      H  S  Q   T  P  A  T   P  R  D   K  V  P   L  E  H   I
     GCACAGCCAG ACGCCCGCCA CACCGCGAGA CAAGGTCCCG CTCGAGCACA
1151 ---------+ ---------+ ---------+ ---------+ ---------+ 1200

       Q  A  W   L  L  L   S  H  Y  E   L  L  R   I  G  V
     TCCAGGCCTG GCTGTTGTTA AGCCACTACG AGCTGCTGCG GATCGGCGTG
1201 ---------+ ---------+ ---------+ ---------+ ---------+ 1250

     H  Q  A  M   L  T  A   G  R  A   F  R  L  V   Q  M  A
     CACCAGGCTA TGCTCACGGC TGGCCGGGCC TTTCGTCTCG TGCAGATGGC
1251 ---------+ ---------+ ---------+ ---------+ ---------+ 1300

      R  L  S   E  L  D  A   G  S  D   R  Q  L   S  P  P   S
     ACGACTGTCA GAGCTGGATG CCGGGTCAGA TCGACAGCTC TCGCCGCCGT
1301 ---------+ ---------+ ---------+ ---------+ ---------+ 1350

       S  S  P   P  S  S   L  T  L  S   P  S  G   E  N  A
     CTTCGTCGCC GCCGTCTTCG CTAACCCTAT CTCCTTCGGG GGAGAATGCT
1351 ---------+ ---------+ ---------+ ---------+ ---------+ 1400
```

```
       E  N  F  V   D  A  E   E  G  R   R  T  F  W   L  A  Y
     GAGAACTTCG TCGACGCCGA AGAAGGCCGG CGGACGTTCT GGCTTGCTTA
1401 ---------+ ---------+ ---------+ ---------+ ---------+ 1450

      C  F  D   R  L  L  C   L  Q  N   E  W  P   L  T  L  Q
     TTGCTTTGAT CGTTTGCTTT GCTTGCAGAA TGAGTGGCCG TTAACGTTAC
1451 ---------+ ---------+ ---------+ ---------+ ---------+ 1500

       E  E  M
     AAGAAGAGAT GGTACGTCGC GCTTCTTTTA TTCTATTTAC CTCAGAATTT
1501 ---------+ ---------+ ---------+ ---------+ ---------+ 1550

                                         I  L  T   R  L  P
     ATATTCAGTT ATTTTTTATT CTAACCCTGC TAGATATTAA CCCGCCTCCC
1551 ---------+ ---------+ ---------+ ---------+ ---------+ 1600

      S  L  E   H  N  Y  Q   N  N  L   P  A  R   T  P  F  L
     CTCCCTCGAA CACAACTACC AGAACAATCT CCCCGCACGC ACGCCCTTTC
1601 ---------+ ---------+ ---------+ ---------+ ---------+ 1650

       T  E  A   M  A  Q   T  G  Q  S   T  M  S   P  F  A
     TCACTGAAGC CATGGCCCAG ACCGGGCAGA GCACAATGTC CCCGTTTGCC
1651 ---------+ ---------+ ---------+ ---------+ ---------+ 1700

     E  C  I  I   M  A  T   L  H  G   R  C  M  T   H  R  R
     GAATGCATTA TCATGGCCAC CCTTCACGGC CGATGTATGA CGCACCGCCG
1701 ---------+ ---------+ ---------+ ---------+ ---------+ 1750

      F  Y  A   N  S  N  S   T  A  S   G  S  E   F  E  S  G
     CTTCTACGCA AACAGCAACT CGACTGCGTC CGGCTCCGAG TTCGAGTCTG
1751 ---------+ ---------+ ---------+ ---------+ ---------+ 1800

       A  A  T   R  D  F   C  I  R  Q   N  W  L   S  N  A
     GCGCCGCGAC GCGAGACTTC TGTATCCGCC AGAATTGGCT GTCGAATGCA
1801 ---------+ ---------+ ---------+ ---------+ ---------+ 1850

     V  D  R  R   V  Q  M   L  Q  Q   V  S  S  P   A  V  D
     GTGGACCGGC GAGTCCAGAT GCTACAGCAG GTCTCCTCGC CCGCTGTTGA
1851 ---------+ ---------+ ---------+ ---------+ ---------+ 1900

      S  D  P   M  L  L  F   T  Q  T   L  G  Y   R  A  T  M
     CAGCGACCCG ATGCTGCTCT TCACGCAGAC GCTCGGCTAC CGCGCGACCA
1901 ---------+ ---------+ ---------+ ---------+ ---------+ 1950

       H  L  S   D  T  V   Q  Q  V  S   W  R  A   L  A  S
     TGCACCTGAG CGATACCGTC CAGCAAGTCT CCTGGCGGGC TCTCGCCAGC
1951 ---------+ ---------+ ---------+ ---------+ ---------+ 2000

     S  P  V  D   Q  Q  L   L  S  P   G  A  T  M   S  L  S
     TCGCCCGTTG ACCAGCAGCT ACTGAGCCCG GGCGCGACGA TGTCGCTGTC
2001 ---------+ ---------+ ---------+ ---------+ ---------+ 2050

      A  A  A   Y  H  Q  M   A  S  H   A  A  G   E  I  V  R
     GGCCGCCGCG TACCACCAGA TGGCCAGCCA CGCAGCCGGC GAGATCGTCC
2051 ---------+ ---------+ ---------+ ---------+ ---------+ 2100
```

```
       L  A  K   A  V  P   S  L  S  P   F  K  A   H  P  F
     GCCTGGCGAA GGCCGTCCCC TCGCTGAGTC CGTTCAAGGC GCACCCGTTC
2101 ---------+ ---------+ ---------+ ---------+ ---------+ 2150

     L  P  D  T   L  A  C   A  A  T   F  L  S  T   G  S  P
     CTACCCGATA CGTTGGCGTG CGCCGCCACG TTCCTCTCGA CGGGCAGTCC
2151 ---------+ ---------+ ---------+ ---------+ ---------+ 2200

      D  P  T   G  G  E  G   V  Q  H   L  L  R   V  L  S  E
     CGATCCCACG GGCGGCGAGG GGGTGCAGCA TCTGCTACGA GTGTTAAGCG
2201 ---------+ ---------+ ---------+ ---------+ ---------+ 2250

       L  R  D   T  H  S   L  A  R  D   Y  L  Q   G  L  S
     AGCTGCGCGA TACACACAGC CTGGCGCGGG ATTATTTGCA GGGGTTGTCG
2251 ---------+ ---------+ ---------+ ---------+ ---------+ 2300

     V  Q  T  Q   D  E  D   H  R  Q   D  T  R  W   Y  C  T
     GTGCAGACGC AGGACGAAGA TCATAGACAG GATACGAGGT GGTATTGTAC
2301 ---------+ ---------+ ---------+ ---------+ ---------+ 2350


     ATAG
2351 ---- 2354
     TATC
```

Figure S2. Corrected annotation of AN11200. Intron is shown in red.

```
ATGTTCTCAA GTACCCGGCG GGTAAGTAAC TCTTTCCATC ATCTGGCCCA TCTTCTTTTC TTTTTTGTTT TCAATTGTAA
GCTCTCGACT AACGACGCCC GGCACCTAGG CAGAAGGCCC CTGTGCAACC GAACTGACGC AGGTATCATC GCTGCTACCT
CCGCGCGGGC CATACGAGTT CAGCCTCCTG CCAACACTCA CTCGACCGTT AGAGGACCTC TCGAAATGCA TCGAAGGTGC
GAGACAGACC TCTGCGACTG CAAATGGTTA CAGCCCCACA GGGCTCGTCC CGCTAGCGGA TTCGATTCTG GAAATCTGTC
AGGCTGCTTG TACAGCTTAT GGTCTTGTTG ACGGTGCTAT TGCTGCAGGT GTGGGTACAG GAAGCAGTGA TAATAGCCCT
ACTGCCACAG GAATAGGAGC AGCAGGACTT ACAGGAGACC GCCCCTCCTC TTCCGGCGCA TCGACCTGGC GCTGTGTAAA
AACCCCCATG ACGCTGGGAT CGCTTACGCT ACAGAATGAA GAAGAGTCGC TGCTCGCAAG GCAGATCGTG TACGCCGTGT
TGACAAGCTT GAGCGCATTA CTGCGAGAAG TTTATGTTCG AGAGAAGGAC GTTGTTTCAG AGACTGATGT GGTGGGGGAA
GGAGGGGTAG GAGCTGGAGC GGCACTGTAT GGGCGTGAAG GGGCTGGAGC CGTTAGTCAG TGTCTCTCGA GGGTTTTAGC
GCTCTTGGGA AAGATAGTAC CTGAGTGA
```

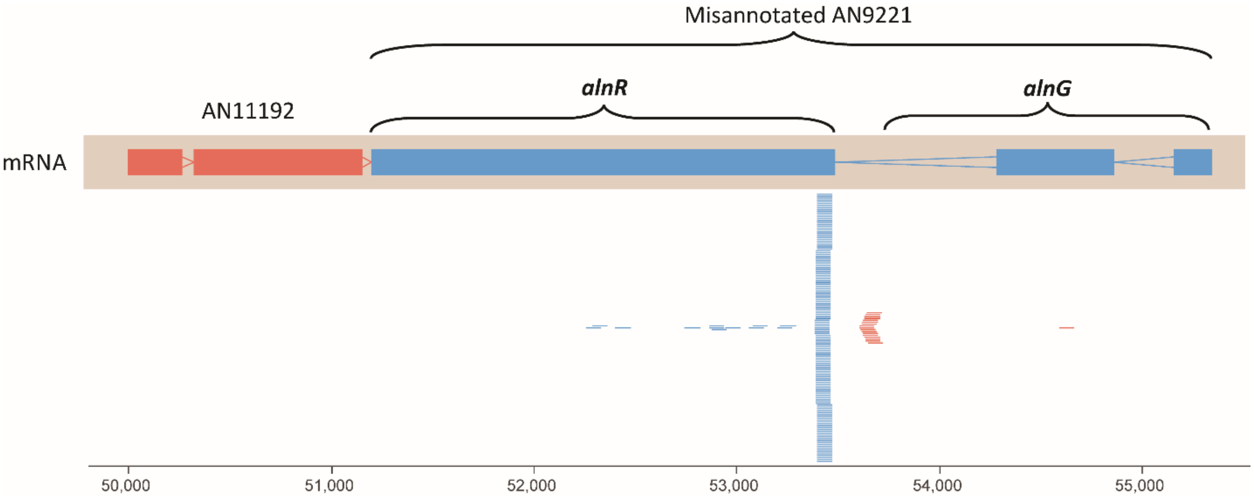


Figure S3. CAGE RNA-seq data showing the transcription start site for *alnR*. The AspGD gene models for AN11192 and AN9221 are shown. Blue genes are transcribed right to left and orange genes are transcribed left to right. In CAGE RNA-seq the 5' cap structures of mRNAs are captured and used for library construction. Sequencing of the library reveals transcription start sites. The CAGE RNA-seq library used in the current study was made from a mixture of mRNAs from two sources, wild-type hyphae cultured for one day at 37ºC in liquid glucose minimal medium and hyphae carrying a deletion of the *mcrA* gene grown for four days at 37ºC in liquid glucose minimal medium. Deletion of *mcrA* upregulates many genes including secondary metabolism genes.[2] Deletion of *mcrA* did not upregulate the (+)-asperlin cluster but there were enough reads in the region of the gene annotated as AN9221 to allow us to determine the transcription start site. The CAGE RNA-seq reads are shown as blue or orange lines below the gene models. The great majority of the reads in AN9221 map to the beginning of what was annotated as the third exon. This result indicates that the third exon is a separate transcription unit (i.e. gene). These and additional RNA-seq data reveal that AN9221 is actually two genes that we are now designating *alnR* and *alnG.*

Reference:

2. Oakley, C. E., M. Ahuja, W. W. Sun, R. Entwistle, T. Akashi, J. Yaegashi, C. J. Guo, G. C. Cerqueira, J. R. Wortman, C. C. Wang, Y. M. Chiang, and B. R. Oakley. 2017. Discovery of McrA, a master regulator of *Aspergillus* secondary metabolism. Mol Microbiol **103**:347-365.

Figure S4. Coding sequence of *alnR* along with predicted amino acid sequence. The position of the Zn(II)2Cys6 zinc binuclear cluster DNA-binding domain is underlined with the cysteine residues in bold.

```
    M  S  T  V   N  Q  S   S  T  R   S  E  L  A   G  N  W
  1 ATGAGCACGG TGAACCAATC TTCCACGCGT TCAGAGCTAG CCGGTAACTG 50

     E  R  L   R  K  S  C   D  T  C   Q  E  A   K  V  K  C
 51 GGAACGCCTG CGCAAGTCCT GCGATACCTG TCAGGAGGCC AAGGTCAAAT 100

      S  Q  H   K  P  S   C  H  R  C   L  R  H   R  Q  P
101 GCAGTCAACA CAAGCCGTCC TGCCACCGAT GCCTTCGACA TCGTCAGCCC 150

    C  V  Y  S   P  Q  R   R  S  G   R  P  P  K   R  P  S
151 TGCGTCTACA GCCCGCAACG TCGGTCGGGA CGTCCTCCCA AGAGGCCCAG 200

     P  S  S   R  L  G  P   E  S  N   N  S  G   D  D  I  H
201 TCCCTCCAGT CGCTTAGGAC CTGAATCAAA CAATTCCGGA GATGACATTC 250

      N  E  N   T  I  Q   R  T  N  L   N  A  N   D  S  A
251 ACAATGAAAA CACCATACAG CGAACGAATC TAAATGCCAA TGACTCTGCC 300

    M  T  D  A   G  A  V   D  P  R   V  L  T  G   D  F  A
301 ATGACTGACG CCGGGGCAGT CGATCCCCGG GTGCTAACCG GCGACTTCGC 350

     A  S  T   G  I  D  P   V  D  D   I  F  Q   T  S  F  E
351 CGCAAGTACT GGCATAGATC CTGTCGACGA TATCTTCCAA ACATCCTTTG 400

      S  F  L   A  A  S   L  S  P  K   G  G  L   L  P  G
401 AATCCTTCCT CGCAGCCTCA TTGTCTCCTA AAGGTGGACT CCTGCCAGGA 450

    S  H  S  N   P  T  T   P  N  G   F  S  M  N   S  P  S
451 TCTCATAGCA ATCCAACCAC ACCCAACGGC TTCTCGATGA ATTCGCCCTC 500

     I  T  D   P  F  G  A   F  P  F   L  I  T   D  H  N  L
501 CATCACTGAT CCATTCGGCG CCTTTCCGTT TCTCATAACG GACCACAACT 550

      P  I  A   A  L  S   S  H  V  P   P  I  D   Q  L  P
551 TGCCTATCGC CGCGCTCTCA TCGCATGTTC CTCCAATTGA TCAGCTACCC 600

    V  L  S  T   G  A  S   N  T  S   S  E  C  G   D  C  G
601 GTACTAAGCA CCGGAGCCTC AAATACAAGC AGCGAGTGCG GCGACTGCGG 650

     A  K  C   Y  S  S  L   L  Q  H   L  L  F   L  R  Q  T
651 TGCGAAGTGC TACAGCTCAC TATTACAGCA CCTTTTGTTC CTCCGCCAGA 700

      L  P  E   S  T  R   P  S  I  D   V  I  M   Q  A  E
701 CGCTCCCCGA GTCCACCAGG CCATCAATAG ACGTGATAAT GCAGGCTGAG 750

    G  H  V  R   A  L  L   D  R  V   L  G  C  N   A  C  L
751 GGCCATGTGC GTGCTTTACT TGATCGGGTA TTAGGCTGCA ACGCATGCCT 800

     G  N  R   S  S  I  L   L  I  S   A  I  T   E  R  I  V
801 TGGCAATCGG TCGTCTATCC TGCTCATATC AGCGATAACA GAGCGCATAG 850
```

```
       Q  M  L   D  W  I   I  E  E  K   T  L  L   D  T  E
 851 TCCAGATGTT AGACTGGATC ATCGAGGAGA AGACTCTTTT GGATACCGAG 900

     N  M  R  Y   N  R  R   T  F  S   S  W  G  R   P  P  R
 901 AATATGCGTT ACAACCGACG AACGTTTAGT TCATGGGGTC GCCCTCCCCG 950

      L  P  P   H  G  L  N   G  M  R   R  N  V   C  H  V  S
 951 GTTACCACCT CATGGCCTTA ATGGTATGCG GAGGAACGTC TGCCACGTTT 1000

       L  R  V   G  N  T   E  L  D  E   D  A  K   Q  Y  F
1001 CACTTCGCGT GGGTAATACT GAATTGGATG AGGACGCCAA ACAGTATTTT 1050
     L  K  N  F   I  L  L   R  L  K   K  L  A  V   K  V  Q

1051 CTTAAGAATT TCATTTTGCT TCGACTAAAG AAACTCGCAG TTAAGGTGCA 1100

      E  V  R   R  T  A  T   T  R  P   G  D  C   I  Y  R  A
1101 GGAAGTGCGA CGGACAGCTA CCACCCGTCC TGGCGATTGC ATATACCGCG 1150

       A  E  L   V  L  A   D  S  I  Q   R  L  D   Y  L  R
1151 CTGCGGAATT GGTGCTGGCG GATTCGATTC AACGACTGGA TTATCTTCGT 1200

     G  Q  C  Q   L  W  E   *
1201 GGCCAGTGTC AGTTATGGGA GTGA 1224
```

Figure S5. Coding sequence of *alnG* and amino acid sequence of its predicted product. The intron is shown in red.

```
     M  T  R  Q   I  P  L   L  A  L   S  W  L  E   L  I  F
  1 ATGACGCGGC AAATCCCGCT CCTAGCGCTA TCGTGGCTTG AATTGATTTT 50

      F  S  C   Y  Y  G  G   L  A  G   L  G  Y   H  S  L  W
 51 CTTCAGCTGC TACTACGGCG GACTAGCGGG ACTGGGATAC CATTCCCTCT 100

       R  I  A   L  R  R   R  N  V  A   P  A  I   K  S  V
101 GGAGGATTGC ACTTCGCCGA AGGAATGTGG CACCCGCTAT CAAGTCTGTT 150

     L  Q  T  G   R  F  A   D  G  T   P  L  T  R   R  Y  T
151 CTGCAGACTG GGCGCTTTGC GGATGGAACG CCCCTAACGC GCCGGTATAC 200

      N  L  E   F  L  D  K   K  L  V   P  A  V   I  F  Y  D
201 TAACTTGGAA TTTTTGGATA AGAAATTGGT TCCTGCGGTA ATCTTCTACG 250

       G  L  L   T  G  A   C  P  L  Y   R  L  L   L  V  D
251 ACGGATTGTT GACTGGAGCA TGCCCACTTT ATCGCTTGTT ACTGGTGGAC 300

     I  H  S  T   M  Q  A   M  A  L   C  M  L  V   S  T  R
301 ATCCATTCGA CCATGCAAGC GATGGCACTC TGCATGCTTG TCAGCACCAG 350

      S  K  S   L  S  T  I   S  L  L
351 ATCCAAGTCG TTATCGACTA TATCTTTGCT GTGAGTCGGG TCCTTCTGCC 400

                                                  L  P  T
401 TTTGAGTATA ACGAAGCTCT AATAATCTAC CGAGGGACAG CTTGCCAACT 450

     F  W  N  V   F  N  Q   F  Y  G   A  A  F  V   Y  P  L
451 TTTTGGAATG TCTTCAACCA GTTTTACGGT GCTGCCTTCG TCTACCCCCT 500

      Y  L  L   L  E  A  V   T  T  G   F  N  P   L  Y  P  V
501 CTACCTCTTA TTAGAGGCAG TAACGACTGG CTTTAACCCT CTGTATCCGG 550

       E  T  E   T  S  R   S  A  L  L   V  S  A   M  I  G
551 TCGAGACCGA GACATCTCGT TCTGCGTTAC TGGTGAGCGC TATGATCGGC 600

     S  F  L  P   F  T  F   L  W  P   A  F  L  R   S  G  T
601 TCTTTTTTAC CGTTCACCTT TCTCTGGCCA GCTTTTCTTC GGTCTGGCAC 650

      E  S  R   Q  R  A  I   A  L  Y   R  F  A   P  V  V  F
651 GGAGAGCCGA CAACGTGCTA TTGCATTATA CCGATTTGCT CCGGTAGTGT 700

       S  L  L   Q  I  V   G  E  K  V   L  G  A   Q  M  I
701 TCTCACTTCT GCAGATTGTT GGAGAGAAGG TGCTGGGCGC GCAGATGATC 750

     P  Q  P  T   S  Q  A   S  P  Y   L  V  A  G   C  A  A
751 CCTCAGCCAA CTTCTCAGGC TAGCCCTTAT TTGGTTGCCG GCTGCGCTGC 800

      T  V  G   H  W  Y  A   L  G  G   A  L  G   L  A  M  R
801 CACAGTGGGG CATTGGTACG CTCTTGGGGG AGCTTTAGGT CTCGCCATGC 850

       L  S  H   R  K  G   R  L  G  A   L  T  L   V  L  K
851 GGCTGTCTCA CAGAAAGGGC CGCTTGGGGG CTCTCACCTT AGTCCTCAAA 900
```

```
     R  L  Y  L   P  R  S   A  E  E   T  T  R  L   D  A  S
 901 CGGCTTTATC TGCCTCGCTC GGCTGAAGAA ACTACTCGCT TGGACGCCTC 950

      V  L  A    R  A  A  H    E  F  L    Q  Y  D    V  L  V  L
 951 TGTACTCGCT CGCGCAGCGC ACGAATTTCT GCAATACGAT GTCCTCGTGC 1000

       I  A  A    Y  I  P    Y  A  Y  Y    L  L  A    P  L  N
1001 TCATTGCAGC TTATATTCCG TACGCATACT ATCTGCTCGC GCCCCTCAAT 1050

     L  A  S  P   F  A  M   V  V  S   L  V  L  G   T  I  F
1051 CTGGCATCGC CCTTTGCGAT GGTTGTGTCC CTTGTACTTG GCACCATTTT 1100

      L  G  P    G  A  V  L    A  F  A    Y  R  V    R  W  H  L
1101 TTTAGGGCCG GGGGCGGTTC TGGCTTTCGC GTACCGGGTT CGCTGGCATC 1150

       A  I  S    D  *
1151 TAGCTATCTC AGATTAG 1167
```

Figure S6. HRESIMS spectra of (2*Z*, 4*Z*, 6*E*)-octa-2,4,6-trienoic acid (**1**) (negative mode) and (+)-asperlin (**2**) (positive mode).

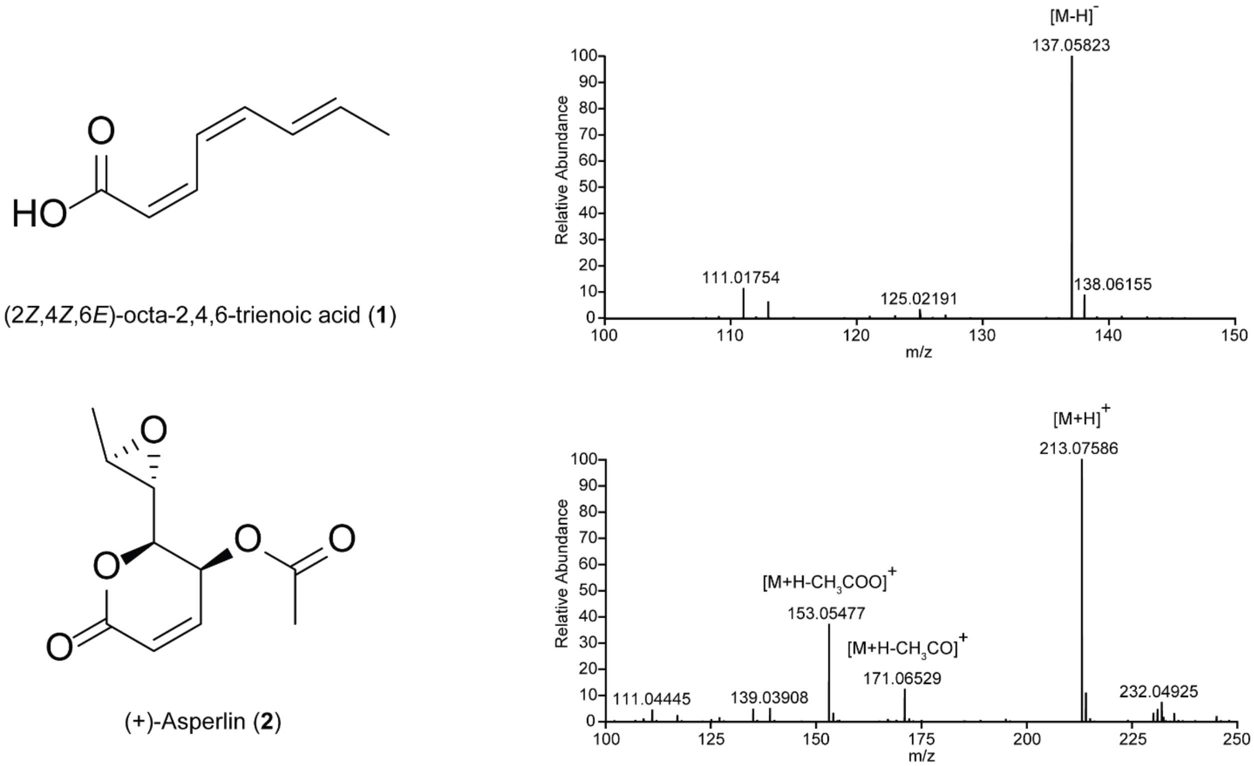
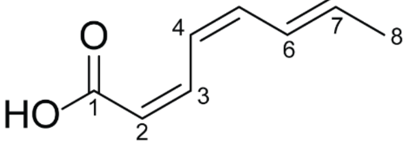
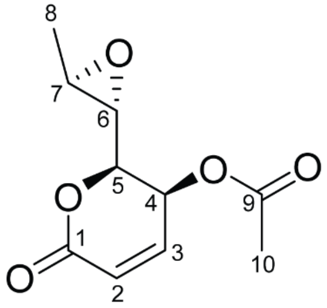


Figure S7. (2*Z*,4*Z*,6*E*)-octa-2,4,6-trienoic acid (**1**) and (+)-asperlin (**2**) ^{1}H and ^{13}C assignments.



1



2

	(2*Z*,4*Z*,6*E*)-octa-2,4,6-trienoic acid (**1**)		(+)-Asperlin (4S,5R,6S,7R) (**2**)	
Position	δ_C, type	δ_H, (*J* in Hz)	δ_C, type	δ_H, (*J* in Hz)
1	172.3, C	–	161.5, C	–
2	116.0, CH	5.67 (1H, d, 11.3)	124.9, CH	6.18 (1H, d, 9.6)
3	141.0, CH	7.18 (1H, dd, 11.3, 10.5)	140.4, CH	7.04 (1H, dd, 9.6, 5.6)
4	126.0, CH	7.15 (1H, dd, 11.1, 10.5)	62.1, CH	5.29 (1H, dd, 5.7, 2.8)
5	138.7, CH	6.36 (1H, dd, 11.1, 10.5)	78.9, CH	4.09 (1H, dd, 7.2, 2.8)
6	122.0, CH	6.62 (1H, dd, 14.6, 10.5)	54.9, CH	3.03 (1H, dd, 7.0, 2.1)
7	136.1, CH	5.97 (1H, dq, 14.6, 6.8)	54.5, CH	3.06 (1H, qd, 5.2, 2.1)
8	18.7, CH_3	1.85 (3H, br d, 6.8)	17.0, CH_3	1.36 (3H, d, 5.2)
9			169.3 C	–
10			20.5, CH_3	2.11 (3H, s)
COO<u>H</u>		10.87 (1H, br s)		

Figure S8. ^{1}H NMR of (2*Z*,4*Z*,6*E*)-octa-2,4,6-trienoic acid (**1**) in $CDCl_3$ (400 MHz).

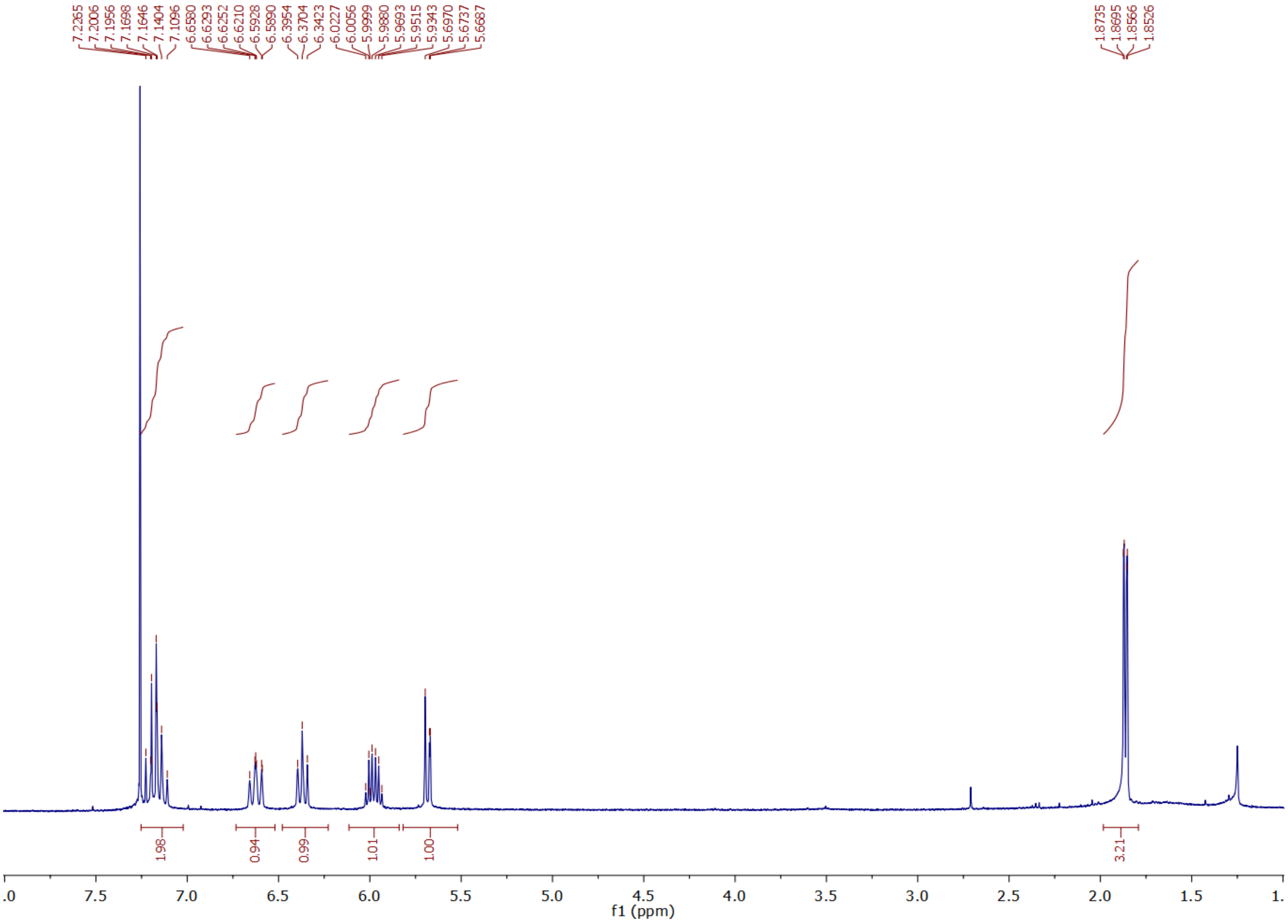


Figure S9. ^{13}C NMR of (2*Z*,4*Z*,6*E*)-octa-2,4,6-trienoic acid (**1**) in CDCl$_3$ (100 MHz).

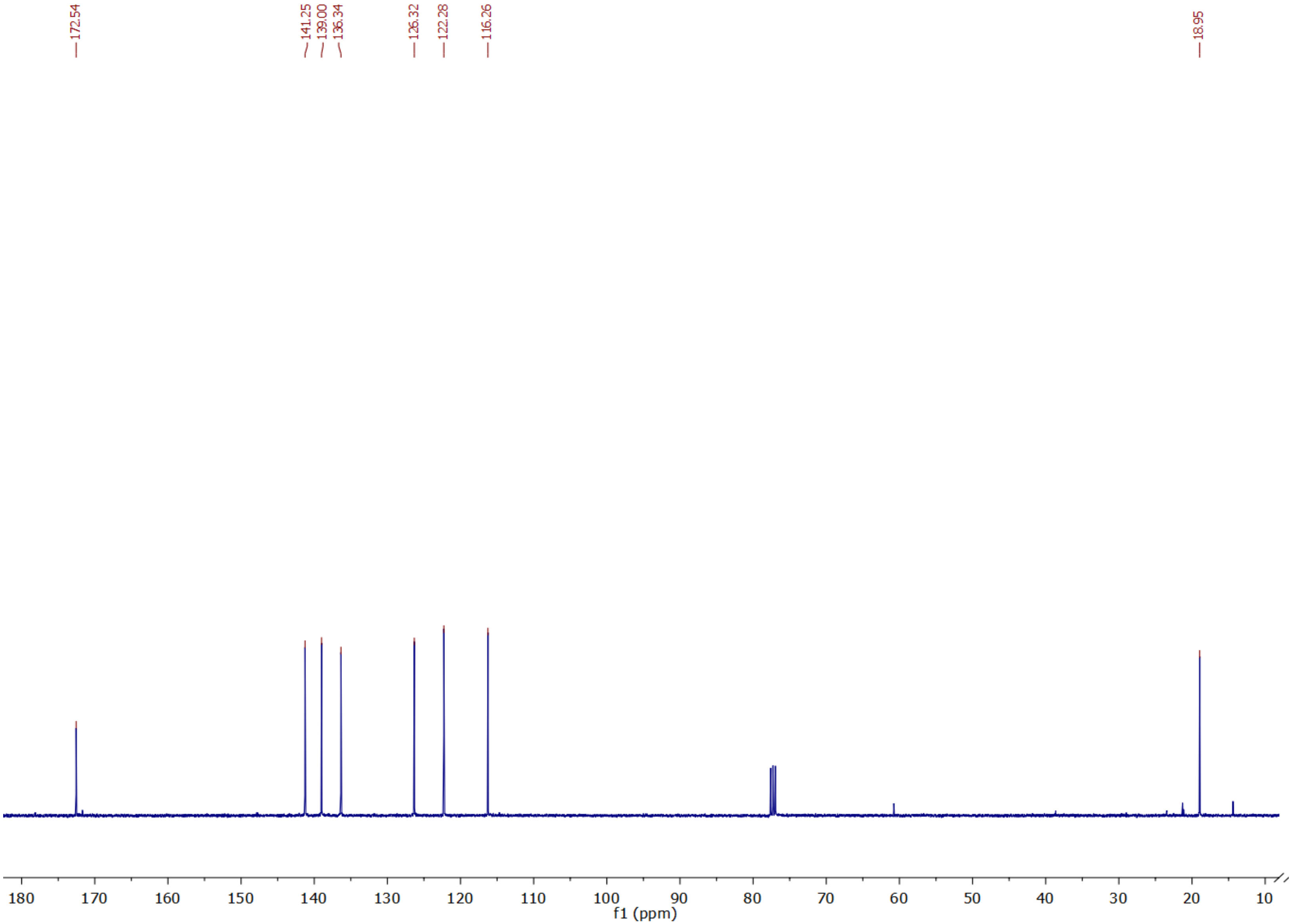


Figure S10. ^{1}H NMR of (+)-asperlin (**2**) in $CDCl_3$ (400 MHz).

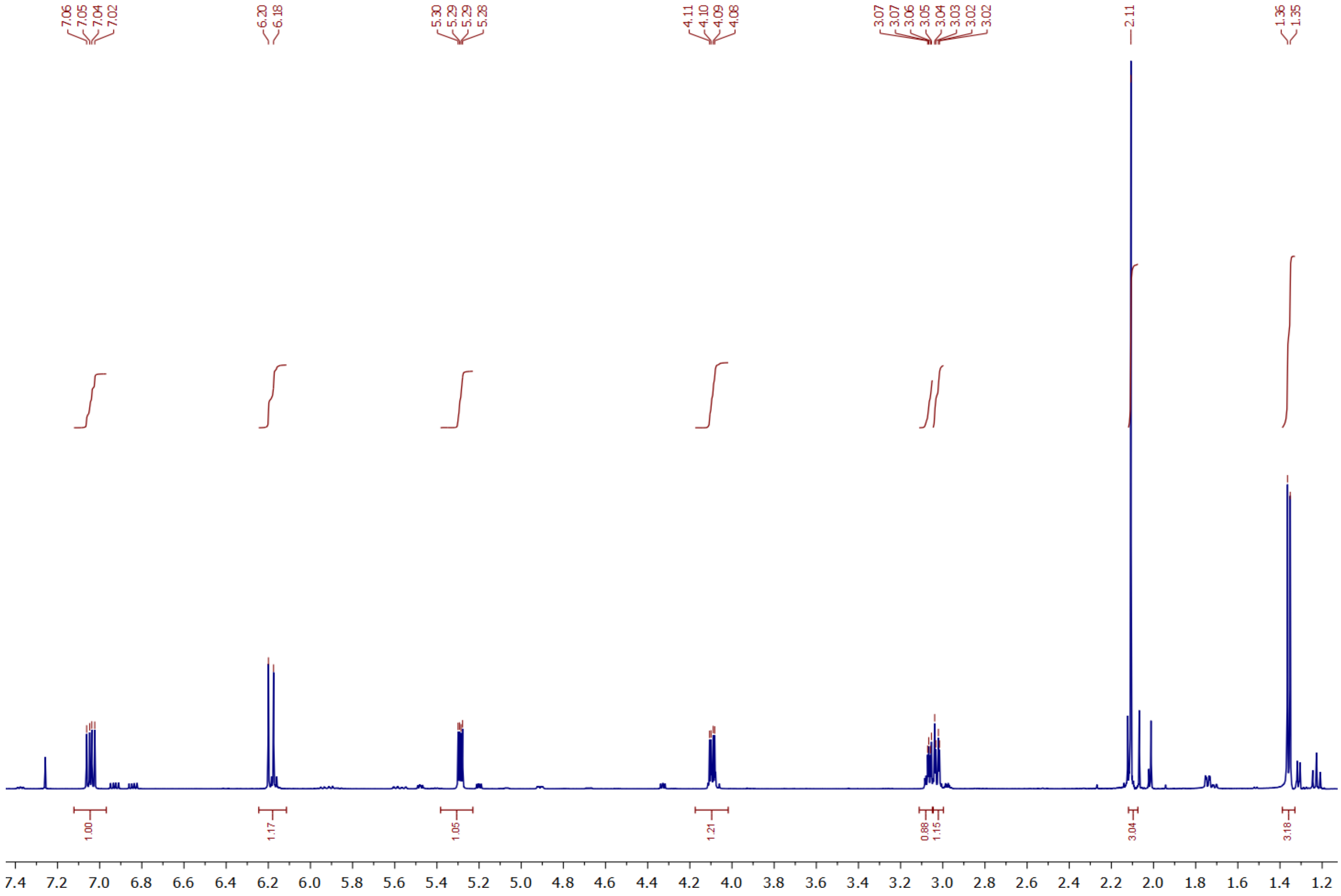


Figure S11. ^{13}C NMR of (+)-asperlin (**2**) in $CDCl_3$ (100 MHz).