Nouvelles approches pour identifier les biomarqueurs présents dans des mélanges complexes de produits naturels

Manon Meunier

THESE DE DOCTORAT

DE
L'UNIVERSITÉ D'ANGERS

SOUS LE SCEAU DE
LA COMUE ANGERS – LE MANS

ECOLE DOCTORALE N° 642
Ecole doctorale Végétal, Animal, Aliment, Mer, Environnement
Spécialité : Pharmacologie, phytochimie et toxicologie

Par

Manon MEUNIER

New approaches to identify (bio)markers in complex mixtures of natural products

Biochemometrics based on laser desorption ionization-(ion mobility)-MS, UPLC-MS and ^{13}C NMR

Thèse présentée et soutenue à Angers, le 5 octobre 2023
Unité de recherche : EA 921 Substances d'Origine Naturelles et Analogues Structuraux (SONAS)

Rapporteurs avant soutenance :

Franz Bucar	Ao. Univ.-Prof., University of Graz, Austria
Emilie Destandau	Professor, University of Orléans, France

Composition du Jury :

Examinateurs :	Mehdi Beniddir	Professor, Paris-Saclay University, France
	Joëlle Quetin-Leclercq	Professor, Louvain drug research institute, Belgium
Dir. de thèse:	Andreas Schinkovitz	Senior lecturer HDR, University of Angers, France
Co-dir. de thèse:	Séverine Derbré	Senior lecturer HDR, University of Angers, France

« La vie n'est facile pour aucun de nous. Mais il faut avoir de la persévérance, et surtout de la confiance en soi. Il faut croire que l'on est doué pour quelque chose, et que, cette chose, il faut l'atteindre coûte que coûte. »

« Chacun de nous doit travailler pour son propre perfectionnement et en même temps partager une responsabilité générale pour toute l'humanité. »

Marie-Curie

Remerciements/Aknowledgment

Je ne peux entamer ces remerciements sans parler de mes encadrants de stage de Master. C'est grâce à eux et à leur passion pour la science que je suis en train d'écrire ces lignes. A l'origine de tout cela, le Dr. Esmeralda Cicchetti. Tu as été la première personne à m'accorder ta confiance et à me partager tes connaissances autour des dérivés terpéniques. J'ai toujours admiré ta persévérance et ton sens du travail et de l'organisation. Mon stage de M2 avec le Dr. Cyrille Santerre m'a également beaucoup appris sur l'analyse par fluides supercritiques. Malgré qu'il ait été impacté par une pandémie mondiale, tu m'as également fait confiance et transmis ta passion.

Grâce à vous, aujourd'hui, j'ai à mon tour la chance de présenter mes travaux de thèse. Ainsi, je tiens à adresser mes sincères remerciements :

A mes encadrants de thèse, le Dr. Andreas Schinkovitz et le Dr. Séverine Derbré,

Merci de m'avoir accordé votre confiance pour mener à bien ce projet de thèse. Grâce à votre expertise très complémentaire en analyse et vos conseils, j'ai appris énormément. Vous avez également été d'un grand soutien et d'une grande compréhension lorsque j'en avais le plus besoin.

Merci Andreas, de m'avoir permis de découvrir un autre laboratoire à Munich et une nouvelle méthode d'analyse. Cette expérience a été très enrichissante. Malgré les quelques difficultés que nous avons rencontrées au départ, nous avons réussi à obtenir de très beaux résultats.

Merci Séverine, pour cette expérience magnifique à l'autre bout du monde en Malaisie où nous avons rencontré de belles personnes toutes aussi passionné que nous pour les sciences. C'était une très belle opportunité pour moi de partager mes travaux et d'en apprendre davantage sur ce pays. Je suis sincèrement reconnaissante pour nos moments d'échanges et d'avoir pu partager cette expérience inoubliable avec toi.

Aux membres de mon CSI,

Le Pr. Mehdi Beniddir et le Dr. Catherine Rouiller qui ont été présents physiquement à chacun de mes CSI. Nos échanges, vos précieux conseils et vos suggestions ont été très constructifs et essentiels à mon travail de recherche.

Aux membres de mon jury,

Je commencerai avec le Pr. Emilie Destandau qui a accepté d'évaluer ce travail en qualité de rapporteur. Étant l'une de vos anciennes étudiantes de Master bioactifs et cosmétiques, je suis ravie de pouvoir partager mes travaux et finaliser mon parcours universitaire en votre présence. Je tiens également à remercier le Pr. Franz Bucar, qui a également accepté d'être rapporteur de mes travaux ainsi que le Pr. Joëlle Quetin-Leclercq et le Pr. Mehdi Beniddir pour prendre part à ce jury en tant qu'examinateurs.

A toute l'équipe du SONAS : permanents, ingénieurs, post-docs, doctorants, et stagiaires,

À vous tous, un grand MERCI. Je suis reconnaissante d'avoir pu croiser votre route et pour tout ce que vous m'avez apporté. Je tiens tout particulièrement à remercier le Pr. David Guillet, directeur du laboratoire SONAS, pour l'opportunité de réaliser mes travaux sur un équipement de qualité. Mais aussi à Patricia Blanchard, pour son aide précieuse à la réalisation des tests anti-AGEs. Merci pour ta bonne humeur et ton travail toujours impeccable. Merci à Ingrid Freuze et Sonia Jerjir de m'avoir aidé et d'avoir partagé vos connaissances sur le LDI-MS. J'ai également pu découvrir le monitorat grâce au Dr. Séverine

Boisard et au Dr. Jean-Jacques Helesbeux entourée de Yannick Abatuci et de l'Ing. David Dallerac. Votre bonne humeur et vos conseils ont été très précieux. Je n'oublierai pas non plus les stagiaires que j'ai eu le plaisir d'encadrer : Emma, Brice, Claire et Vivi. Merci pour le travail que vous avez réalisé.

Et enfin, un très grand merci aux Ing. Dimitri Bréard, Ing. Soprane Suor Cherer et au Dr. Thomas Charpentier. Je suis reconnaissante d'avoir eu l'opportunité de travailler avec des personnes aussi compétentes et bienveillantes. Mon expérience au sein du laboratoire n'aurait pas été la même sans vous. Nos pauses café à tatasser (surtout Toto 😉) et tous nos moments de partage ont rendu cette aventure inoubliable. Merci, Dimitri, pour ton aide précieuse et indispensable au sein du laboratoire (surtout sur le QTOF !).

Aux collaborateurs,

Notamment pour l'accueil par le Pr. Khalijah Awang et son équipe au sein de l'Université de Malaisie, ainsi que les moments de partage scientifique et humain qui sont inestimables.

Merci au Pr. Thomas Brück et à l'Ing. Martina Haack, de l'Université de Munich, pour avoir accepté de m'accueillir pendant deux semaines au sein de leur laboratoire et pour leur temps passé avec moi au laboratoire.

A mon club de patinage artistique, l'ASGA,

J'ai passé deux années sportives pleines de bons moments, grâce aux entraîneurs de l'ASGA. C'est grâce à vous que j'ai trouvé un équilibre de vie et remporté un titre de championne de ligue et de championne de France adulte 2023. Merci à tou(te)s mes ami(e)s patineurs/patineuses qui ont contribué à ce que tout aille bien pendant ces trois ans.

A ma famille et mes amies,

En commençant par mes chats (Sparky et Koda), deux petits êtres trop mignons mais qui ont effacé entièrement deux ou trois fois ma thèse !

Daddy, merci pour tout ce que tu fais pour moi depuis toujours… Sans oublier mes frères, mes belles sœurs et les trois petits bouts de chou de la famille. À tous, je vous suis reconnaissante de m'avoir aidée à devenir ce que je suis aujourd'hui.

Merci à mes amies, Vinciane, Héloïse et Anne. Vous avez été d'une aide et d'une écoute précieuse pendant ces trois ans. Bon courage Anne pour la fin de ta thèse également, on peut le faire !

Enfin, je terminerai ces remerciements avec une profonde pensée pour mes grand parents, Colette et Maurice. Je pense très fort à vous chaque jour, de là où vous êtes. Je suis remplie d'une reconnaissance éternelle pour les valeurs et les enseignements que vous m'avez transmis et qui me serviront pour toujours.

Merci à tous

Table of contents

Communications

ORAL PRESENTATIONS

◊ **METABIODIVEX Thematic School – June 26 to July 1, 2022 - Aussois, France**

Laser desorption ionization mass spectrometry assisted by ^{13}C NMR in applied chemometrics for the targeted isolation of active natural products from complex mixtures

Manon Meunier[1], Séverine Derbré[1], Dimitri Bréard[1], Khalijah Awang[2], Pascal Richomme[1], Andreas Schinkovitz[1]

◊ **Oral presentation and MixONat workshop – September 12 to 16, 2022 - Kuala Lumpur, Malaysia**

Laser desorption ionization mass spectrometry and ^{13}C NMR based dereplication strategies to associate chemical fingerprints of plant extracts with their bioactivity

Manon Meunier[1], Séverine Derbré[1], Andreas Schinkovitz[1]

◊ **PhD students' day SFR Quasav –November 17, 2022 - Angers, France**

Chemometrics based on laser desorption ionization mass spectrometry assisted by ^{13}C NMR dereplication: an alternative approach to LC-MS[2]

Manon Meunier[1], Séverine Derbré[1], Dimitri Bréard[1], Khalijah Awang[2], David Guilet[1], Andreas Schinkovitz[1]

◊ **6e Symposium International de l'AFERP– July 17 to 19, 2023 - Orsay, France**

Matrix-free laser desorption ionization ion mobility mass spectrometry: a complementary approach to chemometrics in natural products research

Manon Meunier[1], Dimitri Bréard[1], Khalijah Awang[2], Martina Haack[3], Dania Awad[3], Thomas Brück[3], Séverine Derbré[1], Andreas Schinkovitz[1]

POSTERS

◊ **American Society of Pharmacognosy (ASP) annual meeting – July 24 to 28, 2021 - Michigan, United-States of America**

Matrix-Free Laser Desorption Ionization Mass Spectrometry assisted by ^{13}C NMR based dereplication strategies using MixONat software to annotate complex mixtures

Manon Meunier[1], Séverine Derbré[1], Dimitri Bréard[1], Khalijah Awang[2], Pascal Richomme[1], Andreas Schinkovitz[1]

◊ **70th International Congress and Annual meeting of the Society for Medicinal Plant and Natural Products Research – August 28 to 31, 2022 - Thessaloniki, Greece**

Laser desorption ionization mass spectrometry assisted by ^{13}C NMR and MixONat software in applied chemometrics to identify bioactive natural products from crude extracts

Manon Meunier[1], Séverine Derbré[1], Dimitri Bréard[1], Khalijah Awang[2], David Guilet[1], Schinkovitz Andreas[1]

◊ **Doctoral school scientific day – June 28, 2023 - Angers, France**

Applied chemometrics in an active molecular network as a combined strategy for the rapid identification of bioactive compounds

Manon Meunier[1], Dimitri Bréard[1], Khalijah Awang[2], Andreas Schinkovitz[1], Séverine Derbré[1]

◊ **XXXI International Conference on Polyphenols (ICP) – July 3 to 6, 2023 - Nantes, France**

Applied chemometrics in an active molecular network as a combined strategy for the rapid identification of bioactive compounds

Manon Meunier[1], Dimitri Bréard[1], Khalijah Awang[2], Andreas Schinkovitz[1], Séverine Derbré[1]

Best poster award

---

[1]Univ Angers, SONAS, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy, Angers, 49070 Beaucouzé, France

[2]Department of Chemistry, Faculty of Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.

[3]Werner Siemens-Chair of Synthetic Biotechnology, Department of Chemistry, Technical University of Munich (TUM), 85748 Garching, Germany

List of abbreviations

A

Advanced chemistry development ..... *ACD/Lab*
Advanced glycation end-products ..... *AGEs*
American Chemical Society ..... *ACS*
Artificial intelligence ..... AI
Atmospheric pressure chemical ionization ..... *APCI*

B

Bioassay-guided fractionation ..... *BGF*
Blocks-based molecular network ..... *BBMN*
Bovine serum albumine ..... *BSA*

C

Caractérisation de mélanges ..... *CARAMEL*
Classical molecular networks ..... *CMN*
Collection of open natural products ..... *COCONUT*
Collision cell ..... *CC*
Collision cross section ..... *CCS*
Collision energy ..... *CE*
Collision-induced dissociation ..... *CID*

D

Data dependent analysis ..... *DDA*
Data independent acquisition ..... *DIA*
Database ..... *DB*
Dictionary of natural products ..... *DNP*
Diode array detector ..... *DAD*

E

Electron impact ..... *EI*
Electrospray ionization ..... *ESI*
Evaporative light scattering detectors ..... *ELSD*

F

Feature-based molecular network ..... *FBMN*
Fluorescence intensity ..... *FI*
Focused-principal component analysis ..... *F-PCA*

G

Garcinia database ..... *GarciniaDB*
Gas chromatography ..... *GC*
Global natural products social molecular networking ..... *GNPS*

H

Hierarchical clustering analysis ..... *HCA*
high collision dissociation ..... *HCD*

I

In-silico MS/MS database ..... *ISDB*
International Chemometrics Society ..... *ICS*
Ion identity molecular network ..... *IIMN*
Ion mobility spectrometry ..... *IMS*

L

Laboratoire d'etude et de recherche en informatique d'Angers ..... *LERIA*
Latent variables ..... *LV*
Liquid chromatography ..... *LC*

M

Mass spectrometry ..... *MS*
Matrix assisted laser desorption ionization ..... *MALDI*
Matrix free laser desorption ionization ..... *LDI*
Metabolomics and dereplication by two-dimensional experiments ..... *MADbyTE*
Molecular network ..... *MN*

N

National Institute of Standards and Technology ..... *NIST*
Natural products ..... *NP*
Network annotation propagation ..... *NAP*
Nuclear magnetic resonance ..... *NMR*

O

Orthogonal partial least squares ..... *OPLS*
Orthogonal partial least squares-discriminant analysis ..... *OPLS-DA*

P

Partial least squares regression ..... *PLS*
Partial least squares-discriminant analysis ..... *PLS-DA*
Predicted carbon-13 NMR data of natural products ..... *PNMRNP*
Principal component ..... *PC*
Principal component analysis ..... *PCA*

Q

Quadrupole time-of-flight mass spectrometry ..... *QTOF*

R

Reanalysis of data user ..... *ReDU*
Regression coefficient ..... *RC*
Retention times ..... *RT*

S

Scatter-plot ..... *S-plot*
Stock solutions ..... *SS*
Substances d'origine naturelles et analogues structuraux ..... *SONAS*
Synthetic compounds ..... *SCs*

T

Tandem mass spectrometry ..... *MS/MS or MS^2*
t-Distributed stochastic neighbor embedding ..... *t-SNE*
Thin layer chromatography ..... *TLC*
Time-of-flight ..... *TOF*
Total ion current ..... *TIC*

U

Ultra-performance liquid chromatography ..... *UPLC*
Universal natural products database ..... *UNPD*

V

Variable importance in projection score ..... *VIP score*

W

Working solutions ..... *WS*

X

Xanthones database ..... *Xanthones-DB*

List of figures

List of tables

Context

The work presented in this manuscript was performed at the "Substances d'Origine Naturelle et Analogues Structuraux" (SONAS) laboratory. This research unit focuses on the analysis, characterization and valorization of plant-based natural products (NPs) and extracts of interest for human [*e.g.* anti-inflammatory: tocotrienols derivatives (Alsabil *et al.*, 2016, 2017; Pein *et al.*, 2018; Dinh *et al.*, 2020), ecological anti-scaling agents: derivatives of cinnamic acid (Richomme *et al.*, 2017), inhibitors of advanced glycation end-products (AGEs): flavonoids and phenolic acids (Derbré *et al.*, 2010)] or animal health (LabCom FeedIn Tech) and plant health [alkaloids and xanthones potentiating plant defenses (Simoneau *et al.*, 2014)].

In this context, one of the objectives of the laboratory is to develop strategies for the rapid identification of NPs in complex mixtures. As a first step the work of A. Bruguière *et al.*, facilitated the development of a dereplication algorithm based on ^{13}C nuclear magnetic resonance (NMR), which resulted in the MixONat software (Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Suor-Cherer, *et al.*, 2020). The latter allows the rapid annotation of NPs in complex mixtures.

In addition, the laboratory has purchased an ultra-performance liquid chromatography coupled to quadrupole time-of-flight mass spectrometry (UPLC-QTOF) in 2020. This instrument allows the detection of trace amounts of metabolites from complex mixtures and is equipped with a high-resolution mass spectrometer detector. The data can then be further processed using the simultaneously purchased ProgenesisQI software. Another advantage of this workflow is the ability to generate a molecular network (MN). This dereplication technique also allows the rapid annotation of NPs in complex mixtures.

As complementary strategies, mass spectrometry (MS) and NMR now make it possible to avoid the long and costly purification steps of NPs in order to identify them by means of dereplication processing. However, despite its powerful analytical performance and broad application, liquid chromatography coupled to mass spectrometry (LC-MS) still has certain limitations such as solvent limitations, time-consuming chromatographic method development, sample preparation.

Therefore, studies on matrix free laser desorption ionization coupled to mass spectrometry (LDI-MS) has been developed in the lab as it allows the rapid and sensitive analysis of a complex mixture without time consuming sample preparation (Le Pogam *et al.*, 2015; Skopikova *et al.*, 2020). Many NPs show close structural similarities to compounds used as matrices in matrix assisted laser desorption ionization (MALDI) and can therefore be easily ionized by UV-LDI lasers. In view of this, the method has been widely used in the laboratory, as it is applicable to a wide range of chemically diverse NPs, such as alkaloids (Schinkovitz *et al.*, 2018; Le Pogam *et al.*, 2020; Skopikova *et al.*, 2020), phenolics or cannabinoids (dos Santos *et al.*, 2019).

In this context, the ultimate goal of my thesis is to discuss about the different workflow available to rapidly identify (bio)active compounds in complex mixtures.

INTRODUCTION

Today, the study of NPs has become an area of interest in many different fields of research. These secondary metabolites are small molecules produced by living organisms (plants, animals, or microorganisms, from any terrestrial or marine source). Unlike primary metabolites, they are involved in various biological functions (*e.g.*, survival or reproduction) that are not related to growth, the maintenance of cellular activities and cell life. This manuscript focuses on NPs produced by plants. They can be divided in three major classes based on the underlying biosynthetic pathway: phenolic compounds (*i.e.* polyphenols), terpenes and steroids, and nitrogen-containing compounds (*i.e.* alkaloids) (Twaij and Hasan, 2022). Compared to synthetic compounds (SCs), NPs are characterized by a larger chemical diversity and their high structural complexity, especially with respect to stereochemical aspects (*e.g.*, NPs tend to have a higher number of chiral centers). In addition, NPs exhibit a large number of carbon and oxygen atoms, but fewer nitrogen and halogen atoms compared to SCs (Clardy and Walsh, 2004).

Thus, NPs have played an important role throughout history. Naturally optimized by evolution, NPs have been known to exert a wide range of highly interesting biological effects and have provided valuable leads for various applications. Most importantly they were the cornerstones of modern pharmaceutical care and played a key role in drug discovery, especially for cancer and infectious diseases (Atanasov *et al.*, 2015; Harvey, Edrada-Ebel and Quinn, 2015). Early lead compounds such as vancomycin, quinine or morphine (Wilson and Danishefsky, 2006) are still very important for modern medicines. NPs are also widely used in cosmetic or food applications (antioxidant (Lourenço, Moldão-Martins and Alves, 2019) or antimicrobial activity (Gyawali and Ibrahim, 2014)). In addition, researchers have found renewed interest in NPs for new biocontrol strategies in agriculture (Jamiołkowska, 2020). These products have bioactive properties that protect plants from pathogens (Jamiołkowska, 2020; Efenberger-Szmechtyk, Nowak and Czyzowska, 2021) and can also control plant growth (*i.e.*, gibberellins (Hernández-García, Briones-Moreno and Blázquez, 2021; Hong, Jang and Ryu, 2021; Wu *et al.*, 2021).

Today, bioassay-guided fractionation (BGF) is conventionally used for the discovery of new (bio)active metabolites (Cunha *et al.*, 2017). Typically, a part of the plant (flowers, fruits, roots, stems, or leaves) is extracted and evaluated for its biological activity. As plant extracts are highly complex and contain different types of metabolites, those extracts that show best (bio)activity then undergo several sequential fractionation steps, each followed by biological evaluation. As a result, the complexity of the initial sample is progressively reduced, ideally yielding one or more single active metabolites. Once isolated, the full chemical characterization of a NP is then usually achieved using a combination of analytical techniques. These include high-resolution mass spectrometry (HRMS) and NMR, as well as methods to determine the absolute configuration of the NP. Techniques such as electronic circular dichroism (ECD) or vibrational circular dichroism (VCD), X-ray crystallography and polarimeter experiments ($[\alpha]_D$ value) can be used for this purpose. This workflow has led to the isolation and structure elucidation of many bioactive compounds (*i.e.*, quinine, morphine, vinblastine, etc.) (Rates, 2001; Christensen, 2021).

However, because BGF requires continuous activity monitoring after each separation step, it significantly reduces the amount of working material and bares the risk of repetitive isolation of previously reported compounds

(Nahrstedt and Butterweck, 2010; Inui *et al.*, 2012). It can also lead to a loss of activity due to irreversible binding of components to the stationary phase of the used chromatographic system. Some (reactive) stationary phases (*e.g.*, silica) may further induce degradation. In addition, BGF is usually biased towards dominant metabolites and thus may omit low yield compounds in obtained fractions (Kellogg *et al.*, 2016a).

Avoiding some of these pitfalls, certain (bio)active NPs can be obtained by total chemical synthesis or by semi-synthesis using a naturally occurring precursor. However, the generation of these structural analogues is often quite challenging and can be prohibitively expensive, especially when synthetic pathways are difficult to implement. Alternatively, biosynthetic engineering, in which the biosynthetic pathways of the producing organism are manipulated to produce NP analogues (Atanasov *et al.*, 2021), may provide another notable option.

Providing an interesting alternative, chemometrics have become a versatile tool for facilitating the early detection of (bio)active metabolites in complex mixtures in recent years. The term first appeared in the beginning of 1970s and is defined as "an interdisciplinary field that involves multivariate statistics, mathematical modelling, computer science, and analytical chemistry" (Brown, 2017). Based on various analytical techniques, chemometrics allows the statistical classification of analytical data. For example, principal component analysis (PCA) (Lever, Krzywinski and Altman, 2017) facilitates the grouping of extracts according to their chemical profile. Partial least squares regression (PLS) analysis (Nica *et al.*, 2013), further allows the incorporation of biological or chemical activity data and consequently the detection of sample (bio)activity markers.

Regardless of the applied statistical model, chemometrics require solid analytical data to produce exploitable results. In this context, (UP)LC coupled to HRMS [(UP)LC-HRMS] is widely considered as the gold standard. Combining the assets of chromatographic separation with HRMS (typically by QTOF detectors), the method can detect a wide range of NPs at low concentrations due to its high sensitivity and resolution (Aydoğan, 2020). Despite these undeniable advantages, (UP)LC-HRMS may encounter certain limitations. For example, sample solubilization is generally limited to mobile phases, mostly mixtures of acetonitrile, MeOH, and water, which are not well adapted for lipophilic compounds. In addition, the development of chromatographic and mass spectrometric methods is often laborious and time-consuming.

As many NPs show very close structural similarities to matrix molecules used in matrices in MALDI, they can be easily ionized without matrix support. In this regard, matrix free laser desorption ionization coupled to HRMS (LDI-HRMS) has shown comparable results to LC-MS (Schinkovitz *et al.*, 2018; Le Pogam *et al.*, 2020; Skopikova *et al.*, 2020) within a fraction of time. In addition, and contrary to electrospray ionization (ESI) sources mostly used in LC-MS, LDI promotes the formation of both radical and quasi-molecular ions (Zenobi and Knochenmuss, 1998; Le Pogam *et al.*, 2015). This may extend the range of detectable compounds. A comparative study on ESI/LDI has shown that certain phenolic compounds (e.g., the lichen quinones haemoventosin and parietin) were detected by their radical ion in LDI-MS but not in ESI (Le Pogam *et al.*, 2015). Similar observations have been made for some vitamin E isomers (Zhu *et al.*, 2016). Moreover, LDI-MS experiments can be completed within a few seconds.

With all this in mind, LDI-HRMS may serve as a valuable complement or potential alternative to certain LC-HRMS applications. The method has been successfully applied to various crude plant extracts, lichen and propolis

extracts (Schinkovitz *et al.*, 2018; Le Pogam *et al.*, 2020; Skopikova *et al.*, 2020) as well as to non-extracted plant material (Islam, Alam and Kim, 2022).

All MS methods (including (UP)LC-HRMS and LDI-HRMS) are most important tools for applied chemometrics as they allow the "putative identification" of NPs within complex mixtures. The term refers a five-level confidence system that was defined by the Metabolomics Society (Schymanski *et al.*, 2014; Blaženović *et al.*, 2018a). The classification criteria are as follows:

- Level 0 corresponds to the unambiguous assignment of a 3D structure including stereochemistry (Blaženović *et al.*, 2018a). It can be obtained by X-ray analyses, circular dichroism and/or alpha-D value (from polarimeter experiments) (Creek *et al.*, 2014)
- Level 1 corresponds to a confirmed structure by two orthogonal methods (*e.g.*, LC-MS, NMR)
- Level 2 corresponds to a probable structure that match with literature data or database (DB)
- Level 3 is a tentative candidate (substance class or substructure match)
- Level 4 is an unknown metabolite (sometimes divided in level 4: molecular formula and level 5 corresponds only to the exact mass (Schymanski *et al.*, 2014))

Complete characterization (level 0 or 1) allows for high quality identification, but the process can be quite time-consuming, especially when known compounds are redundantly characterized and their spectroscopic data have been published several times.

Since the 1990s, scientists use a process called "dereplication" to confirm these previously reported compounds (Hubert, Nuzillard and Renault, 2017a). It involves comparing experimental data with reported spectral data (MS, MS/MS or NMR) from the literature or a DB. Some DBs also provide taxonomic or other structural information (molecular weight, molecular formula, SMILES, etc.), which may help to limit the number of potential candidates. To automate and accelerate dereplication, scientists have developed tools (*i.e.*, DBs, software, or algorithms).

Currently no comprehensive experimental DB comprising all reported NPs exists. For MS/MS DBs, the search for experimental spectra is further limited by the lack of standardized collision energy (CE) settings used for fragmentation. This limitation does not apply for volatile compounds that are analyzed by electron impact (EI) ionization in gas chromatography (GC) coupled to MS (GC-MS). Standard settings of 70 eV produce highly repeatable fragmentation patterns. Similarly, in NMR, although spectral information is much more consistent across platforms, it can vary significantly depending on the nature of the solvent and the power of the magnetic field used for measurements. Ideally, the spectra of a given compound should be measured under different conditions, and the sample preparation and acquisition parameters should be well documented. Considering these pitfalls, several simulated spectra programs such as MetFrag (Ruttkies *et al.*, 2016a), CFM-ID (Allen *et al.*, 2014; Wang *et al.*, 2021a) or MS-FINDER (Tsugawa *et al.*, 2016) have been developed to produce specific MS/MS predicted data.

Another approach to dereplication was developed by M. Wang *et al.* (Wang *et al.*, 2016) who introduced the global natural products social molecular networking (GNPS) (*GNPS - Analyze, Connect, and Network with your Mass Spectrometry Data*, 2023). The GNPS allows the generation of MN that cluster tandem MS/MS data according to similarities of fragmentation patterns thus allowing the visualization of related molecules. In addition, the platform

facilitates the free exchange of experimental MS/MS data in order to rapidly annotate metabolites through the MN and propagate the annotation. Despite these notable advances, an HRMS-based MN can only provide putative annotations up to confidence level 2 because it does not combine orthogonal information from different methods.

In addition, a major limitation of all MS techniques is the differentiation of isomers. Missing chromatographic separation, this drawback particularly applies to (MA)LDI-HRMS. Still in (UP)LC-HRMS certain isomers may not be differentiated but ^{13}C NMR may provide a solution to this problem as even stereoisomers can be differentiated by their ^{13}C chemical shifts. This non-destructive method can be used to annotate known and even unknown compounds with a similar level of confidence as (UP)LC-HRMS. However, the time required to acquire high-quality NMR spectra depends on factors such as sample concentration, magnetic field strength, and the complexity of the studied molecule. In addition, manual interpretation, and analysis of NMR spectra to determine the complete structure of a compound is time-consuming process requiring significant expertise in the field. In this regard, dereplication strategies may save time by annotating already described compounds prior to compound isolation. One such strategy is the ^{13}C NMR based MixONat dereplication algorithm developed by A. Bruguière *et al.* (Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Suor-Cherer, *et al.*, 2020; Bruguière *et al.*, 2021). This algorithm enables confident annotation of metabolites in complex mixtures, with a confidence level of 2. When combined with HRMS data, this strategy allows the structure elucidation of compounds in complex mixtures with a confidence level of 1.

In continuation of this work, my PhD aims to develop a hyphenated approach of MS based chemometrics and ^{13}C NMR dereplication for the identification of potentially active NPs from *Garcinia parvifolia* (*Clusiaceae*). The latter is a native plant from Malaysia. Previous studies on four *G. parvifolia* bark extracts (KL5670, KL5259, KL5248, KL5073) conducted at the SONAS laboratory have found notable inhibiting effects on the formation of advanced glycation end-products (AGEs) in an in-house assay (Derbré *et al.*, 2010). *Garcinia parvifolia* is rich in xanthones, which show close structural similarities to dithranol a reported matrix molecule. Consequently, it's an ideal candidate to develop LDI-MS based chemometrics, which will be compared to conventional LC-MS based methods for identification of active metabolites. At a later stage this approach which will be extended to ion mobility MS as well as the creation of LDI-MS based molecular networks.

In **chapter 1** "State of the art" of my PhD, a review manuscript will outline "current and emerging tools and strategies for identification of bioactive natural products in complex mixtures". Next, **chapter 2** "General Methodology and Experimental Conditions" will describe all the techniques that were applied in practice during my PhD. This includes detailed protocols of the applied anti-AGEs assays as well as all (UP)LC-HRMS, LDI-HRMS and NMR experiments.

Chapter 3, "Early prediction of natural products activity in complex mixtures using chemometric analyses", will focus on the identification of activity markers in complex mixtures by a combined LDI-HRMS and ^{13}C NMR approach. Results were previously published in the journal of Talanta.

Chapter 4, "Dereplication strategies through the use of molecular networks", will analyze improvements, interests, and limitations of the molecular network-based annotations of compounds in complex mixtures. This chapter is divided in two parts. The first part will focus on a draft paper "Looking for actives in the haystack: Merging $HRMS^2$-based molecular networking, chemometrics and ^{13}C NMR-based dereplication approaches" while

the second part will focus on "Matrix free laser desorption ionization ion mobility mass spectrometry: a complementary approach to the differentiation of natural isomers and creation of molecular networks".

A concluding discussion of future perspectives will be provided in **chapter 5**, "conclusions and perspectives".

1. State of the art

1.1. Introduction

This chapter summarizes recent developments in the context of the annotation of active compounds in a complex mixture. Therefore, a review entitled "Current and emerging tools and strategies for identification of bioactive natural products in complex mixtures" has been written to better understand the methods and tools currently used by the scientific community.

[FR]

Ce chapitre vise à répertorier les connaissances actuelles qui ont pour but l'annotation rapide des composés actifs dans un mélange complexe. Par conséquent, une revue intitulée "Outils et stratégies actuels et émergents pour l'identification de produits naturels bioactifs dans des mélanges complexes" a été rédigée pour mieux comprendre les nouvelles méthodes et outils utilisés par la communauté scientifique.

1.2. Review on current and emerging tools and strategies for identification of bioactive natural products in complex mixtures

This review is divided into several parts. The first part (**introduction**) defines the scope of the review and highlights the challenges associated with the early isolation and characterization of NPs in complex mixtures. Subsequently, (bio)chemometrics is discussed as a promising statistical approach for the detection of potentially active metabolites in complex mixtures. The challenges of chemical characterization by conventional MS and NMR methods are then discussed through an in-depth analysis of the strengths and limitations of each method as well as a comparison of their effectiveness. In this way, the introduction provides a clear and comprehensive overview about the complexities of task associated to the identification and characterization of active NPs in complex mixtures.

The second part of the review focuses on **(bio)chemometric approaches** used for the rapid detection of potential (bio)active metabolites in complex mixtures. This section highlights and defines the most commonly used statistics, such as PCA, PLS and OPLS, and provides several examples of their effectiveness. Furthermore, the advantages and limitations of using statistical and (bio)chemometric analyses are outlined.

The third section presents MS- or NMR-based **dereplication tools**, using algorithms, software, or MN for rapid annotation of known compounds in complex mixtures. The section also covers methods that combine both strategies, *i.e.*, (bio)chemometrics and dereplication. Furthermore, the review highlights advanced data analysis tools of GNPS that can improve metabolite annotation.

The fourth part finally provides a detailed discussion about **confidence levels of metabolite identification** in the context of the applied strategy. It explains the different factors that can affect the accuracy and reliability of metabolite annotation, such as the type of data (MS, NMR, or other), the use of orthogonal information (such as

MS/MS and retention time (RT) or collision cross section (CCS)), and the use of reference databases (*in-house*, experimental or in silico). Overall, this section aims to provide a critical view on strengths and limitations of the different presented strategies.

[FR]

Cette étude est divisée en plusieurs parties. La première partie (**introduction**) définit le champ d'application de l'étude et met en évidence les défis associés à l'isolement précoce et à la caractérisation des NPs dans des mélanges complexes. Ensuite, la (bio)chimiométrie est présentée comme une approche statistique prometteuse pour la détection de métabolites potentiellement actifs dans des mélanges complexes. Les défis de la caractérisation chimique par les méthodes conventionnelles de MS et de NMR sont ensuite discutés à travers une analyse approfondie des forces et des limites de chaque méthode, ainsi qu'une comparaison de leur efficacité. De cette manière, l'introduction offre une vue d'ensemble claire et exhaustive des complexités liées à l'identification et à la caractérisation des NP actives dans des mélanges complexes.

La deuxième partie de l'étude se concentre sur les **approches (bio)chimiométriques** utilisées pour la détection rapide de métabolites (bio)actifs potentiels dans des mélanges complexes. Cette section met en évidence et définit les statistiques les plus couramment utilisées, telles que l'ACP, la PLS et l'OPLS, et fournit plusieurs exemples de leur efficacité. De plus, les avantages et les limites de l'utilisation des analyses statistiques et (bio)chimiométriques sont soulignés.

La troisième section présente des **outils de déréplication** basés sur la MS ou la RMN, utilisant des algorithmes, des logiciels ou des MN pour l'annotation rapide de composés connus dans des mélanges complexes. La section couvre également les méthodes qui combinent les deux stratégies, c'est-à-dire la (bio)chimiométrie et la déréplication. De plus, la revue met en évidence les outils avancés d'analyse de données du GNPS qui peuvent améliorer l'annotation des métabolites suite à la création du MN.

La quatrième partie fournit enfin une discussion détaillée sur les **niveaux de confiance** de l'identification des métabolites en fonction de la stratégie appliquée. Elle explique les différents facteurs qui peuvent affecter la précision et la fiabilité de l'annotation des métabolites, tels que le type de données (MS, NMR ou autre), l'utilisation d'informations orthogonales (telles que les spectres MS/MS, le temps de rétention (RT) ou les valeurs CCS), et l'utilisation de bases de données de référence (internes, expérimentales ou *in silico*). Dans l'ensemble, cette section vise à fournir un point de vue critique sur les forces et les limites des différentes stratégies présentées.

Current and emerging tools and strategies for identification of bioactive natural products in complex mixtures

Manon Meunier,[a] Andreas Schinkovitz [a]* and Séverine Derbré[a]*

* Corresponding authors
[a] Univ Angers, SONAS, SFR QUASAV, F-49000 Angers, France
E-mail: andreas.schinkovitz@univ-angers.fr and severine.derbre@univ-angers.fr

The identification of (bio)active natural products (NPs) from complex mixtures poses a significant challenge due to the presence of numerous compounds with diverse structures and (bio)activities. This review provides an overview of current and emerging tools and strategies for the identification of (bio)active NPs in complex mixtures. Traditional approaches, bioassay-guided fractionation (BGF) followed by nuclear magnetic resonance (NMR) and mass spectrometry (MS) analysis for compound structure elucidation, continue to play an important role in active NPs identification. However, recent advances have led to the development of novel techniques such as (bio)chemometric analysis, dereplication or combined approaches, which allow efficient prioritization for the elucidation of (bio)active compounds. This review highlights the strengths and limitations of each technique and provides insights into their combined use to achieve the highest level of confidence in the identification of (bio)active natural products from complex matrices.

1. Introduction

Natural Products (NPs) research aims to discover and characterize novel and/or bioactive metabolites from complex mixtures. Due to their chemical diversity and intrinsic propensity to disturb living organisms (*i.e.*, role in the defence, reproduction or communication of the living organisms that biosynthesize them)[1]. These metabolites, their derivatives and analogues are widely represented in drugs such as anticancer, anti-infective or analgesic agents[2]. They are also attracting the attention of researchers interested in understanding the traditional uses of medicinal plants, as the market for herbal complementary medicines continues to grow[3,4]. Nowadays, NPs which are less persistent in soil than fully synthetic compounds, are also widely studied in the field of phytopharmaceuticals as biostimulants and biocontrol agents[5].

Against this background, the isolation and characterization of these (bio)active NPs remains a considerable challenge. To address this issue, the researchers have traditionally used a strategy called "bioassay-guided fractionation" (BGF)[6], in which extracts are subjected to several sequential separation steps, each followed by biological evaluation. In this way, the complexity of the initial extract is progressively reduced until one or more active metabolites are isolated[7] (**Figure 1, 1-2**).

However, access to sufficient material to isolate novel and/or bioactive metabolites remains a challenging task. It often results in the isolation of major NPs present in complex matrices, while missing minor constituents. To bypass this limitation, alternative methods such as the total chemical synthesis or semi-synthesis using available NPs as precursor compounds are performed by laboratories. However, the generation of these structural analogues is sometimes challenging and can be prohibitively expensive, especially when synthetic routes are difficult[1].

Conventionally, nuclear magnetic resonance (NMR) experiments are then used to characterize the new (bio)active NPs isolated from the BGF process or to confirm the structure of the potential (bio)active (semi)synthetic derivatives. These analyses are systematically complemented by high resolution mass spectrometry (HRMS) analyses, which allow the determination of molecular formulas. Additional X-ray analyses allow the elucidation of the 3D configuration (**Figure 1, 3**). However, these approaches require time, large amounts of organic solvents, and sufficient sample material. In fact, ^{13}C NMR, DEPT or 2D experiments can be very time-consuming for low quantity samples.

Therefore, alternative new strategies have been developed over the last few years to improve the early isolation and identification of minority original and/or (bio)active NPs. One of them are the (bio)chemometric approaches[8], which aim to extract a maximum of relevant information from an analytical data set through statistical analysis. Generally applied after an initial fractionation, this method allows a faster targeting of potentially (bio)active compounds at an early stage of the identification process[9]. Depending on the data and objectives, different statistical analyses can be implemented [10].

However, while (bio)chemometrics can tag a list of potential (bio)active compounds in complex mixtures or fractions, it is not an adapted tool for their conclusive identification and structure elucidation (**Figure 1, 7**). To accelerate the process, researchers have been using dereplication strategies for decades[11]. Dereplication refers to rapidly identifying metabolites from complex mixtures that were previously described in the literature. Many available databases (DBs) can assist the dereplication process and may significantly reduce the time required to elucidate a known structure. In practice, dereplication may be based on either based on experimental or *in-silico* HRMS or NMR data. By today MS fragmentation patterns or NMR chemical shifts may be predicted with a considerable degree of certainty.

In addition, molecular networking (MN), a highly efficient tool, was introduced in 2012 by M. Wang *et al.*[12] This emerging computer-based approach allows the visualization and organization of tandem mass spectrometry (MS/MS) data[12,13]. The annotation of metabolites can then be propagated conducted on the entire MN using experimental spectral libraries available on open access platforms such as GNPS. The obtained MN can also be enriched by multiple information (biological, taxonomical, spectrometric, statistical, or others) through open-source software (*i.e.*, Cytoscape)[14]. Other tools are also available based on different strategies or data[13,15] (**Figure 1, 4-5**).

Interestingly, many successful studies have permitted to merge MN with (bio)chemometrics. As a result, potential (bio)active metabolites from complex mixtures (extracts or fractions) can be rapidly tagged and putatively annotated using these combined workflows (**Figure 1, 7**).

The confidence level of their annotation depends on the available analytical information. According to the criteria defined by the Metabolomics Society, the level of confidence for the identification of these compounds depends largely on the amount and the nature of available information[16]. As a standard approach, NMR experiments (1D and 2D) of complex mixtures are performed as the method of first choice for structural assignment, combined with HRMS data for exact mass and molecular formula determination.

However, depending on the objective the highest level of confidence for structure identification may not always be necessary. In terms of rapid identification of already known compounds present in a complex mixture, modern combined approaches (*e.g.*, (bio)activity-based MN) will always be less reliable than data obtained from a single isolated compound. However, if the primary goal is to avoid isolating NPs from a complex mixture that are already known to be (bio)active, this strategy can save time by focusing only on the NPs of interest.

In recent years, the development of tools (software, algorithms or DBs) has led to the publication of several reviews that provide comprehensive coverage of existing tools and strategies[11,16–19]. Therefore, this review aims to provide a fresh perspective by building on the existing literature and focusing on new strategies that have emerged since 2019.

With this in mind, this review will highlight the new thinking in the rapid isolation and identification of (bio)active NPs in complex mixtures through selected cornerstone studies. The first section covers the most commonly used statistical tools and (bio)chemometric analyses. Emphasis is placed on principal component analysis (PCA), partial least squares (PLS), orthogonal partial least squares (OPLS) and the overall workflow combining multiple statistical models. Then, MS and NMR dereplication

tools are defined and their applications in combination with chemometric methods are discussed. Finally, the level of confidence according to the chosen approaches is discussed. We would like to mention that this review is not comprehensive due to the large number of studies and apologize for omitting certain contributions to this field of research.

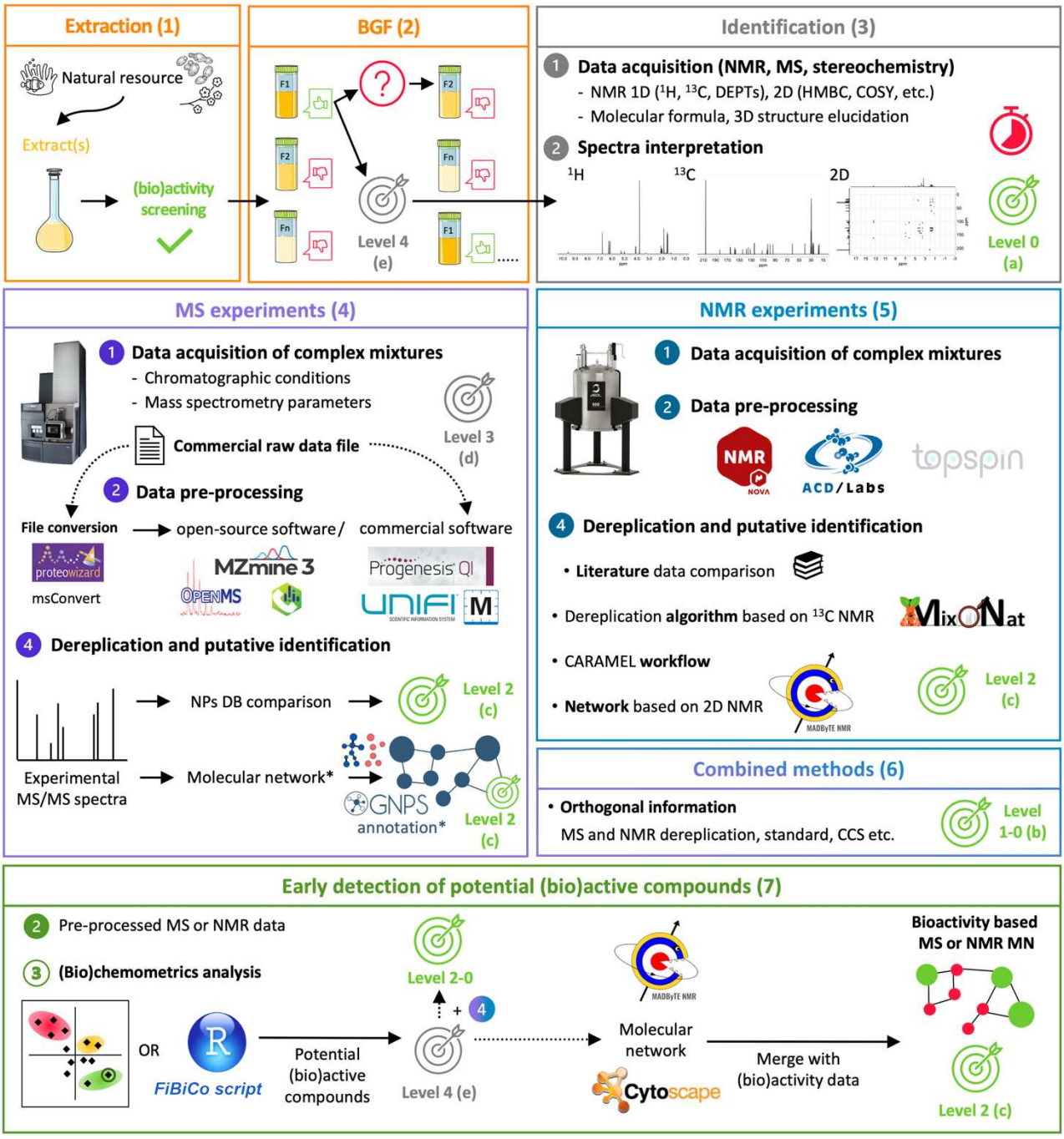


Figure 1. Schematic workflow of metabolite annotation according to the selected pathway and its confidence level. Extraction of the natural resource is typically followed by BGF (2). Traditionally, the identification of isolated metabolites relies on comprehensive interpretation of 2D NMR, MS, and stereochemistry (3). Alternatively, complex mixtures (extracts or fractions) can be directly analysed using MS (4) or NMR (5). After pre-processing, each method allows for dereplication of the data. The combination of orthogonal information allows a better confidence level (6). In addition, early targeting of (bio)active metabolites is possible (7).

2. (Bio)chemometric approaches to detect potential (bio)active metabolites

2.1. Origin and definition of (bio)chemometrics

The first paper using the term 'chemometrics' was published in the beginning of the 1970s by S. Wold[20], who, together with B. Kowalski, founded the International Chemometrics Society (ICS) in 1974. The latter defined chemometrics as "an interdisciplinary field that involves multivariate statistics, mathematical modelling, computer science, and analytical chemistry"[21]. Later, in the 1980s, the first specialized journals (*i.e., Chemometrics and Intelligent Laboratory Systems* and *Journal of Chemometrics*), the first book with "chemometrics" in the title, several American Chemical Society (ACS) symposia, the first book series (Research Studies Press), and the first dedication software (ARTHUR[22], SIMCA[23], and UNSCRAMBLER[24]) appeared. Next chemometrics were extended to biochemistry, biotechnology, and molecular biology, leading to the creation of "biochemometrics", as proposed by Martens *et al*. in 2006[25].

There are two categories of statistical approaches. First, unsupervised methods (PCA), which are mainly descriptive and focus on isolating useful information within a data set. Second, supervised methods (PLS, OPLS) that model the complex relationships between independent (*e.g.*, m/z) and dependent variables (*e.g.*, (bio)activity)). Their main goal is to model the distribution of the data set in order to learn more about the data.

2.2. Experimental design

A fundamental step in (bio)chemometric analysis is experimental design, which is the structured approach used to collect data in order to obtain valid and objective conclusions. This comprises experimental optimization, calibration, data post processing and the extraction of a maximum of chemical information. Indeed, statistical analyses are often complex and require extensive post processing of raw data prior to statistical modelling (see section 3.1.1).

2.3. Applied (bio)chemometric tools in the field of NPs research

This review will focus on the main applied statistical tools in the field of NPs research, *i.e.*, PCA, PLS-(discriminant analysis) (PLS(-DA)), OPLS(-DA), Scatter-plot (S-plot) from OPLS experiment, VIP scores and regression coefficients (RC). Other methods are mentioned but not discussed in detail. In addition, the combined application of several (bio)chemometric tools is also discussed, as it can enhance the reliability of the processed data.

2.3.1. Principal component analysis

2.3.1.1. History and definition

Invented in 1901 by K. Pearson[26], PCA[27] is an unsupervised method for reducing the dimensionality of data while preserving as much information as possible from the original data. It can express similarities as variables.

2.3.1.2. How PCA is working

The objective of PCA is to transform a high-dimensional data set into a lower-dimensional space while preserving important information. For that, original data must be arranged in a two-way matrix where the columns represent variables (*i.e.*, m/z, retention times (RT), chemical shifts, or peaks) and the rows represent the samples in which these variables were measured. This arrangement is the starting point for PCA, a dimensionality reduction technique[28]. The transformation to a lower-dimensional space is accomplished by creating new orthogonal variables called principal components (PC). The first principal component (PC1) is the line that best approximates the data in the least square sense. It represents the maximum variance direction of the data set. However, one PC is insufficient to model the variation of the data set. Therefore, a second PC (PC2) representing another line (orthogonal to the PC1) is required[29]. This continues until a total of PC have been calculated, which is equal to the original number of variables. Typically, a comparison is made between PC1 and PC2, as they provide valuable insights. However, there are cases where comparing PC1 to PC3 can lead to a clearer understanding of the information contained in the data set[30].

2.3.1.3. Pitfalls of PCA

PCA is a powerful method for data reduction or exploring the structure of the data. However, as PCA aims to maximize the explained variance in the data, it cannot be used to explore the relationship between variables. Moreover, PCA is sensitive to outliers, and their presence can bias results. In addition, PCA assumes linearity, which can limit its efficiency in capturing nonlinear relationships. Interpreting the resulting PC can be challenging because they are linear combinations of the original variables. Scaling and standardization of variables are also critical to ensure unbiased results. PCA involves dimensionality reduction, which can result in the loss of valuable information contained in discarded components. Finally, there is a risk of overfitting and limited generalization if the number of retained components is not carefully determined. To avoid these pitfalls, it is important to carefully consider data characteristics, objectives, and alternative methods when applying PCA.

2.3.2. Partial Least Square regression

2.3.2.1. History and definition

PLS is also a multivariate statistical technique that has been developed in the late 1960s by H. Wold. Originally used in the field of econometrics[31,32], PLS found its way into chemistry in the late 1970s by S. Wold, H. Wold's son. Unlike PCA, PLS is a supervised method that also allows

dimensionality reduction with focus on group separation (samples, extracts, etc.).

2.3.2.2. How PLS is working

PLS is based on linear regression that models complex relationships between independent [X, e.g., m/z] and dependent [Y, e.g., (bio)activity] variables. In PLS, the main goal is to find a set of latent variables (LV). These components are linear combinations of the original predictor variables (independent variables) that capture the maximum covariance between X and Y. PLS is particularly useful when dealing with data sets that have a large number of variables, potential multicollinearity among predictors, or when there are more predictors than observations. It can be used to predict a continuous variable based on a set of input variables.

PLS-DA, as an extension of PLS, which is specifically designed for classification problems. It is used when the dependent variable (Y) represents categorical information. The goal of PLS-DA is to build a model that maximizes the separation between different classes in the data. Similar to PLS, PLS-DA also seeks for latent variables or components that capture the maximum covariance between X and Y. However, the focus is on class discrimination rather than regression.

For example, in a recent study, PCA and PLS-DA were used to fingerprint bell pepper samples. As a result, PLS-DA produced perfectly discriminated and compacted groups, while PCA produced more dispersed clusters[33].

2.3.2.3. Pitfalls of PLS

There are several significant pitfalls of PLS. For example, the method can suffer from a strong influence of outliers, which can distort the relationships and affect the calculation of the LV. Outliers can contribute excessive variation or noise[32]. Moreover, it is recommended that the method is used on large data sets, otherwise, it will probably underperform if extreme distributions are present[34].

2.3.3. Orthogonal Partial Least Squares

2.3.3.1. History and Definition

In the literature, OPLS was first introduced in 2002, by J. Trygg and S. Wold in a paper entitled "Orthogonal projections to LV (OPLS)"[35].

2.3.3.2. How OPLS and S-Plot are working

This new approach is also a supervised method close to the PLS model. However, it is suitable for visualizing the differences between two groups (*e.g.*, active/non-active complex mixtures) and highlights the variables that have the greatest discriminative power.

To do this, OPLS tries to extract the maximum amount of information that reflects the variation in the data set, using orthogonal signal correction as described by S. Wold *et al.*[20]. In short, this method maximizes the explained covariance on the first LV, while the remaining LV capture variance in the predictors that is orthogonal, *i.e.*, statistically uncorrelated with the response variables.

In addition to OPLS, another useful tool in chemometrics is the S-Plot. The S-Plot is a graphical representation that helps to identify variables (*e.g.*, spectral peaks, chemical components) and contributes significantly to the separation or correlation between sample groups. It visualizes the correlation between a given variable and the response variable of interest, highlighting influential data points[36]. By examining the S-Plot, researchers can identify important variables driving the separation between two groups and gain insight into the underlying factors responsible for the observed differences.

2.3.3.3. Pitfalls of OPLS and S-plot

While PLS and OPLS models share some common pitfalls, they also present unique challenges. First, the complexity of both models can hinder interpretation, but OPLS models can be more complex due to the inclusion of the orthogonal component. In addition, the choice of orthogonal signal correction methods is unique to OPLS and introduces subjectivity that must be carefully considered. In contrast, PLS models lack this additional layer and are more straightforward in their interpretation.

2.3.4. Definition of VIP scores and regression coefficients

Variable importance in projection score (VIP score) is a measure used in multivariate analysis, particularly in PLS and OPLS. It quantifies the importance of each variable in the model for explaining the variation in the dependent variable. VIP scores are calculated based on the weights and explained variance of the model. Variables with high VIP scores are considered more influential in the model, indicating their relevance in explaining the observed variation. In this way, VIP scores may help to detect potentially active metabolites[37–41].

A regression coefficient is a measure of the relationship between the independent variable and the dependent variable in a regression model. The sign of the RC indicates the direction of the relationship (positive or negative), and the magnitude represents the strength of the association.

2.4. Examples of applications in the quest for bioactive NPs

2.4.1. Basic chemometric analyses of HRMS data

In the field of NPs, MS detector is most widely used. As an example, J.J. Kellogg *et al.*[10] have compared several statistical approaches such as PLS, S-plot, and selectivity ratios for the detection of (bio)active compounds in complex mixtures. Their objective was to demonstrate which of these approaches would be most effective for distinguishing active and inactive compounds.

As a practical case, the authors used crude extract and fractions from the endophytic fungal species *Alternaria* and their antimicrobial activities against *Staphylococcus aureus* as the dependent variable. The subsequent biochemometric analysis of the ultra-performance liquid chromatography (UPLC)-HRMS data set demonstrated the potential of PLS, S-plot and selectivity ratio to understand the fundamental relationship or differences between two samples. While the

S-plot allows the identification of the variables most correlated with bioactivity, the selectivity ratio indicates variables that have a greater contribution to bioactivity. The comparison of the techniques led to the detection of two potentially active metabolites, which were identified as altersetin and macrosphelide A by NMR.

Another recent study applied the PLS model to screen for immunoactive compounds in *Ganoderma lucidum* spores (GLS). Based on HRMS data, this approach facilitated the identification of several immunoactive compounds *i.e.,* 20-hydroxylganoderic acid G, elfvingic acid A and ganohainanic acid C[42].

2.4.2. Combined use of chemometric analyses on an HRMS data set

A case study by L. Ory *et al.*[43] proposed to combine several statistical models (PCA, Spearman, focused-PCA (F-PCA), PLS, and PLS-DA) through a script based on the open-source software R Cran[44]. This NPs drug discovery program, called FiBiCo, highlighted a "super list" of potential bioactive compounds present in an extract of *Penicillium chrysogenum*. One compound was identified by a NMR pharmacophoric deconvolution approach as ergosterol, which showed anti-proliferative activity against breast cancer cells.

Another study by D. P. Demarque *et al.*[45] used a different strategy and combined unsupervised (PCA and hierarchical clustering analysis (HCA)) and supervised methods (PLS) on the Brazilian plant *Annona crassiflora* analysed by liquid chromatography (LC)-MS and LC-MS/MS. VIP scores were then used to spot most important compounds that differentiate the groups and could be responsible for the larvicidal activity against *Aedes aegypti*. As a result, the authors have proven the capacity of chemometrics to predict active metabolites in complex mixtures and were able to identify various annonaceous acetogenins by NMR.

2.4.3. Chemometric analyses of ^{1}H NMR data

As exemplified in the previous two paragraphs, MS is widely used to detect bioactive NPs in complex matrices through chemometrics. However, it has been shown that ^{1}H NMR combined with chemometric tools is also useful for finding active compounds in complex mixtures. In their study, D.S. Alves *et al.*[46] analysed ^{1}H NMR data of different *Duguetia lanceolata* samples using PLS. This facilitated the grouping of sample material according to their metabolomic profiles and insecticidal activity. Finally, the observed activity could be attributed to 2,4,5-trimethoxystyrene.

Data sets from ^{1}H NMR experiments were also used in a PLS-DA to correlate spectra of *Ormosia arborea* leaves with their *Leishmania donovani* nucleoside hydrolase activity. This strategy confirmed the feasibility of NMR-based chemometric methods and yielded the isolation of two active A-type proanthocyanidins[37].

2.5. Advantages and limitations of using statistical and (bio)chemometric analyses

Data interpretation by statistical analyses is of great value for detecting potential (bio)active compounds within complex data sets. However, the effectiveness of multivariate models relies heavily on data processing, which remains a challenging task[47]. As a result, the interpretation of results can be imperfect or incomplete. Furthermore, relying on a single statistical model may compromise the comprehensiveness and accuracy of the compounds detected, as no model is perfectly suited to all situations. Consequently, several studies have used combined chemometric approaches to elucidate potentially (bio)active compounds, recognizing the need for a more comprehensive analysis.

Moreover, in the context of NP research, the use of chemometrics as a complementary strategy to conventional BGF can help reducing the quantities of samples material and solvents by targeting (bio)active compounds earlier on in the process.

However, it is important to note that the use of (bio)chemometrics by itself cannot identify potential (bio)active compounds. The latter necessarily requires further analytical tools and/or dereplication strategies to confirm the chemical identity and the subsequent biological assay to validate the bioactivity of the compound.

3. Dereplication tools used for rapid annotation of known compounds in complex mixtures

Since its introduction in 1990, dereplication is currently widely used by researchers prior to the isolation of potential (bio)active NPs to reduce the time required by conventional methods[11,48,49]. This strategy facilitates the rapid annotation of known compounds in complex mixtures (extracts or fractions). This is facilitated by dedicated algorithms or software that compares acquired data with those from experimental or predicted databases.

Due to the very limited structural information provided by simple UV or DAD detection systems, other more powerful instruments are generally required within a dereplication strategy. Among these, MS and NMR instruments are the most widely used. Another dereplication strategy based on MN is being widely explored by researchers today[12,13,15,50].

With this in mind, this section will focus on both MS and NMR dereplication approaches through the different strategies developed in recent years. Then MN and their different ways of construction and visualization will be defined and discussed.

3.1. Mass spectrometry dereplication workflows

The interpretation of MS data is a critical factor in NPs discovery. The huge number of data sets generated by HRMS measurements require considerable efforts to extract useful information from thousands of MS features. One

strategy to address this issue is automatized and treatment of MS data. Post processing software tools for LC-MS facilitate automatized background subtraction, alignment, peak picking, deconvolution, and organization of data[51].

3.1.1. Data acquisition, software and post processing strategies

Prior to processing the data generated by HRMS detectors, automatic acquisition of MS/MS spectra for all detected metabolites is a prerequisite. The two main strategies used are data dependent acquisition (DDA) and data independent acquisition (DIA).

In DDA mode, the precursor ion is selected in a scan event and fragmented in the subsequent MS/MS by collision induced dissociation (CID) or high collision dissociation (HCD). DDA works by initiating fragmentation based on predefined criteria such as intensity thresholds and the number of selected precursor ions or dynamic exclusion (limiting recurring precursor ions)[52].

In DIA mode, MS/MS spectra do not contain information about the selected precursor ion. All ions detected in an MS scan are fragmented within a predefined range, providing full scan MS/MS spectra. This automated acquisition mode can be operated on several HRMS instruments including MS^E (Waters Corporation), all-ion MS/MS fragmentation (AIF, Agilent Technologies), broadband CID (bbCID, Bruker), multiplexed MS/MS data-independent acquisition (MSX-DIA; ThermoFisher Scientific) or sequential window acquisition of all theoretical mass spectra (SWATH; Sciex)[51]. However, the product ions in these acquisition modes are convoluted and, unlike DDA, the challenge is to produce clean deconvoluted MS/MS spectra that are linked to their precursor ions.

Once acquired, MS and MS/MS data must be processed before entering the metabolite annotation workflow. Some of the most commonly used commercial programs are UNIFI (Waters)[53], ProgenesisQI (Nonlinear Dynamics)[54], MetaboScape (Bruker), MassHunter Profiler (Agilent), Profiling Solution (Shimadzu), and Compound Discoverer (ThermoFisher)[55]. Alternatively, the open-source software MZmine[56] (for DIA mode) and MSdial[57] (for DDA mode) can be used.

These MS data post-processing software packages perform several key steps to improve data quality and prepare it for further analysis, including alignment, noise filtering, peak detection, deisotoping, and deconvolution. Alignment corrects systematic shifts in RT or m/z values across multiple samples, ensuring accurate comparison. Noise filtering eliminates random noise and improves signal-to-noise ratio. Peak detection identifies and localizes peaks of interest based on intensity and shape. Deisotoping removes isotopic peaks, simplifying the data. Deconvolution resolves overlapping peaks and separate co-eluting compounds. These post processing steps improve data quality, enhance peak detection, and provide a refined data set for subsequent analyses.

3.1.2. Experimental databases of natural products for MS-based dereplication processing

Immediately after data post processing, metabolites annotation can be initiated. The process starts by applying a dereplication strategy. The latter plays a crucial role in identifying "known-unknowns" compounds in complex mixtures, providing valuable insights[58].

There are two types of experimental DBs: *in-house* MS/MS DBs and larger shared MS/MS DBs. On the one hand, the *in-house* DBs are created by laboratories and contain MS/MS spectra of isolated natural compounds or mixtures, that were obtained under specific experimental conditions. These DBs can be tailored to the laboratory's research interests or sample types. Researchers have control over data quality and can continually update and expand the DB. However, *in-house* databases can have limitations, such as limited coverage and the need to validate the identification of various known compounds.

On the other hand, more extensive shared experimental MS/MS DBs compile spectra from different sources, including literature reports and community contributions. Examples of commercial experimental DBs include the national institute of standards and technology (NIST) DB[55], METLIN[59], mzCloud[60]. MassBank[61], and the global natural products social molecular networking (GNPS)[12]. The latter are open-access DBs, which benefit from community contributions and promote accessibility and knowledge sharing.

However, the major drawback of these DBs is that MS/MS spectra were acquired from different instruments with varying experimental settings. In addition, the instrument specific conception, such as the ionization source [*e.g.*, electrospray (ESI), atmospheric pressure chemical ionization (APCI) or matrix-assisted laser desorption/ionization (MALDI)] may promote or suppress the ionization of certain compounds[62]. This strongly impairs the creation of exchangeable MS/MS data sets.

3.1.3. *In silico* fragmentation tools for theoretical natural product databases

For the reasons outlined above, *in silico* fragmentation tools namely MS-FINDER[63,64], CFM-ID[65,66], MetFrag[67,68], CSI:FingerID[16], or the commercial feature from advanced chemistry development, Inc. (ACD/labs, MS-Fragmenter) were created. These algorithms can annotate compounds from DBs of structures and iteratively break bonds of each compound to produce a series of possible fragments, which are then compared to each ion in acquired MS/MS spectra. Here is a non-exhaustive list of some of the known DBs dedicated to or including structures of NPs downloadable in .sdf format:

- The Dictionary of Natural Products[69] (DNP), considered the most comprehensive, contains over 300k NPs from plants, animals, and microorganisms. For a license fee, users can download structures of these NPs along with data such as molecular structures and sources.

- Reaxys[70] and SciFinder[71] are commercial DBs, each containing a collection of over 220k NPs. By purchasing a license, users can selectively access the structures of these NPs within these DBs.
- The Universal Natural Products Database (UNPD)[72], with more than 229k NPs, provides 3D structures and downloadable molecular descriptors.
- The collection of open natural products (COCONUT)[73], contains over 411k NPs sourced from multiple accessible DBs. The structures can be freely downloaded in .sdf format, but selection on the basis of chemotaxonomic criteria has not yet been fully developed.
- LOTUS[74] is a dynamic and collaborative DB offering structural search, and taxonomy-oriented queries,
- The *In-silico* MS/MS database (ISDB)[75] is an *in silico* MS/MS DB with more than 170k NPs and their MS spectra in positive mode.

To improve annotation rates, DB type restrictions such as species, genus or plant families can be applied to the most recent databases. However, *in silico* fragmentation algorithms still need to improve. Research has shown that only 17-25% of compounds can be correctly identified using these algorithms (CFM-ID being the best individual tool for correct annotation)[76]. However, best results are obtained by combining *in silico* algorithms with additional, compound specific information (*e.g.*, RT, molecular formula, collision cross section (CCS) value, etc.).

3.2. NMR dereplication workflows

The other common technique used to identify NPs is NMR. This method provides detailed structural information, allowing the determination of connectivity, stereochemistry, and functional groups. In addition, NMR is a non-destructive technique that preserves the sample for further analysis. The combination of these advantages makes NMR a powerful tool for the accurate identification and comprehensive understanding of metabolite structures and properties. NMR spectrometers now provide data sets (1D, 2D) that allow complete structural determination, including stereoisomers, within a reasonable amount of time. Combined with information such as molecular formula or X-ray experiments, NMR ensures a high level of confidence in metabolite annotation at level 1.

Traditionally, NMR requires a complete set of 1D and 2D data (^{1}H, ^{13}C, DEPT, COSY, HSQC, HMBC) to describe a novel metabolite. With enough sample of a purified NP and a powerful NMR instrument, experiments can be performed quickly. However, this purification step can be laborious and lead to the isolation of a NP that is already known and/or in very limited quantity, the structural determination of which will be unnecessarily time-consuming. Therefore, tools based on mixture dereplication have also been developed to rapidly identify known compounds.

3.2.1. ^{13}C NMR-based dereplication strategies and related tools

The first computer-assisted ^{13}C NMR approach for chemical profiling of crude extracts, regardless of the types of NPs they contain, was developed in 2017 and is called “caractérisation de mélanges” in French (CARAMEL[77], meaning mixture characterization). The strategy is based on the organization and visualization by clusters of the entire chemical shifts (δ_C) of the ^{13}C NMR spectrum using HCA. Each cluster theoretically corresponds to the δ_C of a given NP. A DB containing structures of selected NPs together with their experimental or predicted ^{13}C chemical shifts ($\delta_{C\text{-}SDF}$) and data set is used and processed by a software from ACD/Labs called C+H Predictors and DB[78]. Finally, the experimental δ_C are compared with those of NPs ($\delta_{C\text{-}SDF}$) from this DB. This allows the assignment of a NP to specific NMR chemical shifts clusters.

In 2020, another dereplication tool based on ^{13}C NMR (MixONat) was developed by A. Bruguière, *et al.*[79–81]. The software, which is distributed free of charge, enables mixtures to be dereplicated based on the ^{13}C NMR spectrum, optionally combined with DEPT-135 and -90 spectra to discriminate between CH_3, CH_2, CH and Cq. In ^{13}C NMR-based dereplication carbon multiplicity is a discriminating filter[79]. A DB of predicted and/or experimental $\delta_{C\text{-}SDF}$ categorized by carbon types is required. The KnapsackSearch algorithm first described by Nuzillard in 2021 facilitates this task using LOTUS DB and either the software ACD/labs, NMR predictors (C,H) or nmrshiftsdb2[80,82,83]. Results can then be refined based on user-defined parameters.

With these two procedures, the isolation of single NP from complex mixtures are no longer required for the confirmation of known compounds. Each strategy enables the quick view of the major NPs in extracts without any loss of sample, which means that 100% of the starting extract mass can be recovered at the end of the identification process.

3.2.2. Natural products databases for ^{13}C NMR-based dereplication processing

As mentioned in the previous section, NPs DBs are mandatory for dereplication algorithms. As with MS data, these DBs can be generated from either experimental or predicted chemical shifts.

If DBs containing experimental δ_C of the NP of interest in the given deuterated solvent are available, their use obviously increases the chances of better matches. Despite the availability of NMR DBs for NPs (see 3.1.3), none of them, to the best of our knowledge, is optimized for ^{13}C NMR-based dereplication of mixtures. Such a DB would need to be exhaustive, freely, and available in a format that is compatible with the tools mentioned before. It must further include the structures of known NPs and provide filter options for chemotaxonomic data, and experimental chemical shifts of ^{1}H and ^{13}C in different deuterated

solvents. Nevertheless, collaborative initiatives are underway in this direction[74,84,85].

Alternatively, $\delta_{C\text{-}SDF}$ may be predicted[86]. Several commercial and open-access tools have been developed to facilitate these predictions based on the structures of NPs. Therefore, these NPs DBs, filtered or not according to parameters, may be relevant for the user. As mentioned above (see 3.1.3), one such commercial software is ACD/Labs NMR predictors (C,H)[87], for which the variation of $\delta_{C\text{-}SDF}$ predictions is within +/-1.3 ppm in more than 70% of cases.[86]. On the other hand, freely available algorithms such as nmrshiftdb2[88,89] may also be used. These tools use large DBs of known δ_C and employ algorithms to generate prediction based on structural similarity. As an example, in 2022, Kuhn and Nuzillard created the acd_lotusv7 DB (updated version: acd_lotusv9 DB)[90], which is compatible with CARAMEL and MixONat. It is derived from the open-source LOTUS DB of NPs structures and includes predicted $\delta_{C\text{-}SDF}$ of 218,478 NPs last version. This initiative corresponds to an update of the first open access approach called Predicted ^{13}C NMR data of NPs (PNMRNP) DB[91], which includes data from UNPD and KnapsackSearch (a database generator providing taxonomically focused libraries of NPs). The PNMRNP database is based on three pillars of structural dereplication, biological taxonomy, knowledge of the structure of metabolites, and their spectroscopic fingerprints.

3.3. The use and prerequisites of NMR and MS

The combination of NMR and MS can offer significant advantages for metabolite identification in complex mixtures. By exploiting the assets of both techniques, a more comprehensive characterization of metabolites can be achieved. NMR provides detailed structural information allowing the determination of connectivity and stereochemistry. On the other hand, MS provides accurate mass measurements and fragmentation patterns for structural elucidation.

Integrating NMR and MS data can also help resolve ambiguities and distinguish between isomeric compounds. However, the combined use of NMR and MS may require expertise in both techniques and sophisticated data analysis approaches to integrate and interpret the data generated.

The choice between NMR and MS for metabolite identification in complex mixtures also depends on the amount of material available. Due to its lower sensitivity, NMR typically requires a larger amount of sample than MS. In cases where the amount of sample is limited, NMR may not be the preferred method. On the other hand, MS can work with smaller sample sizes and is more suitable for the analysis of trace metabolites.

However, it is important to note that advances in NMR instrumentation and techniques, such as cryogenic probes and microflow NMR, have improved the sensitivity of NMR analysis, allowing the analysis of smaller sample volumes and lower concentrations, thus narrowing the gap between NMR and MS in terms of sample requirements[92,93].

3.4. Molecular network dereplication

Another way to dereplicate data, is the use of MN. This computational approach has revolutionized the field of NPs research in recent years.

3.4.1. Definition

Developed in 2012 by M. Wang et al.[12], the MNs allow the visualization and organization of tandem MS/MS data. Each spectrum is represented by a node. These nodes are connected by an edge if their pairwise similarities are above a similarity threshold. In this way, similar structures are grouped into molecular families. By organizing the molecules, the MNs enables the identification of known compounds, the discovery of new analogs, and the exploration of chemical diversity within complex mixtures. These MNs are also known as classical molecular networks (CMN).

Moreover, a cosine score or modified cosine score is calculated to measures the similarity between two MS/MS spectra. The cosine score focuses on peak m/z matches, while the modified cosine score combines both the matching peak m/z and the peak m/z shifted by the precursor m/z difference. Both methods range from 0 to 1 (1 being a perfect match)[94,95].

In addition to the basic concept of CMN, it is important to emphasize that different molecular networking approaches have been developed to address specific research needs and challenges.

3.4.2. The different types of molecular network based on mass spectrometry data

3.4.2.1. Feature-based molecular network

Based on MS/MS comparison only, CMN may be affected by certain MS clustering limitations. In this regard, A. T. Aron, *et al.*[96] have developed an alternative workflow, known as feature-based molecular network (FBMN) in 2017. The method solves problems associated with assigning constitutional isomers/stereoisomers that have similar MS/MS spectra but different RT or IMS values (**a**). It also reduces the redundancy of duplicate nodes with identical precursor ions (**b**) and facilitates the creation of a unique representative MS/MS spectrum (**c**). In addition, FBMN increases the robustness of the statistical analysis by providing a more accurate estimate of the relative ion intensities. Unlike CMN, the latter is achieved by using peak areas or heights rather than spectral counts or summed precursor ion counts.

FBMN requires pre-processed data from MZmine[97], MS-DIAL, MetaboScape, ProgenesisQI, or others software, that allow the alignment and feature detection[50]. Next an exported list of the feature quantification table (.csv file) and a tandem MS/MS spectral summary (.msp or .mgf file) are uploaded to the GNPS to generate the FBMN.

3.4.2.2. Ion identity molecular network

The ion identity molecular network (IIMN) workflow is a complement to FBMN and was developed by R. Schmid, *et*

al.[98] in 2021. During LC-HRMS ionization, a given compound may generate multiple ion species ($[M-H]^-$, $[M+Na]^+$, and others) that appear as individual nodes in CMN or FBMN, due to the different precursor m/z. The IIMN workflow aims to group them into a single node.

However, some post processing software like ProgenesisQI include this step in their workflow. Ions and their adduct are grouped together, in a process called deconvolution.

3.4.2.3. Building blocks-based molecular network

Another strategy, called building blocks-based molecular network (BBMN) was developed by Q.-F. He, *et al.* in 2021. Compared to the previously discussed methods, this workflow is highly selective and fast. It can recognize characteristic fragments derived from specific biogenetic building blocks, which could selectively identify biogenetically relevant compounds. As a result, bulky MS^2 data can be transformed into a concise and targeted form, allowing for faster data analysis.

3.4.2.4. MetGem

Developed in 2018 by F. Olivon *et al.*, MetGem is based on the generation of CMN in a first window and on the t-distributed stochastic neighbor embedding (t-SNE) based visualization[13]. The latter preserves the interactions between related groups of spectra, while the CMN output allows a clear separation of clusters.

MetGem clusters spectra by focusing on local details within the entire data rather than individual connections between spectra. As a result, it avoids having too many "self-loop nodes" or merging "molecular families" in the molecular network when a similarity cutoff is set[99]. However, the t-SNE based method cannot not provide information about the relationships between "nodes", and it is complementary to the cosine score from GNPS based CMN.

3.4.3. The different types of molecular network based on nuclear magnetic resonance data

As it continues to improve, the MS dereplication workflow is the most widely applied and used workflow. However, it has its limitations, such as non-exhaustive ionization of compounds and sometimes failure to distinguish between isomers.

The metabolomics and dereplication by two-dimensional experiments (MADbyTE) is the first NMR-based MN dereplication and was developed in 2021 by J. M. Egan *et al*.[15]. The algorithm allows the identification of spin system features within a complex mixture by heteronuclear and homonuclear experiments (HSQC and TOCSY). The common features between each of the samples can then be visualized by a chemical similarity network.

3.5. (Bio)activity-based molecular networks

Based on MN, several strategies have been developed to rapidly detect (bio)active NPs in complex mixtures. Additional information such as (bio)chemometrics or taxonomy can be integrated into a MN using tools such as the Cytoscape software[14].

The first combined strategy was described by L.-F. Nothias *et al.*[100] in 2021. In this study, a bioinformatic workflow was applied for the discovery of antiviral compounds from *Euphorbia dendroides*. After a fractionation step of the bioactive extract, LC-MS/MS experiments and bioactivity evaluations were performed on the fractions. This allowed the calculation of the bioactivity score, which is based on the Pearson correlation between feature ion intensities and the bioactivity level of fractions. The results were then combined with the MN, allowing the identification of two bioactive compounds against Chikungunya virus. The workflow was further successfully applied for the discovery of additional active compounds[101,102].

Another workflow, proposed by F. Olivon *et al.*[103], includes additional information such as bioactivity (IC_{50}) or taxonomic data in the MN. The resulting MN quickly detect potential compounds against the Wnt pathway and Chikungunya virus. Identification was performed by NMR experiment after their isolation.

As mentioned above, MADbyTE allows the creation of networks based on NMR data. The workflow can include (bio)activity data as an additional feature. In this way, it can assist in the elucidation of structure-activity relationships, especially when different spin systems have biological relevance. The efficiency of MADbyTE was demonstrated in a study by L. Flores-Bocanegra *et al.*[15]. The biologically based NMR network allowed the detection and discovery of three new palmarumycins with biological activity against transforming growth factor-β-activated kinase 1 (TAK1).

3.6. Advanced data analysis of molecular network

Advanced MN analysis tools have been developed to assist in the annotation of compounds in a complex mixture:

- MolNetEnhancer[104] is a workflow used to enhance molecular networks by incorporating additional data and information from MS2LDA, *in silico* annotation tools (NAP or DEREPLICATOR) and ClassyFire,
- MS2LDA[105–108] is used to identify groups of mass fragments and/or neutral losses, supported by MotifDB, an open database that allows users to efficiently search for characterized motifs in their own experiments,
- Network annotation propagation (NAP)[109,110] is a computational method for annotating molecular networks by propagating annotations from spectral library matching to improve the ranking of in silico fragmentation candidate structures,
- DEREPLICATOR(+)[111] is an in silico database search tool that allows the annotation of metabolites in MS/MS data using in silico fragmentation graph,
- MolDiscovery[112] is an MS DB search method that improves the efficiency and accuracy of small molecule identification by using a learned probabilistic model to

match small molecules to their corresponding mass spectra,
- The Reanalysis data user interface (ReDU)[113] is a community-based approach that facilitates the discovery and reuse of public mass spectrometry data by linking the GNPS platform to the MassIVE data repository. It allows researchers to access, analyse, or integrate public data with their own data, fostering collaborative exploration,
- Merge network polarity[114] is a method for combining molecular networks generated from positive and negative ionization modes in mass spectrometry-based experiments,
- Qemistree[115] is a computational tool that builds a tree based on MS/MS features, enabling the application of phylogeny-based tools to investigate the chemical composition of samples.

4. Confidence level of metabolite identification

The level of confidence in metabolite annotation in complex mixtures remains a challenge in NP research. Traditionally, a new NP is reported with its NMR spectra (1D and 2D), exact mass from HRMS, and complementary spectroscopic data (IR, UV). Absolute configuration of chiral (bio)active NPs is obtained by X-ray crystallography if possible, or alternative techniques such as NMR-based methods (e.g. Mosher approach) or chiroptical spectroscopies (e.g. Optical rotatory dispersion, electronic circular dichroism, vibrational circular dichroism)[116]. Best NP identification is achieved by combining different methods and their results. Despite these efforts misinterpretation of data can not be fully excluded.

With this in mind, a committee called the Metabolite Identification Task Group worked with the scientific community to develop the following final identification reporting standards[16,17]:
- Level 0, unambiguous 3D structure: reported by I. Blaženovic *et al.*[16], this level is assigned when the metabolite has been isolated and the full stereochemistry is provided. In this case, the 1D and 2D NMR data must be fully interpreted according to the NP guidelines,
- Level 1, confident 2D structure: indicates that the annotation is based on comparison with standards and that two types of orthogonal information were used for confirmation two orthogonal methods used for structural elucidation [*e.g.*, MS/MS and RT or CCS]. Thus, this level can only concern unambiguous identifications of known NPs but cannot prove the full characterization of unknows. The latter requires full interpretation of the NMR data, in addition to the accurate mass measured by HRMS, MS/MS data and RT,
- Level 2, putative structure annotation: in this case, the annotation is made by manual or *in silico* comparison of experimental data with the literature or databases. The annotation must be complete with at least two orthogonal pieces of information, such as MS/MS and exact mass, including evidence that excludes all other candidates,
- Level 3, putative compound class: indicates that partial structural annotation is possible based on spectral information only. The partial matches of the metabolites only provide confidence in possible isomers, compound class or substructure. This means that multiple candidates are possible,
- Level 4, unknown compound: only chromatographic or NMR data (uninterpretable alone) are available but with no correct or partial matches to the literature.

These five levels of identification have the advantage of being relatively easy to determine. However, it remains difficult to compare different annotations because not every data processing workflow involves the same instrument, raw data, or annotation strategy.

5. Conclusion

Living organisms biosynthesize a large number of potentially bioactive NPs with very diverse chemical structures. The prioritization of bioactive NPs which identity needs to be determined in these complex matrices is now largely guided by chemometric tools that can be tailored according to specific research objectives. The integration of MS and NMR has proven to be essential for the comprehensive annotation of selected NPs in complex mixtures. By combining the complementary assets of these orthogonal techniques, researchers can achieve a higher level of confidence in the identification and characterization of compounds in complex mixtures[17].

The DBs used for fast and efficient annotation of known metabolites provide access to a vast collection of information, including chemical structures, spectral data, chemical properties, sources, and biological activities of NPs. This saves time and resources, especially when dealing with complex mixtures. In addition, these DBs accelerate the research and discovery process by providing a starting point for scientists to explore compounds with specific properties or activities.

However, NP DBs also have limitations. Information, including spectral data, may be experimental or predictive, incomplete, or outdated, requiring researchers to critically evaluate and cross-reference multiple sources. Data quality and reliability can vary from DB to DB, requiring caution and verification. Limited accessibility and data integration challenges are additional barriers. It is also important to recognize that gaining expertise in these methods and ensuring data robustness through post-processing are critical steps.

In addition, the use of current algorithms includes accelerated compound identification and increased efficiency. These algorithms yield a list of structural proposals ranked by scores. However, these propositions

must be carefully inspected, as they need to be confirmed by orthogonal techniques, experimental data, or reference compounds. This will significantly increase their level of confidence.

Today, significant progress has been made in exploring artificial intelligence (AI) for NPs discovery[117–120]. It could revolutionize the field by accelerating and improving the process of efficiently analysing large amounts of data, predicting chemical structures, and identifying potential (bio)active compounds from complex mixtures. In addition, AI could facilitate the integration of diverse data sources, such as genomic, chemical, and biological information, allowing researchers to gain deeper insights into the complex interactions between NPs and their biological targets. Overall, the use of AI has the potential to open new frontiers and lead to the discovery of a wide range of NPs. This that can significantly impact various industries, including pharmaceuticals, nutraceuticals, and agrochemicals.

6. References

1 A. G. Atanasov, S. B. Zotchev, V. M. Dirsch, The International Natural Product Sciences Taskforce and C. T. Supuran, *Nat. Rev. Drug Discov.*, 2021, **20**, 200–216.

2 H. Yuan, Q. Ma, L. Ye and G. Piao, *Molecules*, 2016, **21**, 559.

3 K. Ahn, *BMB Reports*, 2017, **50**, 111–116.

4 C.-T. Che and H. Zhang, *IJMS*, 2019, **20**, 830.

5 S. P. Snini and F. Mathieu, *Toxins*, 2020, **12**, 353.

6 N. Malviya and S. Malviya, *Int. J. Clin. Pharmacol.*, 2017, 1–6.

7 M. G. Weller, *Sensors*, 2012, **12**, 9181–9209.

8 K. M. Wyss, G. C. Llivina and A. I. Calderón, *Comb. Chem. High Throughput Screen.*, 2019, **22**, 290–306.

9 L. Ory, E.-H. Nazih, S. Daoud, J. Mocquard, M. Bourjot, L. Margueritte, M.-A. Delsuc, J.-M. Bard, Y. F. Pouchus, S. Bertrand and C. Roullier, *Anal. Chim. Acta*, 2019, **1070**, 29–42.

10 J. J. Kellogg, D. A. Todd, J. M. Egan, H. A. Raja, N. H. Oberlies, O. M. Kvalheim and N. B. Cech, *J. Nat. Prod.*, 2016, **79**, 376–386.

11 J. Hubert, J.-M. Nuzillard and J.-H. Renault, *Phytochem. Rev.*, 2017, **16**, 55–95.

12 M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.

13 F. Olivon, N. Elie, G. Grelier, F. Roussi, M. Litaudon and D. Touboul, *Anal. Chem.*, 2018, **90**, 13900–13908.

14 C. T. Lopes, M. Franz, F. Kazi, S. L. Donaldson, Q. Morris and G. D. Bader, *Bioinformatics*, 2010, **26**, 2347–2348.

15 L. Flores-Bocanegra, Z. Y. Al Subeh, J. M. Egan, T. El-Elimat, H. A. Raja, J. E. Burdette, C. J. Pearce, R. G. Linington and N. H. Oberlies, *J. Nat. Prod.*, 2022, **85**, 614–624.

16 I. Blaženović, T. Kind, J. Ji and O. Fiehn, *Metabolites*, 2018, **8**, 31.

17 J.-L. Wolfender, J.-M. Nuzillard, J. J. J. Van Der Hooft, J.-H. Renault and S. Bertrand, *Anal. Chem.*, 2019, **91**, 704–742.

18 P.-M. Allard, G. Genta-Jouve and J.-L. Wolfender, *Curr. Opin. Chem. Biol.*, 2017, **36**, 40–49.

19 A. E. Fox Ramos, L. Evanno, E. Poupon, P. Champy and M. A. Beniddir, *Nat. Prod. Rep.*, 2019, **36**, 960–980.

20 S. Wold and M. Sjöström, *Chemom. Intell. Lab. Syst.*, 1998, **44**, 3–14.

21 P. Gemperline, Ed., *Practical guide to chemometrics*, CRC/Taylor & Francis, Boca Raton, 2nd ed., 2006.

22 S. D. Brown, *J. Chemom.*, 2017, **31**, e2856.

23 S. Wold and M. Sjöström, *Chemometrics: Theory and Application*, , DOI:10.1021/bk-1977-0052.

24 S. D. Brown, *J. Chemom.*, 1995, **9**, 527–529.

25 H. Martens, S. W. Bruun, I. Adt, G. D. Sockalingum and A. Kohler, *J. Chemom.*, 2006, **20**, 402–417.

26 R. W. Farebrother, *J. Multivar. Anal.*, 2022, **188**, 104814.

27 J. Lever, M. Krzywinski and N. Altman, *Nat. Methods*, 2017, **14**, 641–642.

28 A. L. Charles and M. A. Alamsjah, *Food Chem.*, 2019, **289**, 269–277.

29 R. R. Meglen, *J. Chemometrics*, 1991, **5**, 163–179.

30 C. L. C. Meira, C. G. Novaes, F. C. Novais, V. da S. de Jesus, D. M. de Oliveira and R. M. Aguiar, *Microchem. J.*, 2020, **152**, 104284.

31 P. Geladi, *J. Chemom.*, 1988, **2**, 231–246.

32 S. Wold, P. Geladi, K. Esbensen and J. Öhman, *J. Chemom.*, 1987, **1**, 41–56.

33 S. Barbosa, J. Saurina, L. Puignou and O. Núñez, *Foods*, 2020, **9**, 486.

34 S. P. Serbin, J. Wu, K. S. Ely, E. L. Kruger, P. A. Townsend, R. Meng, B. T. Wolfe, A. Chlus, Z. Wang and A. Rogers, *New Phytologist*, 2019, **224**, 1557–1568.

35 J. Trygg and S. Wold, *J. Chemom.*, 2002, **16**, 119–128.

36 T. Rajalahti, R. Arneberg, F. S. Berven, K.-M. Myhr, R. J. Ulvik and O. M. Kvalheim, *Chemom. Intell. Lab. Syst.*, 2009, **95**, 35–48.

37 L. M. Casanova, L. M. Rodrigues, P. F. de Aguiar and L. W. Tinoco, *J. Nat. Prod.*, 2020, 83, 243–254.

38 F. V. Carvalho, L. Fonseca Santana, V. Diogenes A. da Silva, S. L. Costa, L. Zambotti-Villelae, P. Colepicolo, C. G. Ferraz and P. R. Ribeiro, *Food Chem.*, 2021, **364**, 130453.

39 P. Aminfar, M. Abtahi and H. Parastar, *J. Chromatogr. A*, 2019, 1602, 432–440.

40 G. K. More, S. Meddows-Taylor and G. Prinsloo, *Molecules*, 2021, 26.

41 E. Plazas, R. Casoti, M. Avila Murillo, F. B. Da Costa and L. Enrique Cuca, *PHYTOCHEMISTRY*, 2019, 168.

42 Z. Li, Y. Shi, X. Zhang, J. Xu, H. Wang, L. Zhao and Y. Wang, *Front. Pharmacol.*, 2020, **11**, 287.

43 L. Margueritte, P. Markov, L. Chiron, J.-P. Starck, C. Vonthron-Sénécheau, M. Bourjot and M.-A. Delsuc, *Magn. Reson. Chem.*, 2018, **56**, 469–479.

44 R: The R Project for Statistical Computing, https://www.r-project.org/, (accessed August 19, 2022).

45 D. P. Demarque, R. G. Dusi, F. D. M. de Sousa, S. M. Grossi, M. R. S. Silvério, N. P. Lopes and L. S. Espindola, *Sci. Rep.*, 2020, **10**, 1051.

46 D. S. Alves, A. R. T. Machado, V. A. C. Campos, D. F. Oliveira and G. A. Carvalho, *J. Econ. Entomol.*, 2016, 109, 649–659.

47 L. K. Caesar, J. J. Kellogg, O. M. Kvalheim and N. B. Cech, *J. Nat. Prod.*, 2019, **82**, 469–484.

48 K. V. Sashidhara and J. N. Rosaiah, *Nat. Prod. Commun.*, 2007, **2**, 193–202.

49 T. Kind and O. Fiehn, *Phytochem. Lett.*, 2017, **21**, 313–319.

50 L.-F. Nothias, D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. A. Aksenov, O. Alka, P.-M. Allard, A. Barsch, X. Cachet, A. M. Caraballo-Rodriguez, R. R. Da Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kameník, K. B. Kang, N. Kessler, I. Koester, A. Korf, A. Le Gouellec, M. Ludwig, C. Martin H., L.-I. McCall, J. McSayles, S. W. Meyer, H. Mohimani, M. Morsy, O. Moyne, S. Neumann, H. Neuweger, N. H. Nguyen, M. Nothias-Esposito, J. Paolini, V. V. Phelan, T. Pluskal, R. A. Quinn, S. Rogers, B. Shrestha, A. Tripathi, J. J. J. van der Hooft, F. Vargas, K. C. Weldon, M. Witting, H. Yang, Z. Zhang, F. Zubeil, O. Kohlbacher, S. Böcker, T. Alexandrov, N. Bandeira, M. Wang and P. C. Dorrestein, *Nat Methods*, 2020, **17**, 905–908.

51 G. Alvarez-Rivera, D. Ballesteros-Vivas, F. Parada-Alfonso, E. Ibañez and A. Cifuentes, *TrAC*, 2019, **112**, 87–101.

52 S. Tiwary, R. Levy, P. Gutenbrunner, F. Salinas Soto, K. K. Palaniappan, L. Deming, M. Berndl, A. Brant, P. Cimermancic and J. Cox, *Nat. Methods*, 2019, **16**, 519–525.

53 L. Deng, A.-M. Shi, H.-Z. Liu, N. Meruva, L. Liu, H. Hu, Y. Yang, C. Huang, P. Li and Q. Wang, *J. Mass Spectrom.*, 2016, **51**, 1157–1167.

54 H. Pan, H. Zhou, S. Miao, J. Cao, J. Liu, L. Lan, Q. Hu, X. Mao and S. Ji, *J. Chromatogr. A*, 2020, **1613**, 460674.

55 T. Kind, H. Tsugawa, T. Cajka, Y. Ma, Z. Lai, S. S. Mehta, G. Wohlgemuth, D. K. Barupal, M. R. Showalter, M. Arita and O. Fiehn, *Mass Spec Rev*, 2018, **37**, 513–532.

56 T. Pluskal, S. Castillo, A. Villar-Briones and M. Orešič, *BMC Bioinform.*, 2010, **11**, 395.

57 H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn and M. Arita, *Nat Methods*, 2015, **12**, 523–526.

58 M. Sorokina and C. Steinbeck, *J. Cheminform.*, 2020, **12**, 20.

59 C. Guijas, J. R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A. E. Aisporna, D. W. Wolan, M. E. Spilker, H. P. Benton and G. Siuzdak, *Anal. Chem.*, 2018, **90**, 3156–3164.

60 J. Wang, D. A. Peake, R. Mistrik and Y. Huang, *Blood*, 2013, **4**, 2–8.

61 H. Oberacher, M. Sasse, J.-P. Antignac, Y. Guitton, L. Debrauwer, E. L. Jamin, T. Schulze, M. Krauss, A. Covaci, N. Caballero-Casero, K. Rousseau, A. Damont, F. Fenaille, M. Lamoree and E. L. Schymanski, *Environ. Sci. Eur.*, 2020, **32**, 43.

62 J. L. Josephs and M. Sanders, *Rapid Commun. Mass Spectrom.*, 2004, **18**, 743–759.

63 Z. Lai, H. Tsugawa, G. Wohlgemuth, S. Mehta, M. Mueller, Y. Zheng, A. Ogiwara, J. Meissen, M. Showalter, K. Takeuchi, T. Kind, P. Beal, M. Arita and O. Fiehn, *Nat. Methods*, 2018, **15**, 53–56.

64 H. Tsugawa, T. Kind, R. Nakabayashi, D. Yukihira, W. Tanaka, T. Cajka, K. Saito, O. Fiehn and M. Arita, *Anal. Chem.*, , DOI:10.1021/acs.analchem.6b00770.

65 F. Allen, A. Pon, M. Wilson, R. Greiner and D. Wishart, *Nucleic Acids Research*, 2014, **42**, W94–W99.

66 F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner and D. S. Wishart, *Anal. Chem.*, 2021, **93**, 11692–11700.

67 C. Ruttkies, S. Neumann and S. Posch, *BMC Bioinform.*, 2019, **20**, 376.

68 C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender and S. Neumann, *J. Cheminformatics*, 2016, **8**, 3.

69 Dictionary of natural products 31.2, https://dnp.chemnetbase.com/chemical/ChemicalSearch.xhtml?dswid=4743, (accessed March 13, 2023).

70 Reaxys, https://www.elsevier.com/fr-fr/solutions/reaxys, (accessed March 13, 2023).

71 CAS SciFinder^{n}, https://scifinder-n-cas-org.buadistant.univ-angers.fr/, (accessed March 13, 2023).

72 J. Gu, Y. Gui, L. Chen, G. Yuan, H.-Z. Lu and X. Xu, *PLoS ONE*, 2013, **8**, e62839.

73 M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik and C. Steinbeck, *J. Cheminform.*, 2021, **13**, 2.

74 A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, A. Gaudry, J. G. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson and P.-M. Allard, *eLife*, 2022, **11**, e70780.

75 P.-M. Allard, T. Péresse, J. Bisson, K. Gindro, L. Marcourt, V. C. Pham, F. Roussi, M. Litaudon and J.-L. Wolfender, *Anal. Chem.*, 2016, **88**, 3317–3323.

76 I. Blaženović, T. Kind, H. Torbašinović, S. Obrenović, S. S. Mehta, H. Tsugawa, T. Wermuth, N. Schauer, M. Jahn, R. Biedendieck, D. Jahn and O. Fiehn, *J. Cheminform.*, 2017, **9**, 32.

77 A. Bakiri, B. Plainchont, V. de Paulo Emerenciano, R. Reynaud, J. Hubert, J.-H. Renault and J.-M. Nuzillard, *Mol. Inform.*, 2017, **36**, 1700027.

78 J.-M. Nuzillard, *Analytica*, 2021, **2**, 50–56.

79 A. Bruguière, S. Derbré, J. Dietsch, J. Leguy, V. Rahier, Q. Pottier, S. Suor-Cherer and G. Viault, *Anal. Chem.*, 2020, **92**, 8793–8801.

80 A. Bruguière, S. Derbré, D. Bréard, F. Tomi, J.-M. Nuzillard and P. Richomme, *Planta Med.*, 2021, **87**, 1061–1068.

81 A. Bruguière, S. Derbré and P. Richomme, MixONat, https://sourceforge.net/projects/mixonat/, (accessed January 14, 2022).

82 J.-M. Nuzillard, KnapsackSearch https://github.com/nuzillard/KnapsackSearch 2023.

83 S. Kuhn and J.-M. Nuzillard, *Chemistry–Methods*, 2023, **3**, 1–5.

84 M. Pupier, J.-M. Nuzillard, J. Wist, N. E. Schlörer, S. Kuhn, M. Erdelyi, C. Steinbeck, A. J. Williams, C. Butts, T. D. W. Claridge, B. Mikhova, W. Robien, H. Dashti, H. R. Eghbalnia, C. Farès, C. Adam, P. Kessler, F. Moriaud, M. Elyashberg, D. Argyropoulos, M. Pérez, P. Giraudeau, R. R. Gil, P. Trevorrow and D. Jeannerat, *Magn. Reson. Chem.*, 2018, **56**, 703–715.

85 D. S. Wishart, Z. Sayeeda, Z. Budinski, A. Guo, B. L. Lee, M. Berjanskii, M. Rout, H. Peters, R. Dizon, R. Mah, C. Torres-Calzada, M. Hiebert-Giesbrecht, D. Varshavi, D. Varshavi, E. Oler, D. Allen, X. Cao, V. Gautam, A. Maras, E. F. Poynton, P. Tavangar, V. Yang, J. A. van Santen, R. Ghosh, S. Sarma, E. Knutson, V. Sullivan, A. M. Jystad, R. Renslow, L. W. Sumner, R. G. Linington and J. R. Cort, *Nucleic Acids Res.*, 2022, **50**, D665–D677.

86 A. Bruguière, S. Derbré, C. Coste, M. Le Bot, B. Siegler, S. T. Leong, S. N. Sulaiman, K. Awang and P. Richomme, *Fitoterapia*, 2018, **131**, 59–64.

87 Predict NMR spectra with confidence, https://www.acdlabs.com/products/spectrus-platform/nmr-predictors/, (accessed June 23, 2023).

88 C. Steinbeck, S. Krause and S. Kuhn, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1733–1739.

89 C. Steinbeck and S. Kuhn, *Phytochemistry*, 2004, **65**, 2711–2717.

90 J.-M. Nuzillard, 2022.

91 M. Lianza, R. Leroy, C. Machado Rodrigues, N. Borie, C. Sayagh, S. Remy, S. Kuhn, J.-H. Renault and J.-M. Nuzillard, *Molecules*, 2021, **26**, 637.

92 Z.-F. Wang, Y.-L. You, F.-F. Li, W.-R. Kong and S.-Q. Wang, *Molecules*, 2021, **26**, 6308.

93 M. A. Salem, L. Perez De Souza, A. Serag, A. R. Fernie, M. A. Farag, S. M. Ezzat and S. Alseekh, *Metabolites*, 2020, **10**, 37.

94 F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers and J. J. J. van der Hooft, *PLoS Comput Biol*, 2021, **17**, e1008724.

95 W. Bittremieux, R. Schmid, F. Huber, J. J. J. van der Hooft, M. Wang and P. C. Dorrestein, *J. Am. Soc. Mass Spectrom.*, , DOI:10.1021/jasms.2c00153.

96 A. T. Aron, E. C. Gentry, K. L. McPhail, L.-F. Nothias, M. Nothias-Esposito, A. Bouslimani, D. Petras, J. M. Gauglitz, N. Sikora, F. Vargas, J. J. J. van der Hooft, M. Ernst, K. B. Kang, C. M. Aceves, A. M. Caraballo-Rodríguez, I. Koester, K. C. Weldon, S. Bertrand, C. Roullier, K. Sun, R. M. Tehan, C. A. Boya P., M. H. Christian, M. Gutiérrez, A. M. Ulloa, J. A. Tejeda Mora, R. Mojica-Flores, J. Lakey-Beitia, V. Vásquez-Chaves, Y. Zhang, A. I. Calderón, N. Tayler, R. A. Keyzers, F. Tugizimana, N. Ndlovu, A. A. Aksenov, A. K. Jarmusch, R. Schmid, A. W. Truman, N. Bandeira, M. Wang and P. C. Dorrestein, *Nat. Protoc.*, 2020, **15**, 1954–1991.

97 FBMN with Progenesis QI - GNPS Documentation, https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-progenesisQI/, (accessed May 20, 2021).

98 R. Schmid, D. Petras, L.-F. Nothias, M. Wang, A. T. Aron, A. Jagels, H. Tsugawa, J. Rainer, M. Garcia-Aloy, K. Dührkop, A. Korf, T. Pluskal, Z. Kameník, A. K. Jarmusch, A. M. Caraballo-Rodríguez, K. C. Weldon, M. Nothias-Esposito, A. A. Aksenov, A. Bauermeister, A. Albarracin Orio, C. O. Grundmann, F. Vargas, I. Koester, J. M. Gauglitz, E. C. Gentry, Y. Hövelmann, S. A. Kalinina, M. A. Pendergraft, M. Panitchpakdi, R. Tehan, A. Le Gouellec, G. Aleti, H. Mannochio Russo, B. Arndt, F. Hübner, H. Hayen, H. Zhi, M. Raffatellu, K. A. Prather, L. I. Aluwihare, S. Böcker, K. L. McPhail, H.-U. Humpf, U. Karst and P. C. Dorrestein, *Nat. Commun.*, 2021, **12**, 3832.

99 G.-F. Qin, X. Zhang, F. Zhu, Z.-Q. Huo, Q.-Q. Yao, Q. Feng, Z. Liu, G.-M. Zhang, J.-C. Yao and H.-B. Liang, *Molecules*, 2022, **28**, 157.

100 L.-F. Nothias, M. Nothias-Esposito, R. da Silva, M. Wang, I. Protsyuk, Z. Zhang, A. Sarvepalli, P. Leyssen, D.

Touboul, J. Costa, J. Paolini, T. Alexandrov, M. Litaudon and P. C. Dorrestein, *J. Nat. Prod.*, 2018, **81**, 758–767.

101 Z. Feng, J. Chen, C. Chen, L. Feng, R. Wang, J. Zhu, R. Lou, J. Liu, Y. Ye and L. Lin, *Br. J. Pharmacol.*, 2023, **180**, 589–608.

102 L. Buedenbender, F. A. Astone and D. Tasdemir, *Marine Drugs*, 2020, **18**, 311.

103 F. Olivon, P.-M. Allard, A. Koval, D. Righi, G. Genta-Jouve, J. Neyts, C. Apel, C. Pannecouque, L.-F. Nothias, X. Cachet, L. Marcourt, F. Roussi, V. L. Katanaev, D. Touboul, J.-L. Wolfender and M. Litaudon, *ACS Chem. Biol.*, 2017, **12**, 2644–2651.

104 M. Ernst, K. B. Kang, A. M. Caraballo-Rodríguez, L.-F. Nothias, J. Wandy, C. Chen, M. Wang, S. Rogers, M. H. Medema, P. C. Dorrestein and J. J. J. Van Der Hooft, *Metabolites*, 2019, **9**, 144.

105 J. Wandy, Y. Zhu, J. J. J. van der Hooft, R. Daly, M. P. Barrett and S. Rogers, *Bioinformatics*, 2018, **34**, 317–318.

106 S. Rogers, C. W. Ong, J. Wandy, M. Ernst, L. Ridder and J. J. J. Van Der Hooft, *Faraday Discuss.*, 2019, **218**, 284–302.

107 J. J. J. Van Der Hooft, J. Wandy, F. Young, S. Padmanabhan, K. Gerasimidis, K. E. V. Burgess, M. P. Barrett and S. Rogers, *Anal. Chem.*, 2017, **89**, 7569–7577.

108 J. J. J. Van Der Hooft, J. Wandy, M. P. Barrett, K. E. V. Burgess and S. Rogers, *PNAS*, 2016, **113**, 13738–13743.

109 K. B. Kang, E. J. Park, R. R. Da Silva, H. W. Kim, P. C. Dorrestein and S. H. Sung, *J. Nat. Prod.*, 2018, **81**, 1819–1828.

110 R. R. Da Silva, M. Wang, L.-F. Nothias, J. J. J. Van Der Hooft, A. M. Caraballo-Rodríguez, E. Fox, M. J. Balunas, J. L. Klassen, N. P. Lopes and P. C. Dorrestein, *PLoS Comput. Biol.*, 2018, **14**, e1006089.

111 H. Mohimani, A. Gurevich, A. Shlemov, A. Mikheenko, A. Korobeynikov, L. Cao, E. Shcherbin, L.-F. Nothias, P. C. Dorrestein and P. A. Pevzner, *Nat. Commun.*, 2018, **9**, 4035.

112 L. Cao, M. Guler, A. Tagirdzhanov, Y.-Y. Lee, A. Gurevich and H. Mohimani, *Nat. Commun.*, 2021, **12**, 3718.

113 A. K. Jarmusch, M. Wang, C. M. Aceves, R. S. Advani, S. Aguirre, A. A. Aksenov, G. Aleti, A. T. Aron, A. Bauermeister, S. Bolleddu, A. Bouslimani, A. M. Caraballo Rodriguez, R. Chaar, R. Coras, E. O. Elijah, M. Ernst, J. M. Gauglitz, E. C. Gentry, M. Husband, S. A. Jarmusch, K. L. Jones, Z. Kamenik, A. Le Gouellec, A. Lu, L.-I. McCall, K. L. McPhail, M. J. Meehan, A. V. Melnik, R. C. Menezes, Y. A. Montoya Giraldo, N. H. Nguyen, L. F. Nothias, M. Nothias-Esposito, M. Panitchpakdi, D. Petras, R. A. Quinn, N. Sikora, J. J. J. van der Hooft, F. Vargas, A. Vrbanac, K. C. Weldon, R. Knight, N. Bandeira and P. C. Dorrestein, *Nat. Methods*, 2020, **17**, 901–904.

114 Merge Network Polarity - GNPS Documentation, https://ccms-ucsd.github.io/GNPSDocumentation/mergepolarity/#page-contributors, (accessed May 22, 2023).

115 A. Tripathi, Y. Vázquez-Baeza, J. M. Gauglitz, M. Wang, K. Dührkop, M. Nothias-Esposito, D. D. Acharya, M. Ernst, J. J. J. van der Hooft, Q. Zhu, D. McDonald, A. D. Brejnrod, A. Gonzalez, J. Handelsman, M. Fleischauer, M. Ludwig, S. Böcker, L.-F. Nothias, R. Knight and P. C. Dorrestein, *Nat. Chem. Biol.*, 2021, **17**, 146–151.

116 E. Santoro, S. Vergura, P. Scafato, S. Belviso, M. Masi, A. Evidente and S. Superchi, *J. Nat. Prod.*, 2020, **83**, 1061–1068.

117 F. I. Saldívar-González, V. D. Aldas-Bulos, J. L. Medina-Franco and F. Plisson, *Chem. Sci.*, 2022, **13**, 1526–1546.

118 M. Liu, P. Karuso, Y. Feng, E. Kellenberger, F. Liu, C. Wang and R. J. Quinn, *Medchemcomm*, 2019, **10**, 1667–1677.

119 L. Zhang, J. Song, L. Kong, T. Yuan, W. Li, W. Zhang, B. Hou, Y. Lu and G. Du, *Pharmacology & Therapeutics*, 2020, **216**, 107686.

120 H. T. Xue, M. Stanley-Baker, A. W. K. Kong, H. L. Li and W. W. B. Goh, Drug Discovery Today, 2022, 27, 2235–2243.

2. General methodologies and experimental conditions

For a better understanding of the discussion part of this manuscript, this chapter provides a comprehensive description of the methodologies and experimental conditions that were used.

The first part of this chapter defines and explains the **anti-AGEs assays** used. As mentioned in the introduction, previous studies on the four *Garcinia parvifolia* bark extracts (KL5670, KL5259, KL5248, KL5073) used in my PhD have shown significant inhibitory effects on the formation of AGEs in an in-house assay (Derbré *et al.*, 2010).

The second part is dedicated to the presentation of the **MS methods and their experimental conditions** that were applied. First, a detailed description of the recently acquired UPLC-HRMS will be presented, along with a step-by-step description of the post processing of the data using the ProgenesisQI software. The LDI-HRMS technique including the workflow for post-acquisition data treatment is also discussed. In addition, ion mobility spectrometry (IMS) is presented as an alternative method for the detection of isomers that cannot be distinguished by the above techniques.

The third part of this chapter is focuses on the **NMR methods and experiments** that were used. First the experimental NMR settings are discussed, including acquisition parameters such as the type of NMR instrument used, sample preparation, and the type of spectral acquisition. Next, the focus is placed on the data processing used. This includes a detailed description of the MixONat software and its workflow, as well as the used DBs for the annotation of NPs.

[FR]

Pour une meilleure compréhension de la partie discussion de ce manuscrit, ce chapitre fournit une description détaillée des méthodologies et des conditions expérimentales qui ont été utilisées.

La première partie de ce chapitre définit et explique les **tests anti-AGEs** utilisés. Comme mentionné dans l'introduction, des études antérieures sur les quatre extraits d'écorce de *Garcinia parvifolia* (KL5670, KL5259, KL5248, KL5073) utilisés dans le cadre de ma thèse ont montré des effets inhibiteurs significatifs sur la formation des AGEs (Derbré et al., 2010).

La deuxième partie est consacrée à la présentation des **méthodes par MS et de leurs conditions expérimentales** qui ont été appliquées. Tout d'abord, une description détaillée de l'UPLC-HRMS récemment acquise sera présentée, ainsi qu'une description étape par étape du post-traitement des données à l'aide du logiciel ProgenesisQI. La technique d'analyse par LDI-HRMS, y compris le traitement des données après acquisition, est également discutée. De plus, la spectrométrie de mobilité ionique (IMS) est présentée comme une méthode alternative pour la détection des isomères qui ne peuvent pas être différenciés par les techniques précédemment mentionnées.

La troisième partie de ce chapitre est axée sur les **méthodes et les expériences de RMN** qui ont été utilisées. Tout d'abord, les paramètres expérimentaux de la RMN sont discutés, notamment les paramètres d'acquisition tels que le type d'appareil de RMN utilisé, la préparation des échantillons et le type d'acquisition spectrale. Ensuite, l'accent est mis sur le traitement des données utilisé. Cela comprend une description détaillée du logiciel MixONat, ainsi que des bases de données utilisées pour l'annotation des NPs.

2.1. Plant material and extractions

2.1.1. Plant collection

Garcinia parvifolia bark samples were collected in Malaysia at four different locations: **KL5670**, Hutan Simpanan Meranto Gua Musang, Kelantan (date of collection: 13.05.2009). **KL5248**: Hutan Simpanan Sungai Temau Luala Lipis, Pahang (date of collection: 25.05.2006). **KL5259**: Km 11 Sanggang, Kuala Kangsar, Perak (date of collection: 22.06.2006). **KL5073**: 1.2 km kg. Toh Kah, Jerangau, Terengganu (date of collection: 20.08.2004). Plant material was verified by the botanist Mr. Teo Leong Eng and voucher specimens are kept at the herbarium of the Department of Chemistry, University of Malaya, Kuala Lumpur, Malaysia.

2.1.2. Extraction procedure

For initial experiments, 10 g of grinded *G. parvifolia* samples were mixed with 100 mL dichloromethane (DCM) and extracted at room temperature for 30 min using an ultrasound bath (Elma, Singen, Germany). The instrument operated at 280 W and at a frequency of 37 kHz. After filtration and solvent evaporation, the following dry weights were obtained: **KL5670**: 460 mg, **KL5248**: 470 mg; **KL5259**: 444 mg, and **KL5073**: 578 mg.

In order to facilitate additional sample processing, **KL5670** (80 g) was re-extracted using an up-scaled extraction protocol, which yielded 4.48 g of dry extract.

2.2. Reference compounds

Reference compounds amentoflavone (**1**) and betulinic acid (**2**) were purchased from Extrasynthese (Genay, France), while ferulic acid (**3**) was bought from Sigma Aldrich (St. Louis, USA), and 3-hydroxyxanthone (**4**) was obtained from Thermo Fisher (Kandel, Germany). Stock solutions of all compounds were prepared in acetone at a concentration of 10.0 mg/mL. Working solutions were prepared in molar ratios outlined in Table 1. A detailed experimental protocol is provided in section 3. Except for acetone (analytical grade), all solvents used for stock solutions, plant extractions, flash chromatography and thin layer chromatography (TLC), were of HPLC grade and purchased from Carlo Erba reagents, (Val-de-Reuil, France).

Table 1. Standard mixtures of reference compounds and their anti-AGEs activities. Relative mixing ratios are expressed in mM. Inhibition of AGEs formation is expressed as IC_{50} ± (standard deviation). Aminoguanidine (AG) was used as positive control (IC_{50}: 139 ± 21 µg/mL or 1900 ± 300 µM).

Standard mixtures	Compounds				AGE inhibition IC_{50} (µg/mL)
	(1)	(2)	(3)	(4)	
M5	2	3	6	6	26.3 ± 1.5
M2	12	3	6	6	28.3 ± 3.2
M6	7	7	7	7	34.7 ± 3.2
M1	5	5	13	12	38.0 ± 5.0
M4	2	3	32	6	58.0 ± 4.6
M3	2	14	6	6	73.0 ± 5.6
AM					139.1 ± 20.6

2.3. Samples fractionation

2.3.1. Flash chromatography of extracts KL5670 and KL5073

Sample KL5670 (4.20 g) was fractionated on a CombiFlash Rf flash chromatograph (Serlabo Technologies, Entraigues, France) using a Puriflash 50SIHC-F0220 (220 g) silica gel column. The mobile phase consisted of petroleum ether (solvent A), an equivolumetric mixture of ethyl acetate and chloroform (solvent B), and methanol (solvent C). The flow rate was set to 50 mL/min and the following gradient was applied: 0 min, (A/B/C - 100/0/0); 0–30 min, (A/B/C - 70/30/0); 30–60 min, (A/B/C - 70/30/0); 60–65 min, (A/B/C - 65/35/0); 65–78 min, (A/B/C - 65/35/0); 78–80 min, (A/B/C - 61/39/0); 90–110 min, (A/B/C - 61/39/ 0); 110–130 min, (A/B/C - 40/60/0); 130–140 min, (A/B/C - 0/100/0); 140–170 min, (A/B/C - 0/100/0); 170–190 min, (A/B/C - 0/90/10); 190–200 min, (A/B/C - 0/80/20); 200–230 min, (A/B/C - 0/80/20). In order to reasonably reduce the number of samples for consecutive experiments, fractions were roughly combined according to their analytical TLC profile (data not shown). All TLC experiments were conducted on analytical silica gel 60 TLC plates (Merck, Darmstadt, Germany) comprising a fluorescence indicator at λ: 254 nm. Mobile phases consisted of mixtures of DCM/MeOH and were adapted according to sample requirements. Overall, 14 recombined fractions were obtained (See section 3.2).

Sample KL5073 (570 mg) was fractionated on a CombiFlash Rf flash chromatograph (Serlabo Technologies, Entraigues, France) using a Puriflash 50SIHC-JP-F0025 (25 g) silica gel column. The mobile phase consisted of petroleum ether (solvent A), an equivolumetric mixture of ethyl acetate and chloroform (solvent B), and methanol (solvent C). All solvents were purchased from Carlo Erba reagents (Val-de-Reuil, France). The flow rate was set to 20 mL/min and the following gradient was applied: 0 min, (A/B/C - 100/0/0); 0-60 min, (A/B/C - 0/100/0); 60-61 min, (A/B/C - 0/80/20); 61-68 min, (A/B/C - 0/80/20). Fractions were collected every 20 mL and analyzed by TLC (Merck, Darmstadt, Germany) (data not shown). Overall, 15 recombined fractions were obtained (section 5.2).

2.3.2. Experimental preparative HPLC conditions for isolation of selected NPs

Fractions from KL5670: Fraction 3 (40 mg), Fraction 4 (150 mg), Fraction 6 (135 mg) Fraction 9 (50 mg), Fraction 10 (270 mg) and Fraction 10E (8 mg) were dissolved in MeOH (HPLC grade, VWR chemicals, Rosny-sous-Bois, France) before being processed on a Nexera Prep (Shimadzu, Kyoto, Japan) preparative HPLC system using a Pursuit XRs 5 C-18 column (250 x 21.2 mm, 5 µm, Agilent, Santa Clara, California, USA) except for Fraction 4 which was processed using a Hypersil Gold PFP column (150x20 mm, 5 µm, Thermo Scientific, San Jose, USA). Mobile phase consisted of HPLC grade water (Solvent A) and acetonitrile (Solvent B, HPLC grade, VWR chemicals, Rosny-sous-Bois, France) and the flow rate was set to 21.2 mL/min. The injection volume was 2 mL.

Solvent gradient for **F3** was as follows: 0 min, 80% B; 20min, 100% B; 25 min, 100% B. Solvent gradient for **F4** was set as follows: 0 min, 45% B; 18min, 48% B; 20 min, 100% B; 25 min, 100% B. Solvent gradient for **F6** was set as follows: 0 min, 45% B; 25min, 50% B; 30 min, 70% B; 35 min, 100% B; 40 min, 100% B. Solvent gradient for **F9** was set as follows: 0 min, 50% B; 20min, 70% B; 30 min, 100% B; 40 min, 100% B. Solvent

gradient for **F10** was set as follows: 0 min, 60% B; 15min, 65% B; 20 min, 70% B; 30 min, 100% B; 35 min, 100% B; 37 min, 60% B. Solvent gradient for **F10E** was set as follows: 0 min, 60% B; 15min, 70% B; 16 min, 100% B; 20 min, 100% B.

Chromatograms were recorded at 210 and 280 nm for **F3**, 254 and 280 nm for **F4** and **F10E** and 310 and 280 nm for **F9**. For **F6** and **F10**, chromatograms were recorded at 254 and 280 nm. Overall, 3 repetitive injections (RI) of **F3** (50.4 mg, 94.3 mg, and 120.0 mg), 10 RI of **F4** (15.0 mg each), 2 RI (56.4 and 78.6 mg) of **F6**, 3 RI (50.4, 94.3 and 120.0 mg) of **F10**, 2 RI (23.0 mg and 15.0 mg) of **F9** and 1 injection of **F10E** were conducted.

Isocowanin (11.5 mg) was isolated from F3, **TPTQ derivates M1-M4** (1: 4.9 mg, 2: 5.8 mg, 3 : 6.1 mg, 4: 5.2 mg) were isolated from F4, **rubraxanthone** (57.0 mg) was isolated from F6 (rt, 21.0 min), **parvixanthone H** (4.2 mg) was isolated from F9, **isocowanol** (46.49 mg) and **parvixanthone G** (15.43 mg) were isolated from F10 (rt, 26.7 min and 12.5 min, respectively), **garcinone D** (1.2 mg) was isolated from F10-E.

Fraction **10B** (9.6 mg) was dissolved in MeOH (HPLC grade, VWR chemicals, Rosny-sous-Bois, France) before being processed on an Agilent 1200 semi prep-HPLC (Agilent technologies, Les Ulis, France) using a Luna C-18 column (250x4.60 mm, 5 µm, Phenomenex, Torrance, USA). Mobile phase consisted of HPLC grade water (Solvent A) and MeOH (Solvent C, HPLC grade, VWR chemicals, Rosny-sous-Bois, France). Solvent gradient was as follows: 0 min, 75% C; 0-40min, 80% C. The flow rate was 3.70 mL/min and chromatograms were recorded at 210, 240, 254, 280 and 310 nm. **Parvixanthone C** (1.8 mg) and **butyraxanthone D** (1.3 mg) were isolated.

9-hydroxycalabaxanthone was isolated from previous work by A. Bruguière *et al.* (Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Suor-Cherer, *et al.*, 2020). F10F was pure **cratoxylone**.

Spectrum of each isolated compound are displayed in section 4.1.

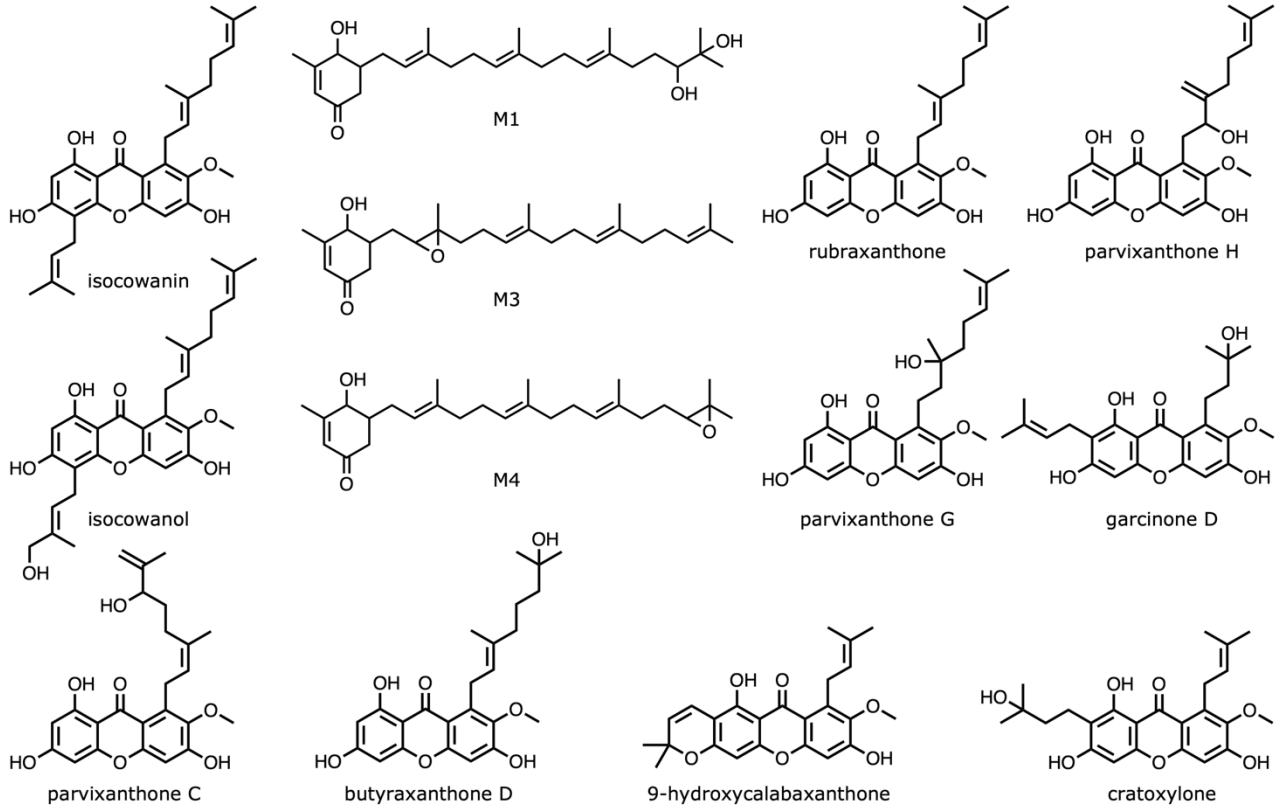


Figure 1. Isolated compounds from *G. parvifolia* bark extract (KL5670)

2.4. Advanced glycation end products inhibition assays

The anti-AGEs assay was developed at the SONAS laboratory by S. Derbré et al. (Derbré et al., 2010) and serves as a screening tool to evaluate the inhibition of AGEs, which are implicated in the pathogenesis of several diseases, including diabetes and its vascular complications, neurological disorders and more generally in tissue aging. The assay was selected due to its simplicity and high throughput capacity. The assays measure the formation of fluorescent AGEs, from albumin and D-ribose in an aqueous milieu (a Maillard reaction). Extracts, fractions, or isolated compounds were tested.

All the experiments were conducted according to a previously described protocol (Derbré et al., 2010; Séro et al., 2013). Stock solutions (SS) of mixtures or extracts were prepared in DMSO at a concentration of 10 mg/mL, while single compounds were dissolved at a concentration of 30 mM. These SS were then diluted with 50 mM phosphate buffer (pH: 7.4) yielding working solutions (WS) at a concentration range of 10 to 10-2 mg/mL for standard mixtures and bark extracts. The WS for single compounds were prepared at a concentration range of 30-3.10^{-2} mM. Next 10 µL of each WS were deposited in 96 black well bottom plates (Fisher Scientific, Illkirch, France) and mixed with 90 µL of a solution containing bovine serum albumine (BSA) (11 mg/L, Sigma Aldrich, St Quentin Fallavier, France), D-ribose (0.25 M, Acros Organics, Geel, Belgium), and phosphate buffer (50 mM, NaN3 0.02%, pH 7.4). Plates were then incubated for 24h at 37°C, before being fluorometrically analyzed (λexc: 335 nm, λem: 385 nm) on an Infinite M200 plate reader (Tecan, Lyon, France). The formation of AGEs was calculated according to the following formula (FI = fluorescence intensity):

Equation 1. AGEs formation formula

$$\frac{[FI\ (BSA + ribose + sample) - FI\ (BSA + sample)]\ \mathrm{x}\ 100}{FI(BSA + ribose) - FI(BSA)}$$

Aminoguanidine (Sigma Aldrich, St Quentin Fallavier, France) was used as positive, a mixture of BSA + ribose as negative control. A solution of BSA served as blank. Results were expressed as IC_{50} values in µg/mL (mixtures and extracts) or µM (single compounds).

For IC_{50} calculations dose-effect curves were best fitted with a sigmoidal dose-response equation using Sigma Plot 12.0 software.

2.5. General information about UPLC-ESI-QTOF

2.5.1. The principle of liquid chromatography

Liquid chromatography is a separation technique that allows the separation of different components of a complex mixture in solution (samples). The metabolites analyzed are called analytes. Typically, samples are dissolved in organic solvents such as methanol or acetonitrile, and mixtures of these solvents with water. The solvent must be miscible with the mobile phase, which is typically a solvent or solvent mixture consisting of water, methanol and/or acetonitrile. The mobile phase passes the dissolved sample through a column filled with a stationary phase, usually a resin or modified silica (*e.g.*, C-18). The analytes interact differently with the stationary and mobile phases, resulting in their separation according to their different affinities towards the two phases.

The SONAS laboratory is equipped with an AcquityUPLC system (Waters, Manchester, UK) shown in Figure 2. Compared to HPLC, UPLC columns use smaller particle sizes (typically less than 2 microns), which allows faster separations and higher chromatographic resolutions. This feature results in a higher-pressure system (up to 15,000 psi) that improves efficiency and reduces analysis time.

Figure 2. Ultra-performance high-resolution chromatography system coupled to a quadrupole time-of-flight (UPLC-QTOF) from Waters Corp.

2.5.2. UPLC-MS detection and instrument settings

Various detection methods, such as UV-Vis spectrophotometry, diode array detector (DAD), evaporative light scattering detector (ELSD) or MS, can then be used to detect and quantify the separated components. The choice of detector depends on the specific requirements of the analysis, such as sensitivity, selectivity, and detection range.

The UPLC system in the SONAS laboratory is coupled to a QTOF mass spectrometer (WatersXevo®G2-S QTOF mass spectrometer). This system features an ESI source that is widely recognized for its ability to detect and analyze both polar and non-polar metabolites.

The ESI source consists of a metal capillary tube held at a high voltage (typically between 2.5 and 6.0 kV). As the solution flows through the capillary, a cone-shaped droplet, known as the Taylor cone, forms at the tip of the needle. The eluted sample is nebulized by nitrogen gas, resulting in the formation of charged droplets due to the high voltage applied to the capillary (Figure 3, 1). The droplets then enter the spray chamber and are vaporized by the countercurrent of the heated dry gas, reducing the droplet diameter, and increasing the surface charge density (Figure 3, 2). As the droplets shrink in size, their charge density increases and eventually the electric field strength within the droplets reaches a critical point where ions on the droplet surface are ejected into the gas phase (Coulombic repulsion). The desolvated ions are then sampled into the mass spectrometer for further analysis (Figure 3, 3). The mass spectrometer separates the ions based on their mass-to-charge ratio (m/z) and

measures their abundance, allowing the determination of the monoisotopic masses and structural information of the analyte (Figure 3, 4).

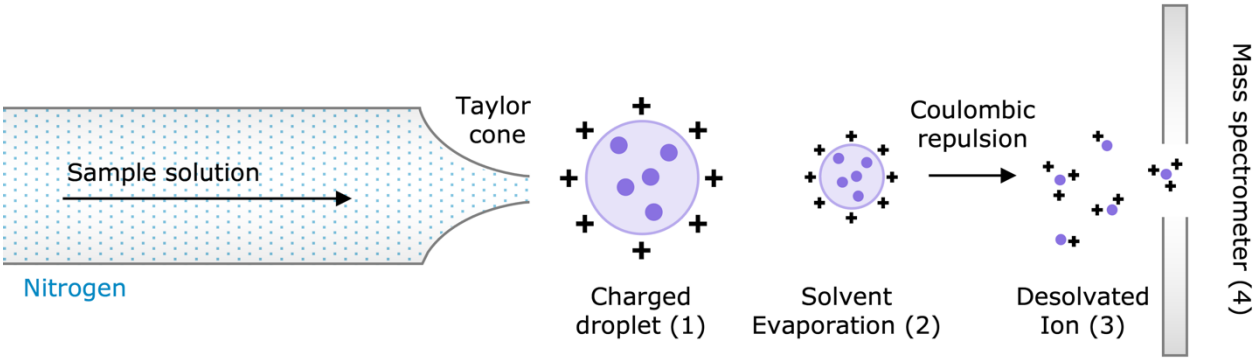


Figure 3. Electrospray ionization principle

The ions are then detected by a QTOF mass spectrometer (also referred to as HRMS in this manuscript). As the name suggests, it combines two key components: a quadrupole mass filter and a time-of-flight (TOF) analyzer. The quadrupole filter selects ions of specific m/z ratios (Figure 3, 5), while the TOF analyzer measures the time an ion requires to migrate from the source to the detector (Figure 3, 6). This "time of flight" is directly related to the ion's m/z ratio. Ions with a lower m/z ratio will reach the detector faster than those with a higher m/z ratio. The longer the distance ions need to pass to reach the detector the higher the resolution. Therefore, modern instruments use electrostatic mirrors (reflectrons) to extend this distance.

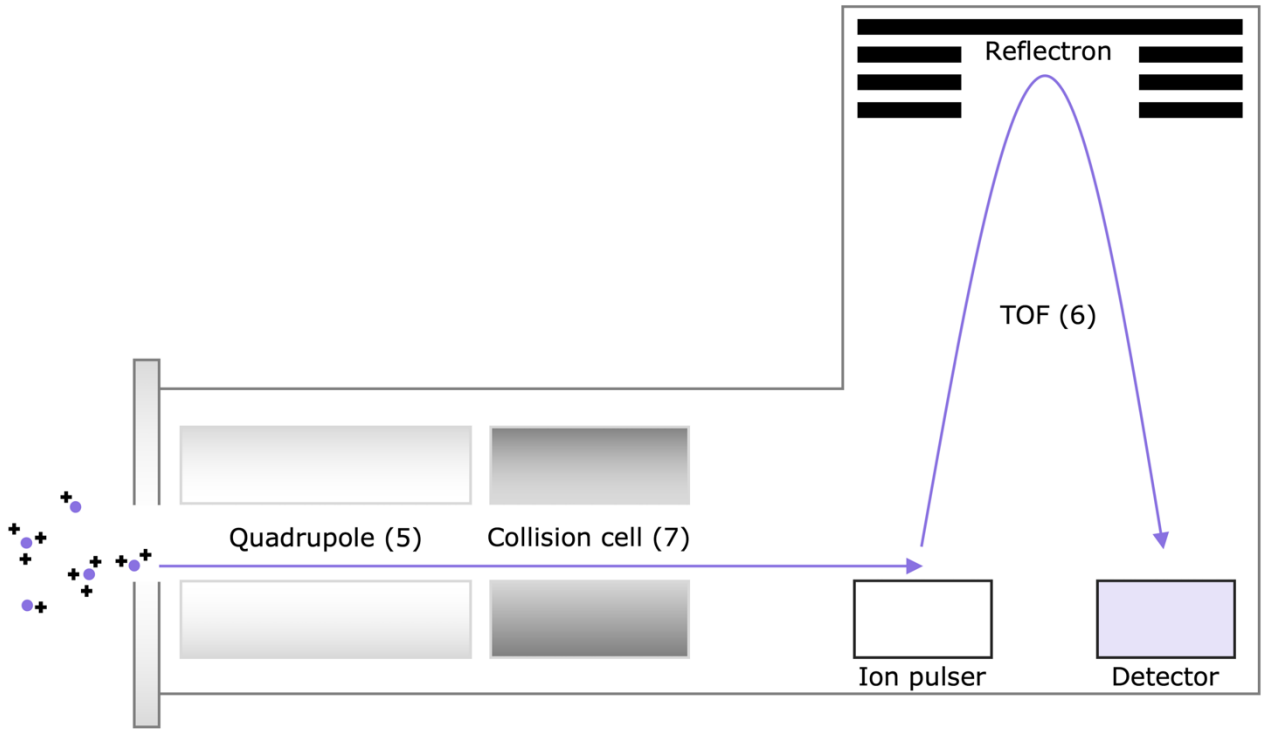


Figure 4. The schema of quadrupole time-of-flight mass spectrometer

Finally, mass spectra that provide detailed information about the masses and relative abundances of the detected ions are obtained. The QTOF setup also allows tandem mass spectrometry (MS/MS). This involves the selection of specific precursor ions using the quadrupole mass filter, followed by their fragmentation in the collision cell

(CC) (Figure 3, 7). The CC induces fragmentation of the selected precursor ions by collision-induced dissociation (CID), in which high-energy collisions with inert gas such as argon or nitrogen, cause the ions to break apart, producing fragmented ions. These fragmented ions are then accelerated into the flight tube where their time of flight is measured based on their mass-to-charge ratio. The resulting data are used to generate a mass spectrum that provides insight into the molecular composition and structure of the sample being analyzed.

2.5.3. Experimental UPLC-MS conditions

General chromatographic conditions: All experiments were conducted at 25°C on an Acquity UPLC system (Waters, Manchester, UK) using a HSS T3 C-18 (2.1 × 100 mm, 1.8 µm) UPLC column (Waters). LC-MS grade water (solvent A) and acetonitrile (solvent B, LC-MS grade, VWR chemicals), both supplemented with 0.1% of formic acid (LC-MS grade, VWR chemicals) were used as mobile phases. Chromatographic conditions for crude extracts and fractions of KL5670 are outlined below.

Chromatographic conditions for crude extracts: Samples were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.5 mg/mL and injected in triplicate (injection volume 1 µL). The mobile phase gradient was as follows: 0 min, 30% B; 0–10 min, 80% B; 10–17 min 100% B; 17–20 min, 100% B. The flow rate was 0.3 mL/min.

Chromatographic conditions for fractions of KL5670: Samples were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.02 mg/mL and injected in triplicate (injection volume 1 µL). The mobile phase gradient was as follows: 0 min, 40% B; 0–5 min, 80% B; 5–10 min 100% B; 10–12 min, 100% B. The flow was set to 0.4 mL/min.

Chromatographic conditions for fractions of KL5073: Samples were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.02 mg/mL and injected in triplicate (injection volume 1 µL). The mobile phase gradient was as follows: 0 min, 60% B; 0–5 min, 20% B; 5–10 min 100% B; 10–13 min, 100% B. The flow was set to 0.4 mL/min.

Mass spectrometry settings for all samples: Mass chromatograms and spectra were recorded on a WatersXevo®G2-S QTOF mass spectrometer (Waters) using negative electrospray ionization. Source parameters were as follows: capillary and cone were set to 0.50 and 40 kV, respectively. Desolvation gas flow was 1000 L/h at 36°C, and cone gas was set to 100 L/h. Source temperature was 120°C and data were collected in continuum mode at 0.5 scan/s. Instrument calibration was conducted using a leucine-enkephalin solution at a flow rate of 5 µL/min (calibration points: m/z 554.26202 for negative mode). All UPLC-MS/MS data were acquired in data dependent analysis (DDA) mode. The mass analyzer mass range was set to 50–1200 Da for full scan mode (scan time: 0.5/s). The MS/MS mode was automatically activated once the total ion current (TIC) exceeded 20,000 intensity/s (crude extracts) or 10,000 intensity/s (fractions of KL5670 and KL5073) but remained in MS mode when TIC was below these thresholds. CE ramp of 20–25 V for low-mass, and 65–85 V for high-mass analytes were employed.

2.5.4. ProgenesisQI, a post-acquisition data processing software

Acquired data must be processed before any interpretation can be performed. There are numerous of data processing softwares (see section 1.2). For the current work, the ProgenesisQI software (Nonlinear Dynamics) was used for post-acquisition UPLC-MS data processing and data analysis during my PhD. ProgenesisQI includes tasks such as alignment, experimental design, peak picking, deconvolution, identification, and some statistics. The following paragraphs will describe the general procedure that was applied. Specific parameters for final post-acquisition data processing will then be outlined in section 2.5.5.

- **Import data**

The first step displayed under the *Import Data* tab, allows data to be imported from Thermo (.raw), Agilent (.d), ABSciex (.wiff), Bruker (.d) instruments. Open formats such as .mzML, .mzXML and .NetCDF are also supported.

Mass spectrometry data can be acquired in profile or centroid mode. In profile mode, a peak is defined by a series of signals over multiple scans. However, this means that the file contains a lot of information and require significant storage space. It is also not suitable for further processing such as MN, which requires centroid data files. Therefore, data are usually transformed in centroid spectra to eliminate redundant information. The signals are then displayed as single m/z value with no line width (Figure 5). The advantage of centroid data is that the file size is significantly smaller because there is less information to describe a signal.

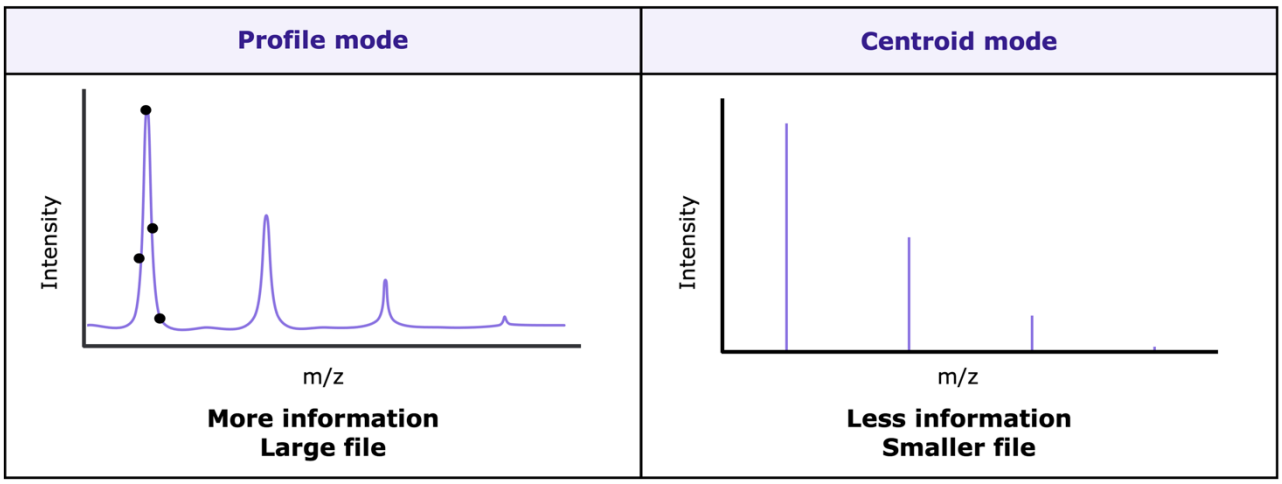


Figure 5. Differences between a profile and centroid mass spectrum

In this way, all UPLC-MS data were first acquired in profile mode for higher m/z accuracy. ProgenesisQI then transformed these data to the centroid format which was exported and used for consecutive MN applications.

- **Alignment and peak picking**

Once the profile data has been correctly imported, alignment is performed against a reference run (usually the pool run, or the raw extract). This is done by automatically or manually placing landmarks called alignment vectors. Each vector connects the position of a particular compound ion on the reference run to the position of the same ion on the run to be aligned.

The result of all aligned runs is then displayed on the *Review Alignment* tab (Figure 6). The score represents the percentage of matched peaks. For each of the small green squares (Figure 6, ion intensity map), the total matched

intensity of the reference is added to the corresponding total matched intensity of the other experiment used to obtain an aligned score (in %).

ProgenesisQI will then combine data from multiple aligned spectra from LC-MS runs on the *Experiment design set up* tab. The "Between-subject" design was used, which allows grouping replicates under the same condition. The subsequent ANOVA calculation considers that the conditions are independent and applies the statistical test that considers the means of the conditions are equal.

Once the experimental design is set up, the *Peak Picking* tab appears (Figure 6). *Peak Picking* can be performed automatically, but detected ions must be thoroughly inspected manually.

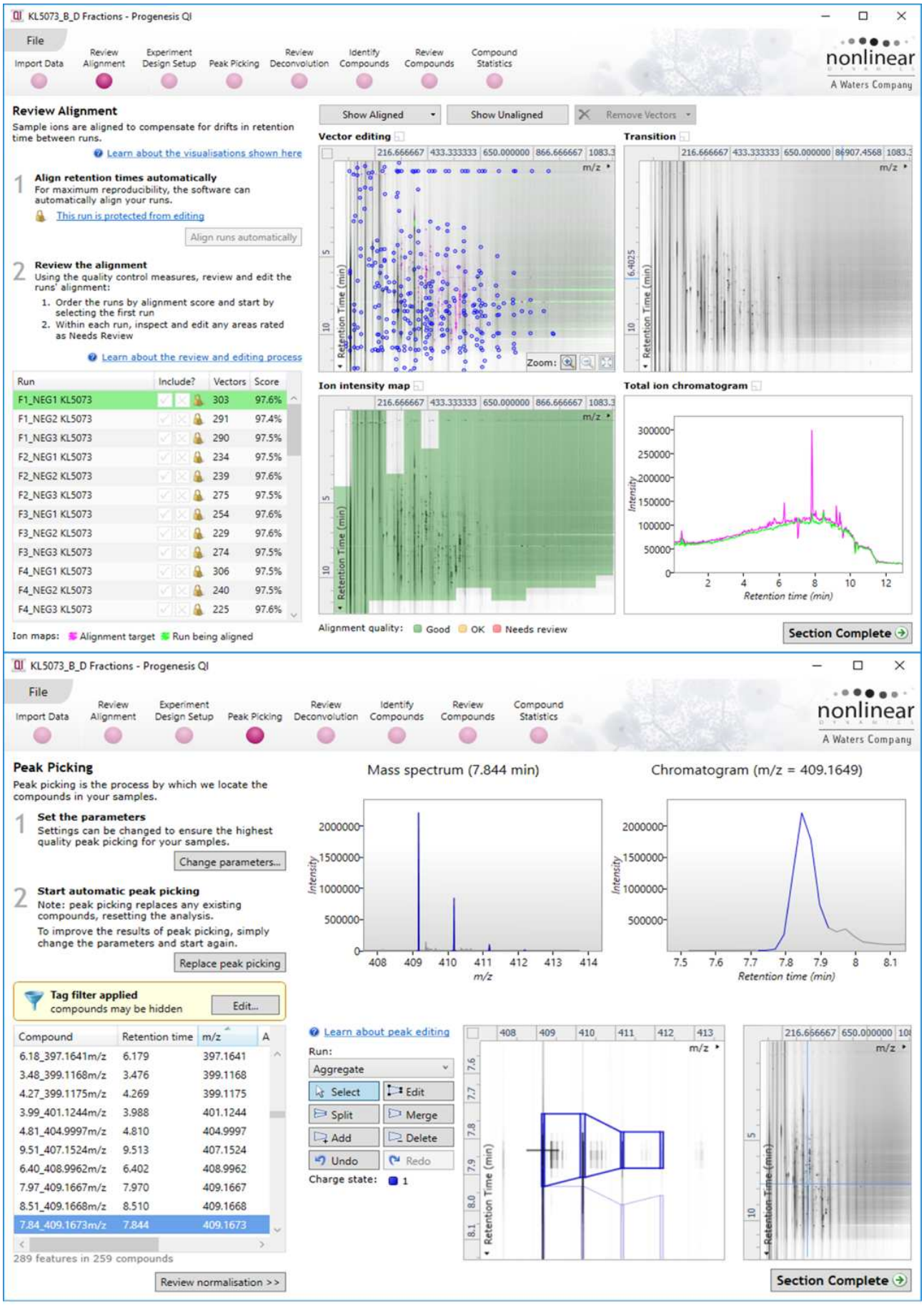


Figure 6. ProgenesisQI alignment and peak picking tabs

- **Deconvolution**

Once this is done, the *Review Deconvolution* tab (Figure 7) (*How are the compound ions grouped into compounds?*, 2023) provides possible identification of adducts by comparing the retention time and the ion mass difference between detected ions. For subsequent processing, the two ions are grouped together as a single compound.

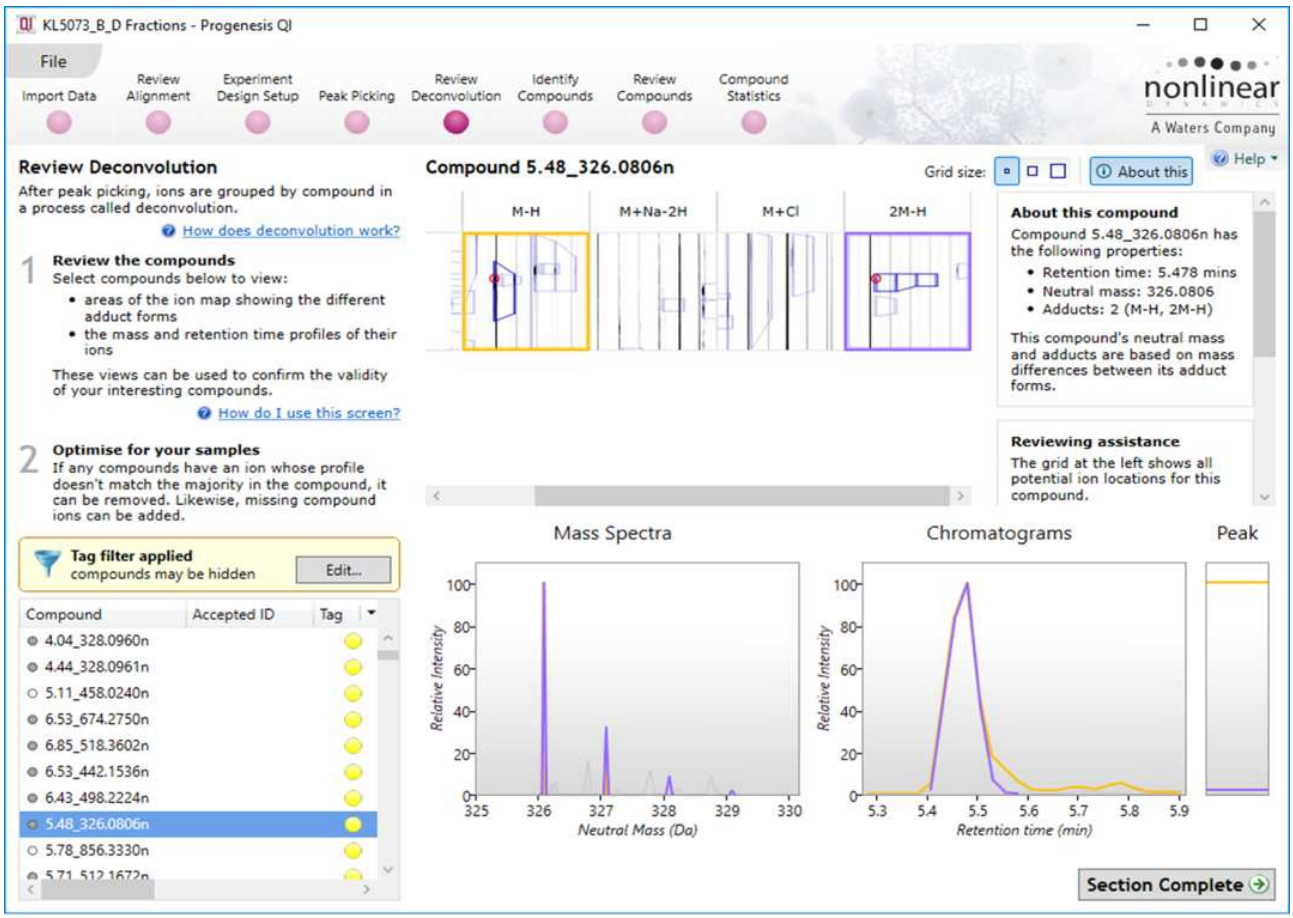


Figure 7. ProgenesisQI deconvolution tab

- **Identification of compounds**

A list of detected compounds is then displayed in the next tab *"Identify Compounds"* (Figure 8). Multiple identification tools are available:

- **Elemental composition** (*Elemental composition plug-in - Progenesis QI*, 2023) calculates the theoretical molecular formulas that matches the measured mass and isotopic distribution of the experimental compound.
- **Progenesis MetaScope** (MetaScope search plug-in - Progenesis QI, no date), which allows experimental data (neutral mass, retention time, collisional cross-sectional area (CCS) and/or MS/MS fragmentation data) to be compared against a DB. The setup displays two different methods to help identifying compounds from MS/MS fragmentation data. First, the experimental MS/MS fragmentation can be compared with the theoretical fragmentation generated by an algorithm called MetFrag (Wolf *et al.*, 2010; Ruttkies *et al.*, 2016a; Ruttkies, Neumann and Posch, 2019). For each structure in an .sdf file, this

algorithm simulates the bond cleavages in the structure and generates their MS/MS patterns. Secondly, it is possible to save fragmentation patterns of previously solved structures and compare these results with new experimental data. The software supports .sdf, .csv, .xls or .xlsx databases.

- **ChemSpider Search** (*ChemSpider Search plug-in - Progenesis QI*, 2023) is a web-based chemical structure DB with access to over 32 million structures from hundreds of data sources. This method compares the experimental mass of compounds of interest to selected DBs within a selected error range in ppm.
- **LipidBlast Search** (*LipidBlast Search plug-in - Progenesis QI*, 2023) is a computer generated positive and negative ionization MS/MS DB containing over 212,000 spectra from over 119,000 lipids (Kind *et al.*, 2013). The DB is available in .msp format for both neutral mass and MS/MS data searching.
- **National Institute of Standards and Technology (NIST) MS/MS Search** (*NIST MS/MS Search plug-in - Progenesis QI*, 2023), which provides MS/MS data for small molecules and peptides. The method uses the same search and scoring as Progenesis MetaScope.

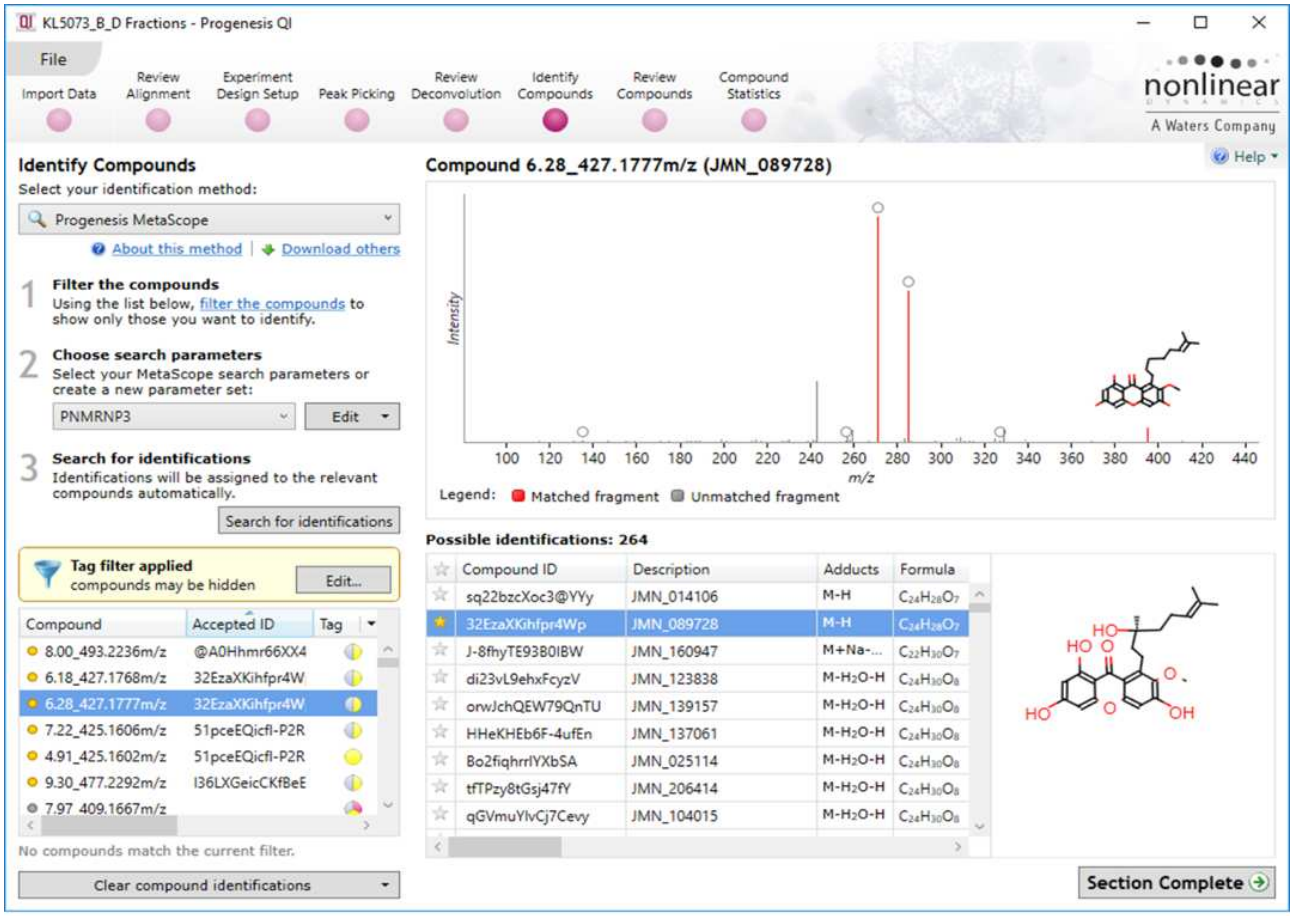


Figure 8. ProgenesisQI identification tab

The annotation of compounds in this manuscript is based on the dereplication strategy using Progenesis MetaScope, along with the PNMRNP DB3 (Lianza *et al.*, 2021).

- **Review compounds**

The following tab allows to *Review Compounds* (Figure 9). In addition, tags can be created to filter the data (*e.g.*, only MS/MS data tags, only significant ions with low p-values, etc.).

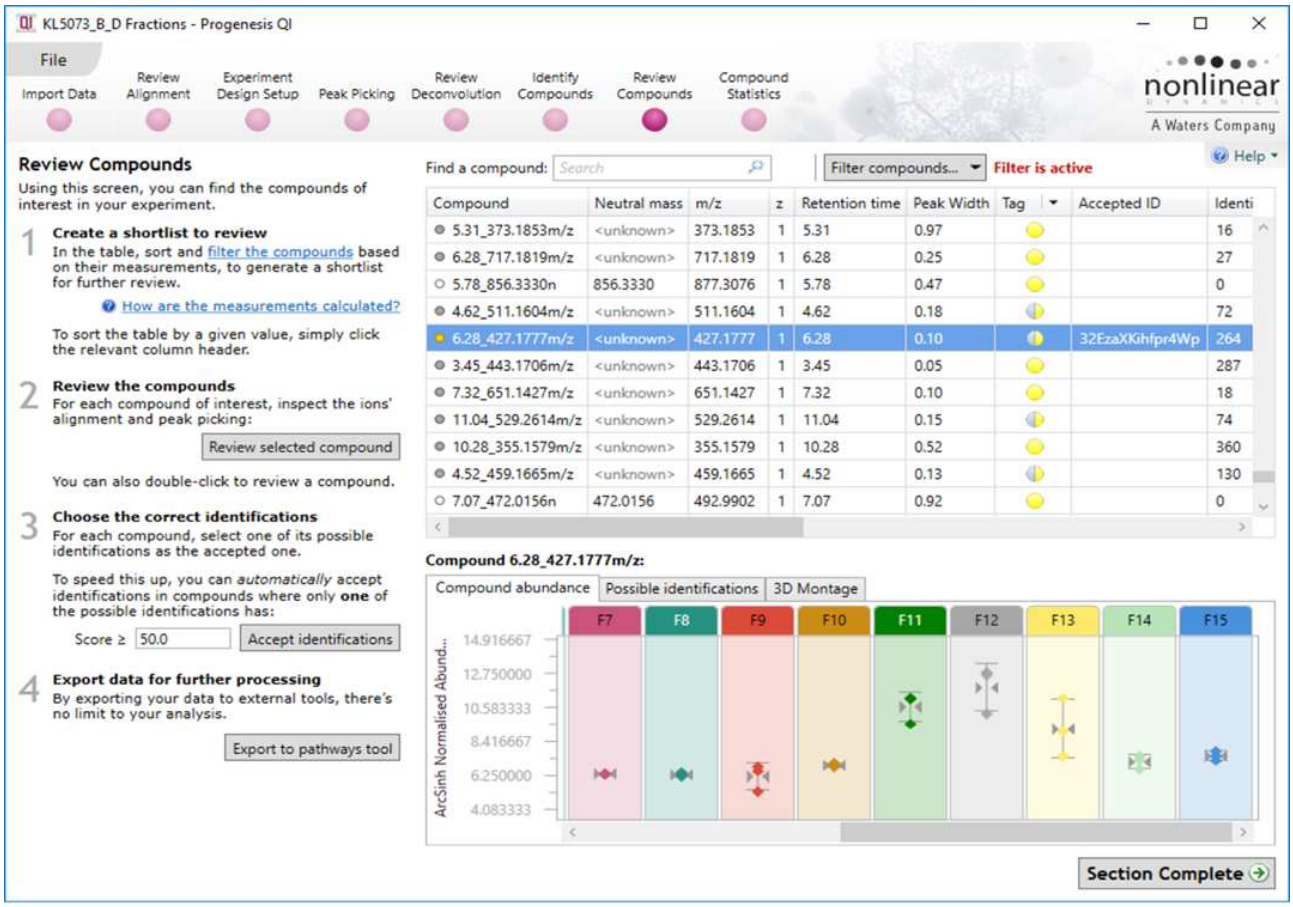


Figure 9. ProgenesisQI review compound tab

- **Compound statistics**

Finally, the list of quantified and identified compounds can enter multivariate statistical analyses, including principal component analysis, correlation analysis, false discovery rate q-values, and an overview how adduct abundance varies between runs. Most importantly, the pre-processed data can be easily exported to a separate statistical tool, the EZinfo software. Once processed, the EZinfo list of compounds of interest can be retransferred to Progenesis QI as tagged groups.

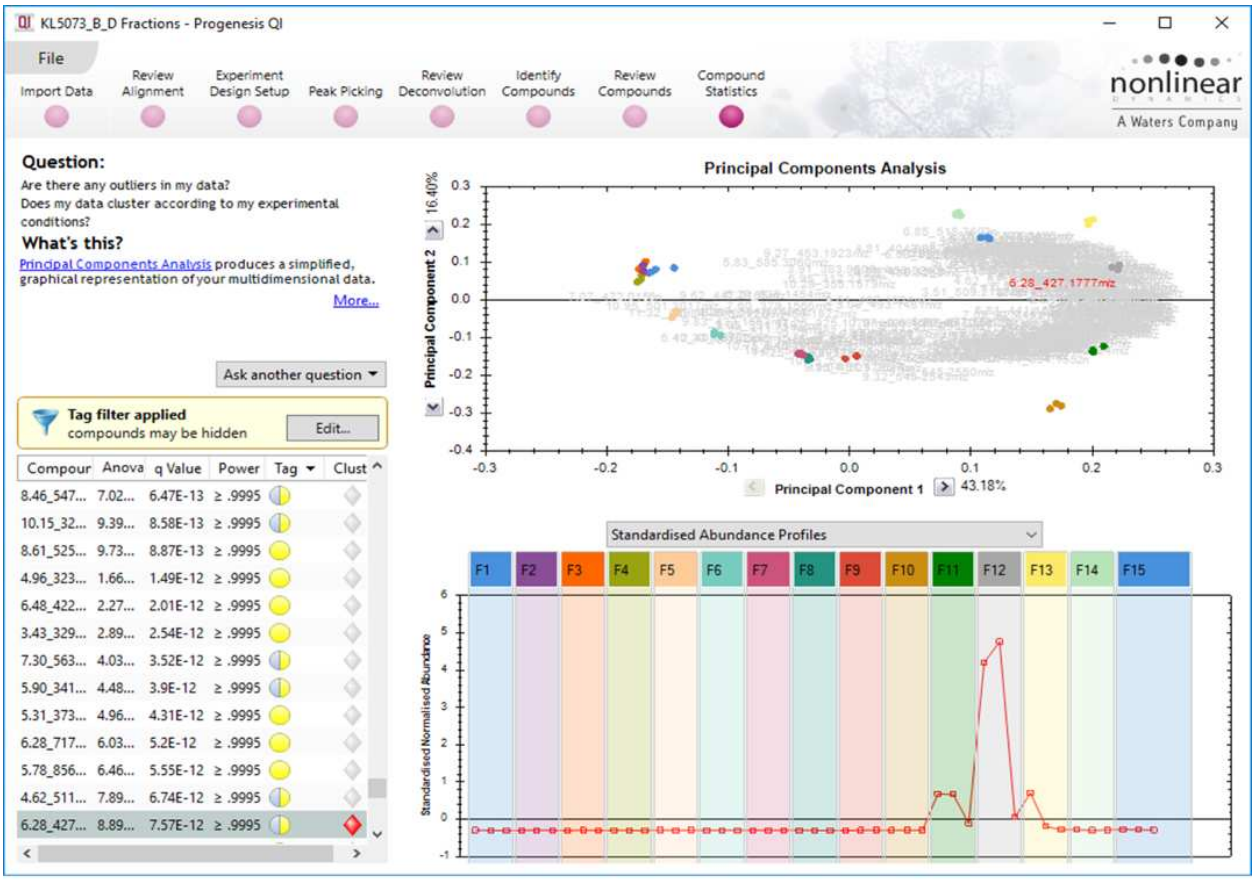


Figure 10. ProgenesisQI compound statistic tab

2.5.5. Applied post acquisition parameters

2.5.5.1. ProgenesisQI post acquisition parameters of KL5670 and KL5073 extract

Automated data alignment and peak-picking of KL5670 and KL5073 analyses were conducted separately by the ProgenesisQI software (Waters, Newcastle, UK). Results were thoroughly inspected and manually corrected if necessary (*e.g.*, regrouping of some ions with identic retention times, peak shapes and/or isotopic distribution). Data were then filtered with "tag" to keep only significative compounds with an intensity over 4000, MS/MS data and an ANOVA p-value ≤ 0.05. Data were then processed by the Progenesis MetaScope dereplication tool. The latter creates theoretical fragmentation patterns from imported structures and compares them with experimental MS/MS data. Putative structural assignments were made using the "Predicted Carbon-13 NMR Data of Natural Products" (PNMRNP) database, which comprises 211,280 NPs. All structures were exported as .sdf, a format that is compatible with ProgenesisQI. Both precursor and fragmentation tolerance were set to 5 ppm. Based on these parameters ProgenesisQI created a list of possible structures with their respective probability scores (Appendix 1). Processed MS data were directly exported and consecutively processed by EZinfo 3.0 for Waters (Umetrics, Umea, Sweden).

2.5.5.2. Statistical parameters applied for KL5670 data

Following exportation, PLS regression analyses based on spectrometric information (m/z values, retention times and peak areas) as well as activity data (AGEs inhibition) were conducted on EZinfo 3. The Pareto scale setting was applied.

2.5.5.3. Molecular networking parameters

First, a feature-based molecular network (FBMN) was created by exporting "fragmentation" data and the "compound measurement" files from ProgenesisQI (.msp and .csv files respectively) of KL5670 to the GNPS. The parameters for the FBMN were set as follow: a precursor ion mass and fragment ion mass tolerance set at 0.02 Da, a cosine score equal to or greater than 0.7, topK was set to 5, minimum matched fragment ions were set to 4 and maximum connected component size was 0. Concerning library search options, the minimum number of matched peaks was 4, and the score threshold set to 0.6. The job record is available online:

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2ae86a52caf04ebcbf397df1bd6946cb

The graphical visualization and treatment of molecular network was performed using the Cytoscape 3.9.0 software (Lopes *et al.*, 2010).

In a second step, a FBMN based on KL5670, KL5073 and selected Pierre Fabre *Garcinia* extracts was generated by merging the respective "fragmentation" data and "compound measurement" files from the ProgenesisQI post-treatment. The parameters for the FBMN were set as follow: a precursor ion mass and fragment ion mass tolerance set at 0.02 Da, a cosine score equal to or greater than 0.7, topK was set to 6, minimum matched fragment ions were set to 6 and maximum connected component size was 0. Concerning library search options, the minimum number of matched peaks was 4, and the score threshold set to 0.6. The job record is available online:

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2cdf823e77df4a9393f123a12128dde9

The graphical visualization and treatment of molecular network was performed using the Cytoscape 3.10.0 software (Lopes *et al.*, 2010).

2.6. General information about LDI-MS

Matrix free laser desorption ionization (LDI) mass spectrometry (Figure 11) is a powerful and method for the analysis of small UV light absorbing molecules. It provides a direct and efficient way for the quick chemical characterization of complex mixtures of NPs. Many NPs show close structural similarities to compounds that are used as matrices in matrix-assisted laser desorption ionization (MALDI)-MS. A MALDI matrix takes up most of the laser energy and transfers part of it towards the analytes. MALDI is generally used on large molecules such as proteins or compounds that do not ionize by matrix-free LDI. However, the extensive formation of matrix ions and the potential risk of signal superposition between matrix and analyte ions limits its use for the analysis of small molecules. In addition, the presence of matrix ions creates artificial similarities in the overall ionization profiles of samples, thus notably impairing correct statistical interpretation. In this respect matrix free LDI is always the better option if sufficient ionization of target compounds can be achieved.

During my PhD general LDI-MS experiments were conducted at SONAS laboratory, on a Spiral-TOF mass spectrometer (JEOL Ltd., Tokyo, Japan) (see section 2.6.1.2). Experiments on LDI-ion mobility MS (LDI-IM-MS) were performed on a timsTOF fleX MALDI-2 (Bruker Daltonics GmbH & Co. KG, Bremen, Germany) (see section 2.7.1.)

Figure 11. Laser desorption ionization coupled to spiral time-of-flight mass spectrometer from JEOL

2.6.1. Sample preparation and MS settings for LDI-MS experiments

2.6.1.1. Sample preparation

Standard mixtures of reference compounds initially prepared in acetone (section 2.2) were dried and redissolved in DCM. Crude *G. parvifolia* extracts as well as fractions of KL5670 were dissolved in a mixture of DCM/MeOH (60/40) at a concentration of 10.0 mg/mL. All samples were deposited in quadruplicates (0.5 µL/each) on the sample carrier.

2.6.1.2. Acquisition parameters

Experiments were conducted on a Spiral-TOF mass spectrometer (JEOL Ltd., Tokyo, Japan), equipped with a pulsed laser operating at 349 nm. Spectra were recorded within a mass range of 50–1200 m/z. Preceding experiments have shown that best results were obtained in negative ionization mode, so all spectra were recorded in high-resolution spiral negative mode. Laser frequency (50–100 Hz), detector gain (49–67%) and laser power (55–80%) were individually adapted to samples requirements.

Acquired LDI-MS spectra provide a sample specific profile of detected ions and their abundances. These data underwent a series of post-acquisition processing steps before being statistically analyzed.

2.6.2. Post processing steps of LDI-MS data

The first step involves MS Tornado (JEOL, JAPAN) which is the operating software of the instrument. It was used for initial spectra visualization, calibration, and data export. However, it does not allow efficient data processing or statistical analysis. Therefore, all spectra were exported as text files (.txt) in order to facilitate further processing by mMass (Strohalm *et al.*, 2008, 2010; Niedermeyer and Strohalm, 2012).

According to G. Ge *et al.* (Ge and Wong, 2008), raw mass spectrometry data can be separated into three components: the baseline value, noise, and real signals. Prior to statistical analysis, it is imperative to eliminate these data to reduce variance between experiments. Otherwise, statistical results can be biased (Ernst et al., 2014).

Therefore, post-acquisition processing steps typically include: noise filtering, normalization, peak picking, deisotoping and alignment (Castillo *et al.*, 2011). Several open source and commercial software may be used for post-acquisition processing. However, the Spiral-TOF mass spectrometer (JEOL Ltd.) can export raw spectra solely as .txt or .tas formats. These file types are incompatible with most software (Table 2), except for mMass (Strohalm et al., 2008) and MassUp (López-Fernández et al., 2015). MSConvert a software from the Proteowizard project, which converts most commercial files to open access files, does not support .tas files either.

Initial post acquisition processing steps were all performed on mMass as outlined below.

- **Noise filtering and peak picking (mMass)**

The first step was to import the raw profile mass spectra in .txt format into mMass (Strohalm *et al.*, 2008, 2010; Niedermeyer and Strohalm, 2012). This open-source software is based on Python library and provides the basic tools for data processing. Thus, all spectra were processed in the same way: cropping, smoothing, baseline correction, calibration (using at least two known masses as reference signals at both ends of the spectrum), peak picking, and manual deisotoping.

These steps significantly reduced the size of data sets, which is crucial for consecutive processing steps using other software (*i.e.,* MassUP) as large files are generally not supported.

- **Normalization**

Normalization is another important step, which reduces signal variations within replicates. Two main strategies are commonly used.

The first strategy uses either an internal standard or the base peak (largest peak of a spectrum) for normalization. As the addition of an internal standard would require several extra steps of sample preparation, the approach was not considered. **The second strategy** uses scaling factors for each sample based on the full data set. For example, MZmine (Pluskal *et al.*, 2010) divides individual peak height or area by the total sum of heights or areas of all selected peaks (Castillo *et al.*, 2011).

For finding the best normalization strategy, the ten most intense ions of four LDI-MS spectra of KL5259 were studied. As shown in Figure 12, A, an average coefficient of variation (CV) of 16.4% was observed without normalization.

- When applying **the first strategy**, signals are expressed in percentage of the base peak (100 %). This method resulted in a lower average CV of 7.1 % (Figure 12, B). However, the base peak will always have a relative intensity of 100% with no variation, which is problematic for successive statistical analyses.
- When applying the **second strategy** relative ion intensities were calculated according to the equation below:

Equation 2. Relative intensity formula for LDI-MS data

$$relative\ intensity = \frac{ion\ intensity}{\sum ion\ intensity} \times 100$$

Next to exhibiting a lower average CV of 6.4% (Figure 12, C), this normalization method yielded relative base peak intensities that were far below 100% together with expected natural variations between replicates. Therefore, all LDI-MS spectra were systematically normalized by the second strategy.

In practice, processed mMass spectra were exported to Excel in centroid format and then normalized. Results were saved in .txt and .cvs formats to facilitate consecutive peak alignment.

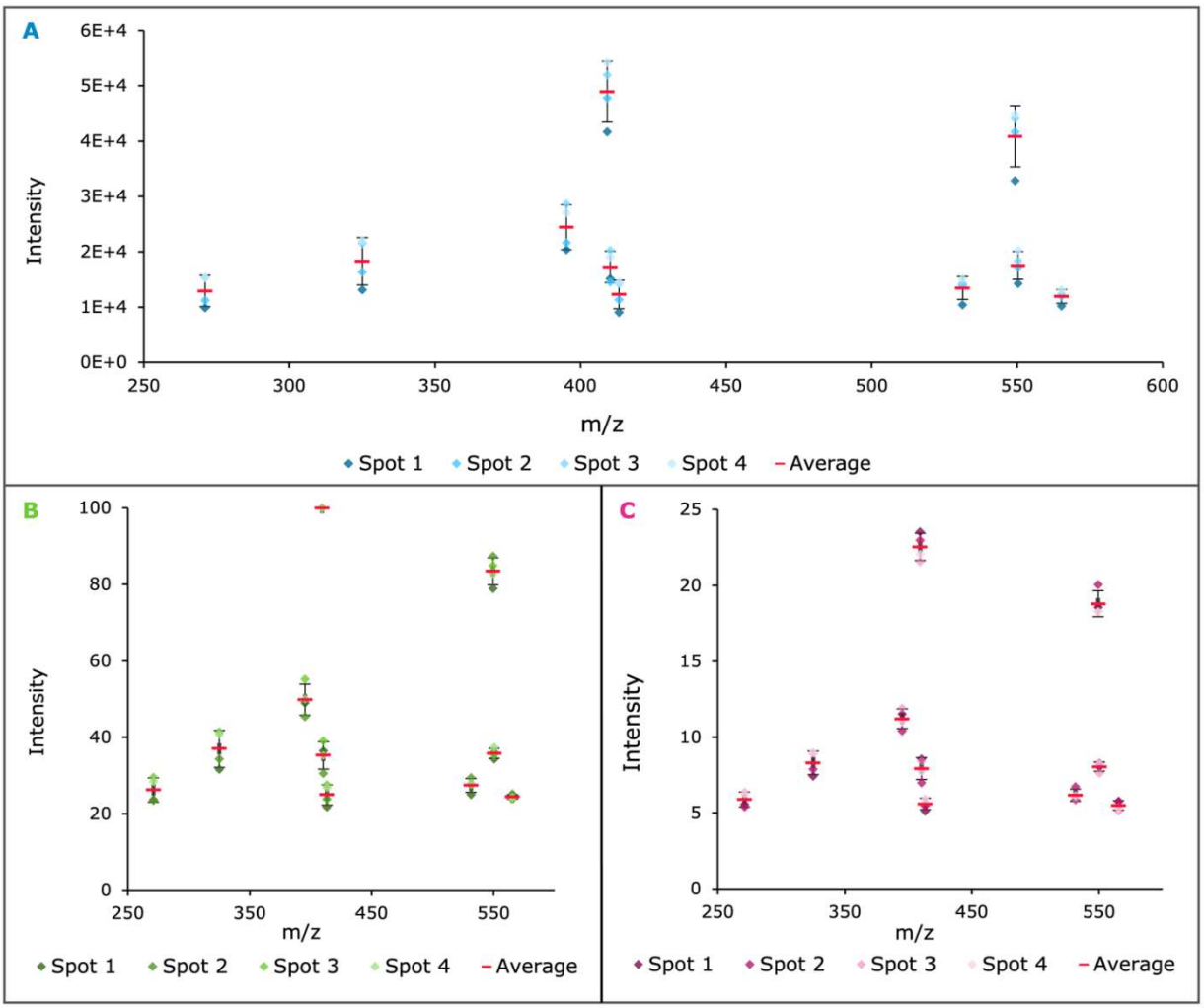


Figure 12. LDI-MS signals intensities of the ten most intense ions of KL5259 before and after normalization: (**A**) Initial signal intensities, (**B**) relative signal intensities (first strategy), (**C**) relative signal intensities (second strategy). Sample were analyzed in quadruplets and error bars correspond to standard deviations.

Table 2. Open source and commercial post acquisition processing softwares (non-exhaustive list)

Software	Open format	Vendors format	Ref.
OpenMS	.mzML, .featureXML, .consensusXML, .idXML		(Sturm *et al.*, 2008)
mMass	.mzML, .mzXML, .mzData, .mgf, .ASCII		(Strohalm *et al.*, 2008)
Mass-Up	.mzML, .mzXML, .csv		(López-Fernández *et al.*, 2015)
XCMS	.mzXML, .netCDF, .mzData		(Smith *et al.*, 2006)
MZmine	.mzML, .mzXML, .imzML, .netCDF, .aird	.raw (Thermo Scientific), .d and .tdf/tsf (Bruker Daltonics)	(Pluskal *et al.*, 2010)
MS-DIAL	.mzML, .CDF	.raw (Thermo and Waters)*, .d (Agilent and Bruker Daltonics)*, .wiff (SCIEX)*, .QGD/.LCD (Shimadzu)*	(Tsugawa *et al.*, 2015)
ProgenesisQI (Nonlinear Dynamics)	.mzML, .mzXML	.raw (Thermo and Waters), .d (Agilent and Bruker Daltonics), .wiff (SCIEX)	(*Nonlinear Dynamics Progenesis software*, no date)

*Requires conversion to the .ABF format by Reifics File Converter

- **Spectral alignment**

Following these steps, the spectra need to be aligned so that the features associated to a particular compound can be compared across all experiments. Two workflows are possible (Figure 13):

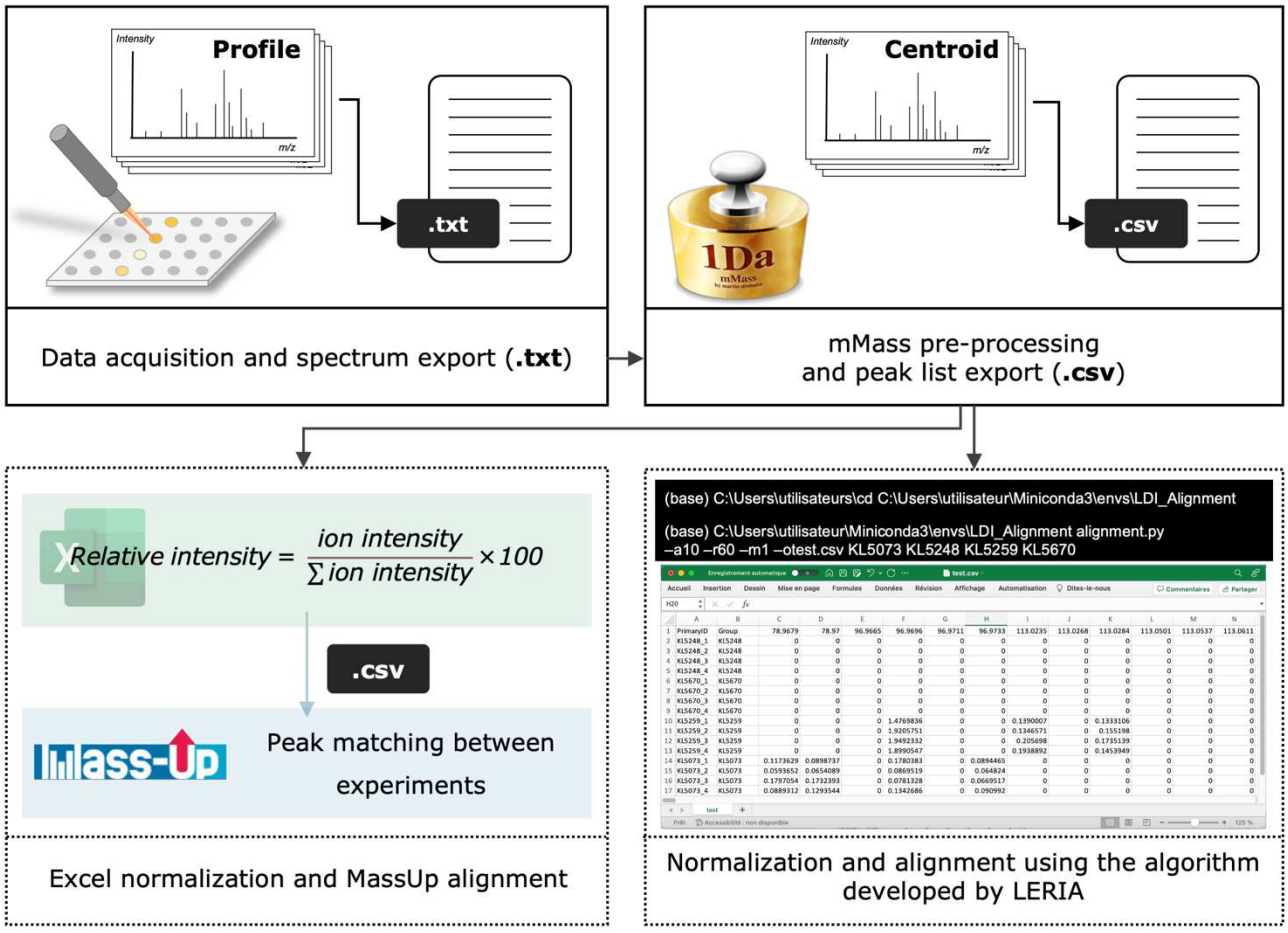


Figure 13. Schema for LDI-MS data processing from acquisition to peak alignment

- **via MassUp software**

As mMass does not allow automatic peak alignment this automatically .txt were (through Excel) transformed in the .csv format, which is compatible with the MassUp software.

The latter can perform normalization, smoothing, baseline correction, peak picking, and data alignment. However, when trying to import several profile mode mass spectra to execute these steps, the software constantly failed to function properly, and data were lost.

For that reason, the program was exclusively used for peak alignment of post-processed centroid data (see section 2.5.4). Based on an algorithm by Kazmi *et al.* (Kazmi *et al.*, 2006), peaks within a defined mass rage are clustered. Signals within a certain set mass range to the averaged m/z values of a cluster are added to this cluster. For peaks outside this range a new cluster is created.

- **via a newly created "in-house" data alignment algorithm**

As data alignment by MassUp software is a time consuming, cumbersome, and tedious procedure a more feasible and user-friendly solution was needed. Therefore, a specifically adapted Python-based script was developed in close collaboration with the "Laboratoire d'Etude et de Recherche en Informatique d'Angers" (LERIA).

The script is compatible with mMass (.csv files) and automatically normalized selected mass signals according to Equation 2. In addition, ions are aligned according to several manually adjustable parameters:

- **a10**: Intra-experience tolerance setting (20ppm is the default value)
- **r60**: Inter-experience tolerance setting (50ppm is the default value)
- **m3**: Specify that 3 missing values are allowed, from the 4th all experiment intensities will be replaced by 0 (1 is the default value)
- **p2**: Indicate that intensities are rounded to 10^{-2} (4 is the default value)
- **otest.csv**: Specify the name of the file (default: output.csv)

Following the same alignment strategy as MassUp, ions from replicates from one sample are first aligned to each other (intra-matching). Then, all spectra from different samples are matched against each other (inter-matching) (Figure 14). In case of missing signals for certain replicates, any of these signals were set to 0 in all replicates. Finally, the program a created .csv files that allowed further statistical processing in EZinfo (Umetrics, Umea, Sweden).

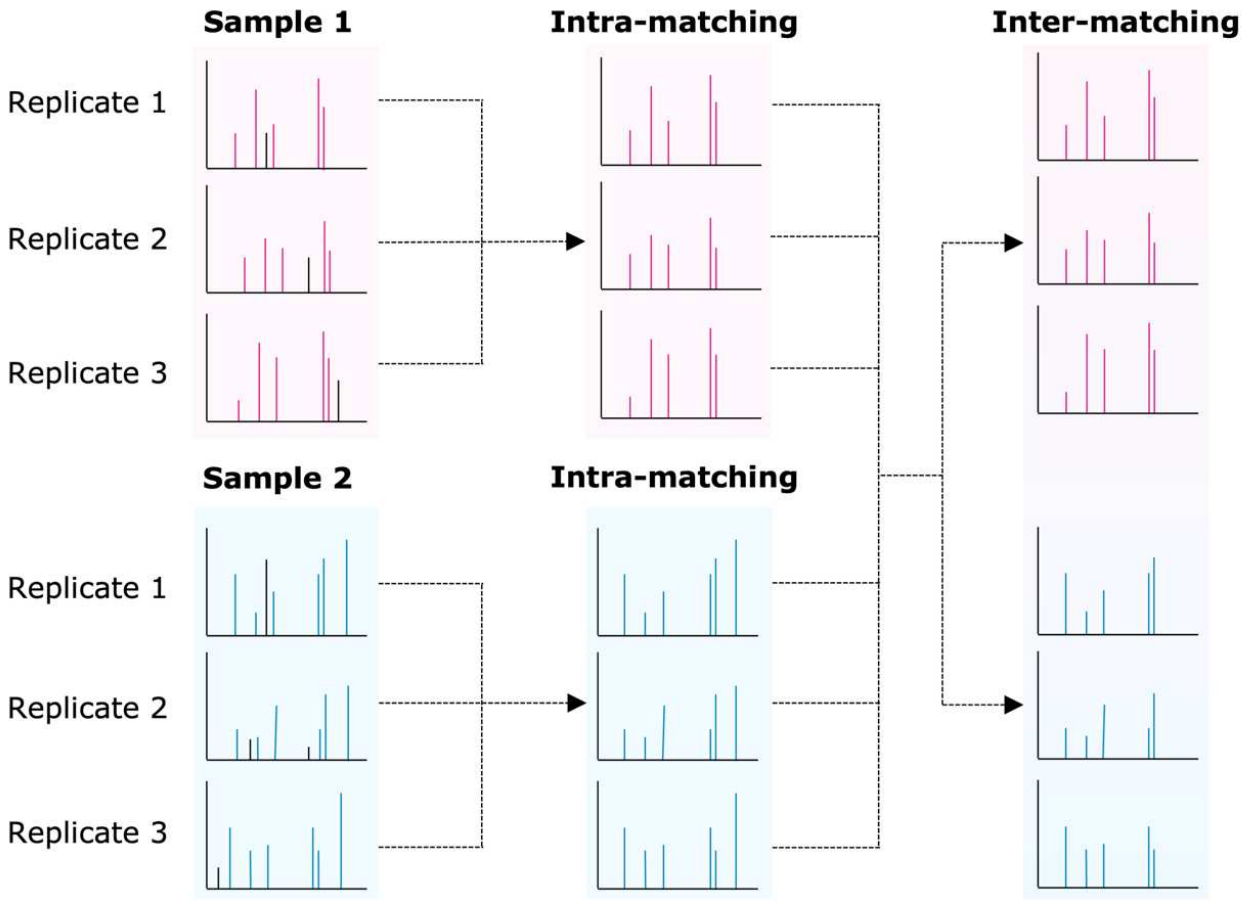


Figure 14. MassUp software alignment strategy used for the script developed by LERIA

2.6.3. Statistical processing of LDI-MS data

Statistical analyses of LDI-MS data was performed using EZinfo (Umetrics, Umea, Sweden). PLS analyses based on LDI-MS (m/z values, normalized signals intensities) and activity data (AGEs inhibition) were performed using the same parameters as applied for UPLC-MS/MS data (section 2.5.5.2).

2.7. General information about LDI-IM-MS

Detection by LDI-MS has many advantages (rapid analysis, simple sample preparation, low solvent consumption). However, the absence of chromatographic separation can cause certain problems. Most importantly, isomers can never be distinguished by classic LDI-MS as they share the same molecular formula and consequently the same mono isotopic masses. Chromatographic separation techniques, such as LC, provide an additional dimension of separation based on the different affinities of the isomers for the stationary phase. Consequently, isomers can be distinguished as long as they can be chromatographically separated.

An alternative approach to the differentiation of isomers is ion mobility spectrometry (IMS). Ions are separated by the interaction with a drift gas based on their size, shape, charge state and spatial orientation. IMS was successfully applied to separate constitutional isomers (Hofmann *et al.*, 2015; Hädener *et al.*, 2018; Zhang *et al.*, 2018). With this in mind a combined LDI-IM-MS may provide an interesting approach and will be discussed in section 5.2.

Several types of IMS technologies are currently available (Figure 15) (Masike, Stander and de Villiers, 2021):

- Drift time ion mobility spectrometry (DTIMS)
- Traveling wave ion mobility spectrometry (TWIMS)
- Differential mobility spectrometry (DMS)
- Differential mobility analyzer (DMA)
- Trapped ion mobility spectrometry (TIMS)

A summary is provided in Figure 15 and is inspired by the work of J. N. Dodds et al. and M. Kliman et al. (Kliman, May and McLean, 2011; Dodds and Baker, 2019), where they describe each method.

2.7.1. The trapped ion mobility spectrometry coupled to LDI-MS

As part of a research exchange program with the Technical University of Munich (BAYFRANCE project FK-04-2022), all LDI-IM-MS experiments were performed on the Bruker Daltonics' TIMS instrument (Figure 16, A).

Ions are first generated by an ionization source, such as ESI, MALDI, or atmospheric pressure chemical ionization (APCI). These ionization sources convert sample analytes into gas-phase ions. These ions are then trapped and separated based on their mobility in a drift tube. In the presence of a weak electric field, they are accelerated against a drift gas towards a detector. Ions with higher mobility reach the detector before ions with lower mobility. This separation is achieved by adjusting the strength of the electric field and the length of the drift region. After passing the drift tube ions directly enter a QTOF mass spectrometer. The combined use of TIMS and QTOF (timsTOF) enhances the ability to characterize and identify metabolites from complex mixtures. In addition, IMS measurements can provide CCS values, which reflect the size and shape of ions in the gas phase. These CCS values allow the direct comparison of data from the different instrument types that are outlined in section 2.7 and can provide an additional information for the identification and characterization of NPs (Schrimpe-Rutledge, Sherrod and McLean, 2018).

2.7.2. Collision cross section definition and databases

The CCS is a physical property of an ion related to its conformation, shape, charge, and mass (Hines *et al.*, 2017). It is determined by the interaction of the ion with the drift gas in the drift tube (Zhang et al., 2018). This information can be used to distinguish isomeric and closely related molecules that exhibit similar or equivalent m/z ratios.

Databases based on CSS values compile experimental and/or predicted CCS values for a wide range of molecules, including small organic compounds, peptides, proteins, and lipids. This can be useful for annotating compounds that are already reported. However, it should be noted that the CCS values in DBs are influenced by the instrument type and experimental parameters (instrumental conditions (in case of LC-IMS-MS), drift gas composition, temperature, and pressure) (Hinnenkamp *et al.*, 2018). Therefore, it is important to consider these factors when using CCS values from DBs because some compounds can show deviations up to 6.2%.

The timsTOF instrument is particularly suitable for the analysis of isomers, conformers, and complex mixtures, providing additional information on the structure of analytes. As said before the approach may overcome certain limitations of LDI-MS, which is linked to the absence of chromatographic separation. This includes the differentiation of constitutional isomers.

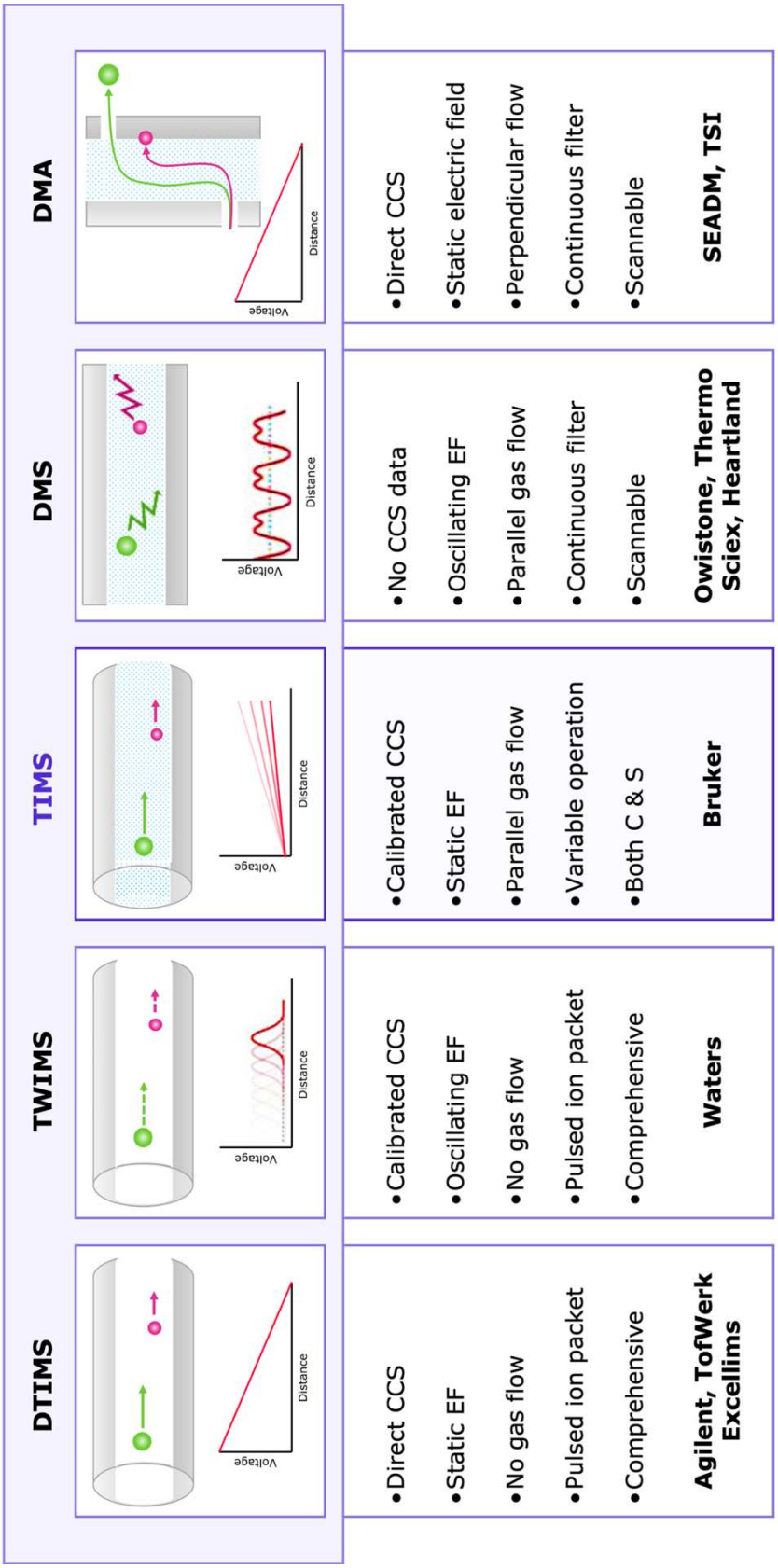


Figure 15. The most common IMS instruments (Kliman, May and McLean, 2011; Dodds and Baker, 2019). Methodological key features and commercial vendors are listed below. The dashed blue background indicates the presence of drift gas.

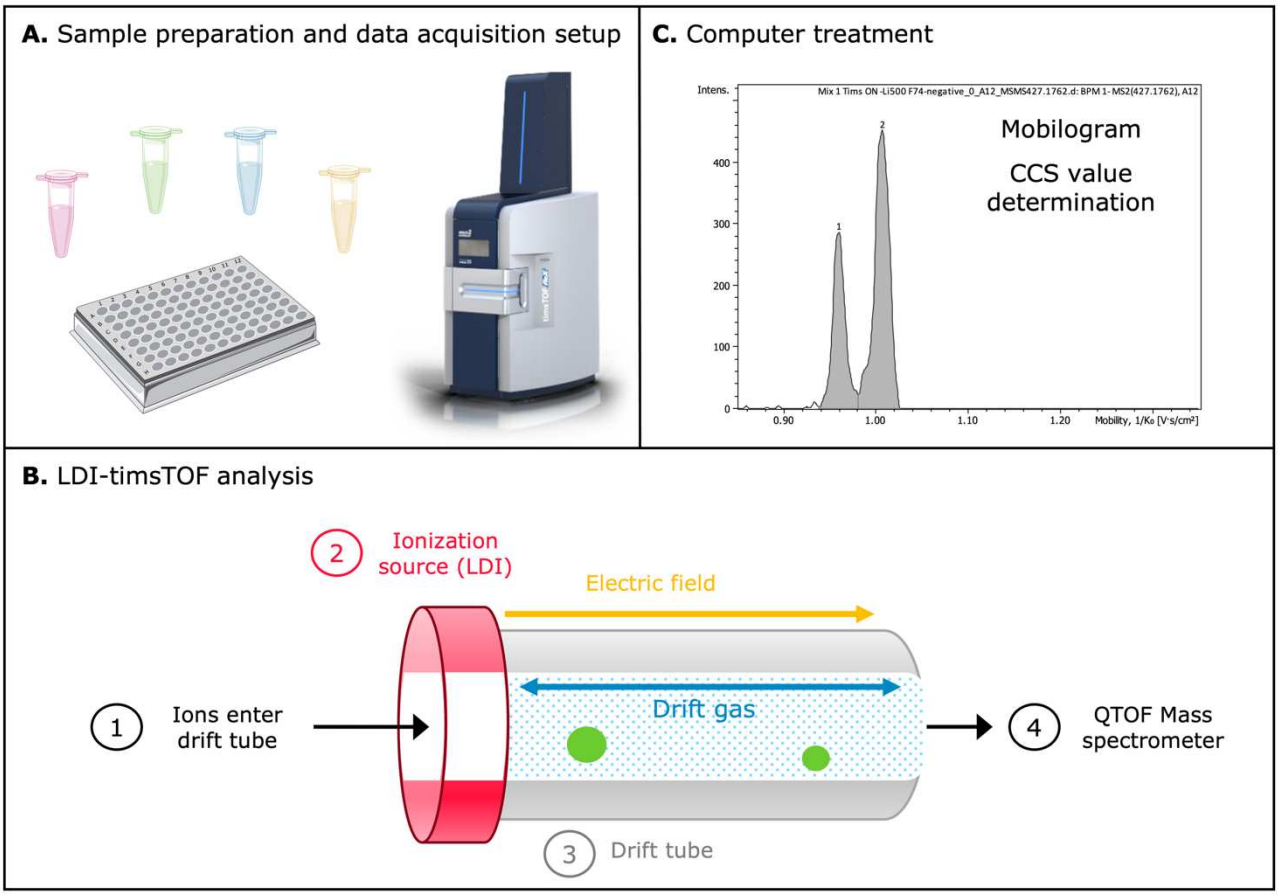


Figure 16. Schema of a LDI-TimsTOF analysis workflow. (A) Sample preparation. (B) The ions enter the drift tube (1) after being ionized by LDI (2). The ions then accelerated by an electric field and moved against a drift gas towards a detector. The interaction with the drift gas depends on the physical properties of the ions. (3). The QTOF mass spectrometer detects the ions (4). (C) Finally, the CCS value for each ion is calculated by a computer (5).

2.7.3. Post-acquisition parameters applied

2.7.3.1. Samples preparation

Stock solutions of all compounds were prepared in acetone at a concentration of 5.0 mg/mL. Working solutions were then prepared by mixing 200 µL of each compound as follow: Mix 1 (butyraxanthone D and cratoxylone), Mix 2 (cratoxylone and garcinone D), Mix 3 (butyraxanthone D, cratoxylone and parvixanthone G), Mix 4 (butyraxanthone D, cratoxylone, garcinone D and parvixanthone G), and Mix 5 (rubraxanthone and α-mangostin).

Mixtures of purified compounds initially prepared in acetone were dried and redissolved in DCM/MeOH (60/40). **Crude *G. parvifolia* extracts** as well as **fractions of KL5073** were dissolved in a mixture of DCM/MeOH (60/40) at a concentration of 10.0 mg/mL. All samples were deposited in triplicates (0.5 µL/each) on the sample carrier.

2.7.3.2. Acquisition parameters of LDI-IM-MS/MS experiments

Experiments were conducted on a timsTOF fleX MALDI-2 (Bruker Daltonics GmbH & Co. KG, Bremen, Germany), equipped with a Bruker scanning SmartBeam 3D Laser and a Bruker MALDI-2 Laser operating at 355 nm and 266 nm, respectively. Spectra were recorded with the following settings: mass range, 50–1200 m/z; ion mobility ramp, 0.8 -1.3 Vs/cm^2. Laser frequency (1000 Hz), laser intensity (70 – 100 %), CE (28 - 55 V), shots 500, were individually adapted to sample and analyte requirements. Preceding experiments have shown that best results were obtained in negative ionization mode, so all spectra were recorded in high-resolution negative mode. Data were processed using the DataAnalysis 5.3 software.

From each fraction of KL5073, the most intense ions were selected for ion mobility and MS/MS fragmentation. In fact, ions with low intensities did not yield exploitable MS/MS results.

2.7.3.3. Molecular networking parameters

The MS/MS spectral data were exported as text files (.mgf) from DataAnalysis 5.3 and manually merged in one file. A classical molecular network (MN) was then created with the following parameters: a precursor ion mass and fragment ion mass tolerance set at 0.02 Da, a cosine score equal to or greater than 0.5, topK was set to 5, minimum matched fragment ions was set to 3 and maximum connected component size was 0. The library search options were set to 4 as the minimum number of matched. The score threshold was set to 0.7. The job record is available at:

https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3a7d8e158cee45e2a8dcba86ce88f147

Graphical visualization and the treatment of molecular network was performed using Cytoscape 3.9.0 software(Lopes *et al.*, 2010).

2.8. Molecular networking and creation of spectral libraries for dereplication

2.8.1. Molecular networking principle and GNPS libraries

Dereplication strategies play a critical role in NP discovery, where the identification of known compounds is a fundamental step in focusing efforts on the characterization of novel molecules. With the increasing availability of HRMS data and advances in computational tools, MN has emerged as a powerful approach for dereplication of complex mixtures. MN uses the concept of spectral similarity to organize MS/MS data into interconnected clusters, where each MS/MS spectrum is represented as a node, typically depicted as a circle. These nodes are connected by edges representing spectral similarities, often quantified using cosine scores or modified cosine scores (see section 1.2). Clusters, on the other hand, are groups of related nodes that are likely to represent the same compound or family of compounds (Figure 17). Additional information, such as compound annotations or metadata, can be incorporated to enhance visualization.

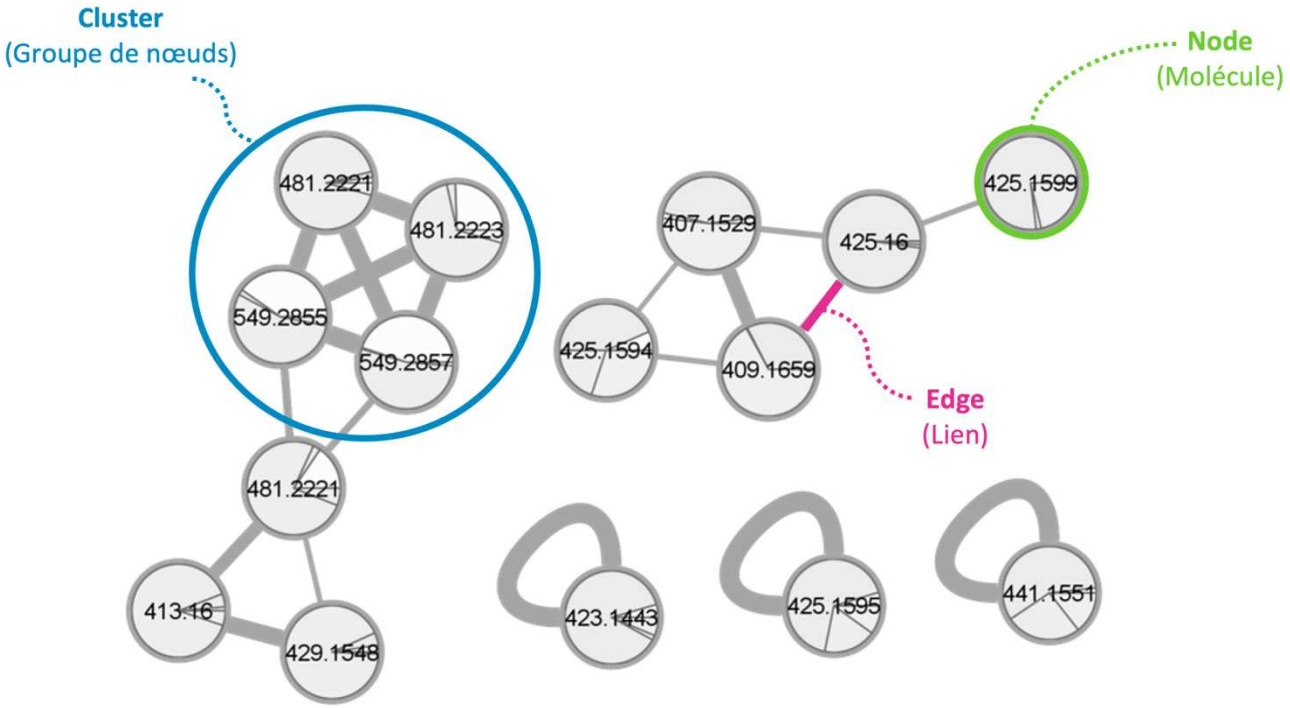


Figure 17. Molecular network principle

In addition, the creation of spectral libraries plays a critical role in MN. Spectral libraries are collections of MS data (experimental or *in silico* (see section 1.2)) that serve as references for the identification of known compounds. By allowing exploration of chemical space, these annotations can aid in the discovery of potential new molecules. This can save researchers time and effort.

However, the current MS/MS spectral libraries available on GNPS are poor in negative ionization and xanthones data (ɣ-mangostin is the only negatively ionized xanthone available). Since the SONAS laboratory being specialized in plant-derived NPs, with a particular focus on xanthones, database called Xanthones-DB was created.

2.8.2. Xanthones analyses by UPLC-MS

2.8.2.1. Samples preparation and chromatographic conditions

Isolated xanthones were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.001 mg/mL (injection volume 1 µL).

All experiments were conducted at 25°C on an Acquity UPLC system (Waters, Manchester, UK) using a HSS T3 C-18 (2.1 × 100 mm, 1.8 µm) UPLC column (Waters). LC-MS grade water (solvent A) and acetonitrile (solvent B, LC-MS grade, VWR chemicals), both supplemented with 0.1% of formic acid (LC-MS grade, VWR chemicals) were used as mobile phases. The mobile phase gradient was as follows: 0 min, 70% A; 0–2 min, 0% A; 2–5 min 0% A; 5–5.50 min, 70% A; 5.50-8 min, 70%A. The flow rate was 0.4 mL/min.

2.8.2.2. Mass spectrometry settings

Mass chromatograms and spectra were recorded on a WatersXevo®G2-S QTOF mass spectrometer (Waters) using negative and positive electrospray ionization. Source parameters were as follows: capillary and cone were set to 0.50 and 40 kV, respectively. Desolvation gas flow was 600 L/h at 36°C, and cone gas was set to 25 L/h.

Source temperature was 80°C and data were collected in continuum mode at 0.3 scan/s. Instrument calibration was conducted using a leucine-enkephalin solution at a flow rate of 5 µL/min (calibration points: 554.26202 Da for negative mode and 556.27658 Da for positive mode). All UPLC-MS/MS data were acquired in data dependent analysis (DDA) mode. The mass analyzer mass range was set to 50–600 Da for full scan mode (scan time: 0.15/s). The MS/MS mode was automatically activated once the total ion current (TIC) exceeded 10,000 intensity/s but remained in MS mode when TIC was below these thresholds. Each xanthone was analyzed according to 3 CE set at 20, 30 and 40 eV. These three different CE provided three different spectra for the same molecule.

2.8.3. Xanthones analyses by LDI-MS

2.8.3.1. Samples preparation

Purified compounds from KL5670 extract were prepared in acetone at a concentration of 5 mg/mL.

2.8.3.2. Instruments settings

Experiments were conducted on a timsTOF fleX MALDI-2 (Bruker Daltonics GmbH & Co. KG, Bremen, Germany), equipped with a Bruker scanning SmartBeam 3D Laser and a Bruker MALDI-2 Laser operating at 355 nm and 266 nm, respectively. Spectra were recorded with the following settings: mass range, 50–1200 m/z; ion mobility ramp, 0.8 - 1.3 Vs/cm^2. Laser frequency (1000 Hz), laser intensity (70 – 100 %), CE (35 V, except for isocowanin and isocowanol: 40 eV), shots 500, were individually adapted to analyte requirements. MS/MS spectra were obtained in negative ionization mode. Data were processed using the DataAnalysis 5.3 software and exported as .mgf files.

2.8.4. Creation of Xanthones-DB

2.8.4.1. LDI-MS data preparation and submission for GNPS library creation

Individual MS/MS spectra have been carefully selected for export to the .mgf format from the DataAnalysis 5.3 software and uploaded with a metadata file as an open access MS/MS database (available at https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=XANTHONES-DB).

The database is called Xanthones-DB.

2.8.4.2. UPLC-MS data preparation and submission for GNPS library creation

Individual runs were imported into ProgenesisQI for post-processing. Spectra were then exported to .msp format and converted to .mgf format using the MSP2MGF algorithm (https://sourceforge.net/projects/msp2mgf/). Each file were uploaded with a metadata file as an open access MS/MS database (available at https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=XANTHONES-DB).

The database is called Xanthones-DB.

2.8.4.3. Xanthones library creation for ProgenesisQI dereplication

The UPLC-MS data were also uploaded into the ProgenesisQI software as an *in-house* library called Xanthones-Manon.

2.9. General information about NMR

Nuclear magnetic resonance is a powerful analytical technique widely used in NPs research. It provides detailed information about the structure, dynamics, and chemical environment of molecules by exploiting the magnetic properties of atomic nuclei. In the context of NPs research, NMR plays a critical role in the identification, structural elucidation, and characterization of organic compounds derived from natural sources.

Samples are prepared in an appropriate deuterated solvent (typically $CDCl_3$, acetone-d_6 or other), carefully selected to ensure sample solubility and to avoid signal overlap. The prepared sample is then transferred to an NMR tube and subjected to NMR analysis. Afterwards, the tube is placed in a strong magnetic field, which aligns the nuclear spins of the atoms in the sample. The application of radio frequency pulses causes the previously selected nuclei to absorb and re-emit energy at characteristic frequencies, producing a unique NMR spectrum. The positions and intensities of the resulting peaks in the spectrum provide valuable information about the molecular structure, including the types of atoms present, their connectivity, and the chemical environments they experience.

In NPs research, NMR is often used in combination with other spectral techniques, such as MS to achieve comprehensive characterization and identification of compounds. The combination of these techniques allows researchers to confidently identify molecules, assess their purity, and study their link with biological activities, making NMR an indispensable tool.

In this manuscript, most of the experiments performed by NMR were ^{1}H, ^{13}C, DEPT-135 and DEPT-90 as data were often processed using MixONat software. A detailed description of the algorithm workflow can be found in the next section (2.9.2).

Figure 11. Nuclear magnetic resonance spectrometer from JEOL

2.9.1. Experimental acquisition by NMR

2.9.1.1. Sample preparation

Experiments (^{1}H NMR, ^{13}C NMR, DEPT-135, DEPT-90 and 2D NMR) were performed on a JEOL 400 MHz NMR spectrometer (JEOL, Tokyo, Japan) equipped with an inverse 5 mm probe (ROYAL RO5). Chemical shifts (δ_H and δ_C) were expressed as ppm and ^{1}H coupling constants (J) are provided in Hz. Depending on their solubility samples were dissolved either in DMSO-d_6 or acetone-d_6 both purchased from Eurisotop, (St-Aubin, France).

For dereplication experiments M1 (10.0 mg) was dissolved in 500 µL DMSO-d_6, while fractions and isolated compounds from KL5670 were dissolved in 500 µL acetone-d6 (F2 and F3 were also dissolved in $CDCl_3$).

2.9.1.2. Acquisition parameters for ^{13}C NMR-based dereplication

For ^{13}C NMR (100 MHz) spectra, a WALTZ-16 decoupling sequence was used with an acquisition time of 1.04 s (32,768 complex data points) and a relaxation delay of 2 s. The number of acquisitions (512–20,000 scans) was individually adapted to sample requirements. A 1 Hz exponential line-broadening filter was applied to each FID prior Fourier transformation. Raw data were processed by MestReNova 12.0.2 (Mestrelab Research, Santiago de Compostela, Spain) and calibrated to solvent peaks at δ_C 39.52 ppm (DMSO-d_6) or 29.84 ppm (acetone-d_6) respectively. Manual phasing and baseline correction were conducted. DEPT experiments were aligned with ^{13}C spectra at a given δ_C. Positive ^{13}C NMR and DEPT-90 as well as positive and negative DEPT-135 signals were collected and exported as .csv files.

2.9.2. MixONat: an overview of the algorithm

Developed in the SONAS lab by A. Bruguière *et al.*, MixONat is a freely available software (open-source algorithm based on Python 3.5) designed to help researchers to identify major NPs in complex mixtures (Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Suor-Cherer, *et al.*, 2020) using their ^{13}C NMR chemical shifts, optionally combined with DEPT-135 and -90 data to discriminate between CH_3, CH_2, CH and Cq. The software automatically compares the data with that of a DB containing selected NPs together with their experimental or predicted chemical shifts (δc) DB.

- **Data importation**

The algorithm has a first tab (Figure 18) that allows the import of the DB in the format of a structure data file (.sdf) sorted by the C-typeGen program if required. The latter allows sorting each δc by carbon type for all NPs in the DB. The user must also import experimental data (^{13}C, DEPT-135 and -90) in .csv format.

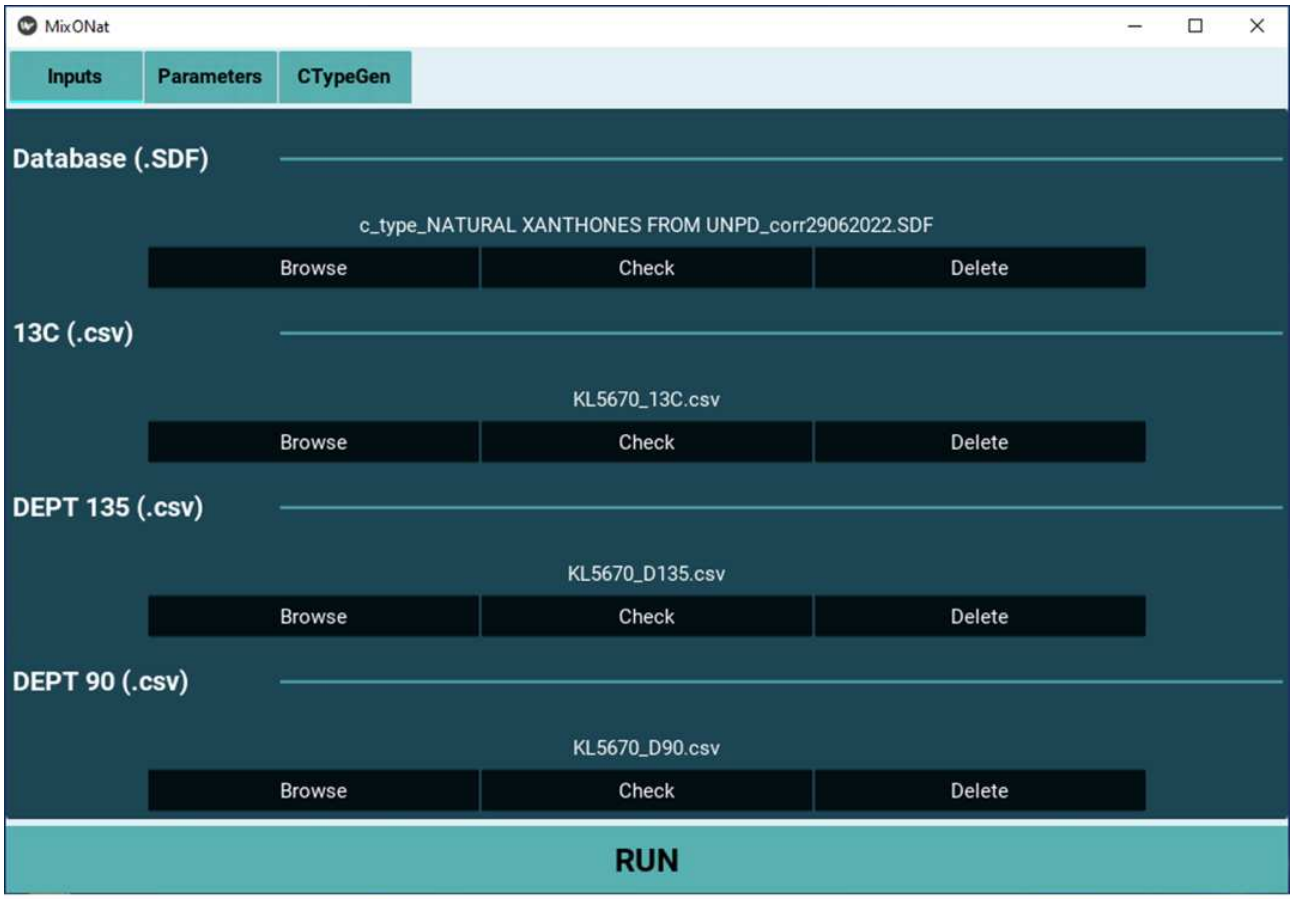


Figure 18. MixONat algorithm: "Inputs" tab. The "Inputs" tab allows to import the database in .sdf format and the spectrum in .csv (DEPT-135 and -90 are optional).

- **Parameters setup**

The second tab (Figure 19) is used to set several parameters, such as the δc tolerance (default: 1.30 ppm), the increment tolerance, the DEPT alignment (default: 0.02 ppm), the equivalent carbon and a molecular weight filter.

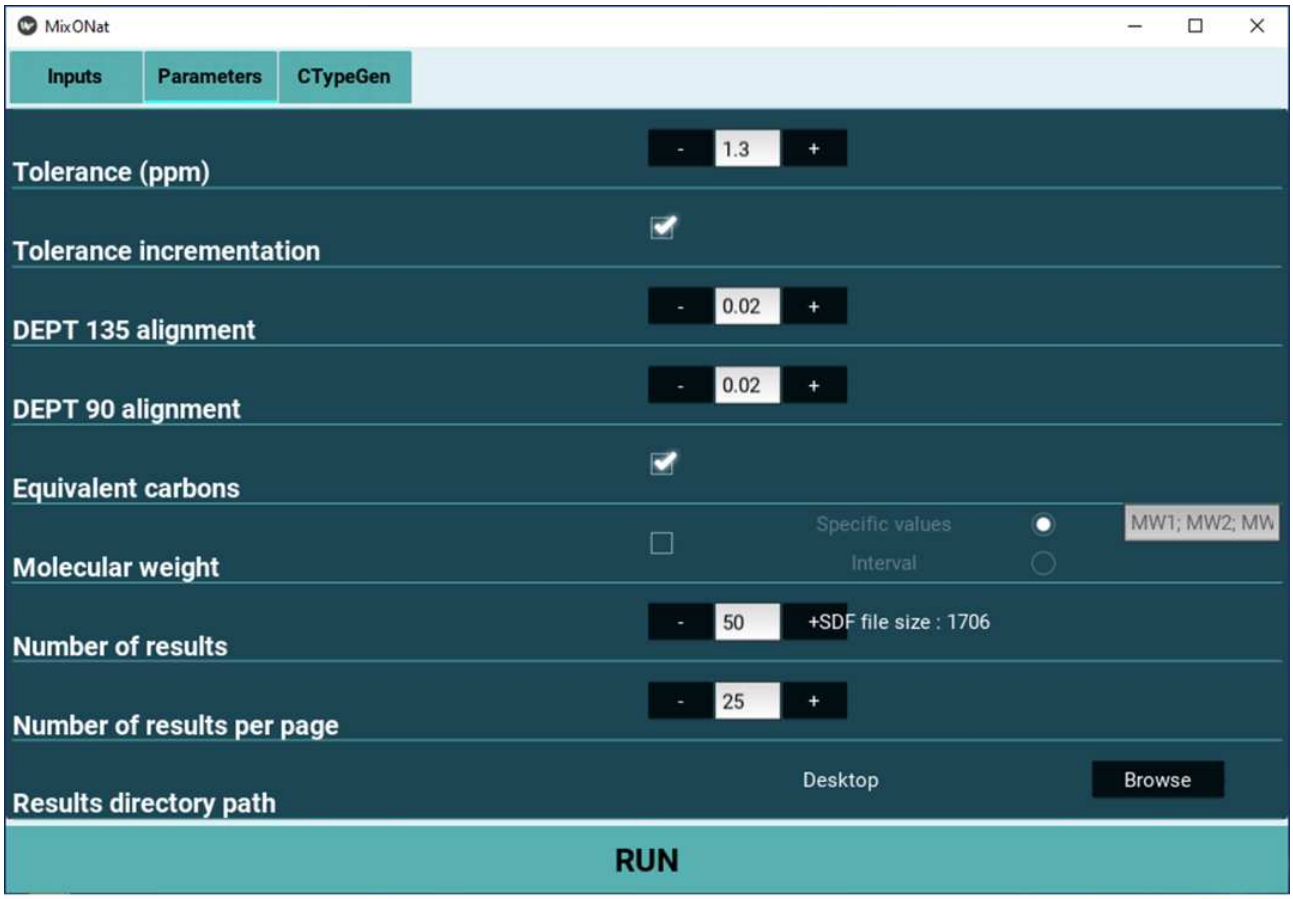


Figure 19. MixONat algorithm: "Parameters" tab. The "Parameters" tab allows to set the different options according to each experiment.

- **Run the algorithm**

Finally, the algorithm can then be run to start the matching process, which consists of comparing the experimental δc with that of the compounds in the DB. When all carbons have been considered, a score is calculated according to the number of δc matched for each compound in the DB.

At the end of the matching process, a list of potential molecules present in the complex mixtures is displayed and ranked by decreasing score. The on-screen results also display the structure, name, molecular weight, score, and error calculated as the cumulative absolute difference between the matched signals. The final step is to review the results (Figure 20). For each molecule, the algorithm displays the matched δc in red in the structure and in a graphical representation (which also displays the δc intensities). This allows the user to compare the information provided by the algorithm with the raw data and decide whether to add or remove δc.

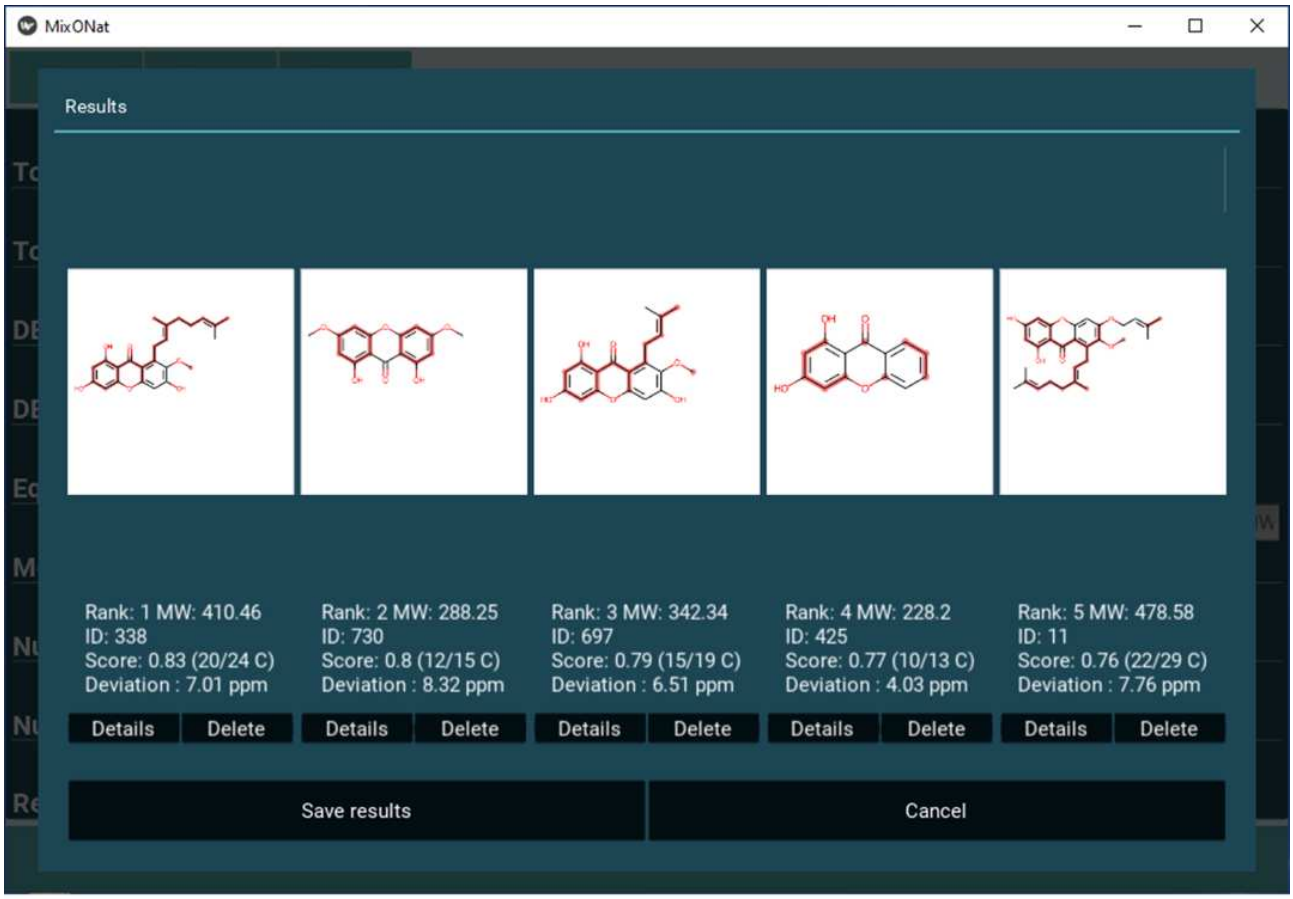


Figure 20. MixONat algorithm: matching screen of experimental δc and the imported database.

2.9.3. Formatted databases for MixONat used in the present work

All ^{13}C NMR and DEPT data were processed by the MixONat ^{13}C NMR dereplication algorithm (Bruguière, Derbré and Richomme, 2020; Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Suor-Cherer, *et al.*, 2020; Bruguière *et al.*, 2021).

In preparation of this analysis, reported structures of the *Garcinia* genus were first downloaded from the dictionary of natural products (DNP) (*Dictionary of natural products 31.2*, 2015), and compiled in a *Garcinia* database (***Garcinia*DB**). The latter comprised 718 entries, which were saved in the .sdf format.

Next, a second, xanthones database (**Xanthones DB**) was created from the universal natural products database (UNPD), which comprises 211,280 NPs (Gu *et al.*, 2013). In order to meaningfully reduce the overall number of compounds, UNPD was filtered by a xanthone scaffold using the ACD/NMR Predictors (C, H) and DB software, version 2019 (Toronto, Canada). Finally, 1694 xanthone structures were obtained from UNPD. Their ^{13}C NMR shifts (δC-SDF) were then simulated by the ACD/NMR Predictors (C, H) and DB software as well as the CType-Gen routine from MixONat used for differentiating methyl, methylene, methine or quaternary δC-SDF to get DBs compatible with MixONat (Bruguière *et al.*, 2018; Silva-Castro *et al.*, 2021).

Comparing simulated with experimental data (^{13}C NMR, DEPT-135 and DEPT-90), MixONat then ranked associated structures with a score from 0 to 1 (where 1 indicates a full match and 0 the absence of any similarity between

simulated and experimental spectra). Finally, experimental data of highest ranked candidates were compared with the literature in order to confirm proposed structural assignments.

3. Prediction of natural products activity in complex mixtures using chemometric analyses

3.1. Introduction

The identification of (bio)active constituents in complex mixtures of natural products (NPs) is an important area of NPs research. It requires a thorough chemical analysis of the sample material, which is generally facilitated by LC-HRMS. However, a complementary approach was developed to simplify and accelerate this process. Thus, the following publication (published in Talanta) presents the use of LDI-HRMS within a chemometric model to identify activity markers in complex mixtures.

The publication consists of four major parts. First, the development of a statistical model based on well-defined mixtures of active and inactive reference compounds was established. Then, analysis of real samples consisting of four *G. parvifolia* bark extracts were performed. Results obtained by LDI-HRMS were compared with those obtained by simultaneous UPLC-HRMS^2 analyses. Finally, potentially active compounds detected in the mixture by LDI-HRMS and UPLC-HRMS^2 were identified using the ^{13}C NMR dereplication tool MixONat.

Through these analyses, the publication aims to highlights the complementarity of UPLC-HRMS^2, LDI-HRMS and ^{13}C NMR as a versatile and holistic approach towards chemometrics in NPs research and other areas of analytical chemistry.

[FR]

L'identification des constituants (bio)actifs à partir de mélanges complexes de NPs est un domaine important de la recherche sur les NPs. Cela nécessite une analyse chimique approfondie du matériau d'échantillon, qui est généralement facilitée par LC-HRMS. Cependant, une approche complémentaire a été développée pour simplifier et accélérer ce processus. Ainsi, la publication suivante (publiée dans Talanta) présente l'utilisation du LDI-HRMS dans le cadre d'un modèle chimiométrique pour identifier des marqueurs d'activité dans des mélanges complexes.

La publication est divisée en quatre parties. Tout d'abord, le développement d'un modèle statistique basé sur des mélanges bien définis de composés de référence actifs et inactifs a été établi. Ensuite, une analyse d'échantillons réels composés de quatre extraits d'écorce de *G. parvifolia* a été réalisée. Les résultats obtenus par LDI-HRMS ont été comparés à ceux obtenus par des analyses par UPLC-HRMS^2. Enfin, les composés potentiellement actifs en mélange détectés par LDI-HRMS et UPLC-HRMS^2 ont été identifiés à l'aide de l'outil de déréplication MixONat.

À travers ces analyses, la publication vise à mettre en évidence la complémentarité de l'UPLC-HRMS^2, du LDI-HRMS et de la RMN dans une approche globale, polyvalente basée sur la de la chimiométrie qu'il s'agisse de recherche sur les PNs ou dans d'autres domaines de la chimie analytique.

3.2. Matrix free laser desorption ionization assisted by ^{13}C NMR dereplication: a complementary approach to LC-MS^2 based chemometrics

Talanta 253 (2023) 123998

Contents lists available at ScienceDirect

Talanta

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

Matrix free laser desorption ionization assisted by ^{13}C NMR dereplication: A complementary approach to LC-MS^2 based chemometrics

Manon Meunier [a], Dimitri Bréard [a,b], Khalijah Awang [c], Séverine Boisard [a,b], David Guilet [a], Pascal Richomme [a], Séverine Derbré [a,*], Andreas Schinkovitz [a,*]

[a] Univ Angers, SONAS, SFR QUASAV, F-49000 Angers, France
[b] PHYTO Platform, SFR QUASAV, Campus du Végétal, 49045, Angers CEDEX 01, France
[c] University of Malaya, Faculty of Sciences, Department of Chemistry, Malaysia

ARTICLE INFO

Keywords:
Chemometrics
Matrix free laser desorption ionization mass spectrometry
LC-MS^2
^{13}C NMR dereplication
Anti-AGE activity
Garcinia parvifolia

ABSTRACT

The identification of bioactive constituents from complex mixtures of Natural Products (NPs) is an important field of applied chemometrics. It requires the thorough chemical analysis of sample material, which is generally facilitated by liquid chromatography coupled to high-resolution tandem mass spectrometry (LC-MS^2). Providing a simple and time efficient complementary approach, the current work presents the use of matrix free laser desorption ionization mass spectrometry (LDI-MS) within a chemometric model that allows the identification of activity markers in complex mixtures. As a working example the correct prediction and consecutive isolation of NPs with notable anti-AGEs effects from crude bark extracts of *Garcinia parvifolia* is discussed. Results were thoroughly compared to those obtained by a concurrently performed LC-MS^2 analyses. In addition, the ^{13}C NMR dereplication tool MixONat was successfully used to confirm the structures of high yield activity markers directly from complex mixtures prior to their isolation. Overall LDI-MS yielded comparable results to LC-MS^2 and should be considered as complementary technique in applied chemometrics.

1. Introduction

Chemometric analyses have become an important tool in natural products (NPs) research. These techniques permit the chemical characterization of sample material, but also spotting activity markers within complex mixtures by various statistical means. While *e.g.* Principal Component Analysis (PCA) [1] facilitates the grouping of sample material according to its chemical profile, Partial Least Squares (PLS) regression analysis may further include biological or chemical activity data [2], thus allowing the identification of sample-specific activity markers [3]. Independently from the applied statistical model, all chemometrics do require sound analytical data to yield exploitable results. In this respect, liquid chromatography coupled to high-resolution tandem mass spectrometry (LC-MS^2) is considered a gold standard method. Combining the assets of chromatographic separation with the sensitivity of mass spectrometry, the technique permits the detection of a wide range of NPs at very low concentrations [4]. However, despite these indubitable benefits, LC-MS^2 may still encounter certain limitations. For example, sample solubilization is generally limited to mobile phases, mostly comprising mixtures of acetonitrile, methanol, and water, which are not well adapted to lipophilic compounds. Commonly applied electrospray ionization (ESI) typically promotes the formation of quasi-molecular ions such as $[M+H]^+$ or $[M-H]^-$. However, certain constituents, *e.g.* the lichen quinones haemventosin and parietin, preferably form radical ions and may fail to ionize in ESI [5]. Similar observations were also made for some vitamin E isomers [6]. Facilitating the formation of both, radical and quasimolecular ions [5–7], matrix free laser desorption ionization mass spectrometry (LDI-MS) may therefore provide a valuable supplement to LC-ESI-MS [8]. As many NPs show good absorption at the emission wavelength of common UV-LDI lasers, the method covers a wide range of chemically diverse compounds such as alkaloids [9], phenolics [10,11] or cannabinoids [12]. Moreover, LDI-MS may be directly applied on crude extracts of different polarities [9,10] without any sample prepurification steps and even works on powdered non-extracted plant material [13]. In addition, experimental times are much shorter than in LC-MS^2. However, despite these promising results and to the best of the authors' knowledge, LDI-MS has never been included into a chemometric workflow targeting

* Corresponding authors.
E-mail addresses: severine.derbre@univ-angers.fr (S. Derbré), andreas.schinkovitz@univ-angers.fr (A. Schinkovitz).

https://doi.org/10.1016/j.talanta.2022.123998
Received 11 July 2022; Received in revised form 3 October 2022; Accepted 5 October 2022
Available online 11 October 2022

the prediction of (bio)active compounds.

With this in mind, the current manuscript describes the step-by-step development of an LDI-MS based chemometric PLS regression model for the identification of activity markers from complex mixtures of NPs. A fluorometric assay measuring the inhibition of the formation of advanced glycation end-products (AGEs) was used for activity evaluation [14,15]. In the process, a statistical model based on well-defined mixtures of active and inactive reference compounds was first established (**I**). Next, real-life samples comprising four *Garcinia parvifolia* bark extracts were studied (**II**). Results were then compared to concurrently performed LC-MS^2 based chemometrics (**III**). Finally, selected activity markers previously identified by LDI-MS and LC-MS^2 were analyzed by MixONat, a free ^{13}C NMR dereplication tool (**IV**) that was developed in the SONAS laboratory [16].

2. Material and methods

2.1. Standard mixtures reference compounds, extraction solvents

Reference compounds amentoflavone (**1**) and betulinic acid (**2**) were purchased from Extrasynthese (Genay, France), while ferulic acid (**3**) was bought from Sigma Aldrich (St. Louis, USA), and 3-hydroxyxanthone (**4**) was obtained from Thermo Fisher (Kandel, Germany). Stock solutions of all compounds were prepared in acetone at a concentration of 10.0 mg/mL. Working solutions were prepared in molar ratios outlined in Table 1. A detailed experimental protocol is provided in Table S1 (supplementary material). Except for acetone (analytical grade), all solvents used for stock solutions, plant extractions, flash chromatography and thin layer chromatography (TLC), were of HPLC grade and purchased from Carlo Erba reagents, (Val-de-Reuil, France).

2.2. Plant collection

Garcinia parvifolia bark samples were collected in Malaysia at four different locations: KL5670, Hutan Simpanan Meranto Gua Musang, Kelantan (date of collection: 13.05.2009). KL5248: Hutan Simpanan Sungai Temau Luala Lipis, Pahang (date of collection: 25.05.2006). KL5259: Km 11 Sanggang, Kuala Kangsar, Perak (date of collection: 22.06.2006). KL5073: 1.2 km kg. Toh Kah, Jerangau, Terengganu (date of collection: 20.08.2004). Plant material was verified by the botanist Mr. Teo Leong Eng and voucher specimens are kept at the herbarium of the Department of Chemistry, University of Malaya, Kuala Lumpur, Malaysia.

2.3. Extraction procedure

For initial experiments, 10 g of grinded *G. parvifolia* samples were mixed with 100 mL dichloromethane (DCM) and extracted at room temperature for 30 min using an ultrasound bath (Elma, Singen, Germany). The instrument operated at 280 W and at a frequency of 37 kHz. After filtration and solvent evaporation, the following dry weights were obtained: KL5670: 460 mg, KL5248: 470 mg; KL5259: 444 mg, and KL5073: 578 mg. In order to facilitate additional sample processing, KL5670 (80 g) was re-extracted using an up-scaled extraction protocol, which yielded 4.48 g of dry extract.

Table 1
Standard mixtures of reference compounds and their anti-AGEs activities. Relative mixing ratios are expressed in mM. Inhibition of AGEs formation is expressed as IC_{50} ± (standard deviation). Aminoguanidine (AG) was used as positive control (IC_{50}: 139 ± 21 μg/mL or 1900 ± 300 μM).

Standard mixtures	Compounds				AGEs inhibition IC_{50} (μg/mL)
	(1)	(2)	(3)	(4)	
M5	2	3	6	6	26.3 ± 1.5
M2	12	3	6	6	28.3 ± 3.2
M6	7	7	7	7	34.7 ± 3.2
M1	5	5	13	12	38.0 ± 5.0
M4	2	3	32	6	58.0 ± 4.6
M3	2	14	6	6	73.0 ± 5.6
AG					139.0 ± 21.0

2.4. Flash chromatography

Sample KL5670 (4.20 g) was fractionated on a CombiFlash Rf flash chromatograph (Serlabo Technologies, Entraigues, France) using a Puriflash 50SIHC-F0220 (220 g) silica gel column. The mobile phase consisted of petroleum ether (solvent A), an equivolumetric mixture of ethyl acetate and chloroform (solvent B), and methanol (solvent C). The flow rate was set to 50 mL/min and the following gradient was applied: 0 min, (A/B/C - 100/0/0); 0–30 min, (A/B/C - 70/30/0); 30–60 min, (A/B/C - 70/30/0); 60–65 min, (A/B/C - 65/35/0); 65–78 min, (A/B/C - 65/35/0); 78–80 min, (A/B/C - 61/39/0); 90–110 min, (A/B/C - 61/39/0); 110–130 min, (A/B/C - 40/60/0); 130–140 min, (A/B/C - 0/100/0); 140–170 min, (A/B/C - 0/100/0); 170–190 min, (A/B/C - 0/90/10); 190–200 min, (A/B/C - 0/80/20); 200–230 min, (A/B/C - 0/80/20). In order to reasonably reduce the number of samples for consecutive LDI-MS and UPLC-MS^2 experiments, fractions were roughly combined according to their analytical TLC profile (data not shown). All TLC experiments were conducted on analytical silica gel 60 TLC plates (Merck, Darmstadt, Germany) comprising a fluorescence indicator at λ: 254 nm. Mobile phases consisted of mixtures of DCM/MeOH and were adapted according to sample requirements. Overall, 14 recombined fractions were obtained (Fig. S1, supplementary material).

2.5. Experimental preparative HPLC conditions

Fraction 6 (135 mg) and Fraction 10 (270 mg) were dissolved in MeOH (HPLC grade, VWR chemicals, Rosny-sous-Bois, France) before being processed on a Nexera Prep (Shimadzu, Kyoto, Japan) preparative HPLC system using a Pursuit XRs 5 C-18 column (250 × 21.2 mm, 5 μm, Agilent, Santa Clara, California, USA). The mobile phase consisted of HPLC grade water (Solvent A) and acetonitrile (Solvent B, VWR chemicals) and the flow rate was set to 21.2 mL/min. The injection volume was 2 mL.

Chromatograms were recorded at 254 and 280 nm. For F6 the solvent gradient was as follows: 0 min, 45% B; 0–25 min, 50% B; 25–30 min, 70% B; 30–35 min, 100% B; 35–40 min, 100% B. Two repetitive injections (56.4 and 78.6 mg) were performed and yielded 57.0 mg of rubraxanthone (**5**) at a retention time (rt) of 21.0 min.

Fraction 10 was processed using the following solvent gradient: 0 min, 60% B; 0–15min, 65% B; 15–20 min, 70% B; 20–30 min, 100% B; 30–35 min, 100% B. Overall, three repetitive injections (50.4, 94.3 and 120.0 mg) were performed yielding 46.5 mg of isocowanol (**6**, rt: 26.7 min) and 15.4 mg of parvixanthone G (**7**, rt: 12.5 min).

2.6. Experimental UPLC-MS conditions

All experiments were conducted at 25 °C on an Acquity UPLC system (Waters, Manchester, UK) using a HSS T3 C-18 (2.1 × 100 mm, 1.8 μm) UPLC column (Waters). LC-MS grade water (solvent A) and acetonitrile (solvent B, LC-MS grade, VWR chemicals), both supplemented with 0.1% of formic acid (LC-MS grade, VWR chemicals) were used as mobile phases. Chromatographic conditions for crude extracts and subfractions of KL5670 are outlined below.

Crude extracts: Samples were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.5 mg/mL and injected in triplicate (injection volume 1 μL). The mobile phase gradient was as follows: 0 min, 30% B; 0–10 min, 80% B; 10–17 min 100% B; 17–20 min, 100% B. The flow rate was 0.3 mL/min.

Subfractions of KL5670. Samples were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.02 mg/mL and injected in triplicate (injection volume 1 μL). The mobile phase gradient

was as follows: 0 min, 40% B; 0–5 min, 80% B; 5–10 min 100% B; 10–12 min, 100% B. The flow was set to 0.4 mL/min.

Mass spectrometry settings for all samples: Mass chromatograms and spectra were recorded on a WatersXevo®G2-S QTOF mass spectrometer (Waters) using negative electrospray ionization. Source parameters were as follows: capillary and cone were set to 0.50 and 40 kV, respectively. Desolvation gas flow was 1000 L/h at 36 °C, and cone gas was set to 100 L/h. Source temperature was 120 °C and data were collected in continuum mode at 0.5 scan/s. Instrument calibration was conducted using a leucine-enkephalin solution at a flow rate of 5 μL/min (calibration points: 554.26202 for negative mode). All LC-MS^2 data were acquired in Data Dependent Analysis (DDA) mode. The mass analyzer mass range was set to 50–1200 Da for full scan mode (scan time: 0.5/s). The MS^2 mode was automatically activated once the Total Ion Current (TIC) exceeded 20,000 intensity/s (crude extracts) or 10,000 intensity/s (subfractions of KL5670) but remained in MS mode when TIC was below these thresholds. Collision energy ramps of 20–25 V for low-mass, and 65–85 V for high-mass analytes were employed.

2.7. Experimental LDI-MS conditions

Experiments were conducted on a Spiral-TOF mass spectrometer (JEOL Ltd., Tokyo, Japan), equipped with a pulsed laser operating at 349 nm. Spectra were recorded within a mass range of 50–1200 m/z. Preceding experiments have shown that best results were obtained in negative ionization mode, so all spectra were recorded in high-resolution spiral negative mode. Laser frequency (50–100 Hz), detector gain (49–67%) and laser power (55–80%) were individually adapted to samples requirements.

Standard mixtures of reference compounds initially prepared in acetone were dried and redissolved in DCM. Crude *G. parvifolia* extracts as well as subfractions of KL5670 were dissolved in a mixture of DCM/MeOH (60/40) at a concentration of 10.0 mg/mL. All samples were deposited in quadruplicates (0.5 μL/each) on the sample carrier.

2.8. UPLC-MS^2 data processing and chemometric analysis

Automated data alignment and peak-picking were conducted by the ProgenesisQI software (Waters, Newcastle, UK). Results were thoroughly inspected and manually corrected if necessary (e.g., regrouping of some ions with identic retention times, peak shapes and/or isotopic distribution). An ANOVA tag (p-value ≤ 0.05) was applied to eliminate non-reproducible data. All ions without MS^2 fragmentation were removed and data were then processed by the Progenesis MetaScope dereplication tool. The latter creates theoretical fragmentation patterns from imported structures and compares them with experimental MS^2 data [17]. Putative structural assignments were made using the "Predicted Carbon-13 NMR Data of Natural Products" (PNMRNP) database, which comprises 211,280 NPs [18]. All structures were exported as .sdf, a format that is compatible with ProgenesisQI. Fragmentation tolerance was set to 5 ppm. Based on these parameters ProgenesisQI created a list of possible structures with their respective probability scores.

Finally, a PLS regression analysis based on spectrometric information (*m/z* values, retention times and peak areas) as well as activity data (AGEs inhibition) was conducted on EZinfo 3.0 for Waters (Umetrics, Umea, Sweden). The Pareto scale setting was applied.

2.9. LDI-MS data processing and chemometric analysis

Spectral data were exported as text files (.txt) using the Tornado MS (JEOL) software. Post-acquisition calibration and automated peak-picking were conducted by the mMass software [19]. Results were thoroughly inspected and manually corrected if necessary. All spectra were normalized by dividing the intensities of individual mass signals by the overall sum of all detected signals within a spectrum. Spectra were then aligned using the MassUp software [20]. Finally, PLS analyses based on LDI-MS (*m/z* values, normalized signals intensities) and activity data (AGEs inhibition) were performed on EZinfo (Umetrics, Umea, Sweden) using the same parameters as applied for LC-MS^2 data.

2.10. Experimental NMR conditions

Experiments (^{1}H NMR, ^{13}C NMR, DEPT-135, DEPT-90 and 2D NMR) were performed on a JEOL 400 MHz NMR spectrometer (JEOL, Tokyo, Japan) equipped with an inverse 5 mm probe (ROYAL RO5). Chemical shifts (δ_H and δ_C) were expressed as ppm and ^{1}H coupling constants (*J*) are provided in Hz. Depending on their solubility samples were dissolved either in deuterated DMSO-d_6 or acetone-*d*6 both purchased from Eurisotop, (St-Aubin, France). For dereplication experiments M1 (10.0 mg) was dissolved in 500 μL DMSO-d_6, while F6 (50.0 mg) and F10 (40.0 mg) were dissolved in 500 μL acetone-*d*6.

For ^{13}C NMR (100 MHz) spectra, a WALTZ-16 decoupling sequence was used with an acquisition time of 1.04 s (32,768 complex data points) and a relaxation delay of 2 s. The number of acquisitions (512–20,000 scans) was individually adapted to sample requirements. A 1 Hz exponential line-broadening filter was applied to each FID prior Fourier transformation. Raw data were processed by MestReNova 12.0.2 (Mestrelab Research, Santiago de Compostela, Spain) and calibrated to solvent peaks at δ_C 39.52 ppm (DMSO-d_6) or 29.84 ppm (acetone-*d*6) respectively. Manual phasing and baseline correction were conducted. DEPT experiments were aligned with ^{13}C spectra at a given δ_C. Positive ^{13}C NMR and DEPT-90 as well as positive and negative DEPT-135 signals were collected and exported as .csv files.

2.11. ^{13}C NMR dereplication

All ^{13}C NMR and DEPT data were processed by the MixONat ^{13}C NMR dereplication tool [21,22]. In preparation of this analysis, reported structures of the *Garcinia* genus were first downloaded from the Dictionary of Natural Products [23], and compiled in a Garcinia database (Garcinia DB). The latter comprised 718 entries, which were saved in the SDF format. Next, a second, xanthones specific database (Xanthone DB) was created from the Universal Natural Products Database (UNPD), which comprises 211,280 NPs [24]. In order to meaningfully reduce the overall number of compounds, UNPD was filtered by a xanthone scaffold using the ACD/NMR Predictors (C and H) and DB software, version 2019 (Toronto, Canada). Finally, 1694 xanthone structures were obtained from UNPD. Their ^{13}C NMR shifts ($\delta_{C\text{-}SDF}$) were then simulated by the ACD/NMR Predictors (C and H) and DB software as well as the CTypeGen routine from MixONat used for differentiating methyl, methylene, methine or quaternary $\delta_{C\text{-}SDF}$ to get DBs compatible with MixONat [25,26]. Comparing simulated with experimental data (^{13}C NMR, DEPT-135 and 90), MixONat then ranked associated structures with a score from 0 to 1 (where 1 indicates a full match and 0 the absence of any similarity between simulated and experimental spectra). Finally, experimental data of highest ranked candidates were compared with the literature in order to confirm proposed structural assignments.

2.12. Advanced glycation end-products inhibition assay

Experiments were conducted according to a previously described protocol [14,15]. In short: Stock solutions (SS) of standard mixtures (M1-M6) and *G. parvifolia* bark extracts were prepared in DMSO at a concentration of 10 mg/mL, while single compounds were dissolved at a concentration of 30 mM. These SS were then diluted with 50 mM phosphate buffer (pH: 7.4) yielding working solutions (WS) at a concentration range of 10^{-2} to 10 mg/mL for standard mixtures and bark extracts. The WS of single compounds were prepared at a concentration range of 3×10^{-2} to 30 mM. Next 10 μL of each WS were deposited in 96 black well bottom plates (Fisher Scientific, Illkirch, France) and mixed with 90 μL of a solution containing bovine serum albumine (11 mg/L, Sigma Aldrich, St Quentin Fallavier, France), D-ribose (0.25 M, Acros

Organics, Geel, Belgium), and phosphate buffer (50 mM, NaN_3 0.02%, pH 7.4). Plates were then incubated for 24 h at 37 °C, before being fluorometrically analyzed (λ_{exc}: 335 nm, λ_{em}: 385 nm) on an Infinite M200 plate reader (Tecan, Lyon, France). The formation of AGEs was calculated according to the following formula (FI = fluorescence intensity):

$$\frac{[FI\ (BSA + ribose + sample) - FI\ (BSA + sample)]\ \text{x}\ 100}{FI(BSA + ribose) - FI(BSA)}$$

Aminoguanidine (Sigma Aldrich, St Quentin Fallavier, France) was used as positive, a mixture of BSA + ribose as negative control. A BSA solution served as blank. Results were expressed as IC_{50} values in μg/mL (mixtures and extracts) or μM (single compounds).

All experiments except for KL5259 (insufficient amount of material) were conducted in triplicate. For IC_{50} calculations dose-effect curves were best fitted with a sigmoidal dose-response equation using the Sigma Plot 12.0 software (Systat, Delaware, USA).

3. Results and discussion

3.1. Development of a LDI-MS based chemometric model using defined mixtures of reference compounds

In order to create a working statistical model, six experimental mixtures (M1-M6) (Fig. S2, supplementary material) comprising different quantities of four structurally diverse reference compounds were used. All compounds were detectable by LDI-MS (Fig. S3, supplementary material) and showed very different inhibitory effects on AGEs formation (Fig. 1). Highest activity was observed for the biflavonoid amentoflavone (**1**) (IC_{50}: 23 ± 2 μM), followed by ferulic acid (**3**) (111 ± 12 μM), and 3-hydroxyxanthone (**4**) (124 ± 14 μM), while the IC_{50} of the triterpene betulinic acid (**2**) exceeded 1.0 mM.

Interestingly, M5, which exhibited the best activity of all mixtures, did not contain the highest amount of **1** (M2) (Table 1), which was the most active among all single compounds. Moreover, four mixtures M5, M2, M6, and M1 all exhibited rather similar IC_{50} values within a range of 26–38 μg/mL despite their very different relative composition. The situation is quite similar to what is often observed for complex mixtures of NPs such as crude (plant) extracts. These data, however, do not reveal the samples' underlying active principle without further statistical analysis. For that reason, PLS regression analysis [2] was selected as statistical model to reprocess these data. Unlike PCA, which solely permits the grouping of sample material according to its chemical composition [27], PLS may further include (bio)activity data. The method correlates independent [e.g. spectral data (X)], with dependent variables [activity data, response variable (H)] as well as variables between themselves [28]. Results from this analysis are summarized in Fig. 2.

Highlighting the good repeatability of LDI-MS, all samples were correctly grouped according to their specific composition (Fig. 2a). Samples of highest activities (M5, M2, M6, M1, IC_{50}: 26–38 μM) systematically showed up on the right side of the PLS scores plot, while less active M4 and M3 were found on the left. Most active M5 was shifted to the upper-right corner far off the rest of the samples.

The PLS loadings plot (Fig. 2b) then allowed estimating the specific contribution of individual reference compounds to observed anti-AGEs activities of experimental mixtures. According to current results, the farther a marker ion (diamonds) is shifted from the center towards the right side of the graph, the higher its potential contribution to anti-AGEs effects. Moreover, the closer the distance of a marker ion to a specific mixture, the higher its correlation to this sample and its activity. Several marker ions were found in close vicinity of specific mixtures. For example, **4** (211 m/z, IC_{50}: 110.7 μM, upper-right corner) is strongly correlated to M5.

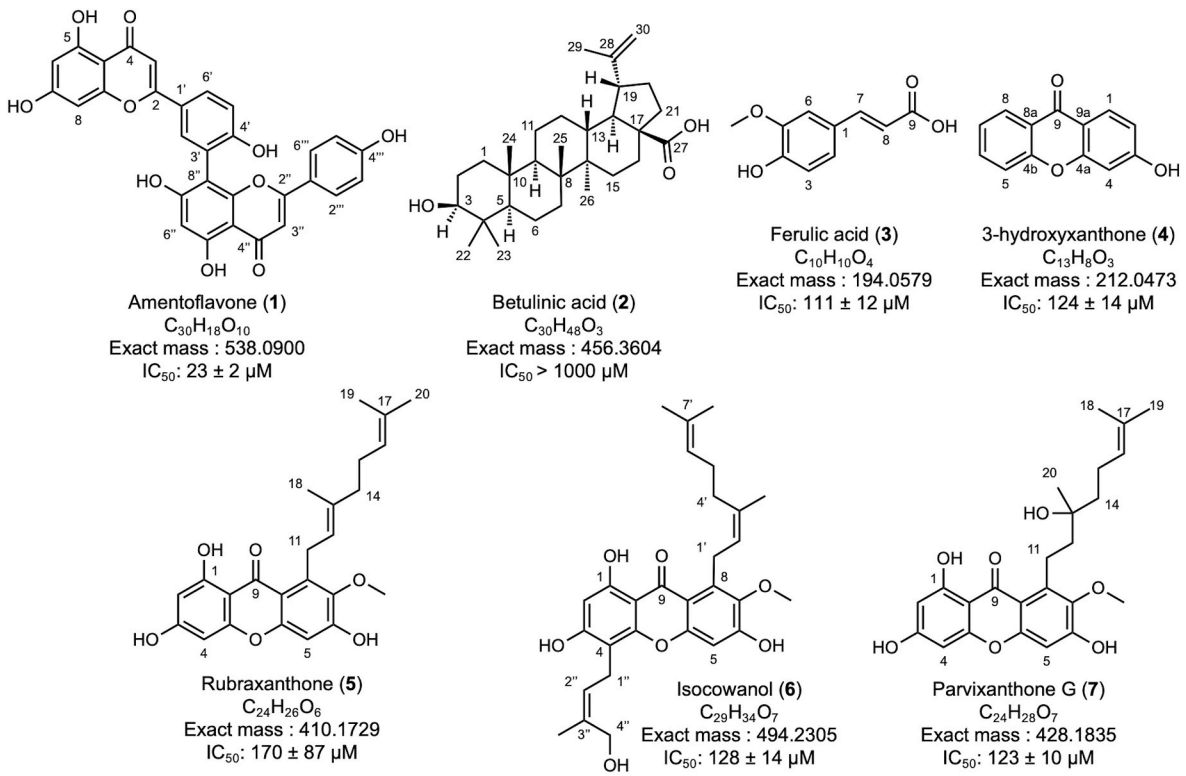


Fig. 1. Structures and anti-AGEs activities of reference (**1**–**4**) and isolated (**5**–**7**) compounds amentoflavone (**1**), betulinic acid (**2**), ferulic acid (**3**), 3-hydroxyxanthone (**4**), rubraxanthone (**5**), isocowanol (**6**) and parvixanthone G (**7**). Averaged IC_{50} values and standard deviations were calculated from three independent experiments. All compounds were detected as quasi-molecular ions $[M - H]^-$ in LDI and ESI-MS. An exhaustive list of theoretical and detected masses is provided in the supplementary material (Table S2).

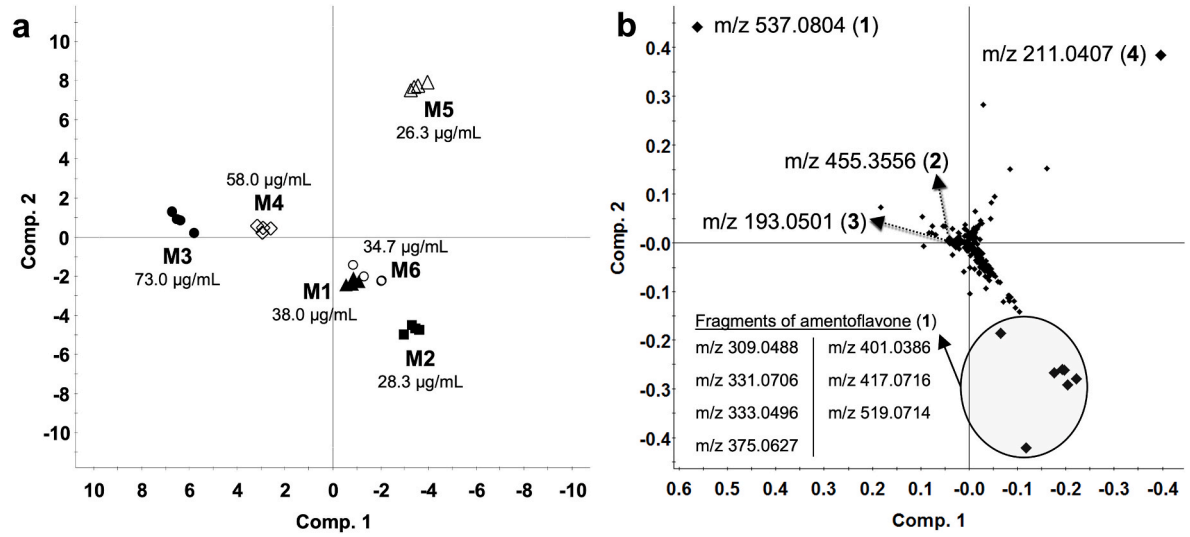


Fig. 2. Partial least square scores plot (**a**) of mixtures M1 to M6 and their anti-AGEs IC_{50} activities. Aminoguanidine was used as positive control (IC_{50}: 139 ± 21 µg/mL). Overall, four principal components accounted for 95.1% explained variations (1st component: 54.18%; 2nd component: 23.37%; 3rd component: 15.66%; 4th component: 1.89%). The PLS loadings plot (**b**) shows variables of each mixture and molecular ions of reference compounds **1**–**4** are marked as large diamonds and were all detected as $[M - H]^-$. Large diamonds within the cycle correspond to fragmentation ions of **1**. The rest of the variables (small diamonds) correspond to fragmentation products of compounds **2**–**4**.

On the other hand, most active **1** (537 m/z, IC_{50}: 22.7 µM) is negatively correlated to M2, which on first sight appears contradictory, as M2 contains the highest quantity of **1** of all mixtures. However, **1** easily fragments in LDI producing fragmentation ions of much higher intensities than the initial non-fragmented compound. These ions are marked by a circle in Fig. 2b and correlate very well with M2. Consequently, PLS loadings plots must be always carefully examined with respect to potentially occurring LDI-induced fragmentation.

Compound **3** (193 m/z, IC_{50}: 110.7 µM) is correlated to M4, which is in line with the fact that the compound exhibits highest concentration in this specific sample (Table 1). For samples M1, M3, M4 and M6, activity correlations to **1**–**4** were generally weak, proposing synergistic effects to explain the observed anti-AGEs effects.

Overall, the LDI-MS based PLS loadings plot quickly identified **1** as a major activity marker, despite its strong in-source fragmentation. With regards to unknown samples, the latter may, however, easily lead to misinterpretation of results. Moreover, high resolution LDI-MS provides exact molecular formulas, but it does not allow final structural assignments. Therefore, ^{13}C NMR dereplication may assist the process and provide essential hints for conclusive structure elucidation of major constituents in multi-compound mixtures. In the present case ^{13}C NMR spectra (^{13}C NMR, DEPT-135 and 90) were recorded for M1 (Figs. S4–S12, supplementary material) and consecutively processed by the MixONat dereplication software [16]. Based on ^{13}C NMR data of 718 compounds compiled in the Garcinia DB, MixONat correctly predicted the presence of all four reference compounds with the following matching scores: 0.97 for amentoflavone (**1**, rank 3), 0.87 for betulinic acid (**2**, rank 2), 1.00 for ferulic acid (**3**, rank 1), and 0.92 for 3-hydroxyxanthone (**4**, rank 4) (Fig. S14, supplementary material). The dereplication analysis was completed within 30 min and experimental data were in line with previous reports (Table S3, supplementary material) [29–32].

3.2. The real-life example of Garcinia parvifolia

In a next step the combined LDI-MS/^{13}C NMR approach was extended to the chemometric analysis of four bark extracts (KL5670, KL5248, KL5073 and KL5259) of *G. parvifolia*, which all exhibited notable anti-AGEs effects (IC_{50} 79–174 µg/mL, Fig. 3). Preceding analyses have further revealed that all extracts comprise significant amounts of xanthones [33]. The latter show close structural similarities to dithranol, an anthraquinone, which is used as a commercial matrix in matrix assisted laser desorption ionization (MALDI) (Fig. S15, supplementary material). Therefore, xanthones should easily ionize upon laser irradiation. All extracts were concurrently analyzed by LDI-MS and UPLC-MS^2, and LDI-MS spectra as well as UPLC-MS chromatograms are show in Fig. S16 and Fig. S17 respectively (supplementary material). A direct comparison of LDI-MS and UPLC-MS^2 based PLS scores plots presented in Fig. 3a and **c** shows that both methods worked equally well for the differentiation of sample material according to its chemical composition. Overall, three principal components accounted for 98.63% (LDI-MS) and 99.91% (UPLC-MS) of explained variations and 406 (LDI-MS) as well as 7726 (UPLC-MS^2) marker ions were detected. The PLS loadings plot for LDI-MS (Fig. 3b) and UPLC-MS^2 (Fig. 3d) further revealed that most important activity markers (large diamonds) were commonly detected by the two methods. Noteworthy, while LDI-MS detected one marker ion at 427.1778 m/z (Fig. 3b), four ions with practically the same mass (427.1756, 427.1757, 427.1758 and 427.1760 m/z) all accounting for $C_{24}H_{28}O_7$ were observed by UPLC-MS^2 (Fig. 3d). This strongly suggests the presence of isomers. The latter can be differentiated by UPLC-MS^2 but not LDI-MS due to the absence of chromatographic separation. Nevertheless, both methods linked activity markers at 409, 427 and 493 m/z to KL5670, as well as 531 m/z to KL5073 (Fig. 3b and d).

Considering that KL5670 (IC_{50}: 79.0 ± 7.0 µg/mL) exhibited best anti-AGEs effects of all extracts, the sample was further processed by flash chromatography in order to refine its chemometric analysis. Overall, fourteen subfractions were obtained, evaluated on their anti-AGEs activities, and consecutively analyzed by LDI-MS and UPLC-MS^2. The PLS analyses of these experiments are summarized in Fig. 4. As previously observed for crude extracts, LDI-MS and UPLC-MS^2 yielded a very similar grouping of all subfractions in PLS scores plots (Fig. 4a and c). Moreover, loadings plots of both methods (Fig. 4b & d) commonly depicted previously identified activity markers 427 and 493 m/z next to F10 (IC_{50} : 82 ± 18.1 µg/mL), as well as 409 m/z next to F6 (IC_{50}: 58 ± 6.6 µg/mL). In order to obtain further structural information about these compounds, F6 and F10 were analyzed by ^{13}C NMR and MixONat [16,22,32]. Next to comparing simulated with reported ^{13}C NMR shifts, the program further comprises a mass filter, which was set to 410 Da when studying F6. MixONat then ranked rubraxanthone (**5**) (hit score

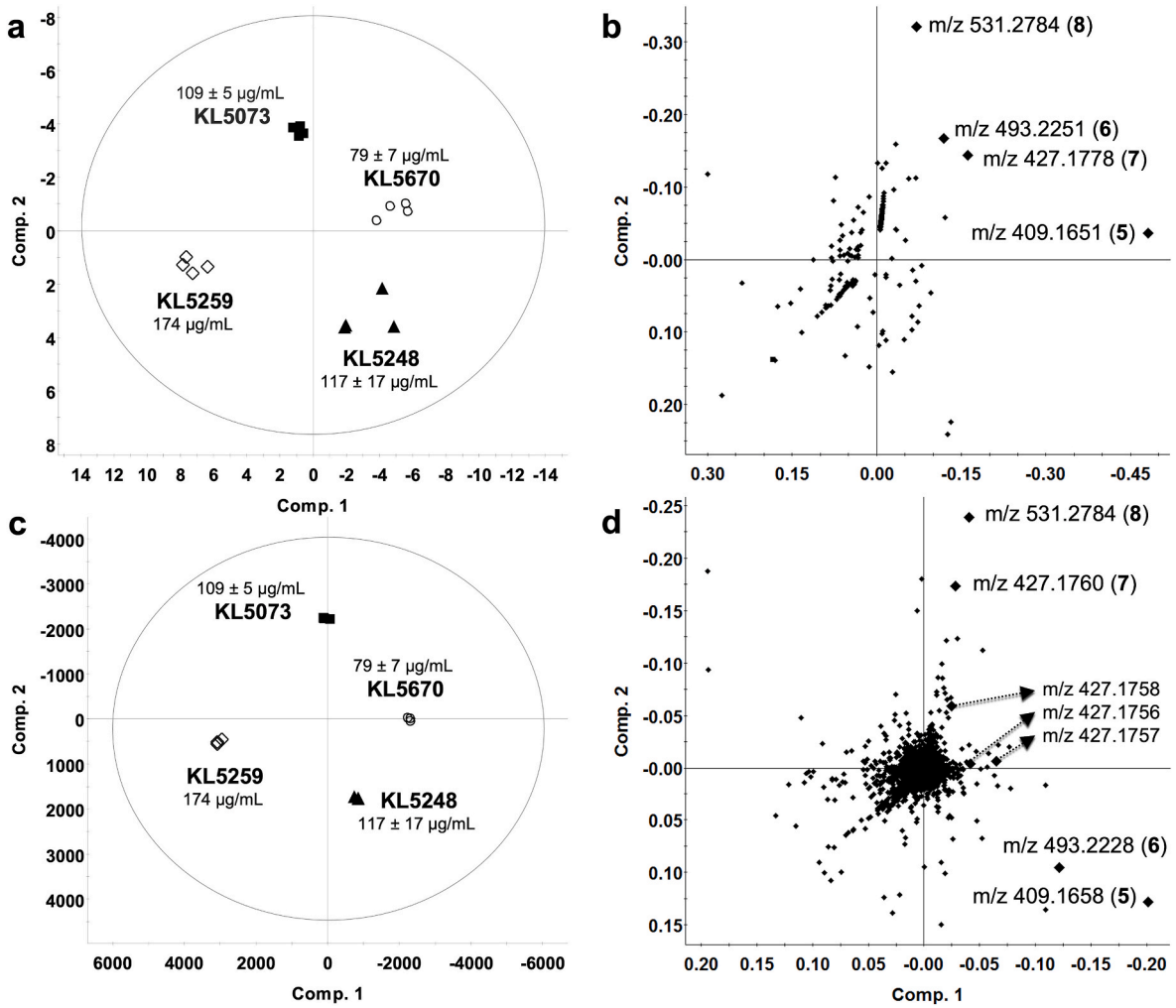


Fig. 3. Partial least squares regression analysis of four *G. parvifolia* bark extracts (KL5670, KL5073, KL5259 and KL5248). Both LDI-MS (**a**) and UPLC-MS^2 (**c**) scores plots yielded comparable group separation. Overall, three principal components accounted for 98.63% (LDI-MS) and 99.91% (UPLC-MS^2) of explained variations [1st component: 82.23% (LDI); 94.15% (UPLC-MS^2), 2nd component: 13.88%; 5.64%, and 3rd component: 2.52%; 0.12%]. All data showed up within the Hoteling's 95% confidence ellipse. The PLS loadings plots [LDI-MS (**b**) and UPLC-MS^2 (**d**)] proposed identical signals for most prominent activity markers (large diamonds). Based on this information, compounds **5**, **6** and **7** were consecutively isolated and tested (Fig. 1).

92%) first of the 718 NPs found in the Garcinia DB database. Results were then crosschecked with experimental NMR data from the literature (Fig. S18–S22, Table S4, supplementary material) and further confirmed by ProgenesisQI assisted UPLC-MS^2 data processing. There **5** was ranked in 6th place with a matching score of 55.1% out of 211,000 entries of PNMRNP database (Table S5, supplementary material).

For F10 the mass filter was set to 428 and 494 Da and MixONat proposed isocowanol (**6**) and parvixanthone G (**7**) at rank 1 and 3 with hit scores of 86 and 83% respectively (Fig. S23-S34, Table S6, supplementary material). The ProgenesisQI analysis ranked **6** and **7** in positions 4 and 2 with matching scores of 56.1 and 58.1 respectively (Table S7-S8, supplementary material). Finally, both compounds (**6**, 46.5 mg and **7**, 15.4 mg) were isolated by preparative HPLC and analyzed by NMR for final structure confirmation. Results are summarized in the supplementary material (for **6**: Fig. S35-41, Table S7,S9; for **7**: Fig. S34 and Table S6, S8). All isolated marker compounds (**5**–**7**) exhibited notable anti-AGEs activities in the range 123–170 µM (Fig. 1), confirming the proposed LDI-MS/^{13}C NMR based chemometric model.

3.3. ^{13}C NMR assisted LDI-MS complementing UPLC-MS^2 based chemometrics

The example of *G. parvifolia* showed that LDI-MS may yield comparable results to UPLC-MS^2 for the chemical and statistical differentiation of complex mixtures of NPs, provided target analytes ionize well upon laser irradiation. In such a case, LDI-MS may offer several advantages over LC based methods of mass spectrometry: LDI-MS experiments are completed within a few seconds instead of some tens of minutes (LC-MS), hardly require any sample preparations, and are not limited by solvent restriction. On the other hand, in-source fragmentation my complicate the correct interpretation of LDI-MS spectra, a problem that is also known for ESI sources. In absence of chromatographic separation, LDI-MS does not permit the differentiation of isomers, which is exemplified by activity marker ions at 427 m/z in Fig. 3b and d. Recent technical advances such as the development of ion mobility mass spectrometry may help bypassing this limitation in the future [34]. At a sufficiently high concentration within sample material, some isomers may further be differentiated ^{13}C NMR.

Compared to LDI-MS, LC-MS^2 generally facilitates the detection of more constituents and compound specific retention times as well as MS^2 data provide additional discriminatory factors for statistical analysis. On the other hand, both methods may detect different compounds [35] within the same sample, some of which exclusively ionizing in LDI but not in ESI sources [5]. This clearly makes both methods complementary.

Eventually current experiments have shown that LDI-MS data may be easily integrated in multivariate statistics such as PLS, which is a core

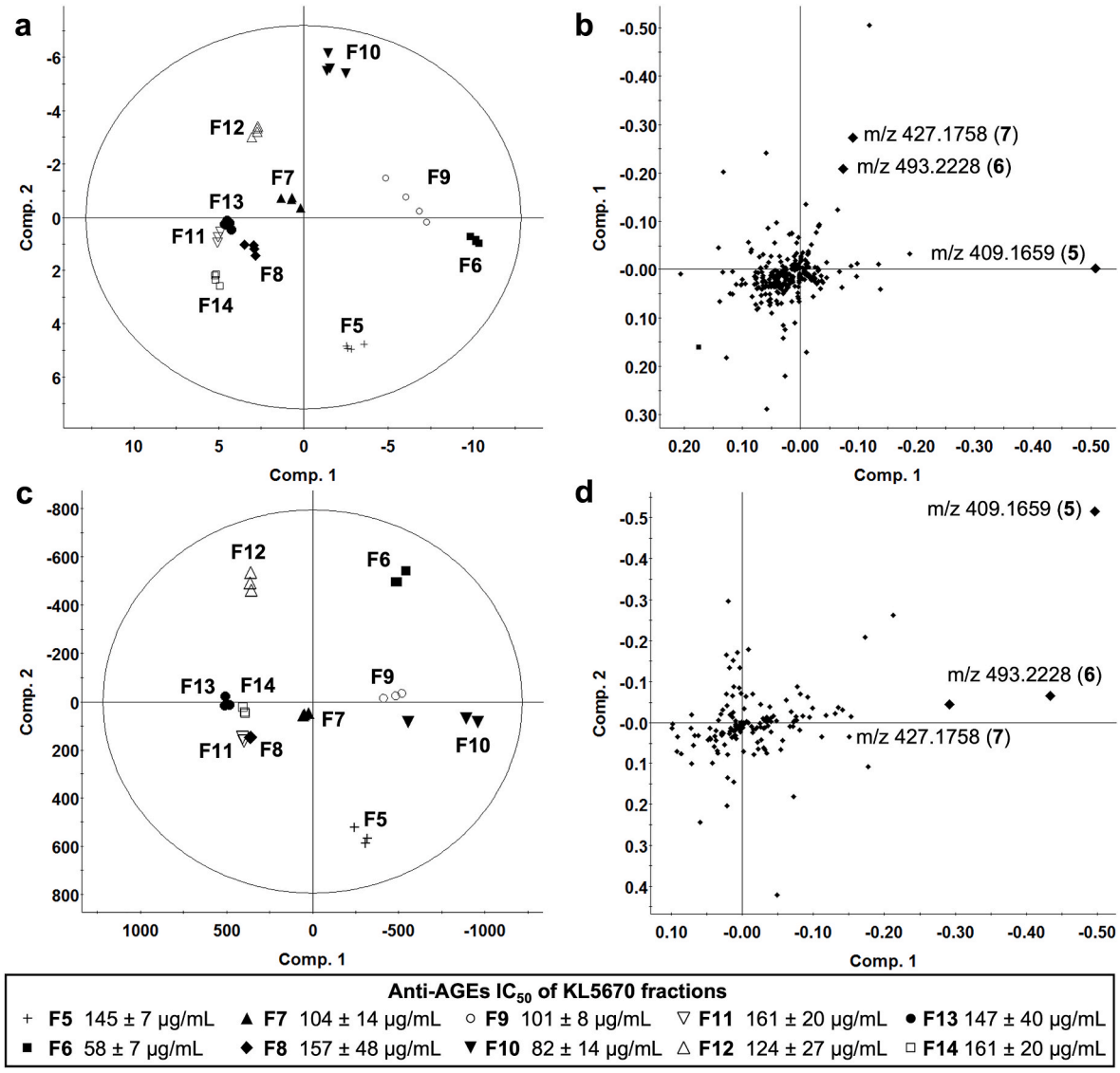


Fig. 4. Partial least square analysis of most active subfraction (F5-F14) obtained after the separation of KL5670 by flash chromatography. As previously observed for crude bark extracts, both LDI-MS (**a**) and UPLC-MS^2 (**c**) yielded very similar sample grouping of all analyzed samples (scores plots). The PLS loadings plots (**b**, **d**) of both methods mutually linked activity markers isocowanol (**6**) and parvixanthone G (**7**) to F10 and rubraxanthone (**5**) to F6. Overall, three principal components accounted for 98.63% (LDI-MS) and 99.91% (UPLC-MS^2) of explained variations [1st component: 82.23% (LDI); 94.15% (UPLC-MS^2), 2nd component: 13.88%; 5.64%, and the 3rd component: 2.52%; 0.12%)]. All data showed up within the Hoteling's 95% confidence ellipse.

technique in applied chemometrics. Moreover, predicted activity markers, once isolated, actually showed notable anti-AGEs effects and consequently confirmed the applied statistical model. Next to identifying common activity markers, scores plots (Fig. 3b,d and 4b,d) may further highlight those constituents that are selectively detected by LDI or LC-MS. Consequently, a combined LC-LDI-MS approach may largely extend the range of potential activity markers identified by chemometrics. Provided sufficient material is available material, ^{13}C NMR dereplication tools like MixONat may further assist the process by providing structural evidence for these markers in complex mixtures prior to their isolation. Overall, the presented work highlights the complementarity of LC-MS^2, LDI-MS and ^{13}C NMR as a versatile and holistic approach towards chemometrics in NPs research and other areas of analytical chemistry.

Credit roles

Manon Meunier: Investigation, Methodology, Writing - Original Draft, Visualization. Dimitri Bréard: Investigation. Khalijah Awang: Resources. Séverine Boisard: Formal analysis, Investigation. David Guilet: Funding acquisition, Validation. Pascal Richomme: Visualization. Séverine Derbré: Conceptualization, Supervision, Writing - Review & Editing. Andreas Schinkovitz: Conceptualization, Supervision, Writing - Review & Editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.

org/10.1016/j.talanta.2022.123998.

References

[1] J. Lever, M. Krzywinski, N. Altman, Principal component analysis, Nat. Methods 14 (2017) 641–642, https://doi.org/10.1038/nmeth.4346.

[2] D.V. Nica, D.M. Bordean, I. Pet, E. Pet, S. Alda, I. Gergen, A novel exploratory chemometric approach to environmental monitoring by combining block clustering with partial least square (PLS) analysis, Chem. Cent. J. 7 (2013) 145, https://doi.org/10.1186/1752-153X-7-145.

[3] J.J. Kellogg, D.A. Todd, J.M. Egan, H.A. Raja, N.H. Oberlies, O.M. Kvalheim, N. B. Cech, Biochemometrics for natural products research: comparison of data analysis approaches and application to identification of bioactive compounds, J. Nat. Prod. 79 (2016) 376–386, https://doi.org/10.1021/acs.jnatprod.5b01014.

[4] C. Aydoğan, Recent advances and applications in LC-HRMS for food and plant natural products: a critical review, Anal. Bioanal. Chem. 412 (2020) 1973–1991, https://doi.org/10.1007/s00216-019-02328-6.

[5] P. Le Pogam, A. Schinkovitz, B. Legouin, A.C. Le Lamer, J. Boustie, P. Richomme, Matrix-free UV-laser desorption ionization mass spectrometry as a versatile approach for accelerating dereplication studies on lichens, Anal. Chem. 87 (2015) 10421–10428, https://doi.org/10.1021/acs.analchem.5b02531.

[6] L. Zhu, S. Yang, G. Li, X. Zhang, J. Yang, X. Lai, G. Yang, Simultaneous analysis of tocopherols, tocotrienols, phospholipids, γ-oryzanols and β-carotene in rice by ultra-high performance liquid chromatography coupled to a linear ion trap-orbitrap mass spectrometer, Anal. Methods 8 (2016) 5628–5637, https://doi.org/10.1039/C6AY00556J.

[7] R. Zenobi, R. Knochenmuss, Ion formation in MALDI mass spectrometry, Mass Spectrom. Rev. 17 (1998) 337–366, https://doi.org/10.1002/(SICI)1098-2787.

[8] F. Chemat, M. Abert-Vian, A.S. Fabiano-Tixier, J. Strube, L. Uhlenbrock, V. Gunjevic, G. Cravotto, Green extraction of natural products. Origins, current status, and future challenges, Trends Anal. Chem. 118 (2019) 248–263, https://doi.org/10.1016/j.trac.2019.05.037.

[9] M. Skopikova, M. Hashimoto, P. Richomme, A. Schinkovitz, Matrix-free laser desorption ionization mass spectrometry as an efficient tool for the rapid detection of opiates in crude extracts of Papaver somniferum, J. Agric. Food Chem. 68 (2020) 884–891, https://doi.org/10.1021/acs.jafc.9b05153.

[10] P. Le Pogam, P. Richomme, M.A. Beniddir, T.H. Duong, G. Bernadat, A. Schinkovitz, A thorough evaluation of matrix-free laser desorption ionization on structurally diverse alkaloids and their direct detection in plant extracts, Anal. Bioanal. Chem. 412 (2020) 7405–7416, https://doi.org/10.1007/s00216-020-02872-6.

[11] A. Schinkovitz, S. Boisard, I. Freuze, J. Osuga, N. Mehlmer, T. Brück, P. Richomme, Matrix-free laser desorption ionization mass spectrometry as a functional tool for the analysis and differentiation of complex phenolic mixtures in propolis: a new approach to quality control, Anal. Bioanal. Chem. 410 (2018) 6187–6195, https://doi.org/10.1007/s00216-018-1225-1.

[12] N.A. dos Santos, L.M. de Souza, F.E. Pinto, C. de J. Macrino, C.M. de Almeida, B. B. Merlo, P.R. Filgueiras, R.S. Ortiz, R. Mohana-Borges, W. Romão, LDI and MALDI-FT-ICR imaging MS in Cannabis leaves: optimization and study of spatial distribution of cannabinoids, Anal. Methods 11 (2019) 1757–1764, https://doi.org/10.1039/C9AY00226J.

[13] S. Islam, R. Alam, S. Kim, Improved coverage of plant metabolites using powder laser desorption/ionization coupled with Fourier-transform ion cyclotron mass spectrometry, Food Chem. 373 (2022), 131541, https://doi.org/10.1016/j.foodchem.2021.131541.

[14] S. Derbré, J. Gatto, A. Pelleray, L. Coulon, D. Séraphin, P. Richomme, Automating a 96-well microtiter plate assay for identification of AGEs inhibitors or inducers: application to the screening of a small natural compounds library, Anal. Bioanal. Chem. 398 (2010) 1747–1758, https://doi.org/10.1007/s00216-010-4065-1.

[15] L. Séro, L. Sanguinet, P. Blanchard, B. Dang, S. Morel, P. Richomme, D. Séraphin, S. Derbré, Tuning a 96-well microtiter plate fluorescence-based assay to identify AGE inhibitors in crude plant extracts, Molecules 18 (2013) 14320–14339, https://doi.org/10.3390/molecules181114320.

[16] A. Bruguière, S. Derbré, J. Dietsch, J. Leguy, V. Rahier, Q. Pottier, S. Suor-Cherer, G. Viault, MixONat, a software for the dereplication of mixture based on ^{13}C NMR spectroscopy, Anal. Chem. 92 (2020) 8793–8801, https://doi.org/10.1021/acs.analchem.0c00193.

[17] X. Shi, W. Yang, S. Qiu, J. Hou, W. Wu, D. Guo, Systematic profiling and comparison of the lipidomes from Panax ginseng , P. quinquefolius , and P. notoginseng by ultrahigh performance supercritical fluid chromatography/high-resolution mass spectrometry and ion mobility-derived collision cross section measurement, J. Chromatogr. A 1548 (2018) 64–75, https://doi.org/10.1016/j.chroma.2018.03.025.

[18] M. Lianza, R. Leroy, C. Machado Rodrigues, N. Borie, C. Sayagh, S. Remy, S. Kuhn, J.H. Renault, J.M. Nuzillard, The three pillars of natural product dereplication. Alkaloids from the bulbs of Urceolina peruviana (C. Presl) J.F. Macbr. as a preliminary test case, Molecules 26 (2021) 637, https://doi.org/10.3390/molecules26030637.

[19] M. Strohalm, M. Hassman, B. Košata, M. Kodíček, mMass data miner: an open source alternative for mass spectrometric data analysis, Rapid Commun. Mass Spectrom. 22 (2008) 905–908, https://doi.org/10.1002/rcm.3444.

[20] H. López-Fernández, H.M. Santos, J.L. Capelo, F. Fdez-Riverola, D. Glez-Peña, M. Reboiro-Jato, Mass-up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery, BMC Bioinf. 16 (2015) 318, https://doi.org/10.1186/s12859-015-0752-4.

[21] A. Bruguière, S. Derbré, D. Bréard, F. Tomi, J.M. Nuzillard, P. Richomme, ^{13}C NMR dereplication using MixONat software: a practical guide to decipher natural products mixtures, Planta Med. 87 (2021) 1061–1068, https://doi.org/10.1055/a-1470-0446.

[22] A. Bruguière, S. Derbré, P. Richomme, MixONat, SourceForge, 2020. https://sourceforge.net/projects/mixonat/. (Accessed 14 January 2022).

[23] Dictionary of Natural Products 26.2 Chemical Search, ChemNetBase, 2015. https://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml. (Accessed 1 November 2015).

[24] P.M. Allard, T. Péresse, J. Bisson, K. Gindro, L. Marcourt, V.C. Pham, F. Roussi, M. Litaudon, J.L. Wolfender, Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication, Anal. Chem. 88 (2016) 3317–3323, https://doi.org/10.1021/acs.analchem.5b04804.

[25] A. Bruguière, S. Derbré, C. Coste, M. Le Bot, B. Siegler, S.T. Leong, S.N. Sulaiman, K. Awang, P. Richomme, ^{13}C-NMR dereplication of Garcinia extracts: predicted chemical shifts as reliable databases, Fitoterapia 131 (2018) 59–64, https://doi.org/10.1016/j.fitote.2018.10.003.

[26] L.F. Silva-Castro, S. Derbré, A.M. Le Ray, P. Richomme, K. García-Sosa, L.M. Peña-Rodriguez, Using ^{13}C-NMR dereplication to aid in the identification of xanthones present in the stem bark extract of Calophyllum brasiliense, Phytochem. Anal. 32 (2021) 1102–1109, https://doi.org/10.1002/pca.3051.

[27] E.R. Britton, J.J. Kellogg, O.M. Kvalheim, N.B. Cech, Biochemometrics to identify synergists and additives from botanical medicines: a case study with Hydrastis canadensis (Goldenseal), J. Nat. Prod. 81 (2018) 484–493, https://doi.org/10.1021/acs.jnatprod.7b00654.

[28] L. Ory, E.-H. Nazih, S. Daoud, J. Mocquard, M. Bourjot, L. Margueritte, M. A. Delsuc, J.M. Bard, Y.F. Pouchus, S. Bertrand, C. Roullier, Targeting bioactive compounds in natural extracts - development of a comprehensive workflow combining chemical and biological data, Anal. Chim. Acta 1070 (2019) 29–42, https://doi.org/10.1016/j.aca.2019.04.038.

[29] Z. Qunfang, X. Zhanhui, T. Pengfei, L. Gansun, C. Hongming, A new triterpene from rosemary (Rosmarinus officinalis), J. Chin. Pharmaceut. Sci. 9 (2000) 131–133.

[30] M. Xiang, H. Su, J. Hu, Y. Yan, Isolation, identification and determination of methyl caffeate, ethyl caffeate and other phenolic compounds from Polygonum amplexicaule var. sinense, J. Med. Plants Res. 5 (2011) 1685–1691.

[31] S.H. Goh, I. Jantan, P.G. Waterman, Neoflavonoid and biflavonoid constituents of Calophyllum inophylloide, J. Nat. Prod. 55 (1992) 1415–1420, https://doi.org/10.1021/np50088a005.

[32] E.G.R. Fernandes, A.M.S. Silva, A.S. Cavaleiro, F.M. Silva, M.F.M. Borges, M. M. Pinto, ^{1}H and ^{13}C NMR spectroscopy of mono-, di-, tri- and tetrasubstituted xanthones, Magn. Reson. Chem. 36 (1998) 305–309, https://doi.org/10.1002/(SICI)1097-458X(199804)36:4<305::AID-OMR193>3.0.CO;2-N.

[33] Y.J. Xu, Y.H. Lai, Z. Imiyabir, S.H. Goh, Xanthones from Garcinia parvifolia, J. Nat. Prod. 64 (2001) 1191–1195, https://doi.org/10.1021/np0101393.

[34] K. Masike, M.A. Stander, A. de Villiers, Recent applications of ion mobility spectrometry in natural product research, J. Pharm. Biomed. Anal. 195 (2021), 113846, https://doi.org/10.1016/j.jpba.2020.113846.

[35] G. Petroselli, M.K. Mandal, L.C. Chen, G.T. Ruiz, E. Wolcan, K. Hiraoka, H. Nonami, R. Erra-Balsells, Mass spectrometry of rhenium complexes: a comparative study by using LDI-MS, MALDI-MS, PESI-MS and ESI-MS: LDI, MALDI, PESI and ESI-MS, J. Mass Spectrom. 47 (2012) 313–321, https://doi.org/10.1002/jms.2965.

Matrix free laser desorption ionization assisted by ^{13}C NMR dereplication: A complementary approach to LC-MS^2 based chemometrics

Manon Meunier[1], Dimitri Bréard[1,2], Khalijah Awang[3], Séverine Boisard[1,2], David Guilet[1], Pascal Richomme[1], Séverine Derbré[1,*], Andreas Schinkovitz[1,*]

[1]Univ Angers, SONAS, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy, 16 Bd Daviers, 49045 Angers CEDEX 01, France

[2]University of Malaya, Faculty of sciences, Department of Chemistry, Malaysia

*Corresponding authors

Table of contents

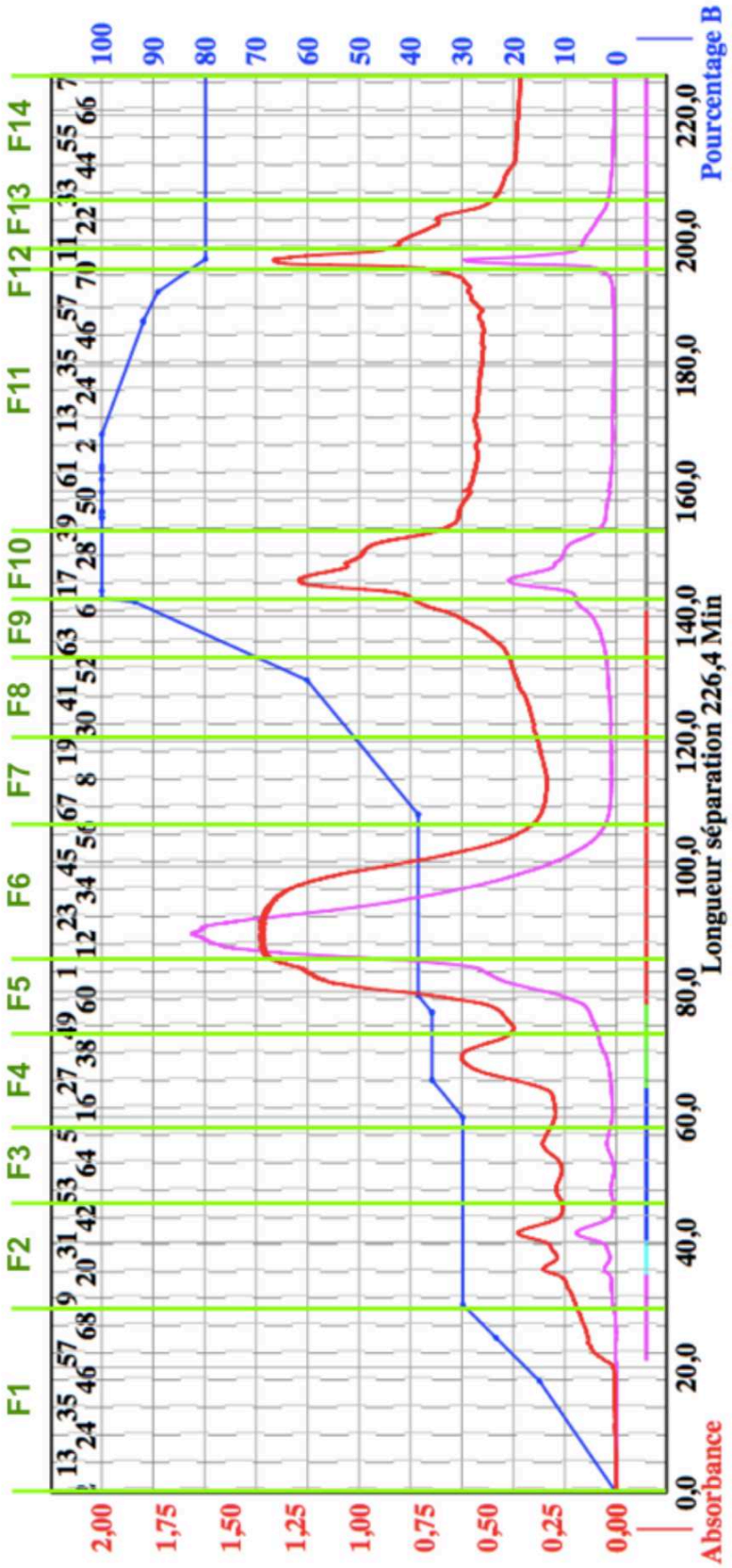


Figure S1. Flash chromatogram of KL5670 recorded at 310 (pink) and 240 nm (red). The mobile phase gradient (solvent B) is marked in blue, fraction cuts are marked in green.

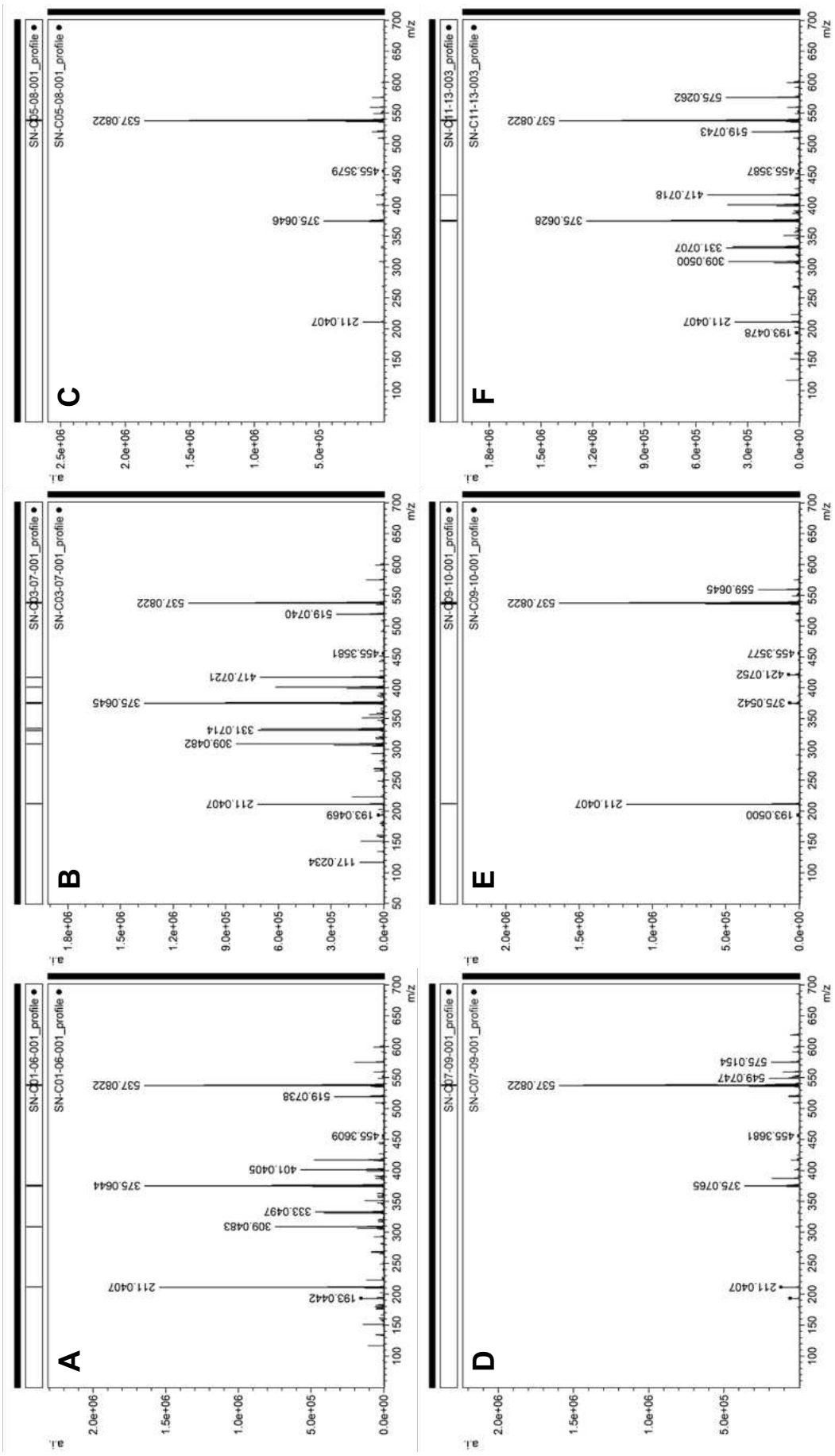


Figure S2. LDI-MS chemical fingerprint of M1 (**A**), M2 (**B**), M3 (**C**), M4 (**D**), M5 (**E**), M6 (**F**).

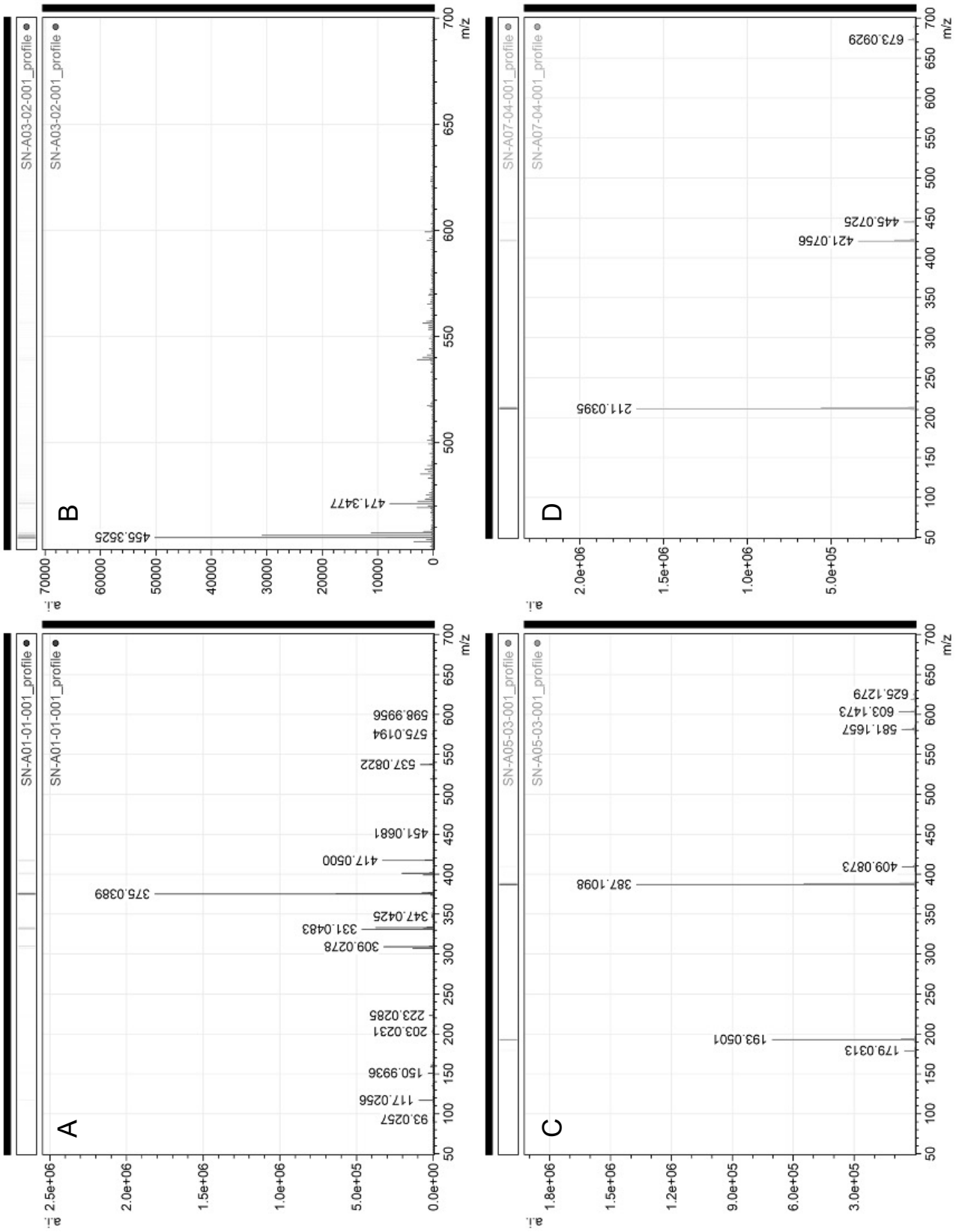


Figure S3. **A**: LDI spectra of reference compounds amentoflavone (**1**) 537.0822 m/z, **B**: betulinic acid (**2**) 455.3525 m/z, **C**: ferulic acid (**3**) 193.0501 m/z and **D**: 3-hydroxyxanthone (**4**) 211.0395 m/z. All compounds were detected as $[M-H]^-$.

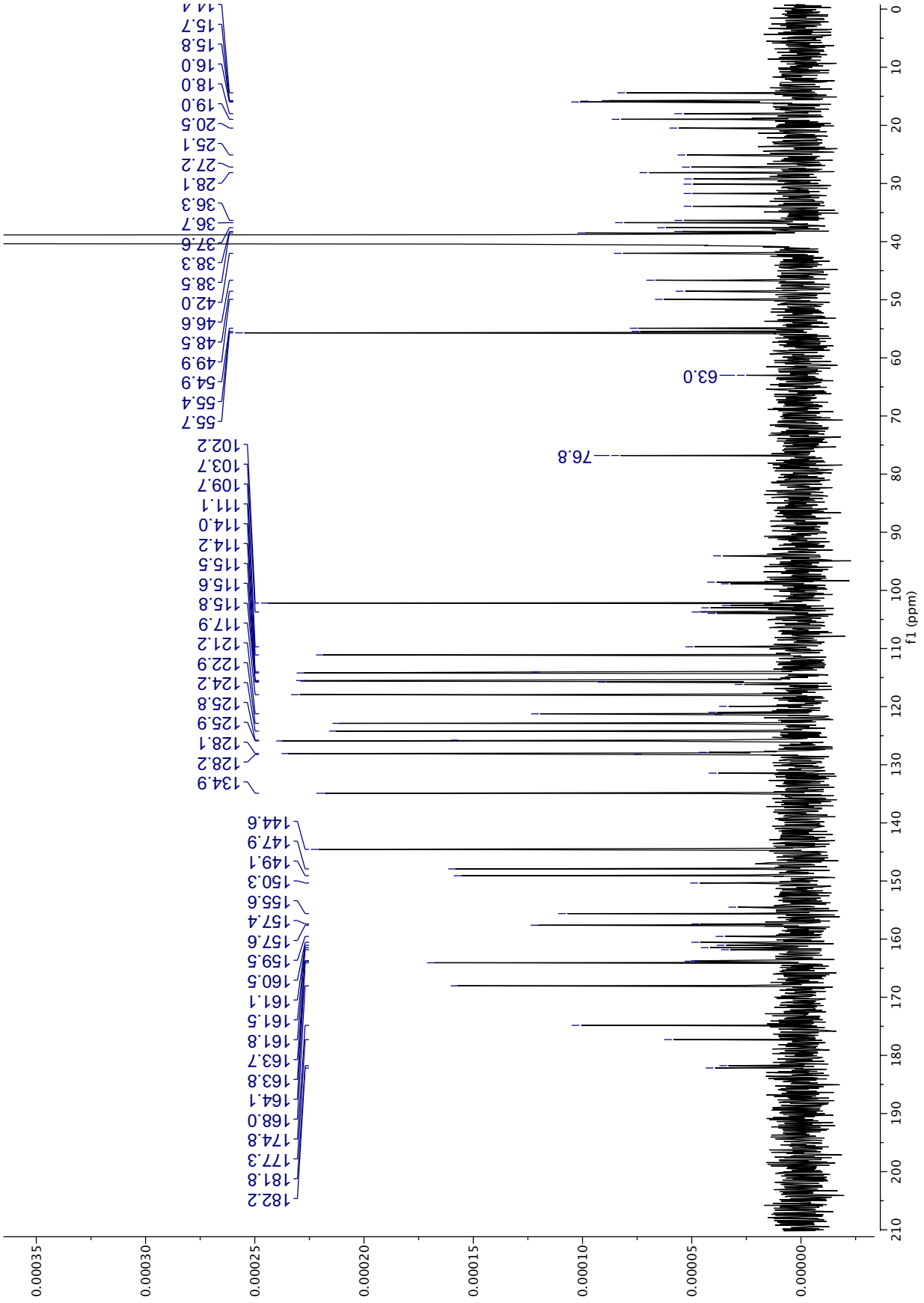


Figure S4. ^{13}C-NMR spectrum (20 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

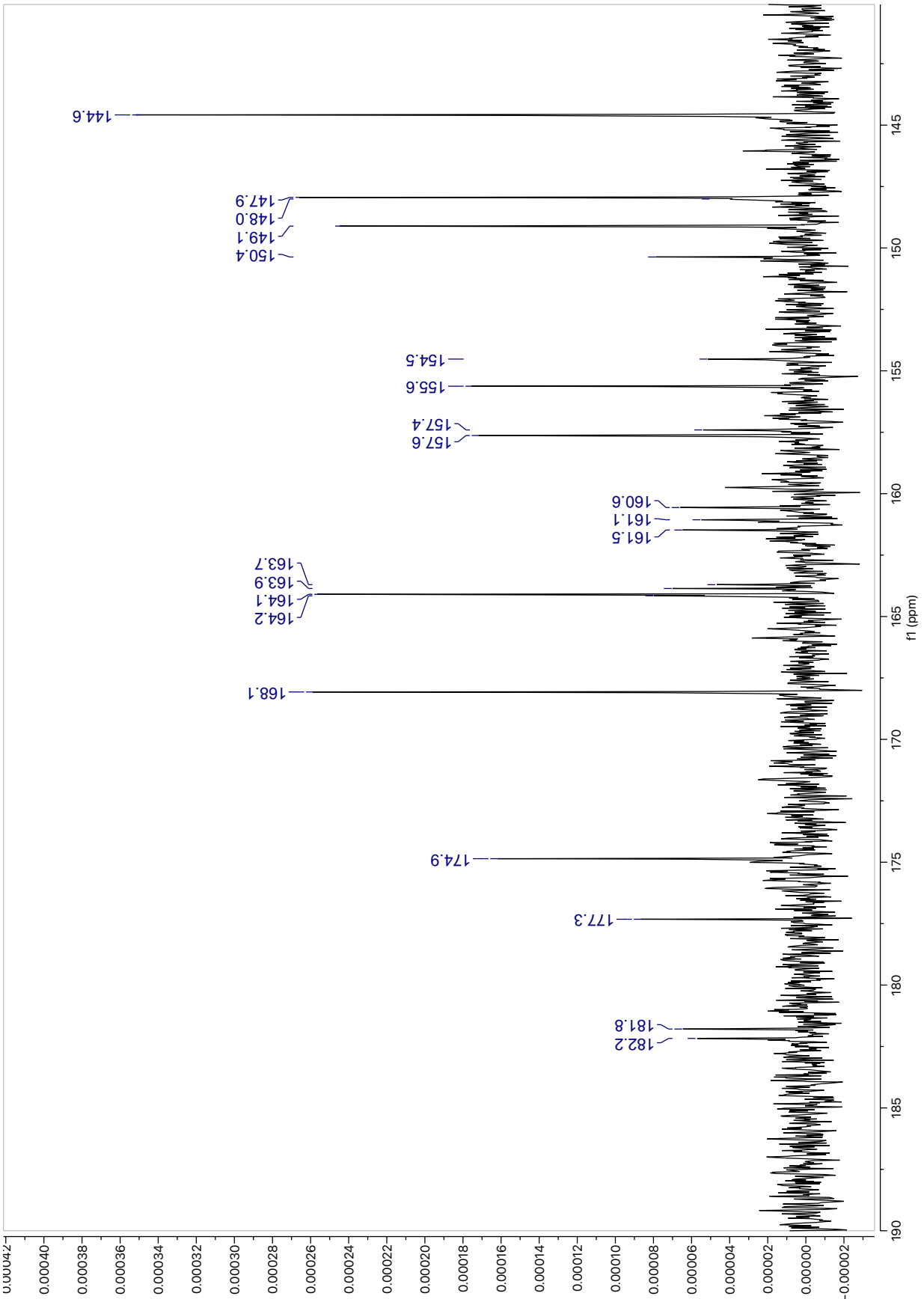


Figure S5. ^{13}C-NMR spectrum (Zoom 1: 190-140ppm) (20 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

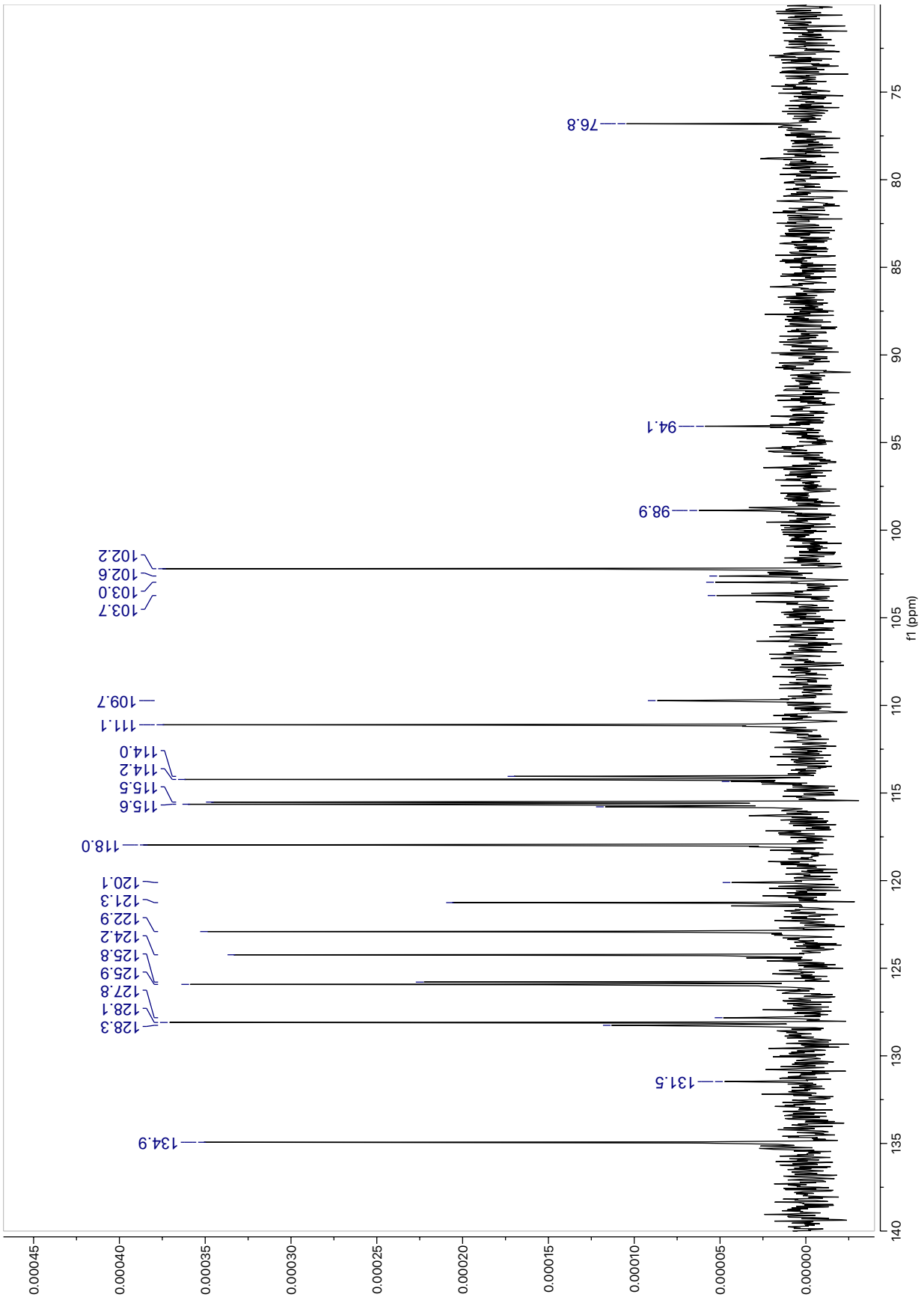


Figure S6. ^{13}C-NMR spectrum (Zoom 2: 140-70ppm) (20 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

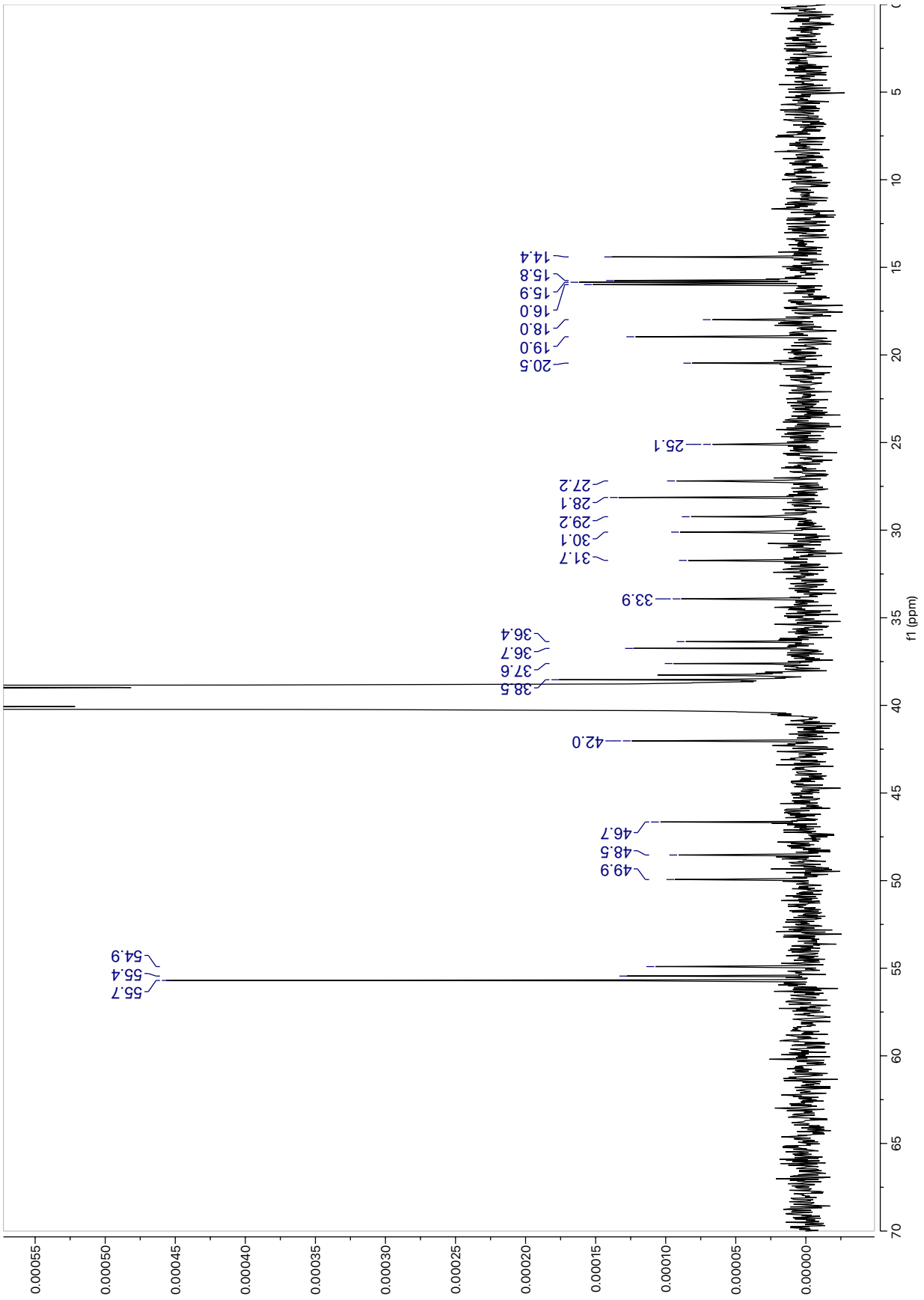


Figure S7. ^{13}C-NMR spectrum (Zoom 3: 70-0ppm) (20 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

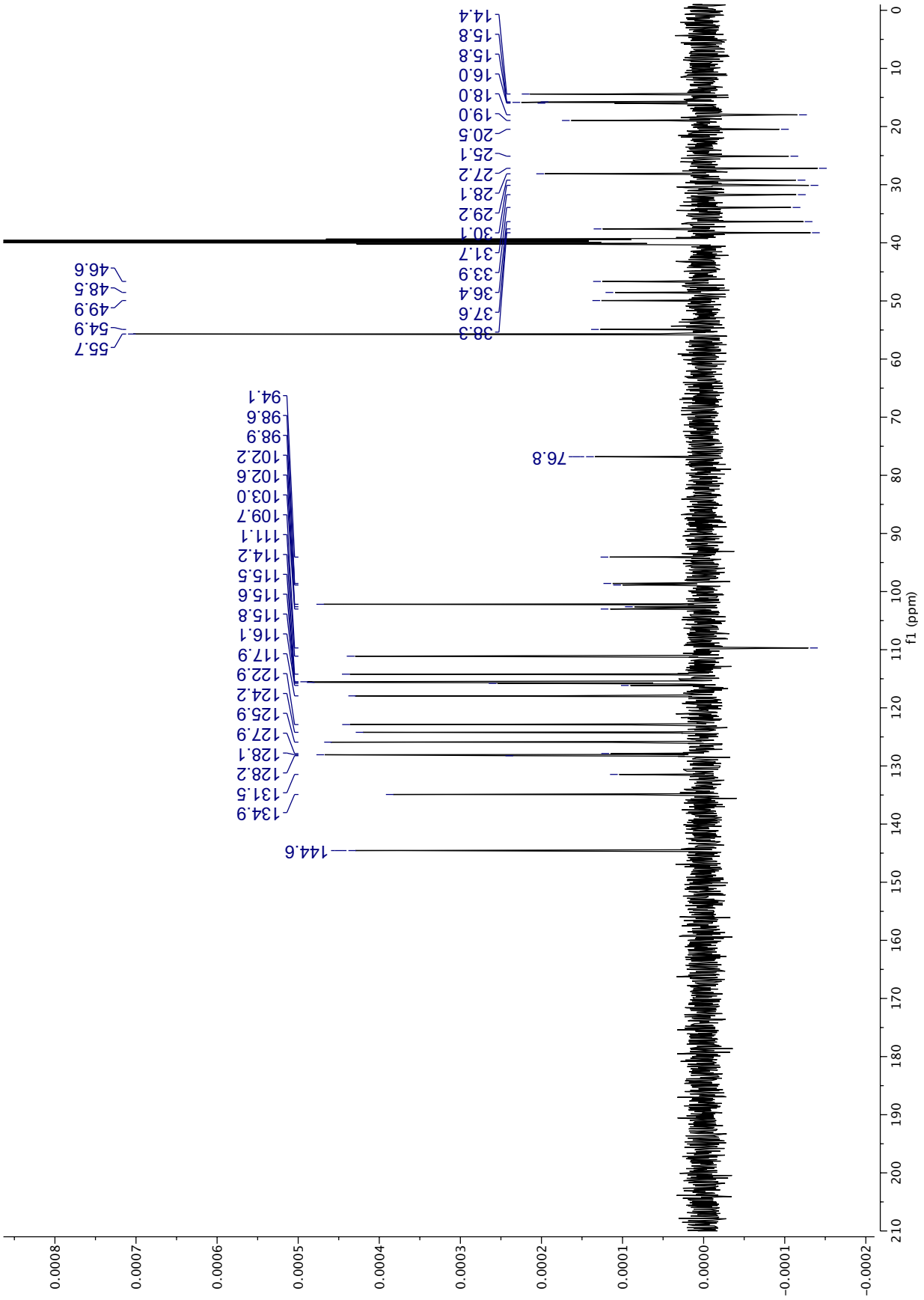


Figure S8. DEPT 135 spectrum (10 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

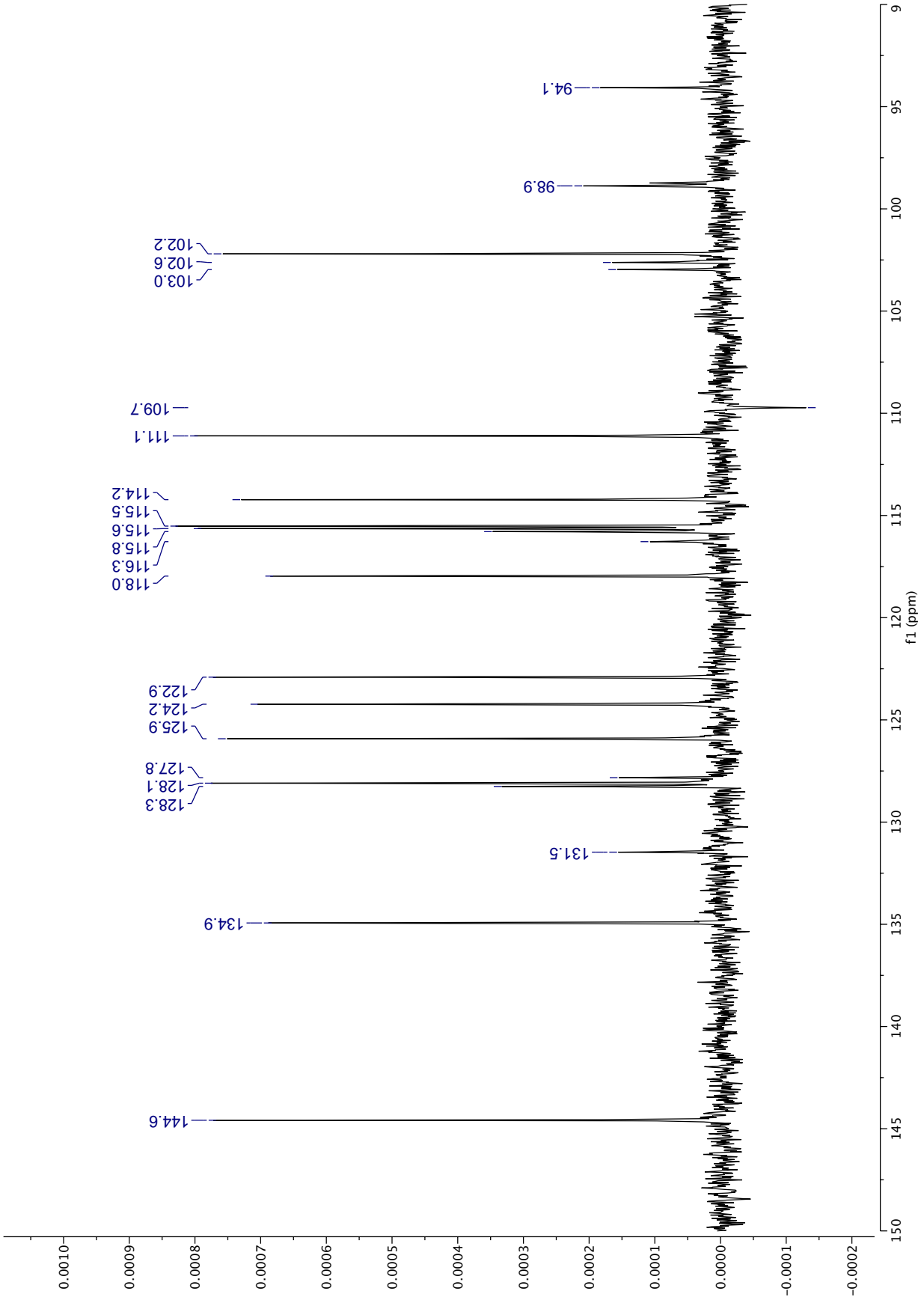


Figure S9. DEPT 135 spectrum (Zoom 1: 150-90ppm) (10 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

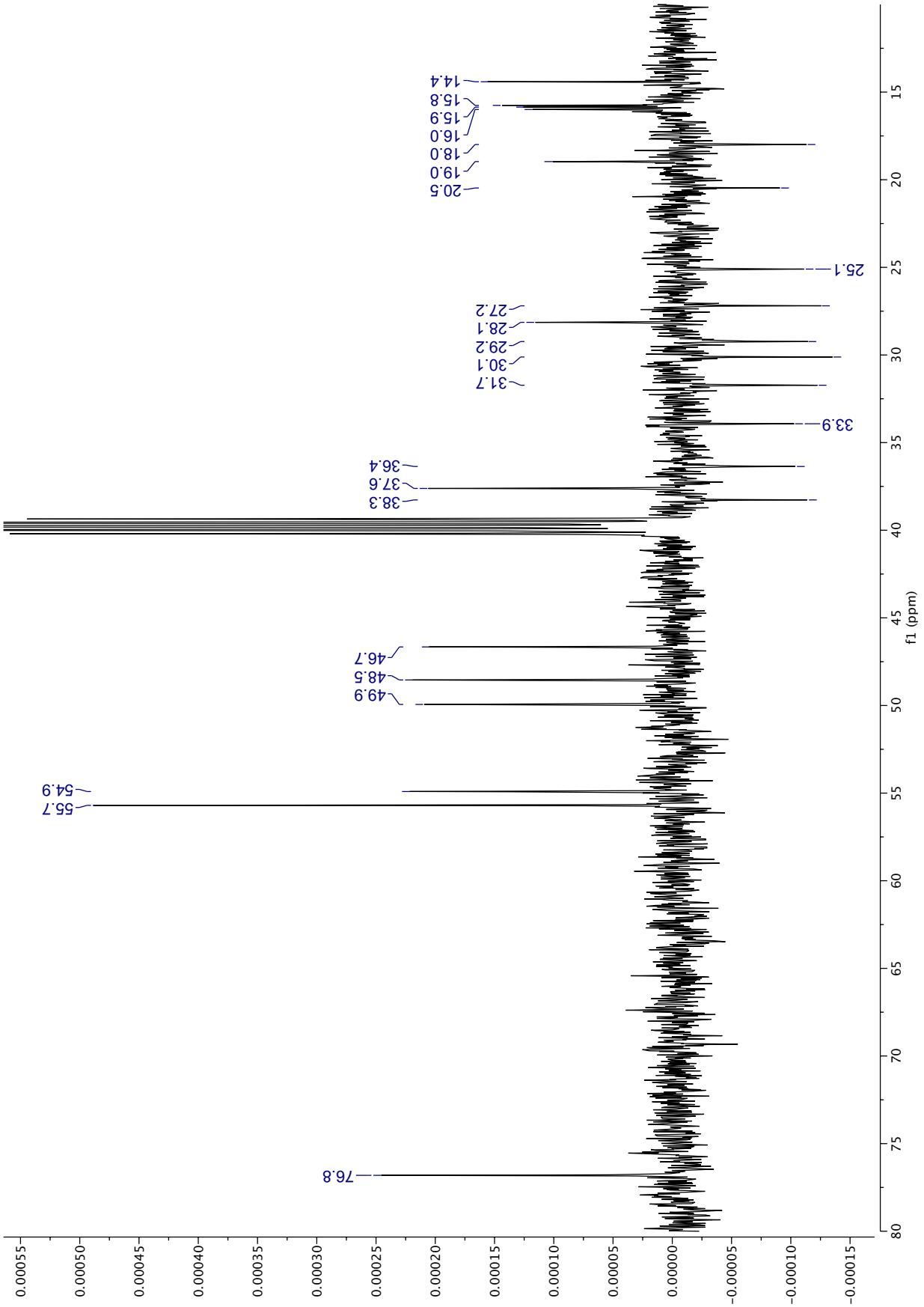


Figure S10. DEPT 135 spectrum (Zoom 2: 80-0ppm) (10 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

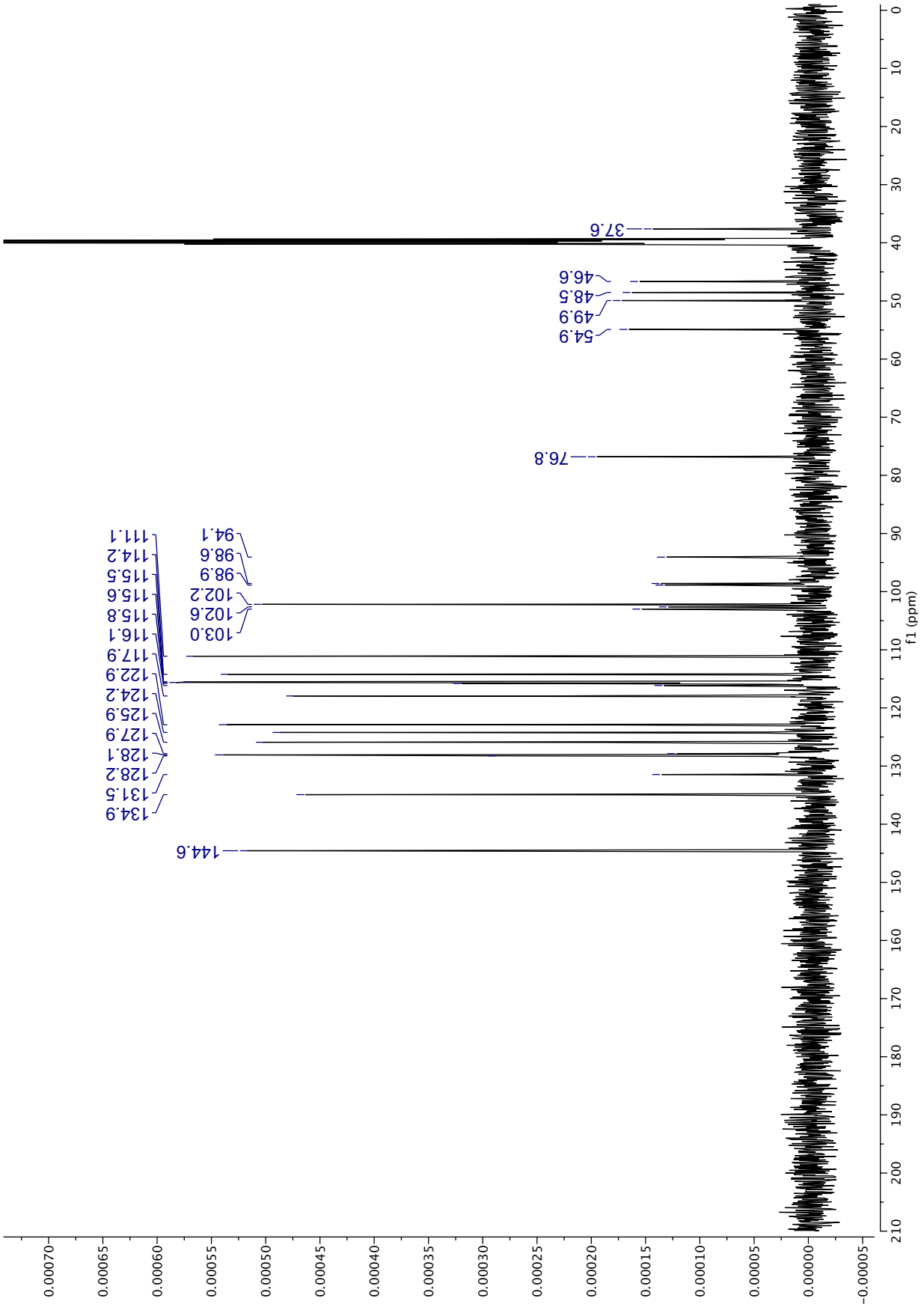


Figure S11. DEPT 90 spectrum (8 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

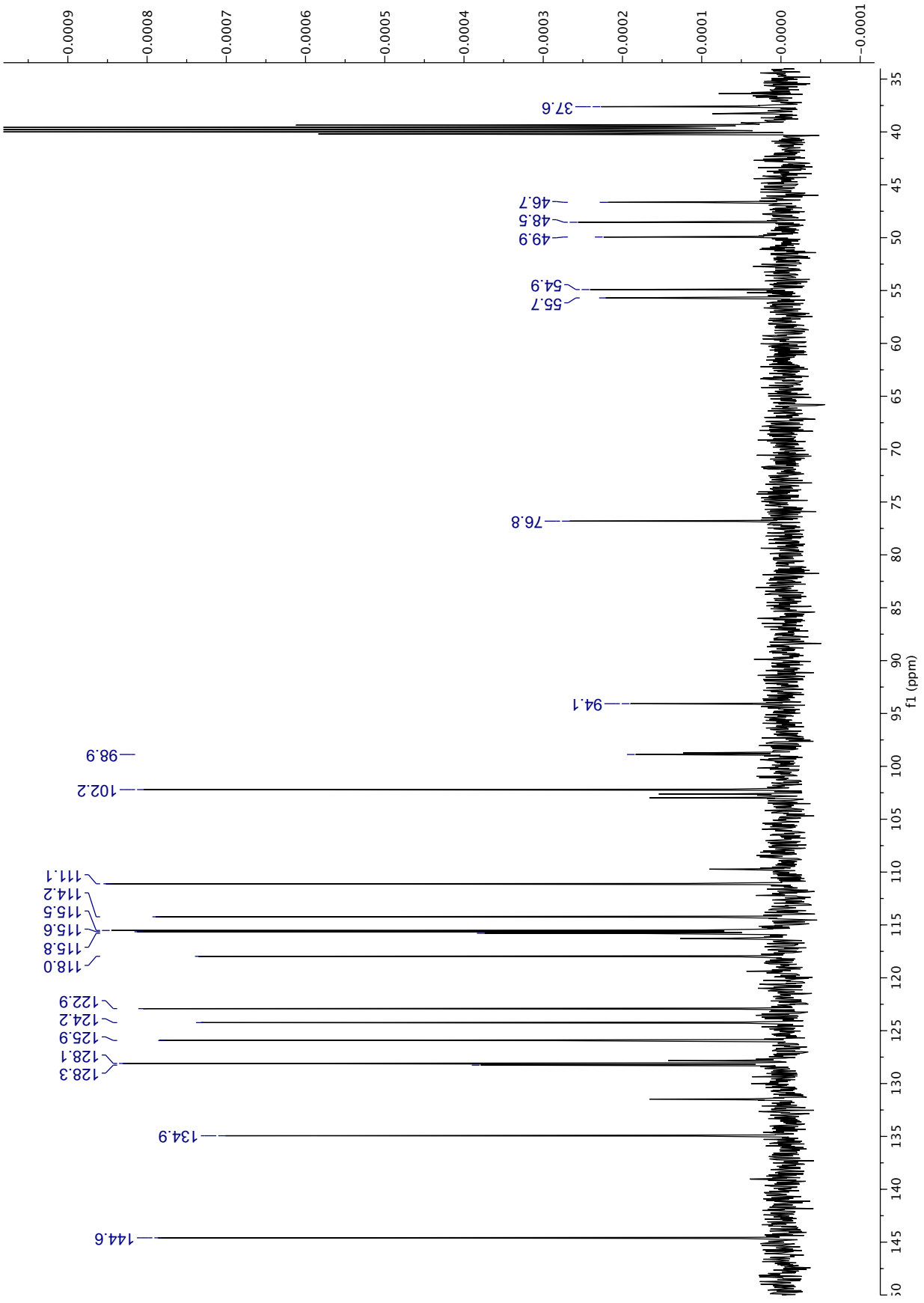


Figure S12. DEPT 90 spectrum (Zoom: 150-35ppm) (8 000 scans) of M1 (10.0 mg) recorded in DMSO-*d*6.

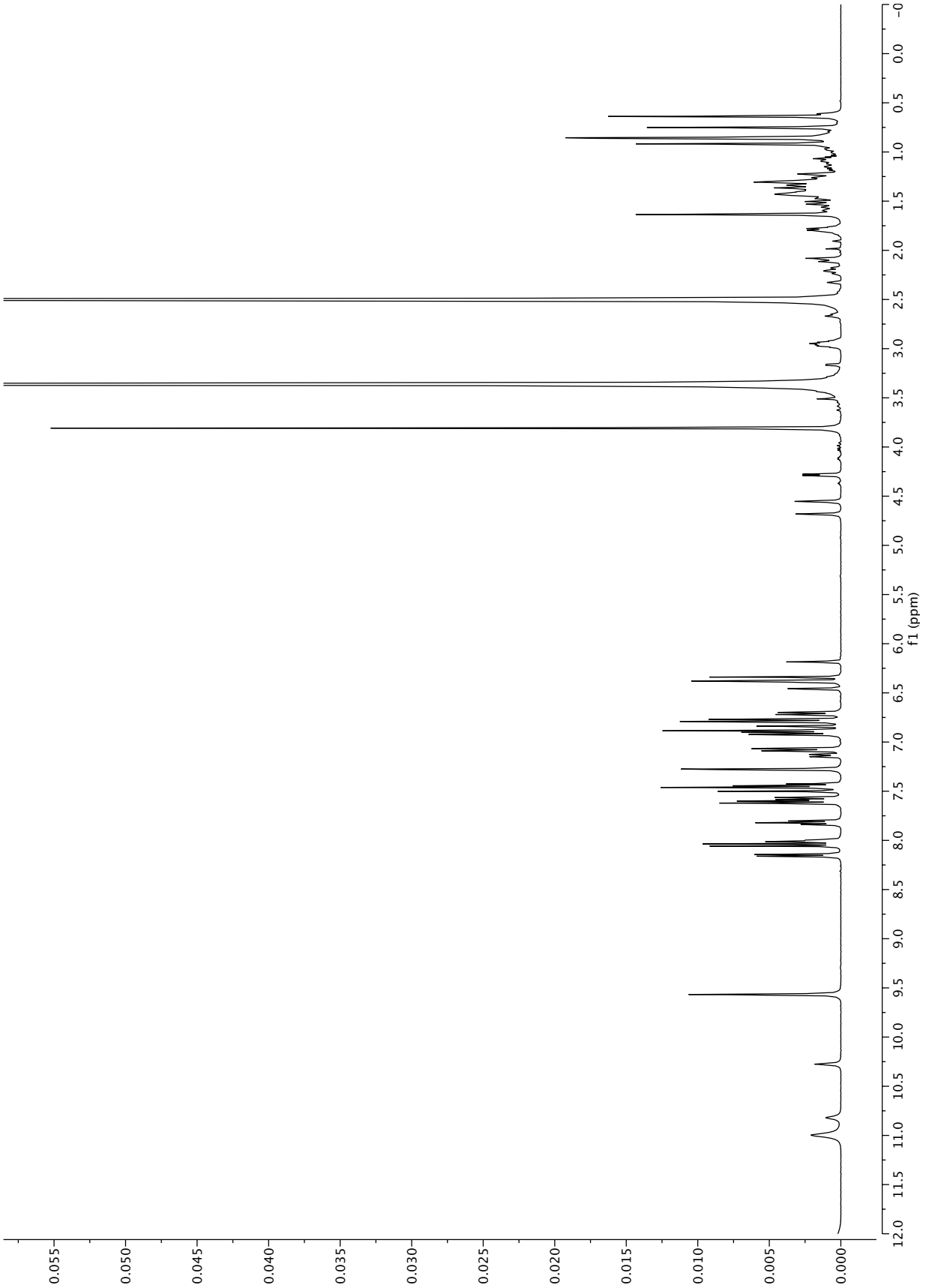


Figure S13. ^{1}H-NMR spectrum of M1 (10.0 mg) recorded in DMSO-*d*6.

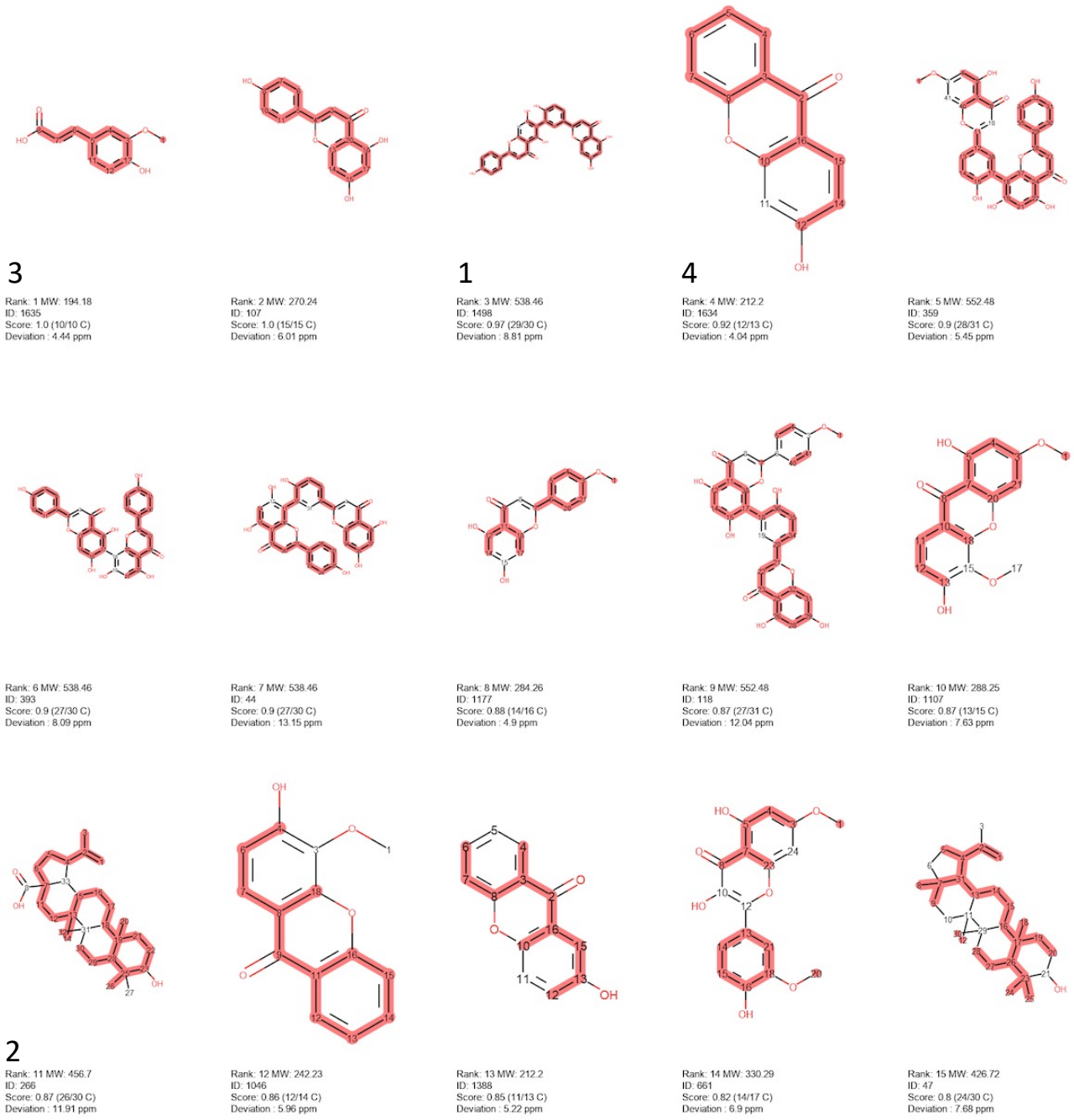


Figure S14. Excerpt of results from the MixONat[13]C-NMR dereplication (+DEPT 90 and 135) of M1 containing standard compounds (**1**, **2**, **3** and **4**) and using the c-type *Garcinia*DB. Equivalent carbons were allowed.

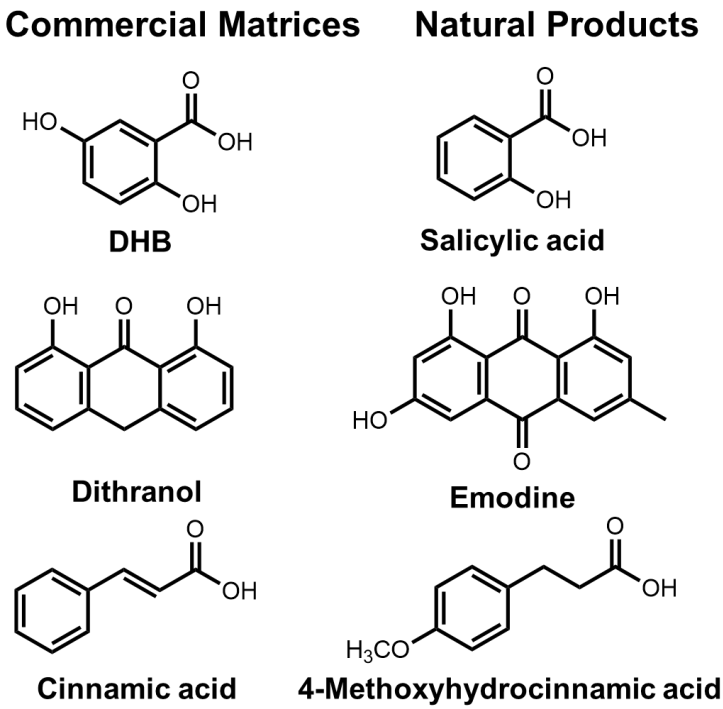


Figure S15. Structural comparison of NPs and comercial MALDI matrices.

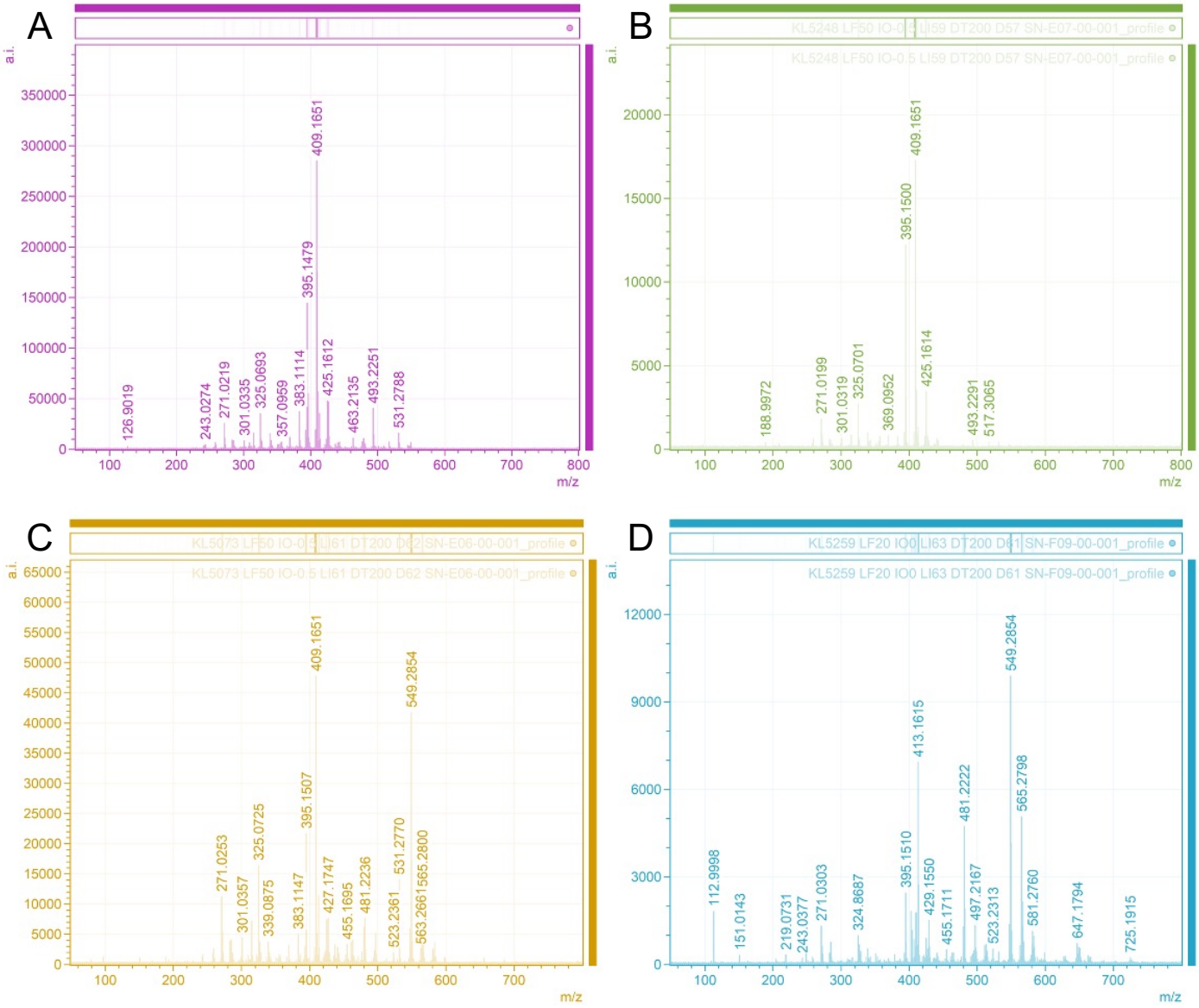


Figure S16. LDI-HRMS spectra of *Garcinia parvifolia* bark extracts KL5670 (**A**) pink, KL5248 (**B**) green, KL5073 (**C**) yellow and KL5259 (**D**) blue.

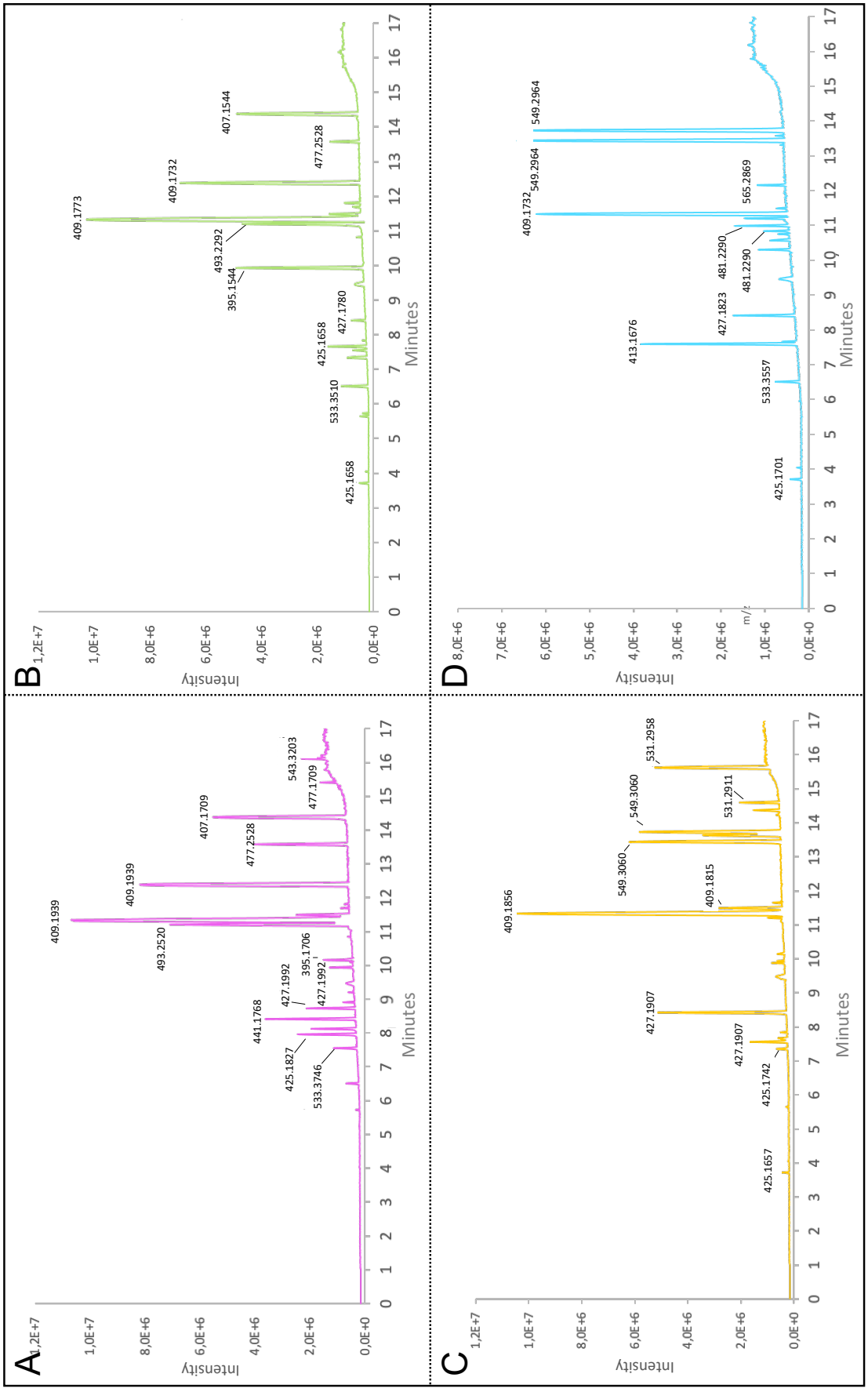


Figure S17. UPLC-HRMS chromatograms (not calibrated) of *Garcinia parvifolia* bark extracts KL5670 (**A**) pink, KL5248 (**B**) green, KL5073 (**C**) yellow and KL5259 (**D**) blue.

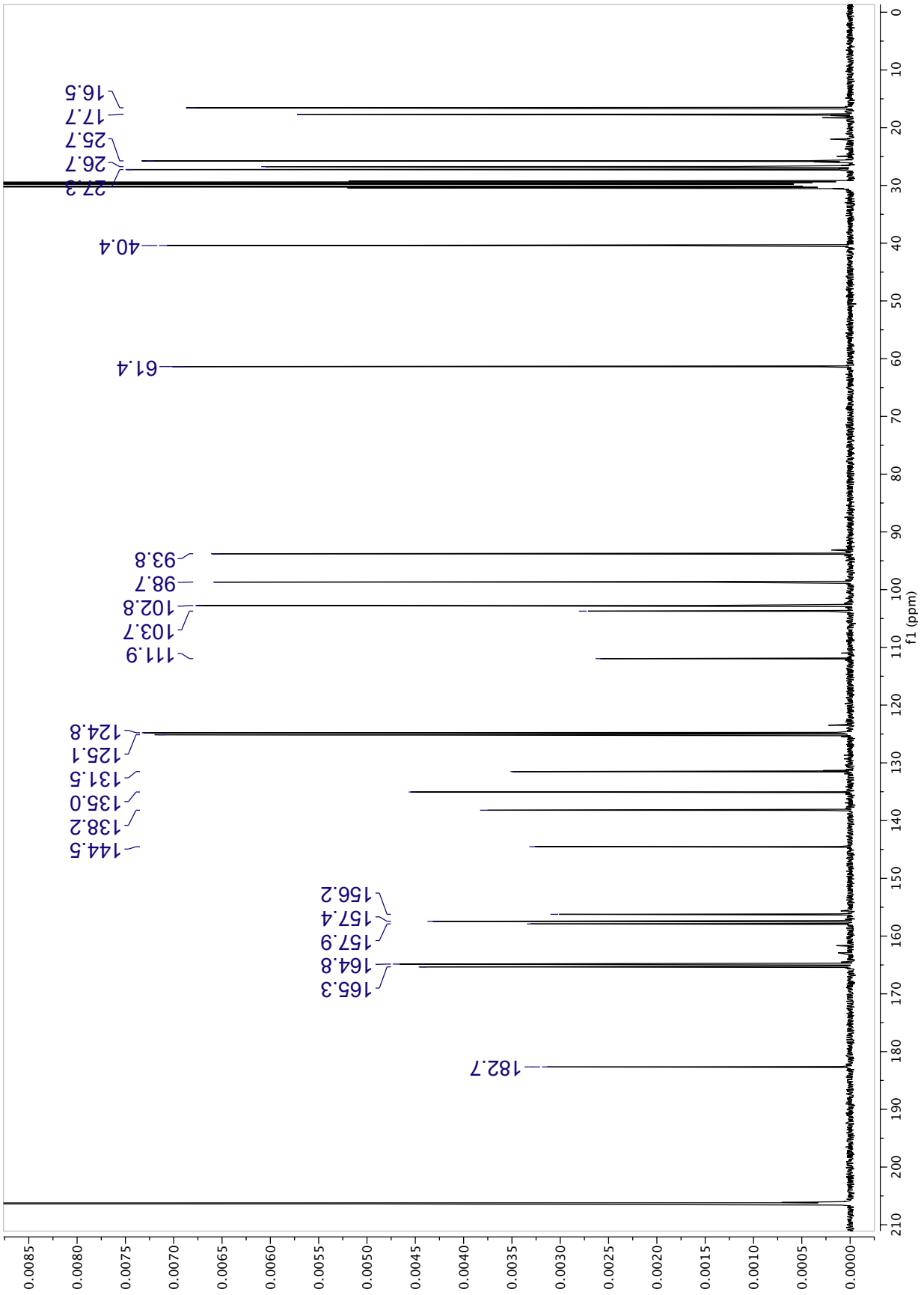


Figure S18. ^{13}C NMR spectrum (5163 scans) of F6 (50.0 mg) recorded in acetone-*d*6.

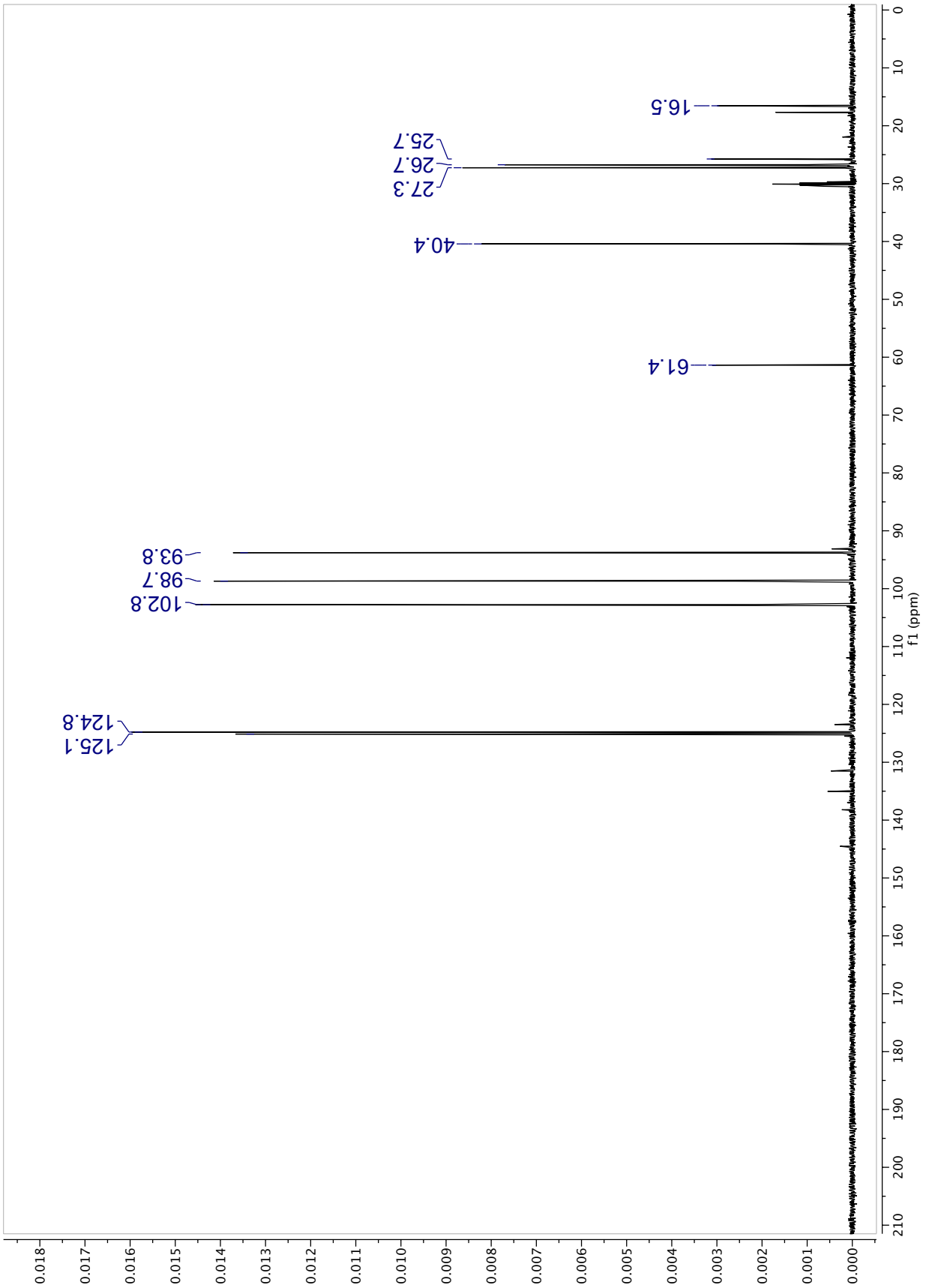


Figure S19. DEPT-135 spectrum (2000 scans) of F6 (50.0 mg) recorded in acetone-*d*6.

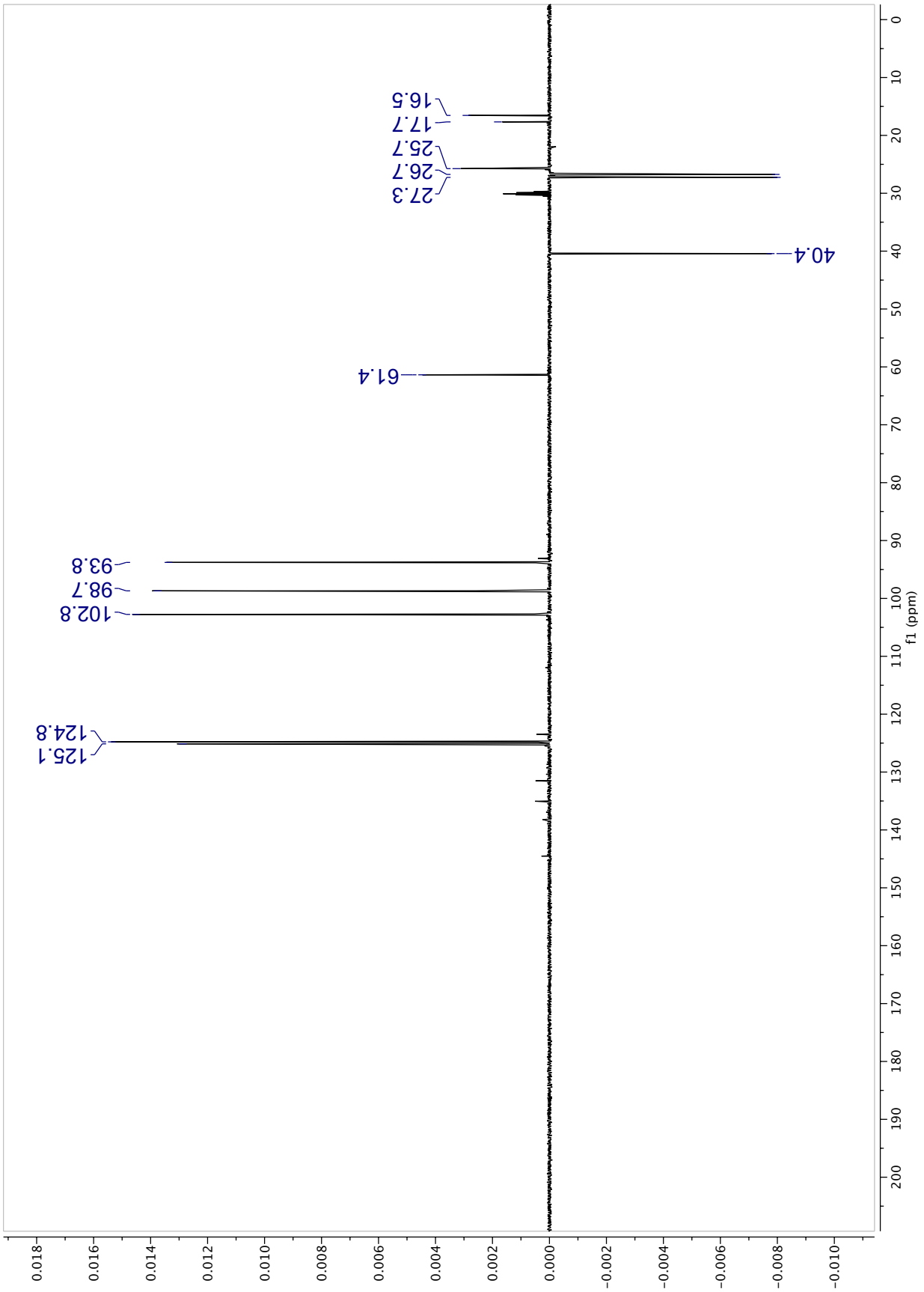


Figure S20. DEPT-90 spectrum (2000 scans) of F6 (50.0 mg) recorded in acetone-*d*6.

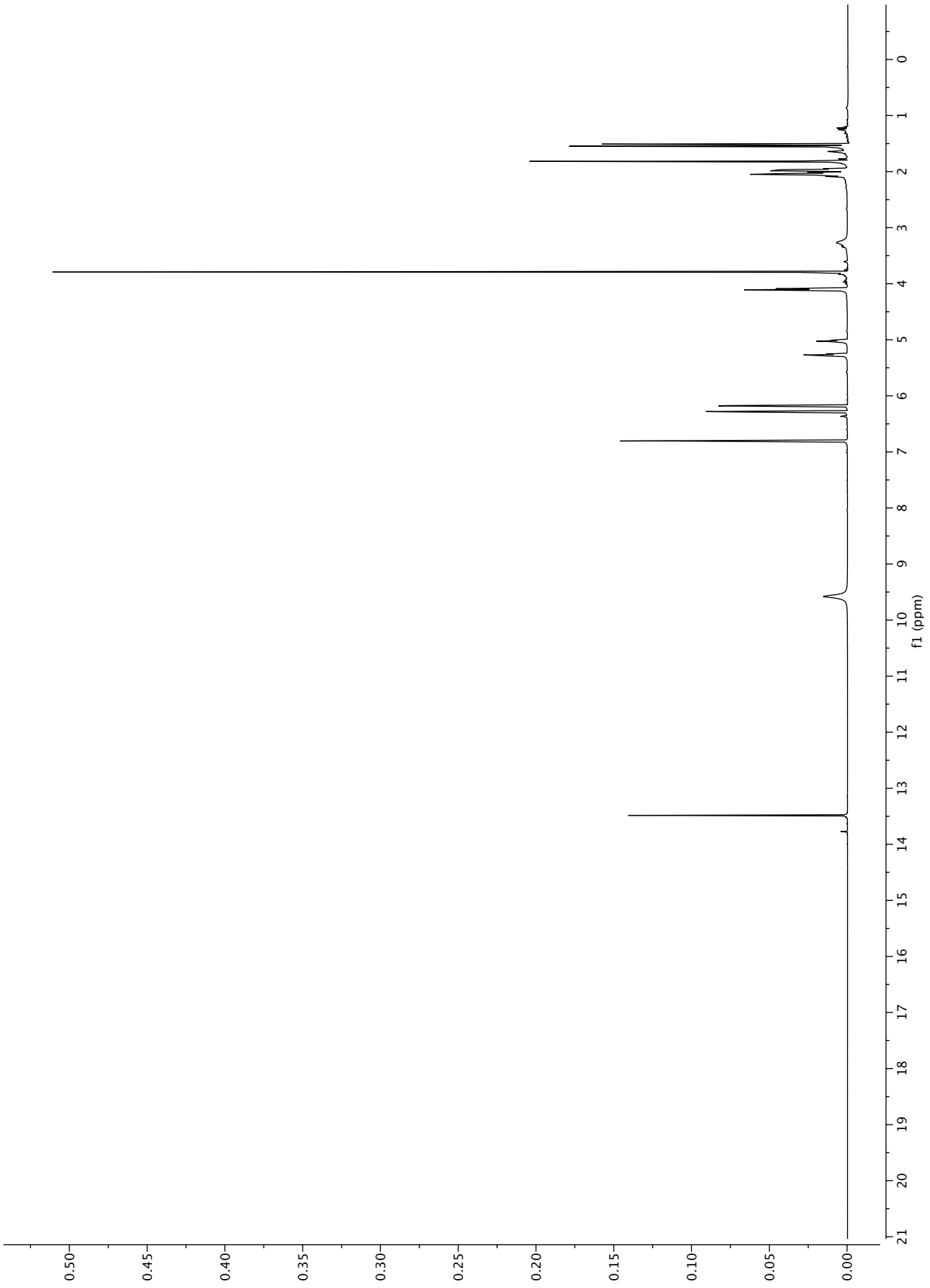


Figure S21. ^{1}H-NMR spectrum of F6 (50.0 mg) recorded in acetone-*d*6.

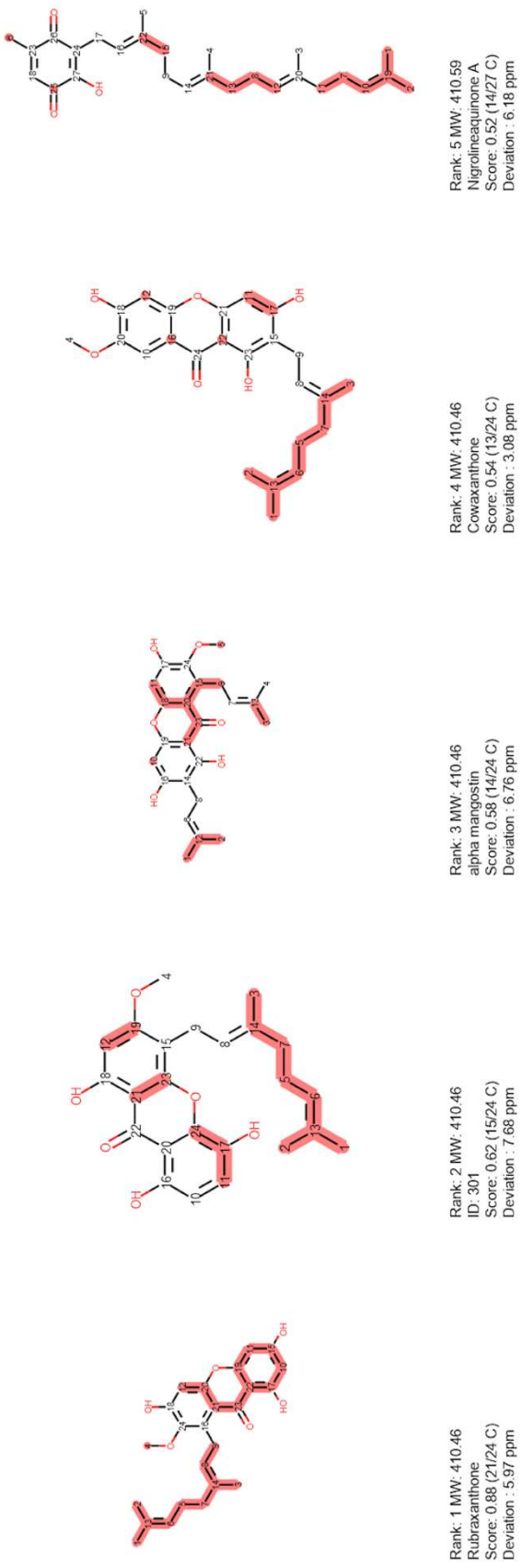


Figure S22. Excerpt of results from the MixONat ^{13}C-NMR dereplication (+ DEPT-90 and 135) of F6 using the c-type *Garcinia*DB. Equivalent carbons were allowed. The mass filter was set to 410 Da.

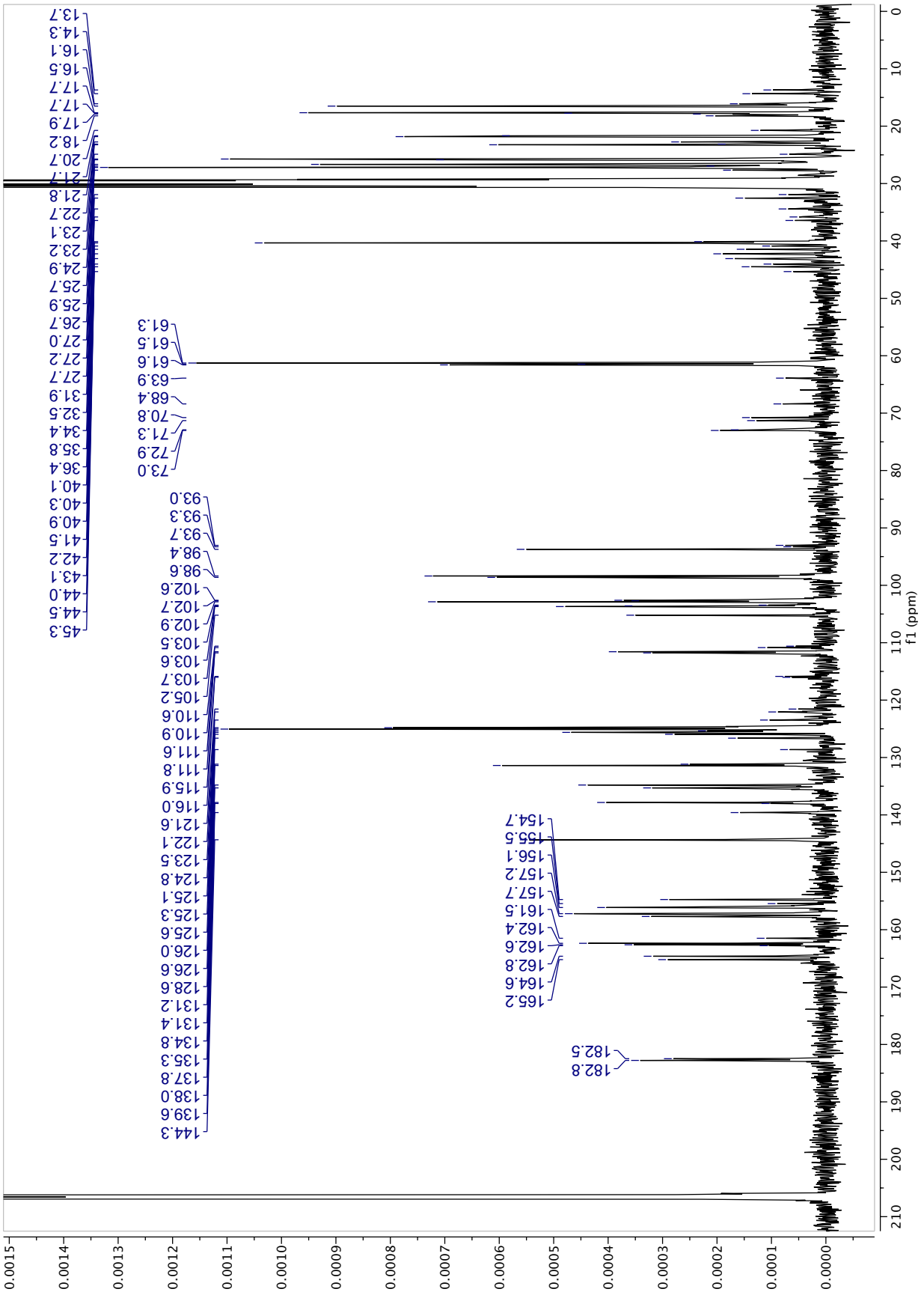


Figure S23. ^{13}C-NMR spectrum (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

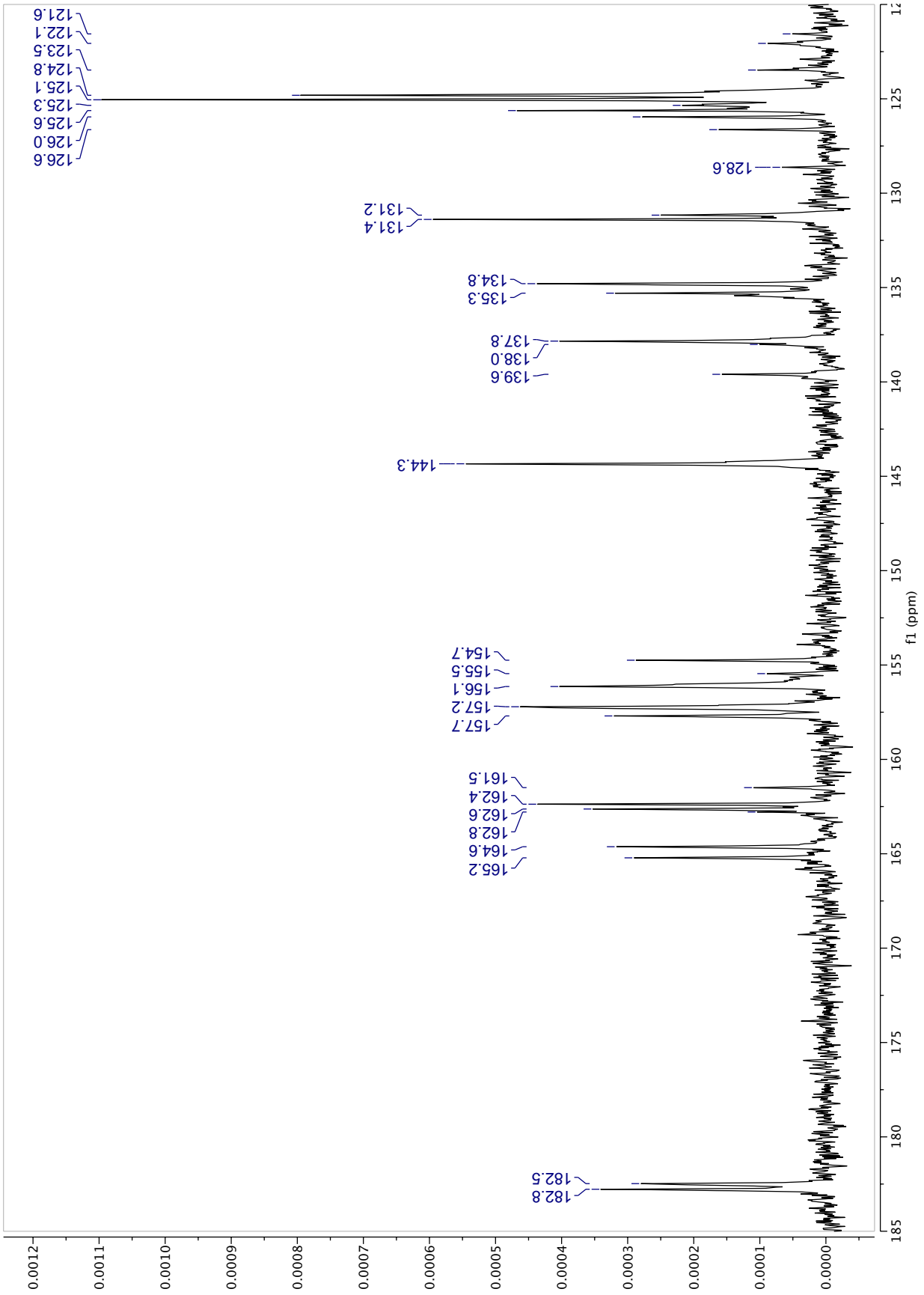


Figure S24. ^{13}C-NMR spectrum (Zoom 1: 185-120ppm) (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

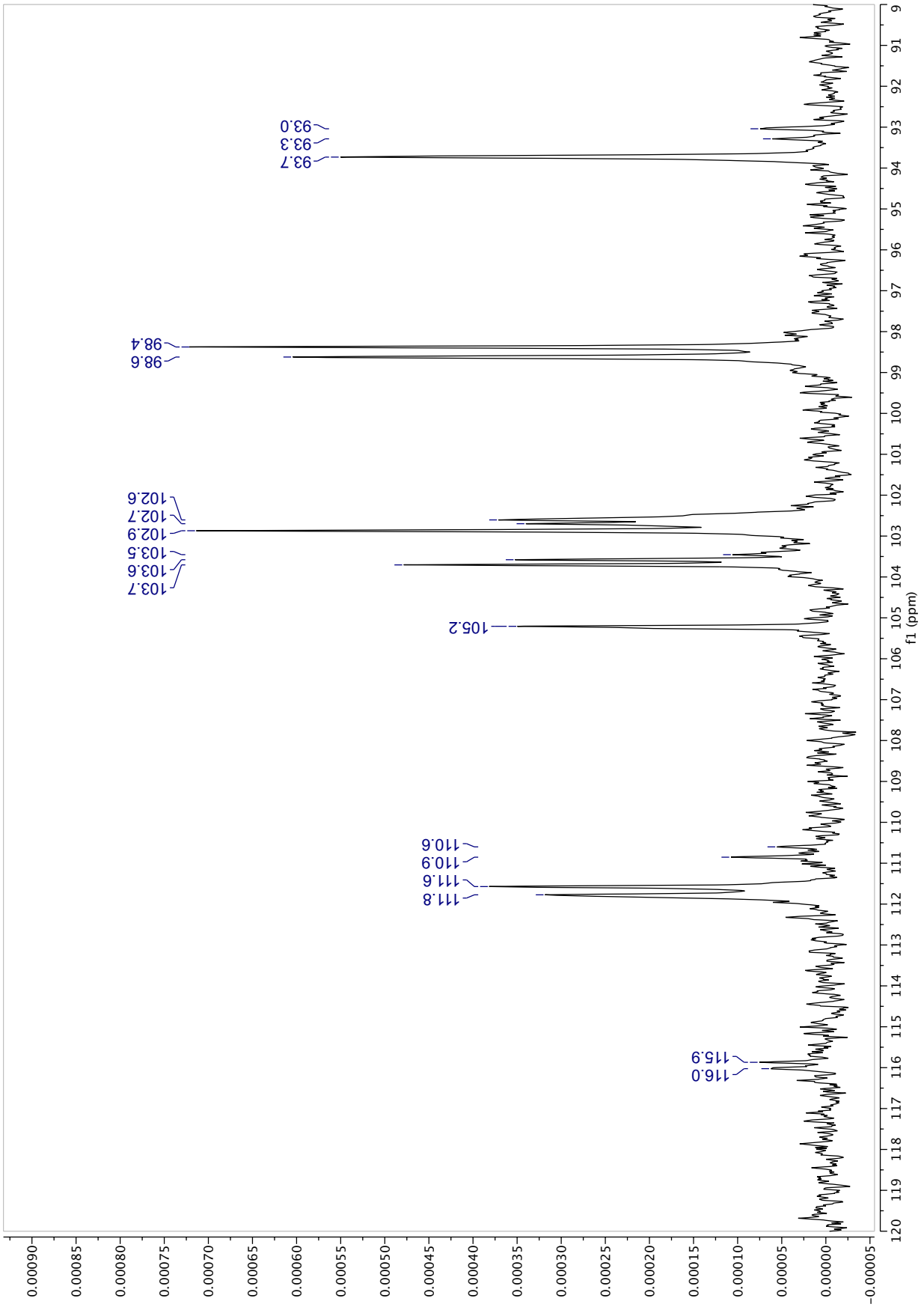


Figure S25. ^{13}C-NMR spectrum (Zoom 2: 120-90ppm) (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

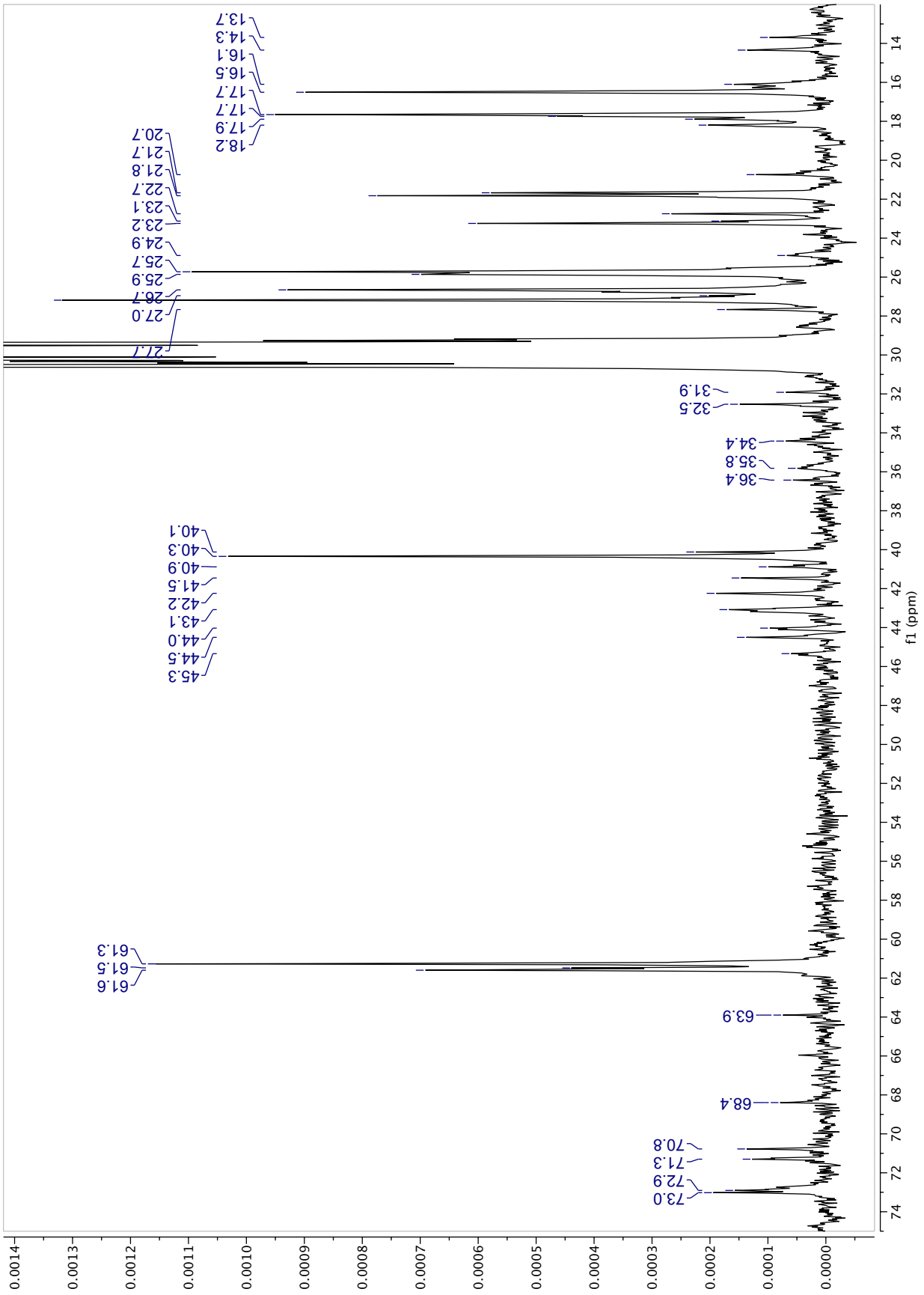


Figure S26. ^{13}C-NMR spectrum (Zoom 3: 75-12ppm) (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

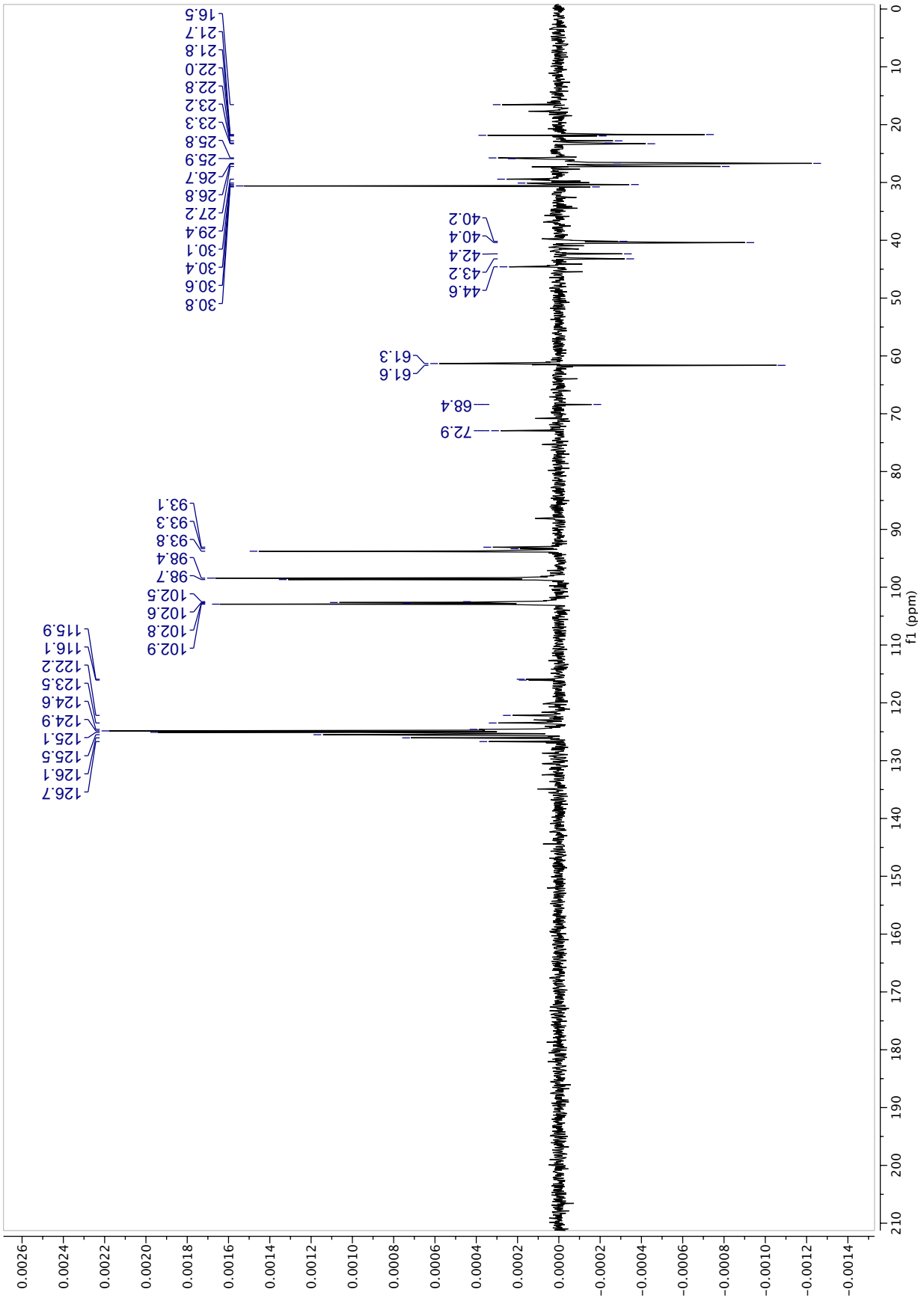


Figure S27. DEPT-135 spectrum (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

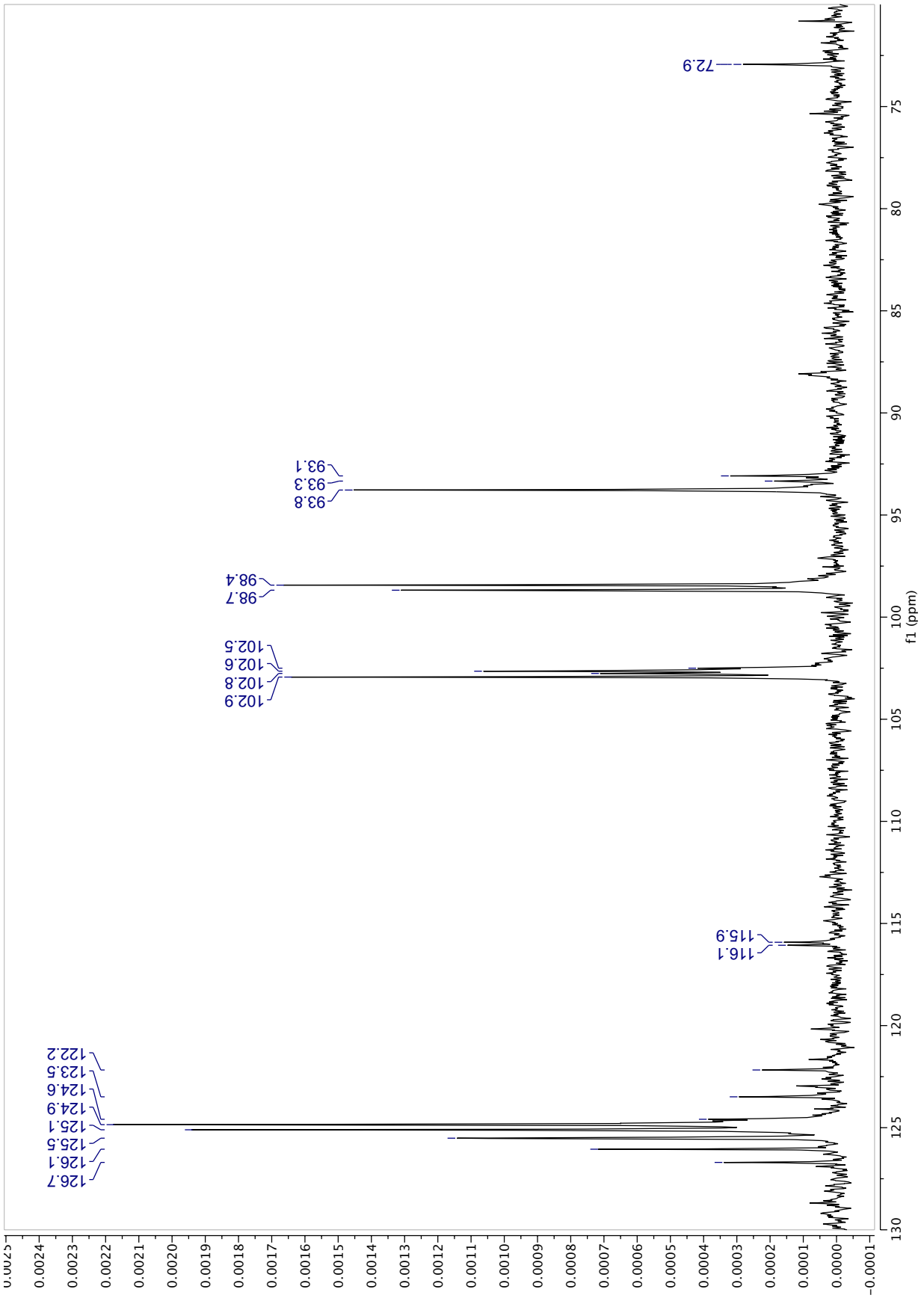


Figure S28. DEPT-135 spectrum (Zoom 1: 130-70ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

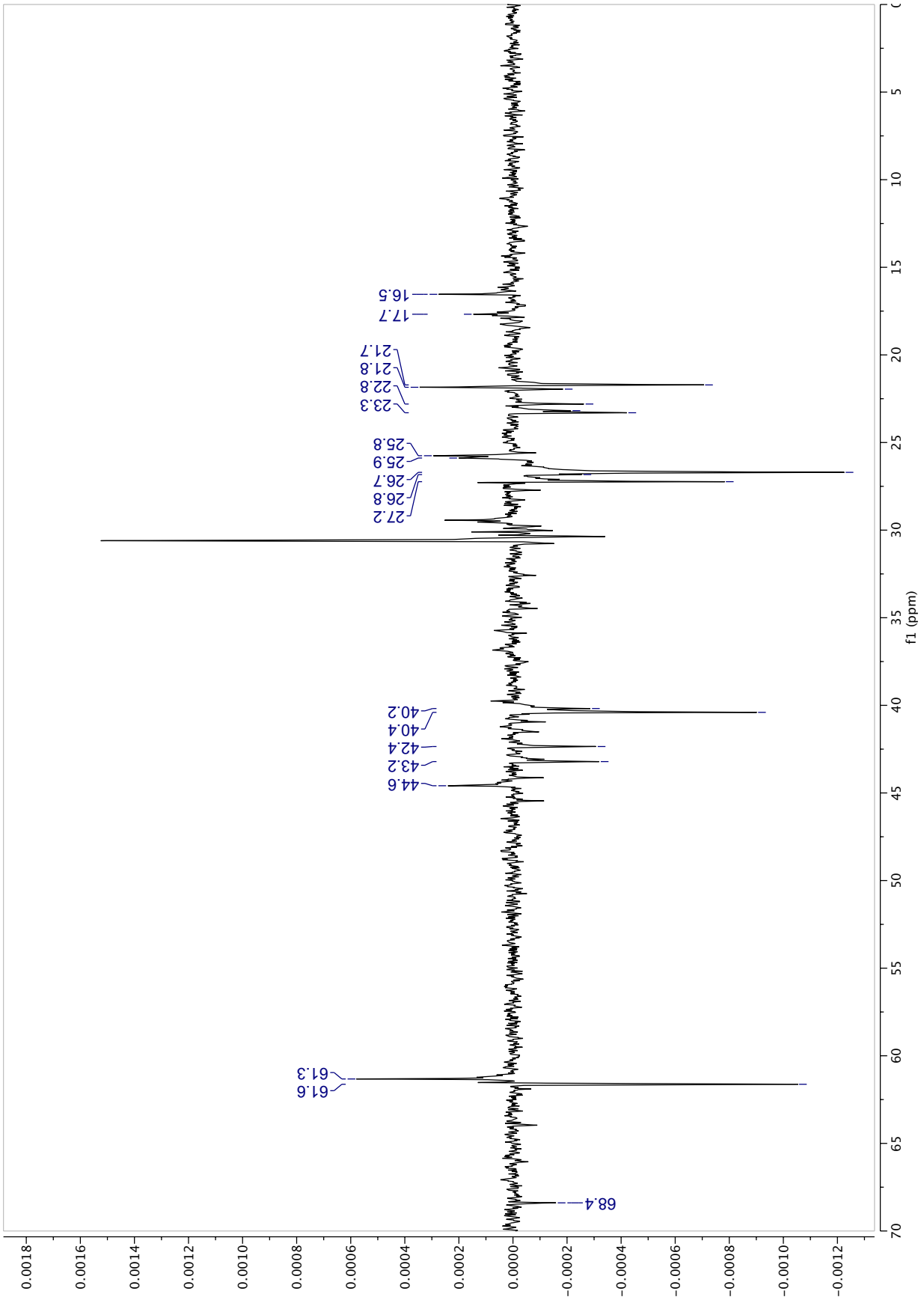


Figure S29. DEPT-135 spectrum (Zoom 2: 70-0ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

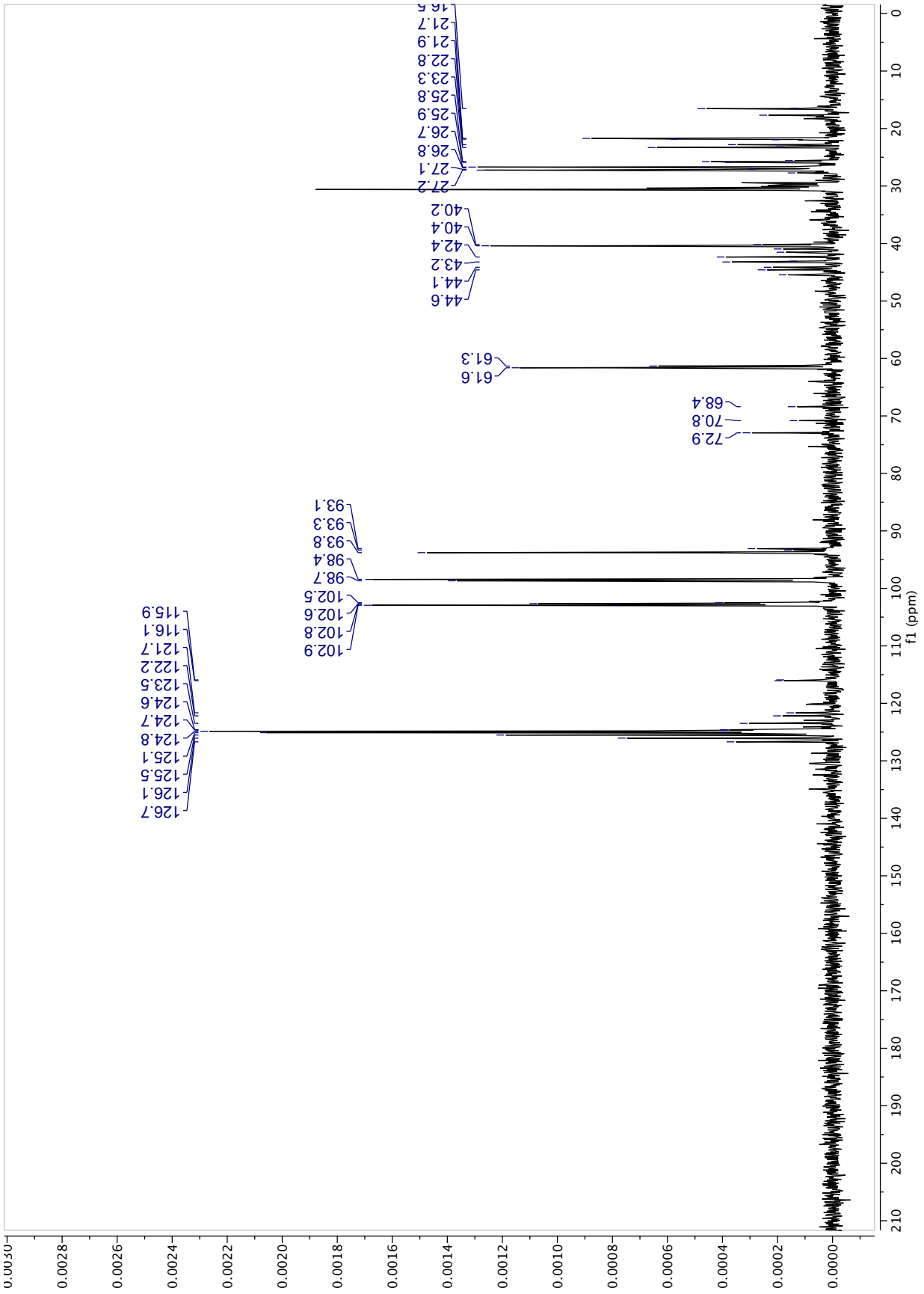


Figure S30. DEPT-90 spectrum (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

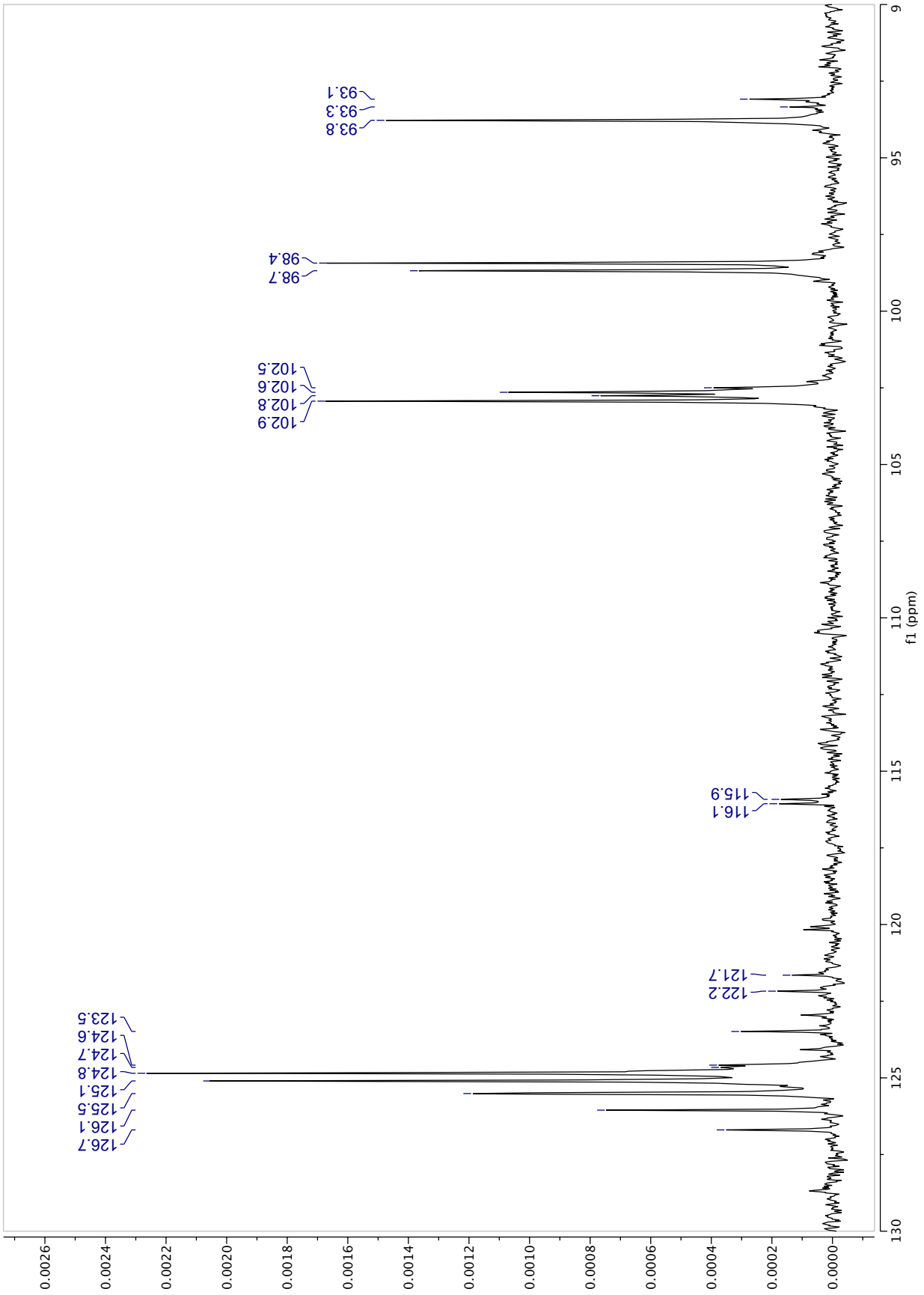


Figure S31. DEPT-90 spectrum (Zoom 1: 130-90ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

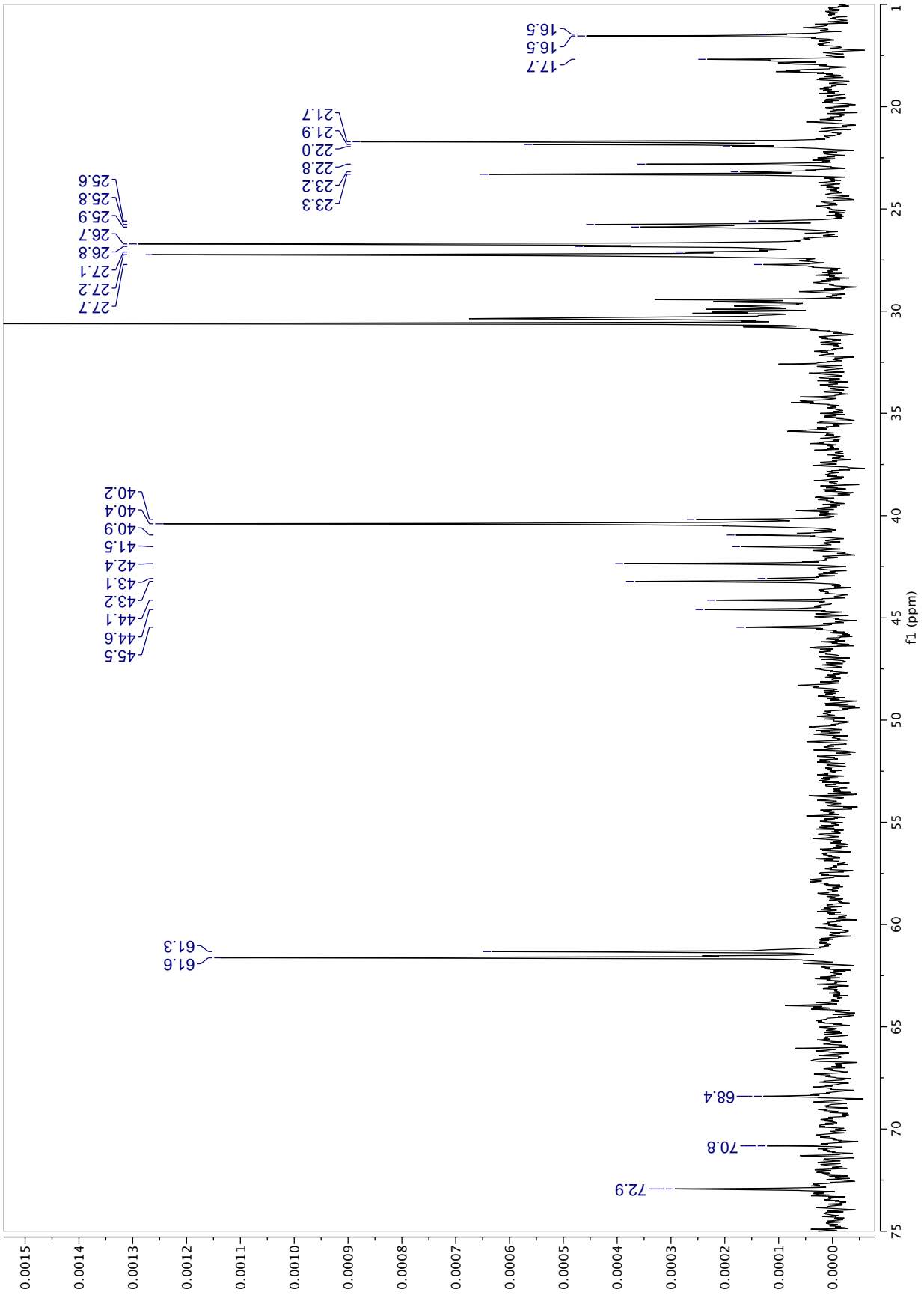


Figure S32. DEPT-90 spectrum (Zoom 2: 75-10ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

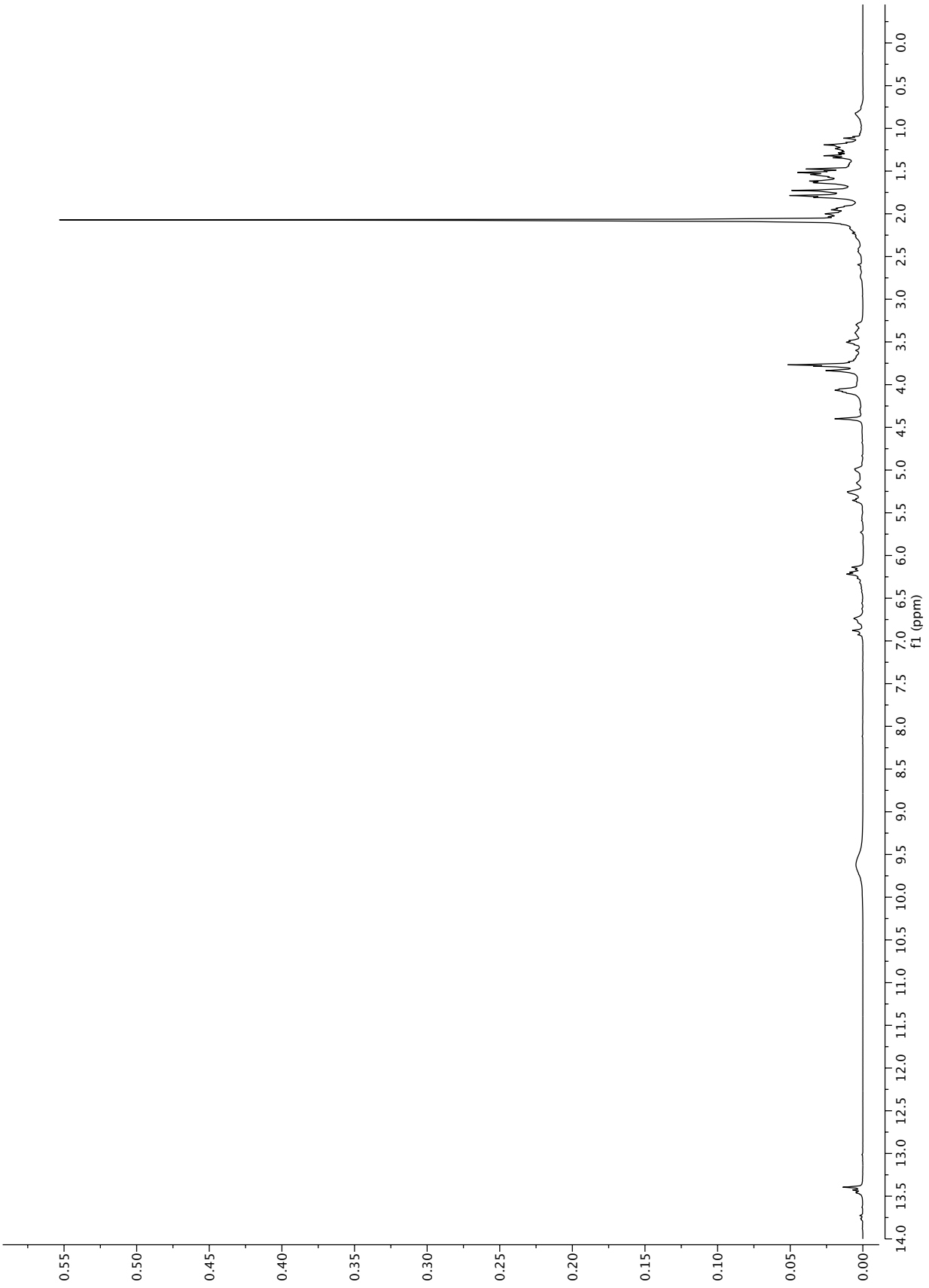


Figure S33. ^{1}H-NMR spectrum of F10 (40.0 mg) recorded in acetone-*d*6.

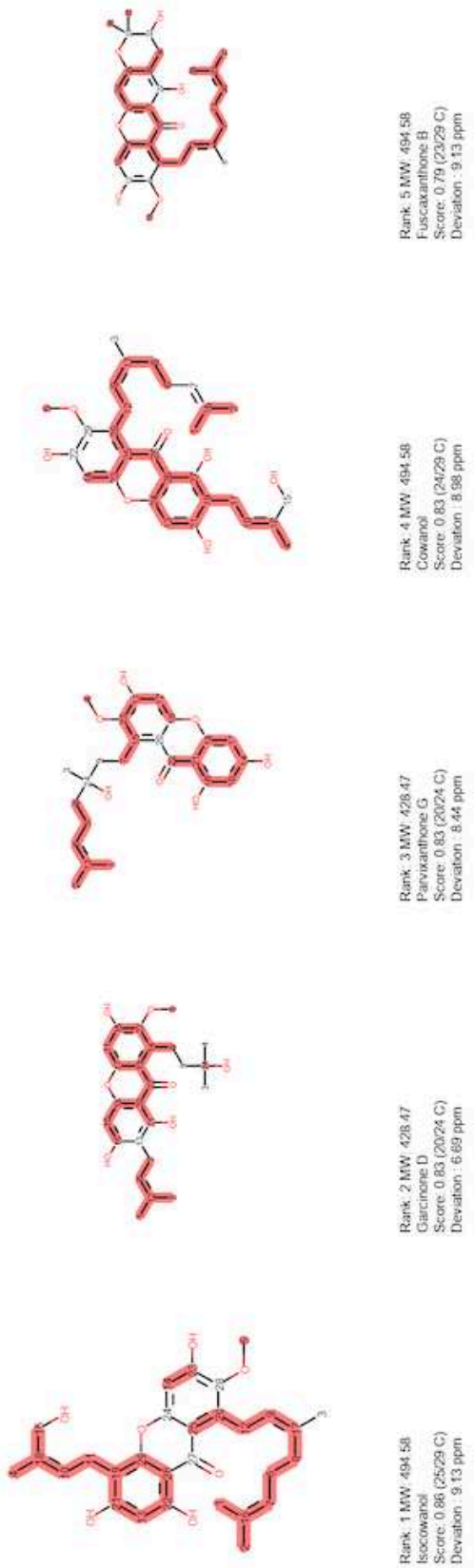


Figure S34. Excerpt of results from the MixONat ^{13}C-NMR dereplication (+ DEPT 90 and 135) of F10 using the c-type *Garcinia*DB. Equivalent carbons were allowed, and the molecular weight filter was set to 494 (Isocowanol, **6**) and 428 Da (Parvixanthone G, **7**).

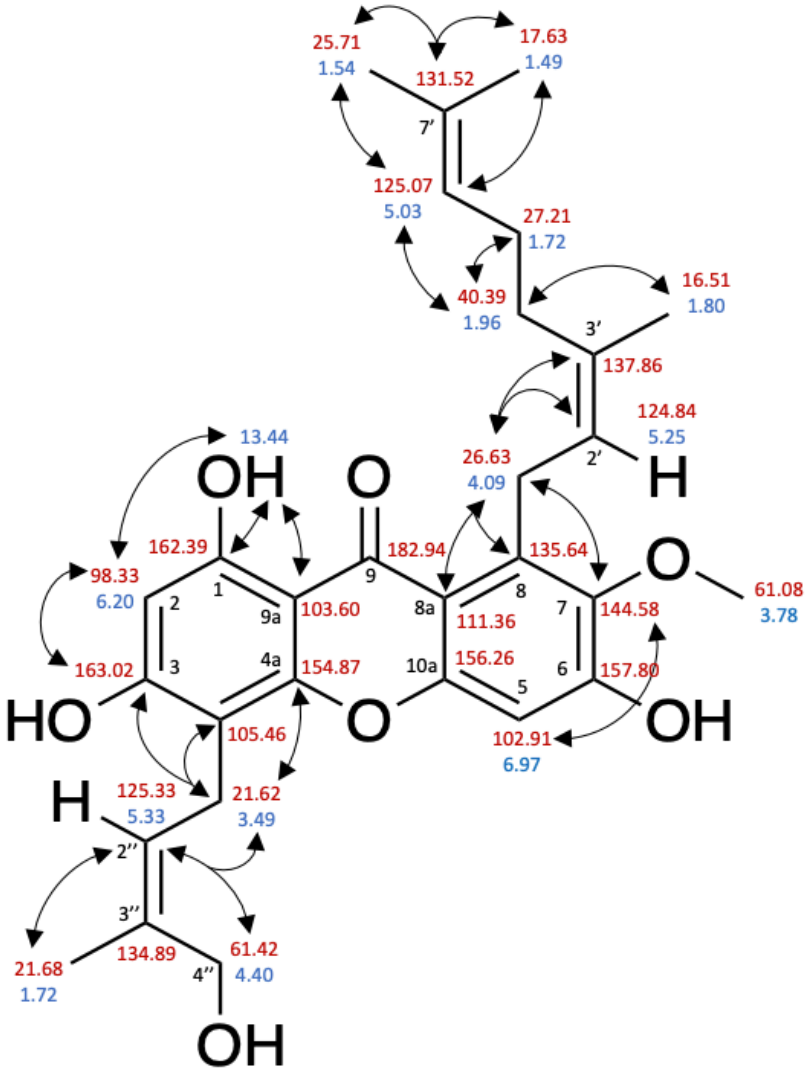


Figure S35. Structure of isocowanol (**6**) and its NMR shifts. δ_C shifts are highlighted in red whereas δ_H are marked in blue.

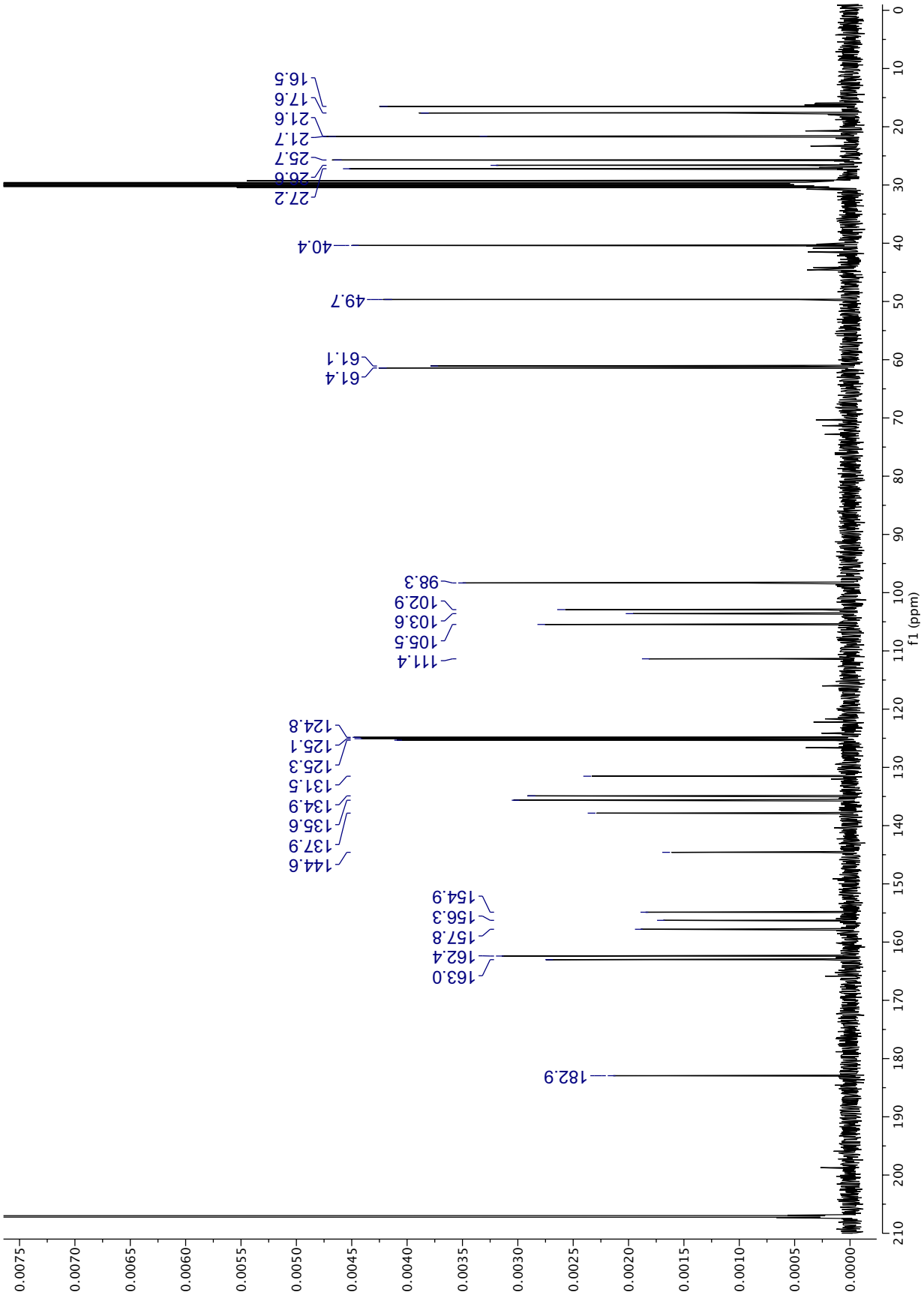


Figure S36. ^{13}C-NMR spectrum (900 scans) of isocowanol (**6**, 46.49 mg) recorded in acetone-*d*6.

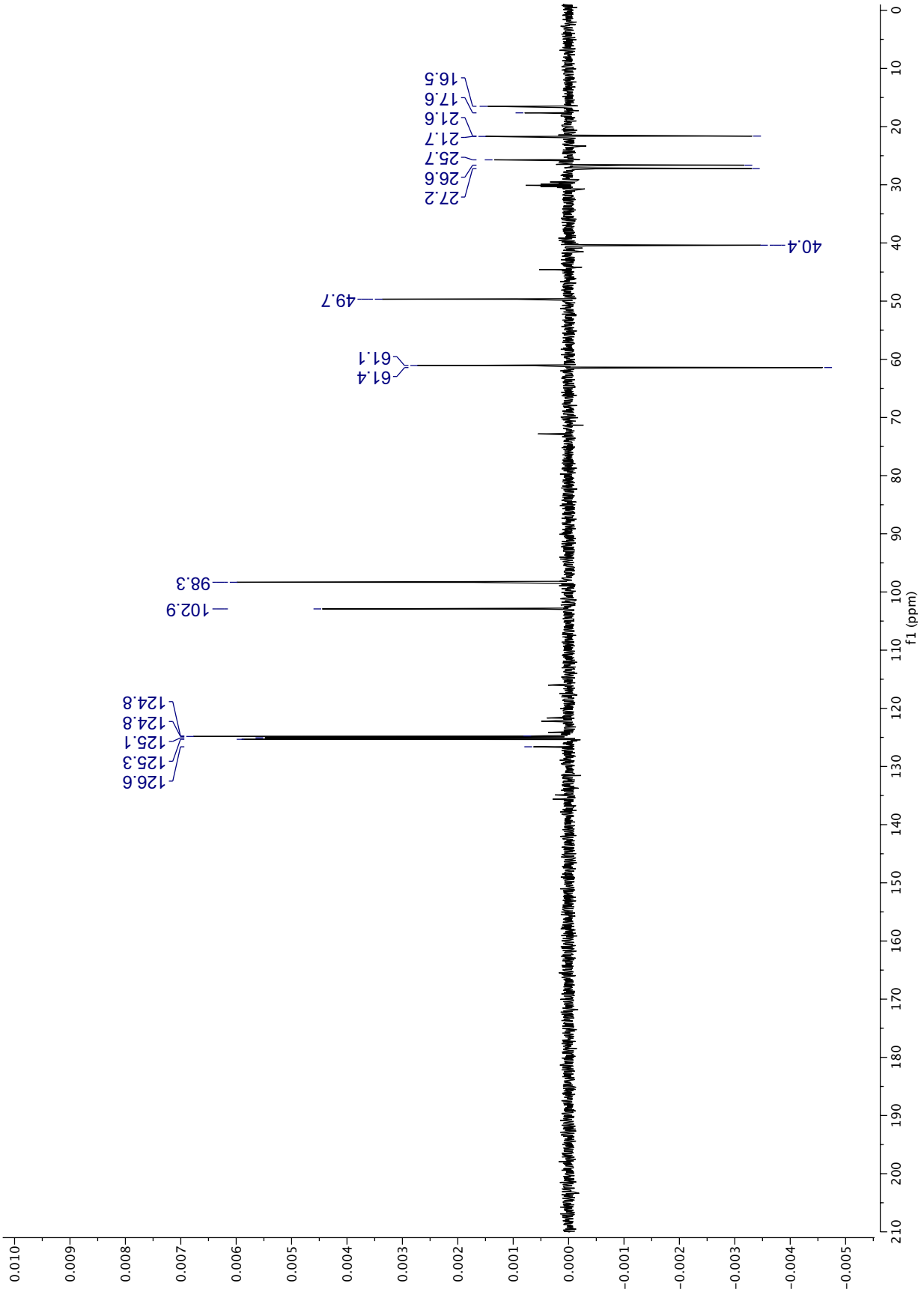


Figure S37. DEPT-135 spectrum (400 scans) of isocowanol (**6**, 46.49 mg) recorded in acetone-*d*6.

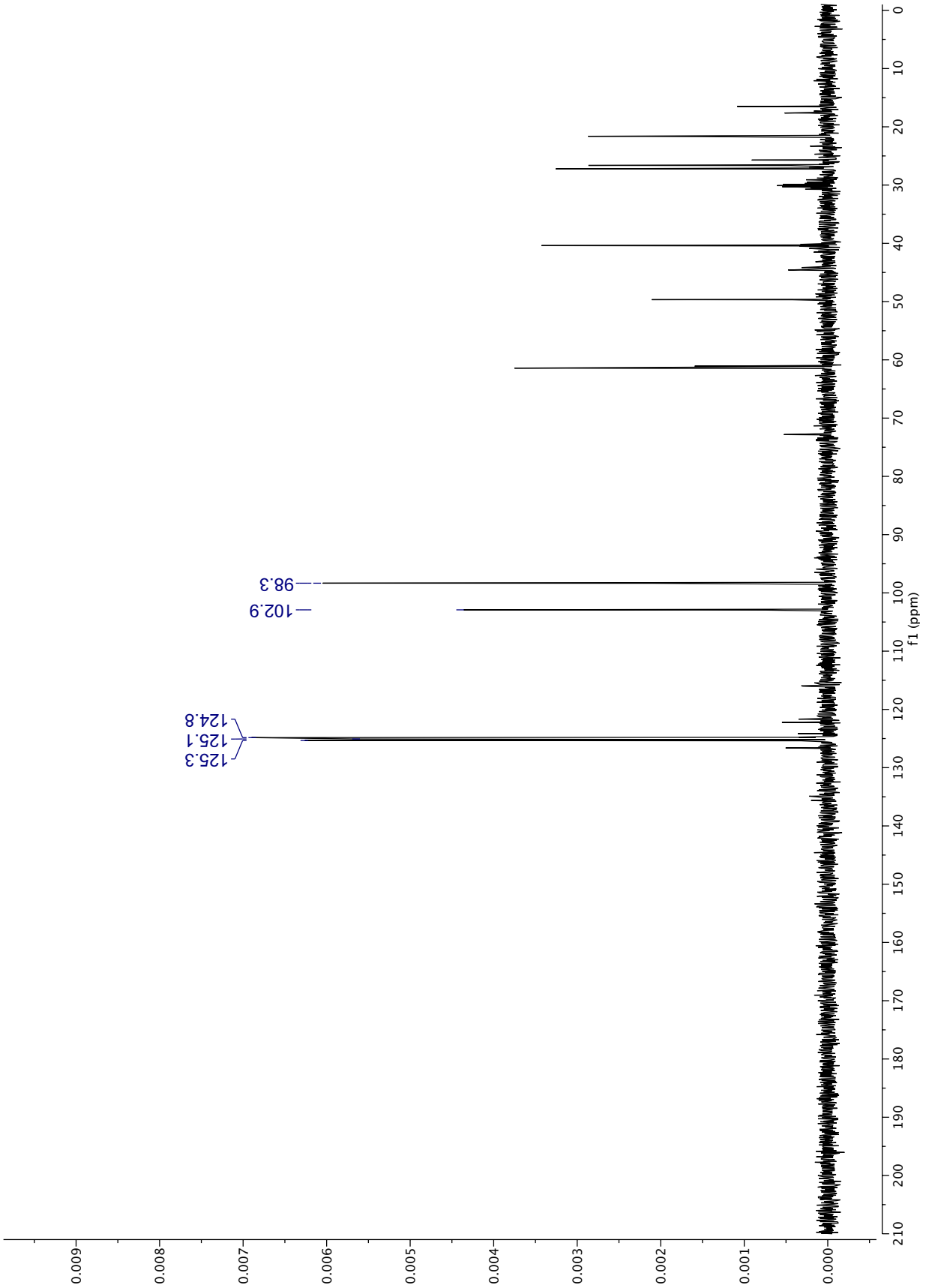


Figure S38. DEPT-90 spectrum (400 scans) of isocowanol (**6**, 46.49 mg) recorded in acetone-*d*6.

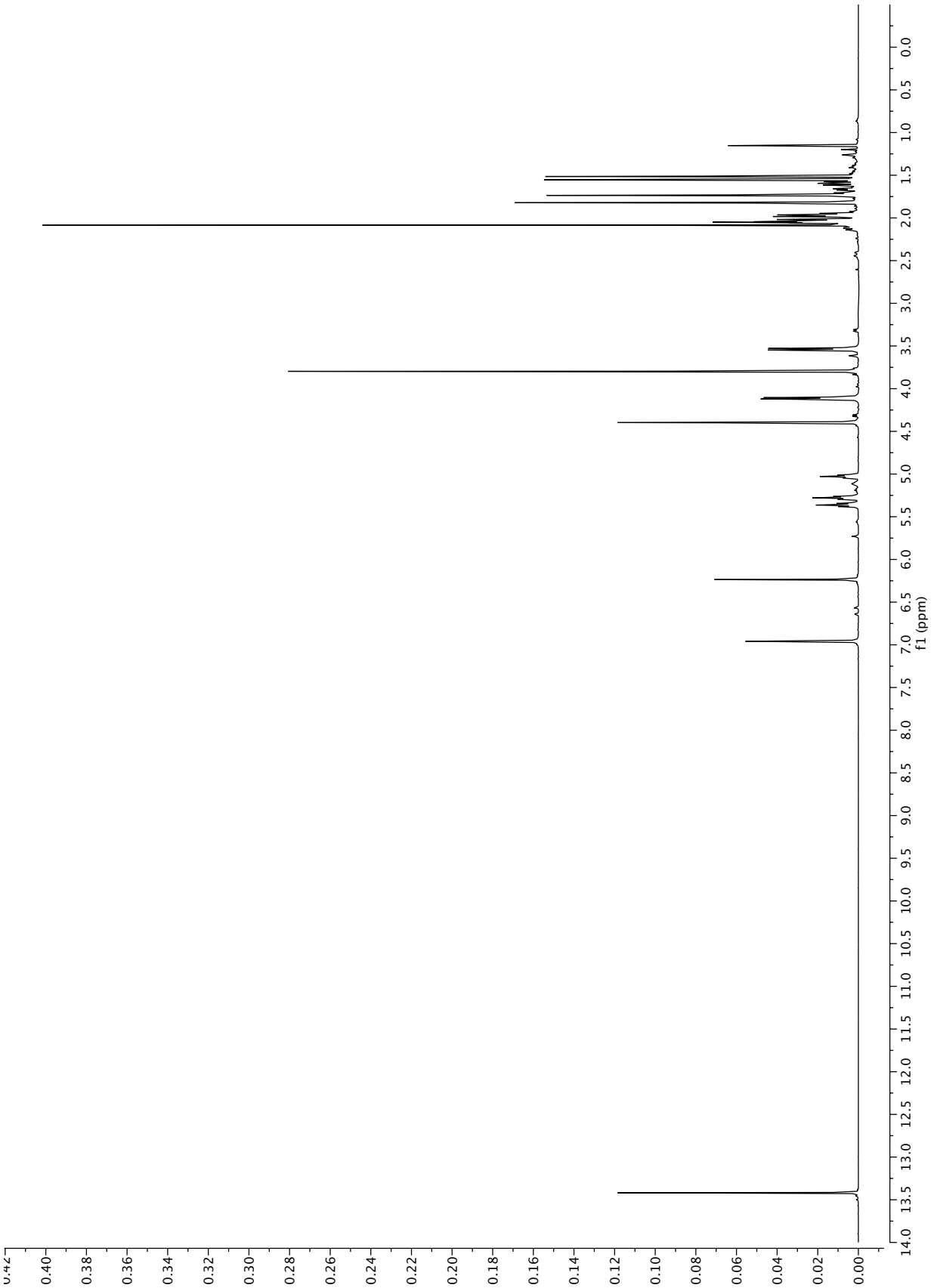


Figure S39. ^{1}H-NMR spectrum of isocowanol (**6**, 46.49 mg) recorded in acetone-*d*6.

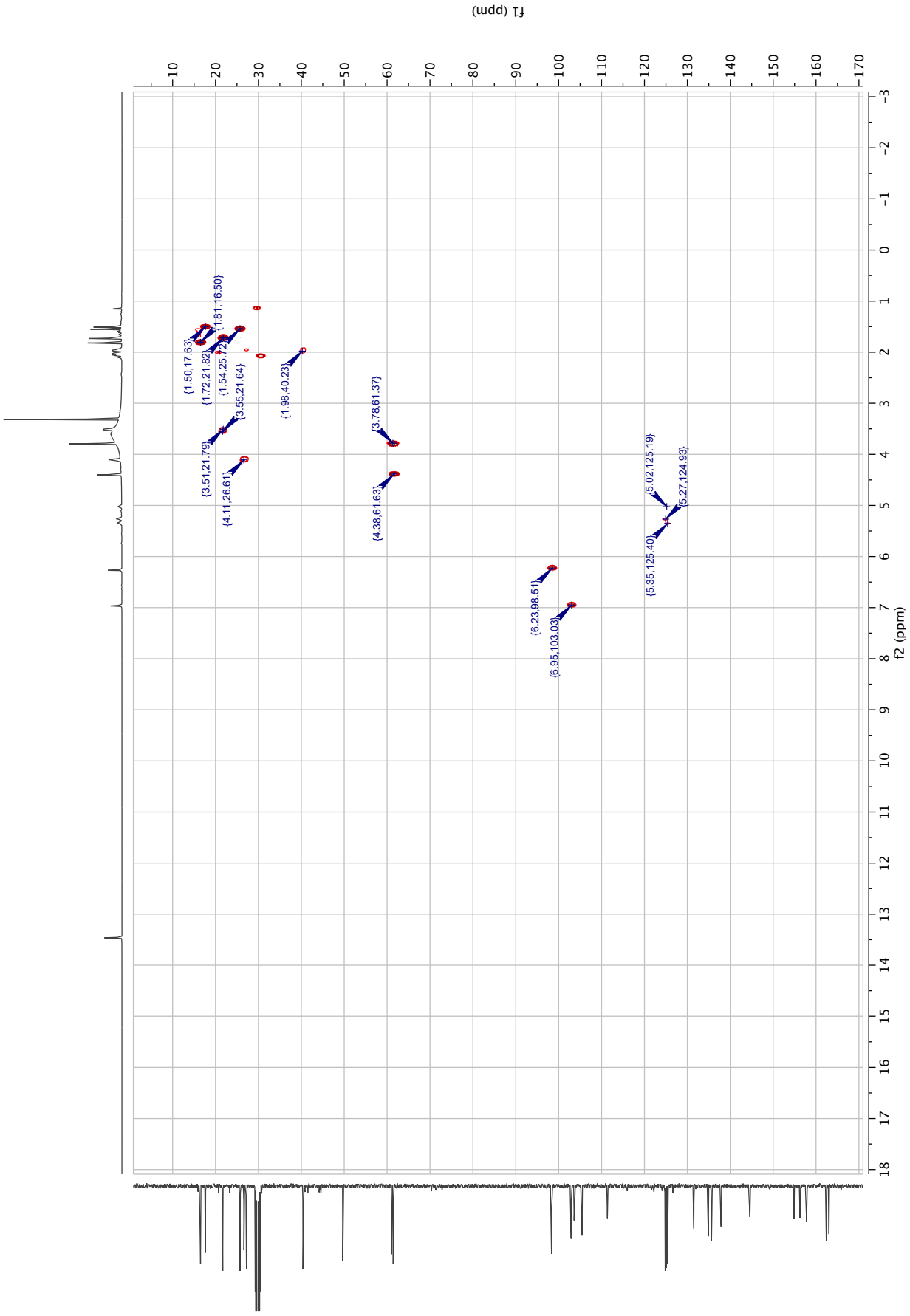


Figure S40. HSQC spectrum (4 scans) of isocowanol (**6**, 46.49 mg) recorded in acetone-*d*6.

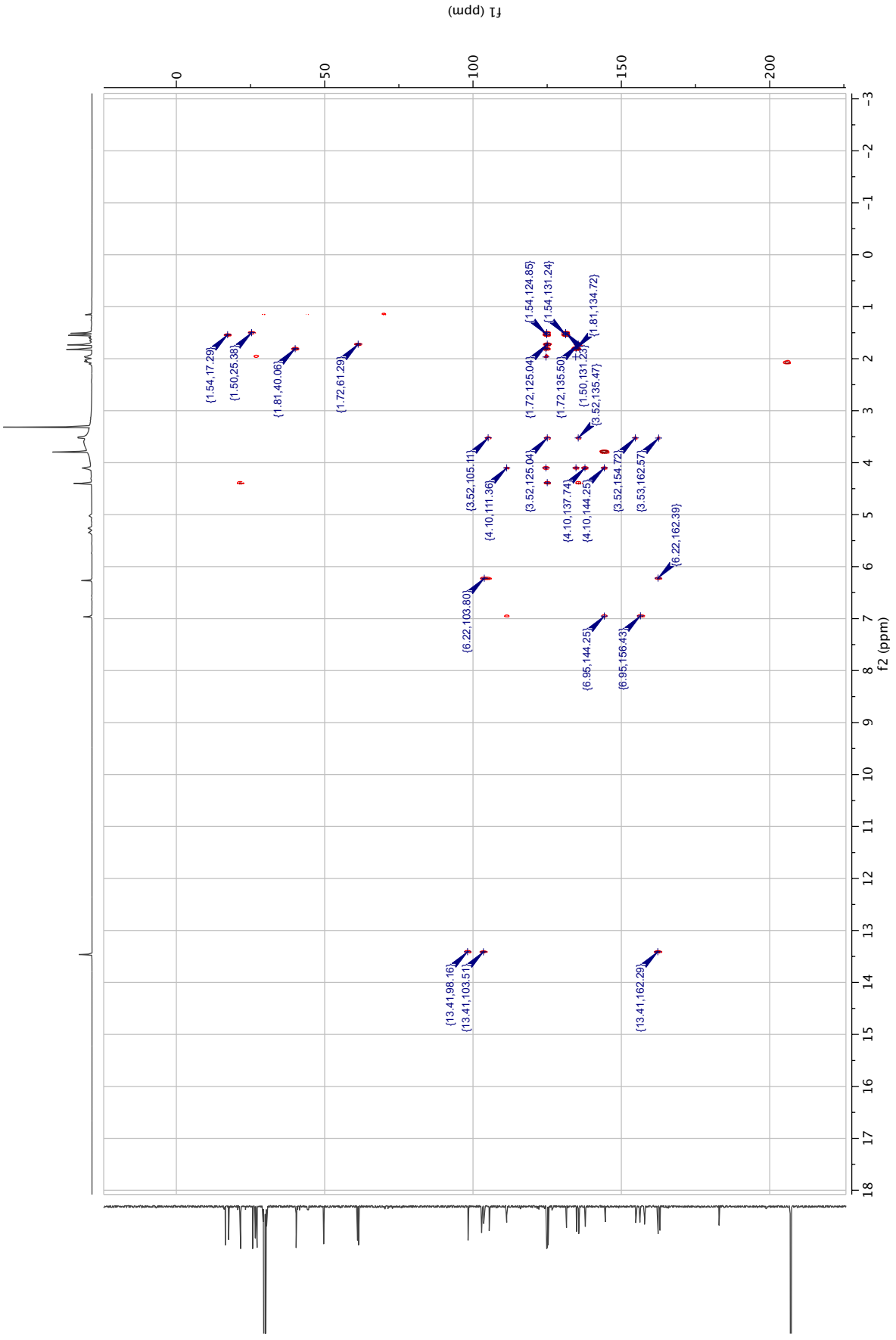


Figure S41. HMBC spectrum (4 scans) of isocowanol (**6**, 46.49 mg) recorded in acetone-*d*6.

Table S1. Composition of standard mixtures M1-M6. Stock solutions of each reference compound [amentoflavone (**1**), betulinic acid (**2**), ferulic acid (**3**) and 3-hydroxyxanthone (**4**)] were prepared at a concentration of 10 mg/mL. Each mixture (M1 to M6) was prepared at a final volume Vf = 200 µL. Mixing ratios (µL) were as follows:

	Amentoflavone (*1*)	**Betulinic acid (*2*)**	**Ferulic acid (*3*)**	**3-hydroxyxanthone (*4*)**
M1	50	50	50	50
M2	125	25	25	25
M3	25	125	25	25
M4	25	25	125	25
M5	25	25	25	125
M6	54	46	19	21

Table S2. List of theoretical and detected masses for compound by UPLC-MS and/or by LDI-MS.

Compound	Monoisotopic mass and Molecular formula	Experimental m/z (UPLC-MS)	Experimental m/z (LDI-MS)	Calculated m/z $[M-H]^-$	Mass error (ppm) (UPLC/LDI)
amentoflavone (1)	538.0900 ($C_{30}H_{18}O_{10}$)	/	537.0822 $[M-H]^-$	537.0827	0.93
betulinic acid (2)	456.3603 ($C_{30}H_{48}O_3$)	/	455.3556 $[M-H]^-$	455.3531	5.49
ferulic acid (3)	194.0579 ($C_{10}H_{10}O_4$)	/	193.0501 $[M-H]^-$	193.0506	2.59
3-hydroxyxanthone (4)	212.0473 ($C_{13}H_8O_3$)	/	211.0407 $[M-H]^-$	211.0401	2.84
Rubraxanthone (5)	410.1729 ($C_{24}H_{26}O_6$)	409.1654 $[M-H]^-$	409.1651 $[M-H]^-$	409.1657	0.73/1.46
Isocowanol (6)	494.2305 ($C_{29}H_{34}O_7$)	493.2228 $[M-H]^-$	493.2251 $[M-H]^-$	493.2232	0.81/3.85
Parvixanthone G (7)	428.1835 ($C_{24}H_{28}O_7$)	427.1758 $[M-H]^-$	427.1778 $[M-H]^-$	427.1762	0.93/3.74

Table S3. Predicted and experimental δ_C matching of refrence compounds 1-4 from mixture M1. [ACD/Labs NMR Predictors (C,H)]; experimental values[4]. *Carbons not matched and exceeding the 1.30 ppm tolerance of the in-silico database.

ppm	amentoflavone (1)			betulinic acid (2)			ferulic acid (3)			3-hydroxyxanthone (4)		
	Carbon numbering	Literature δ_c	δc predicted and matched by MixONat	Carbon numbering	Literature δ_c	δc predicted and matched by MixONat	Carbon numbering	Literature δ_c	δc predicted and matched by MixONat	Carbon numbering	Literature δ_c	δc predicted and matched by MixONat
182.2	C-4"	182.2	181.8									
181.8	C-4	181.9	181.7									
177.3				C-28	177.9	179.1*						
174.8										C-9a	174.8	174.9
168.0							C-9	167.7	169.0			
164.1	C-2"	164.3	164.1									
163.8	C-2	164.1	164.1							C-3	164.0	164.2
163.7	C-7	163.9	163.1									
161.8	C-7"	161.9	163.7									
161.5	C-5	161.6	161.6									
161.1	C-4""	161.0	161.5									
160.6	C-5"	160.8	160.1									
159.5	C-4'	159.6	159.7									158.8
157.6	C-9	157.6	157.5									
157.4										C-4a	157.6	
155.6										C-4b	155.6	155.7
154.5	C-8"	154.7	156.6									
150.4				C-20	150.1	150.7						
149.1							C-4	148.8	149.7			
147.9							C-3	148.3	148.5			
144.6							C-7	140.1	145.4			
134.9										C-6	134.9	134.9
131.5	C-6'	131.6	132.0									
128.2	C-2""; C-6""	128.3	128.3									129.3
128.1			128.8							C-1	128.0	
127.9	C-2'	127.9										
125.9							C-1	126.8	126.3	C-8	125.9	126.1
124.2										C-7	124.2	124.1
122.9							C-6	122.1	123.3			
121.4	C-3'	121.7	121.6									
121.2	C-1""	121.4	121.4							C-8a	121.2	121.2
120.0	C-1'	120.3										
117.9			116.9							C-5	117.9	117.8
116.2	C-5'	116.4										
115.8	C-3""; C-5""	116.0	116.1				C-8	116.1	115.9			
115.6							C-2	115.6	115.6			
114.2										C-2	114.2	113.8
114.0										C-9a	114.0	
111.1							C-5	111.2	111.4			110.9
109.7				C-30	109.4	109.7						
104.0	C-8"	104.1	103.9									
103.7	C-4a; C-4"a	104.0	103.5									
103.0	C-3	103.2										
102.6	C-3"	102.8										
102.2			102.9							C-4	102.1	104.4*
98.9	C-6"	99.1	99.0									
98.6	C-6	98.8	93.9*									
94.1	C-8	94.2	94.0									
76.8				C-3	76.7	77.8						
55.7							C-10	56.0	56.0			
55.4				C-17	55.3	56.2						
54.9				C-5	54.9	54.4						
49.9				C-9	49.9	50.6						
48.5				C-19	48.5	48.3						
46.6				C-18	46.5	48.2*						
DMSO-d6				C-8	40.2	40.8*						
42.0				C-14	41.9	42.5						
38.5				C-1	38.9	38.9						
38.3				C-4; C-13	38.4	37.9						
37.6				C-10	37.5	38.3						
36.7				C-7	36.7	37.3						
36.4				C-22	36.3	37.1						
33.9				C-16	33.9	34.4						
30.1				C-15	30.1	30.2						
29.2				C-21	29.1	30.3						
28.1				C-2	28.0	28.4						
27.2				C-23	27.1	26.9						
25.1				C-12	25.0	25.6						
20.5				C-11	20.4	20.9						
19.0				C-6	18.9	19.3						
18.0				C-29	17.9	18.4						
15.8				C-26	15.9	16.0						
15.8				C-24; C-25	15.6	16.1; 17.2*						
14.4				C-27	14.3	14.7						

[1](**1**) Fernandes, E. G. R. *et al.*, *Magn. Reson. Chem.* **1998**, *36*, 305–309, (**2**) Qunfang, Z. *et al.*, *J. Chin. Pharm. Sci.* **2000**, *9* (3), 131–133, (**3**) Xiang, M. *et al.*, *J. Med. Plants Res.* **2011**, *5* (9), 1685–1691, (**4**) Goh, S. H. *et al.*, *J. Nat. Prod.* **1992**, *55*, (10), 1415–1420.

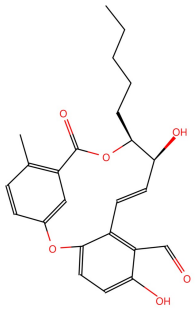
Table S4. Predicted and experimental δ_C matching of rubraxanthone (**5**) from fractions F6. δ_C were matched with predited chemical shifts [ACD/Labs NMR Predictors (C,H)] of compounds from the *Garcinia*DB. The comparison of experimental data with the literature[5] and chemical shifts proposed by MixONat confirmed the identification of rubraxanthone (**5**). *Carbons not matched and exceeding the 1.30 ppm tolerance of the in-silico database.

Carbon numbering	**Experimental**		**Literature**	**Dereplication by MixONat with "*Garcinia*DB"**
	δ_H (ppm)	**δ_C (ppm)**	**δ_C (ppm)**	**δ_C (ppm)**
C-9		182.7	182.8	182.8
C-3	9.44 (2H, s)	165.3	165.5	165.1
C-1	13.48 (1H, s)	164.9	165.0	165.1
C-4a		157.9	158.1	157.9
C-6	9.44 (2H, s)	157.4	157.6	154.1*
C-10a		156.2	156.4	155.4
C-7		144.5	144.7	142.5*
C-8		138.2	138.4	136.5*
C-18	1.83 (3H, s)	135.1	135.2	135.5
C-13		131.5	131.6	131.4
C-17	5.27 (1H, m)	125.1	125.3	125.2
C-12	5.02 (1H, m)	124.8	124.9	124.2
C-8a		111.9	112.2	112.3
C-9a		103.7	103.9	104.2
C-5	6.81 (1H, s)	102.8	102.9	101.6
C-2	6.18 (1H, d, *J* = 2.20 Hz)	98.7	98.9	98.5
C-4	6.28 (1H, d, *J* = 2.22 Hz)	93.8	93.9	93.8
OMe	3.79 (3H, s)	61.4	61.5	62.0
C-14	1.98 (2H, m)	40.4	40.5	40.5
C-16		27.3	27.4	27.4
C-11	4.10 (2H, d, *J* = 6.51 Hz)	26.8	26.9	26.8
C-19	1.55 (3H, s)	25.8	25.8	25.7
C-20	1.51 (3H, s)	17.7	17.8	17.6
C-15	2.02 (2H, m)	16.6	16.7	16.6

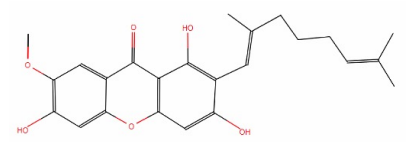
[2]Munekazu, I. *et al.*, *Chem. Pharm. Bull.* **1995** *44*, (1), 232–234.

Table S5. Excerpt of ProgenesisQI structure proposals for ion 409.1658 m/z (**5**). Suggestions are based on in-silico fragmentation of the PNMRNP database. The mass tolerance for precursor and theoretical fragmentation ions was set to 5ppm.

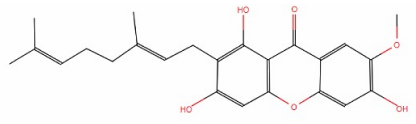
Rank	Compound (RT_MW)	Adducts	Formula	Score	Fragmentation Score	Mass Error (ppm)	Neutral mass (Da)	m/z	Structure
1	11.41_409.1658m/z	M-H	C24H26O6	49.8	62.7	0.44	/	409.1658	Hypothesis a
2	11.41_409.1658m/z	M-H	C24H26O6	48.5	55.9	0.44	/	409.1658	Hypothesis b
3	11.41_409.1658m/z	M-H	C24H26O6	48.4	55.7	0.44	/	409.1658	Hypothesis c
4	11.41_409.1658m/z	M-H	C24H26O6	48.4	55.5	0.44	/	409.1658	Hypothesis d
5	11.41_409.1658m/z	M-H	C24H26O6	48.3	55.1	0.44	/	409.1658	Hypothesis e
6	**11.41_409.1658m/z**	**M-H**	**C24H26O6**	**48.3**	**55.1**	**0.44**	**/**	**409.1658**	Rubraxanthone (**5**)
7	11.41_409.1658m/z	M-H	C24H26O6	48.3	55.0	0.44	/	409.1658	Hypothesis g
8	11.41_409.1658m/z	M-H2O-H	C24H28O7	47.5	50.8	0.43	/	409.1658	Hypothesis h



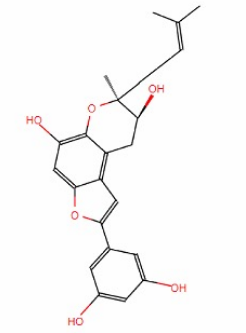
Hypothesis a



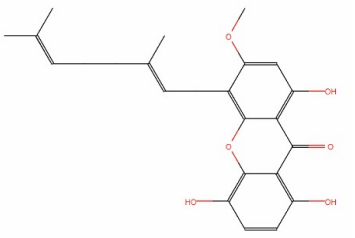
Hypothesis b



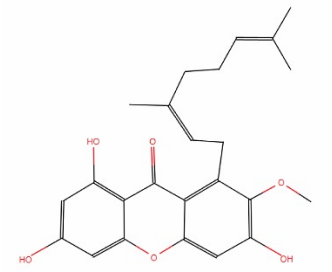
Hypothesis c



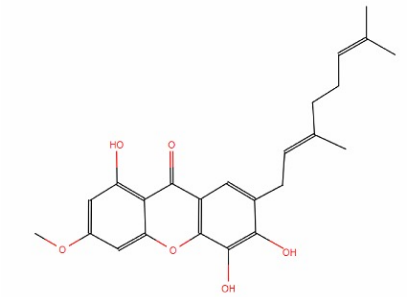
Hypothesis d



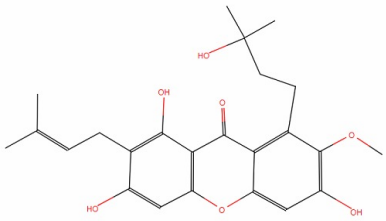
Hypothesis e



Rubraxanthone (**5**)



Hypothesis g



Hypothesis h

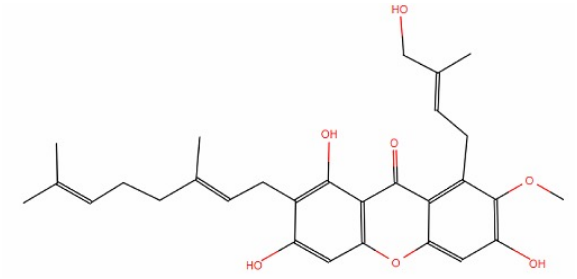
Table S6. Predicted and experimental δ_C matching of parvixanthone G (**7**) from fractions F10. δ_C were matched with predited chemical shifts [ACD/Labs NMR Predictors (C,H)] of compounds from the *Garcinia*DB. The comparison of experimental data with the literature[6] and chemical shifts proposed by MixONat confirm the identification of parvixanthone G (**7**). *Carbons not matched in the range of 1.30 ppm with the in-silico database.

Carbon numbering	**Experimental**		**Literature**	**Dereplication by MixONat with "*Garcinia*DB"**
	δ_H (ppm)	**δ_C (ppm)**	**δ_C (ppm)**	**δ_C (ppm)**
9		182.5	182.5	183.8
3		165.2	165.1	165.1
1	13.49 (1H, s, OH)	164.6	164.6	163.2
4a		157.7	157.8	158.0
6		157.2	157.3	156.7
10a		156.2	156.1	155.6
7		144.3	144.3	143.6
8		139.6	139.9	138.9
17		131.2	131.0	132.5
16	5.18 (1H, brt ?, *J* = 7.17 Hz)	126.0	126.0	123.9
8a		111.8	111.8	113.2
9a		103.6	103.5	103.7
5	6.83 (1H, s)	102.6	102.4	102.7
2	6.19 (1H, d, *J* = 2.11 Hz)	98.6	98.5	97.9
4	6.31 (1H, d, *J* = 2.06 Hz)	93.7	93.6	93.8
13		71.3	72.0	76.3*
7-OMe	3.86 (3H, s)	61.3	61.4	62.0
12	1.76 (2H, m)	43.1	43.3	42.9
14	1.58 (2H, m)	42.2	42.3	39.2
19	1.30 (3H, s)	27.2	27.2	25.7
18	1.67 (3H, s)	25.7	25.7	24.0*
15	2.19 (2H, m)	23.2	23.2	22.9
11	3.43 (3H, dd, *J* = 6.31 ; 10.55 Hz)	22.8	22.7	22.2
20	1.65 (3H, s)	17.7	17.6	17.6

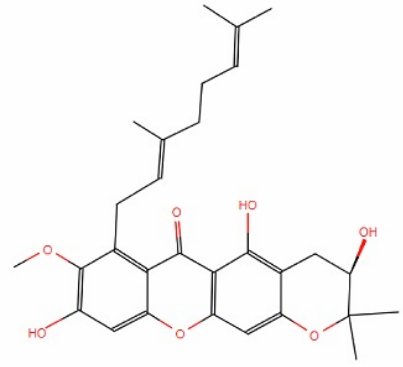
[3]Xu, Y.-J. *et al.*, *J. Nat. Prod.* **2001**, *64* (9), 1191–1195.

Table S7. Excerpt of ProgenesisQI structure proposals for ion 493.2228 m/z (**6**) based on in-silico fragmentation of the PNMRNP database. (precursor and theoretical fragmentation was set at 5ppm).

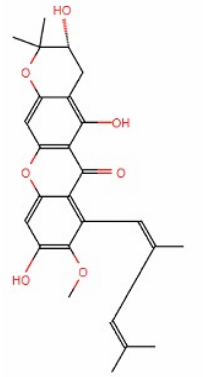
Rank	Compound (RT_MWn)	Adducts	Formula	Score	Fragmentation Score	Mass Error (ppm)	Neutral mass (Da)	m/z	Structure
1	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	50.1	58.4	-0.83	494.2300	493.2228	Hypothesis i
2	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	49.7	56.6	-0.83	494.2300	493.2228	Hypothesis j
3	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	49.7	56.6	-0.83	494.2300	493.2228	Hypothesis k
4	**11.22_494.2300n**	**M-H, 2M-H, M-H2O-H**	**C29H34O7**	**49.6**	**56.1**	**-0.83**	**494.2300**	**493.2228**	Isocowanol (**6**)
5	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	49.6	56.1	-0.83	494.2300	493.2228	Hypothesis m



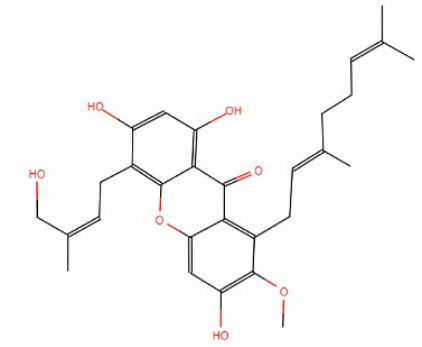
Hypothesis i



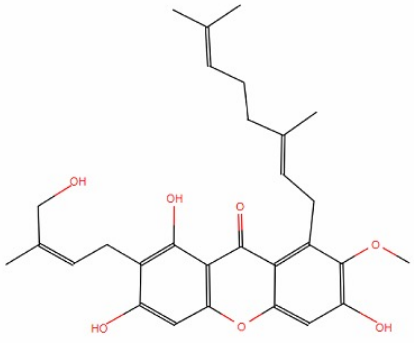
Hypothesis j



Hypothesis k



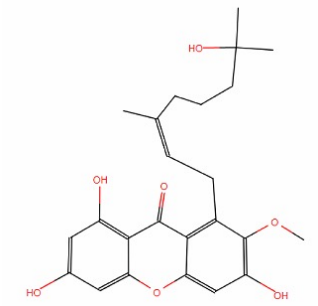
Isocowanol (**6**)



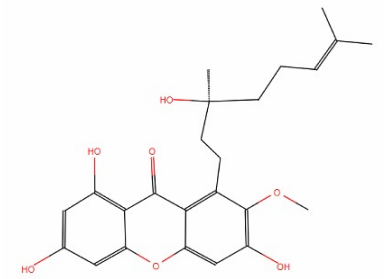
Hypothesis m

Table S8. Excerpt of ProgenesisQI structures propsals for ion 427.1760 m/z (**7**) based on in-silico fragmentation of the PNMRNP database. The mass tolerance for precursor and theoretical fragmentation ions was set to 5ppm.

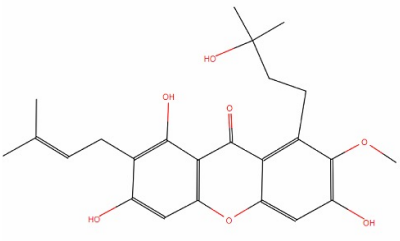
Rank	Compound	Adducts	Formula	Score	Fragmentation Score	Mass Error (ppm)	Neutral mass (Da)	m/z	Structure
1	8.42_428.1833n	M-H, 2M-H	C24H28O7	51.9	62.1	-0.47	428.1833	427.1760	Hypothesis n
2	**8.42_428.1833n**	**M-H, 2M-H**	**C24H28O7**	**51.2**	**58.6**	**-0.47**	**428.1833**	**427.1760**	Parvixanthone G (**7**)
3	8.42_428.1833n	M-H, 2M-H	C24H28O7	44.1	22.7	-0.47	428.1833	427.1760	Hypothesis p
4	8.42_428.1833n	M-H, 2M-H	C24H28O7	43.5	19.8	-0.47	428.1833	427.1760	Hypothesis q
5	8.42_428.1833n	M-H, 2M-H	C24H28O7	42.9	16.8	-0.47	428.1833	427.1760	Hypothesis r



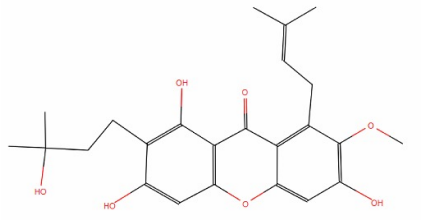
Hypothesis n



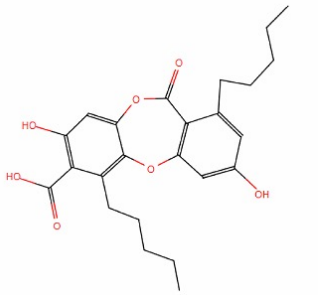
Parvixanthone G (**7**)



Hypothesis p



Hypothesis q



Hypothesis r

Table S9. ^{13}C and ^{1}H NMR data of isocowanol (**6**). Spectra were recorded in acetone-*d*6. Signals marked with (*) are interchangeable.

Carbon numbering	δ_H (ppm)	δ_C (ppm)	Dereplication by MixONat with "*Garcinia*DB"
9		182.9	183.5
3		163.0	163.1
1		162.4	162.3
6		157.8	154.0
10a		156.3	154.6
4a		154.9	153.7
7		144.6	142.7
3'		137.9	137.6
8		135.6	136.5
3''		134.9	136.0
7'		131.5	132.0
2''	5.33 (1H, t, *J* = 7.5 Hz)	125.3	123.0
6'	5.03 (1H, t, *J* = 6.9 Hz)	125.1	125.7
2'	5.25 (1H, t, *J* = 6.5 Hz)	124.8	124.5
8a		111.4	111.1
4		105.5	106.8
9a		103.6	101.2
5	6.97 (1H, s)	102.9	104.3
2	6.20 (1H, s)	98.3	98.7
4''	4.40 (2H, s)	61.4	61.9
O-Me	3.78 (3H, s)	61.1	61.2
4'	1.96 (2H, t)	40.4	40.1
5'	1.72 (2H, s)	27.2	26.8
1'	4.09 (2H, d, *J* = 6.4 Hz)	26.6	25.8
7'Me	1.54* (3H, s)	25.7*	26.9
3''Me	1.72 (3H, s)	21.7	21.7
1''	3.49 (2H, d, *J* = 7.7 Hz)	21.6	22.3
7'Me	1.49* (3H, s)	17.6*	17.9
3'Me	1.80 (3H, s)	16.5	16.5
1-OH	13.44 (1H, s)		

4. Dereplication strategies using molecular networks

This section provides an overview of dereplication strategies using MNs, highlighting their importance for the identification of known compounds, elucidating structural relationships, and the prioritization of uncharacterized molecules.

In a first step, MS/MS-based MN is merged with chemometrics and ^{13}C NMR-based dereplication to detect active compounds in complex mixtures prior to the time-consuming isolation process. By combining these different strategies, several potentially active molecules in *G. parvifolia* (KL5670) were annotated and consecutively isolated. Eventually the predicted activity was confirmed by an anti-AGE assay thus providing proof of concept for the approach.

A notable drawback of LDI-MS experiments is the fact that isomers cannot be distinguished (see section 3.2). At least for the studied xanthones this limitation could be bypassed by LDI-IMS-MS^2. In addition, the creation of a Xanthones-DB repository on the GNPS platform has facilitated the putative annotation of compounds of the KL5073 extract. This integrated approach enhances the ability to distinguish isomeric compounds and aids in the characterization of NPs.

4.1. Looking for actives in the haystack: merging HRMS^2-based molecular networking, chemometrics and ^{13}C NMR-based dereplication approaches

This publication will be submitted in the Journal of Natural Products.

4.1.1. Publication

Looking for actives in the haystack: Merging HRMS²-based molecular networking, chemometrics and ^{13}C NMR-based dereplication

Manon Meunier,[a] *Dimitri Bréard,*[a,b] *Séverine Boisard,*[a,b] *Patricia Blanchard,*[a] *Marc Litaudon,*[c,d] *Khalijah Awang,*[d] *Andreas Schinkovitz,*[a,*] *Séverine Derbré*[a,*]

[a] Univ Angers, SONAS, SFR QUASAV, F-49000 Angers, France

[b] PHYTO Platform, SFR QUASAV, Campus du Végétal, 49045, Angers CEDEX 01, France

[c] Institut de Chimie des Substances Naturelles, CNRS-ICSN, UPR 2301, Université Paris-Saclay, 91198 Gif-sur-Yvette, France

[d] IFM NatProLab, University of Malaya, Faculty of Sciences, Department of Chemistry, Malaysia

KEYWORDS. Advanced glycation end-products, Annotation, Bioactivity-based molecular networking, Clusiaceae, Dereplication, *Garcinia parvifolia*, *In silico* databases, MixONat, Tetraprenyltoluquinone, Xanthones

ABSTRACT. The identification of bioactive natural products (NPs) in complex mixtures is a challenging task. To address this challenge, the present manuscript proposes an integrated strategy that combines tandem mass spectrometry (MS/MS) based molecular network (MN),

chemometrics [variable importance of projection (VIP) score and regression coefficient (RC) from partial least squares (PLS) model], and ^{13}C NMR based dereplication. The aim was to reliably identify active NPs against advanced glycation end products (AGEs), in complex mixtures. A *Garcinia parvifolia* bark extract was chosen as a working example because of its content of prenylated xanthones and its inhibitory effect on AGEs formation. Results from this study demonstrate the effectiveness of the integrated approach in selecting metabolites of interest and annotating them with higher level of confidence when combined with NMR data. Overall, this comprehensive approach provides a powerful and efficient solution for targeting and annotating of active compounds in complex NPs mixtures.

Natural products (NPs) comprise a wide range of chemically diverse compounds that interact with different targets of therapeutic or agronomic interest. Therefore NPs chemists focus on the identification of bioactive compounds in complex matrices such as medicinal plants or microorganisms extracts.[1–3] For a long time, a classical approach, known as "bioassay-guided fractionation" (BGF) has been used to address this challenge.[4] However, the successive fractionation steps associated with systematic bioactivity evaluation are costly and require large amounts of raw material. In addition, the isolation of previously reported NP, sometimes already known for this kind of activity, is not uncommon so considerable efforts are made to avoid this situation.[5] By default, the strategy favors the isolation and structural characterization of most abundant bioactive NPs, while low yield compounds are easily missed out. To bypass these limitations, BGF was extended by simultaneously conducted untargeted metabolomics. The latter are either based on mass spectrometry (MS) or ^{1}H nuclear magnetic resonance (NMR) spectroscopy, which are further linked to bioactivities resulting in a technique that is known as "biochemometrics".[6–8] In this case, the bioactive NPs are directly spotted in complex mixtures and then selectively isolated for full spectral characterization and identification.

Indeed, independently from the applied analytical method, the annotation of selected compounds in complex mixtures remains a major challenge in untargeted metabolomics. This last step relies on search algorithms that compare the spectral data obtained with those of metabolites in a given database (DB) and finally assign the best scores to the most appropriate structural hypotheses. These strategies used either experimental MS² [*e.g.*, global natural product social (GNPS)] / NMR (*e.g.*, NP-MRD) repositories or *in silico* DBs.[9,10] In particular, as part of non-targeted metabolomics, the dereplication – *i.e.,* the unambiguous identification of known NPs responsible for the activity of an extract before BGF[11,12] – of compounds in complex matrices from

living organisms has been the focus of researchers involved in the identification of bioactive NPs for almost twenty-five years.[13] Today, it is estimated that there are more than 500,000 NPs from microbial, marine and terrestrial sources described in the literature[14]. Consequently, and even though started, the comprehensive collection of experimental NMR[10,15] and MS[16–19] data is an enormous task that requires a continuous and collective effort. Alternatively, DBs providing computationally generated *in silico* spectra become increasingly important with regards to the constant progress of spectral prediction algorithms and software [*e.g.* CFM-ID4[20,21], MetFrag[22], ACD/NMR Predictors (C and H)[23], nmrshiftdb2[24,25]]. Consequently, the use of such DBs containing predicted fragmentation spectra or ^{13}C-NMR chemical shifts (δ_C) allow relevant structural assignments.[26–32] When considering the dereplication of complex mixtures of NPs, it should be remembered that, according to the metabolomics standard initiative (MSI)[33], metabolites are annotated with a confidence level. While level 0 refers to the 3D structure of a NP that was unambiguously confirmed by analytical methods, level 1 requires either a reference standard or at least two orthogonal techniques to propose a confident 2D structure. Level 2 indicates a probable metabolite by matching experimental with literature data or DBs. Level 3 is assigned to a possible structure or class of NPs, while an unknown feature receives level 4 or 5 (*i.e.*, 4: molecular formula; 5: exact mass).[9,34] As a result, level 2 can be achieved by annotating NPs using *in–silico* DBs, while level 1 necessarily requires two identification methods (*e.g.*, MS and NMR).

Over the past decade, automated dereplication tools and software based on either high-resolution MS² (HRMS²) associated with molecular networking (MN)[35] or ^{13}C-NMR[36,37] have emerged to efficiently identify known NPs in mixtures. While the main advantage of HRMS² is its sensitivity, less sensitive ^{13}C-NMR may discriminate between stereoisomers. As previously

stated, *in-silico* and experimental DBs exist for both methods and may be used to putatively assign known NPs in complex mixtures. However, this can only be achieved at a lower confidence level (*i.e.*, 2). With this in mind, the present work highlights the interest of combining MS- and NMR-based approaches to increase confidence in annotation towards level 1.

Aiming at the reliable identification of active NPs in crude mixtures, an efficiency strategy merging HRMS² MN, chemometrics, ^{13}C-NMR-based dereplication approaches as well as *in silico* DBs is presented. As a working example, an extract of *Garcinia parvifolia* (*Clusiaceae*) bark (KL5670) was selected among four chemotypes based on its content of prenylated xanthones (*i.e.*, stereoisomers) and its inhibitory effects on the formation of advanced glycation end-products (AGEs).[38]

In 2018, L.-F. Nothias *et al.* introduced the concept of bioactivity-based MN.[7] After fractionation of the raw extract to get relevant biological assay, the bioactivity of the fractions was evaluated. Then, the approach combined tandem mass spectrometry (MS/MS) MN with bioactivity scoring to spot active NPs. In this inaugural publication the authors calculated a bioactivity score based on the Pearson correlation between feature intensity across samples and the bioactivity level associated with each sample. It should be noted that another common and effective method to select metabolites/NPs of interest is partial least squares (PLS) analysis, which highlights the correlation between independent variables (spectral data (X)) and a dependent variable (activity data (Y)). Regarding the present work, PLS regression has previously successfully identified AGEs inhibitors directly from crude bark extracts of *G. parvifolia*.[39] In a publication by Nothias *et al.*, a bioactivity-based MN was obtained using Cytoscape before the targeted isolation of predicted bioactive NPs and their unambiguous identification. Alternatively, the present workflow first selects metabolites of interest using HRMS² and a

bioactivity-based MN using chemometrics [*i.e.*, variable importance of projection (VIP) score and regression coefficient (RC) from PLS model][7]. Secondly, the annotation is performed using the *in silico* fragmenter MetFrag included in the Progenesis QI software[22], ^{13}C-NMR based dereplication of fractions of KL5670 using the open access MixONat software[37] and chemotaxonomically based DBs of NPs including their predicted δ_C (Figure 1).[40] Finally, the accuracy of the proposed model was evaluated by isolating those activity markers that were identified and annotated in extracts and fractions.

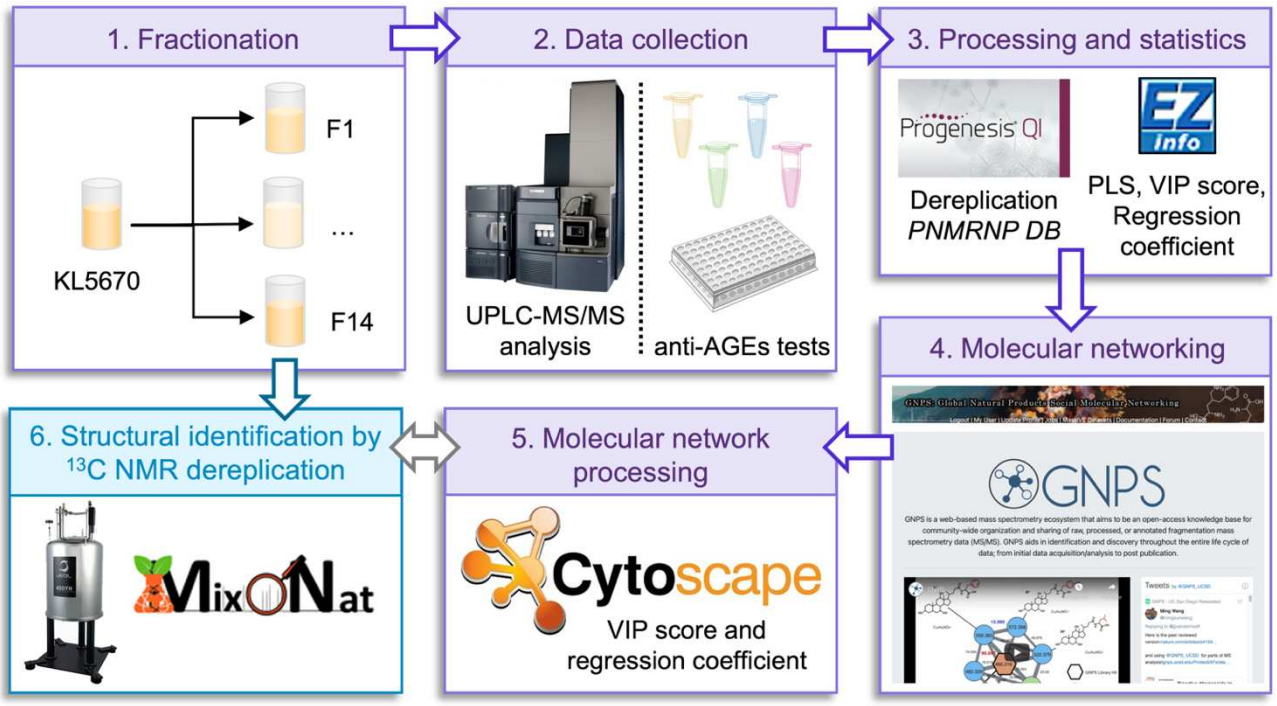


Figure 1. Workflow for an activity-based molecular network approach integrating statistical data (VIP score and RC). (**1**) The crude extract KL5670 was fractionated into 14 fractions. (**2**) Fractions were analyzed by UPLC-MS/MS and evaluated by an anti-AGEs assay. (**3**) Data were pre-processed by ProgenesisQI (DB: PNMRNP) and successively exported into EZinfo for statistical processing. (**4**) Pre-processed data were also exported to the GNPS web-platform for MN. (**5**) The obtained MN was then formatted and merged with statistical data by using Cytoscape. (**6**) Fractions containing NPs of interest systematically underwent ^{13}C NMR dereplication to annotate major constituents (MixONat software, DB: PNMRNP).

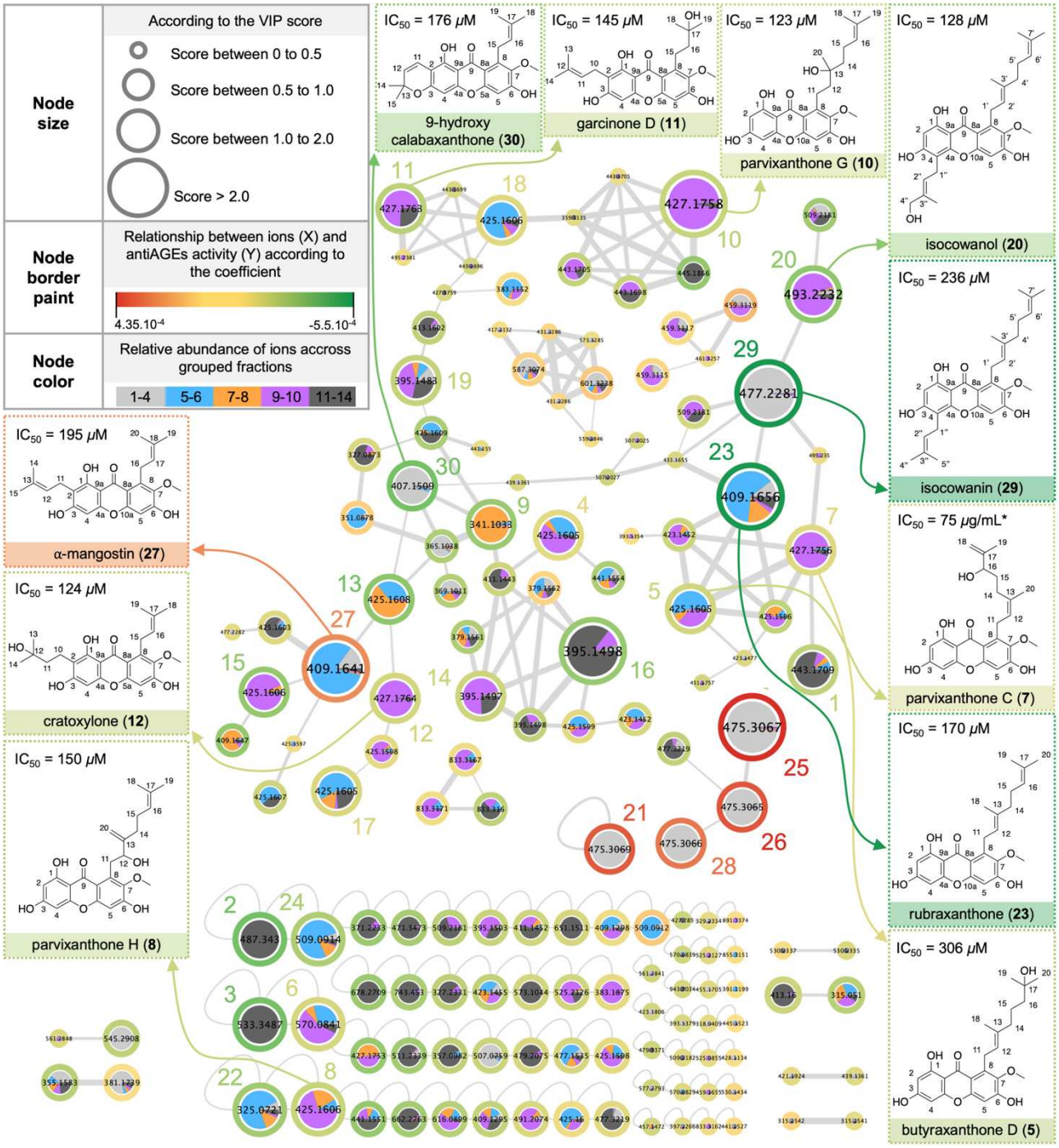


Figure 2. Active-based MN of the *Garcinia parvifolia* bark extract KL5670 after its fractionation and the evaluation of AGE inhibition. The node partitioning corresponds to the relative abundance of ions across grouped fractions (F1-F4: light grey; F5-F6: blue; F7-F8: orange; F9-F10: purple; F11-F14: dark grey). Nodes/ions/NPs that indicate predicted activity markers are surrounded by a green circle. Red cycles highlight putatively inactive compounds. The most abundant ions are the largest nodes. As a proof of concept, major NPs were further systematically

isolated to confirm their and anti-AGE activities (expressed as IC_{50}) For additional structures, see supporting information, Figure S2.

RESULTS AND DISCUSSION

Selection of the main active NPs in fractions of *G. parvifolia*. The main objective of the present study is to propose a workflow to identify *active* NPs within a crude extract as early as possible together with confident annotation levels without isolation. In order to select NPs to focus on, an active MN was generated. The latter was based on UPLC-MS/MS data of the fractions obtained from the previously described active crude extract of *G. parvifolia*, KL5670.[38] Fractions were evaluated on their inhibitory effect on AGE formation (supporting information, Table S1). In this regard, four processing steps were required: First, raw UPLC-MS/MS data were imported to ProgenesisQI (Waters) to conduct alignment, peak picking, and deconvolution. Optionally, a first annotation of data using the predicted carbon-13 NMR data of natural products database (PNMRNP3) DB[40,41] (**i**) was performed . Then, processed data from ProgenesisQI were exported to the GNPS web-platform by using the feature-based MN workflow (FBMN) (**ii**). Once created, MS/MS data were also exported from ProgenesisQI to the EZinfo statistical software allowing to obtain VIP scores and regression coefficients (RC) (supporting information, Table S2) for each feature of a PLS model[38] (**iii**). A VIP score measures the variable's importance in the PLS model. The higher the VIP score of a feature, the more important it is for the applied model and therefore interesting to study. On the contrary, the lower the VIP score of an ion, the less interesting it is. In the obtained MN (Figure 2), the VIP score is represented by the node size. Important nodes are bigger. On the other hand, the RC indicates the direction of the relationship between variables (m/z) and the response variable (activity). Nodes in green are ions representing

potentially active metabolites, whereas red nodes correspond to potentially non-active metabolites. Finally, chemometrics data (VIP scores and RC) of each feature were imported into Cytoscape to obtain the final MN shown in Figure 2 (**iv**). It contains 154 major NPs distributed in fourteen fractions. The node partitioning corresponds to the relative abundance of ions in the grouped fractions. According to the PLS model from a previous project (supporting information, Figure S1)[38], those NPs with highest predicted activity against AGE formation are found in F5-F6 (blue) and F9-F10 (purple). The fractions containing the NPs least involved in the activity are fractions F11 to F14 (grouped in dark grey) and fractions F1 to F4 (grouped in light grey). Fraction F7-F8 (orange) comprise NPs with moderate activity. Concerning the MN in Figure 2, NPs most likely to be active are those with a high VIP score (> 2.0) and negative RC (as the most active fractions are those with the smallest IC_{50}). Consequently, the nodes/ions/NPs expected to contribute the most to the activity are the largest and are surrounded by darker or lighter green circles. Red circles indicate inactive ions. In the present work, our attention was initially focused on active NPs, *i.e.*, nodes **23** and **29** (dark green circle, VIP score > 2.0) respectively in fractions F6 and F3 and then nodes **1**, **9**-**11**, **13**, **15**-**16**, **19**-**20**, **30** (light green circle, VIP score > 1.0) mostly from fractions F2, F7, F9, F10 and F12. (Figure 2). As a result, our annotation efforts focused mainly on these fractions, and in particular on these NPs.

It should also be noted that the MN also points out abundant NPs that are not involved in the activity (red circle, VIP score > 1.0), *i.e.*, nodes **21** and **25**-**28** in fractions F4-F5.

Annotation of active NPs. Once the NPs of interest have been selected, the step of annotating them with a sufficient level of confidence is one of the most challenging. Structural annotation consists of matching experimental spectral data with those of NPs collected in DBs using dedicated algorithms. At this stage, an untargeted UPLC-$HRMS^2$ analyses has been carried out.

The acquired MS/MS data were then used for a first annotation on the GNPS platform.[42] These initial results were quite disappointing: from all detected ions, only seven were assigned to a xanthone, namely γ-mangostin with cosine scores between 0.61 and 0.81 (Node **19**, Figure 2), while fourteen were attributed to α-mangostin with cosine scores varying between 0.62 and 0.88 (Node 27, Figure 2) (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2ae86a52caf04ebcbf397df1bd6946cb). At the time of the study, current community libraries provided very little information on negatively ionized NPs and no DB for MS/MS spectra on natural xanthones existed. Moreover, the large variability of data concerning experimental parameters and MS instrument settings significantly impairs structural annotation.[43]

With this in mind, a second approach was based on *in silico* MS/MS DBs. The annotation of the NPs (Table 1, NPs highlighted in bold) was carried out using the acquired experimental UPLC-HRMS² data, the ProgenesisQI (Waters) software and a large DB the predicted NMR natural products (PNMRNP3) DB[40,41]. The latter was chosen because it contains both the structures of 211,280 NPs as well as their predicted carbon-13 chemical shifts, which are classified by carbon type (*i.e.*, quaternary carbons, methine, methylene and methyl). It is therefore compatible with the MixONat software, which was subsequently used (see below). The dereplication processing of ProgenesisQI uses the MetFrag algorithm[22]. For node **23** (retention time (RT) 6.28 min), 64 possible structures were suggested with fragmentation scores ranging from 62.7 to 3.53 (precursor tolerance: 5 ppm; theoretical fragment tolerance: 10 ppm), including 38 xanthones. The first hypotheses proposed stereoisomers of a geranylated 1,3,6-trihydroxy-7-methoxyxanthone scaffold, where the geranyl substituent is attached to different carbons (non-exhaustive list in supporting information, Table S3). Considering that node **29** (RT 7.78 min) is

also a putative inhibitor of AGE formation, it was also analyzed. Thirty-four structures were proposed including 12 xanthones. Those with the best scores were stereoisomers of the same 1,3,6-trihydroxy-7-methoxyxanthone substituted by a prenyl and a geranyl group (non-exhaustive list in supporting information, Table S4). For nodes **23** and **29**, it was impossible to assign a specific structure based on MS/MS data with a confidence level of 1 (*i.e.*, confident 2D structure). Similar conclusions (confidence levels from 2 to 4) were drawn for the proposals for other nodes of interest (*i.e.*, the largest nodes encircled in green).

While the use of a very large DB of NPs such as PNMRNP allows exploring a wide chemodiversity, its extensive number of entries also makes annotation decisions more difficult. This issue may be addressed by using smaller DBs and by including chemotaxonomic data. For example, the features of the open access LOTUS DB allow downloading the structures of previously isolated NPs from organisms of the genus or family under study.[44] In addition, ^{13}C NMR data may be used to refine results for major NPs. This aspect will be discussed in the following sections.

Annotations strategies based on NMR. If a standard is not available, the confidence level can be improved by combining at least two orthogonal techniques to propose a relevant and confident 2D structure.[9,34] In 2019, in their manuscript dedicated to the problem of identifying bioactive NPs in complex mixtures, Wolfender *et al.*,[45] proposed to combine LC-HRMS² and NMR data with predicted spectra from DBs and chemometrics. Available since 2020 the open-source software MixONat can be used to identify the main NPs of a complex mixture (crude extracts or fractions) based on ^{13}C NMR and, optionally, DEPT-135 and 90 data. DEPT experiments permit to distinguish between carbon types (quaternary carbons, methines, methylenes, methyles).[37,46] It can use any DB containing the structures of NPs of interest (*e.g.*, on the basis of

chemotaxonomic or structural criteria) for which the chemical shifts are experimental or, alternatively, have been predicted.[30,47] Contrary to MS, ^{13}C NMR data are, except for solvents, far less affected by experimental and instrument parameter. This allows distinguishing stereoisomers, which is a crucial benefit for the current study. In addition, and despite being less sensitive than MS, ^{13}C NMR is a global method that allows the simultaneous analysis of all the major organic compounds in a mixture independently from their chemical nature. It therefore is an important complement to UPLC-HRMS², as NMR spectrometers now provide useful data sets within a reasonable timeframe. With that in mind, a ^{13}C NMR-based dereplication strategy using MixONat and ^{13}C NMR DB information on NPs that were previously described for the *Garcinia* genus was applied on fractions of KL5670. The idea was to refine the putative annotation of NPs that is based on UPLC-MS² experiments. Results from these experiments are discussed in the next paragraph.

The hyphenation of MS² and ^{13}C NMR for refined annotation. As nodes **23** and **29** show major compounds from fractions F6 and F3 (Figure 1 and Table 1), ^{13}C NMR and DEPT-135 and 90 experiments (supporting information, Figures S3 to S23) of these samples were carried out. Of the 718 *Garcinia* DB NPs, using molecular weight filters at 410 and 478 Da respectively, MixONat suggests rubraxanthone (**23**, score 0.88) in F6 and isocowanin (**29**, score 0.86) at rank 1 and 2 respectively (supporting information, Figures S24-S25). The MN (Figure 2) and a careful comparison of the experimental spectral data with those from the literature (Supporting information, Tables S5-S6) confirms this assignment.[48] Orthogonal MS and ^{13}C NMR data lead to an improved level of confidence from 2 to 1. We propose to assign intermediate confidence levels: level 1b refers to confirmed 2D structure by two orthogonal methods (*i.e.*, HRMS and ^{13}C NMR) while level 1a corresponds to a more confident level when data are confirmed by an additional

comparison with literature. In this case, rubraxanthone (**23**) and isocowanin (**29**) were annotated with level of confidence 1a as δ_C of these major xanthones in fractions F6 and F3 were those previously described. In F3, MixONat also suggests tetraprenyltoluquinone (TPTQ, score 0.81, rank 16) as a major NP when no molecular weight filter is used (Supporting information, Figure S26). This was confirmed by a meticulous analysis of the spectral data (Supporting information, Table S7).[49] It should be noted that TPTQ does not appear in the MN produced from UPLC-HRMS² data in negative mode (Figure 2, Table 1). This highlights the benefits of a complementary ^{13}C NMR analysis.

Following the same strategy, major NPs in fractions F2, F7, F9, F10 and F12 were putatively identified by merging MS/MS and ^{13}C NMR-based annotations with various confidence level (Table 1). For F2, MixONat places 9-hydroxycalabaxanthone (**30**) in rank 1 (score 0.88) in line with results from ProgenesisQI using the PNMRNP DB (rank 6/30 data not shown. Noteworthy: Eleven of the thirty candidates 3,3-dimethylpyranoxanthones substituted by two hydroxyl groups, one methoxyl group and one prenyl group). The assignment was further confirmed by δ_C from the literature[50] (Supporting information, Figures S27-S45, Table S8) allowing a level of confidence 1a (Table 1). F7 contains two NPs potentially involved in the activity (**9** and **13**) for which the molecular formulas $C_{19}H_{18}O_6$ and $C_{24}H_{26}O_7$ have been deduced from the HRMS data. Following both the hypotheses of ProgenesisQI [data not shown: trihydroxy,methoxy,prenylxanthone (**9**) and dihydroxy,methoxyxanthone bearing either one geranyl or two prenyl groups (**13**)] and MixONat (molecular weight filters set at 342 and 426 $g.mol^{-1}$), these NPs can be tentatively annotated as dulxanthone (**9**) and mangostanin (**13**) respectively (Supporting information, Figures S46-S58). As no comparison was possible with the δ_C of these products analyzed in the same deuterated solvent, a confidence level of 1b is proposed

for these NPs which were suggested by both UPLC-HRMS² and ^{13}C NMR data and their associated annotation tools (Table 1). **19** (molecular formula: $C_{23}H_{24}O_6$, tetrahydroxygeranylxanthone) is a putative active NP found in F9. While using a molecular filter at 396 g.mol^{-1}, MixONat proposed smeathxanthone A (**14**), and cheffouxanthone (**19**) as first options in F9 (scores: 0.78 and 0.74) leading to an annotation 1b for these NPs. Two other major NPs were in F9 with $C_{24}H_{26}O_7$ as a molecular formula. While applying a molecular weight filter of 426 g.mol^{-1}, MixONat suggested parvixanthone H (**8**, 1st rank, score 0.83 respectively, level 1a) (Table S9)[51] and mangostenol (**18**, 5th rank, score 0.67, level 1b) (Supporting information, Figure S59-S72). Also putatively active, NPs corresponding to nodes **10**, **11** (molecular formula: $C_{24}H_{28}O_7$) and **20** (molecular formula: $C_{29}H_{34}O_7$) were mainly in F10 together with nodes **5**, **7** and **12**. Thus, the MixONat mass filter was set to 428 and 494 g.mol^{-1}. In agreement with ProgenesisQI first proposals (Supporting information, Tables S10-S11), the software proposed parvixanthone G (**10**, 3rd rank, score 0.83, level 1a)[51] (Table S12) and isocowanol (**20**, 1st rank, score 0.86, level 1)* as major NPs (Supporting information, Figures S73-S85). Garcinone D was also suggested (2nd rank, score 0.83). After a careful examination of ^{13}C NMR data, both garcinone D (node **11**) and its stereoisomer cratoxylone (node **12**) were putatively identified with a level of confidence 1a in F10 (Table S14). Finally, in F12, nodes **1** ($C_{24}H_{28}O_8$) and **16** ($C_{23}H_{24}O_6$) were predicted by ProgenesisQI as a dihydroxylated 3,6,8-trihydroxy-2-méthoxy-1,7-diprenylxanthone and a tetrahydroxyxanthone substituted by a geranyl group respectively (data not shown). While setting MixONat mass filter was set to 396 and 444 g.mol^{-1}, the software ranked 1,3,5,6-tetrahydroxy-7-geranylxanthone (**16**)

* Isocowanol was annotated with a level of confidence 1. The ^{13}C NMR spectrum was acquired in acetone-*d*6 but no data were available in the literature consequently it was purified and characterized using 1D and 2D NMR experiments.

and mangostenone E (**1**) at position 1 and 11 respectively (scores 0.96 and 0.71). These NPs were annotated with a level of confidence 1b (Supporting information, Figures S86-97).

Table 1 summarizes the levels of confidence in the annotations by crosslinking the proposals from MS data and ProgenesisQI (*i.e.*, Metascope tool using MetFrag) with the PNMRNP DB and those from experimental δ_C and MixONat software with the chemotaxonomic *Garcinia*DB.

To evaluate the approach, major and putatively active NPs were isolated from the fractions. This allowed confirming the putatively assigned structures of **10**, **11**, **20**, **23**, **29**, and **30** at a confidence level of 1. In addition, further NPs, such as mentioned examples were annotated with a confidence level 1a (nodes **5**, **7**).

Concerning the predicted inhibitory on AGE formation results for isolated compounds are summarized in Figure 2. As expected, among selected, purified and tested xanthones, **10**-**11**, **20**, **23**, **29**-**30** inhibit AGE formation with an IC_{50} below 250 µM [from parvixanthone G (**10**): 123 µM to isocowanin (**29**): 236 µM; aminoguanidine (positive reference): 1.3 mM]. It should also be noted that parvixanthones C and H (**6**, **8**) and cratoxylone (**12**) are strong inhibitors (IC_{50} < 150 µM). On the active molecular network, the corresponding nodes were medium-sized but encircled by light green, which suggests moderate activity. This observation illustrates the significance of both, the importance of the variable in the effect (VIP scores, node size), and the direction of this effect (RC, circle color). The selection of nodes of interest is based on the independent variables (m/z) involved in the effect on the dependent variable (activity). Consequently, it should be based both on VIP scores and RC and not limited to large node encircled in dark green. In fact, it is necessary to consider that, in complex mixtures, each of the NPs present at a given concentration can exert its own effect to contribute to the global biological effect of a fraction/extract. Multivariate analyses such as the PLS approach used in this study offer an advantage over traditional BGF by

simultaneously annotating a set of potentially active independent variables/NPs involved in an activity. However, it doesn't completely prevent missing some active NPs. As an example, α-mangostin (**27**, IC_{50} 195 µM) from F4 appears as a large node surrounded by an orange circle suggesting an inactive compound. This inconsistency in the assignment is probably related to the presence of inactive NPs **21**, **25**, **26**, **28** in large quantities in the same fraction (F4). This may decrease in the activity of the fraction.

Table 1. List of the most abundant ions/nodes with m/z ratio and retention time in fractions and their AGE inhibition. Annotation levels were proposed according to the putative identification by ProgenesisQI (PNMRNP DB containing predicted spectral data) and / or MixONat (GarciniaDB). The additional colour code (putatively active: green; putatively inactive: red) is the one of the RC. NPs highlighted in bold are those putatively involved in the inhibition of AGEs of fractions according to VIP score (> 1.0) and RC (green circle) in Figure 2.

Node (MN)	Compound name	m/z	Molecular formula	Retention time (min)	Fraction	Activity (µg/mL) of the corresponding fraction	Progenesis score***	MixONat score	Annotation level
1	**Mangostenone E**	**443.1709**	**$C_{24}H_{28}O_8$**	**2.66**	**F12**	**124**	**60.6**	**0.71**	**1b**
2	Glyyunnansapogenin B	487.3430	$C_{30}H_{48}O_5$	3.25	F13	147	42.9	/	2
3	Passifloric acid	533.3487	$C_{31}H_{50}O_7$	3.25	F13	147	84.7	/	2
4	Garcinisidone A	425.1605	$C_{24}H_{26}O_7$	3.75	F5	145	56.5	0.75	1b
5	Butyraxanthone D	427.1756	$C_{24}H_{28}O_7$	3.92	F10	82	47.1	/	2
6	NI	570.0841	$C_{27}H_{21}NO_{11}$	3.97	F10	82	12.2	/	4
7	Parvixanthone C	425.1605	$C_{24}H_{26}O_7$	3.97	F10	82	61.2	0.75	2
8	Parvixanthone H	425.1606	$C_{24}H_{26}O_7$	4.09	F9	101	44.4	0.83	1a
9	**Dulxanthone D**	**341.1033**	**$C_{19}H_{18}O_6$**	**4.41**	**F7**	**105**	**32.4**	**0.63**	**1b**
10	**Parvixanthone G**	**427.1758**	**$C_{24}H_{28}O_7$**	**4.44**	**F10**	**82**	**57.7**	**0.83**	**1a**
11	**Garcinone D**	**427.1763**	**$C_{24}H_{28}O_7$**	**4.66**	**F10**	**82**	**62.1**	**0.83**	**1a**
12	Cratoxylone	427.1764	$C_{24}H_{28}O_7$	4.78	F10	82	41.0	0.75	1a
13	**Mangostanin******	**425.1608**	**$C_{24}H_{26}O_7$**	**5.33**	**F7**	**105**	**37.1**	**0.67**	**1b**
14	Smeathxanthone A	395.1497	$C_{23}H_{24}O_6$	5.40	F9	101	58.6	0.78	1b
15	Parvixanthone B	425.1606	$C_{24}H_{26}O_7$	5.43	F5	145	45.1	0.83	1b
16	**1, 3, 5, 6-tetrahydroxy-7-geranylxanthone**	**395.1498**	**$C_{23}H_{24}O_6$**	**5.55**	**F12**	**124**	**53.1**	**0.96**	**1b**
17	Mangostanol	425.1605	$C_{24}H_{26}O_7$	5.60	F5	145	36.0	0.75	1b
18	Mangostenol	425.1606	$C_{24}H_{26}O_7$	5.77	F9	101	64.4	0.67	2
19	**Cheffouxanthone**	**395.1483**	**$C_{23}H_{24}O_6$**	**6.11**	**F9**	**101**	**35.6**	**0.74**	**1b**
20	**Isocowanol****	**493.2232**	**$C_{29}H_{34}O_7$**	**6.23**	**F10**	**82**	**41.4**	**0.86**	**1**

21	NI	475.3069	$C_{27}H_{42}O_4$	6.28	F4	615	/	/	3
22	Toxyloxanthone B	325.0721	$C_{18}H_{14}O_6$	6.28	F6	58	40.6	/	2
23	**Rubraxanthone**	**409.1656**	**$C_{24}H_{26}O_6$**	**6.28**	**F6**	**58**	**58.3**	**0.88**	**1a**
24	Oxidized derivative of isocowanol	509.0914	$C_{29}H_{34}O_8$	6.28	F6	58	/	/	3
25	NI	475.3067	$C_{27}H_{42}O_4$	6.43	F4	615	/	/	3
26	NI	475.3065	$C_{27}H_{42}O_4$	6.57	F4	615	/	/	3
27	α-mangostin	409.1641	$C_{24}H_{26}O_6$	6.98	F4	145	30.0	0.75	1a
28	NI	475.3066	$C_{27}H_{42}O_4$	7.17	F4	615	/	/	3
29	**Isocowanin**	**477.2281**	**$C_{29}H_{34}O_6$**	**7.78**	**F3**	**264**	**54.1**	**0.72**	**1a**
30	**9-hydroxycalabaxanthone**	**407.1509**	**$C_{24}H_{24}O_6$**	**8.33**	**F2**	**340**	**24.8**	**0.88**	**1a**
-	Tetraprenyltoluquinone	-	-	-	F3	264	/	0.89	2

NI: Non identified. *CAS number (no name). **2D NMR in acetone-*d*6. ***Fragmentation score. **** Structure error in SciFinder, where the name 9-hydroxycalabaxanthone is associated to the structure of mangostanin [two CAS numbers: 35349-68-9 et 463342-39-4 (m/z 426)].

CONCLUSION

The presented work merges MS/MS based molecular networking, chemometrics and ^{13}C NMR-based dereplication, in order to facilitate the efficient and rapid identification of active NPs from a complex mixture. Compared to a traditional BGF approach, the rapid selection annotation of potentially active NPs of interest is conducted by combining MS/MS and activity data in a PLS-type multivariate analyses. Results were compiled in the bioactive molecular network shown in figure 2.[7] The originality of this work is not based on the isolation and structure elucidation of NPs but in combining different annotation tools based on MS/MS and ^{13}C NMR. By this strategy, confidence level could be notably increased. Level 1 is tentatively proposed, when at least two orthogonal techniques are used. However, we suggest two sub-levels 1a or 1b, depending on the consistency of previously described experimental data. In this work, the subsequent isolation of products demonstrates the effectiveness of the workflow in reliably identifying xanthone stereoisomers. This type of annotation could fill the current gaps in spectral data currently observed on collaborative platforms. It additionally permit the use of customized DBs of NPs based on criteria such as chemotaxonomic, structural, or molecular formula, along with their predicted spectral data.

EXPERIMENTAL SECTION

Plant material. *Garcinia parvifolia* bark (batch KL5670) was collected in Malaysia at Hutan Simpanan Meranto Gua Musang, Kelantan (date of collection: 13.05.2009) by Marc Litaudon. Plant material was verified by the botanist Mr. Teo Leong Eng and voucher specimens are kept at the herbarium of the Department of Chemistry, University of Malaya, Kuala Lumpur, Malaysia.

Extraction procedure. 80.0g of grinded *G. parvifolia* bark were mixed with dichloromethane (DCM) (Carlo Erba reagents, Val-de-Reuil, France) and sonicated at room temperature for 30 min. This step was repeated three times. After filtration and solvent evaporation, 4.48g of dry DCM extract were obtained.

Flash Chromatography. KL5670 (4.20 g) was fractionated on a CombiFlash Rf-200 system (Serlabo Technologies, Entraigues, France) equipped with binary pumps, multiwavelength UV detectors, and fraction collectors using a Puriflash 50SIHC-F0220 (220 g, Interchim, Montluçon, France) silica gel column. The mobile phase consisted of petroleum ether (solvent A), an equivolumetric mixture of ethyl acetate and chloroform (solvent B), and methanol (solvent C). All solvents were purchased from Carlo Erba reagents (Val-de-Reuil, France). The flow rate was set to 50 mL/min and the following gradient was applied: 0 min (A/B/C – 100:0:0); 30 min (70:30:0); 60 min (70:30:0); 65 min (65:35:0); 78 min (65:35:0); 80 min (61:39:0); 110 min (61:39:0); 130 min (40:60:0); 140 min (0:100:0); 170 min (0:100:0); 190 min (0:90:10); 200 min (0:80:20); 230 min (0:80:20). Fractions were collected every 20 mL and

analyzed by thin layer chromatography (data not shown). Overall, 14 recombined fractions were obtained (F1: 127.0 mg; F2: 205.0 mg, F3: 101.0 mg, F4: 366.0 mg, F5: 288.0 mg, F6: 118.0 mg, F7: 42.0 mg, F8: 58.0 mg, F9: 98.8 mg, F10: 504.3 mg, F11: 162.8 mg, F12: 92.3 mg, F13: 185.1 mg and F14: 111.7 mg).

Experimental UPLC-MS Conditions. All experiments were conducted at 25°C on an Acquity UPLC system (Waters, Manchester, UK) using a HSS T3 C-18 (2.1 x 100 mm, 1.8 μm) UPLC column (Waters). LC-MS grade water (solvent A) and acetonitrile (solvent B, LC-MS grade, VWR chemicals), both supplemented with 0.1% of formic acid (LC-MS grade, VWR chemicals) were used as mobile phases.

Fractions of KL5670. Samples were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.02 mg/mL and injected in triplicates (injection volume 1 μL). The mobile phase gradient was as follows: 0 min, 40% B; 5 min, 80% B; 10 min 100% B; 12 min, 100% B. The flow was set to 0.4 mL/min.

Mass Spectrometry Settings for all samples. Experiments were recorded on a WatersXevo®G2-S QTof mass spectrometer (Waters) using negative electrospray ionization. Source parameters were as follows: capillary and cone were set to 0.50 and 40kV respectively. Desolvation gas flow was 1000 L/h at 36 °C, and cone gas was set to 100 L/h. Source temperature was 120 °C and data were collected in continuum mode at 0.5 scan/s. Instrument calibration was conducted using a leucine-enkephalin solution at a flow rate of 5 μL/min (calibration points: 554.2620 for negative mode). All LC-MS/MS data were acquired in Data Dependent Analysis (DDA) mode. The mass analyzer mass range was set to 50-1200 Da for full scan mode

(scan time: 0.5/s). The MS/MS mode was automatically activated once the Total Ion Current (TIC) exceeded 10,000 intensity/s but remained in MS mode when TIC was below this threshold. Collision energy ramps of 20–25 V for low-mass, and 65–85 V for high-mass analytes were employed.

Data pre-processing by ProgenesisQI. Data were then pre-processed by using ProgenesisQI software (Waters). After being aligned, runs were grouped by triplicate and an auto peak-picking was done. Ions with the same retention time, peak shape and isotope distribution were thoroughly checked and grouped. Data were then filtered with "tag" to keep only significative compounds with an intensity over 4000, MS/MS data and an ANOVA p-value ≤ 0.05. This procedure allows targeting fragmentation information from deconvoluted compounds. The later were then annotated using the Progenesis MetaScope tool, a dereplication based strategy using theoretical fragmentations from databases in .SDF format. In this study, the Predicted Carbon-13 NMR Data of Natural Products (PNMRNP) DB[40] [211,280 structures and predicted δ_C of NPs from the Universal Natural Products Database (UNPD)].

UPLC-MS/MS data Processing and Chemometric Analysis. Automated data alignment and peak-picking were conducted by the ProgenesisQI software (Waters, Newcastle, UK). Results were thoroughly inspected and manually corrected if necessary (e.g., regrouping of some ions with identic retention times, peak shapes and/or isotopic distribution). An ANOVA tag (p-value ≤ 0.05) was applied to eliminate non-reproducible data. All ions without MS/MS fragmentation were removed and data were then processed by the Progenesis MetaScope dereplication tool. The latter creates theoretical fragmentation patterns from imported

structures (.SDF files) and compares them with experimental MS/MS data.[52] Putative structural assignments were made using the PNMRNP DB[41]. In order to perform fragmentation matching between experimental MS/MS data and those predicted by ProgenesisQI, both precursor and fragmentation tolerance were set at 5 ppm. Only compounds with structural proposals were studied, *i.e.*, 154 ions. Based on these parameters ProgenesisQI created a list of possible structures with their respective probability scores. Finally, a PLS regression analysis based on spectrometric information (m/z values, retention times and peak areas) as well as activity data (AGE inhibition) was conducted on EZinfo 3.0 for Waters (Umetrics, Umea, Sweden). The Pareto scale setting was applied.

Feature-Based Molecular Network. A feature-based molecular network (FBMN) was created by exporting the "fragmentation" data and the "compound measurement" files as from ProgenesisQI .msp and .csv files respectively to the GNPS. The parameters used for FBMN were set as follow: a precursor ion mass and fragment ion mass tolerance set at 0.02 Da, a cosine score equal to or greater than 0.7, topK set at 5, minimum matched fragment ions set to 4 and maximum connected component size equal to 0. The library search options were set with a minimum matched peak at 4, a score threshold set at 0.6. The job record is available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2ae86a52caf04ebcbf397df1bd6946cb. Visualization and treatment of molecular network was performed using Cytoscape 3.9.0 software.[53]

Experimental NMR conditions. Experiments (^{1}H NMR, ^{13}C NMR, DEPT-135, DEPT-90 and 2D NMR) were performed on a JEOL 400 MHz NMR spectrometer (JEOL, Tokyo, Japan)

equipped with an inverse 5 mm probe (ROYAL RO5). Chemical shifts (δ_H and δ_C) were expressed in ppm and ^{1}H coupling constants (J) in Hz. Depending on their solubility samples were dissolved in deuterated DMSO-*d*6 or acetone-*d*6 both purchased from Eurisotop, (St-Aubin, France). For dereplication, F2 (48.3 mg), F3 (40.4 mg), F6 (50.0 mg), F7 (42.0 mg), F9 (40.0 mg), F10 (40.0 mg) and F12 (39.0 mg) were dissolved in 500 µL acetone-*d*6 and/or $CDCl_3$.

For ^{13}C NMR (100 MHz) spectra, a WALTZ-16 decoupling sequence was used with an acquisition time of 1.04 s (32768 complex data points) and a relaxation delay of 2 s. The number of acquisitions (512 – 20,000 scans) was individually adapted to sample requirements. A 1 Hz exponential line-broadening filter was applied to each FID prior Fourier transformation. Raw data were processed by the MestReNova 12.0.2 (Mestrelab Research, Santiago de Compostela, Spain) and calibrated to solvent peaks at 29.84 ppm (acetone-*d*6) and δ_C 77.16 ppm ($CDCl_3$) and Manual phasing and baseline correction were applied. Alignments of DEPT experiments were done with ^{13}C spectra using a given δ_C. Positive ^{13}C NMR and DEPT-90 as well as positive and negative DEPT-135 signals were collected.

^{13}C-NMR Dereplication. ^{13}C NMR and DEPT data were processed by MixONat[37,46], an open-source ^{13}C NMR- dereplication tool. In preparation of this analysis reported structures of the *Garcinia* genus were first downloaded from the Dictionary of Natural Products[37,54] and compiled in a *Garcinia* database (*Garcinia*DB). The latter comprised 718 entries, which were saved in the SDF format. Their ^{13}C NMR shifts ($\delta_{C\text{-}SDF}$) were then simulated by the ACD/NMR Predictors (C and H) and DB as well as the CTypeGen routine from MixONat for differentiating methyl, methylene, methine or quaternary $\delta_{C\text{-}SDF}$. Then MixONat compared

simulated with experimental data (^{13}C NMR, DEPT-135 and 90) and ranked associated structures with a score from 0 to 1 (1 indicating a full match and zero the absence of any similarity between simulated and experimental spectra). Finally, experimental data of highest ranked candidates were compared with the literature in order to confirm proposed structural assignments.

Experimental preparative HPLC conditions for isolation of selected NPs Fraction 3 (40 mg), Fraction 4 (150 mg), Fraction 6 (135 mg) Fraction 9 (50 mg), Fraction 10 (270 mg) and Fraction 10E (8 mg) were dissolved in MeOH (HPLC grade, VWR chemicals, Rosny-sous-Bois, France) before being processed on a Nexera Prep (Shimadzu, Kyoto, Japan) preparative HPLC system using a Pursuit XRs 5 C-18 column (250 x 21.2 mm, 5 µm, Agilent, Santa Clara, California, USA) except for Fraction 4 which was processed using a Hypersil Gold PFP column (150x20 mm, 5 µm, Thermo Scientific, San Jose, USA). Mobile phase consisted of HPLC grade water (Solvent A) and acetonitrile (Solvent B, HPLC grade, VWR chemicals, Rosny-sous-Bois, France) and the flow rate was set to 21.2 mL/min. The injection volume was 2 mL. Solvent gradient for **F3** was as follows: 0 min, 80% B; 20min, 100% B; 25 min, 100% B. Solvent gradient for **F4** was set as follows: 0 min, 45% B; 18min, 48% B; 20 min, 100% B; 25 min, 100% B. Solvent gradient for **F6** was set as follows: 0 min, 45% B; 25min, 50% B; 30 min, 70% B; 35 min, 100% B; 40 min, 100% B. Solvent gradient for **F9** was set as follows: 0 min, 50% B; 20min, 70% B; 30 min, 100% B; 40 min, 100% B. Solvent gradient for **F10** was set as follows: 0 min, 60% B; 15min, 65% B; 20 min, 70% B; 30 min, 100% B; 35 min, 100% B; 37 min, 60% B to give 11 subfractions (F10A to F10K). Solvent gradient for **F10E** was set as follows: 0 min, 60% B;

15min, 70% B; 16 min, 100% B; 20 min, 100% B. Chromatograms were recorded at 210 and 280 nm for F3, 254 and 280 nm for F4 and F10E and 310 and 280 nm for F9. For F6 and F10, chromatograms were recorded at 254 and 280 nm. Overall, 3 repetitive injections (RI) of F3 (50.4 mg, 94.3 mg, and 120.0 mg), 10 RI of F4 (15.0 mg each), 2 RI (56.4 and 78.6 mg) of F6, 3 RI (50.4, 94.3 and 120.0 mg) of F10, 2 RI (23.0 mg and 15.0 mg) of F9 and 1 injection of F10E were conducted.

Fraction **10B** (9.6 mg) was dissolved in MeOH (HPLC grade, VWR chemicals, Rosny-sous-Bois, France) before being processed on an Agilent 1200 semi prep-HPLC (Agilent technologies, Les Ulis, France) using a Luna C-18 column (250x4.60 mm, 5 µm, Phenomenex, Torrance, USA). Mobile phase consisted of HPLC grade water (Solvent A) and MeOH (Solvent C, HPLC grade, VWR chemicals, Rosny-sous-Bois, France). Solvent gradient was as follows: 0 min, 75% C; 0-40min, 80% C. The flow rate was 3.70 mL/min and chromatograms were recorded at 210, 240, 254, 280 and 310 nm.

Isocowanin (node **29**, 11.5 mg)[55] was isolated from F3 (supporting information, Figures S98), α-mangostin (27, 20.0 mg)[56] from F4 (supporting information, Figures S99) and rubraxanthone (node **23**, 57.0 mg)[48] from F6 (supporting information, Figures S100) (rt, 21.0 min). Parvixanthone H (node **8**, 4.2 mg)[51] was isolated from F9. Parvixanthone G (node **10**, F10D, 15.43 mg)[51] (supporting information, Figures S101 and S102) and isocowanol (node **20**, F10H, 46.5 mg) (supporting information, Figures S103 and S104) were isolated from F10 (12.5 min and rt, 26.7 min respectively) as well as cratoxylone (node **12**, 4.2 mg, F10F)[57] (supporting information, Figures S105 and S106). Garcinone D (node **11**, 1.2 mg)[57] isolated from F10E

(supporting information, Figures S107 and S108) and butyraxanthone D (node **5**, 51.3 mg) and parvixanthone C (node **7**, 9.6 mg) from F10B. 9-hydroxycalabaxanthone (node **30**) was isolated from previous work by Bruguière *et al.*[58].

Advanced glycation end-products (AGEs) inhibition assay. Experiments were conducted according to a previously described protocol.[39,59] In short: Stock solutions (SS) of fractions or purified compound from *G. parvifolia* bark extracts (KL5670) were prepared in DMSO at a concentration of 10 mg/mL, while single compounds were dissolved at a concentration of 30 mM. These SS were then diluted with 50 mM phosphate buffer (pH: 7.4) yielding working solutions (WS) at a concentration range of 10 to 10^{-2} mg/mL for standard mixtures and bark extracts. The WS for single compounds were prepared at a concentration range of 30-3.10^{-2} mM. Next 10 µL of each WS were deposited in 96 black well bottom plates (Fisher Scientific, Illkirch, France) and mixed with 90 µL of a solution containing bovine serum albumine (11 mg/L, Sigma Aldrich, St Quentin Fallavier, France), D-ribose (0.25 M, Acros Organics, Geel, Belgium), and phosphate buffer (50 mM, NaN_3 0.02%, pH 7.4). Plates were then incubated for 24h at 37°C, before being fluorometrically analyzed (λ_{exc}: 335 nm, λ_{em}: 385 nm) on an Infinite M200 plate reader (Tecan, Lyon, France). The formation of AGEs was calculated according to the following formula (FI = fluorescence intensity):

$$\frac{[FI\ (BSA + ribose + sample) - FI\ (BSA + sample)]\ \text{x}\ 100}{FI(BSA + ribose) - FI(BSA)}$$

Aminoguanidine (Sigma Aldrich, St Quentin Fallavier, France) was used as positive, a mixture of BSA + ribose as negative control. A BSA solution served as blank. Results were

expressed as IC_{50} values in µg/mL (mixtures and extracts) or µM (single compounds). All experiments were conducted in triplicates for fractions but only once for purified compounds. For IC_{50} calculations dose-effect curves were best fitted with a sigmoidal dose-response equation using Sigma Plot 12.0 software.

ASSOCIATED CONTENT

The Supporting Information is available free of charge at:

PLS of most active fractions, VIP score and RC from PLS model of features detected in KL5670 fractions, additional structures, part of structures proposition of selected ions by ProgenesisQI dereplication, ^{13}C NMR, DEPT-135 and -90 of selected fractions (F2, F3, F6, F7, F9, F10, F12), part of displayed results for the 13C NMR dereplication, tables of comparison of experimental δ_C with those of previously described NPs in *Garcinia* species IC_{50} (inhibition of AGE formation) of fractions

AUTHOR INFORMATION

Corresponding Author

* AS: andreas.schinkovitz@univ-angers.fr; SD: severine.derbre@univ-angers.fr

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡These authors contributed equally. (Match statement to author names with a symbol)

Funding Sources

This work was carried out in the framework of the International French Malaysian Natural Product Laboratory (IFM-NatPro-Lab) established between CNRS-ICSN and the University Malaya. The authors also thank Le ministère de l'Europe et des Affaires Etrangères (MEAE) and Campus France for their financial support [PANASIA project, BioAsie, 2016-2017 and DEMAiN project, PHC-Hibiscus, 2020-2022].

REFERENCES

(1) Cordell, G. A. Biodiversity and Drug Discovery. A Symbiotic Relationship. *Phytochemistry* **2000**, *55* (6), 463–480. https://doi.org/10.1016/S0031-9422(00)00230-2.

(2) Gwinn, K. D. Chapter 7 - Bioactive Natural Products in Plant Disease Control. In *Studies in Natural Products Chemistry*; Atta-ur-Rahman, Ed.; Elsevier: 2018, 2018; Vol. 56, pp 229–246. https://doi.org/10.1016/B978-0-444-64058-1.00007-8.

(3) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83* (3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.

(4) Malviya, N.; Malviya, S. Bioassay Guided Fractionation-an Emerging Technique Influence the Isolation, Identification and Characterization of Lead Phytomolecules. *Int. J. Clin. Pharmacol.* **2017**, No. 2:5, 1–6. https://doi.org/10.28933/ijhp-2017-07-0901.

(5) Inui, T.; Wang, Y.; Pro, S. M.; Franzblau, S. G.; Pauli, G. F. Unbiased Evaluation of Bioactive Secondary Metabolites in Complex Matrices. *Fitoterapia* **2012**, *83* (7), 1218–1225. https://doi.org/10.1016/j.fitote.2012.06.012.

(6) Kellogg, J. J.; Todd, D. A.; Egan, J. M.; Raja, H. A.; Oberlies, N. H.; Kvalheim, O. M.; Cech, N. B. Biochemometrics for Natural Products Research: Comparison of Data Analysis Approaches and Application to Identification of Bioactive Compounds. *J. Nat. Prod.* **2016**, *79* (2), 376–386. https://doi.org/10.1021/acs.jnatprod.5b01014.

(7) Nothias, L.-F.; Nothias-Esposito, M.; da Silva, R.; Wang, M.; Protsyuk, I.; Zhang, Z.; Sarvepalli, A.; Leyssen, P.; Touboul, D.; Costa, J.; Paolini, J.; Alexandrov, T.; Litaudon, M.; Dorrestein, P. C. Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. *J. Nat. Prod.* **2018**, *81* (4), 758–767. https://doi.org/10.1021/acs.jnatprod.7b00737.

(8) Langeder, J.; Döring, K.; Schmietendorf, H.; Grienke, U.; Schmidtke, M.; Rollinger, J. M. 1H NMR-Based Biochemometric Analysis of Morus Alba Extracts toward a Multipotent Herbal Anti-Infective. *J. Nat. Prod.* **2023**, *86* (1), 8–17. https://doi.org/10.1021/acs.jnatprod.2c00481.

(9) Blaženović, I.; Kind, T.; Ji, J.; Fiehn, O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. *Metabolites* **2018**, *8* (2), 31. https://doi.org/10.3390/metabo8020031.

(10) Wishart, D. S.; Sayeeda, Z.; Budinski, Z.; Guo, A.; Lee, B. L.; Berjanskii, M.; Rout, M.; Peters, H.; Dizon, R.; Mah, R.; Torres-Calzada, C.; Hiebert-Giesbrecht, M.; Varshavi, D.; Varshavi, D.; Oler, E.; Allen, D.; Cao, X.; Gautam, V.; Maras, A.; Poynton, E. F.; Tavangar, P.; Yang, V.; van Santen, J. A.; Ghosh, R.; Sarma, S.; Knutson, E.; Sullivan, V.; Jystad, A. M.; Renslow, R.; Sumner, L. W.; Linington, R. G.; Cort, J. R. NP-MRD: The Natural Products Magnetic Resonance Database. *Nucleic Acids Research* **2022**, *50* (D1), D665–D677. https://doi.org/10.1093/nar/gkab1052.

(11) Appendino, G.; Fontana, G.; Pollastro, F. 3.08 - Natural Products Drug Discovery. In *Comprehensive natural products II: chemistry and biology*; Mander, L. N., Liu, H., Eds.; Elsevier Science: Oxford, 2010.

(12) 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.* **2017**, *16* (1), 55–95. https://doi.org/10.1007/s11101-015-9448-7.

(13) Van Middlesworth, F.; Cannell, R. J. P. Dereplication and Partial Identification of Natural Products. In *Natural Products Isolation*; Cannell, R. J. P., Ed.; Walker, J. M., Series Ed.; Methods in Biotechnology; Humana Press: Totowa, NJ, 1998; Vol. 4, pp 279–327. https://doi.org/10.1007/978-1-59259-256-2_10.

(14) Ntie-Kang, F.; Svozil, D. An Enumeration of Natural Products from Microbial, Marine and Terrestrial Sources. *Phys. Sci. Rev.* **2020**, *5* (8). https://doi.org/10.1515/psr-2018-0121.

(15) Kuhn, S.; Wieske, L. H. E.; Trevorrow, P.; Schober, D.; Schlörer, N. E.; Nuzillard, J.-M.; Kessler, P.; Junker, J.; Herráez, A.; Farès, C.; Erdélyi, M.; Jeannerat, D. NMReDATA: Tools and Applications. *Magn. Reson. Chem.* **2021**, *59* (8), 792–803. https://doi.org/10.1002/mrc.5146.

(16) Fox Ramos, A. E.; Le Pogam, P.; Fox Alcover, C.; Otogo N'Nang, E.; Cauchie, G.; Hazni, H.; Awang, K.; Bréard, D.; Echavarren, A. M.; Frédérich, M.; Gaslonde, T.; Girardot, M.; Grougnet, R.; Kirillova, M. S.; Kritsanida, M.; Lémus, C.; Le Ray, A.-M.; Lewin, G.; Litaudon, M.; Mambu, L.; Michel, S.; Miloserdov, F. M.; Muratore, M. E.; Richomme-Peniguel, P.; Roussi, F.; Evanno, L.; Poupon, E.; Champy, P.; Beniddir, M. A. Collected Mass Spectrometry Data on Monoterpene Indole Alkaloids from Natural Product Chemistry Research. *Sci Data* **2019**, *6* (1), 15. https://doi.org/10.1038/s41597-019-0028-3.

(17) Olivier-Jimenez, D.; Chollet-Krugler, M.; Rondeau, D.; Beniddir, M. A.; Ferron, S.; Delhaye, T.; Allard, P.-M.; Wolfender, J.-L.; Sipman, H. J. M.; Lücking, R.; Boustie, J.; Le Pogam, P. A Database of High-Resolution MS/MS Spectra for Lichen Metabolites. *Sci Data* **2019**, *6* (1), 294. https://doi.org/10.1038/s41597-019-0305-1.

(18) Agnès, S. A.; Okpekon, T.; Kouadio, Y. A.; Jagora, A.; Bréard, D.; Costa, E. V.; da Silva, F. M. A.; Koolen, H. H. F.; Le Ray-Richomme, A.-M.; Richomme, P.; Champy, P.; Beniddir, M. A.; Le Pogam, P. Implementation of a MS/MS Database for Isoquinoline Alkaloids and Other Annonaceous Metabolites. *Sci Data* **2022**, *9* (1), 270. https://doi.org/10.1038/s41597-022-01345-y.

(19) Vargas, F.; Weldon, K. C.; Sikora, N.; Wang, M.; Zhang, Z.; Gentry, E. C.; Panitchpakdi, M. W.; Caraballo-Rodríguez, A. M.; Dorrestein, P. C.; Jarmusch, A. K. Protocol for Community-Created Public MS/MS Reference Spectra within the Global Natural Products Social Molecular Networking Infrastructure. *Rapid Commun. Mass Spectrom.* **2020**, *34* (10), e8725. https://doi.org/10.1002/rcm.8725.

(20) Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S. CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **2021**, *93* (34), 11692–11700. https://doi.org/10.1021/acs.analchem.1c01465.

(21) Wang, F.; Allen, D.; Tian, S.; Oler, E.; Gautam, V.; Greiner, R.; Metz, T. O.; Wishart, D. S. CFM-ID 4.0-a Web Server for Accurate MS-Based Metabolite Identification. *Nucleic Acids Res.* **2022**, *50* (W1), W165–W174. https://doi.org/10.1093/nar/gkac383.

(22) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation. *J. Cheminformatics* **2016**, *8* (1), 3. https://doi.org/10.1186/s13321-016-0115-9.

(23) ACD Labs. *NMR Spectroscopy Software*. https://www.acdlabs.com/products/adh/nmr/ (accessed 2023-01-15).

(24) Kuhn, S. *NMRshiftDB2*. https://nmrshiftdb.nmr.uni-koeln.de/ (accessed 2023-01-15).

(25) Kuhn, S.; Schlörer, N. E. Facilitating Quality Control for Spectra Assignments of Small Organic Molecules: Nmrshiftdb2 – a Free in-House NMR Database with Integrated LIMS for

Academic Service Laboratories. *Magn. Reson. Chem.* **2015**, *53* (8), 582–589. https://doi.org/10.1002/mrc.4263.

(26) Allard, P.-M.; Péresse, T.; Bisson, J.; Gindro, K.; Marcourt, L.; Pham, V. C.; Roussi, F.; Litaudon, M.; Wolfender, J.-L. Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. *Anal. Chem.* **2016**, *88* (6), 3317–3323. https://doi.org/10.1021/acs.analchem.5b04804.

(27) Bertrand, S.; Guitton, Y.; Roullier, C. Successes and Pitfalls in Automated Dereplication Strategy Using Liquid Chromatography Coupled to Mass Spectrometry Data: A CASMI 2016 Experience. *Phytochem. Lett.* **2017**, *21*, 297–305. https://doi.org/10.1016/j.phytol.2016.12.025.

(28) Kind, T.; Fiehn, O. Strategies for Dereplication of Natural Compounds Using High-Resolution Tandem Mass Spectrometry. *Phytochem. Lett.* **2017**, *21*, 313–319. https://doi.org/10.1016/j.phytol.2016.11.006.

(29) Abedini, A.; Colin, M.; Hubert, J.; Charpentier, E.; Angelis, A.; Bounasri, H.; Bertaux, B.; Kotland, A.; Reffuveille, F.; Nuzillard, J.-M.; Renault, J.-H.; Gangloff, S. C. Abundant Extractable Metabolites from Temperate Tree Barks: The Specific Antimicrobial Activity of Prunus Avium Extracts. *Antibiotics* **2020**, *9* (3), 111. https://doi.org/10.3390/antibiotics9030111.

(30) Bruguière, A.; Derbré, S.; Bréard, D.; Tomi, F.; Nuzillard, J.-M.; Richomme, P. ^{13}C NMR Dereplication Using MixONat Software: A Practical Guide to Decipher Natural Products Mixtures. *Planta Med.* **2021**, *87* (12/13), 1061–1068. https://doi.org/10.1055/a-1470-0446.

(31) Silva-Castro, L. F.; Derbré, S.; Le Ray, A. M.; Richomme, P.; García-Sosa, K.; Peña-Rodriguez, L. M. Using 13C-NMR Dereplication to Aid in the Identification of Xanthones Present in the Stem Bark Extract of Calophyllum Brasiliense. *Phytochem. Anal.* **2021**, *32* (6), 1102–1109. https://doi.org/10.1002/pca.3051.

(32) Azonwade, F.; Mabanza-Banza, B. B.; Le Ray, A.-M.; Bréard, D.; Blanchard, P.; Goubalan, E.; Lamine, B.-M.; Banga-Mboko, H.; Richomme, P.; Derbré, S.; Boisard, S. Chemodiversity of Propolis Samples Collected in Various Areas of Benin and Congo: Chromatographic Profiling and Chemical Characterization Guided by 13C NMR Dereplication. *Phytochem. Anal.* **2023**, Accepted manuscript. https://doi.org/10.1002/pca.3227.

(33) Sumner, L. W.; Amberg, A.; Barrett, D.; Beale, M. H.; Beger, R.; Daykin, C. A.; Fan, T. W.-M.; Fiehn, O.; Goodacre, R.; Griffin, J. L.; Hankemeier, T.; Hardy, N.; Harnly, J.; Higashi, R.; Kopka, J.; Lane, A. N.; Lindon, J. C.; Marriott, P.; Nicholls, A. W.; Reily, M. D.; Thaden, J. J.; Viant, M. R. Proposed Minimum Reporting Standards for Chemical Analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **2007**, *3* (3), 211–221. https://doi.org/10.1007/s11306-007-0082-2.

(34) Viant, M. R.; Kurland, I. J.; Jones, M. R.; Dunn, W. B. How Close Are We to Complete Annotation of Metabolomes? *Curr. Opin. Chem. Biol.* **2017**, *36*, 64–69. https://doi.org/10.1016/j.cbpa.2017.01.001.

(35) Yang, J. Y.; Sanchez, L. M.; Rath, C. M.; Liu, X.; Boudreau, P. D.; Bruns, N.; Glukhov, E.; Wodtke, A.; De Felicio, R.; Fenner, A.; Wong, W. R.; Linington, R. G.; Zhang, L.; Debonsi, H. M.; Gerwick, W. H.; Dorrestein, P. C. Molecular Networking as a Dereplication Strategy. *J. Nat. Prod.* **2013**, *76* (9), 1686–1699. https://doi.org/10.1021/np400413s.

(36) Hubert, J.; Nuzillard, J.-M.; Purson, S.; Hamzaoui, M.; Borie, N.; Reynaud, R.; Renault, J.-H. Identification of Natural Metabolites in Mixture: A Pattern Recognition Strategy Based on 13C NMR. *Anal. Chem.* **2014**, *86* (6), 2955–2962. https://doi.org/10.1021/ac403223f.

(37) Bruguière, A.; Derbré, S.; Dietsch, J.; Leguy, J.; Rahier, V.; Pottier, Q.; Bréard, D.; Suor-Cherer, S.; Viault, G.; Le Ray, A.-M.; Saubion, F.; Richomme, P. MixONat, a Software for the Dereplication of Mixtures Based on 13C NMR Spectroscopy. *Anal. Chem.* **2020**, *92* (13), 8793–8801. https://doi.org/10.1021/acs.analchem.0c00193.

(38) Meunier, M.; Bréard, D.; Awang, K.; Boisard, S.; Guilet, D.; Richomme, P.; Derbré, S.; Schinkovitz, A. Matrix Free Laser Desorption Ionization Assisted by 13C NMR Dereplication: A Complementary Approach to LC-MS2 Based Chemometrics. *Talanta* **2023**, *253*, 123998. https://doi.org/10.1016/j.talanta.2022.123998.

(39) Derbré, S.; Gatto, J.; Pelleray, A.; Coulon, L.; Séraphin, D.; Richomme, P. Automating a 96-Well Microtiter Plate Assay for Identification of AGEs Inhibitors or Inducers: Application to the Screening of a Small Natural Compounds Library. *Anal. Bioanal. Chem.* **2010**, *398* (4), 1747–1758. https://doi.org/10.1007/s00216-010-4065-1.

(40) Nuzillard, J.-M.; Leroy, R.; Kuhn, S. Predicted Carbon-13 NMR Data of Natural Products (PNMRNP), 2021. https://doi.org/10.5281/zenodo.4420849.

(41) Lianza, M.; Leroy, R.; Machado Rodrigues, C.; Borie, N.; Sayagh, C.; Remy, S.; Kuhn, S.; Renault, J.-H.; Nuzillard, J.-M. The Three Pillars of Natural Product Dereplication. Alkaloids from the Bulbs of *Urceolina Peruviana* (C. Presl) J.F. Macbr. as a Preliminary Test Case. *Molecules* **2021**, *26* (3), 637. https://doi.org/10.3390/molecules26030637.

(42) Aron, A. T.; Gentry, E. C.; McPhail, K. L.; Nothias, L.-F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J. M.; Sikora, N.; Vargas, F.; van der Hooft, J. J. J.; Ernst, M.; Kang, K. B.; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Koester, I.; Weldon, K. C.; Bertrand, S.; Roullier, C.; Sun, K.; Tehan, R. M.; Boya P., C. A.; Christian, M. H.; Gutiérrez, M.; Ulloa, A. M.; Tejeda Mora, J. A.; Mojica-Flores, R.; Lakey-Beitia, J.; Vásquez-Chaves, V.; Zhang, Y.; Calderón, A. I.; Tayler, N.; Keyzers, R. A.; Tugizimana, F.; Ndlovu, N.; Aksenov, A. A.; Jarmusch, A. K.; Schmid, R.; Truman, A. W.; Bandeira, N.; Wang, M.; Dorrestein, P. C. Reproducible Molecular Networking of Untargeted Mass Spectrometry Data Using GNPS. *Nat Protoc* **2020**, *15* (6), 1954–1991. https://doi.org/10.1038/s41596-020-0317-5.

(43) Morehouse, N. J.; Clark, T. N.; McMann, E. J.; van Santen, J. A.; Haeckl, F. P. J.; Gray, C. A.; Linington, R. G. Annotation of Natural Product Compound Families Using Molecular Networking Topology and Structural Similarity Fingerprinting. *Nat Commun* **2023**, *14* (1), 308. https://doi.org/10.1038/s41467-022-35734-z.

(44) Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Gaudry, A.; Graham, J. G.; Stephan, R.; Page, R.; Vondrášek, J.; Steinbeck, C.; Pauli, G. F.; Wolfender, J.-L.; Bisson, J.; Allard, P.-M. The LOTUS Initiative for Open Knowledge Management in Natural Products Research. *eLife* **2022**, *11*, e70780. https://doi.org/10.7554/eLife.70780.

(45) Wolfender, J.-L.; Nuzillard, J.-M.; Van Der Hooft, J. J. J.; Renault, J.-H.; Bertrand, S. Accelerating Metabolite Identification in Natural Product Research: Toward an Ideal Combination of Liquid Chromatography–High-Resolution Tandem Mass Spectrometry and NMR Profiling, *in Silico* Databases, and Chemometrics. *Anal. Chem.* **2019**, *91* (1), 704–742. https://doi.org/10.1021/acs.analchem.8b05112.

(46) Bruguière, A.; Derbré, S. *MixONat*. SourceForge. https://sourceforge.net/projects/mixonat/ (accessed 2022-01-14).

(47) Nuzillard, J.-M. Taxonomy-Focused Natural Product Databases for Carbon-13 NMR-Based Dereplication. *Analytica* **2021**, *2* (3), 50–56. https://doi.org/10.3390/analytica2030006.

(48) Munekazu, I.; Hideki, T.; Toshiyuki, T.; Soedarsono, R. Three New Xanthones from the Bark of Garcinia Dioica. *Chem. Pharm. Bull.* **1995**, *44* (1), 232–234. https://doi.org/10.1248/cpb.44.232.

(49) Wahyuni, F. S.; Byrne, L. T.; Dachriyanus; Dianita, R.; Jubahar, J.; Lajis, N. H.; Sargent, M. V. A New Ring-Reduced Tetraprenyltoluquinone and a Prenylated Xanthone from Garcinia Cowa. *Aust. J. Chem.* **2004**, *57* (3), 223. https://doi.org/10.1071/CH03175.

(50) Al-Massarani, S.; El Gamal, A.; Al-Musayeib, N.; Mothana, R.; Basudan, O.; Al-Rehaily, A.; Farag, M.; Assaf, M.; El Tahir, K.; Maes, L. Phytochemical, Antimicrobial and Antiprotozoal Evaluation of Garcinia Mangostana Pericarp and α-Mangostin, Its Major Xanthone Derivative. *Molecules* **2013**, *18* (9), 10599–10608. https://doi.org/10.3390/molecules180910599.

(51) Xu, Y.-J.; Lai, Y.-H.; Imiyabir, Z.; Goh, S.-H. Xanthones from Garcinia Parvifolia. *J. Nat. Prod.* **2001**, *64* (9), 1191–1195. https://doi.org/10.1021/np0101393.

(52) Shi, X.; Yang, W.; Qiu, S.; Hou, J.; Wu, W.; Guo, D. Systematic Profiling and Comparison of the Lipidomes from Panax Ginseng , P. Quinquefolius , and P. Notoginseng by Ultrahigh Performance Supercritical Fluid Chromatography/High-Resolution Mass Spectrometry and Ion Mobility-Derived Collision Cross Section Measurement. *J. Chromatogr. A* **2018**, *1548*, 64–75. https://doi.org/10.1016/j.chroma.2018.03.025.

(53) Lopes, C. T.; Franz, M.; Kazi, F.; Donaldson, S. L.; Morris, Q.; Bader, G. D. Cytoscape Web: An Interactive Web-Based Network Browser. *Bioinformatics* **2010**, *26* (18), 2347–2348. https://doi.org/10.1093/bioinformatics/btq430.

(54) *Dictionary of Natural Products 29.2 Chemical Search*. ChemNetBase. https://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml (accessed 2021-06-01).

(55) Ampofo, S. A.; Waterman, P. G. Xanthones from Three Garcinia Species. *Phytochemistry* **1986**, *25* (10), 2351–2355. https://doi.org/10.1016/S0031-9422(00)81694-5.

(56) Ha, L. D.; Hansen, P. E.; Vang, O.; Duus, F.; Pham, H. D.; Nguyen, L.-H. D. Cytotoxic Geranylated Xanthones and O-Alkylated Derivatives of α-Mangostin. *Chem. Pharm. Bull.* **2009**, *57* (8), 5.

(57) Bennett, G. J.; Harrison, L. J.; Sia, G.-L.; Sim, K.-Y. Triterpenoids, Tocotrienols and Xanthones from the Bark of Cratoxylum Cochinchinense. *Phytochemistry* **1993**, *32* (5), 1245–1251. https://doi.org/10.1016/S0031-9422(00)95100-8.

(58) Bruguière, A.; Derbré, S.; Dietsch, J.; Leguy, J.; Rahier, V.; Pottier, Q.; Suor-Cherer, S.; Viault, G. MixONat, a Software for the Dereplication of Mixture Based on ^{13}C NMR Spectroscopy. *Anal. Chem.* **2020**, *92* (13), 8793–8801. https://doi.org/10.1021/acs.analchem.0c00193.

(59) Séro, L.; Sanguinet, L.; Blanchard, P.; Dang, B.; Morel, S.; Richomme, P.; Séraphin, D.; Derbré, S. Tuning a 96-Well Microtiter Plate Fluorescence-Based Assay to Identify AGE

Inhibitors in Crude Plant Extracts. *Molecules* **2013**, *18* (11), 14320–14339. https://doi.org/10.3390/molecules181114320.

Looking for actives in the haystack: Merging HRMS²-based molecular networking, chemometrics and ^{13}C NMR-based dereplication

Manon Meunier,[a] Dimitri Bréard,[a,b] Séverine Boisard,[a,b] Patricia Blanchard,[a] Marc Litaudon,[c,d] Khalijah Awang,[d] Andreas Schinkovitz,[a,*] Séverine Derbré [a,*]

[a] Univ Angers, SONAS, SFR QUASAV, F-49000 Angers, France

[b] PHYTO Platform, SFR QUASAV, Campus du Végétal, 49045, Angers CEDEX 01, France

5. [c] Institut de Chimie des Substances Naturelles, CNRS-ICSN, UPR 2301, Université Paris-Saclay, 91198 Gif-sur-Yvette, France

[d] IFM NatProLab, University of Malaya, Faculty of Sciences, Department of Chemistry, Malaysia

Table of contents

List of Figures

List of Tables

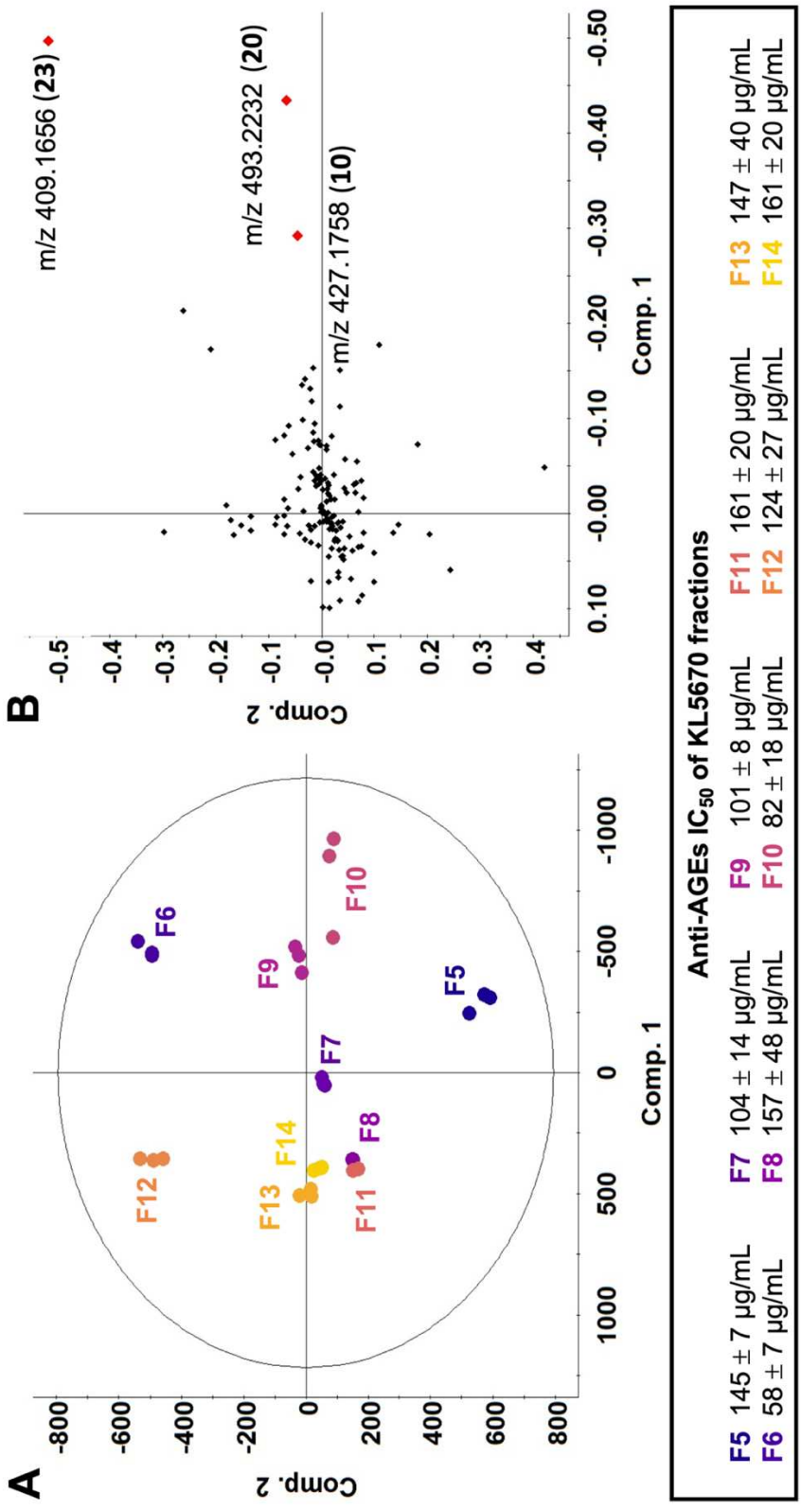


Figure S42. Partial least square analysis of most active fractions (F5-F14) obtained after flash separation of KL5670 using UPLC-MS^2 and bioactivity data. **A**. Score plots. **B**. Loading plots. Overall, three principal components accounted for 99.91% (UPLC-MS) of explained variations [1st component: 94.15% (UPLC), 2nd component: 5.64% and the 3rd component: 0.12%]. All data showed up within the Hoteling's 95% confidence ellipse.

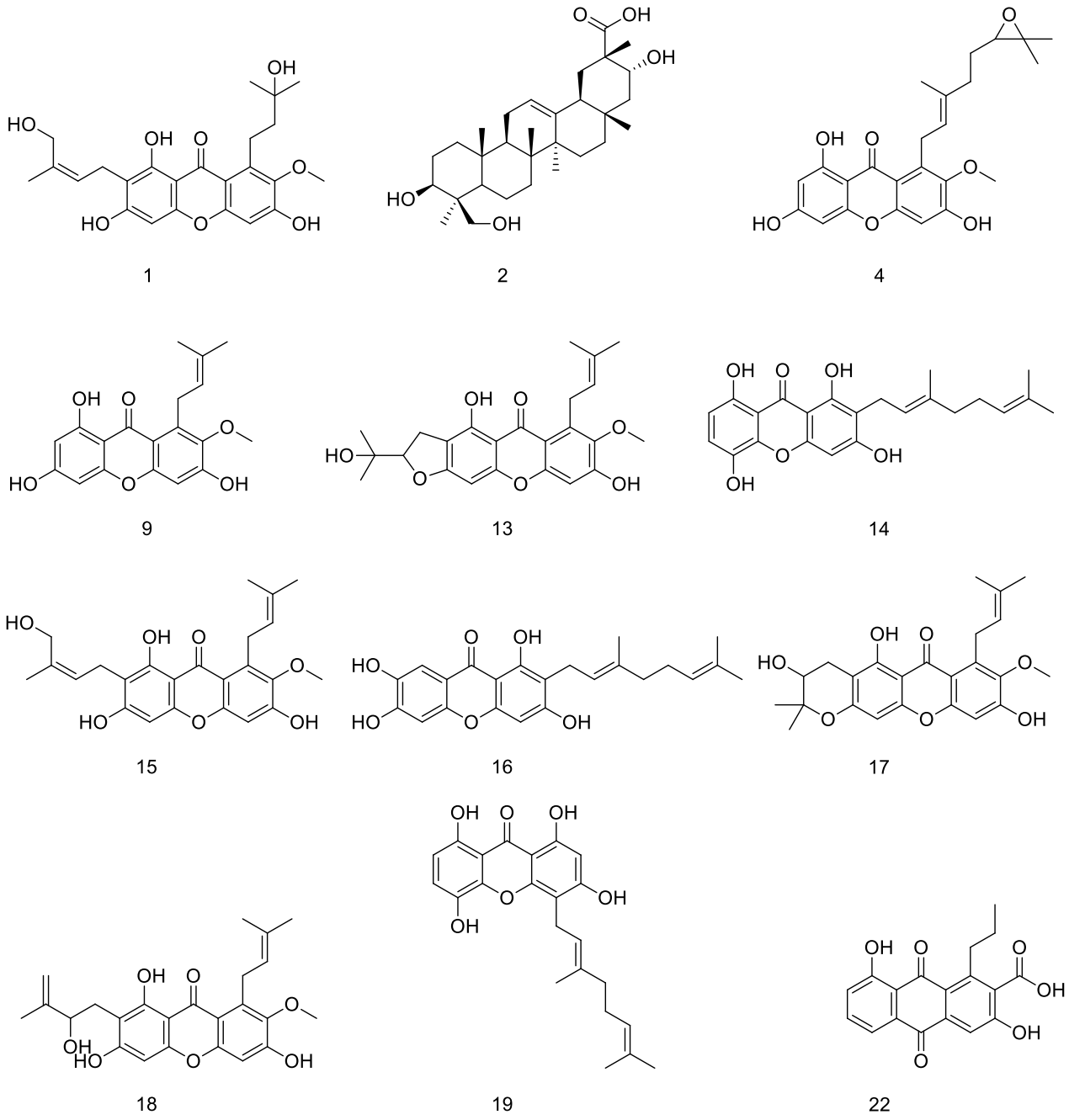


Figure S43. Additional structure complementary to Figure 2

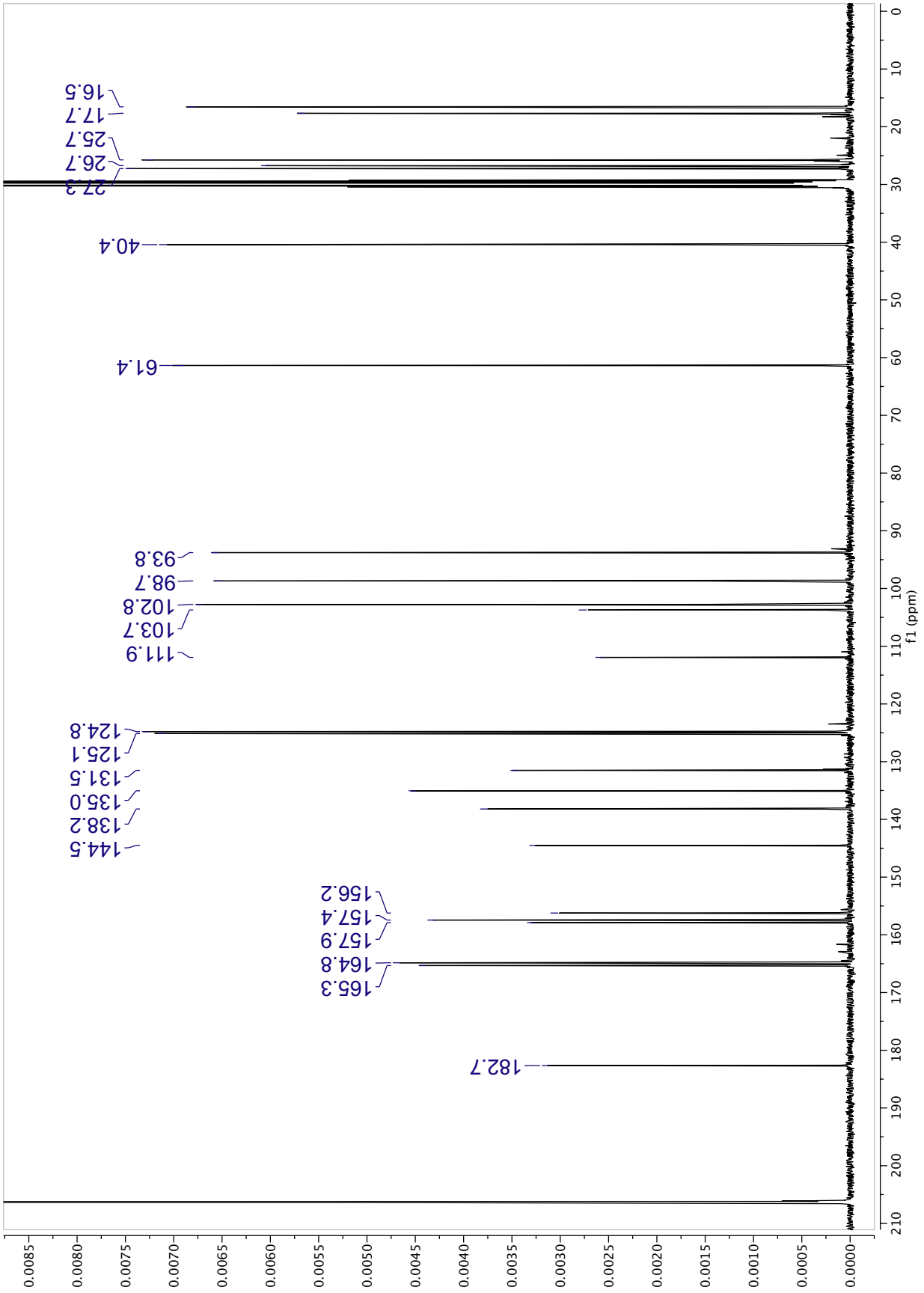


Figure S44. ^{13}C NMR spectrum (5163 scans) of F6 (50.0 mg) recorded in acetone-*d*6.

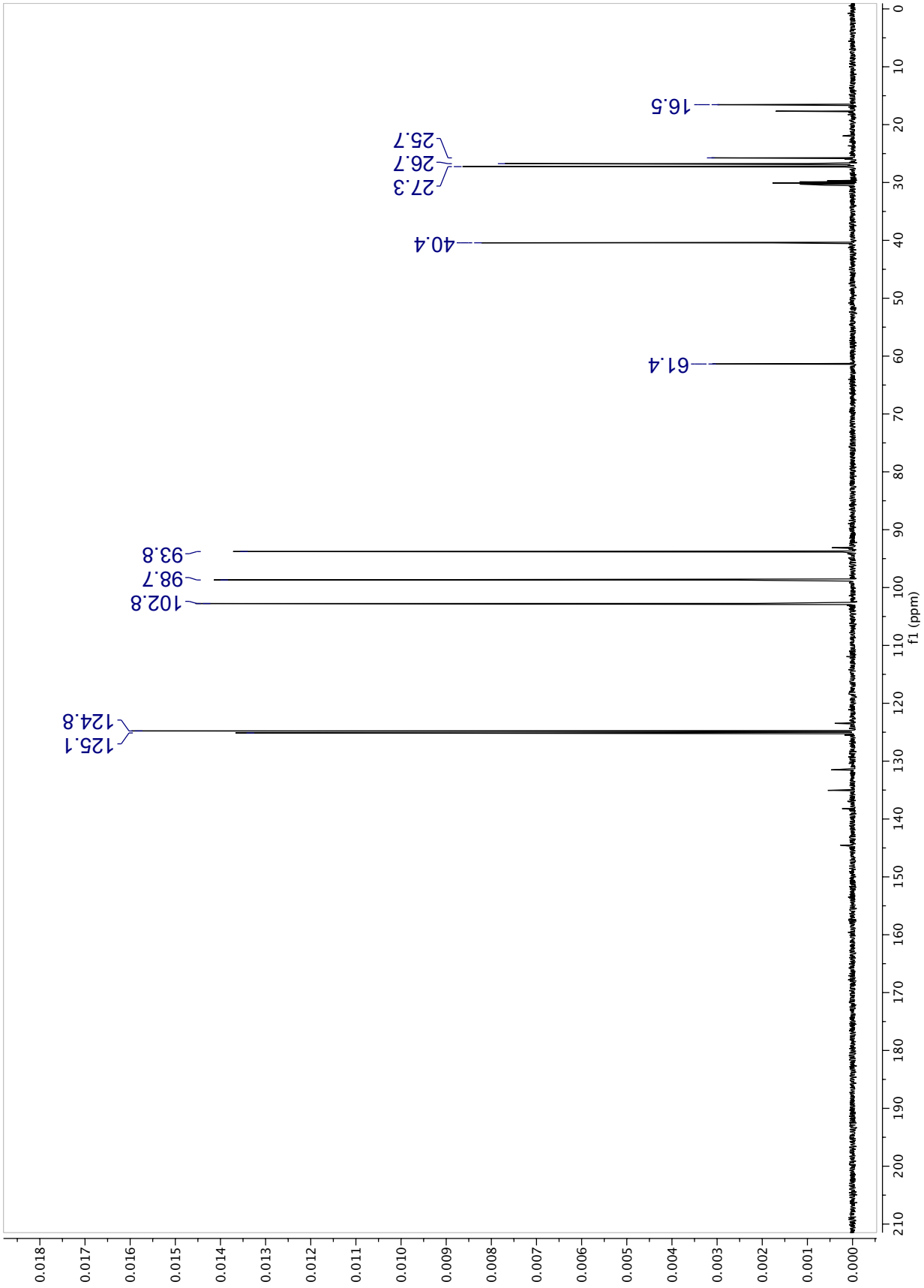


Figure S45. DEPT-135 spectrum (2000 scans) of F6 (50.0 mg) recorded in acetone-*d*6.

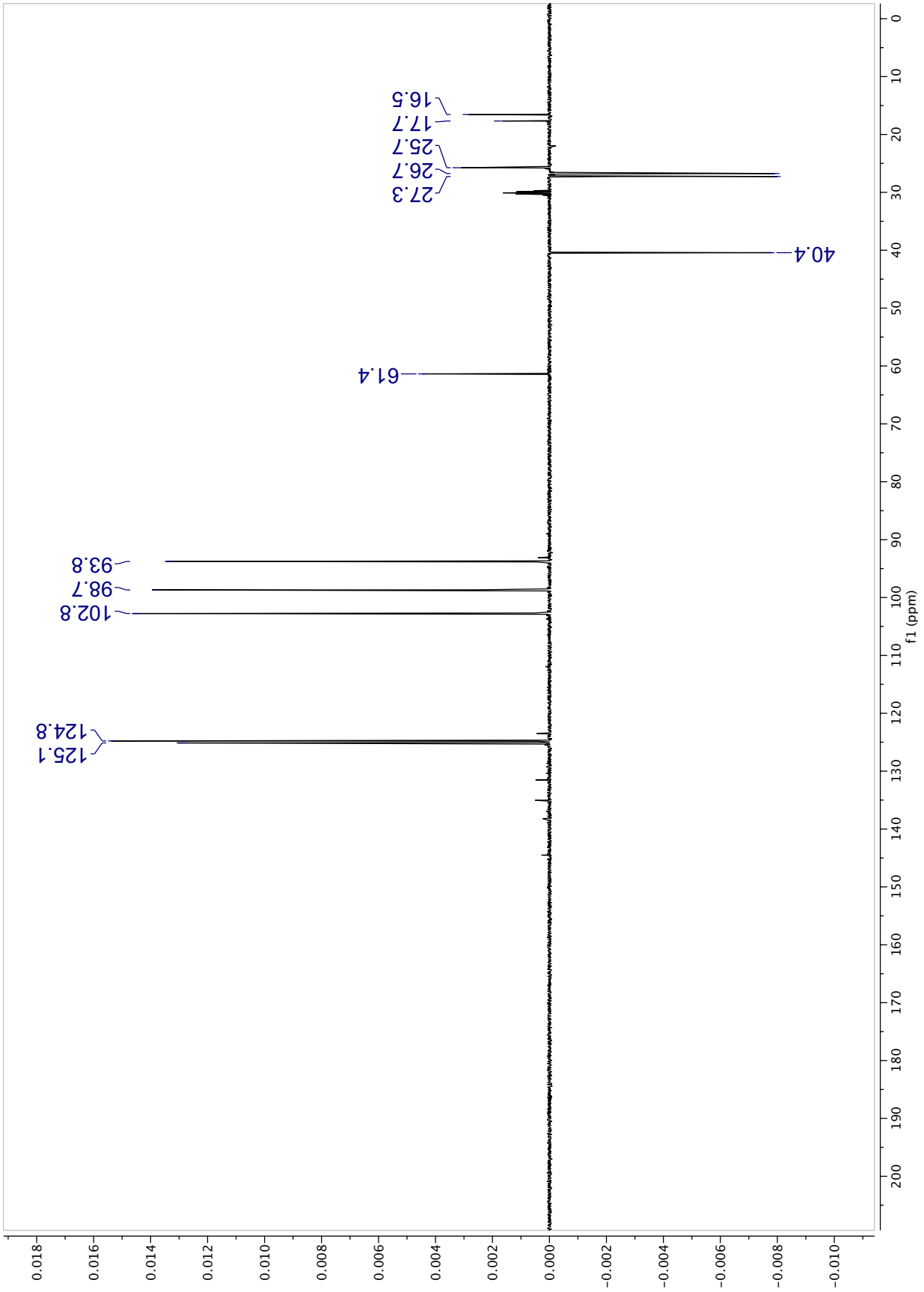


Figure S46. DEPT-90 spectrum (2000 scans) of F6 (50.0 mg) recorded in acetone-*d*6.

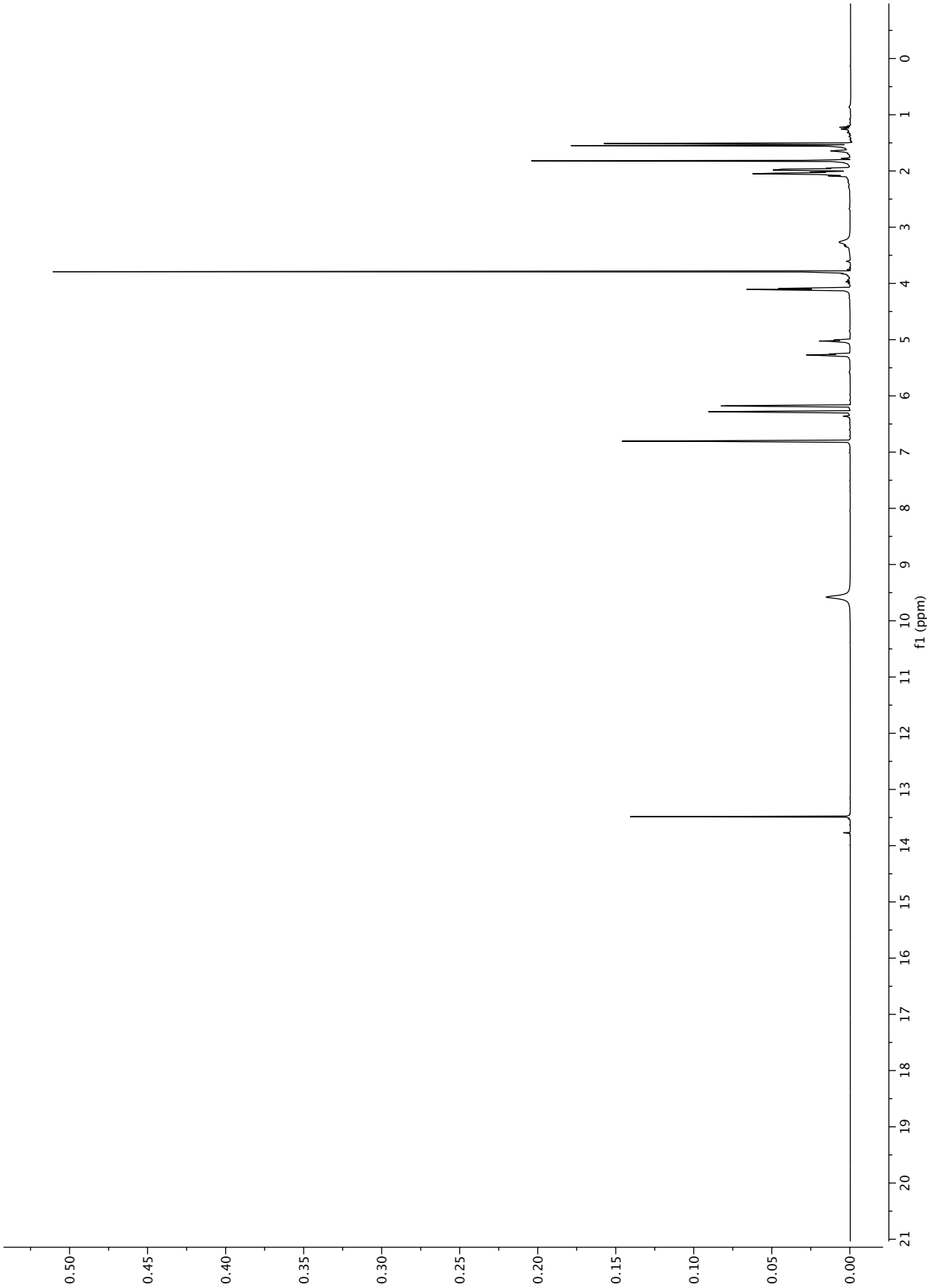


Figure S47. ^{1}H-NMR of F6 (50.0 mg) recorded in acetone-*d*6

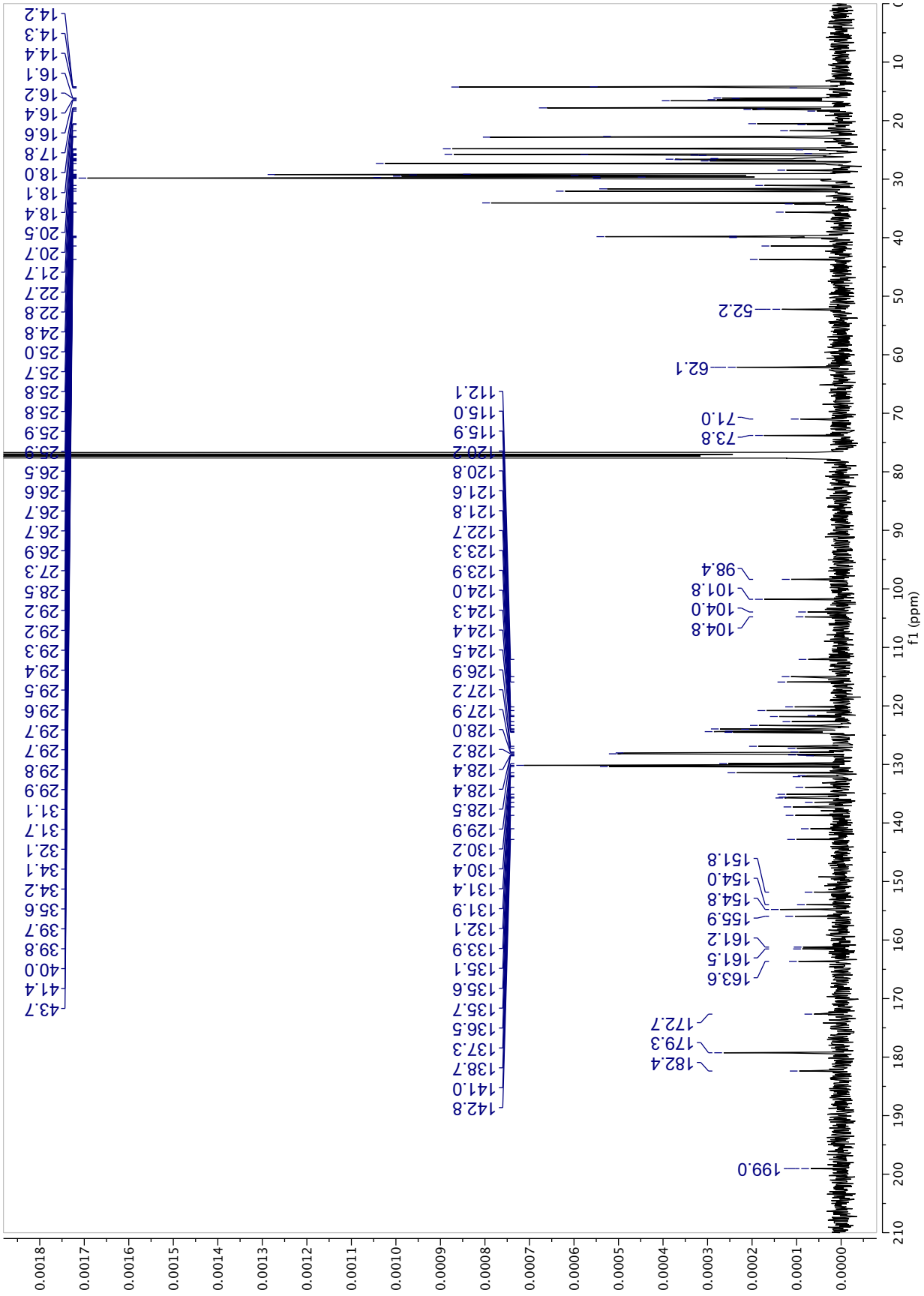


Figure S48. ^{13}C spectrum (5000 scans) of F3 (40.38 mg) recorded in $CDCl_3$.

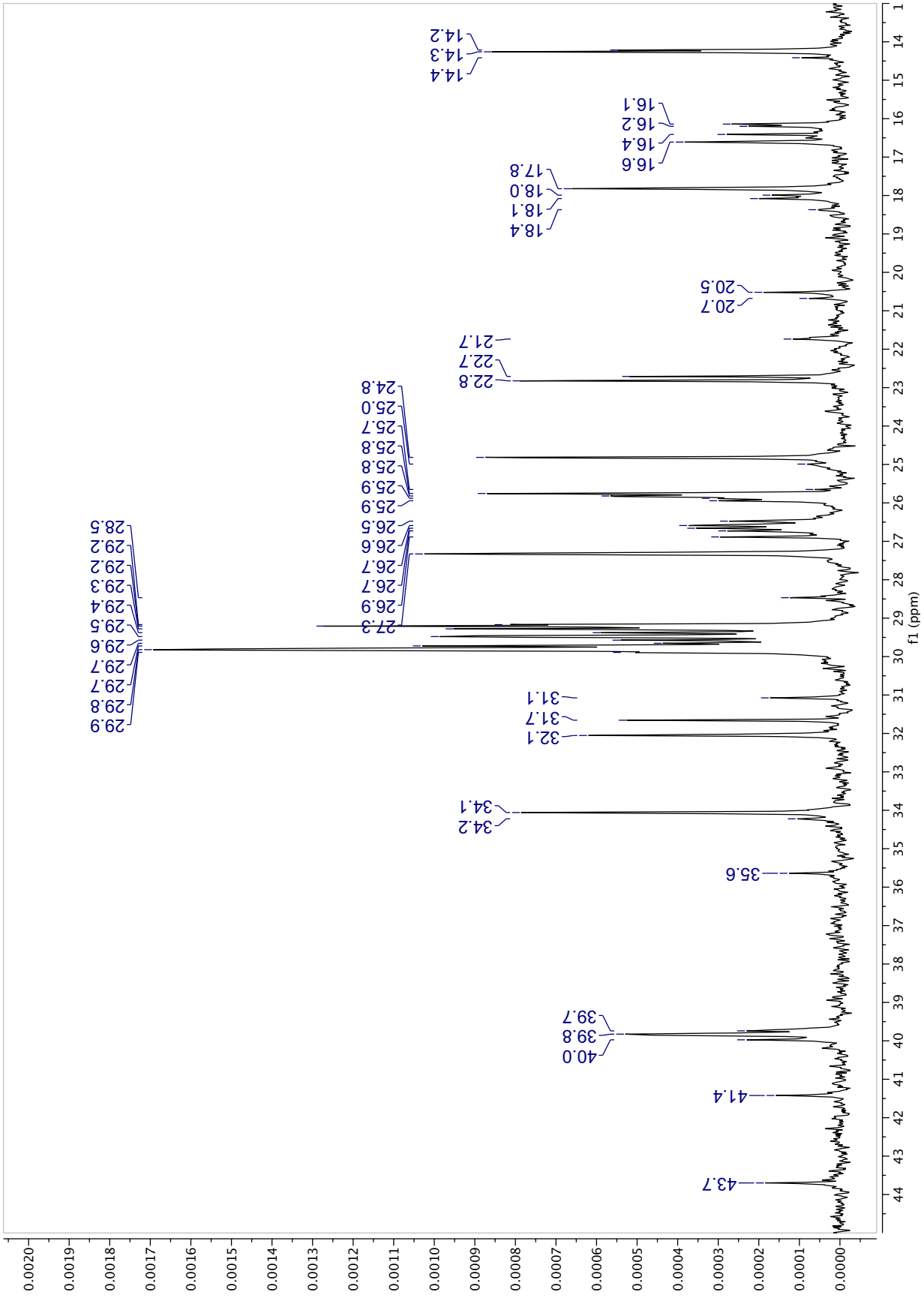


Figure S49. ^{13}C spectrum (Zoom 1: 13-45 ppm) (5000 scans) of F3 (40.4 mg) recorded in $CDCl_3$.

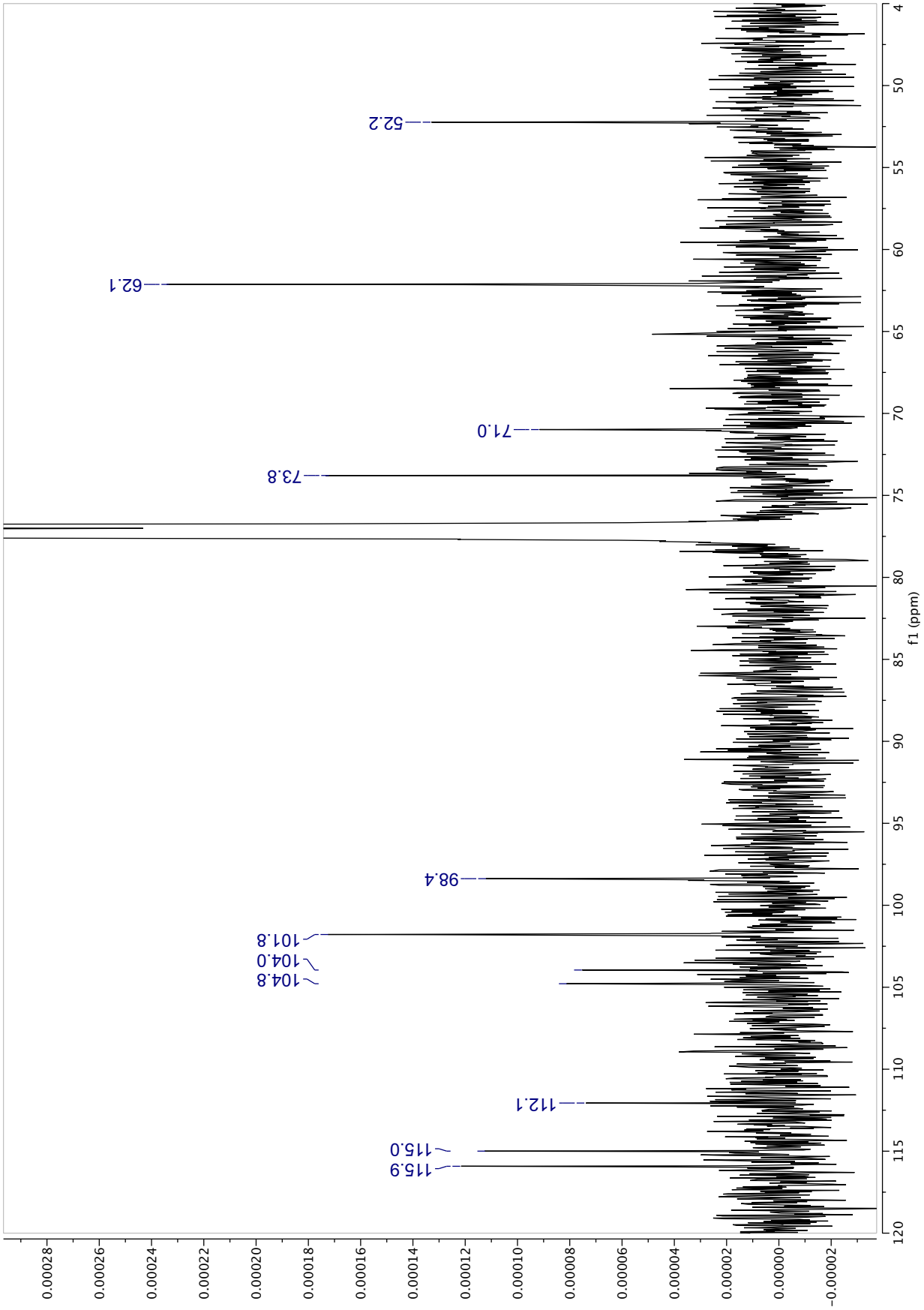


Figure S50. ^{13}C spectrum (Zoom 2: 45-120 ppm) (5000 scans) of F3 (40.4 mg) recorded in $CDCl_3$.

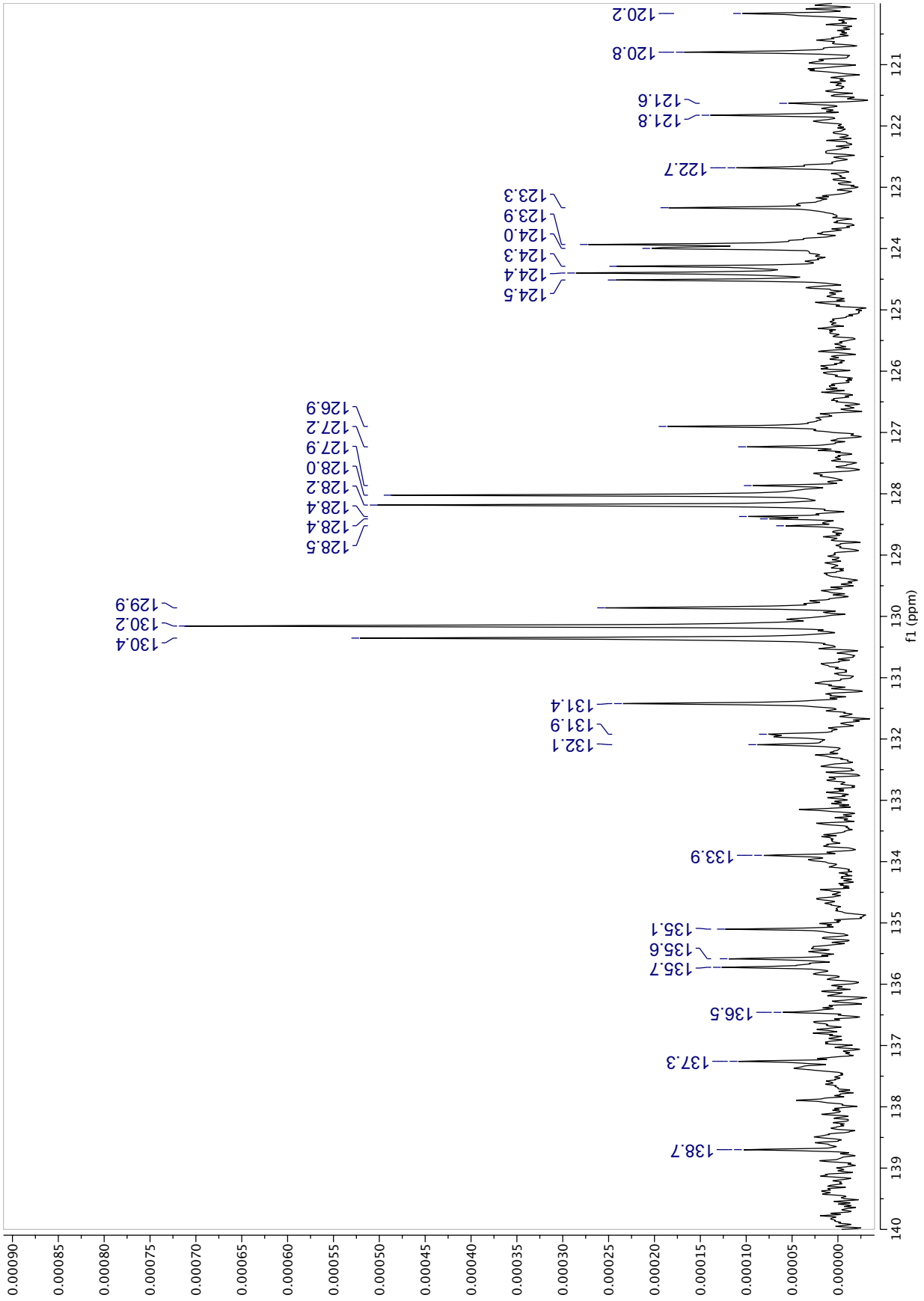


Figure S51. ^{13}C spectrum (Zoom 3: 120-140 ppm) (5000 scans) of F3 (40.4 mg) recorded in $CDCl_3$.

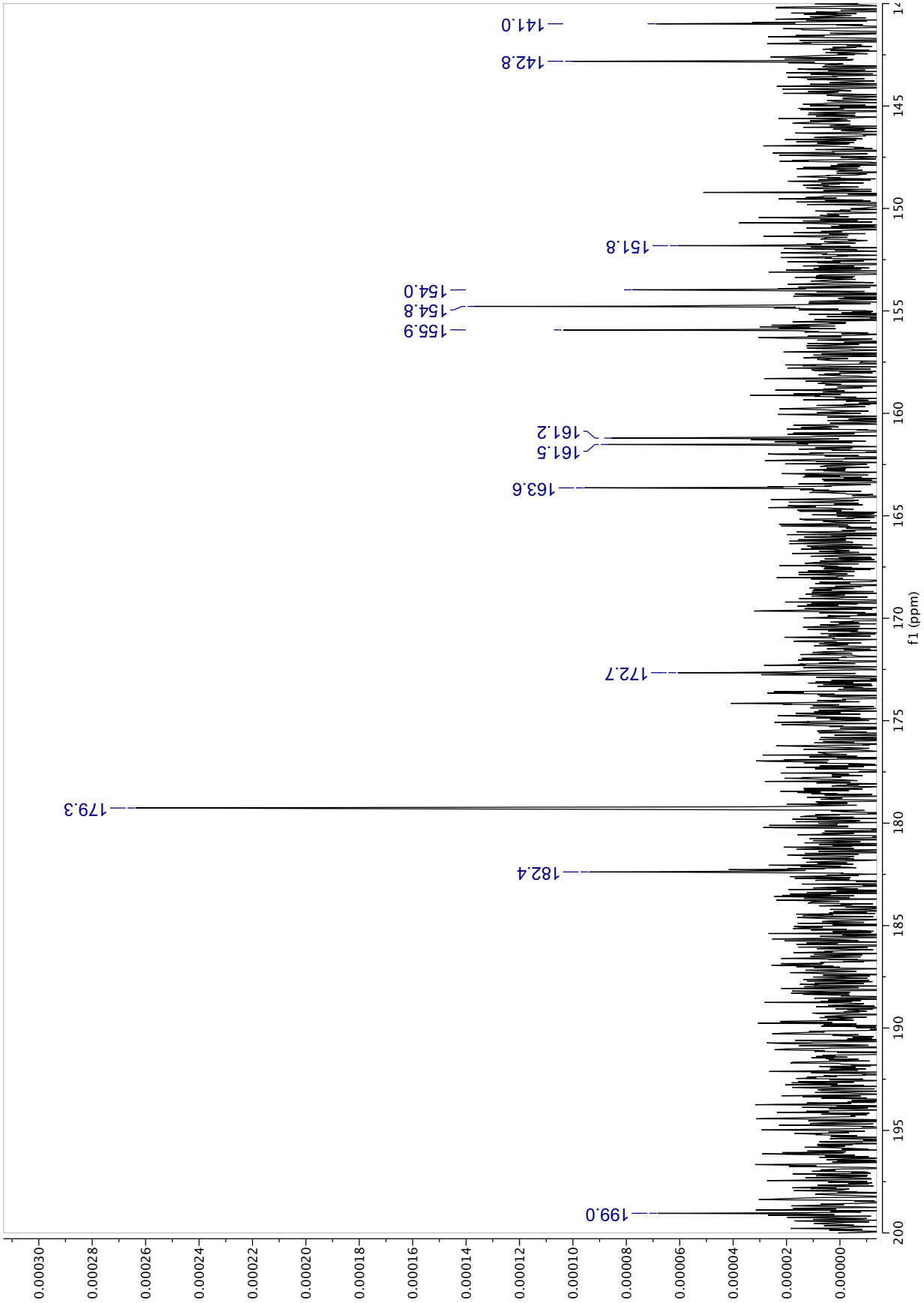


Figure S52. ^{13}C spectrum (Zoom 4: 140-200 ppm) (5000 scans) of F3 (40.4 mg) recorded in $CDCl_3$.

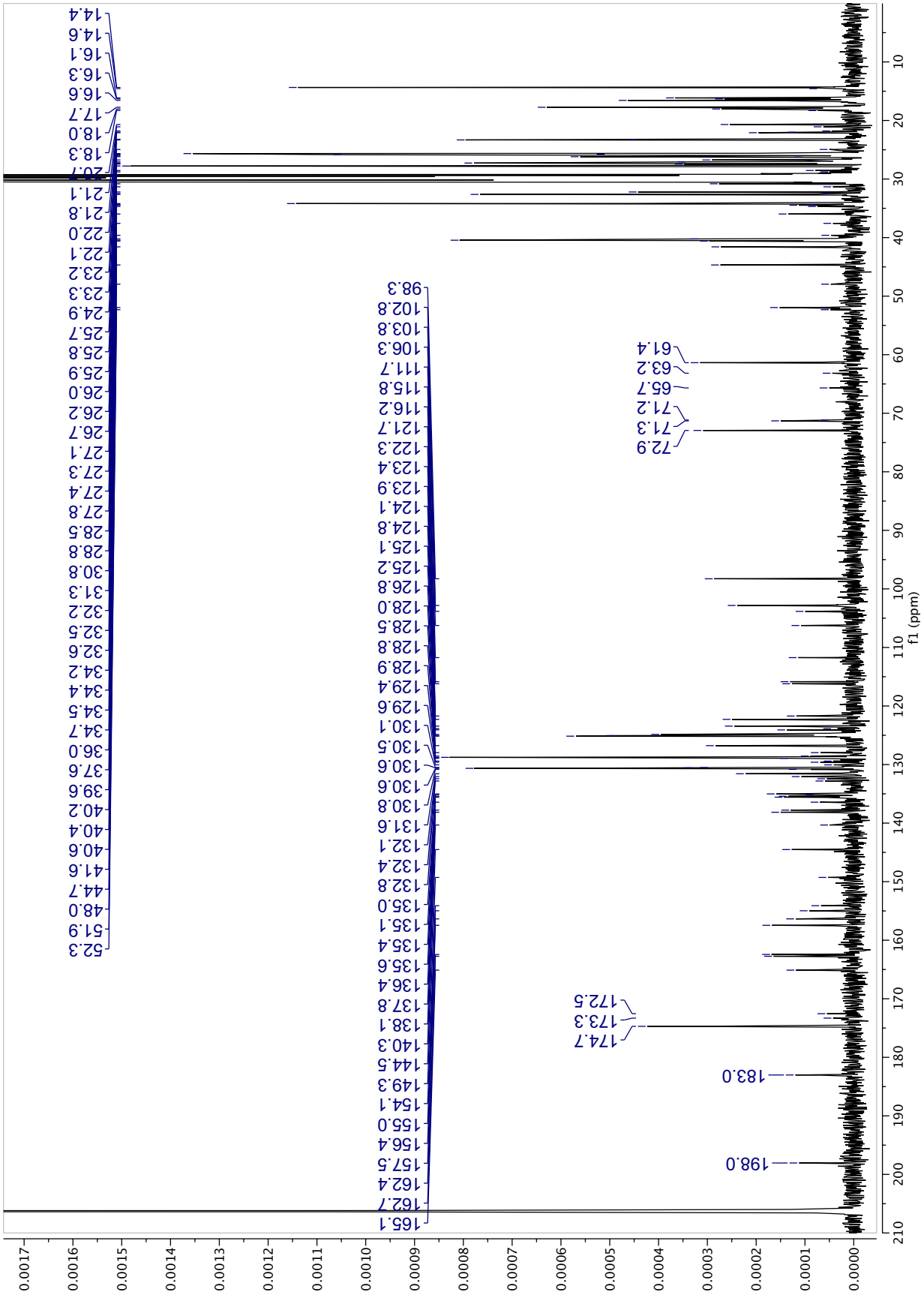


Figure S53. ^{13}C spectrum (8000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

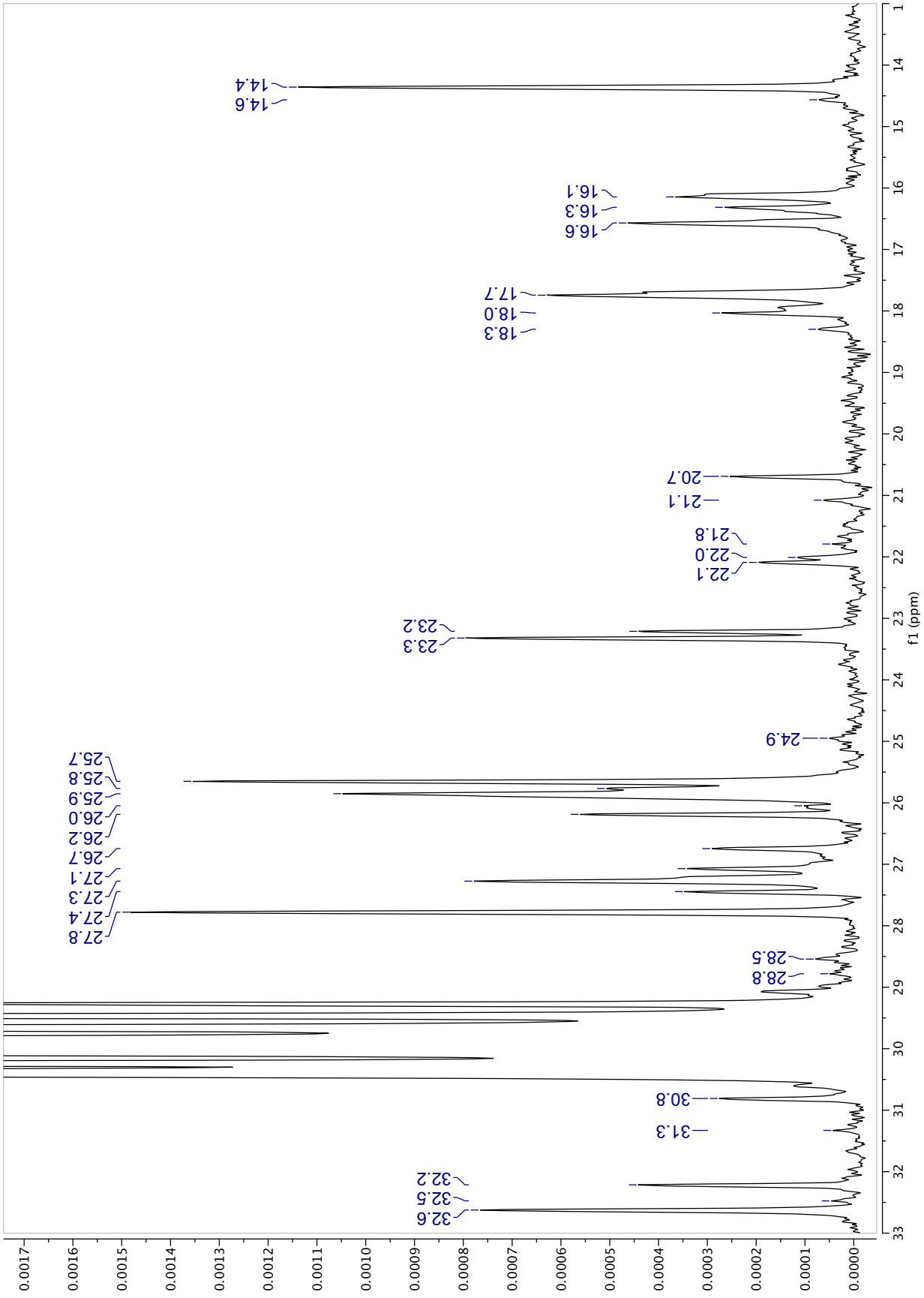


Figure S54. ^{13}C spectrum (Zoom 1: 13-33 ppm) (8000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

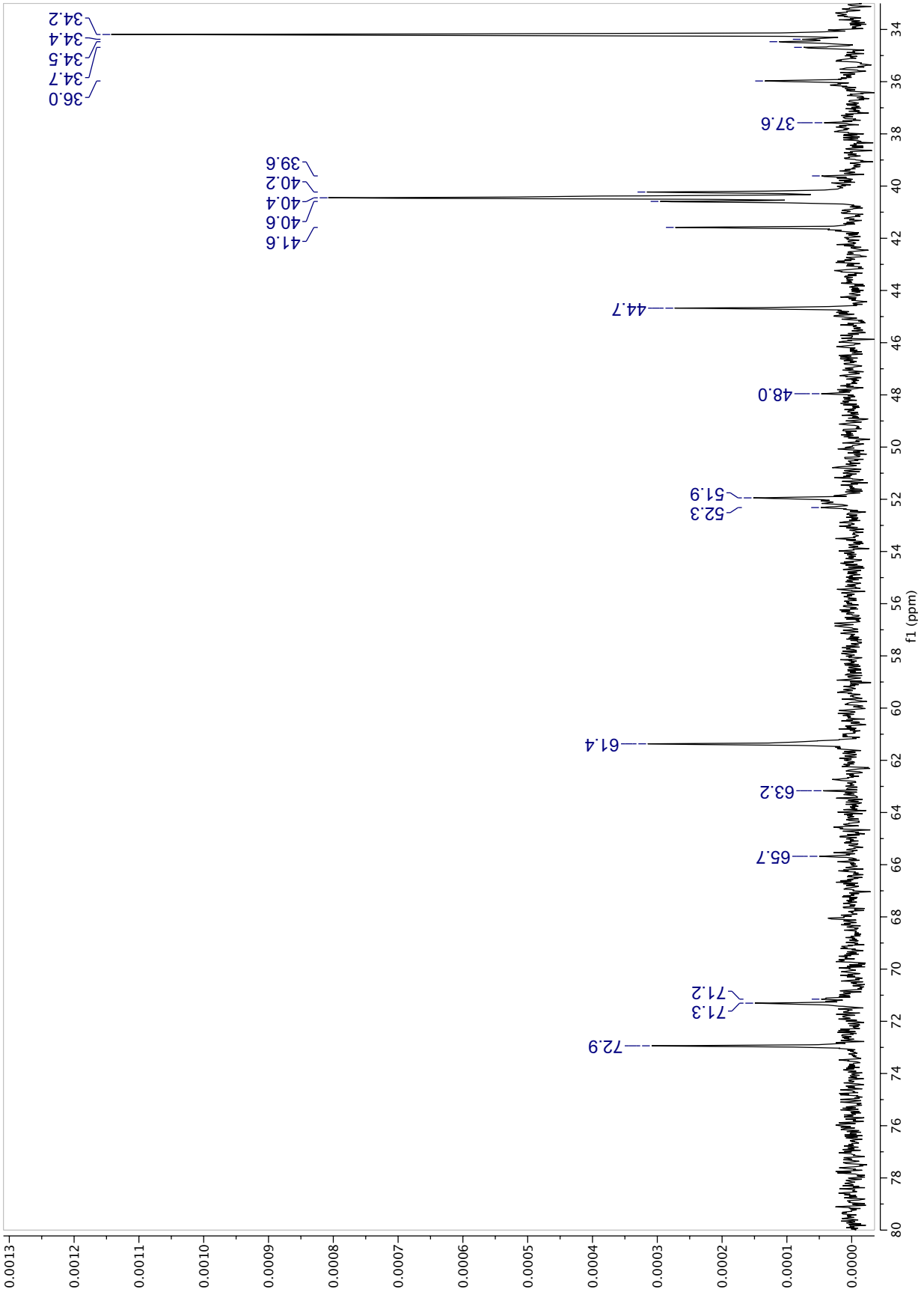


Figure S55. ^{13}C spectrum (Zoom 2: 33-80 ppm) (8000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

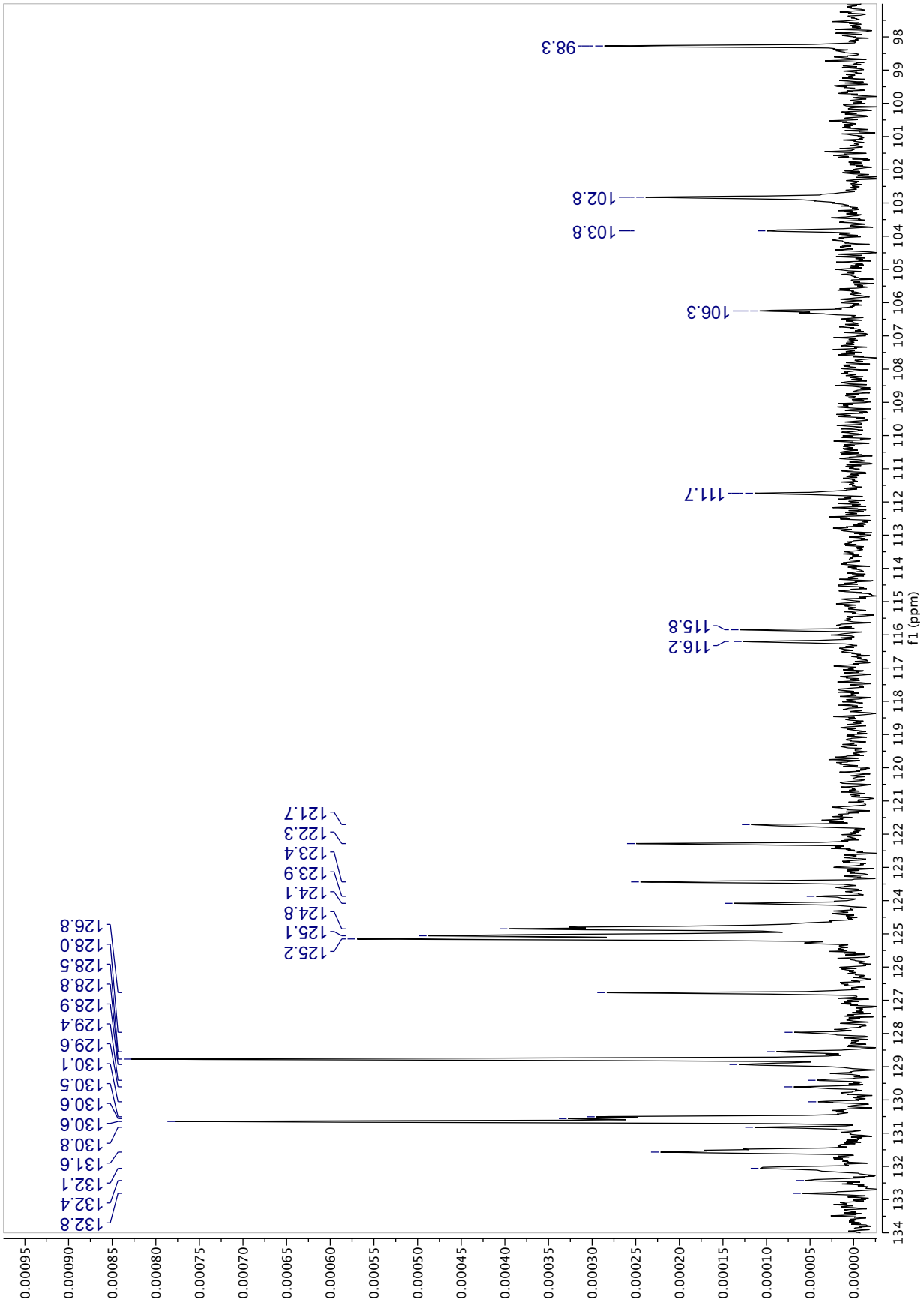


Figure S56. ^{13}C spectrum (Zoom 3: 97-134 ppm) (8000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

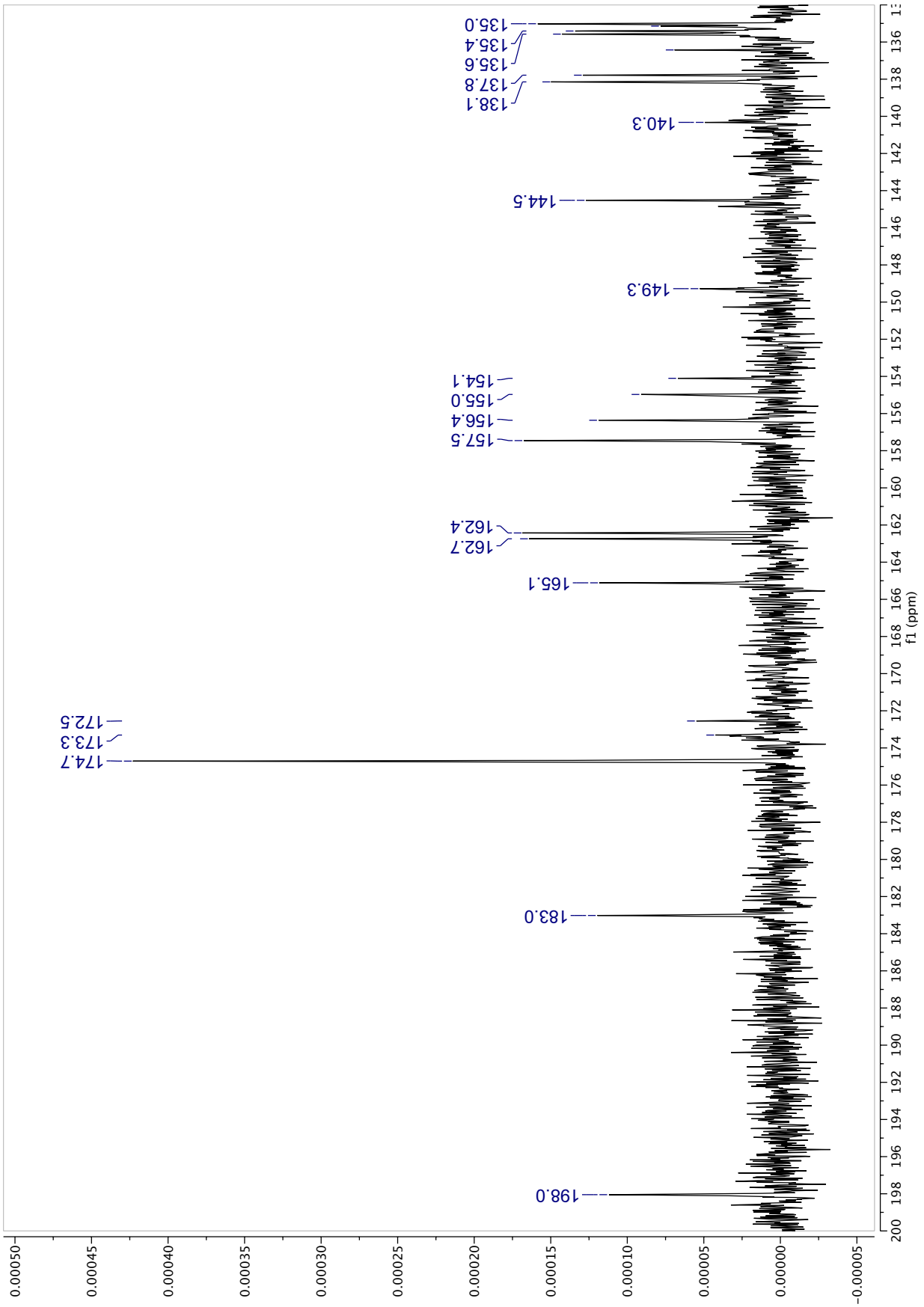


Figure S57. ^{13}C spectrum (Zoom 4: 135-200 ppm) (8000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

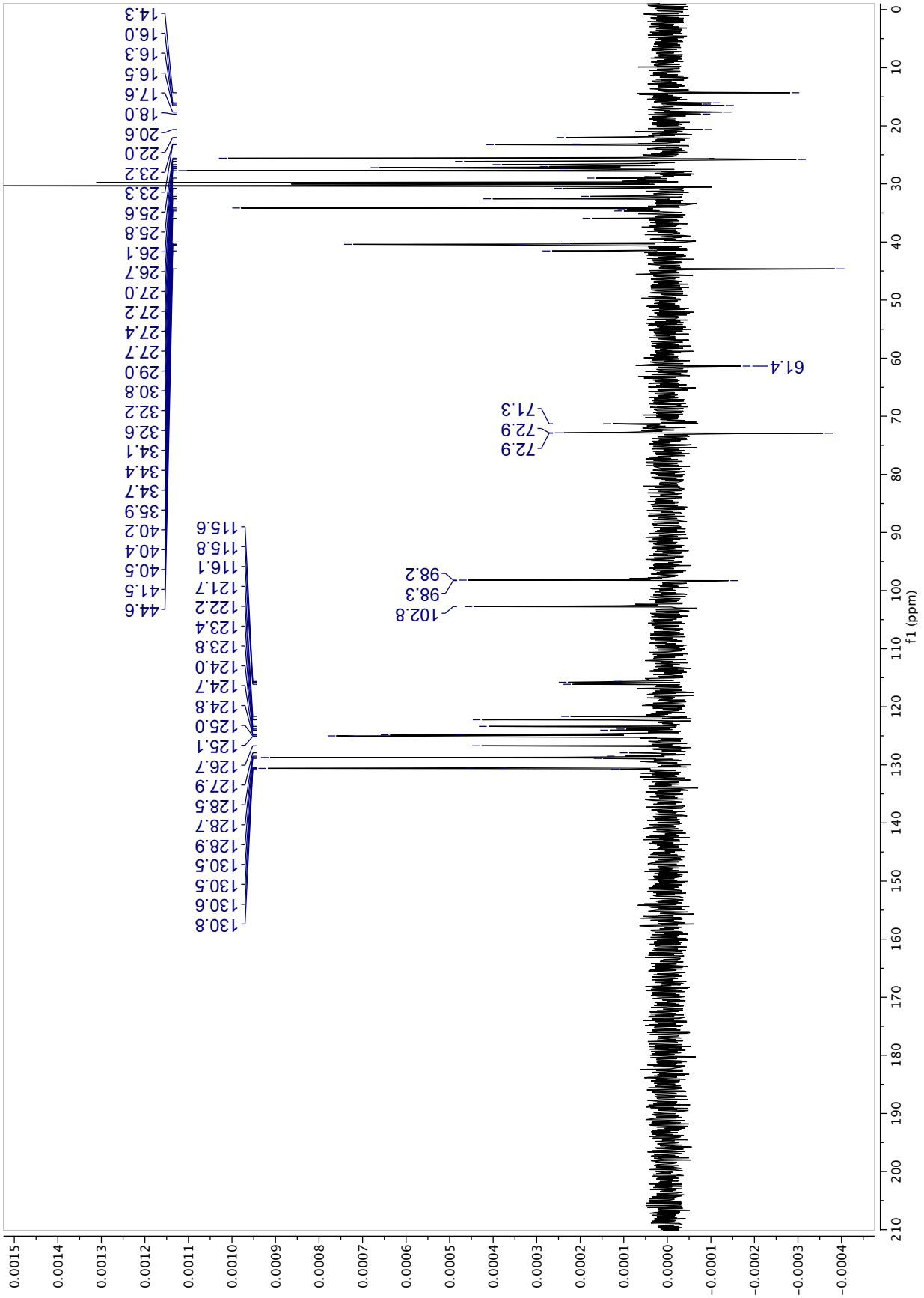


Figure S58. DEPT-135 spectrum (2000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

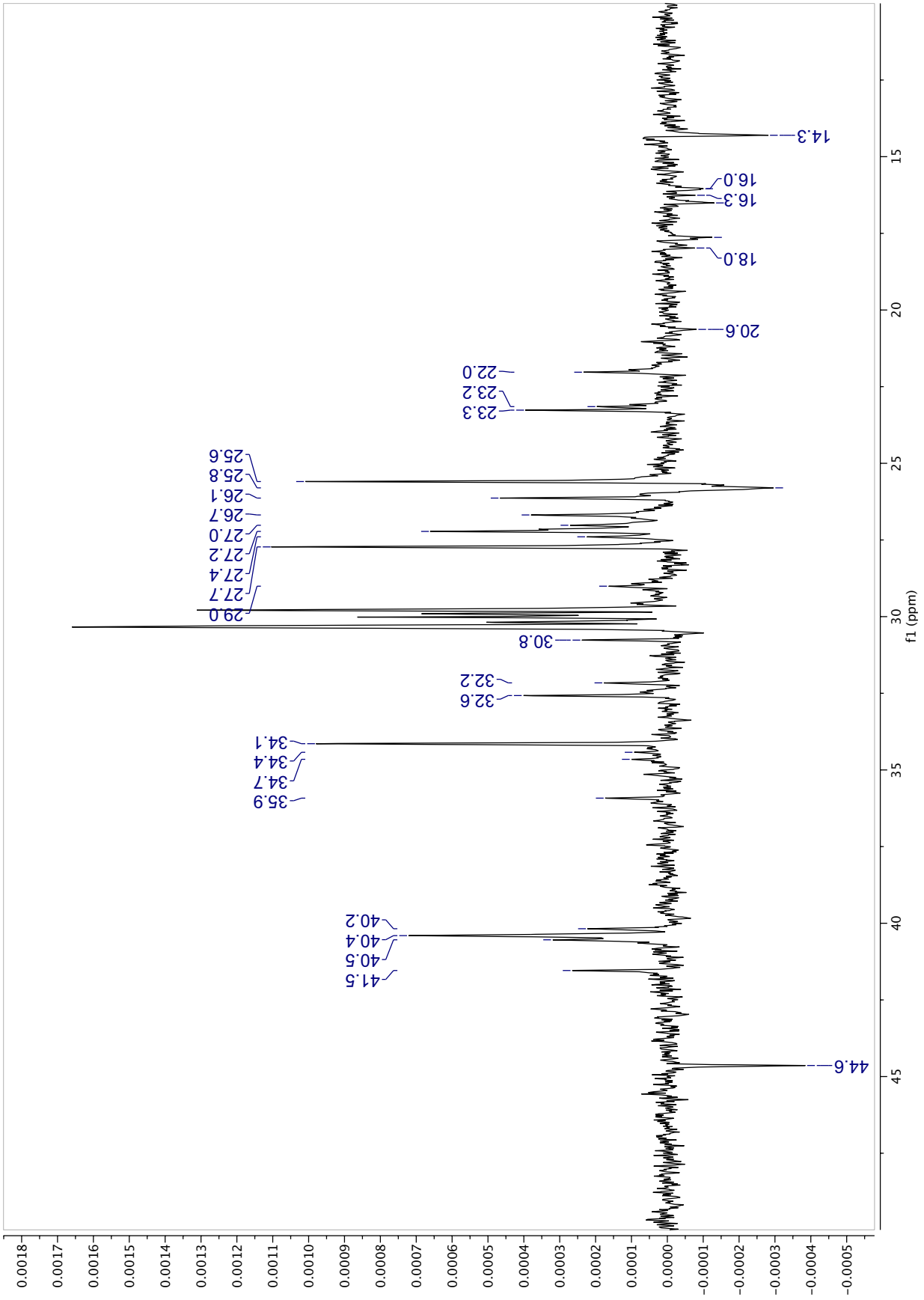


Figure S59. DEPT-135 spectrum (Zoom 1: 14-50 ppm) (2000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

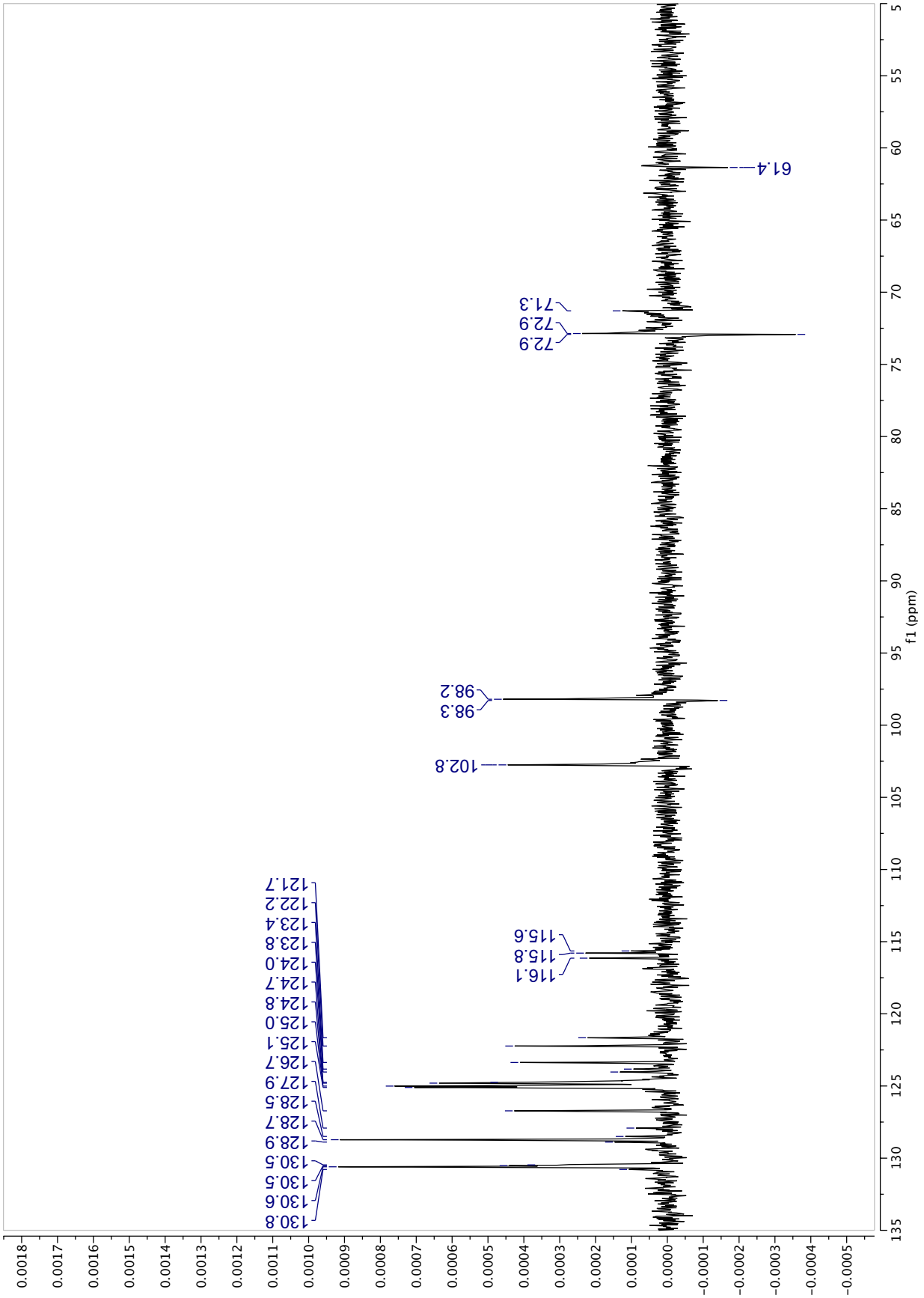


Figure S60. DEPT-135 spectrum (Zoom 2: 50-135 ppm) (2000 scans) of F3 (40.38 mg) recorded in acetone-*d*6.

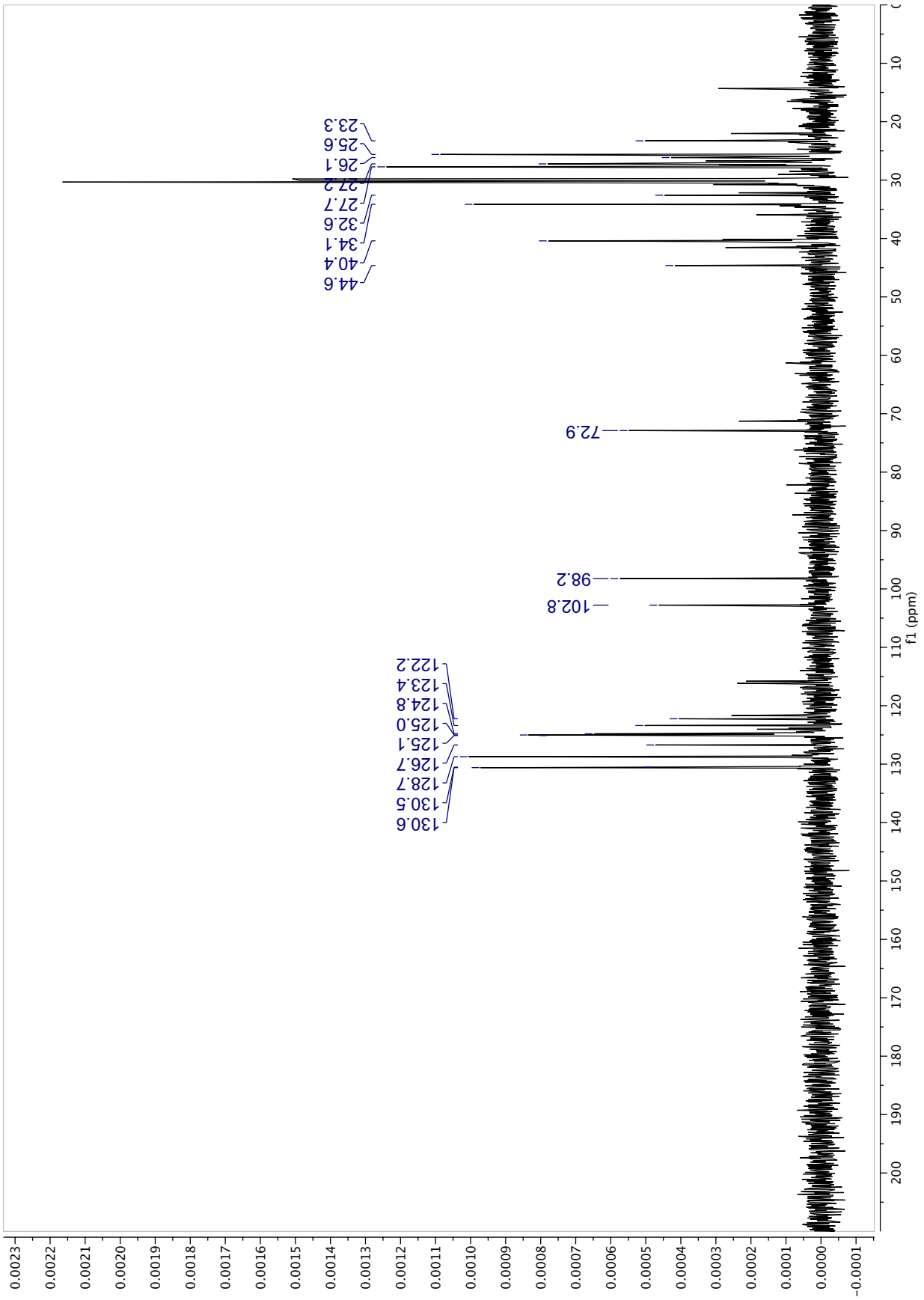


Figure S61. DEPT-90 spectrum (2000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

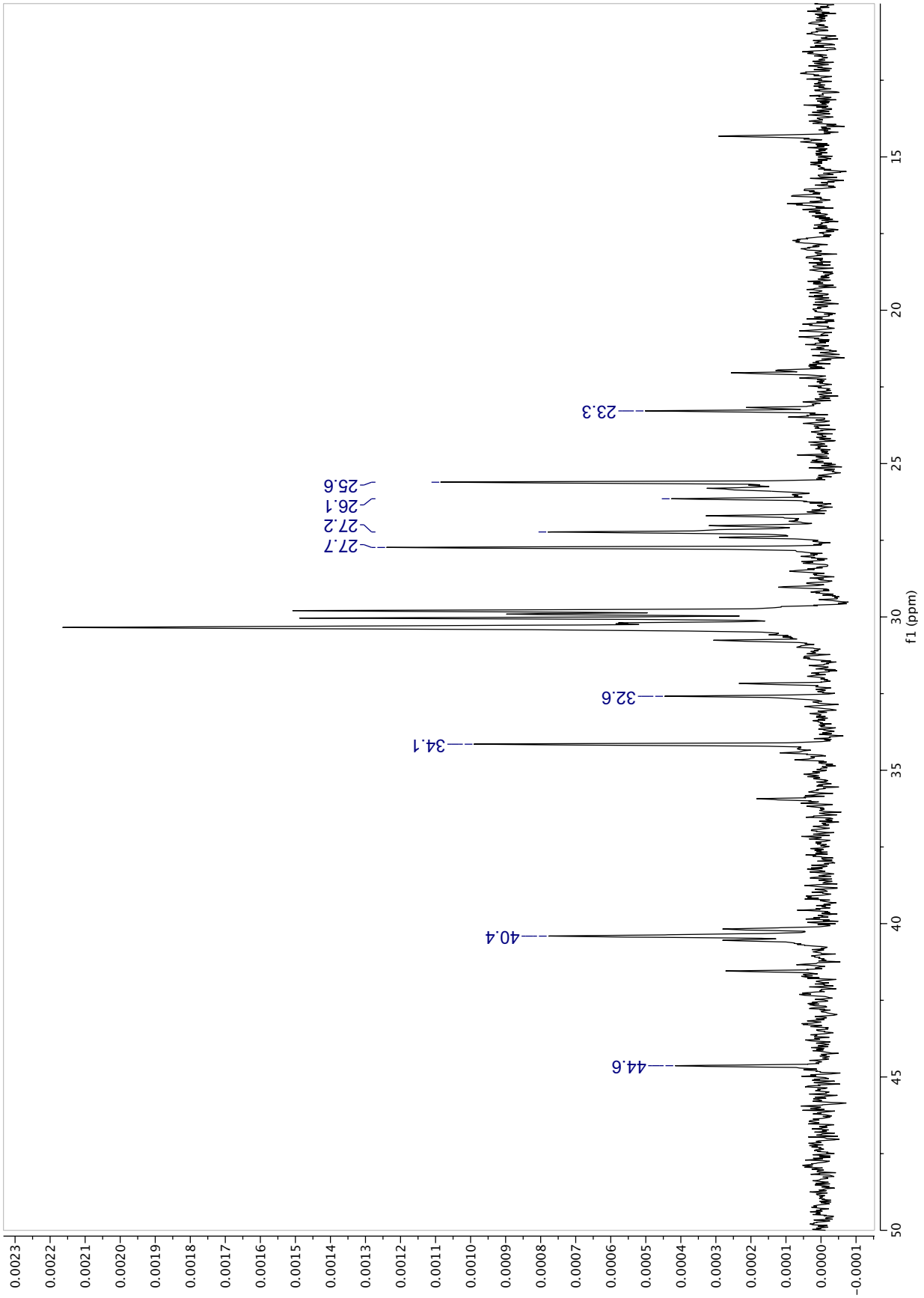


Figure S62. DEPT-90 spectrum (Zoom 1: 10-50 ppm) (2000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

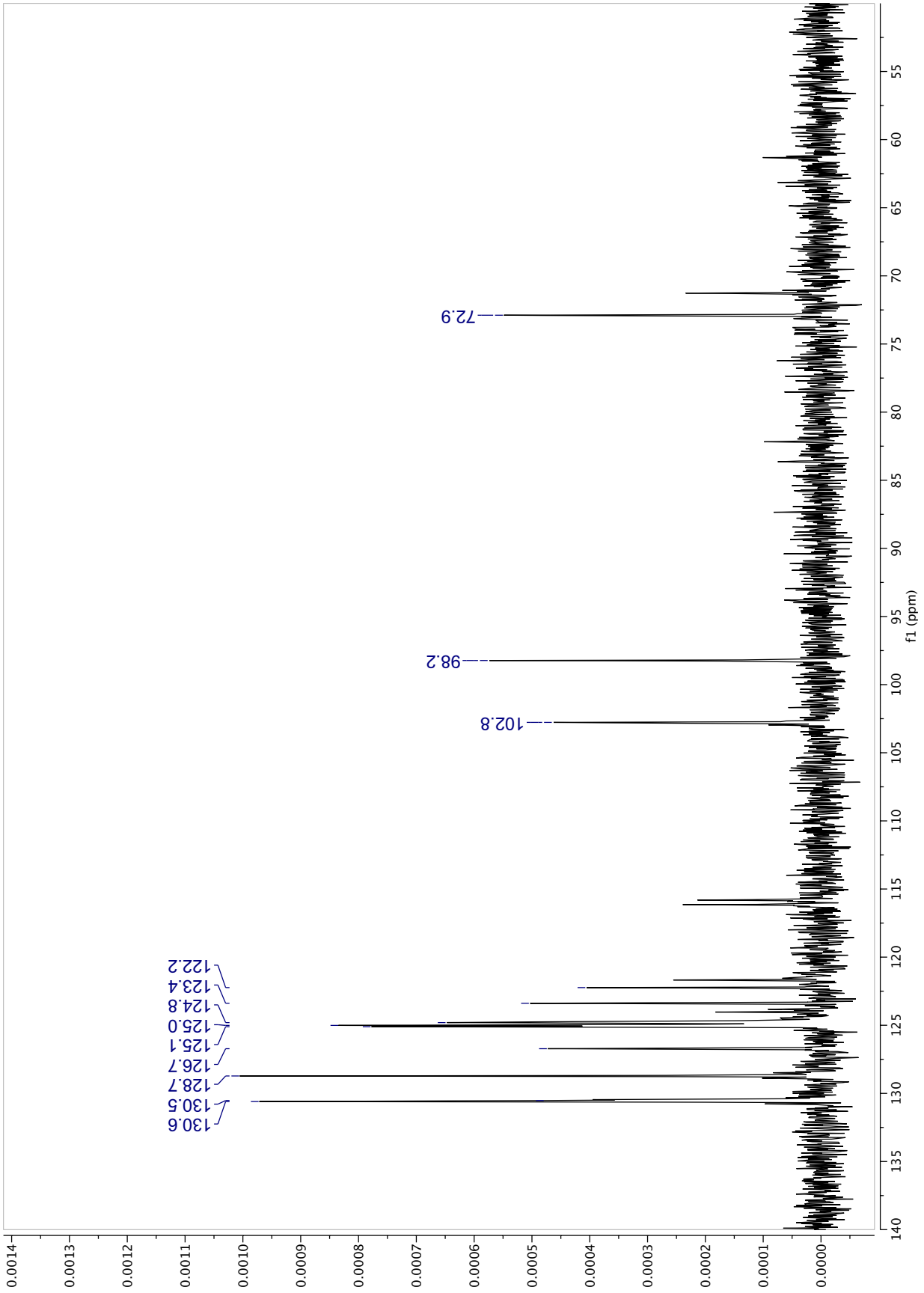


Figure S63. DEPT-90 spectrum (Zoom 2: 50-140 ppm) (2000 scans) of F3 (40.4 mg) recorded in acetone-*d*6.

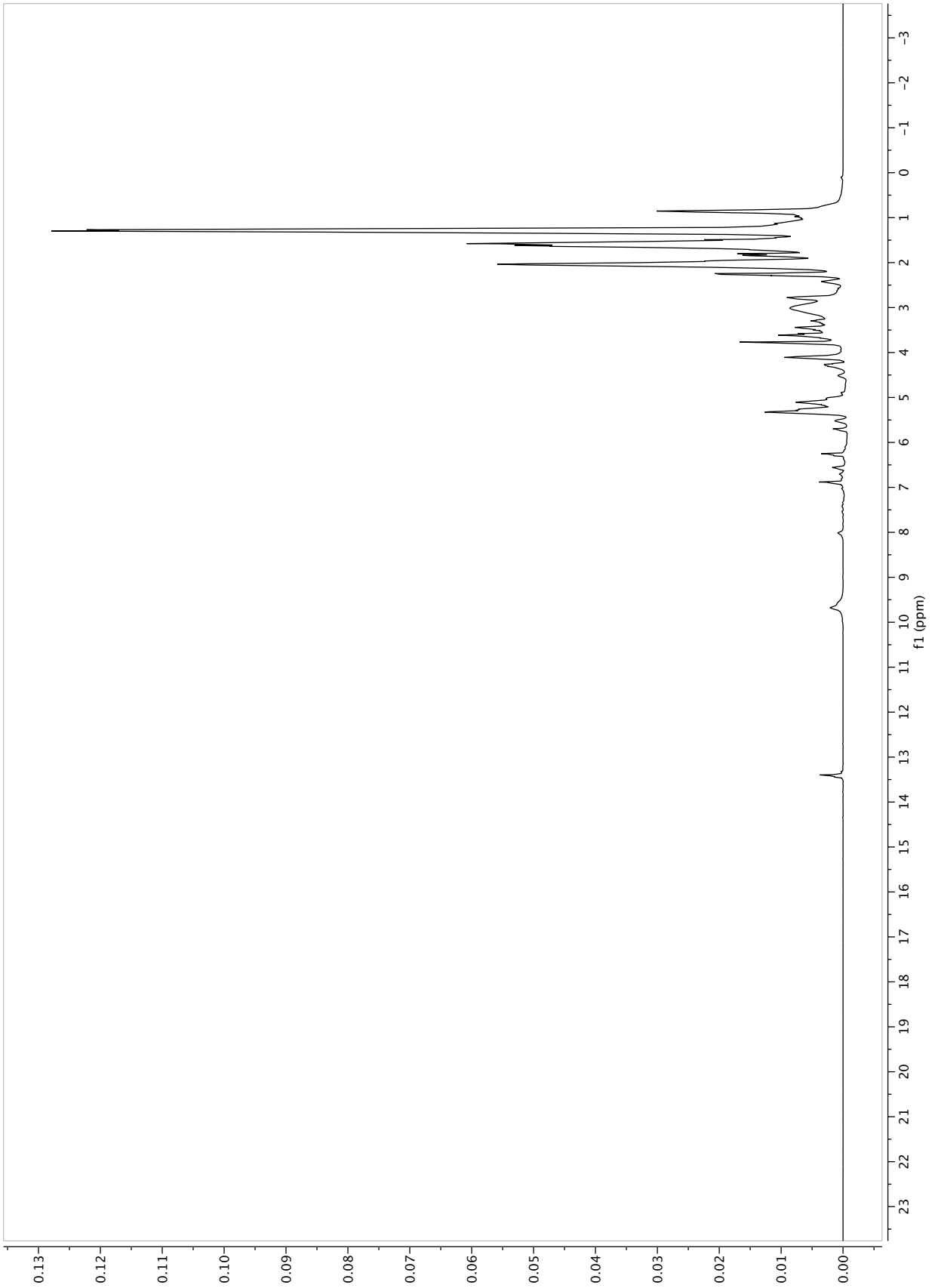


Figure S64. ^{1}H spectrum of F3 (40.4 mg) recorded in acetone-*d*6.

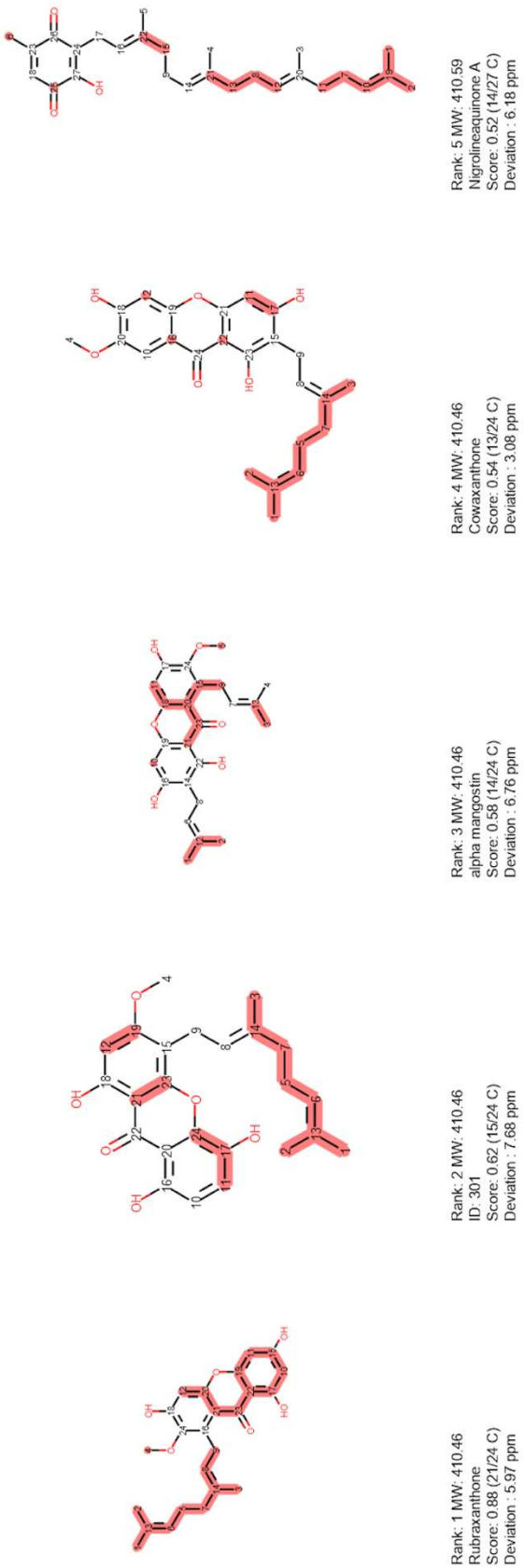


Figure S65. Part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and -135) of F6 using c-type *Garcinia* DB. Equivalent carbons were allowed. Only NPs with MW 410 g.mol^{-1} were allowed.

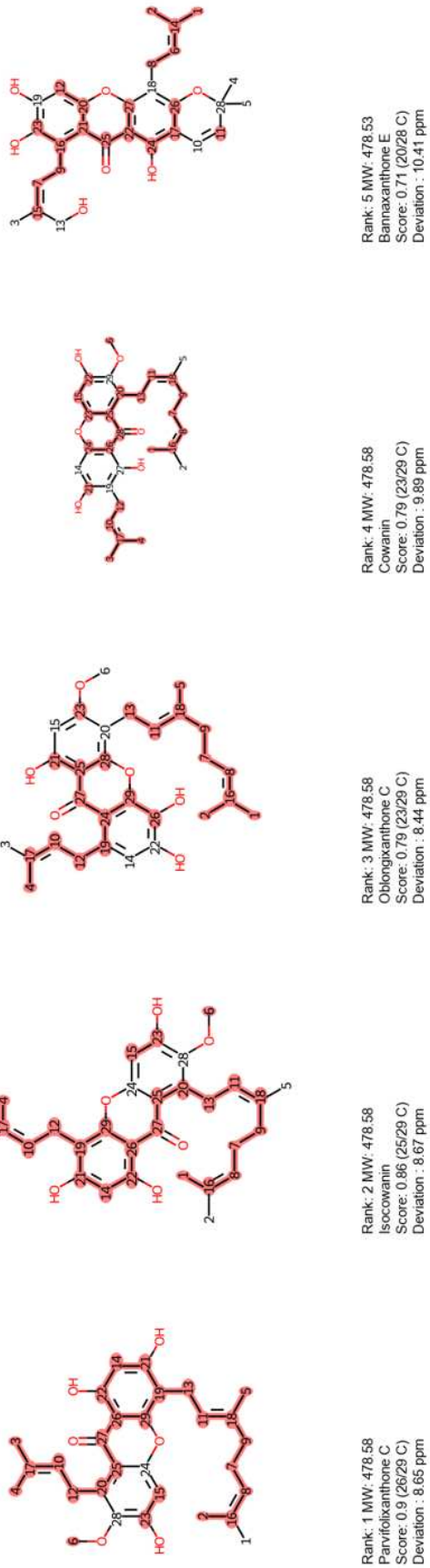


Figure S66. Part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and -135) of F3 using c-type *Garcinia* DB. Equivalent carbons were allowed. Only NPs with MW 478 g.mol^{-1} were allowed. Tolerance was set at 1.5 ppm.

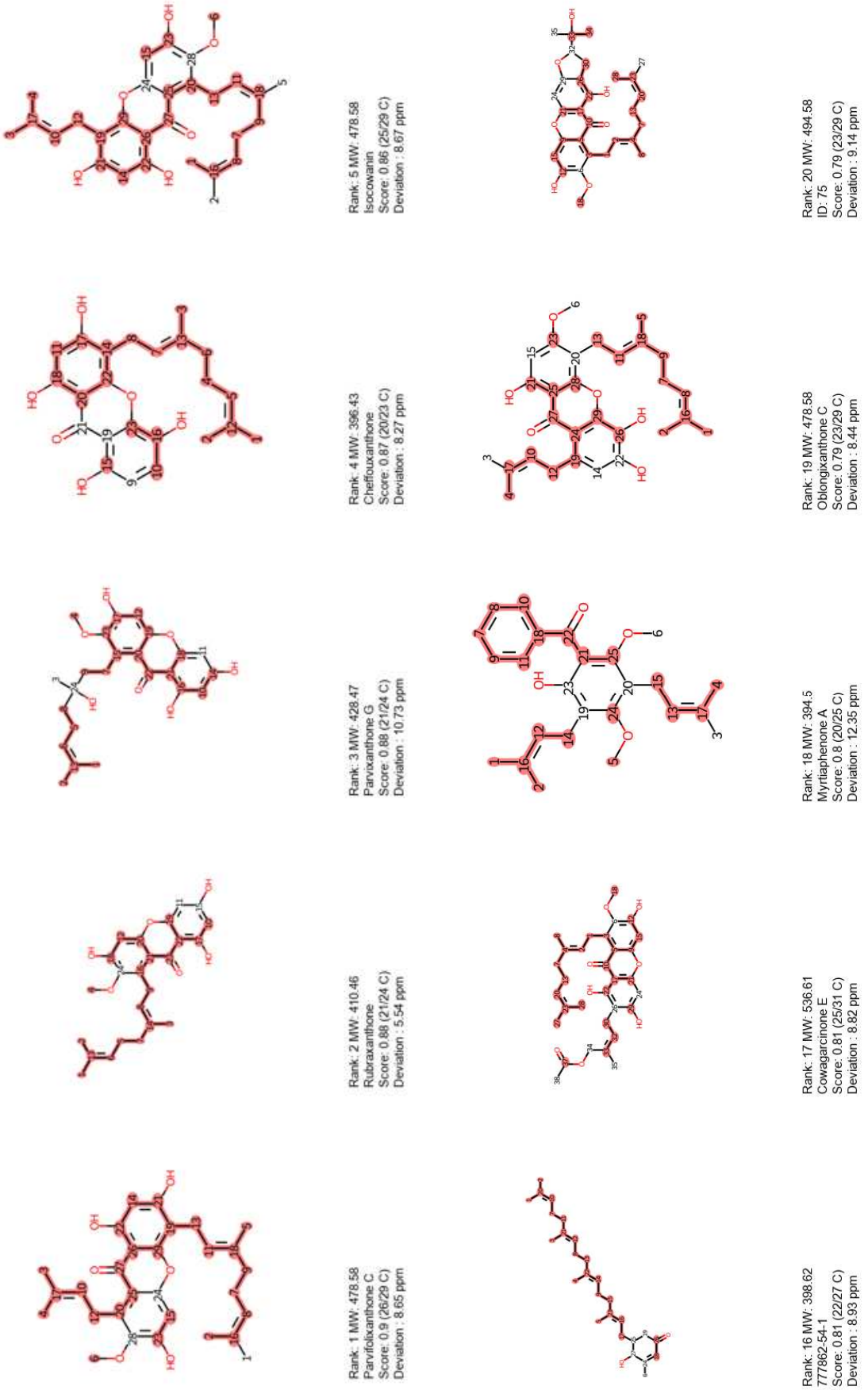


Figure S67. Part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and -135) of F3 using c-type *Garcinia* DB. Equivalent carbons were allowed. Tolerance was set at 1.5 ppm.

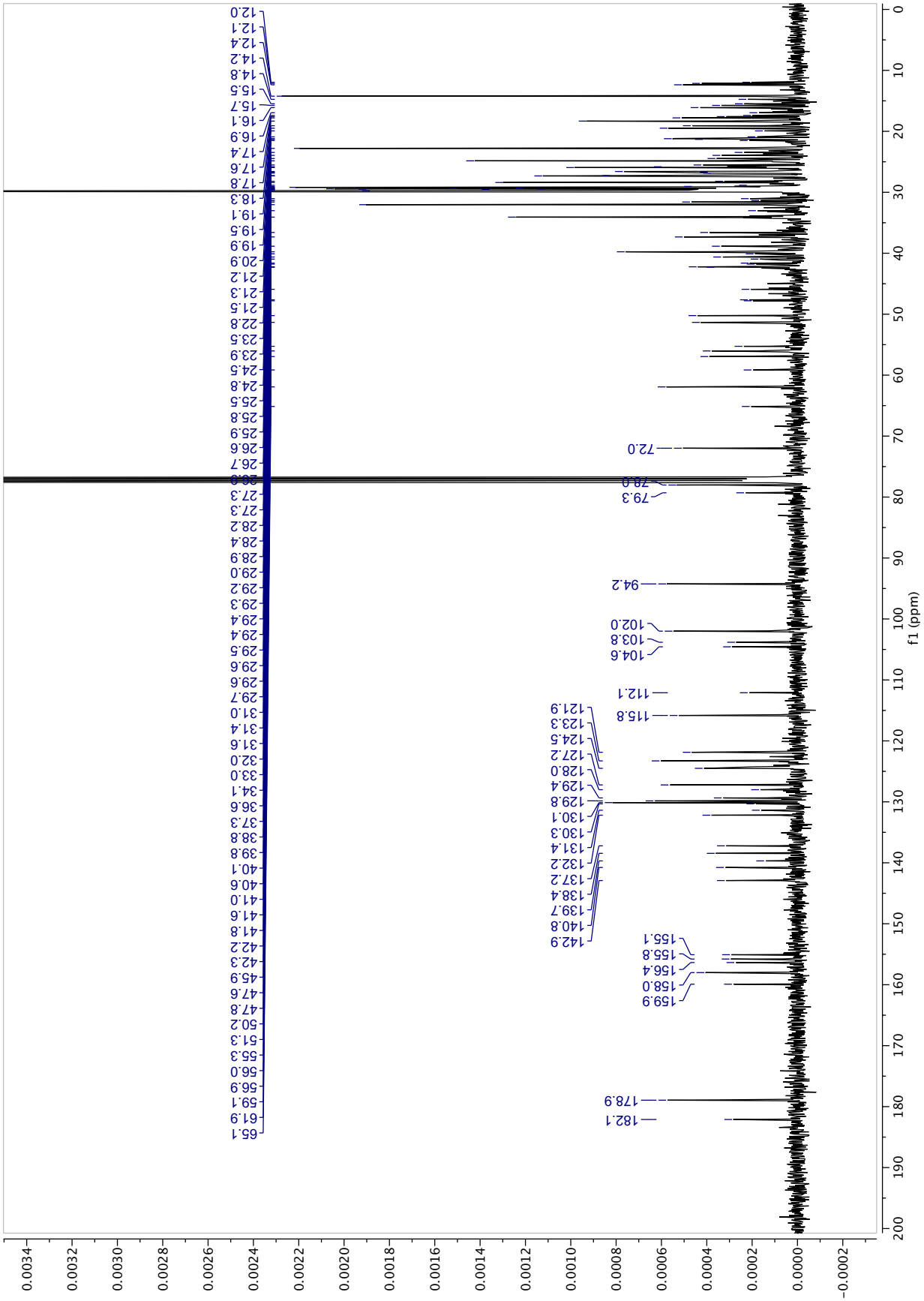


Figure S68. ^{13}C-NMR spectrum (2000 scans) of F2 (48.3 mg) recorded in $CDCl_3$.

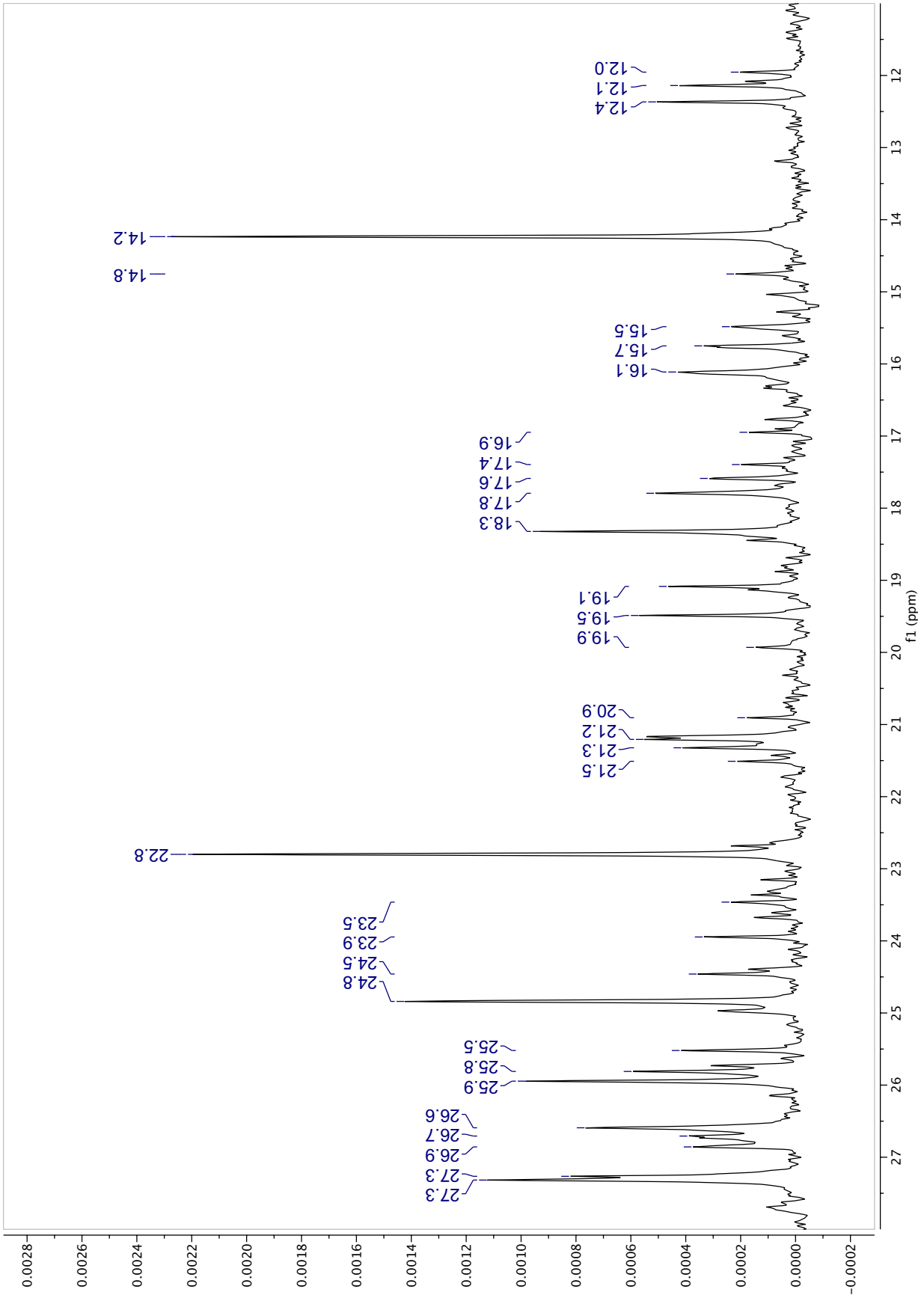


Figure S69. ^{13}C spectrum (Zoom 1: 11-28ppm) (2000 scans) of F2 (48.3 mg) recorded in $CDCl_3$.

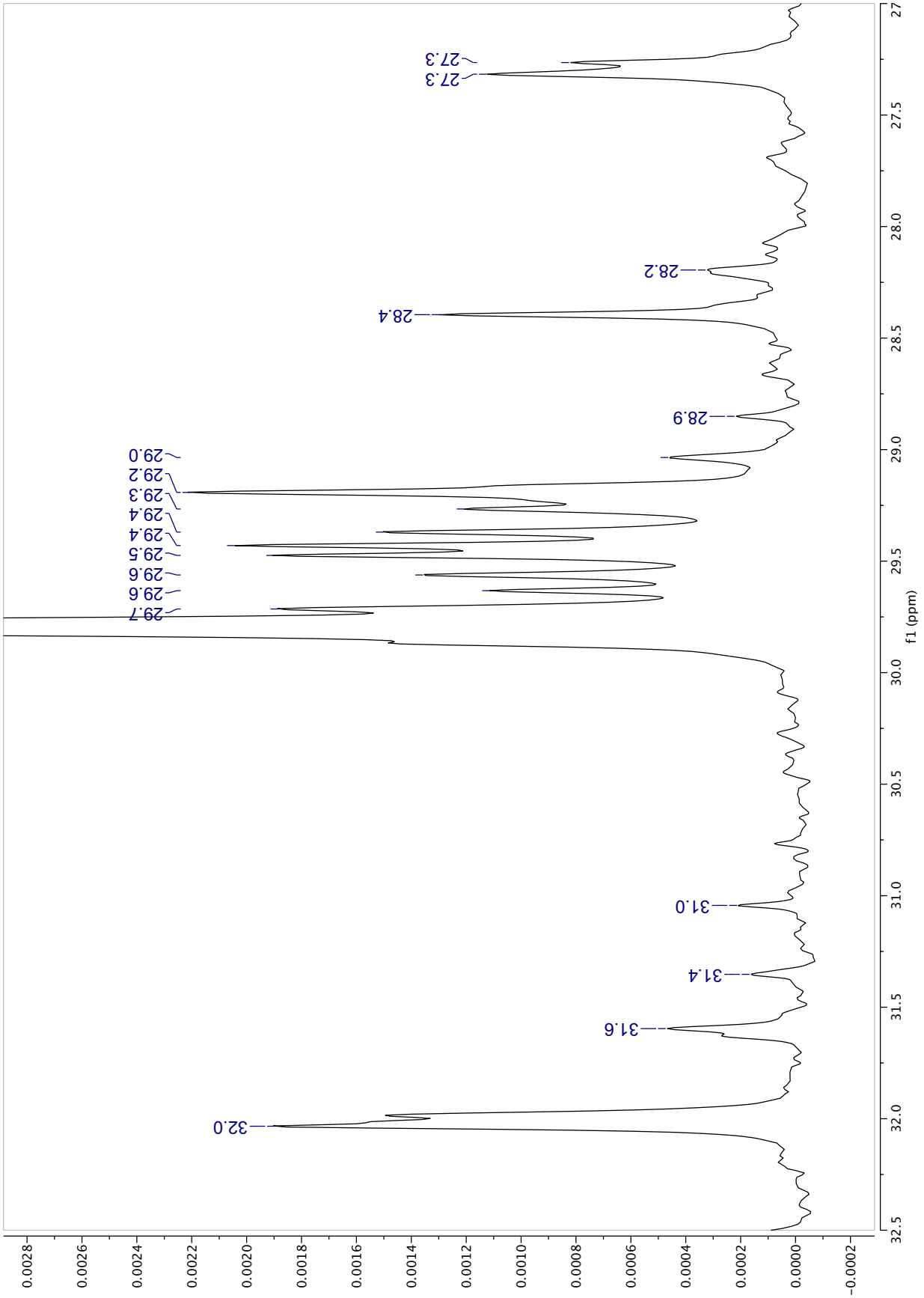


Figure S70. ^{13}C spectrum (Zoom 2: 27-32.5 ppm) (2000 scans) of F2 (48.3 mg) recorded in $CDCl_3$.

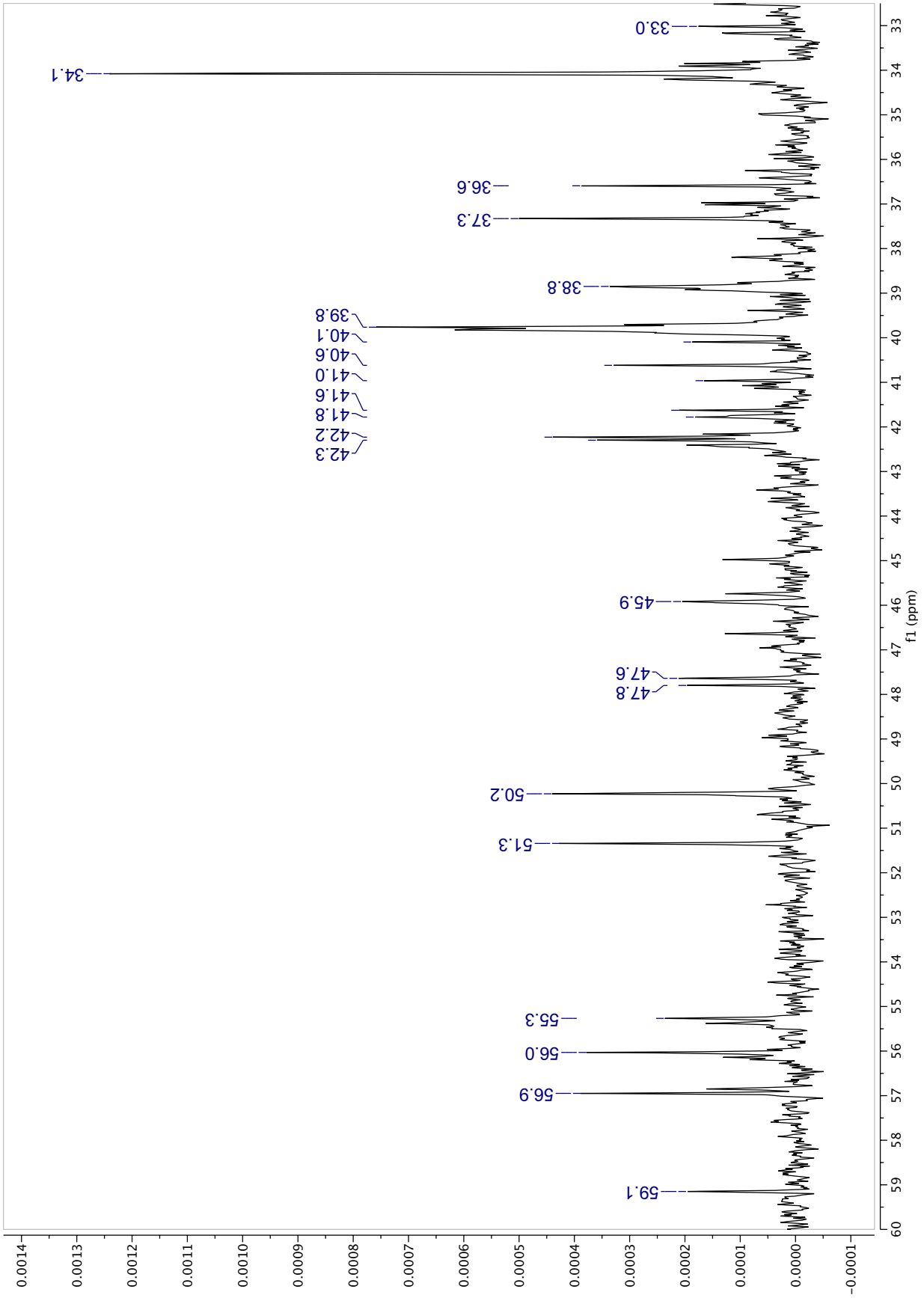


Figure S71. ^{13}C spectrum (Zoom 3: 32.5-60 ppm) (2000 scans) of F2 (48.3 mg) recorded in $CDCl_3$.

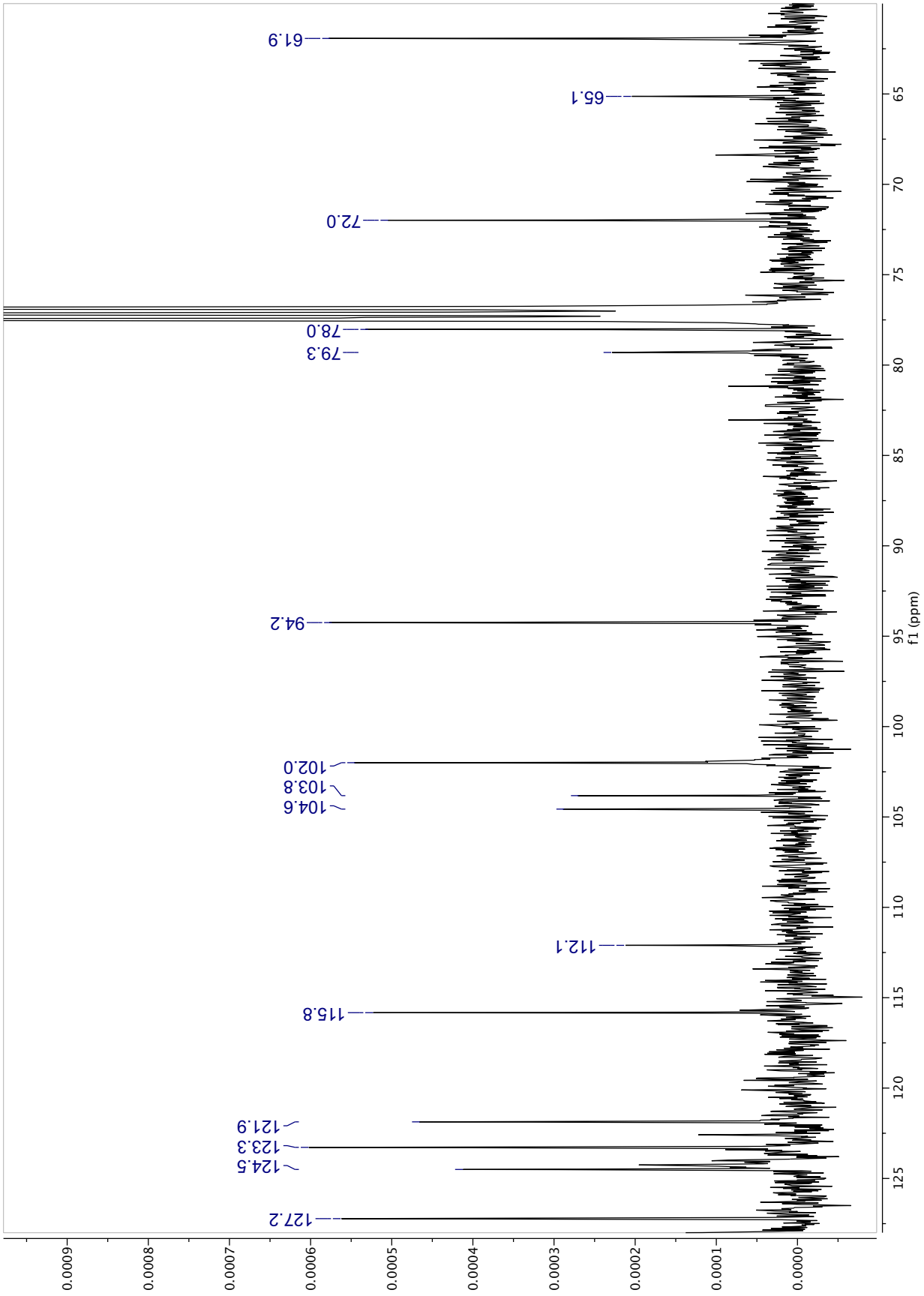


Figure S72. ^{13}C spectrum (Zoom 4: 60-128 ppm) (2000 scans) of F2 (48.3 mg) recorded in $CDCl_3$.

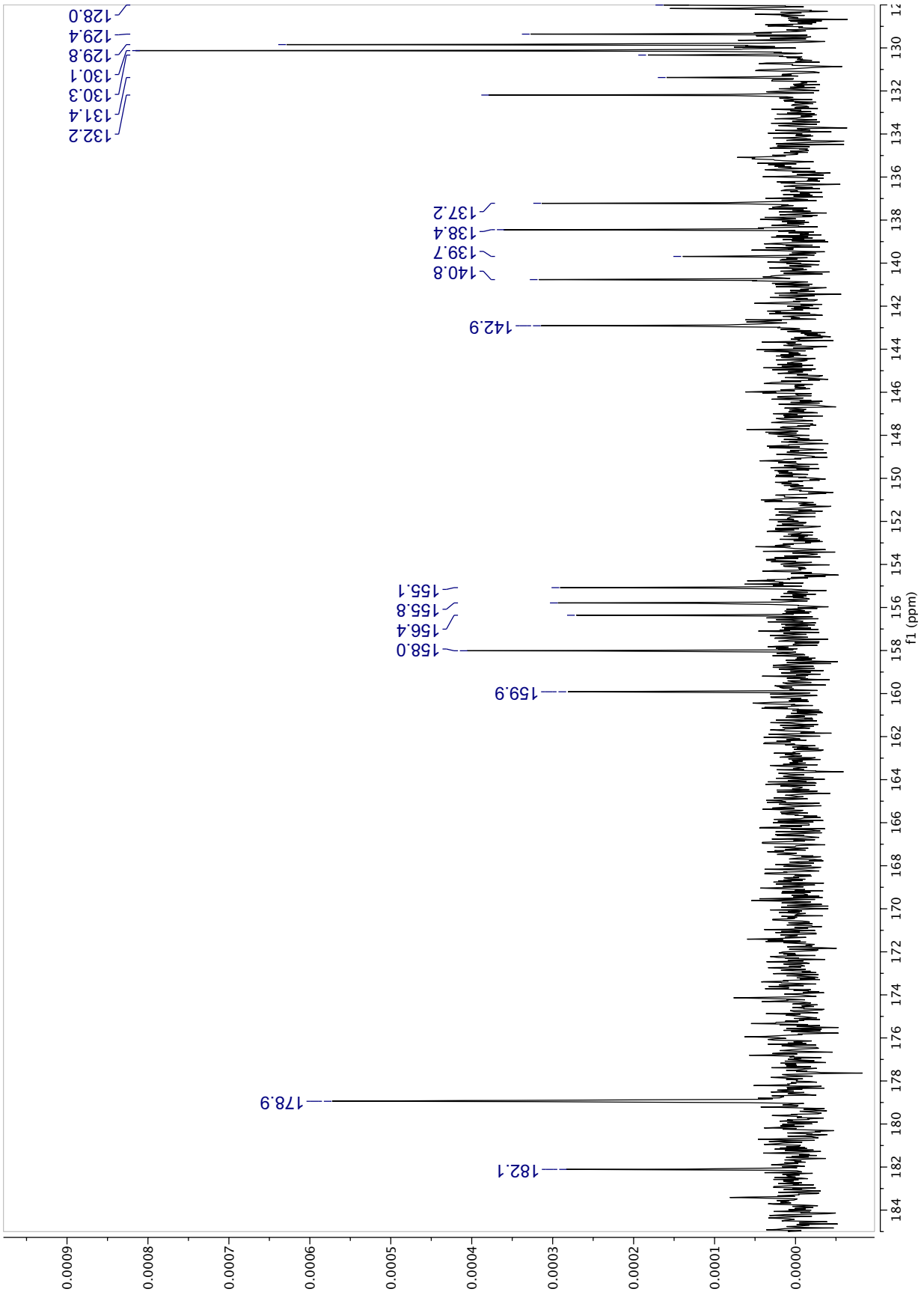


Figure S73. ^{13}C spectrum (Zoom 5: 128-185 ppm) (2000 scans) of F2 (48.3 mg) recorded in $CDCl_3$.

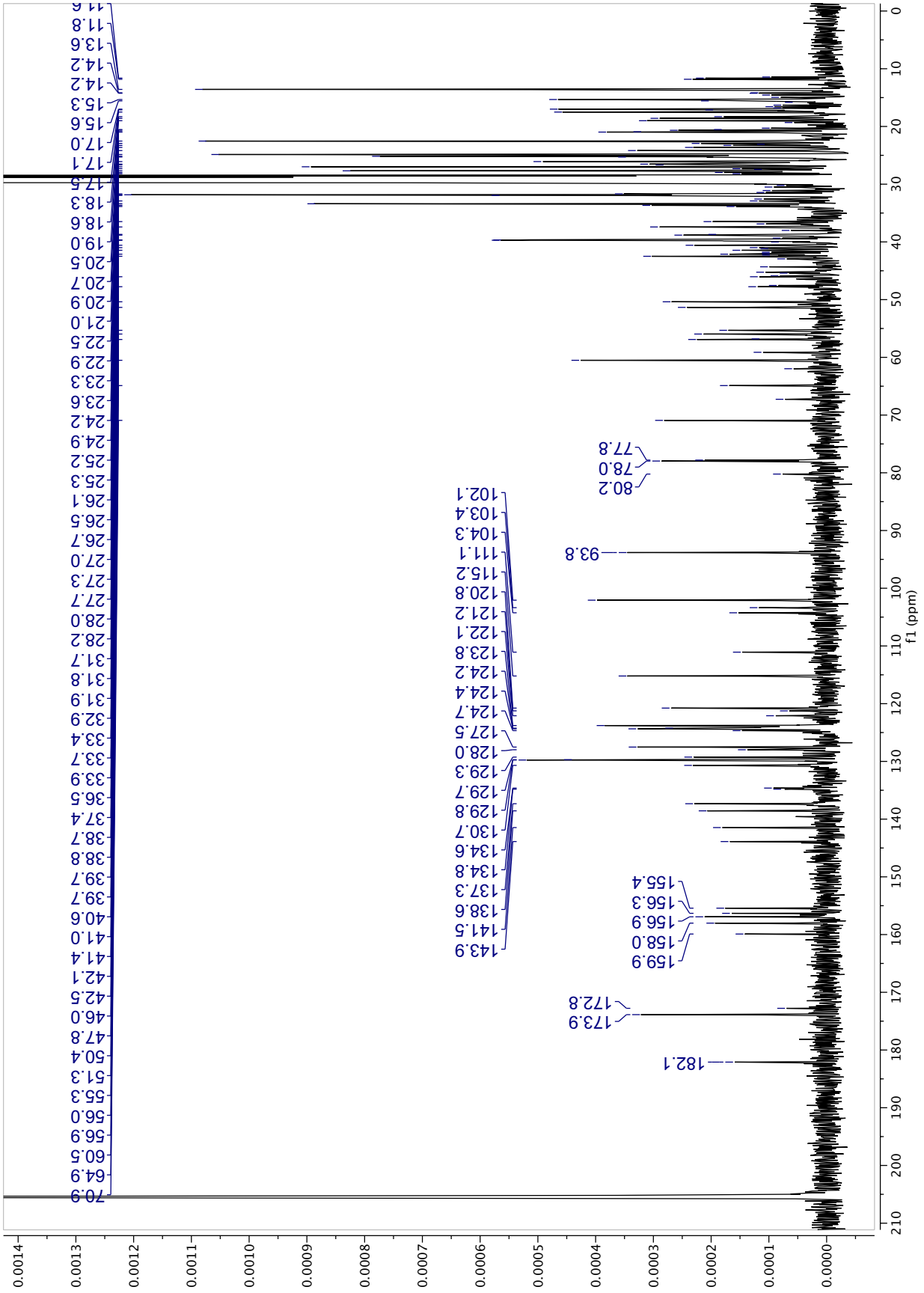


Figure S74. ^{13}C spectrum (8000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

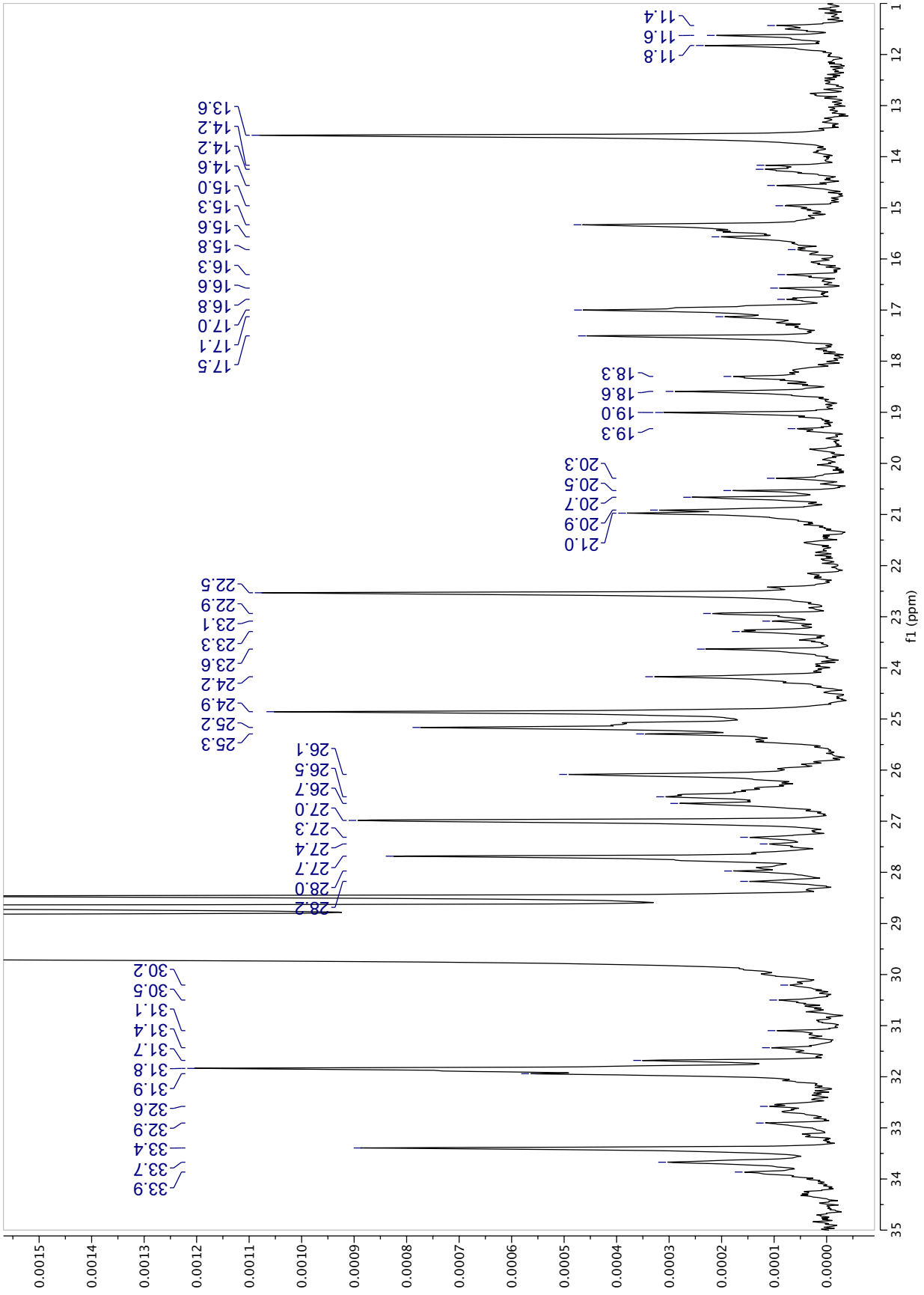


Figure S75. ^{13}C spectrum (Zoom 1: 10-35 ppm) (8000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

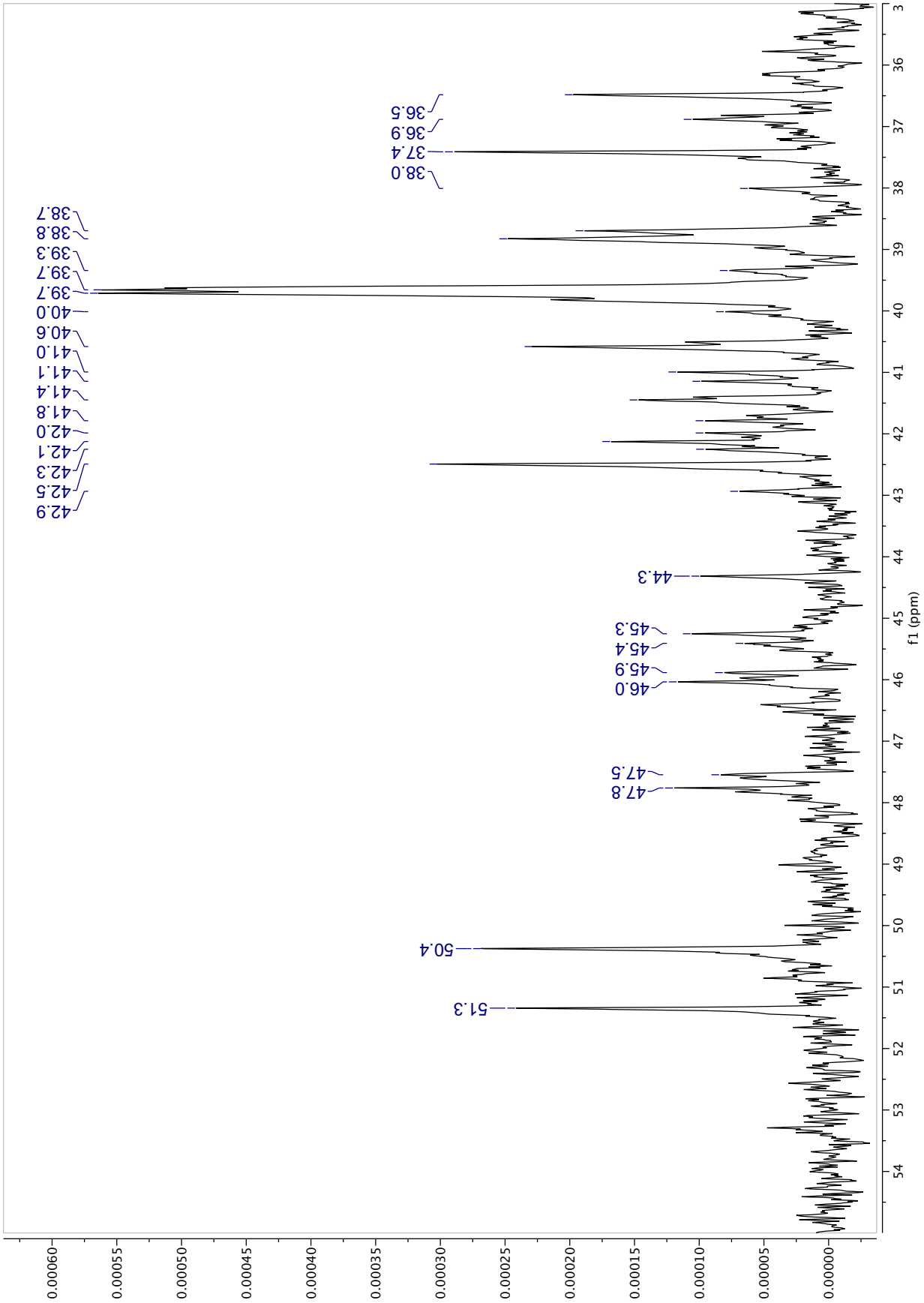


Figure S76. ^{13}C spectrum (Zoom 2: 35-55 ppm) (8000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

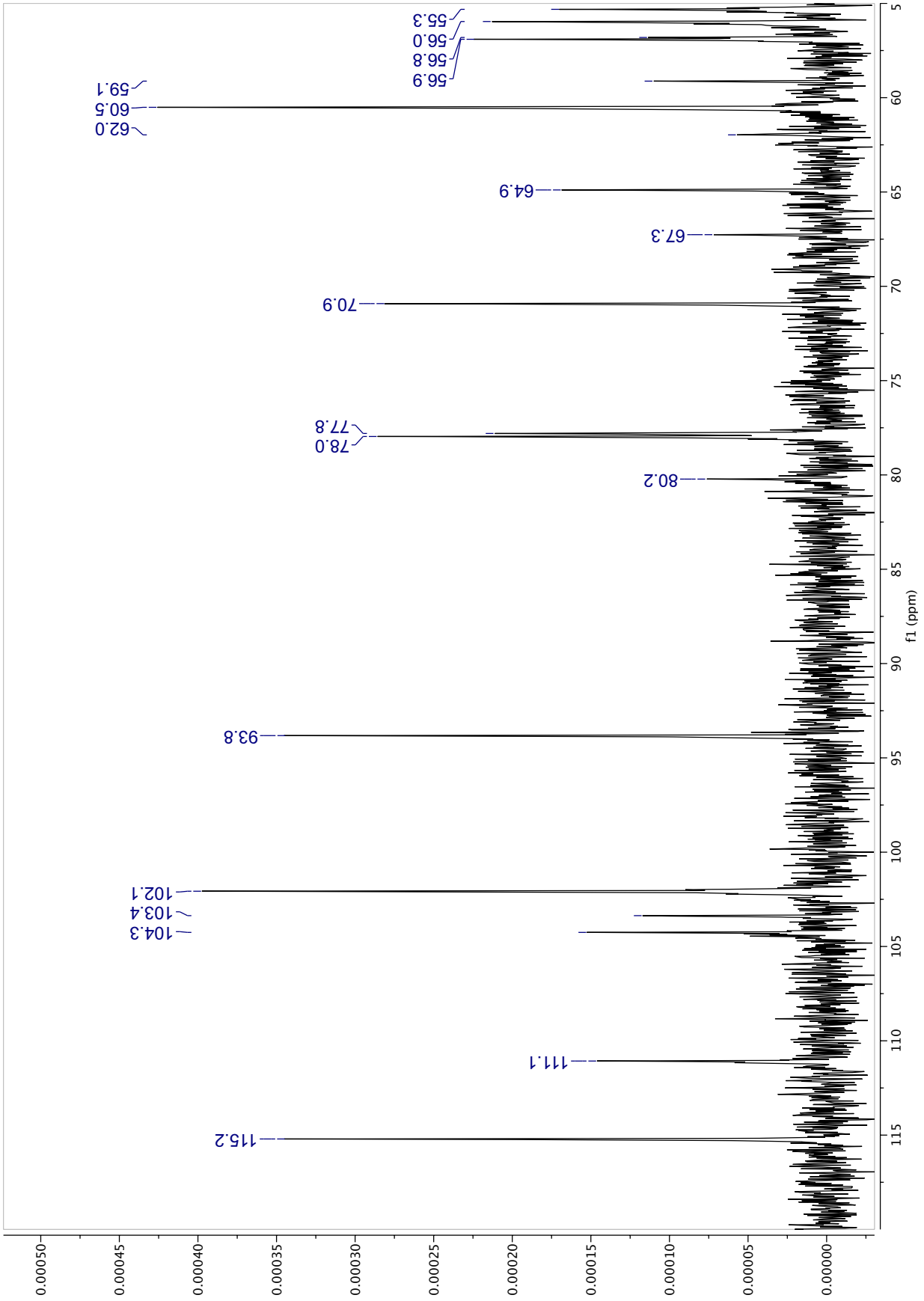


Figure S77. ^{13}C spectrum (Zoom 3: 55-120 ppm) (8000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

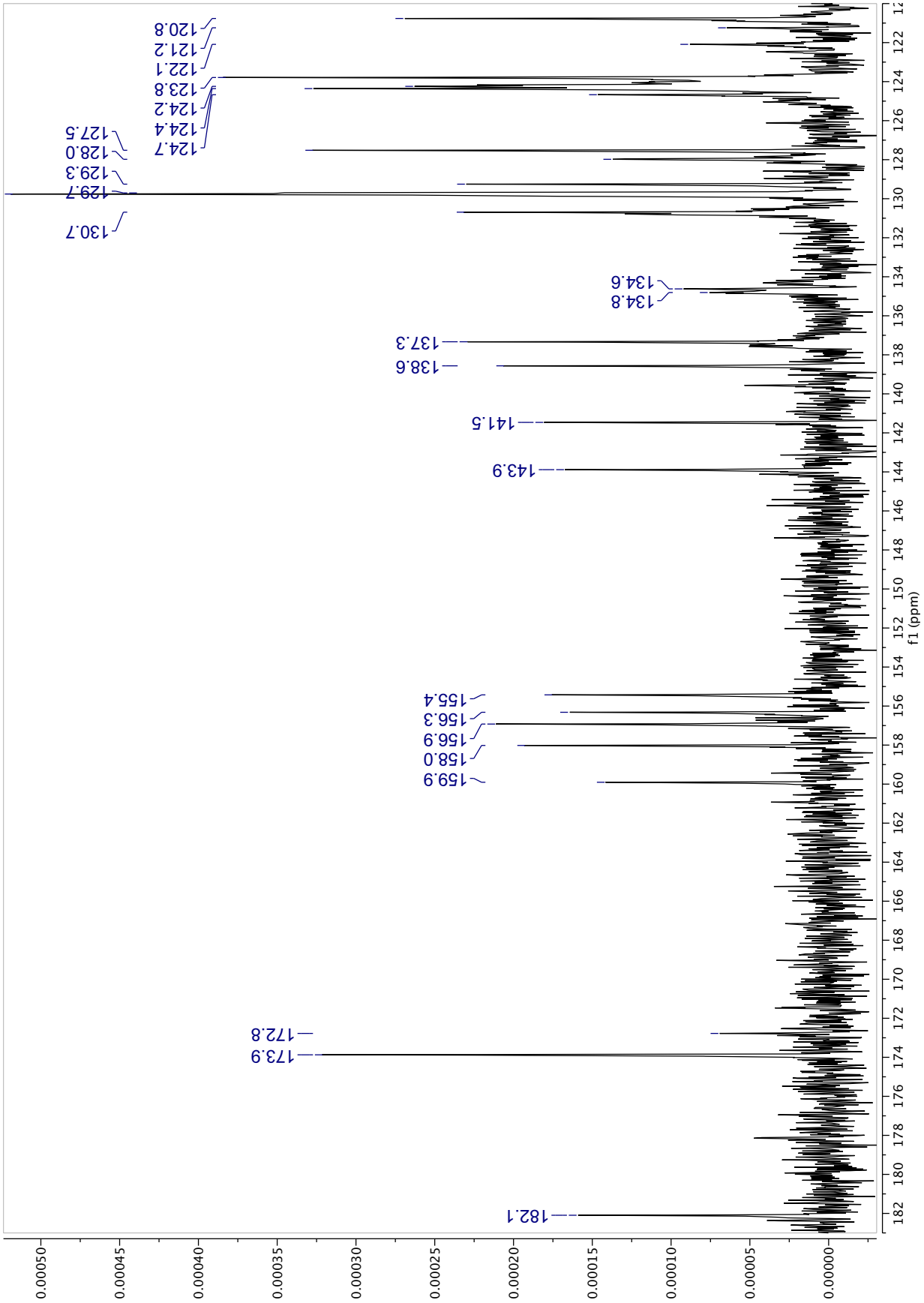


Figure S78. ^{13}C spectrum (Zoom 4: 120-182 ppm) (8000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

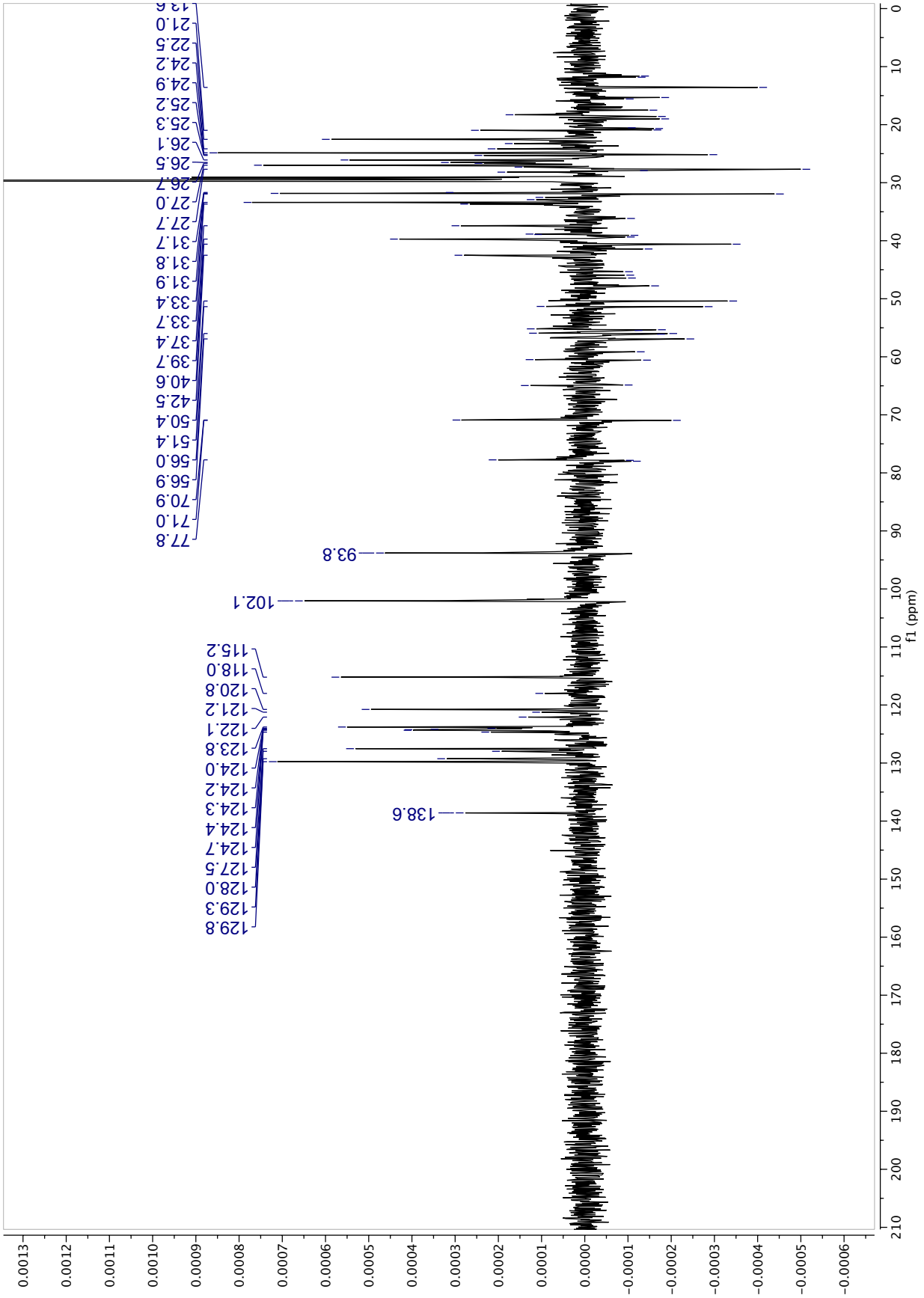


Figure S79. DEPT-135 spectrum (2000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

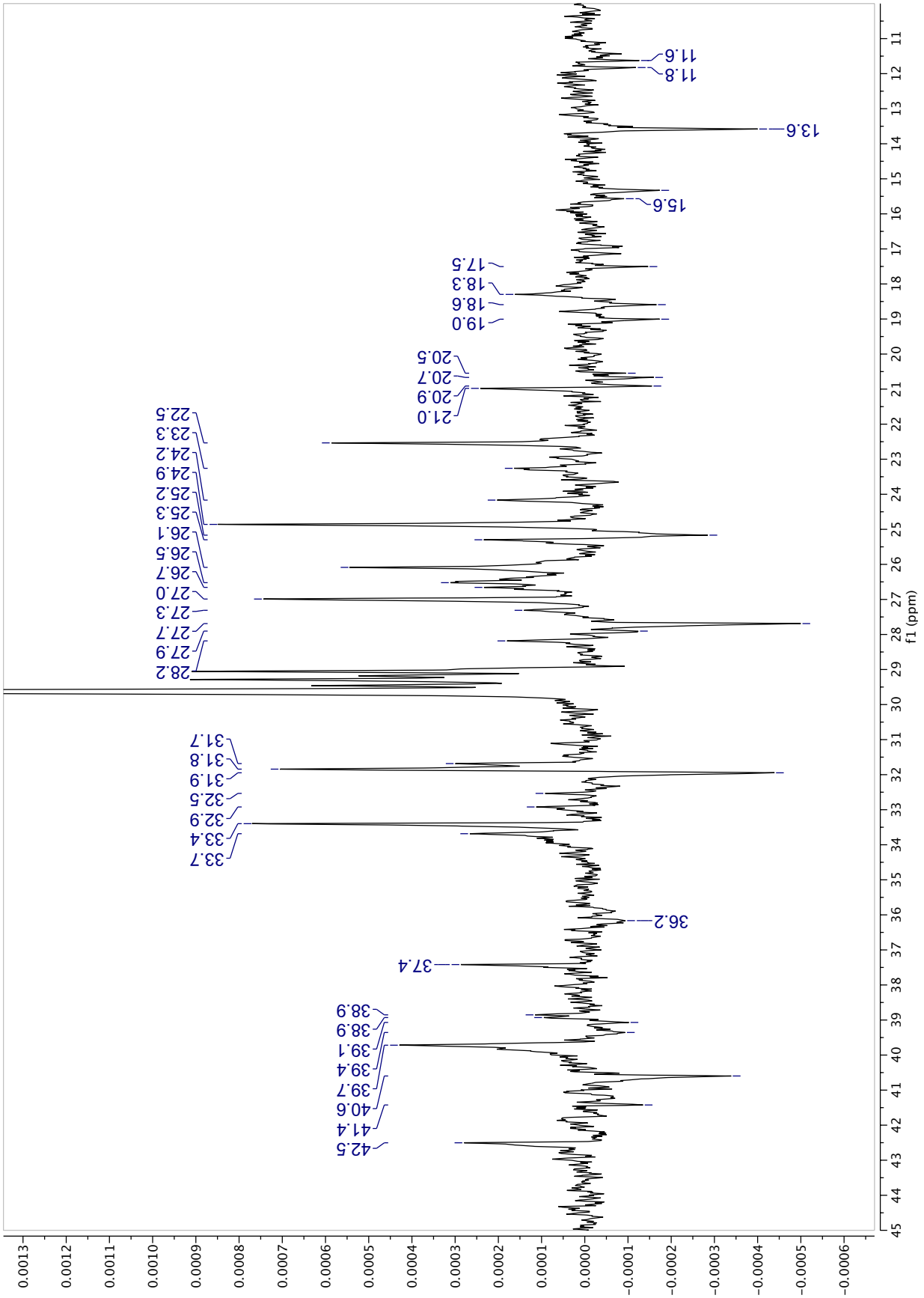


Figure S80. DEPT-135 spectrum (Zoom 1: 10-45 ppm) (2000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

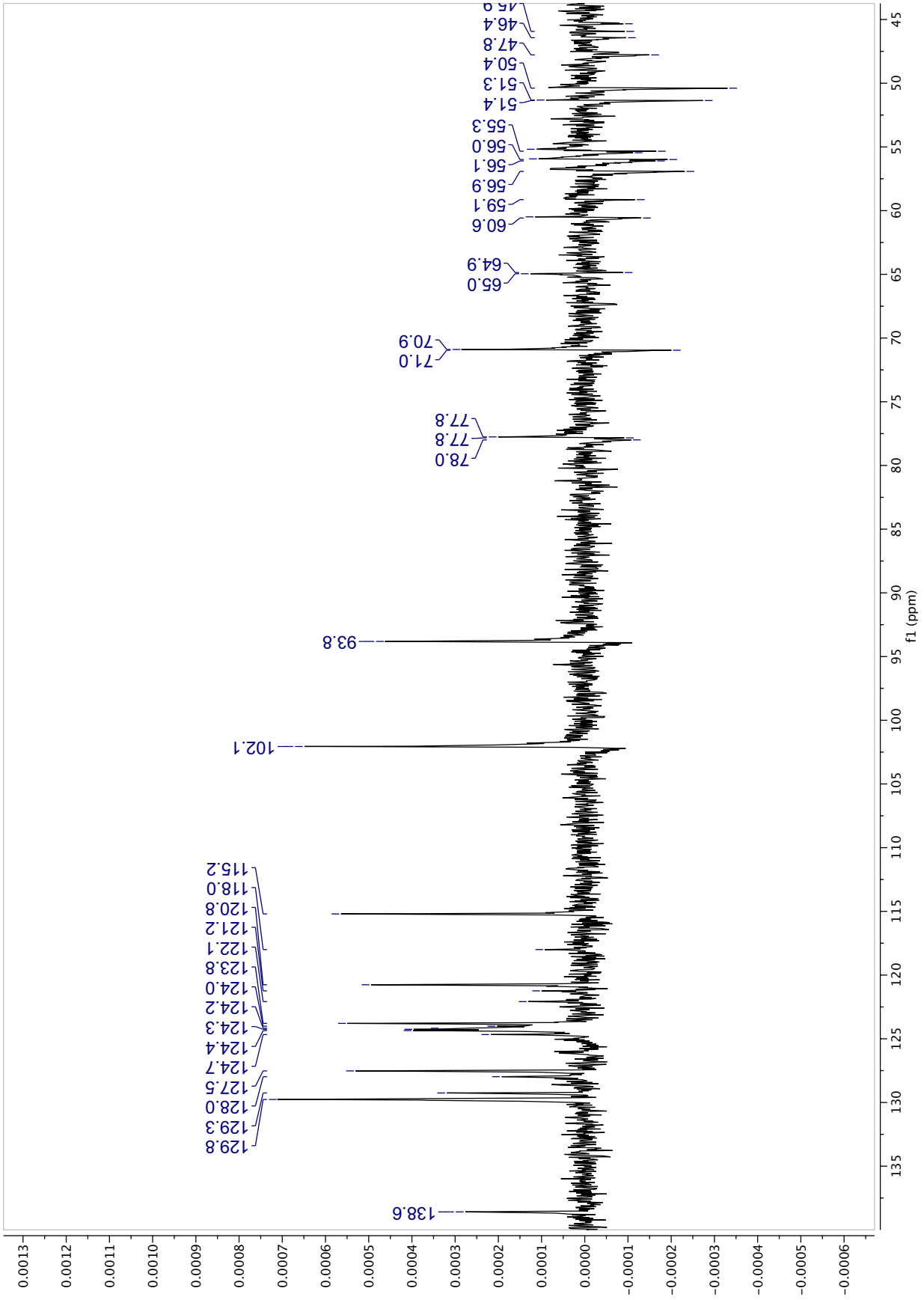


Figure S81. DEPT-135 spectrum (Zoom 2: 45-140 ppm) (2000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

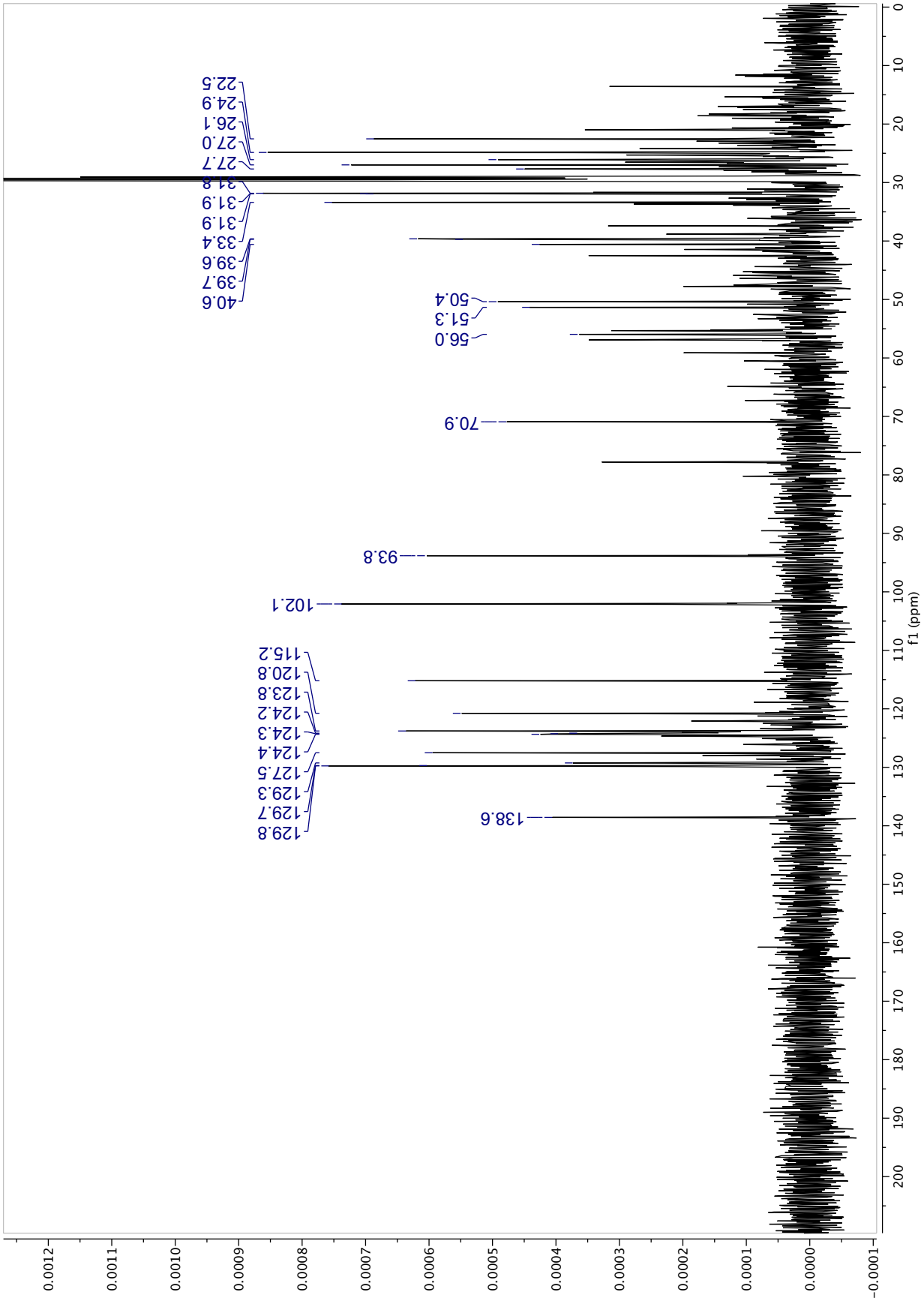


Figure S82. DEPT-90 spectrum (2000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

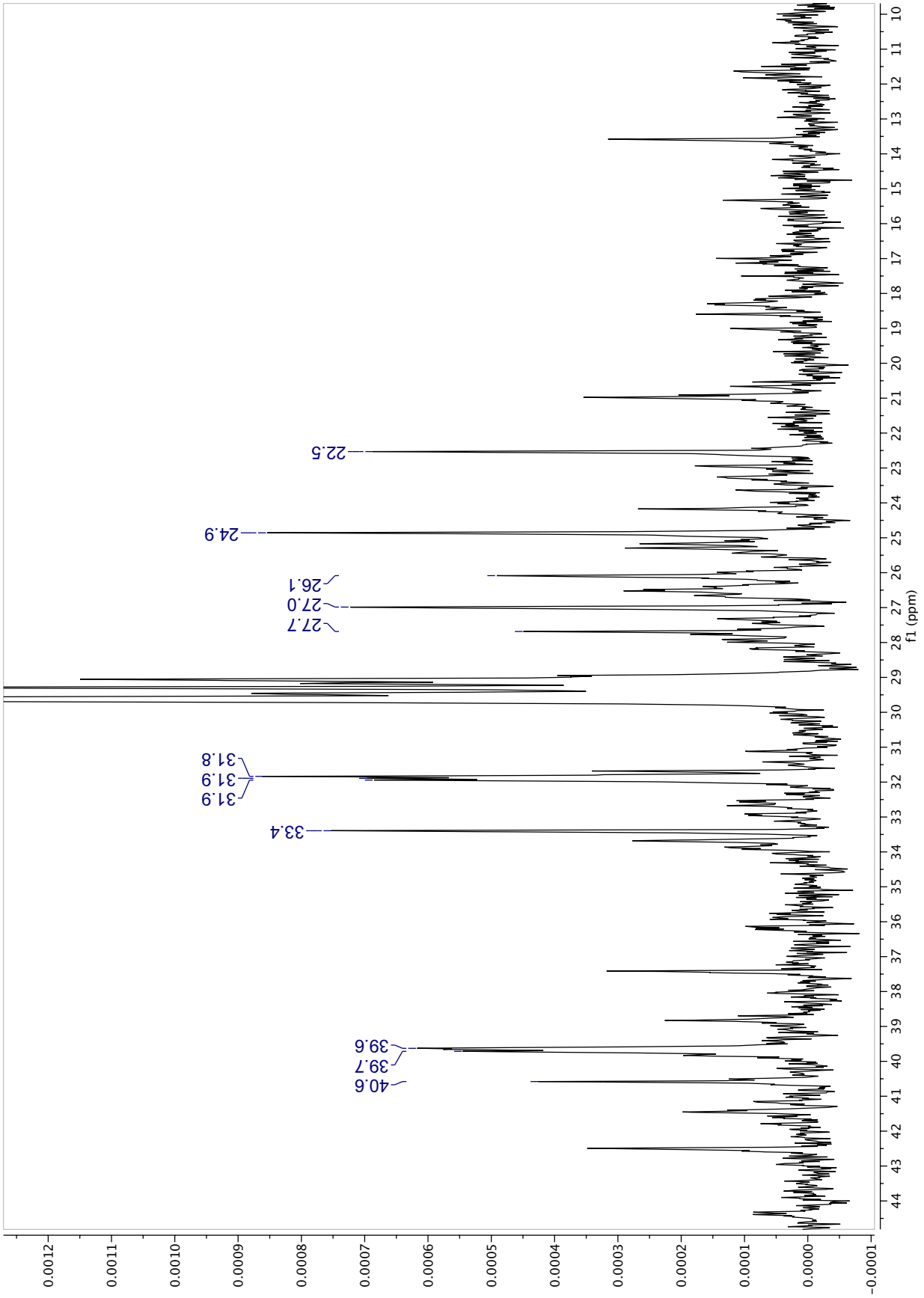


Figure S83. DEPT-90 spectrum (Zoom 1: 10-45 ppm) (2000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

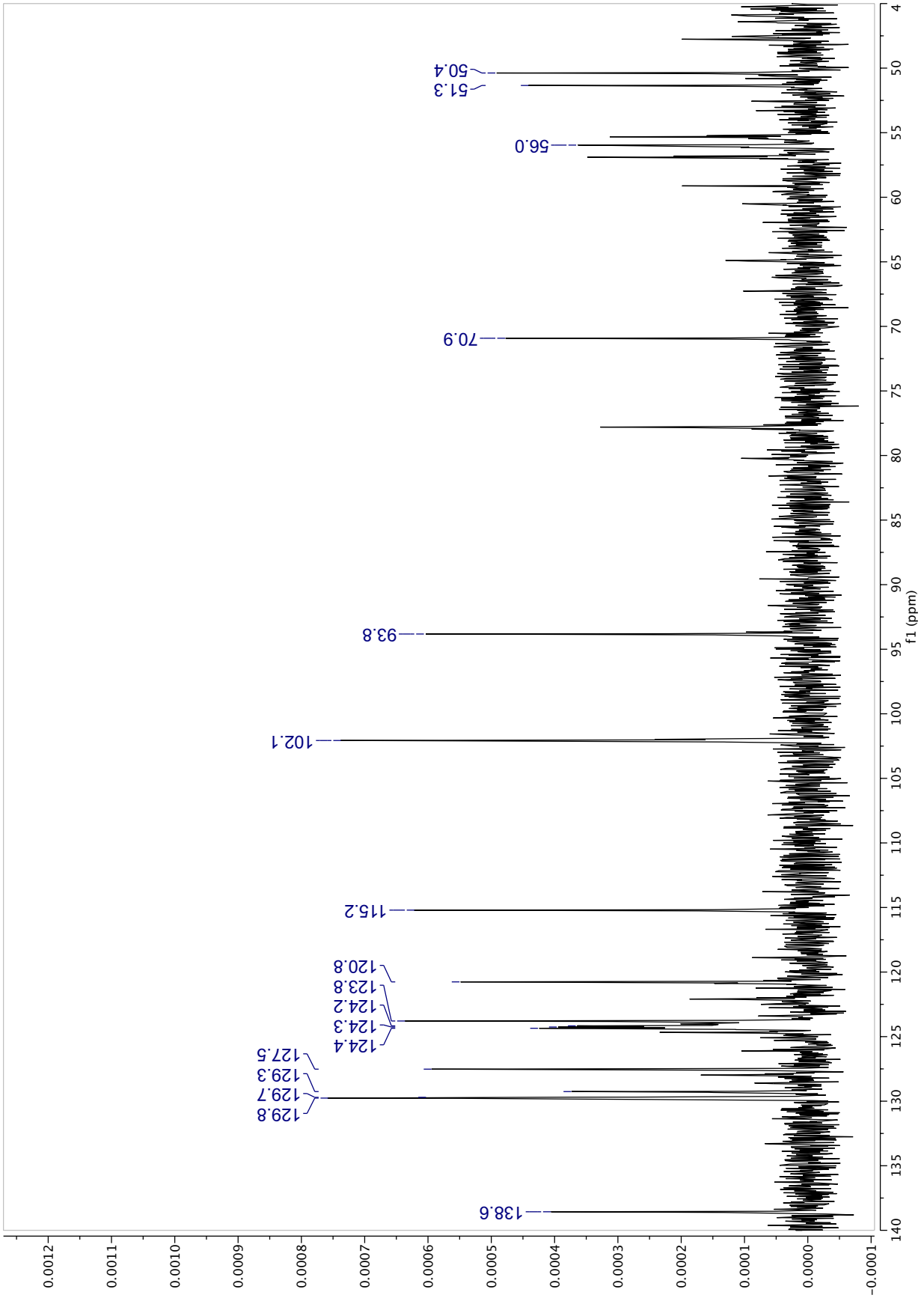


Figure S84. DEPT-90 spectrum (Zoom 1: 45-140 ppm) (2000 scans) of F2 (48.3 mg) recorded in acetone-*d*6.

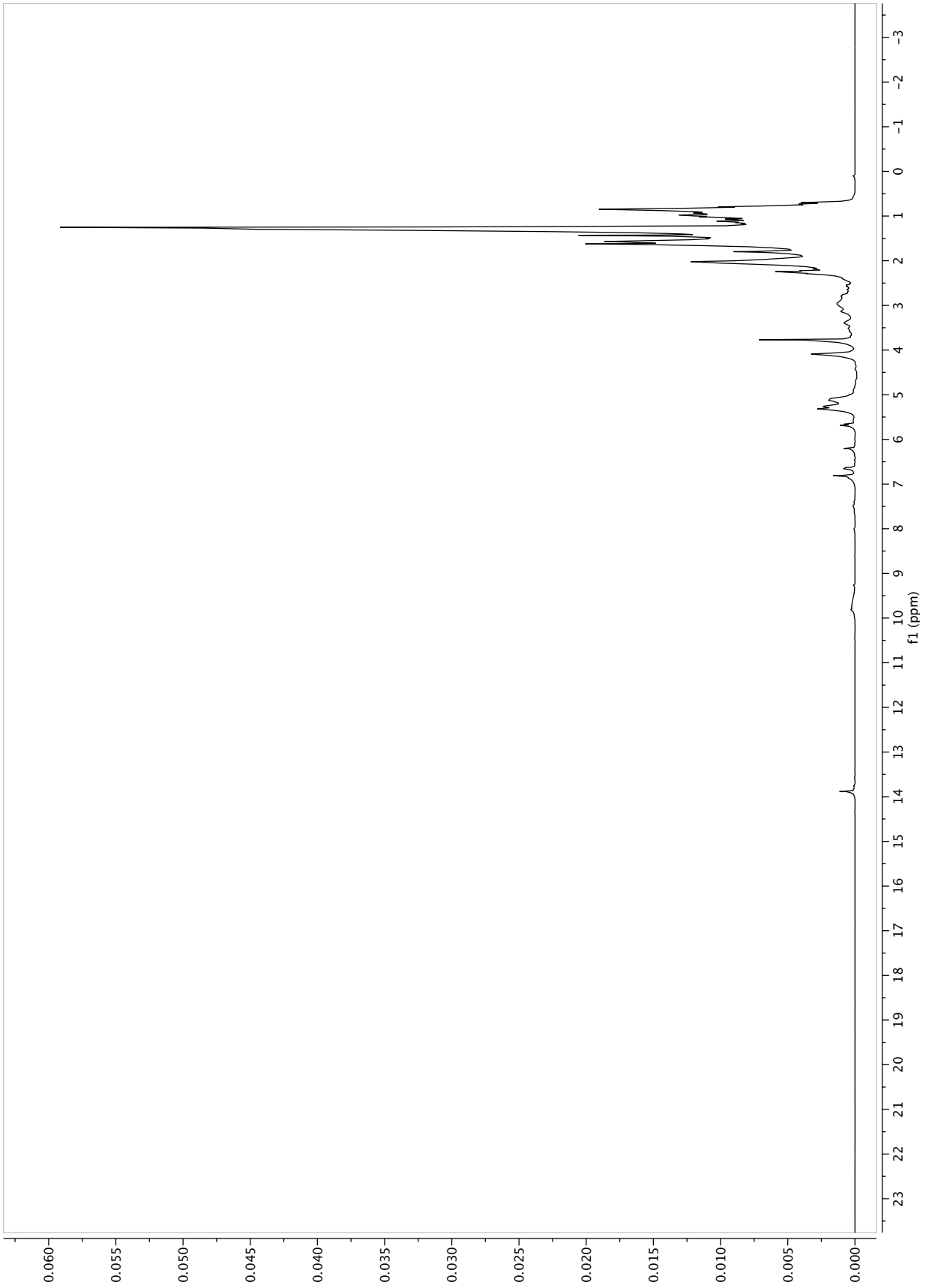


Figure S85. ^{1}H-NMR of F2 (48.3 mg) recorded in acetone-*d*6.

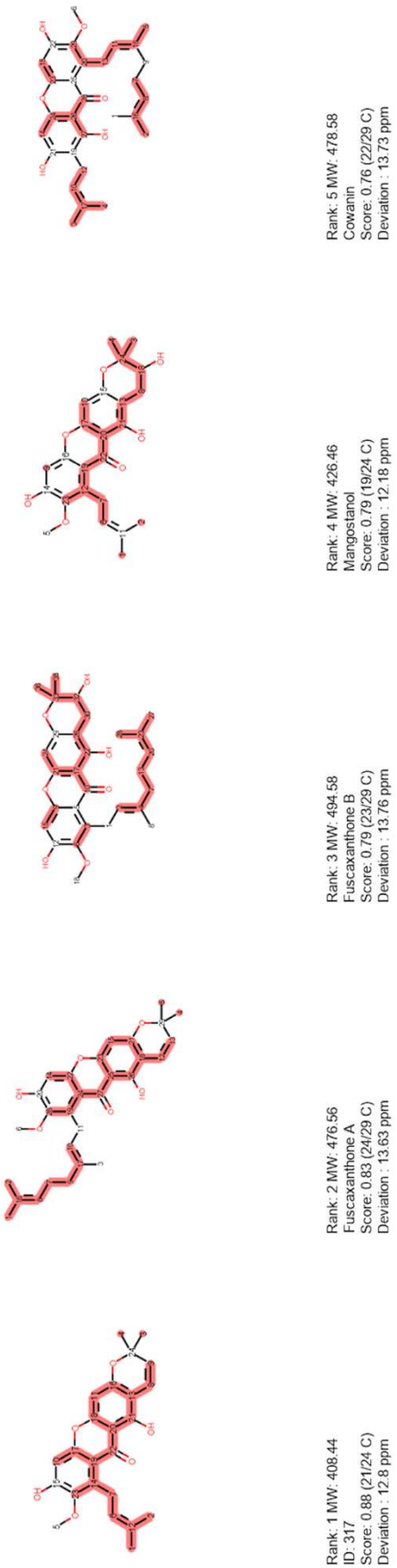


Figure S86. Part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and -135) of F2 using c-type *Garcinia* DB. Equivalent carbons were allowed.

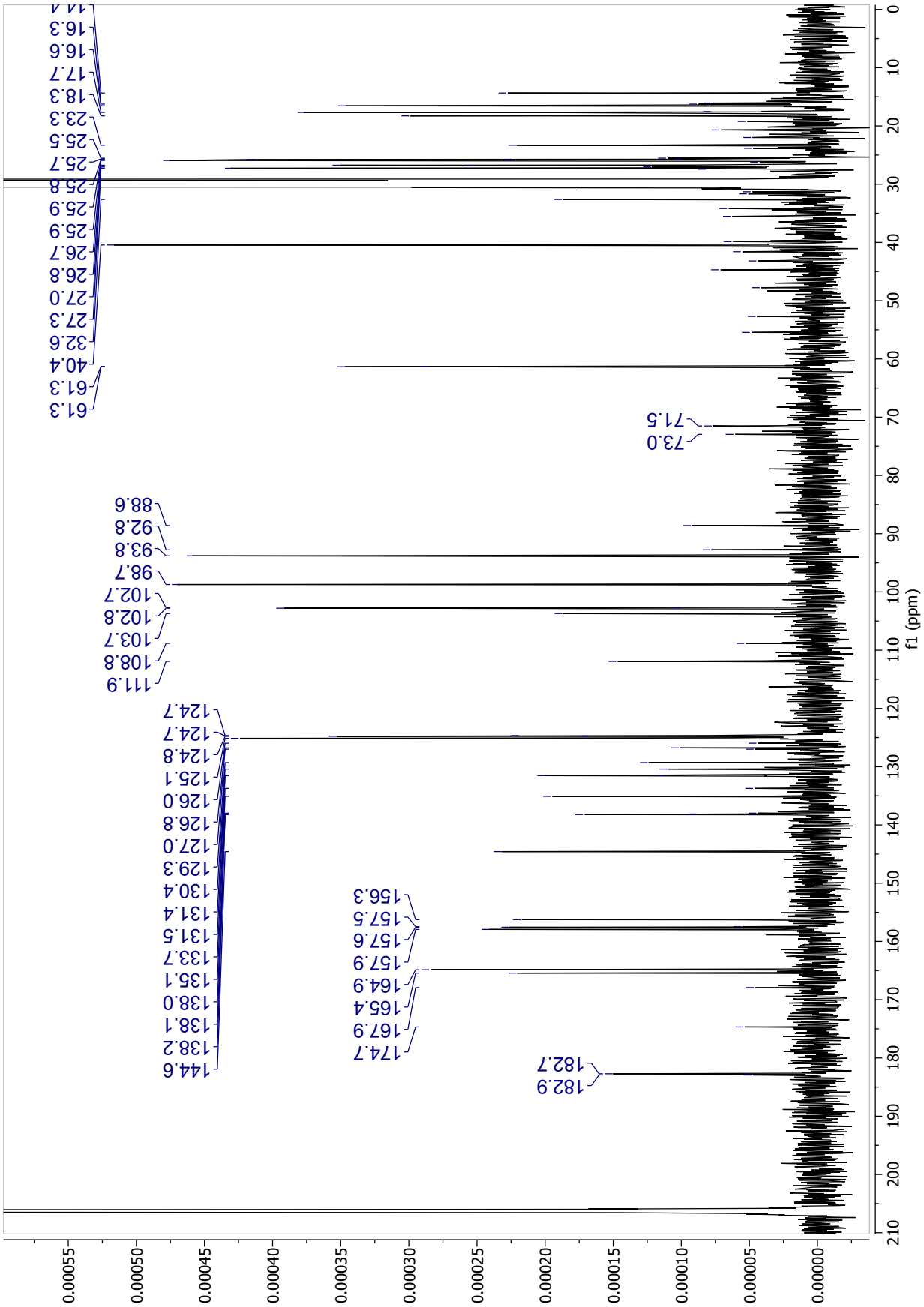


Figure S87. ^{13}C-NMR (8000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

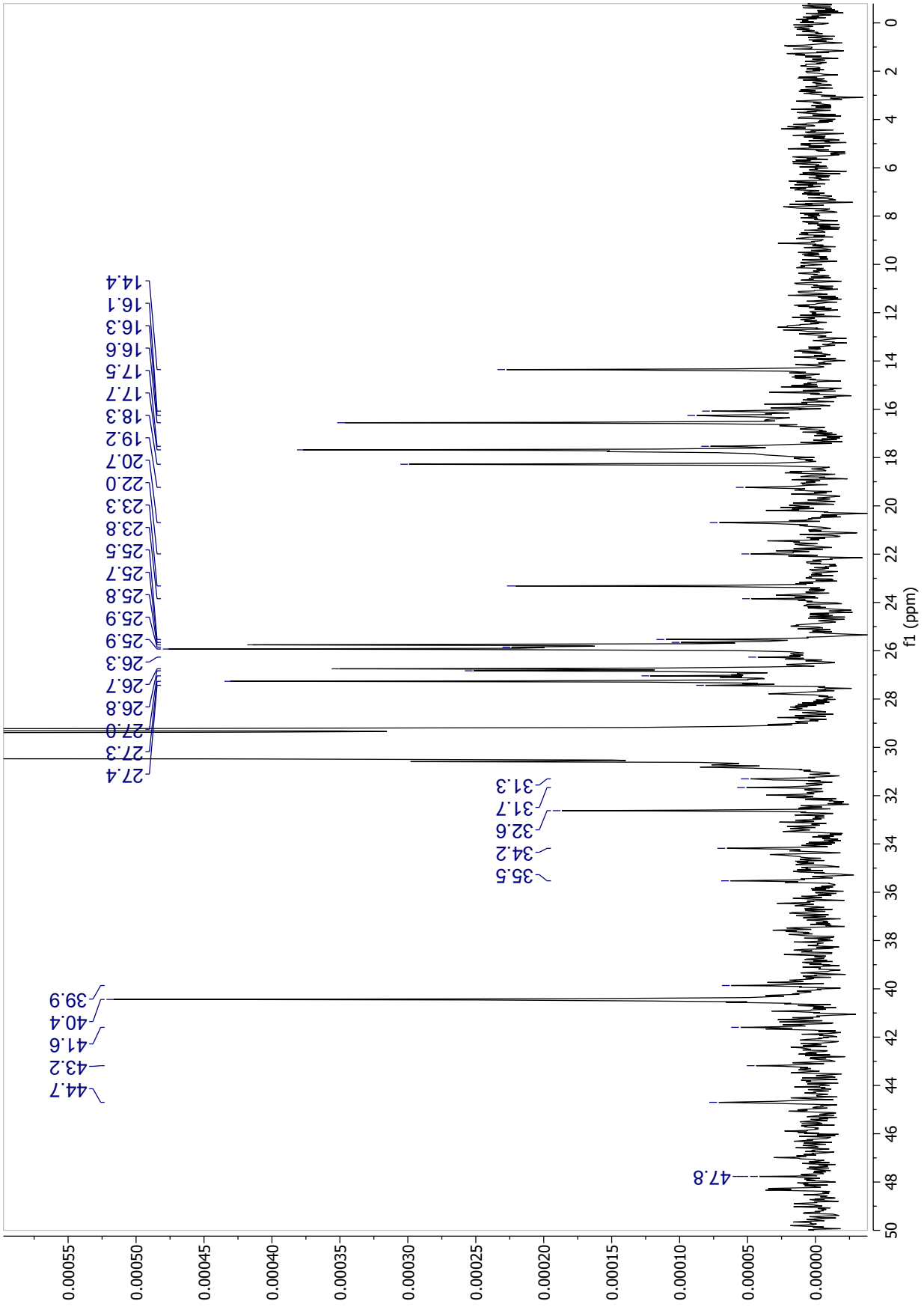


Figure S88. ^{13}C spectrum (Zoom 1: 0-50 ppm) (8000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

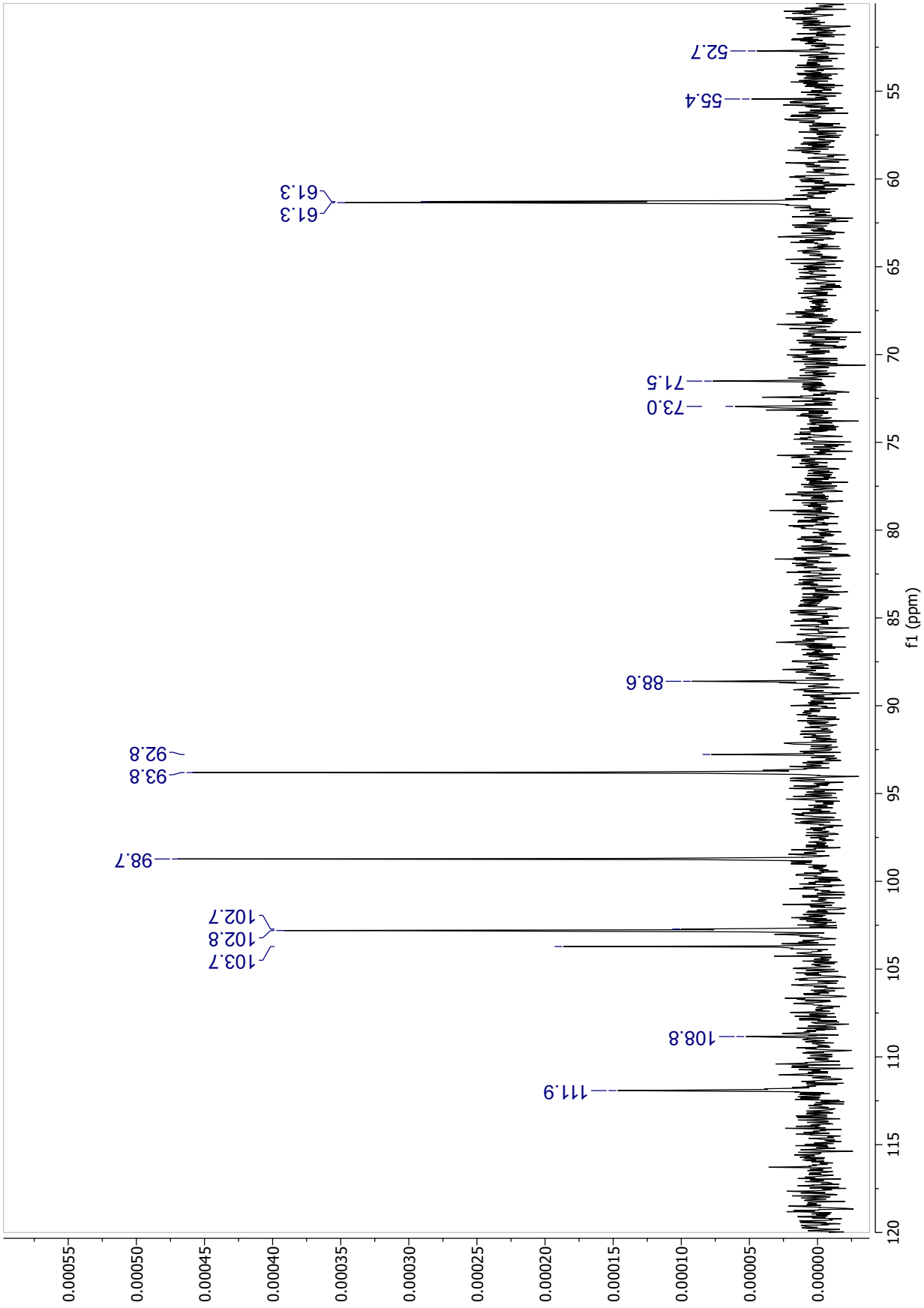


Figure S89. ^{13}C spectrum (Zoom 2: 50-120 ppm) (8000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

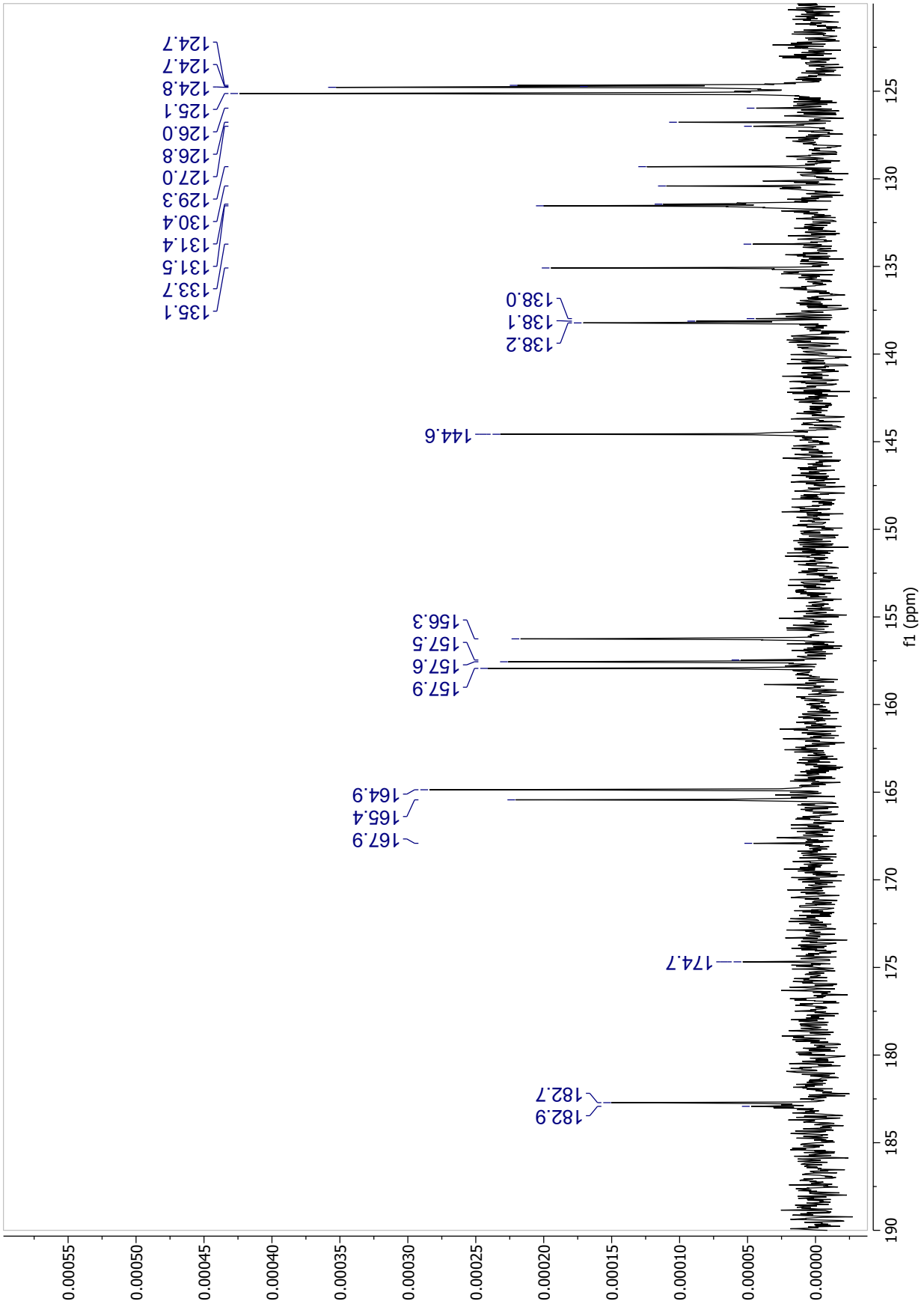


Figure S90. ^{13}C spectrum (Zoom 3: 120-190 ppm) (8000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

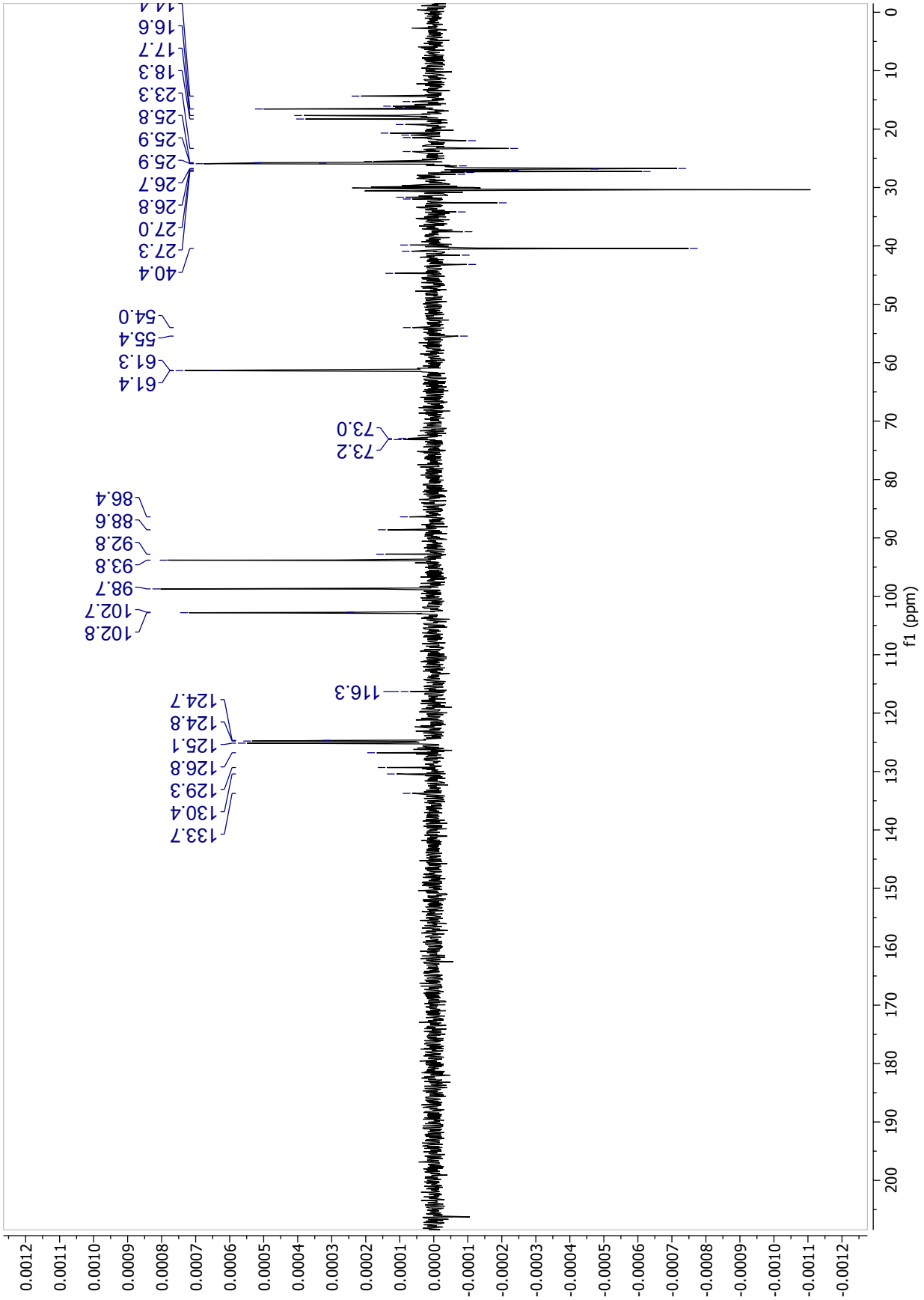


Figure S91. DEPT-135 spectrum (3000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

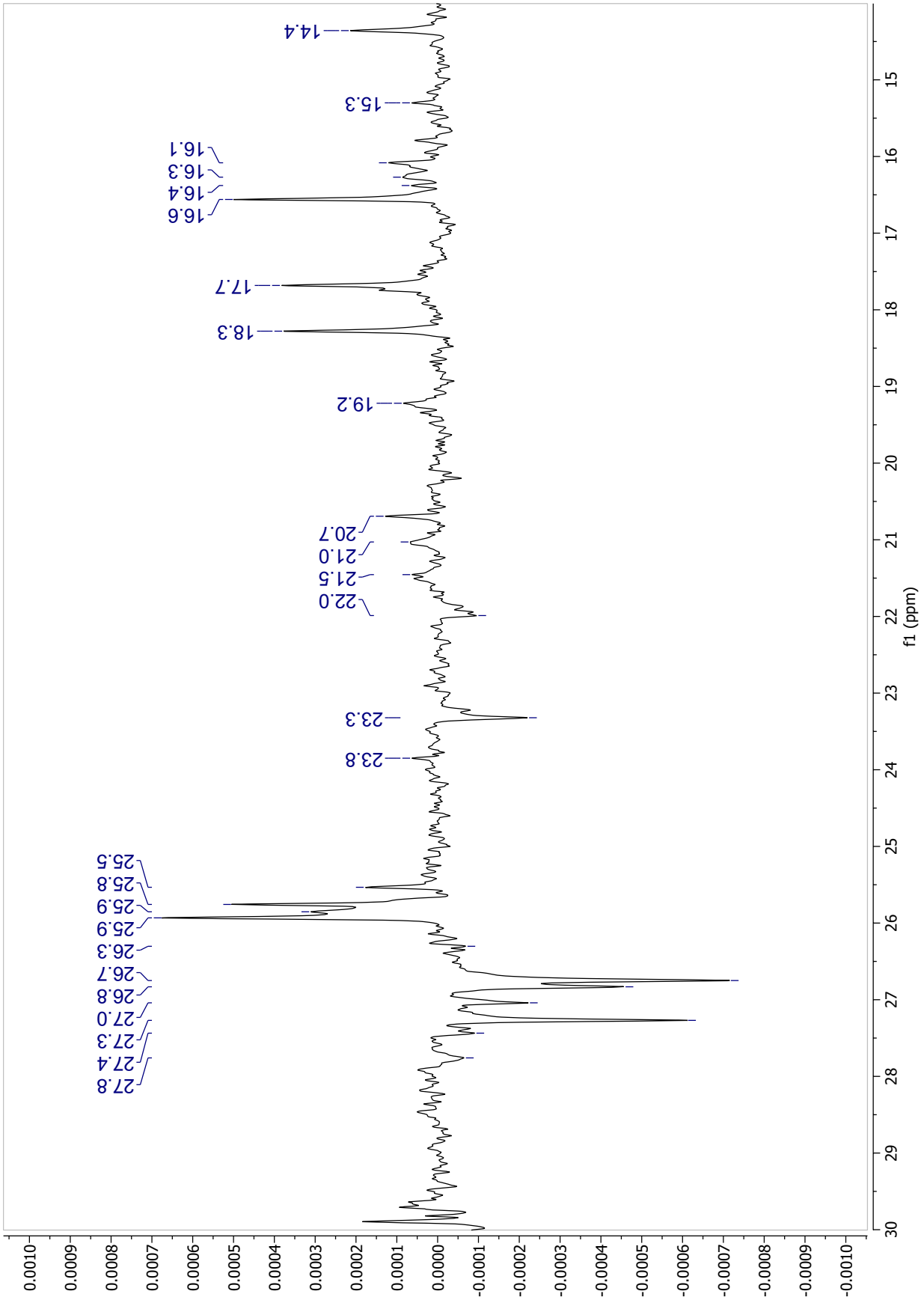


Figure S92. DEPT-135 (Zoom 1: 14-30 ppm) spectrum (3000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

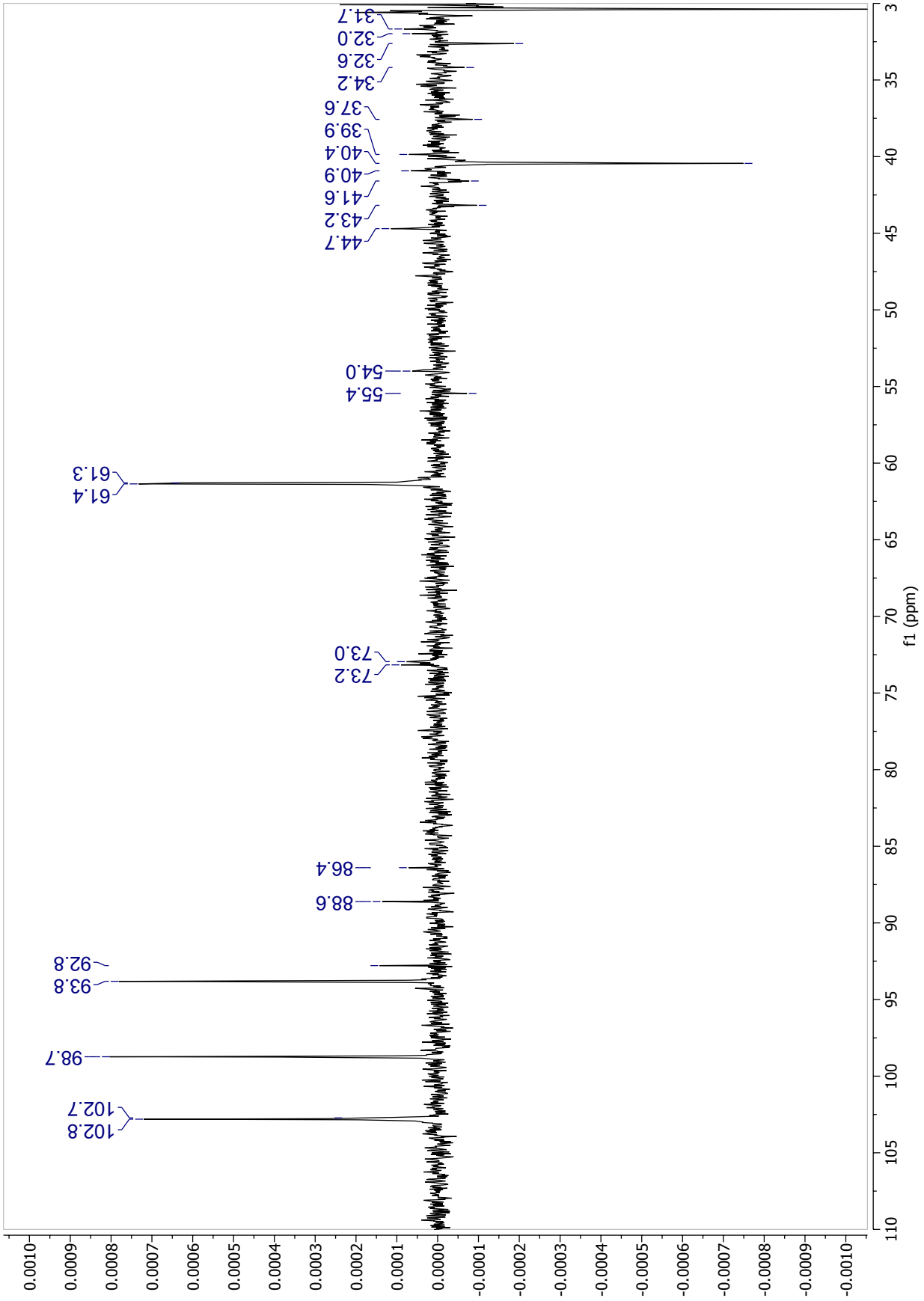


Figure S93. DEPT-135 (Zoom 2: 30-110 ppm) spectrum (3000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

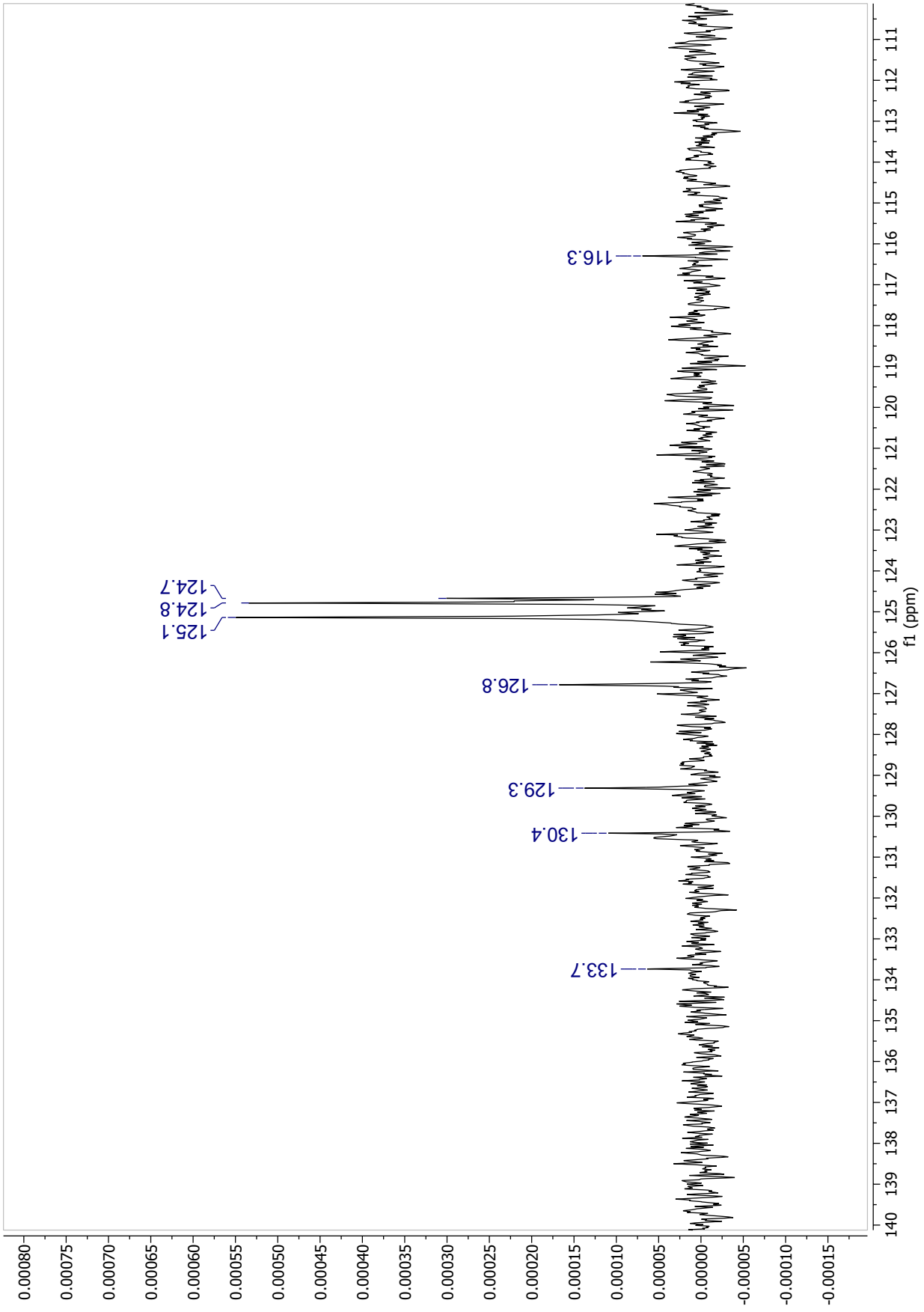


Figure S94. DEPT-135 (Zoom 3: 110-140 ppm) spectrum (3000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

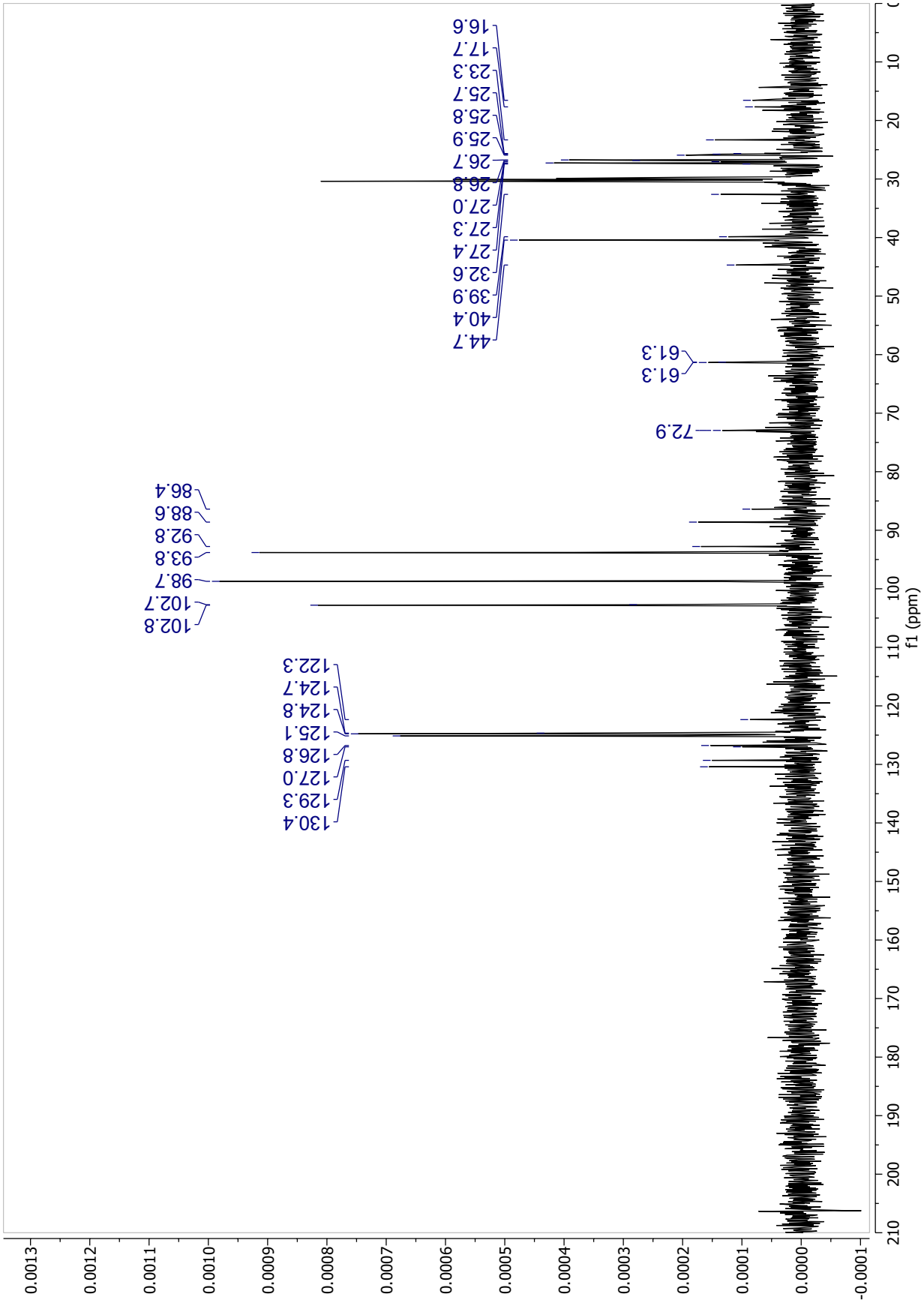


Figure S95. DEPT-90 spectrum (3000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

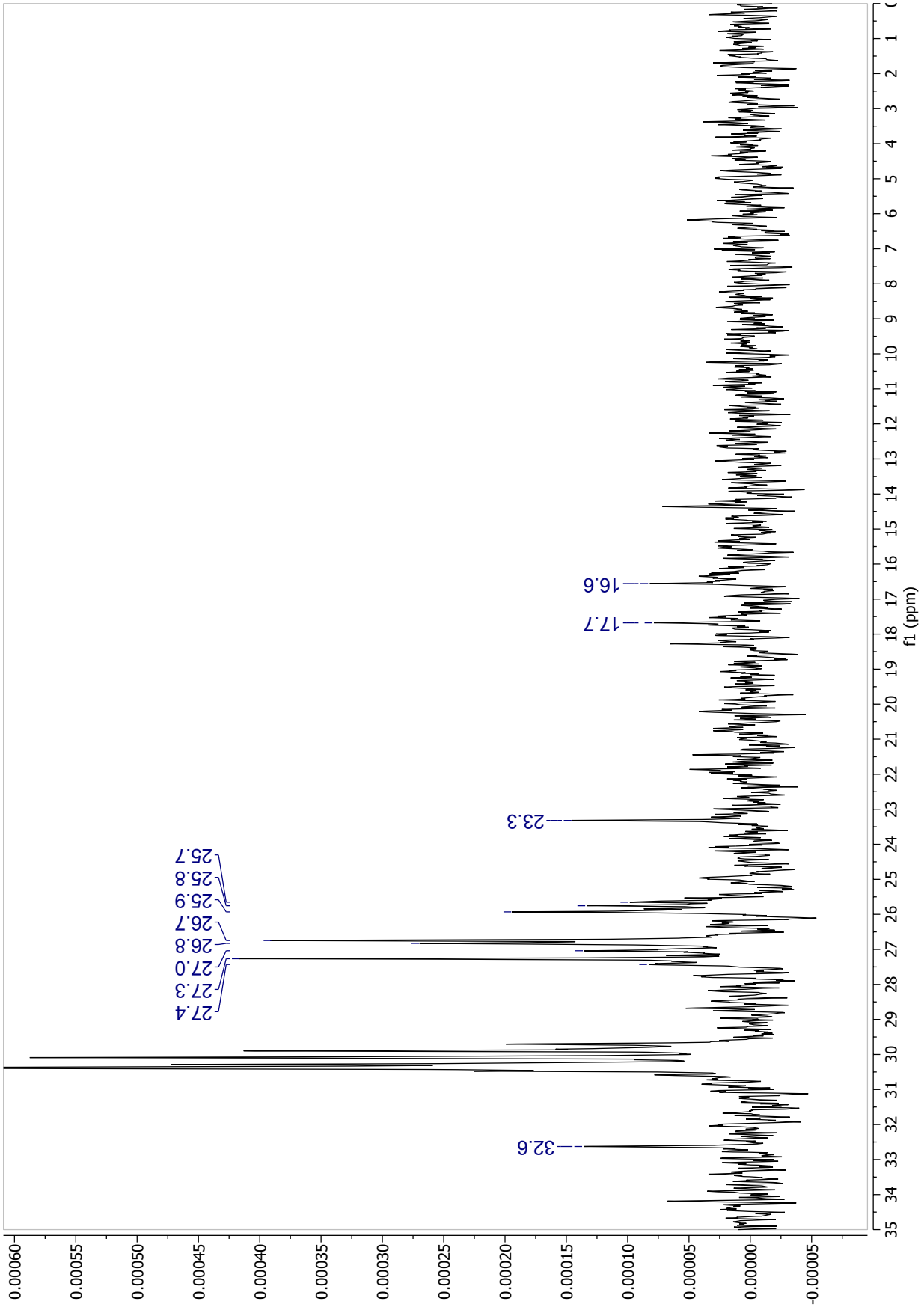


Figure S96. DEPT-90 (Zoom 1: 0-35 ppm) spectrum (3000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

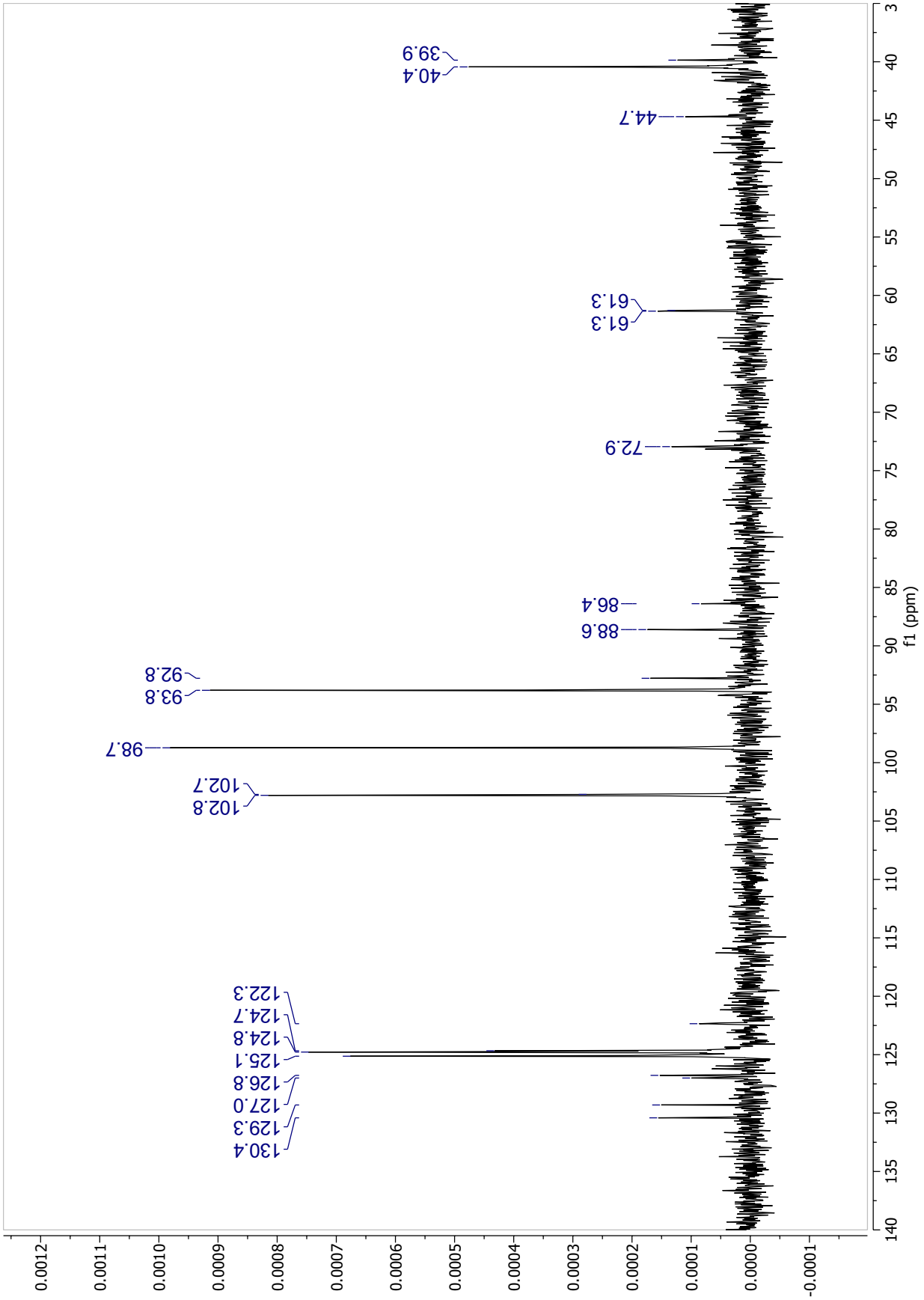


Figure S97. DEPT-90 (Zoom 2: 30-140 ppm) spectrum (3000 scans) of F7 (42.0 mg) recorded in acetone-*d*6.

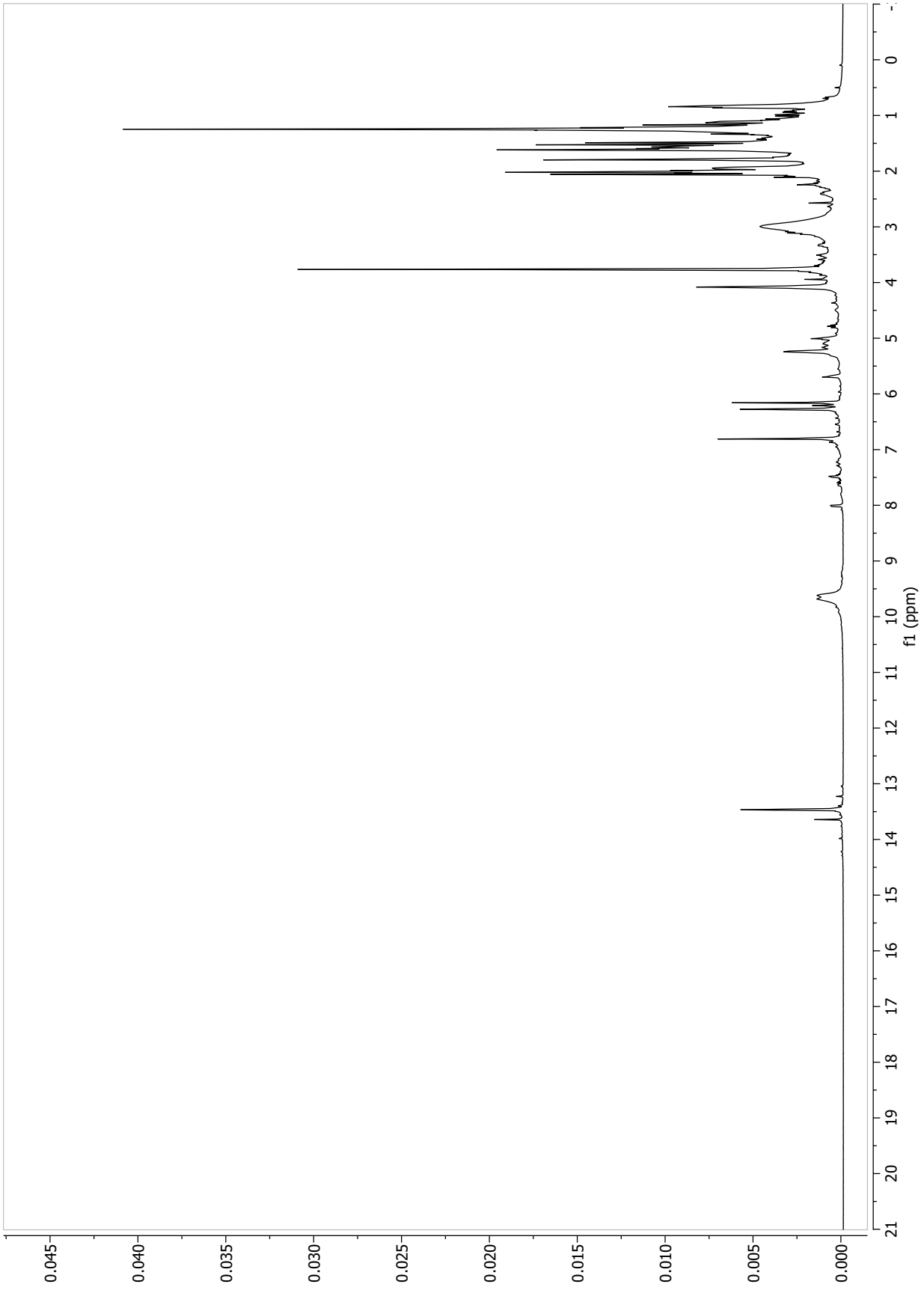


Figure S98. ^{1}H spectrum of F7 (42.0 mg) recorded in acetone-*d*6

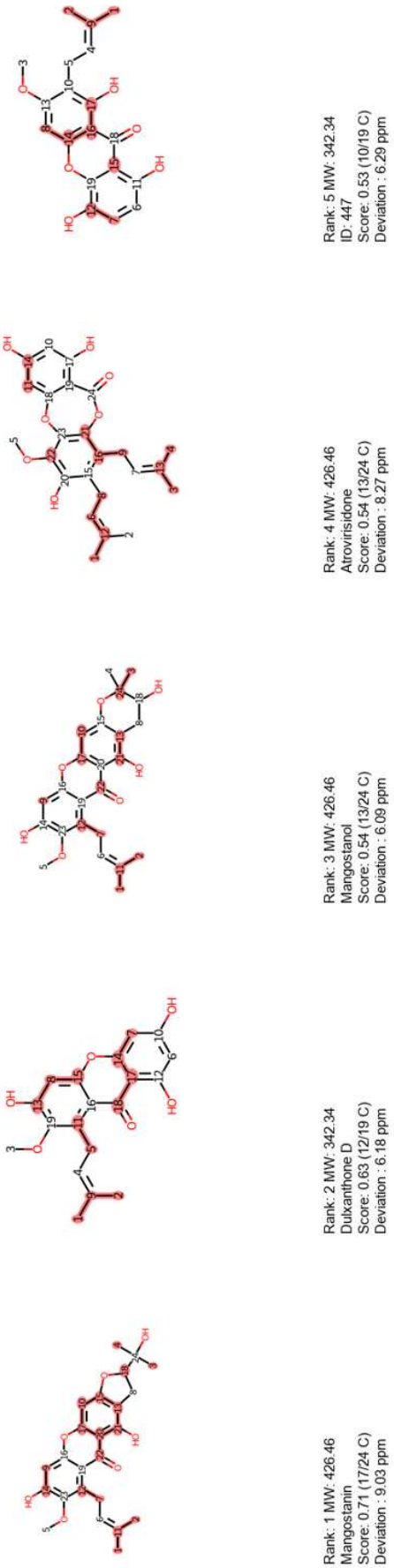


Figure S99. Part of displayed results for the ^{13}C-NMR dereplication (+ DEPT 90 and 135) of F7 using c-type *Garcinia* DB. Only NPs with MW 342 and 426 g.mol^{-1} were allowed. Equivalent carbons were allowed.

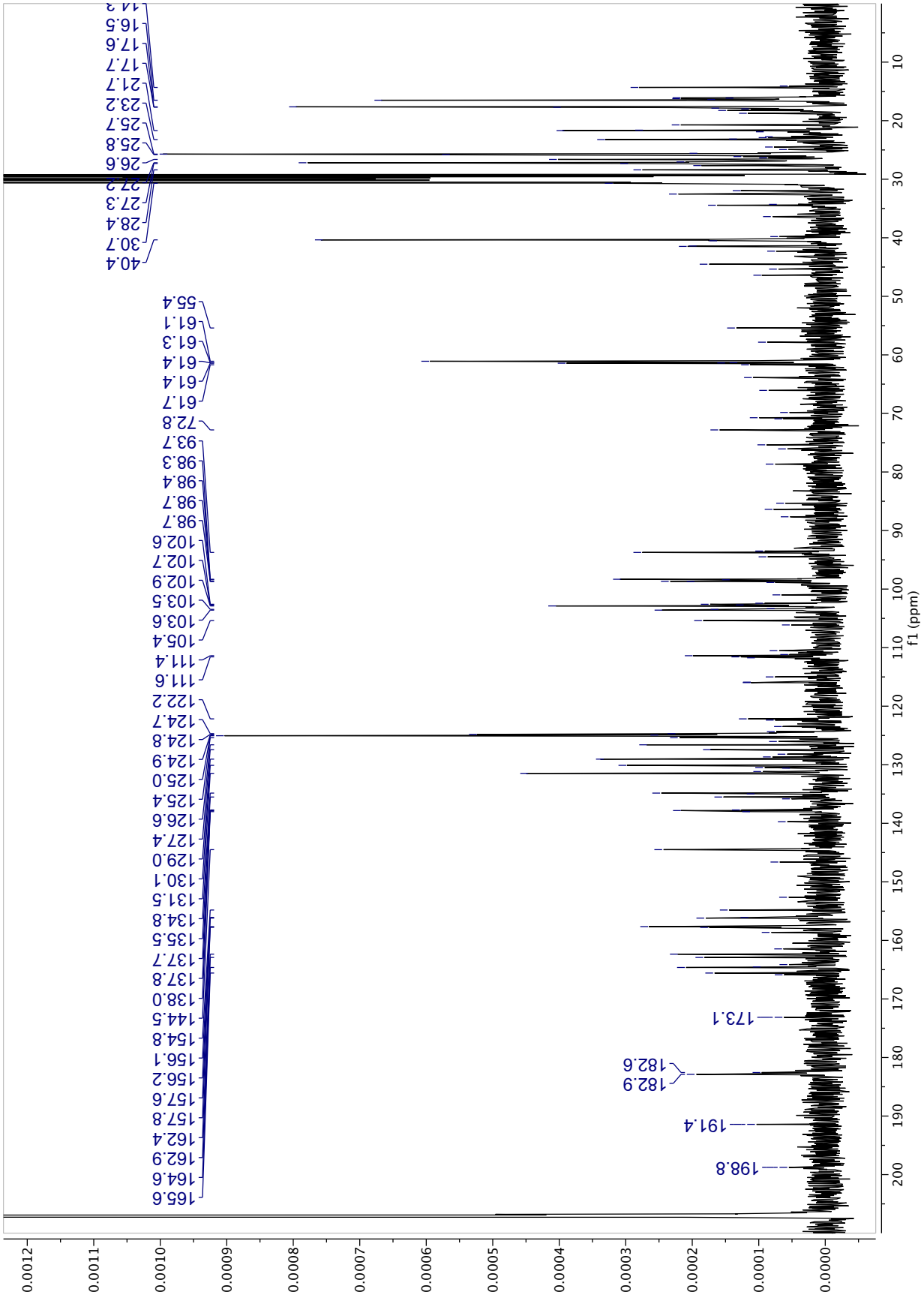


Figure S100. ^{13}C spectrum (2000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

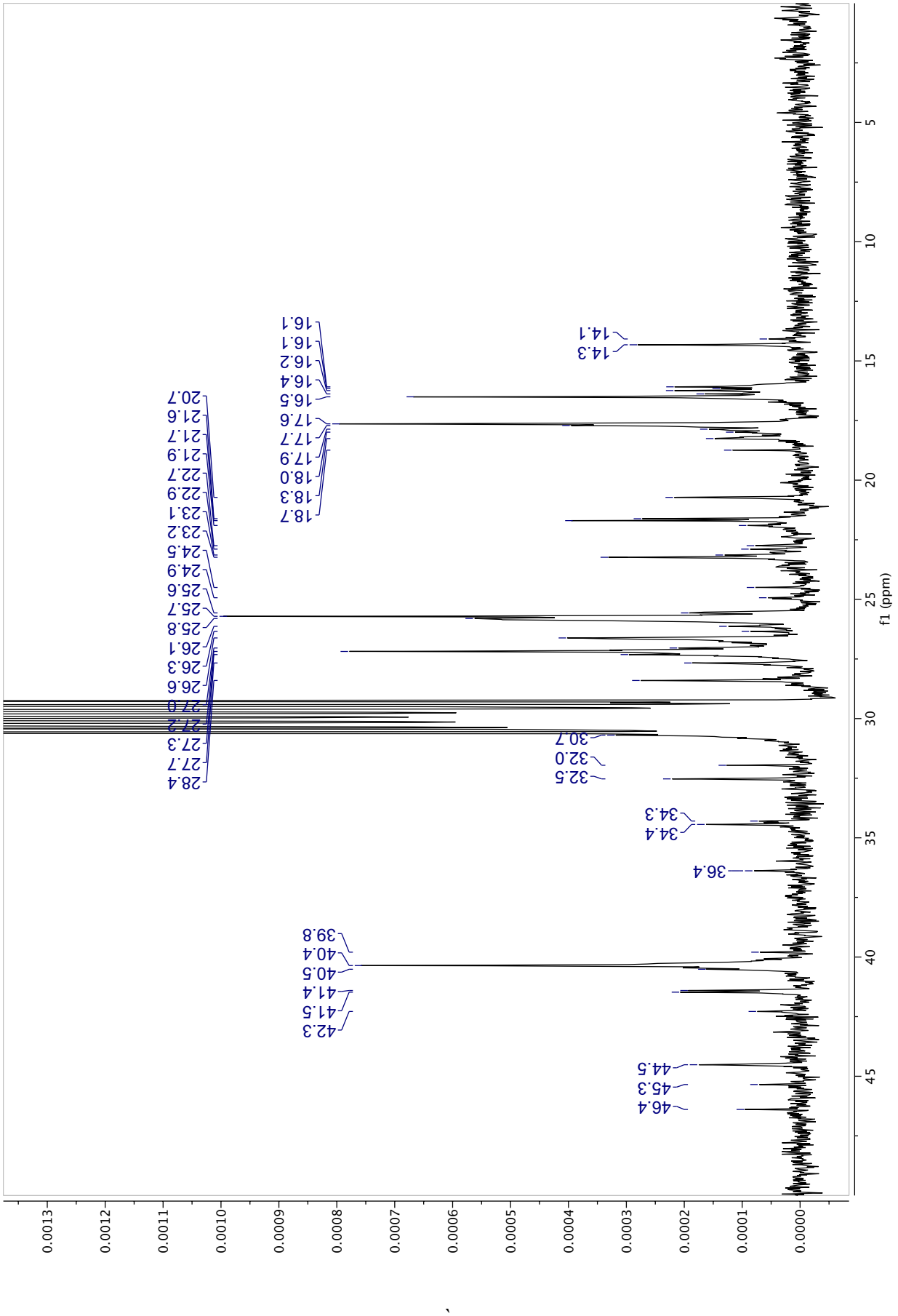


Figure S101. ^{13}C spectrum (Zoom 1: 0-50 ppm) (2000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

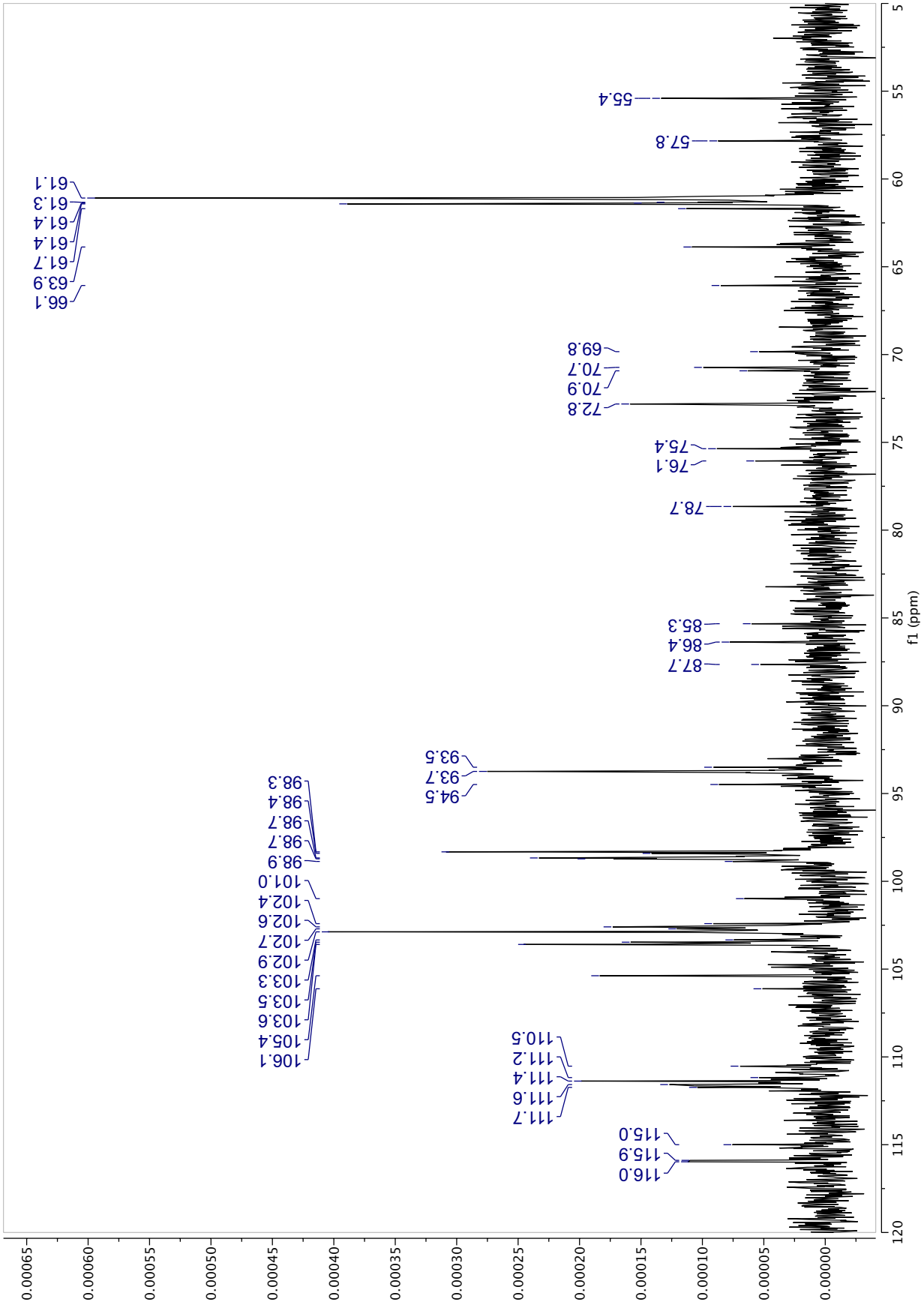


Figure S102. ^{13}C spectrum (Zoom 2: 50-120 ppm) (2000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

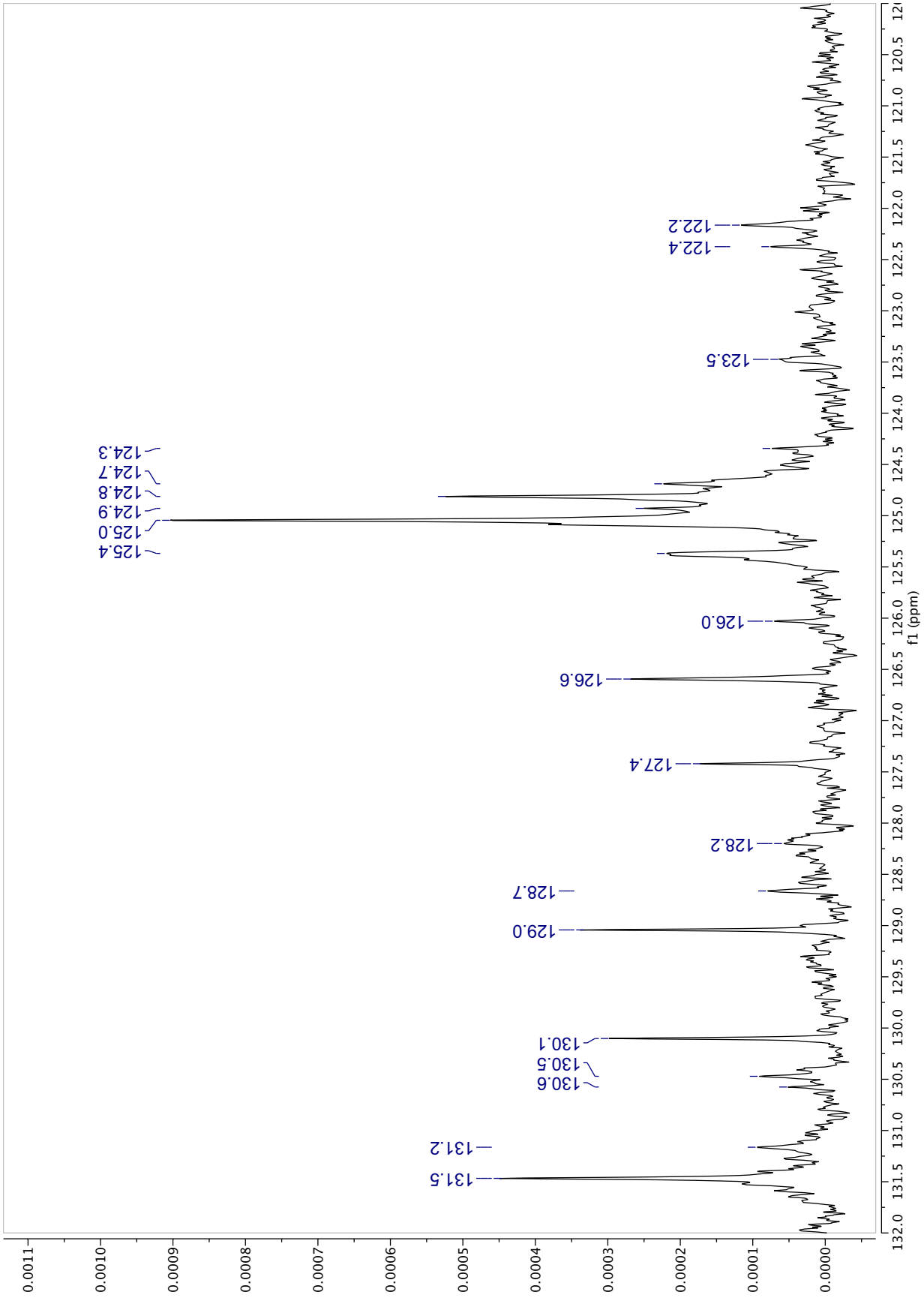


Figure S103. ^{13}C spectrum (Zoom 3: 120-132 ppm) (2000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

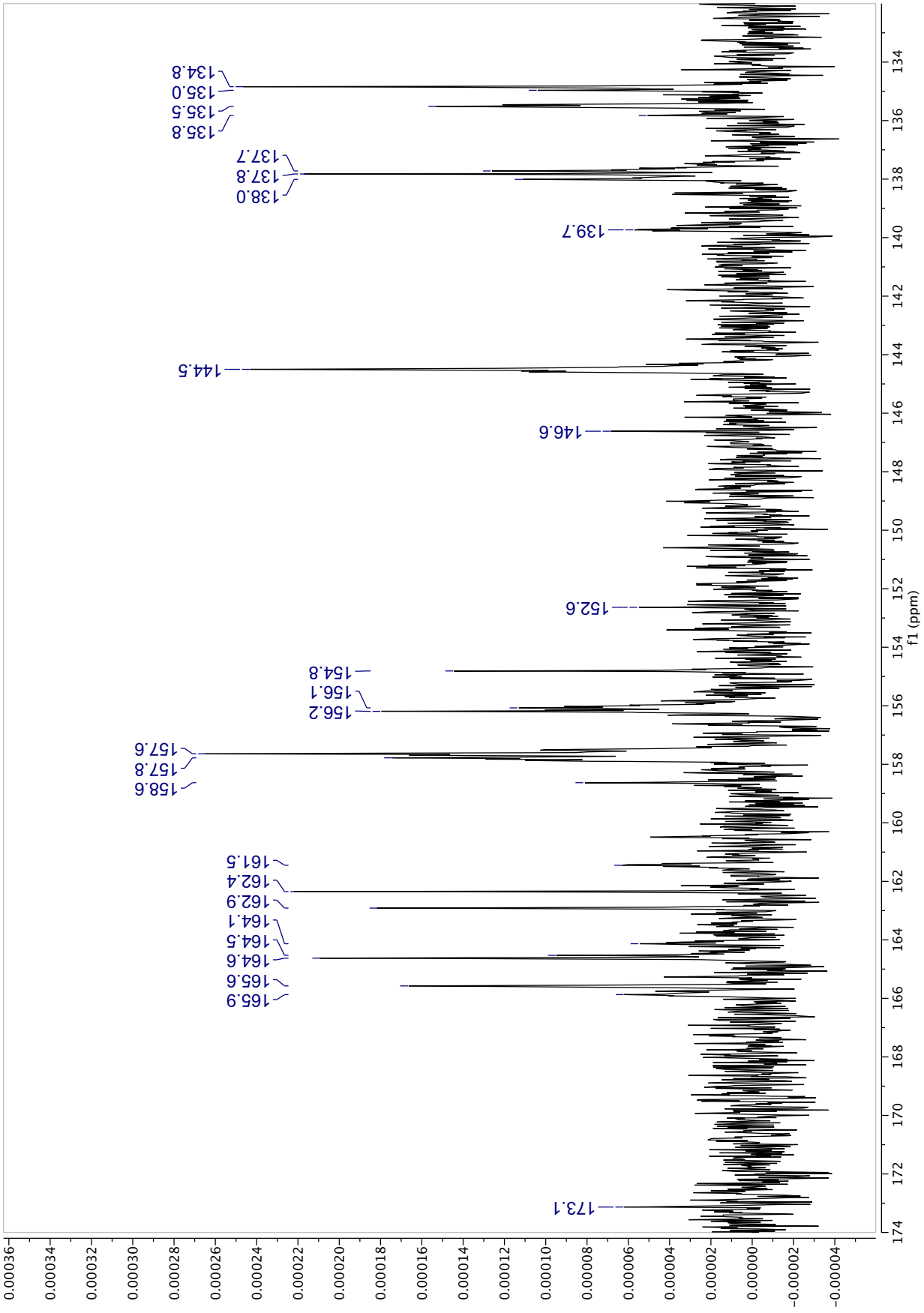


Figure S104. ^{13}C spectrum (Zoom 4: 135-174 ppm) (2000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

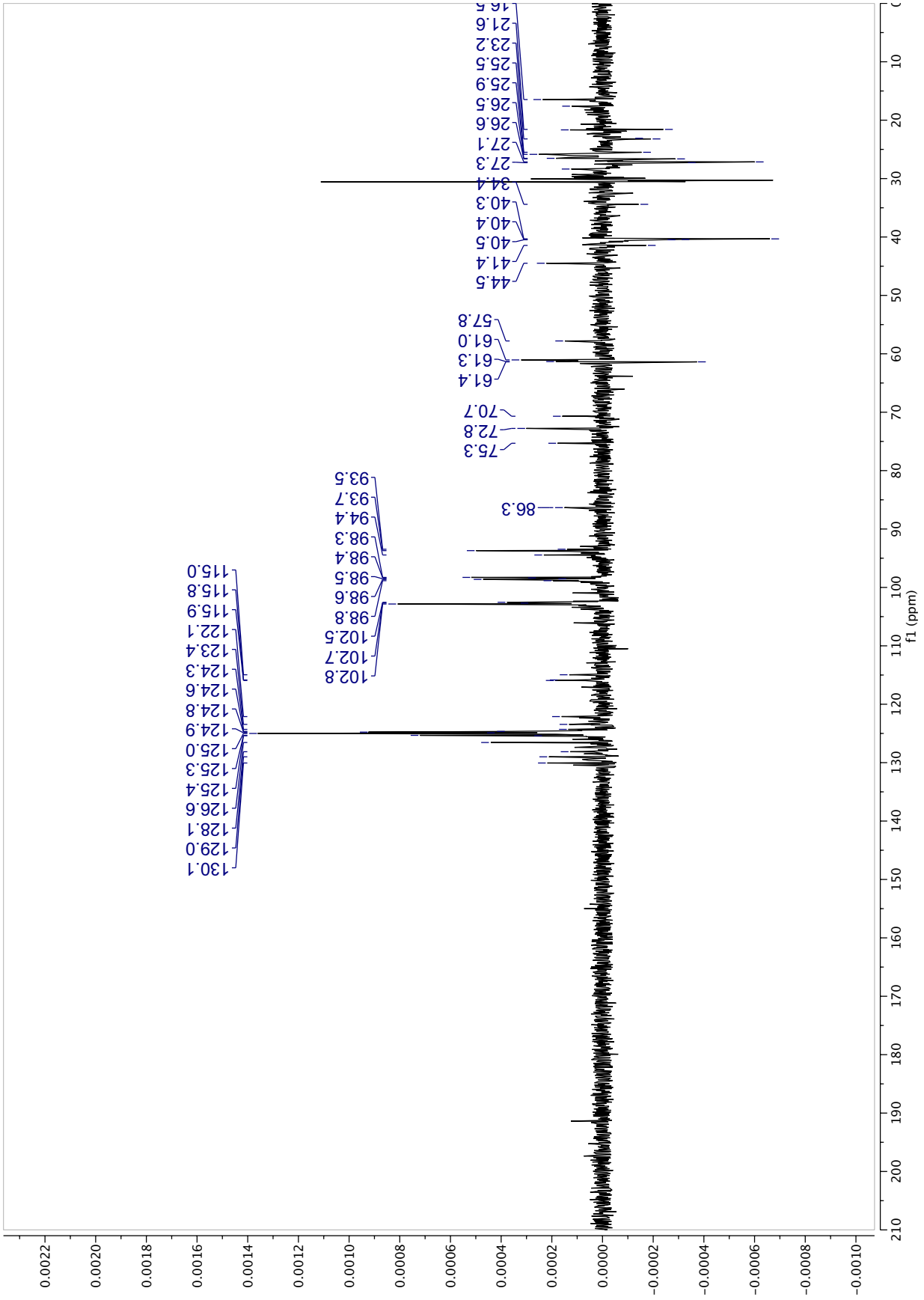


Figure S105. DEPT-135 spectrum (1000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

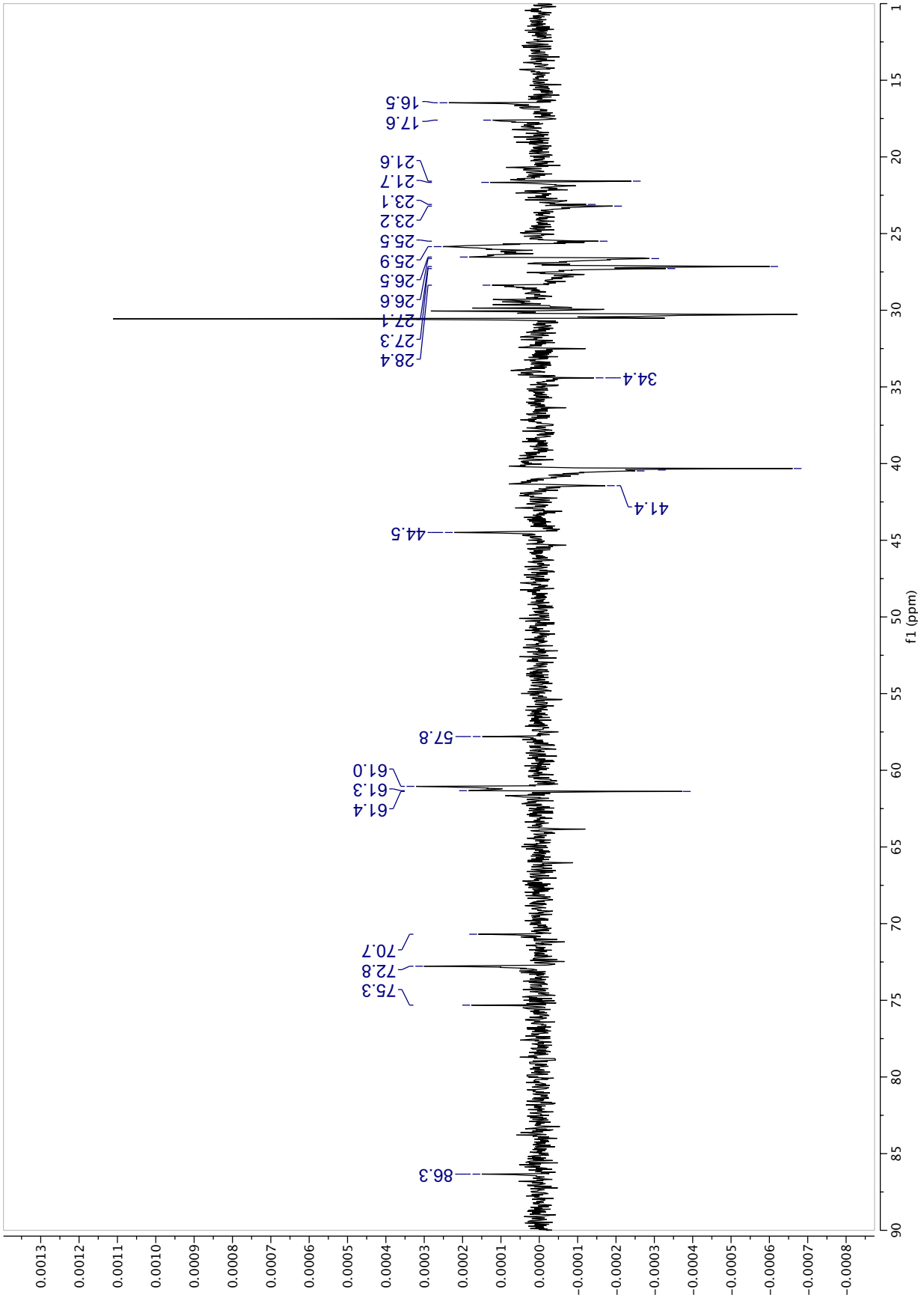


Figure S106. DEPT-135 spectrum (Zoom 1: 10-90 ppm) (1000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

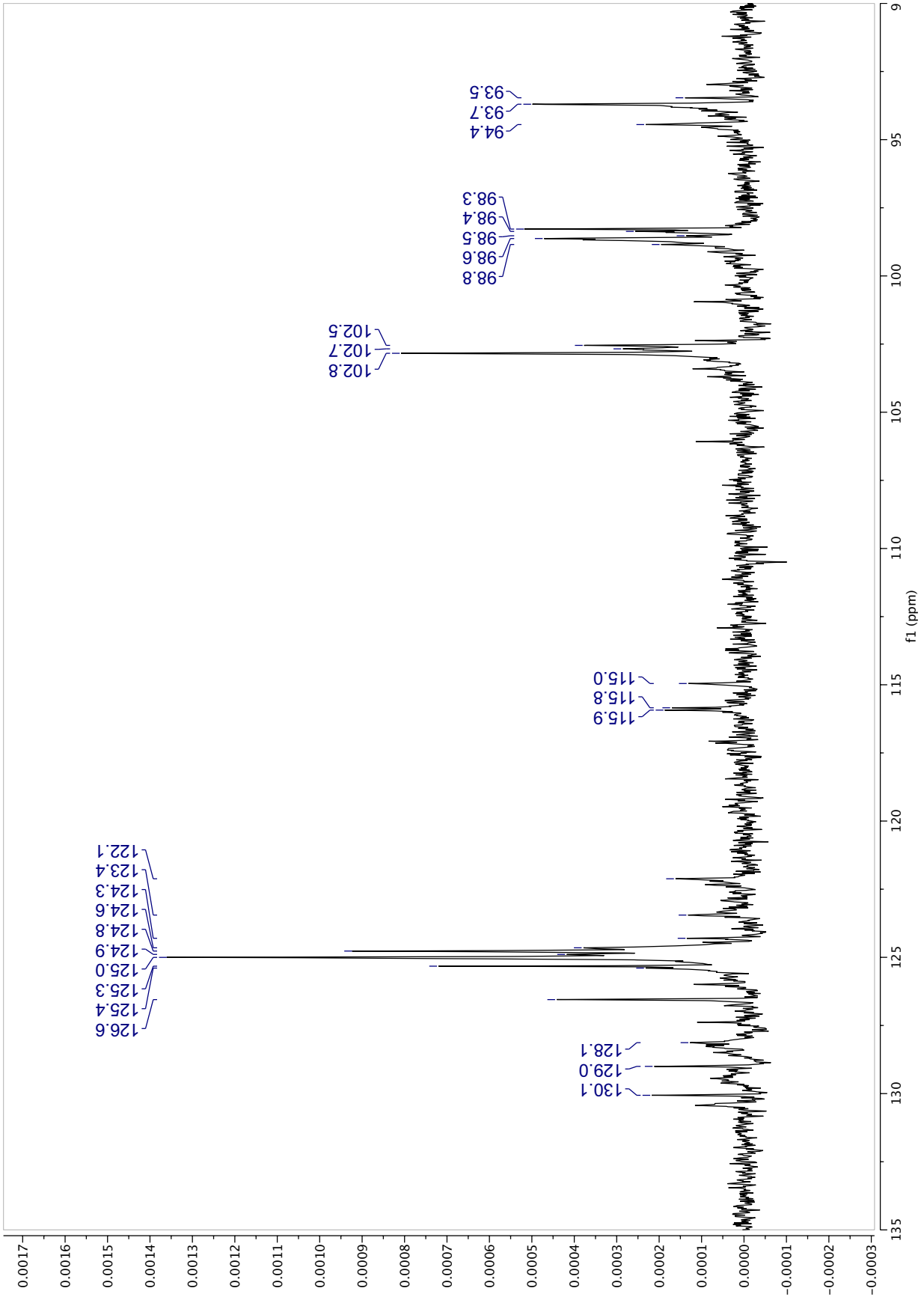


Figure S107. DEPT-135 spectrum (Zoom 2: 90-135 ppm) (1000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

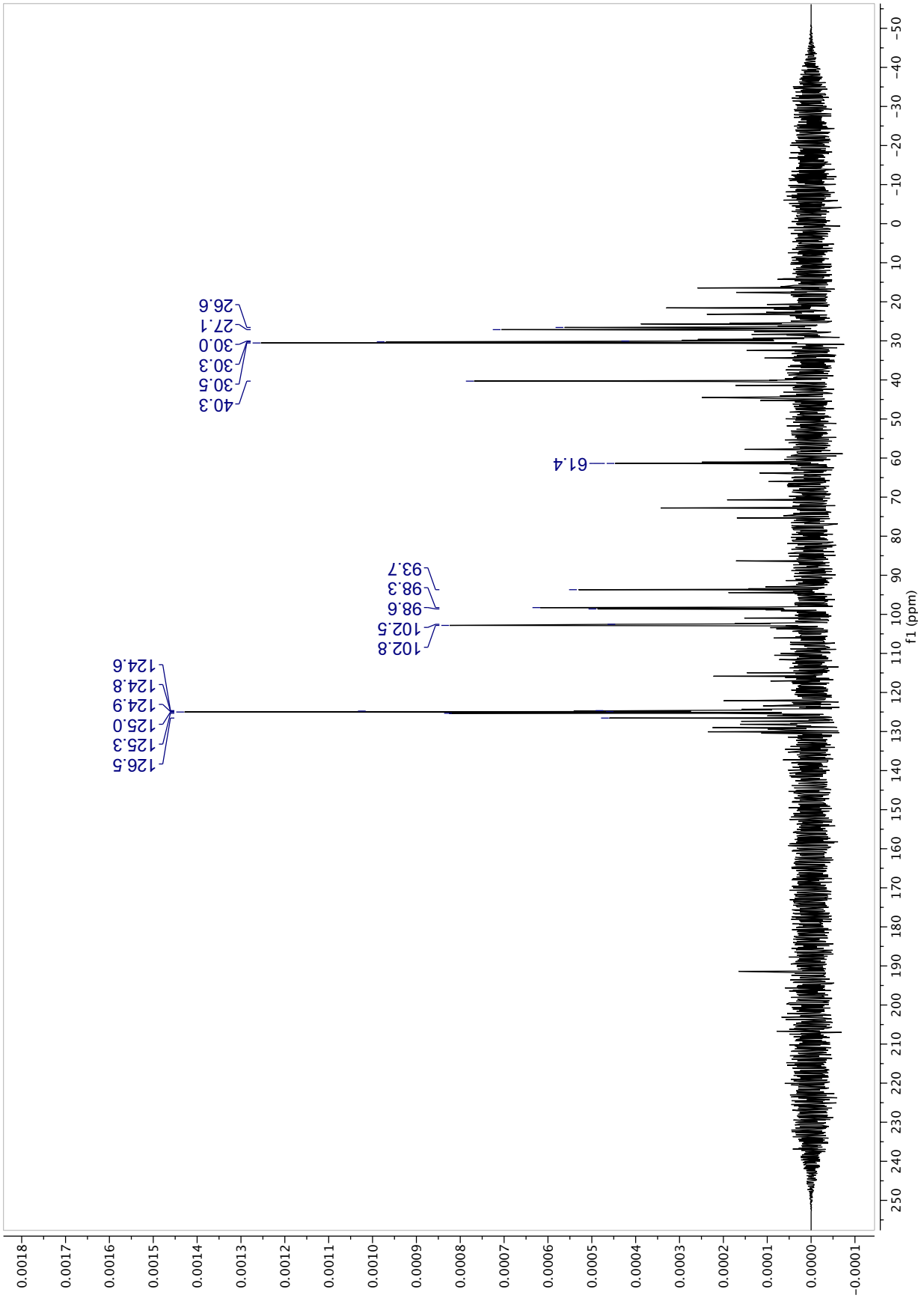


Figure S108. DEPT-90 spectrum (1000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

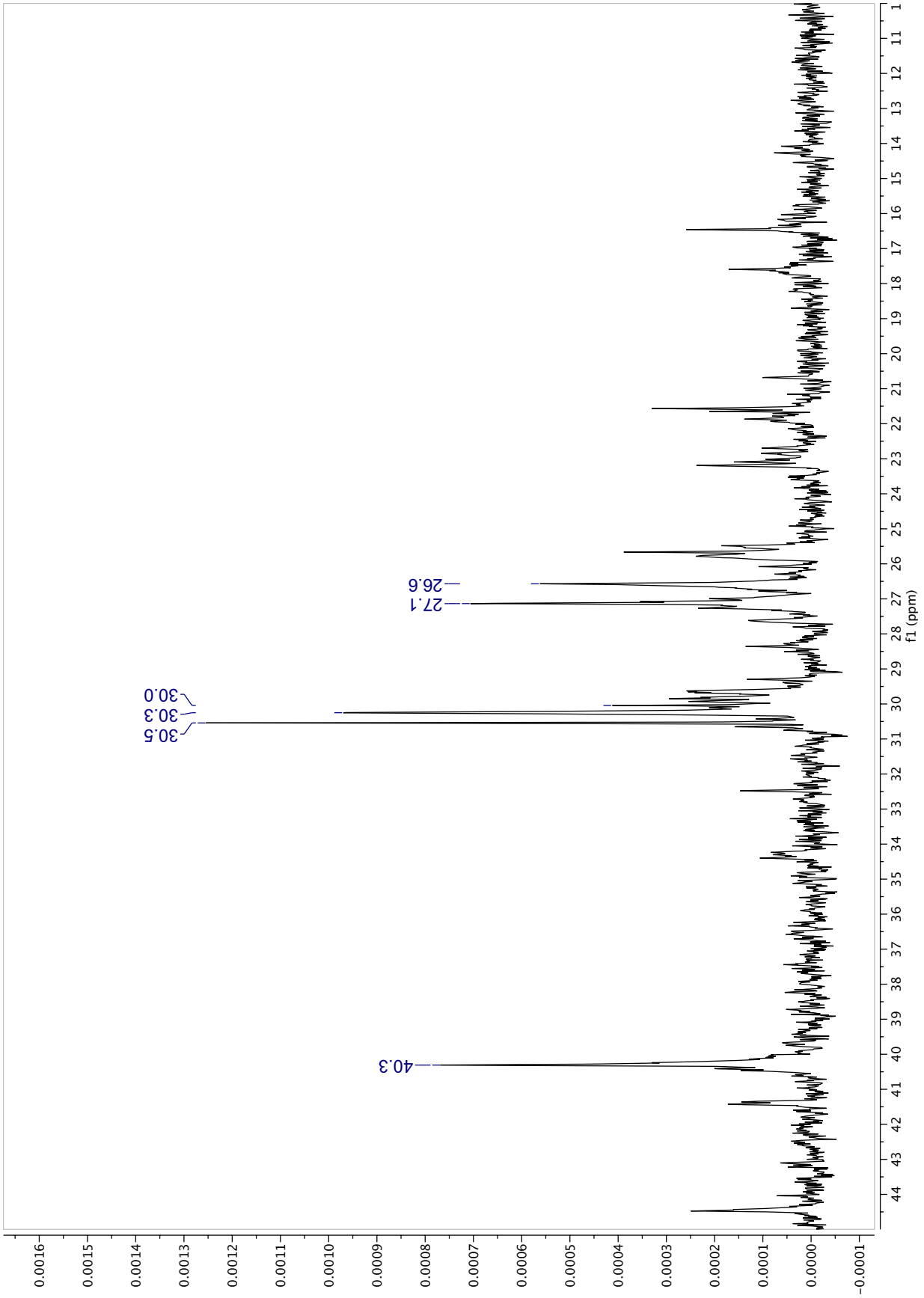


Figure S109. DEPT-90 spectrum (Zoom 1: 10-45 ppm) (1000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

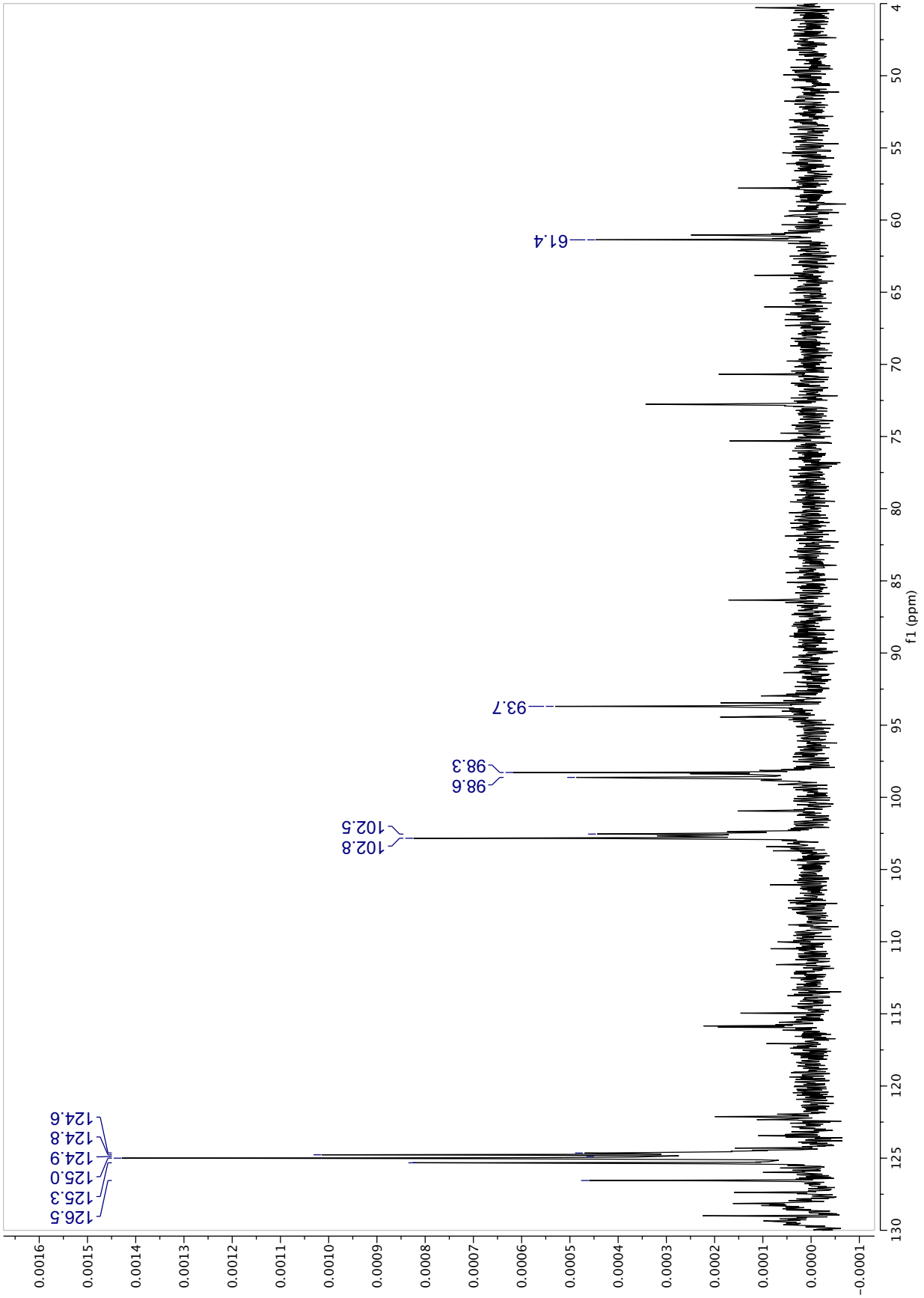


Figure S110. DEPT-90 spectrum (Zoom 2: 45-130 ppm) (1000 scans) of F9 (40.0 mg) recorded in acetone-*d*6.

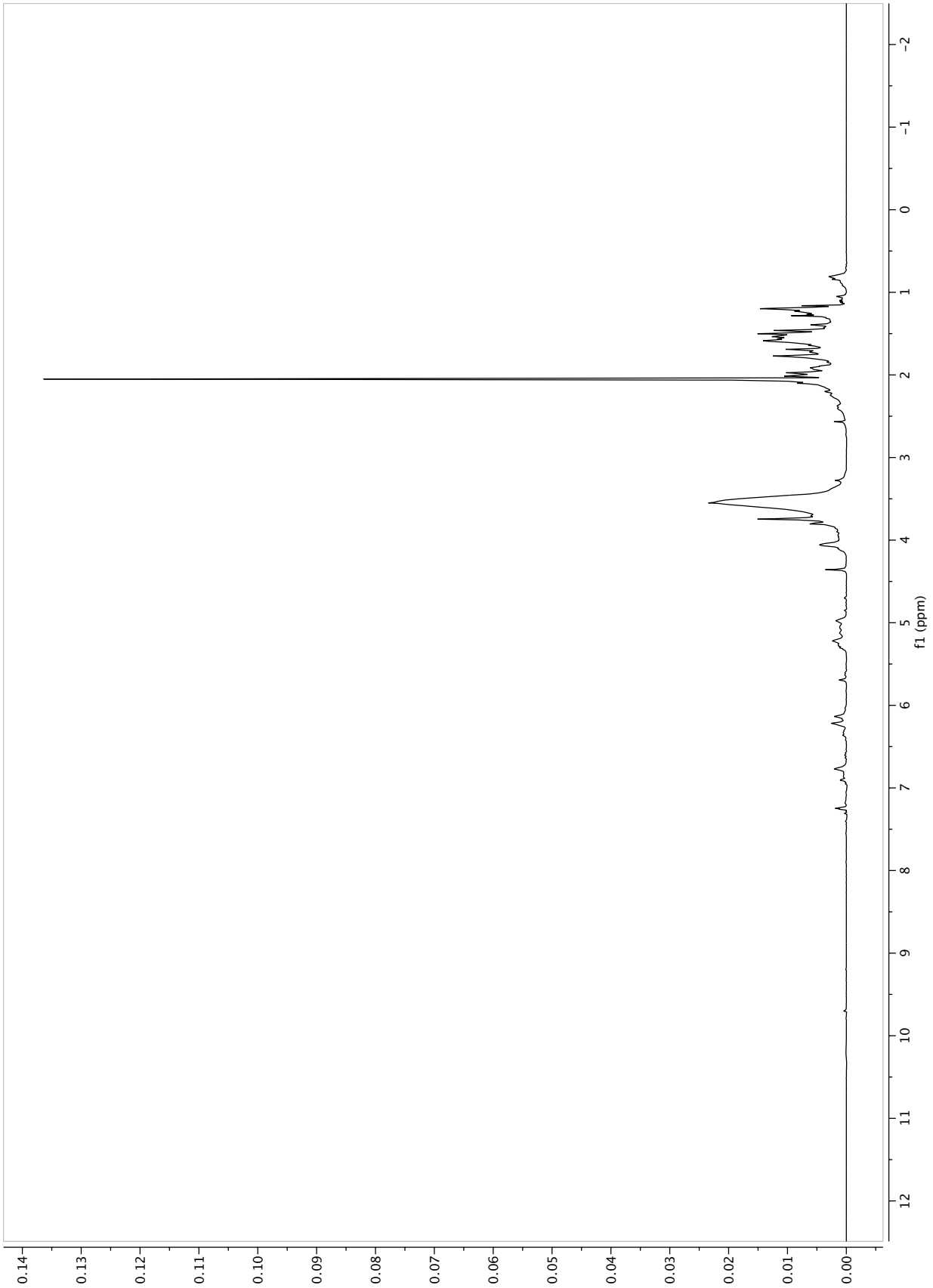


Figure S111. ^{1}H spectrum of F9 (40.0 mg) recorded in acetone-*d*6.

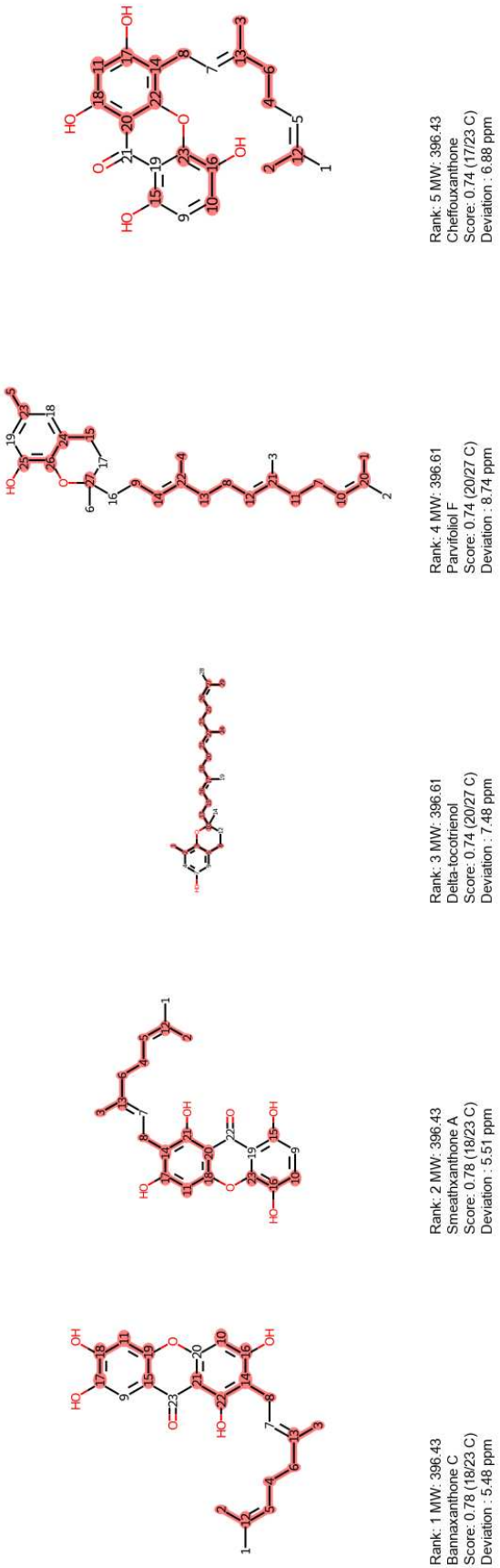


Figure S112. Second part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and 135) of F9 using c-type *Garcinia* DB. A molecular filter (396 g.mol^{-1}) was applied. Equivalent carbons were allowed.

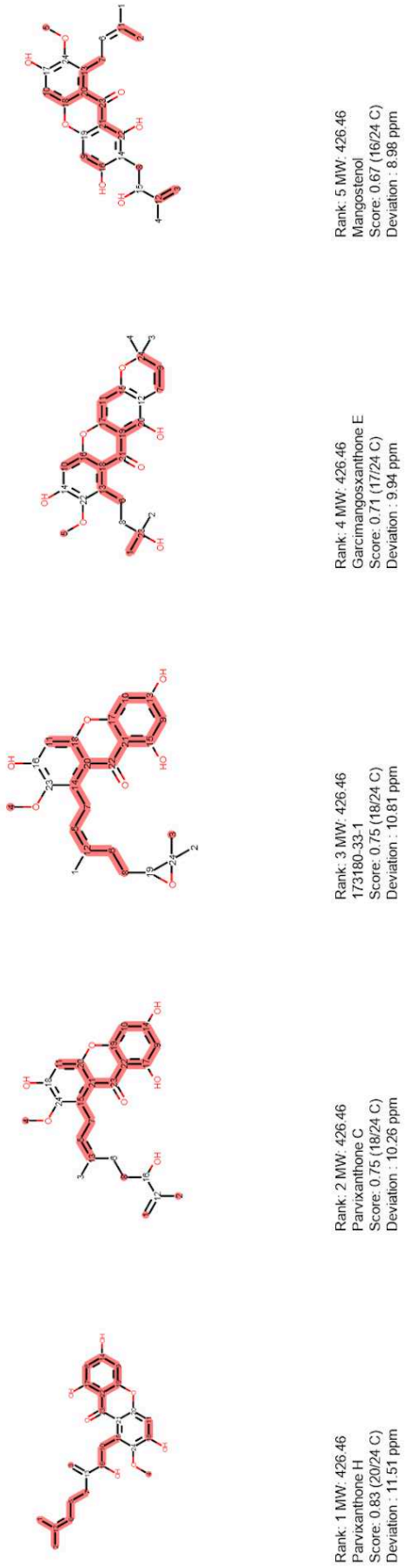


Figure S113. First part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and 135) of F9 using c-type *Garcinia* DB. A molecular filter (426 g.mol^{-1}) was applied. Equivalent carbons were allowed

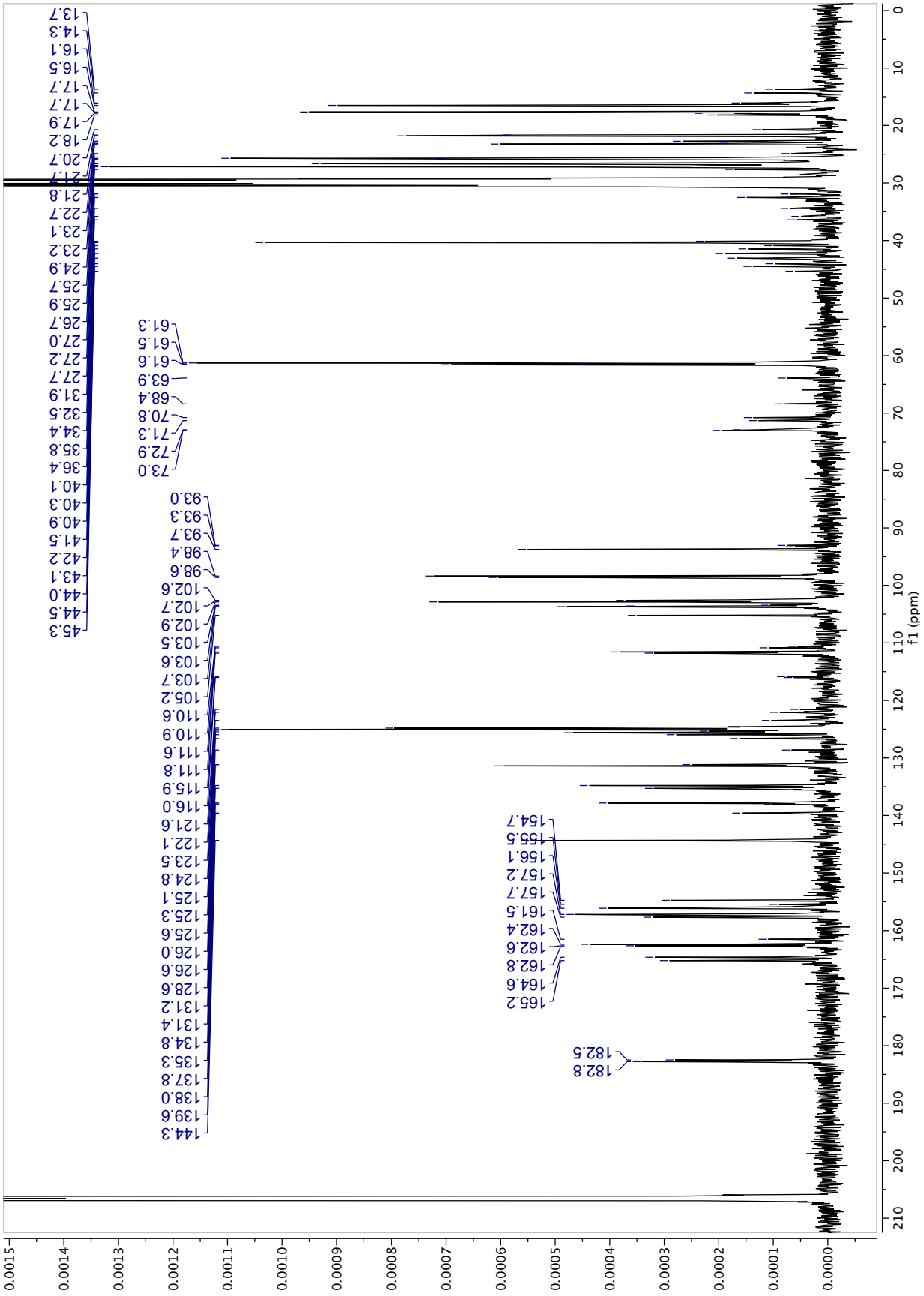


Figure S114. ^{13}C-NMR spectrum (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

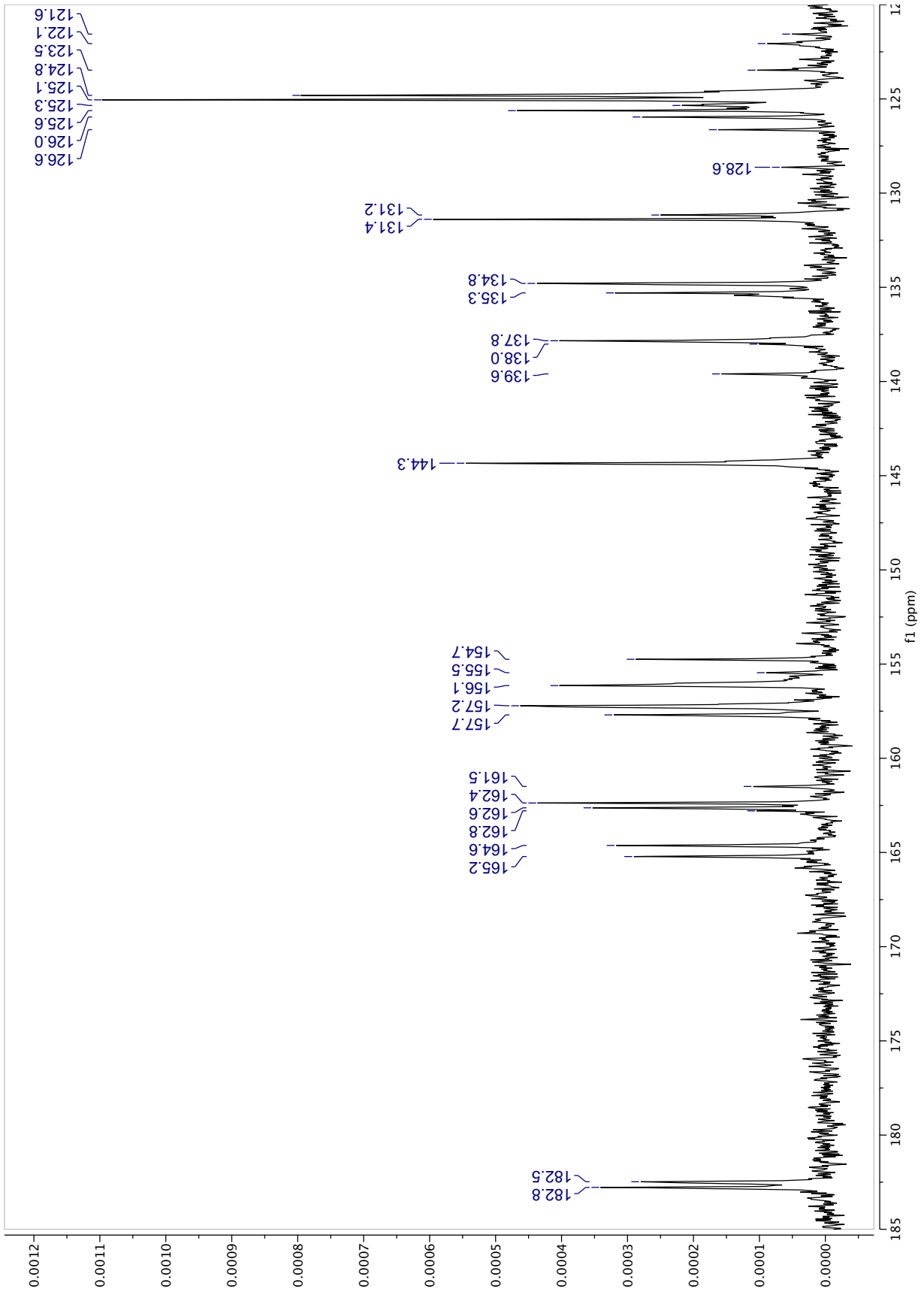


Figure S115. ^{13}C-NMR spectrum (Zoom 1: 185-120ppm) (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

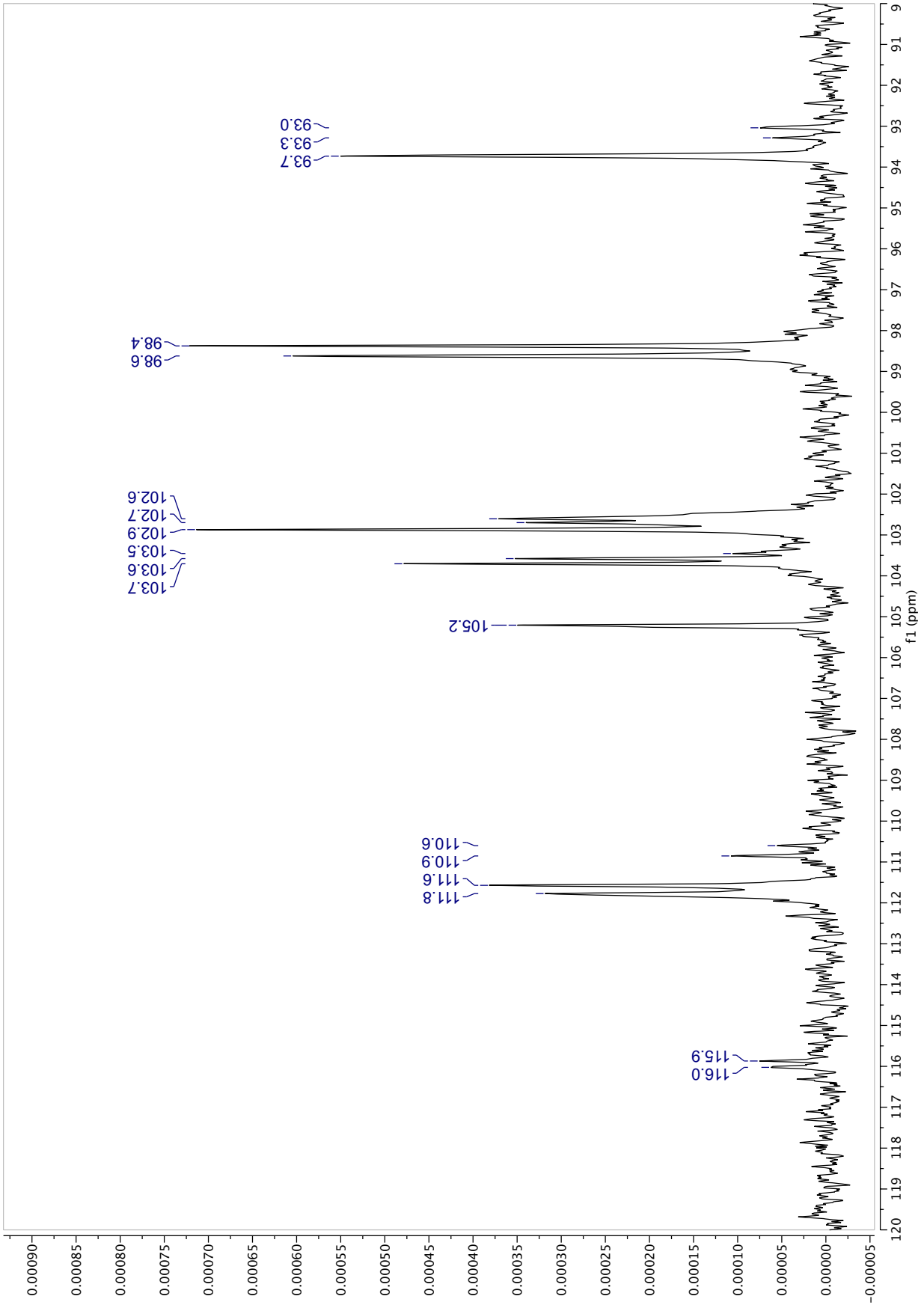


Figure S116. ^{13}C-NMR spectrum (Zoom 2: 120-90ppm) (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

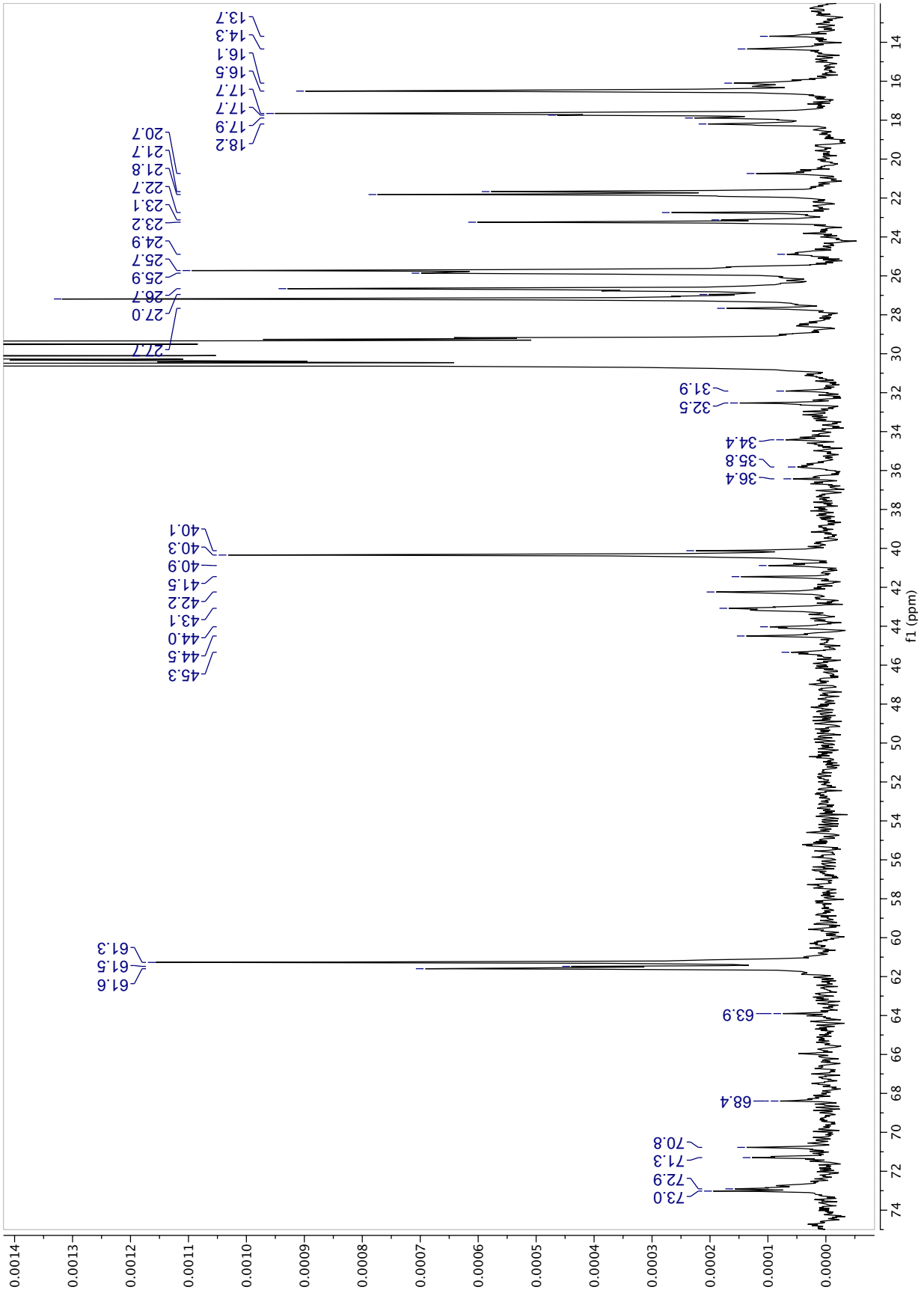


Figure S117. ^{13}C-NMR spectrum (Zoom 3: 75-12ppm) (1024 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

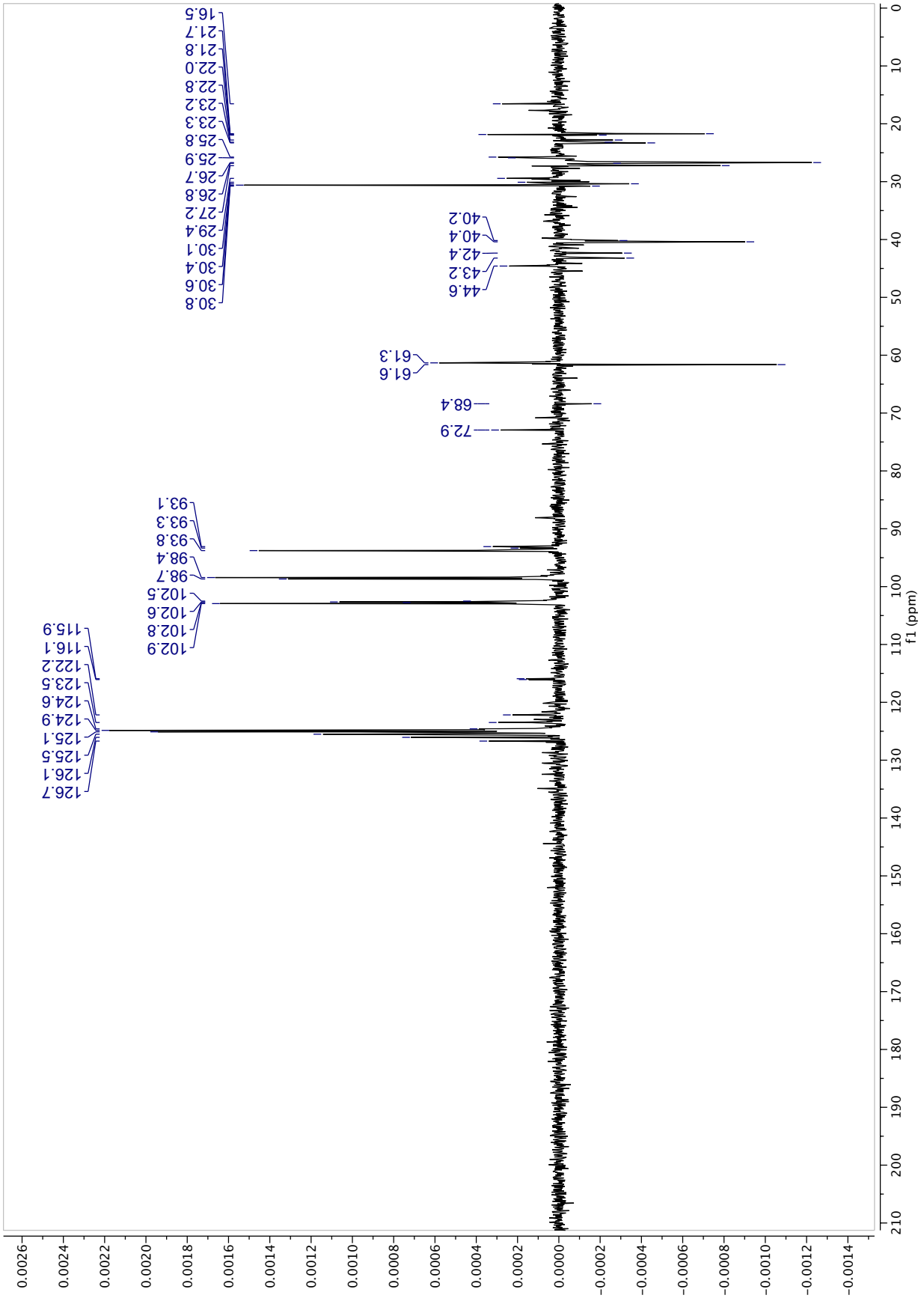


Figure S118. DEPT-135 spectrum (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

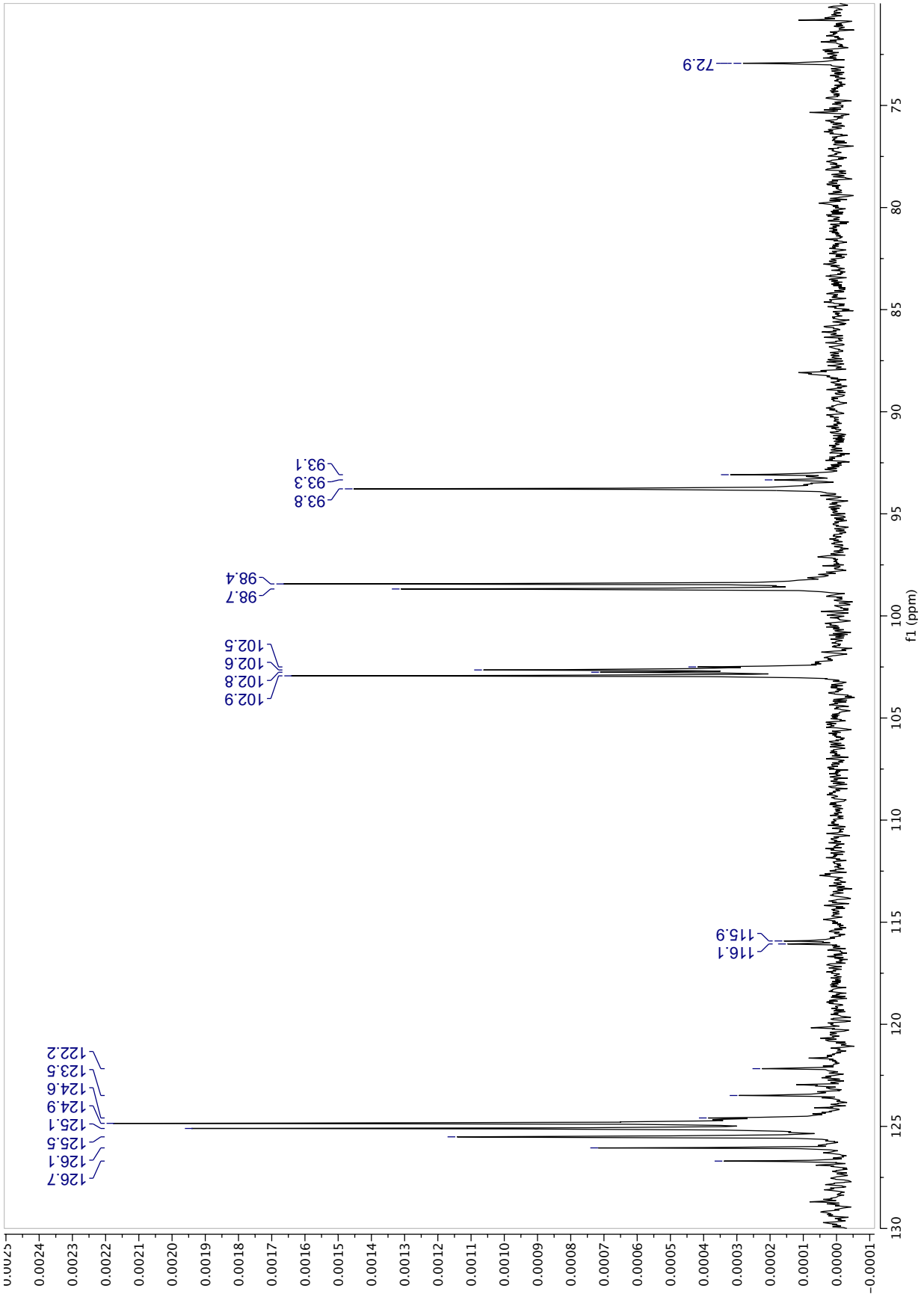


Figure S119. DEPT-135 spectrum (Zoom 1: 130-70ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

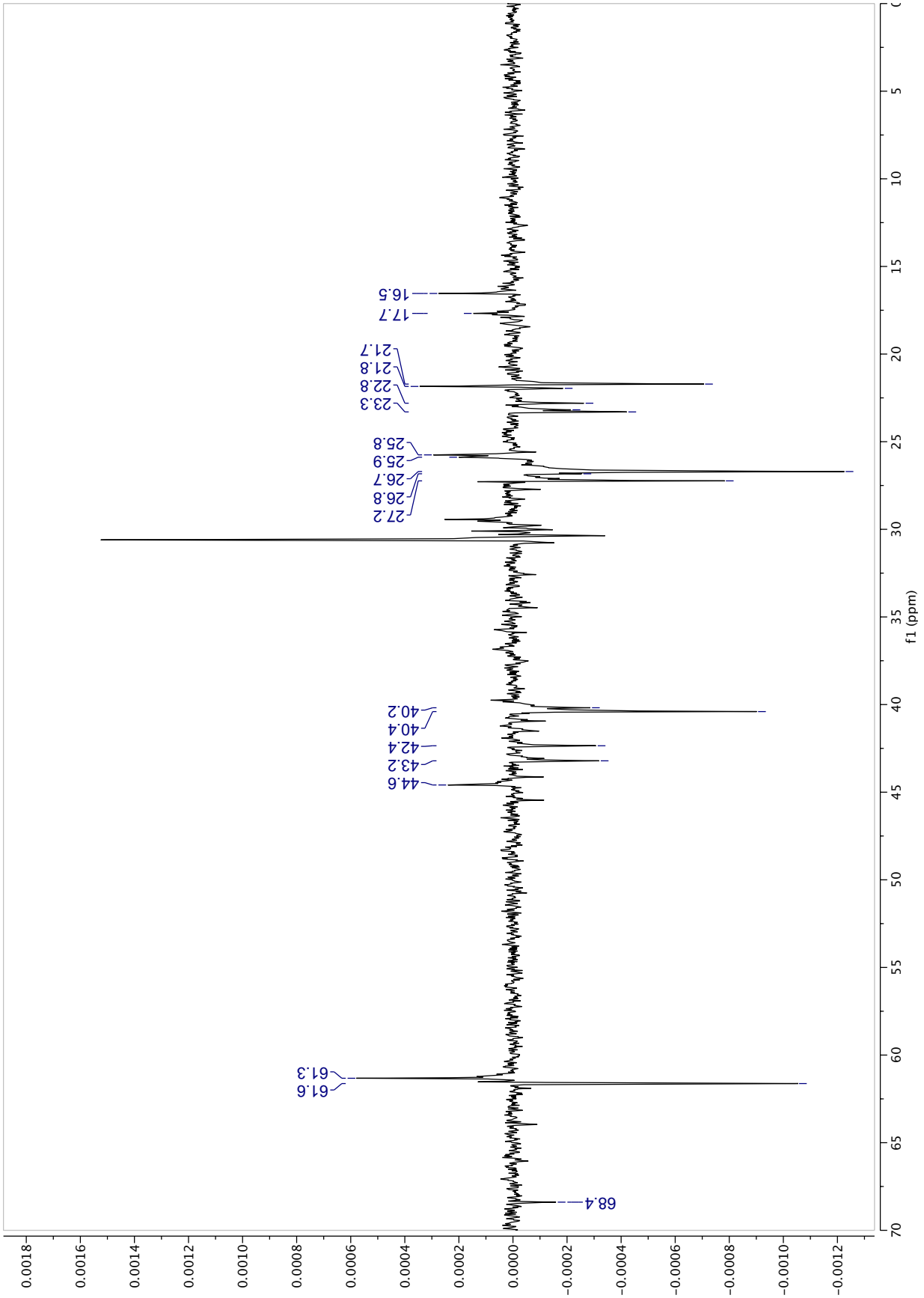


Figure S120. DEPT-135 spectrum (Zoom 2: 70-0ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

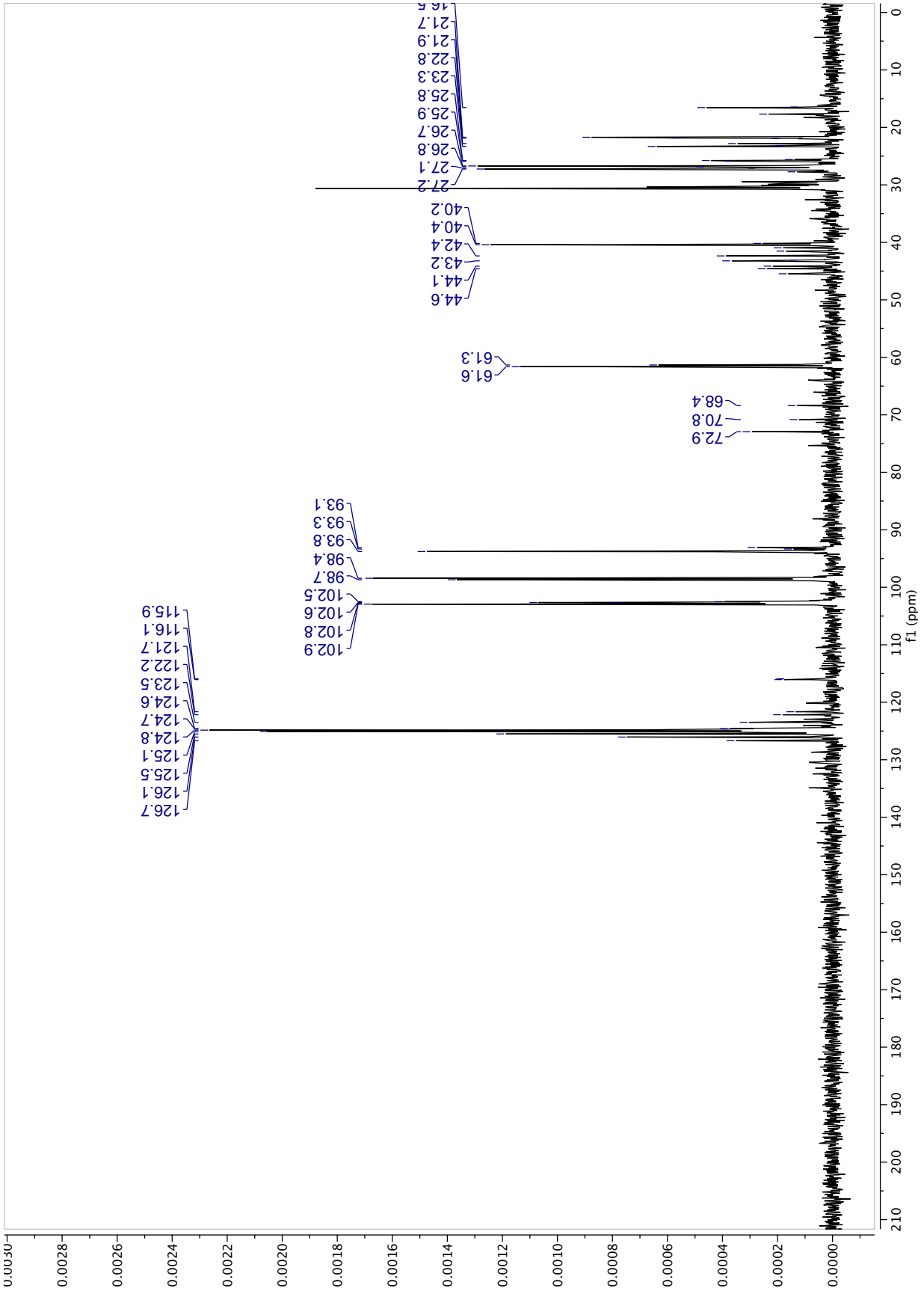


Figure S121. DEPT-90 spectrum (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

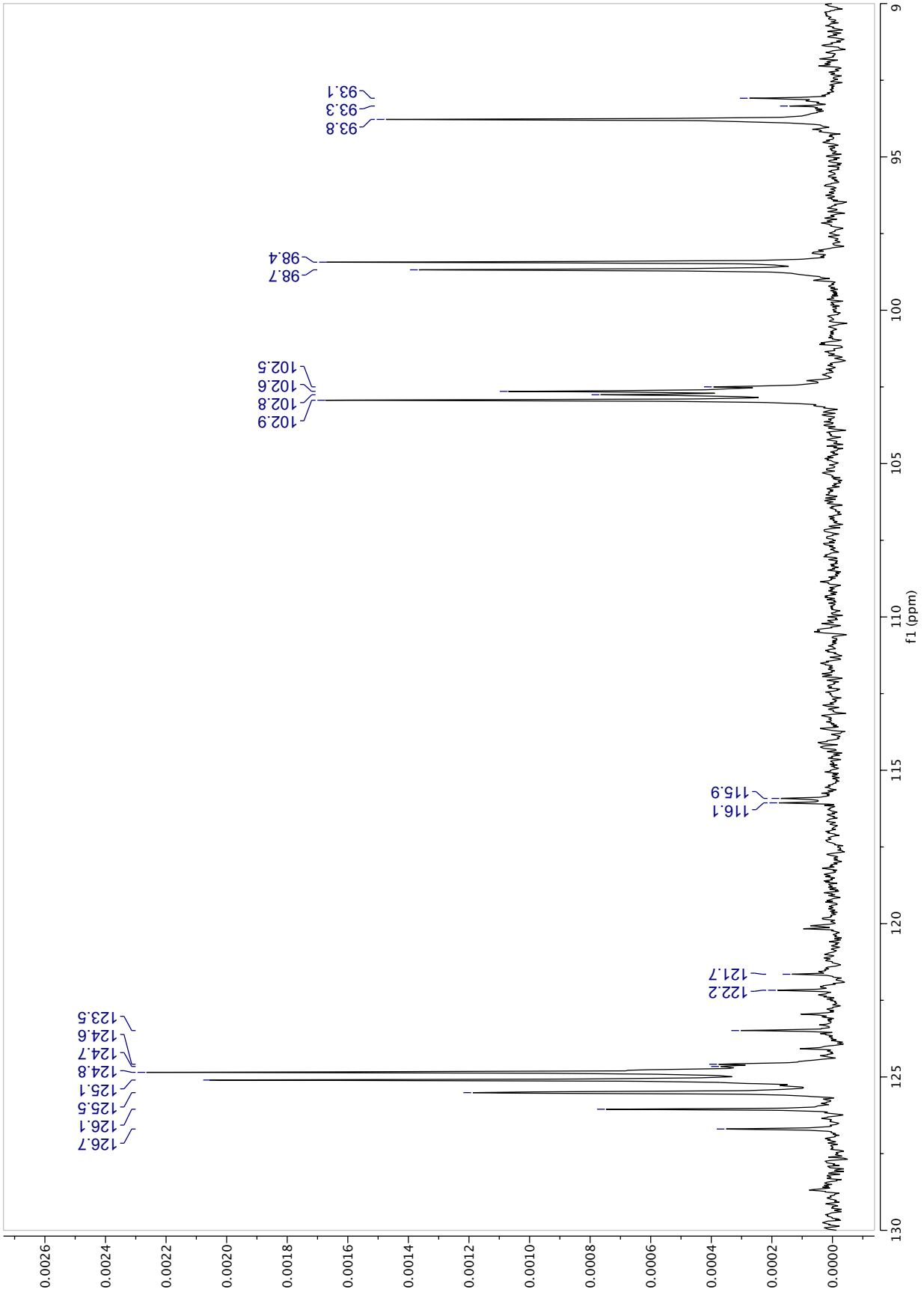


Figure S122. DEPT-90 spectrum (Zoom 1: 130-90ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

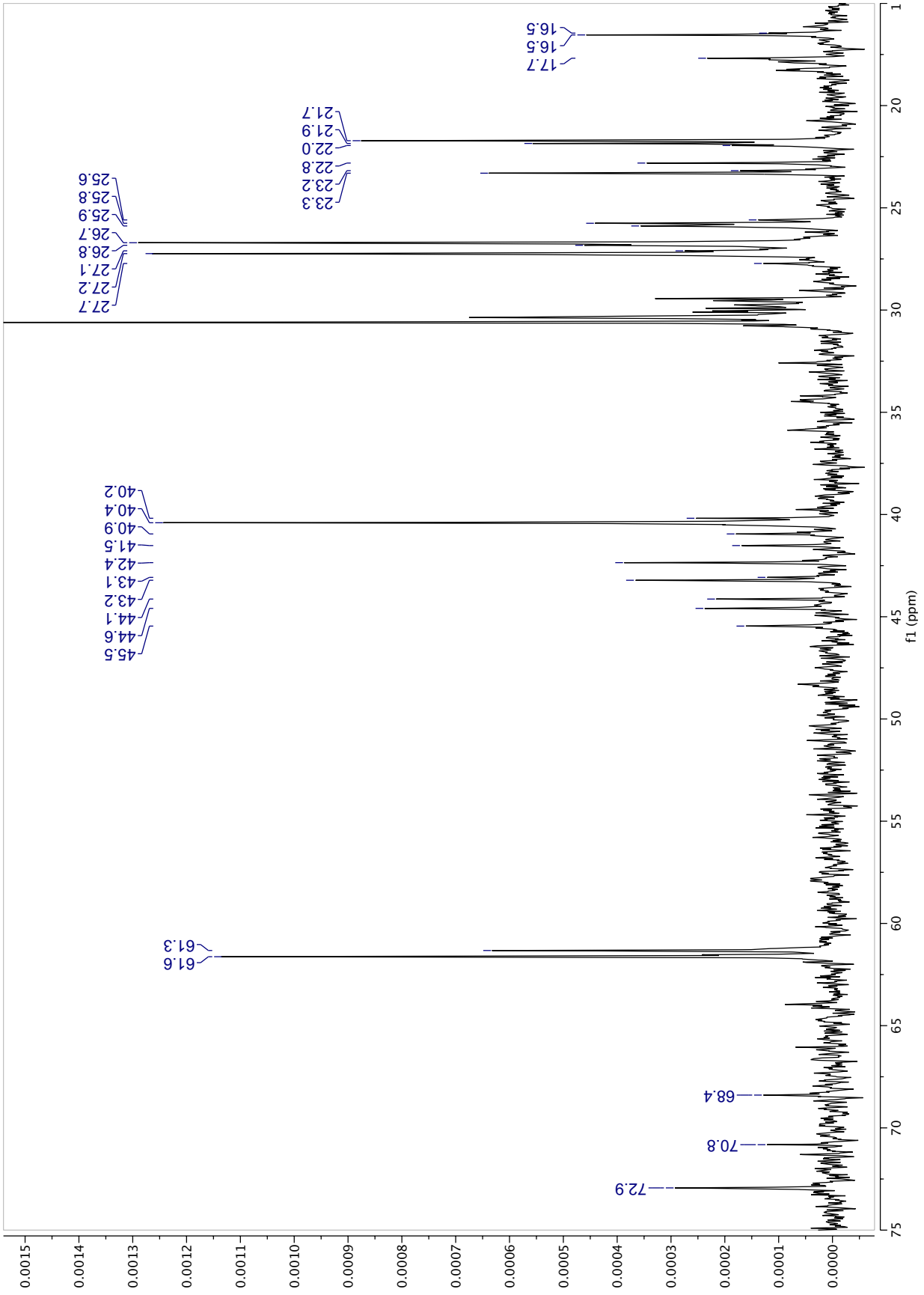


Figure S123. DEPT-90 spectrum (Zoom 2: 75-10ppm) (512 scans) of F10 (40.0 mg) recorded in acetone-*d*6.

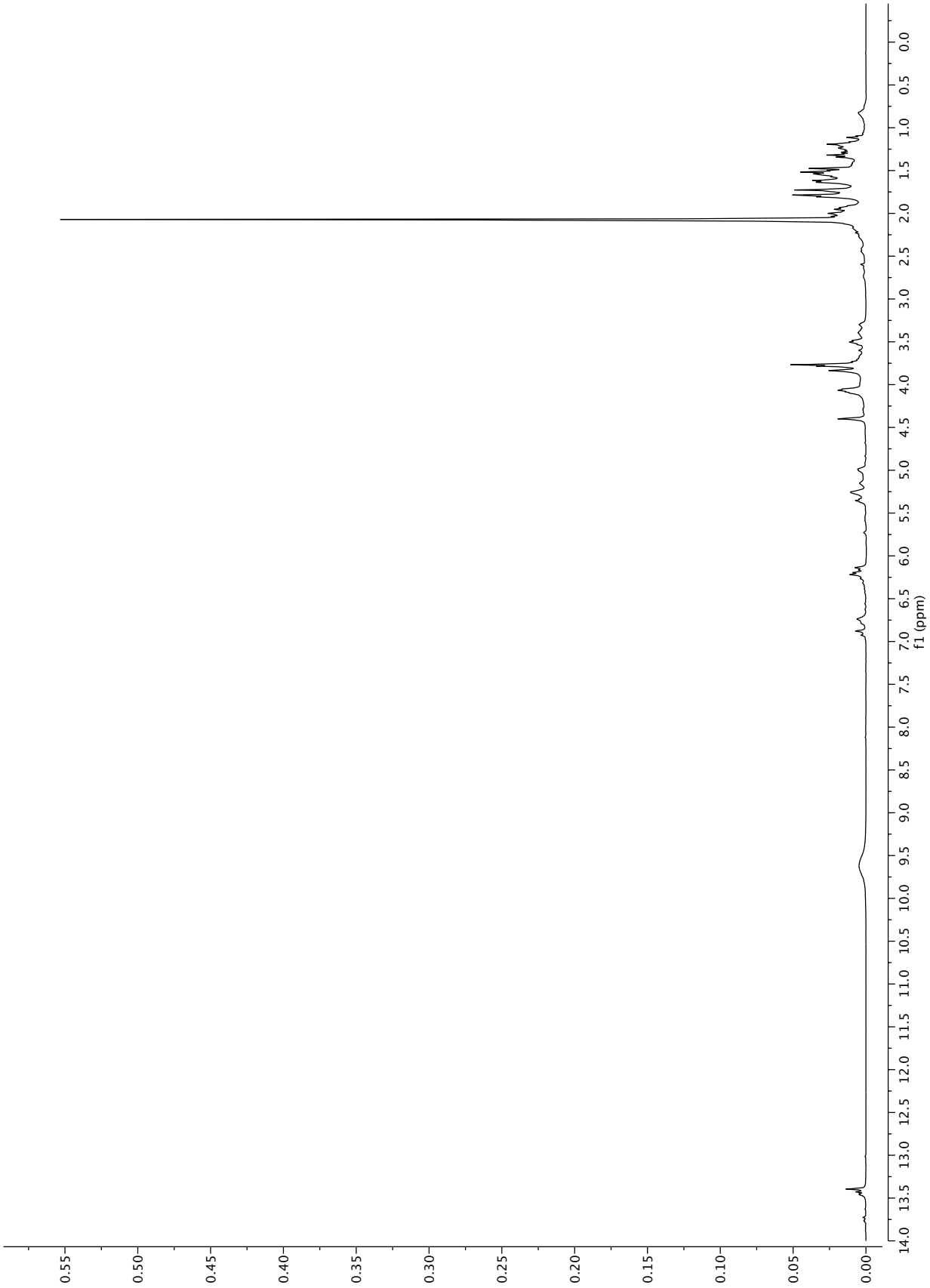


Figure S124. [1]H-NMR of F10 (40.0 mg) recorded in acetone-*d*6.

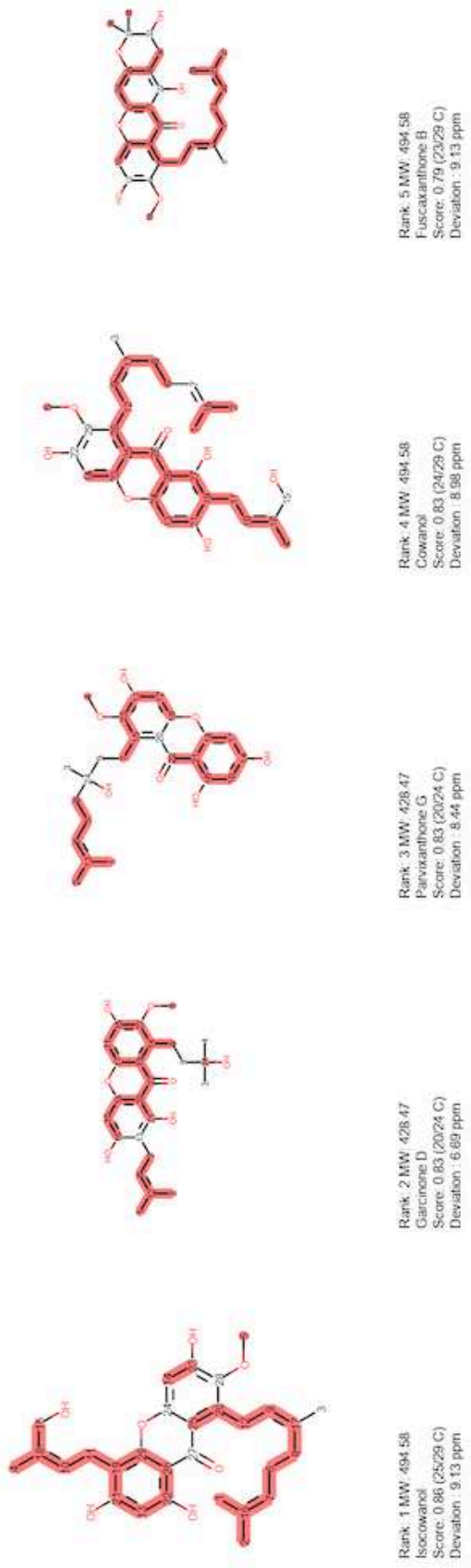


Figure S125. Part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and -135) of F10 using c-type *Garcinia* DB. Equivalent carbons were allowed, and a molecular weight filter (494 and 428 g.mol^{-1}) was applied.

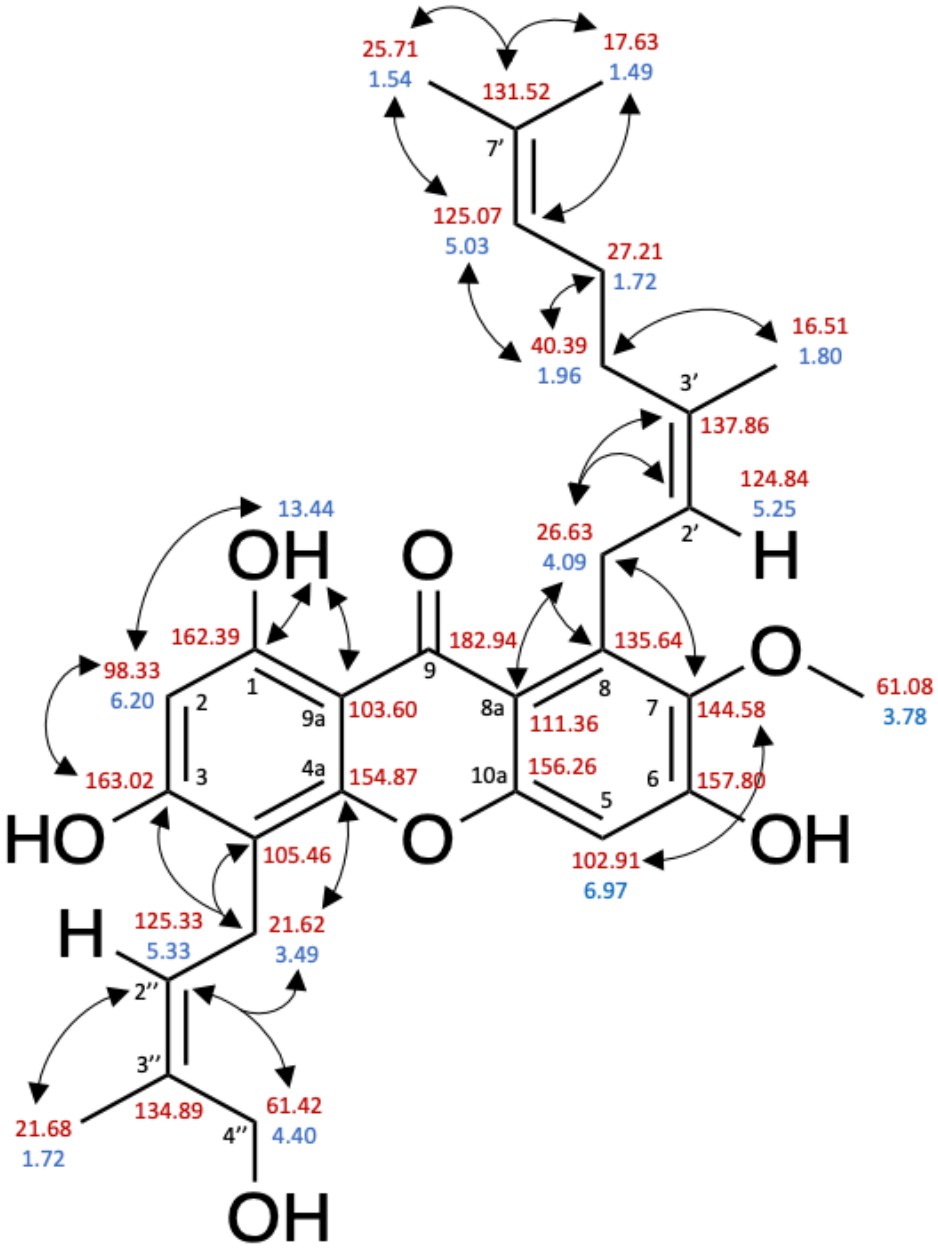


Figure S126. Structure and numbering of the isocowanol (**20**). δ_C are written in red whereas δ_H are written in blue. The analysis was performed in acetone-*d*6.

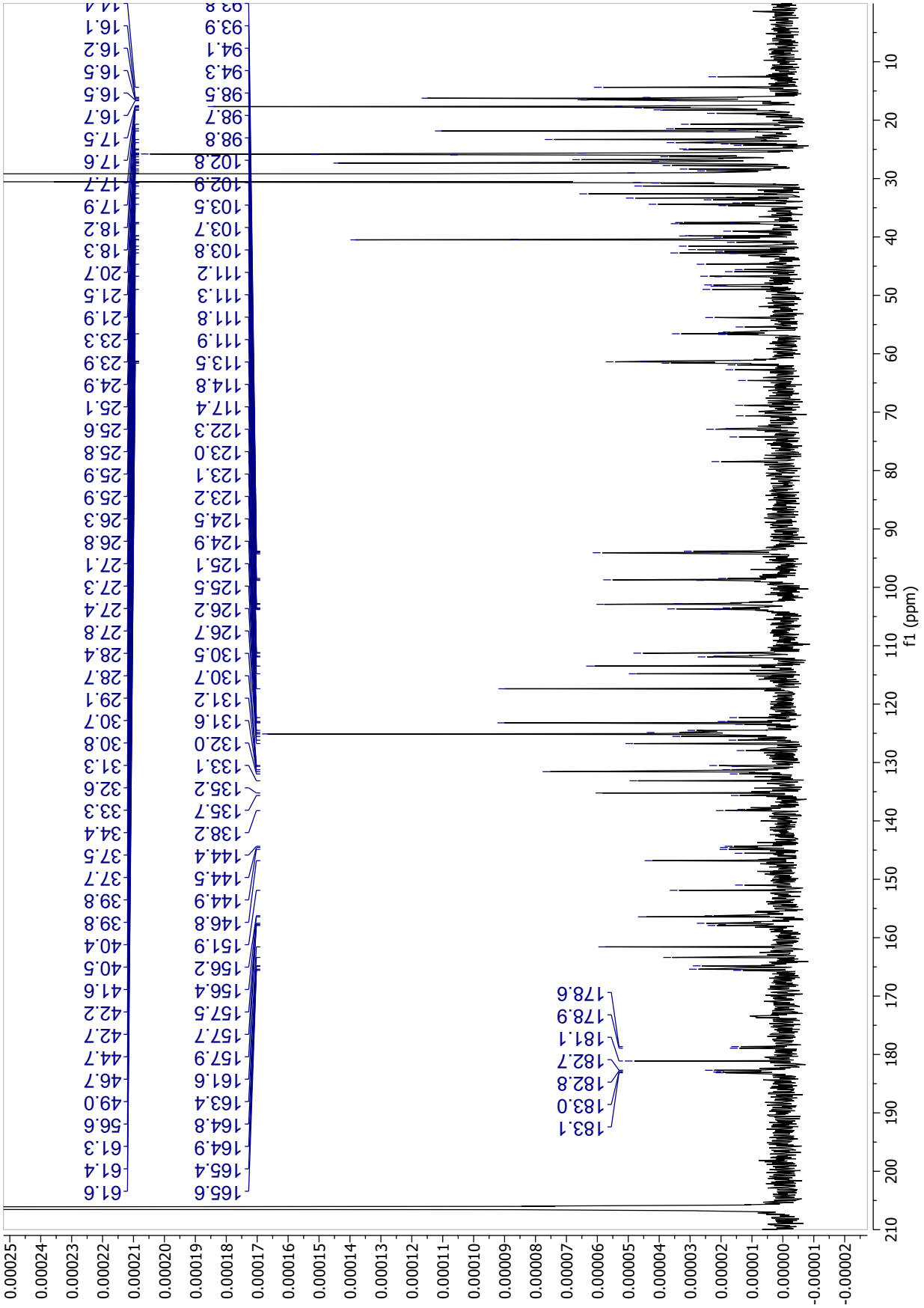


Figure S127. ^{13}C-NMR spectrum (8000 scans) of F12 (39.0 mg) recorded in acetone-*d*6.

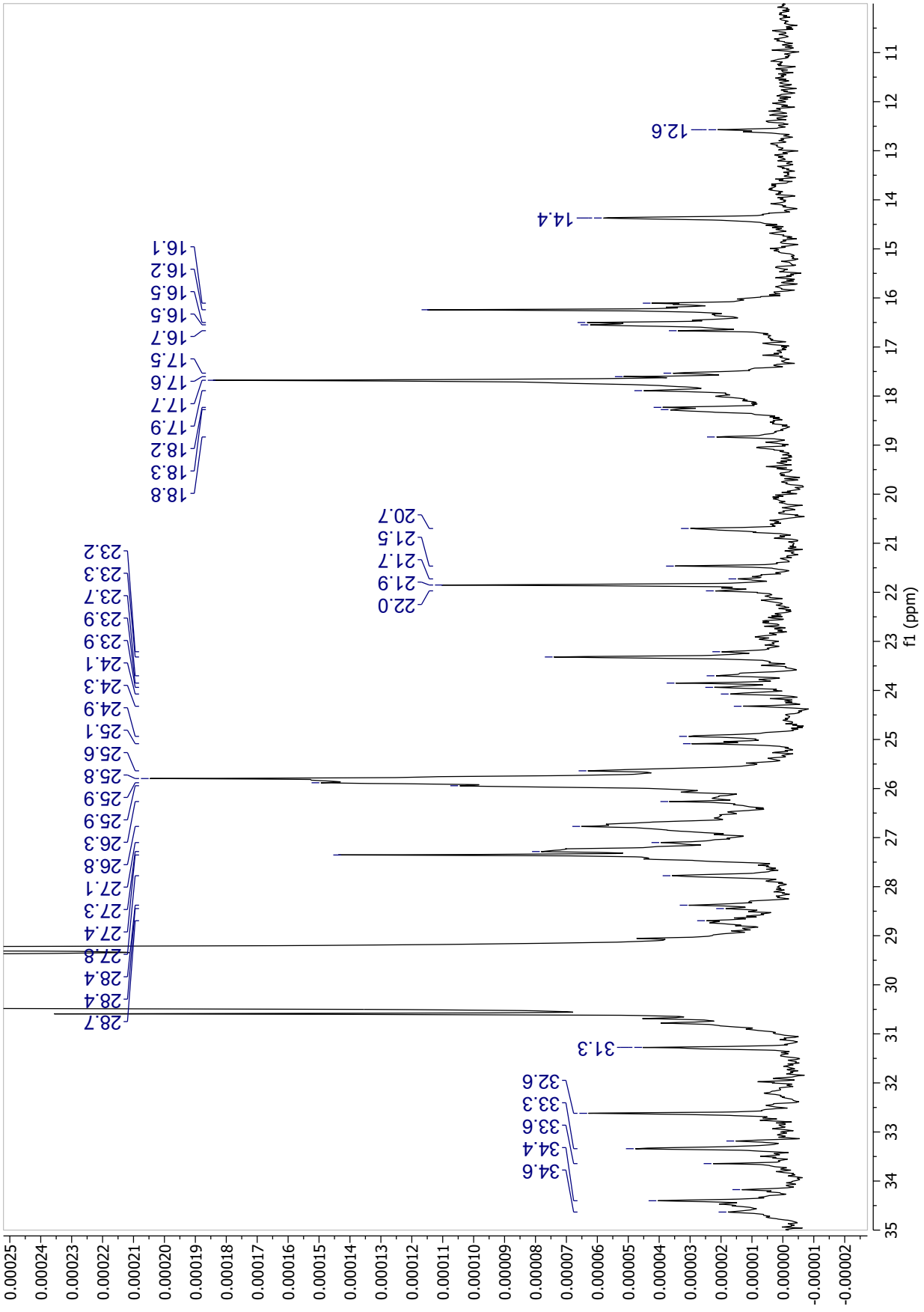


Figure S128. ^{13}C-NMR spectrum (Zoom1: 10-35 ppm) (8000 scans) of F12 (39.0 mg) recorded in acetone-*d*6.

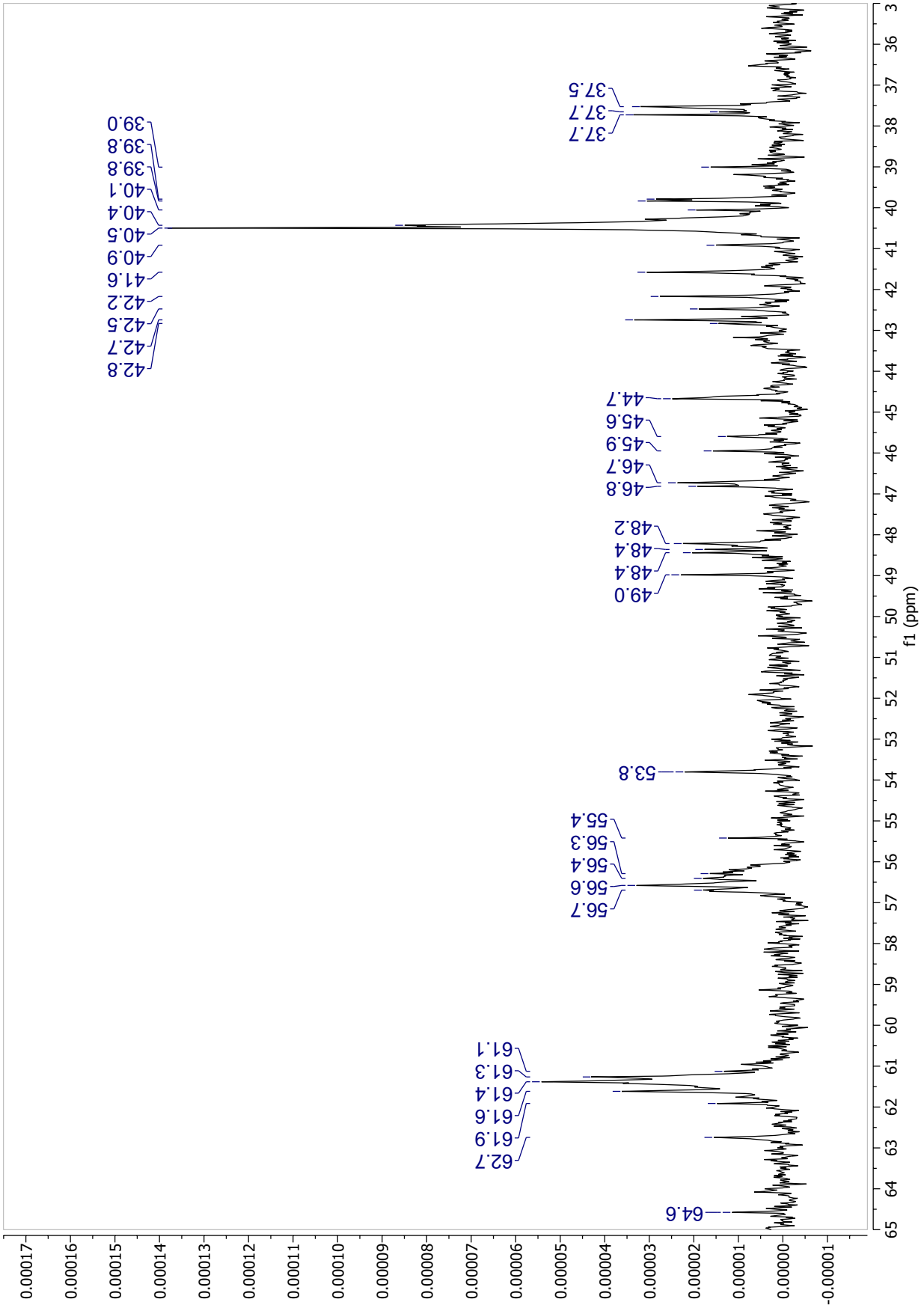


Figure S129. ^{13}C-NMR spectrum (Zoom2: 35-65 ppm) (8000 scans) of F12 (39.0 mg) recorded in acetone-*d*6.

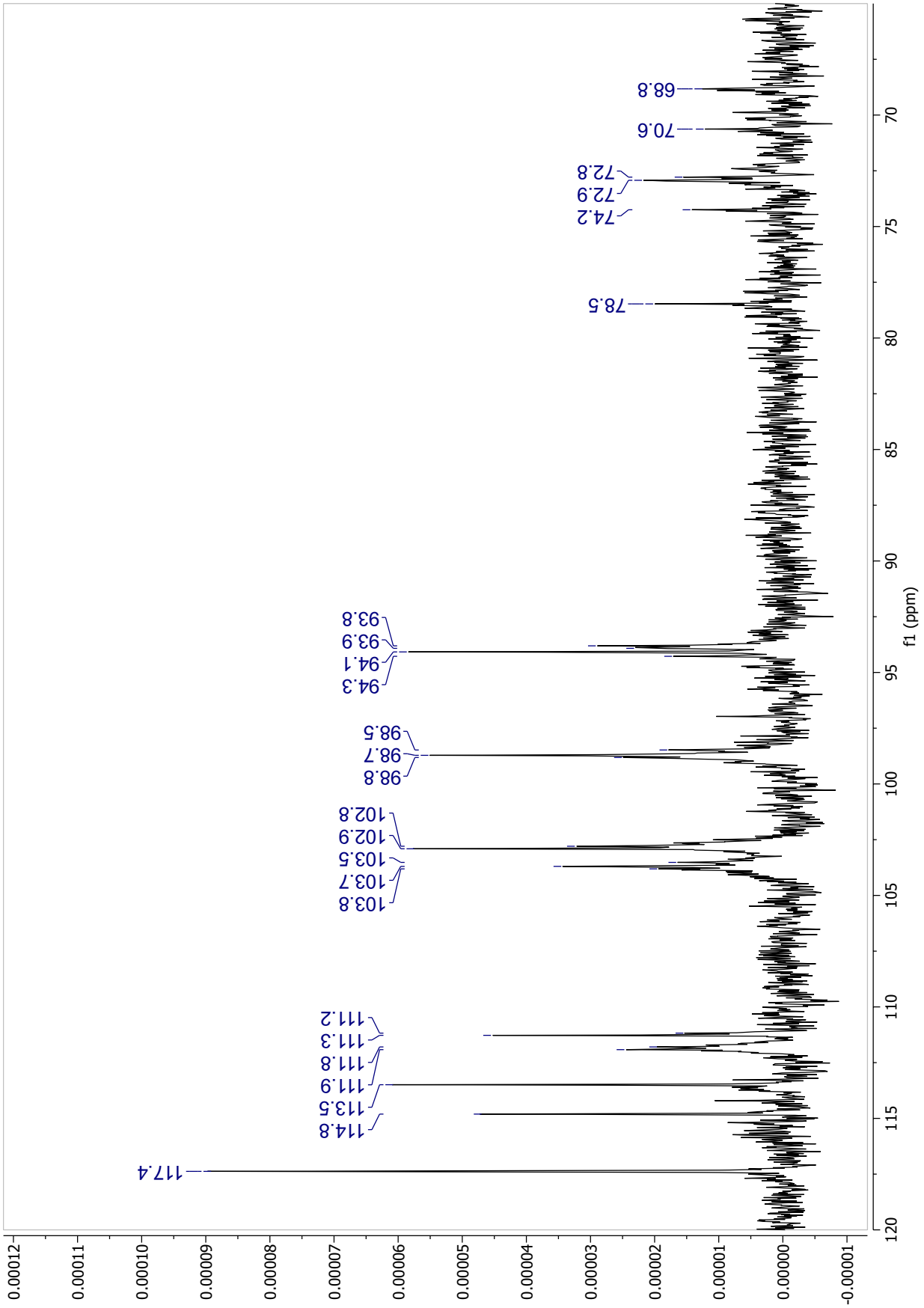


Figure S130. ^{13}C-NMR spectrum (Zoom3: 65-120 ppm) (8000 scans) of F12 (39.0 mg) recorded in acetone-*d*6

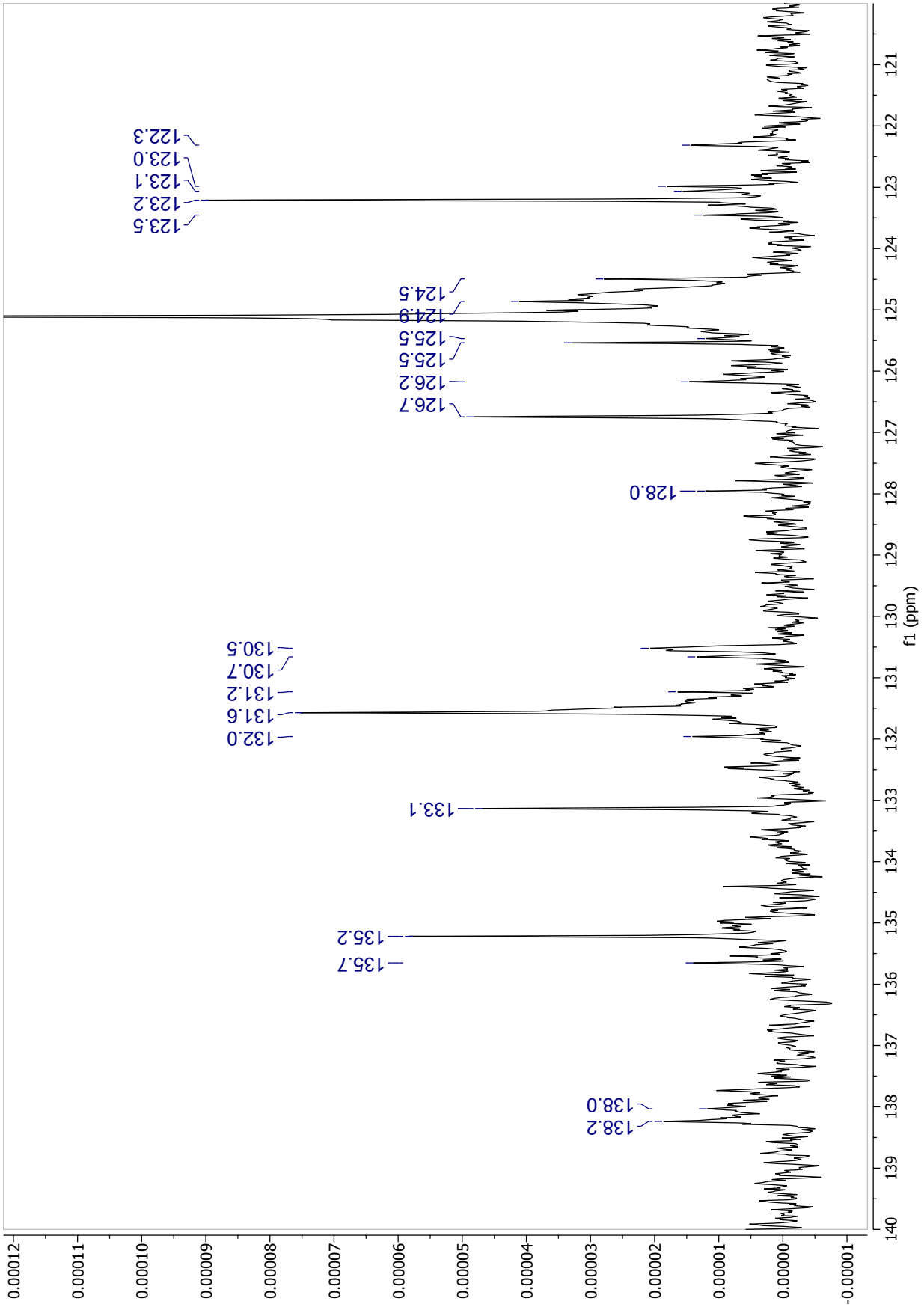


Figure S131. ^{13}C-NMR spectrum (Zoom4: 120-140 ppm) (8000 scans) of F12 (39.0 mg) recorded in acetone-*d*6

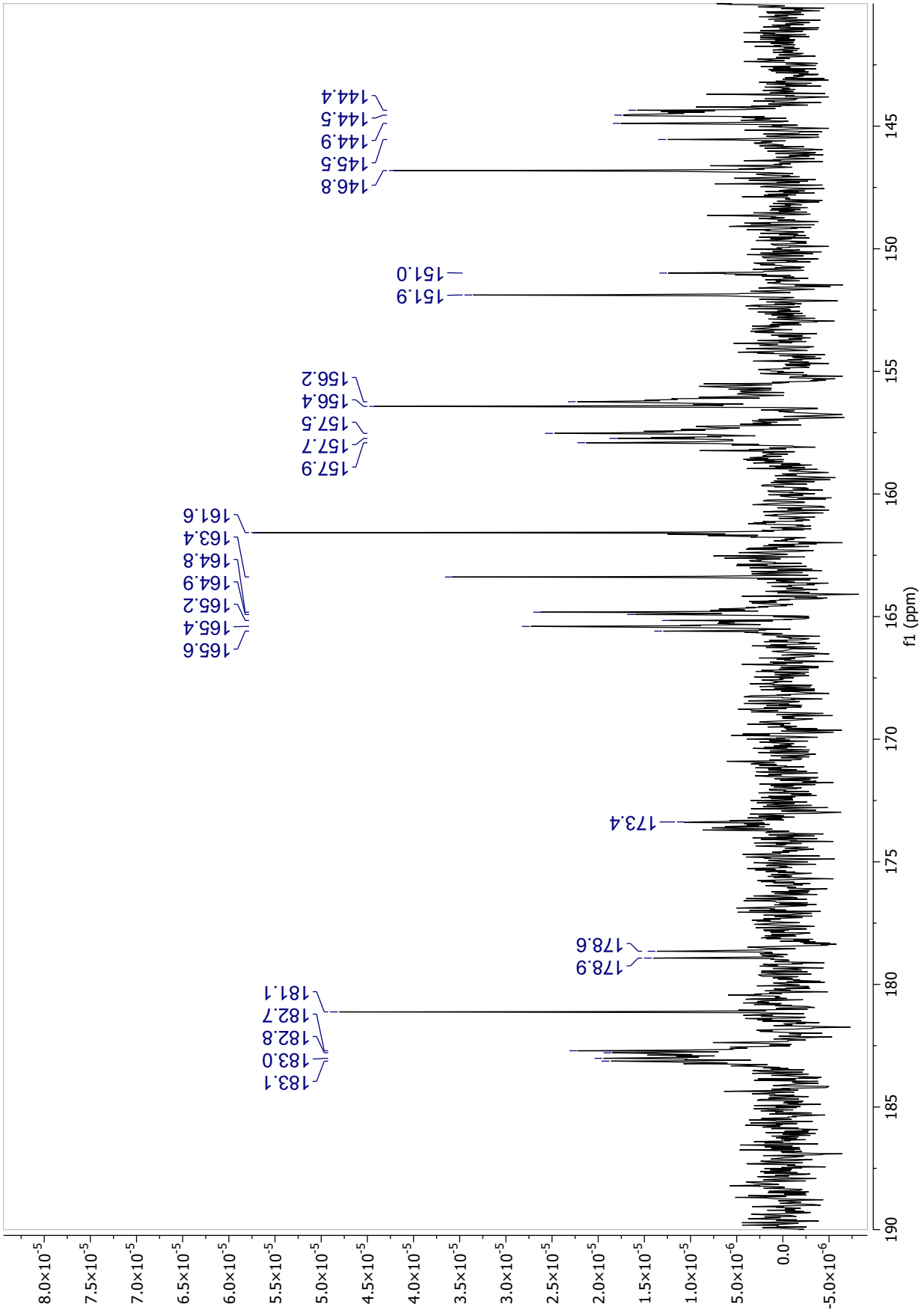


Figure S132. ^{13}C-NMR spectrum (Zoom5: 140-190 ppm) (8000 scans) of F12 (39.0 mg) recorded in acetone-*d6*

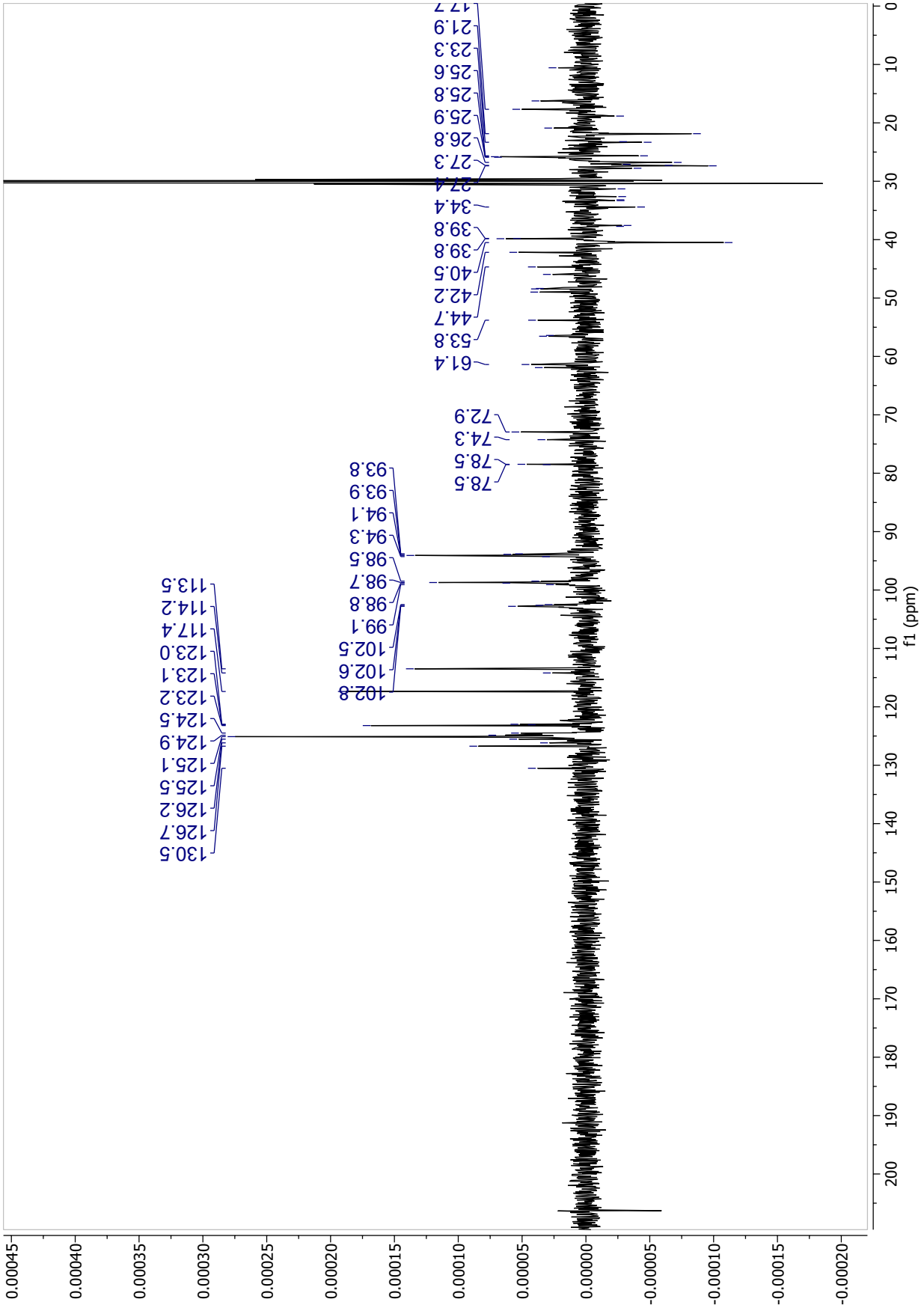


Figure S133. DEPT-135 spectrum (2000 scans) of F12 (39.0 mg) recorded in acetone-*d*6

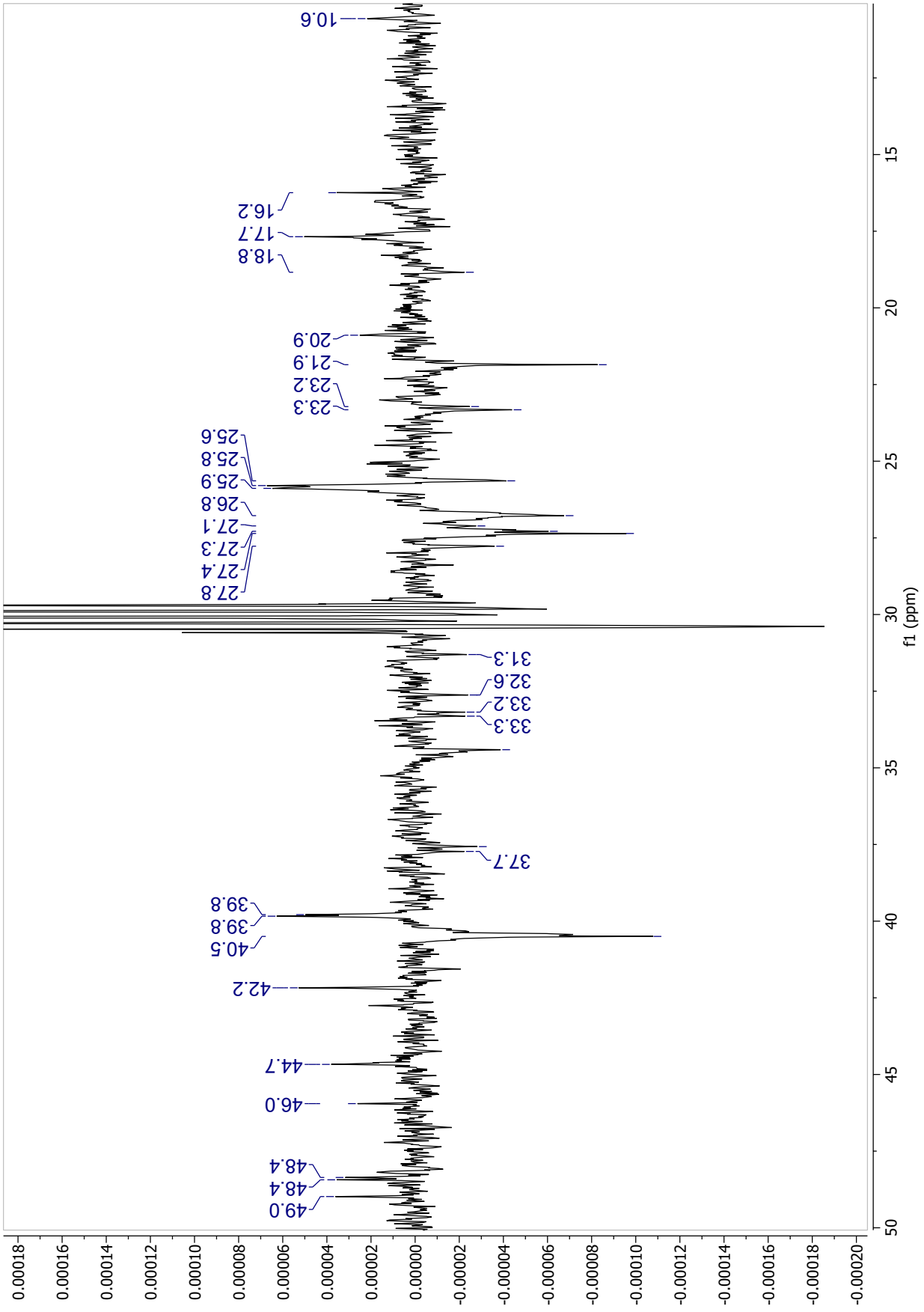


Figure S134. DEPT-135 spectrum (Zoom1: 10-50 ppm) (2000 scans) of F12 (39.0 mg) recorded in acetone-*d6*

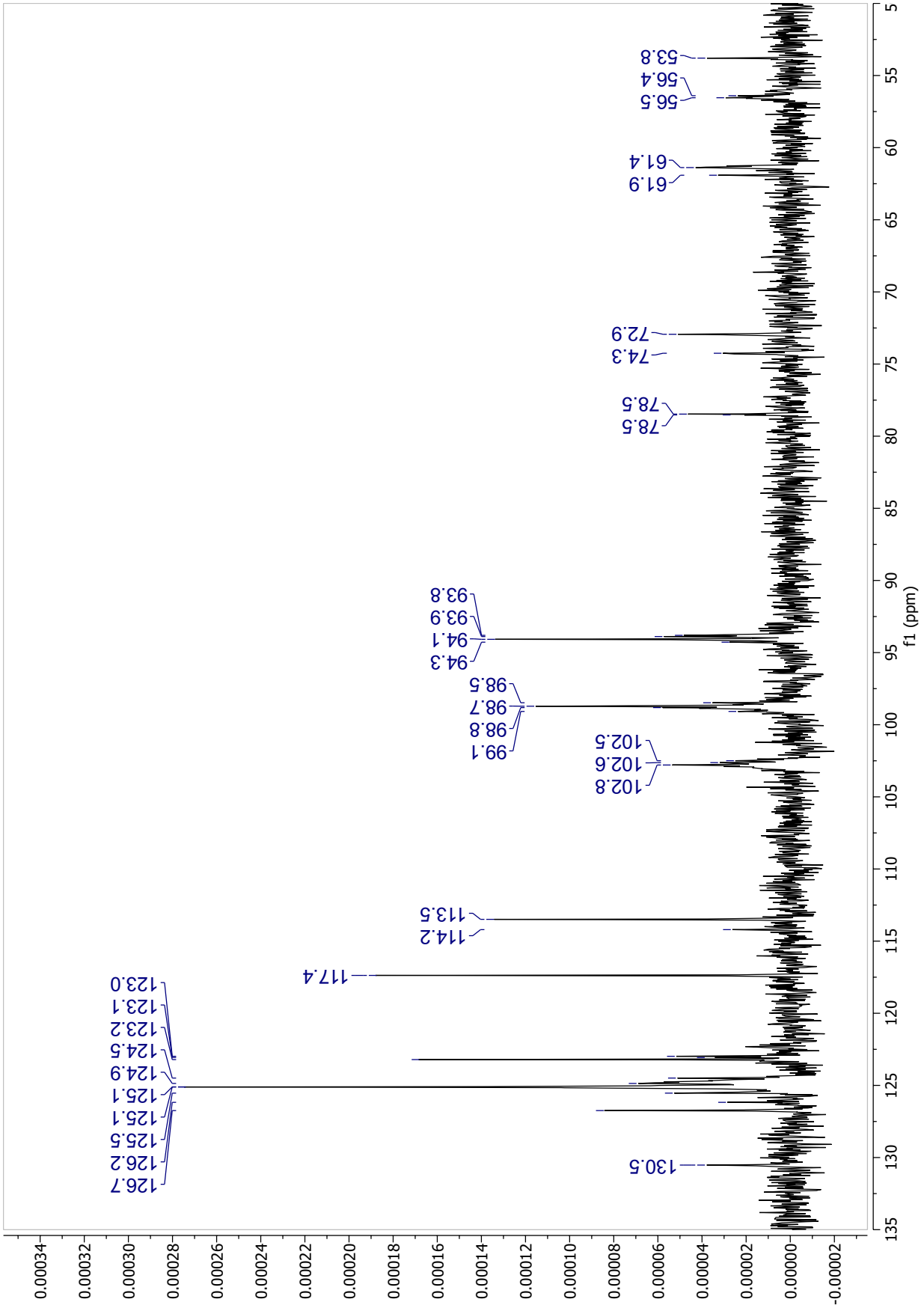


Figure S135. DEPT-135 spectrum (Zoom2: 50-135 ppm) (2000 scans) of F12 (39.0 mg) recorded in acetone-*d6*

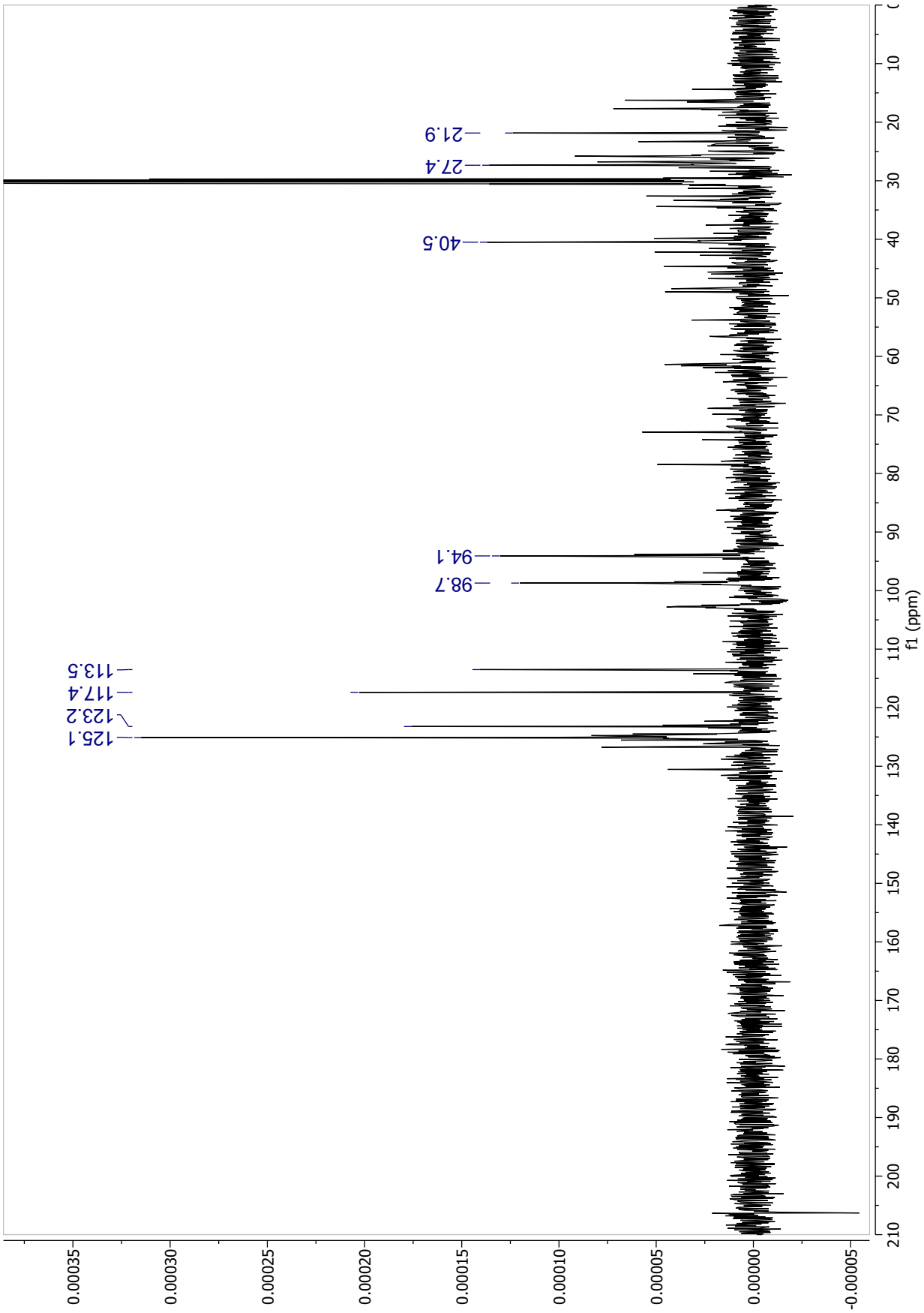


Figure S136. DEPT-90 spectrum (2000 scans) of F12 (39.0 mg) recorded in acetone-*d*6

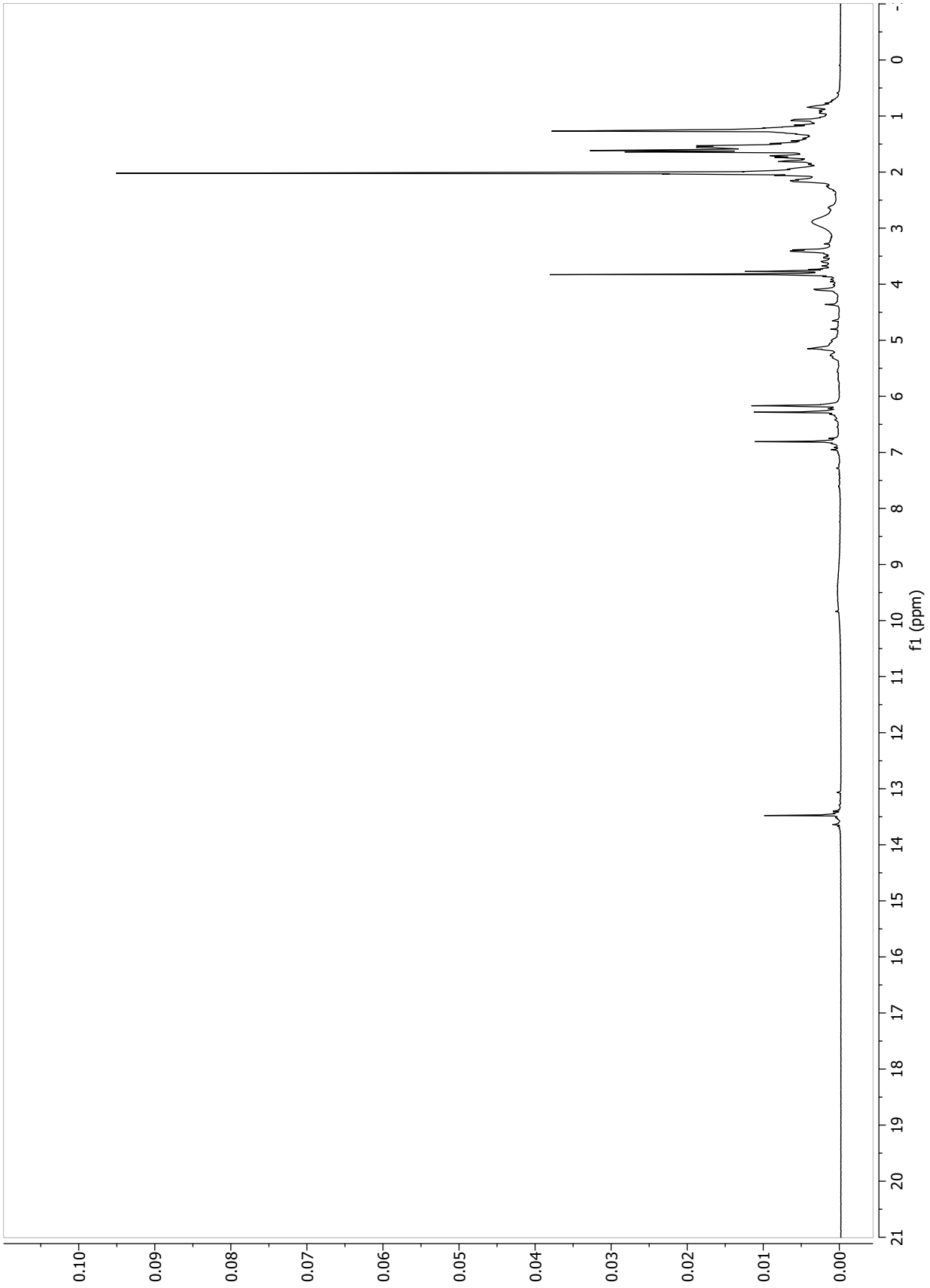


Figure S137. ^{1}H spectrum of F12 (39.0 mg) recorded in acetone-*d*6

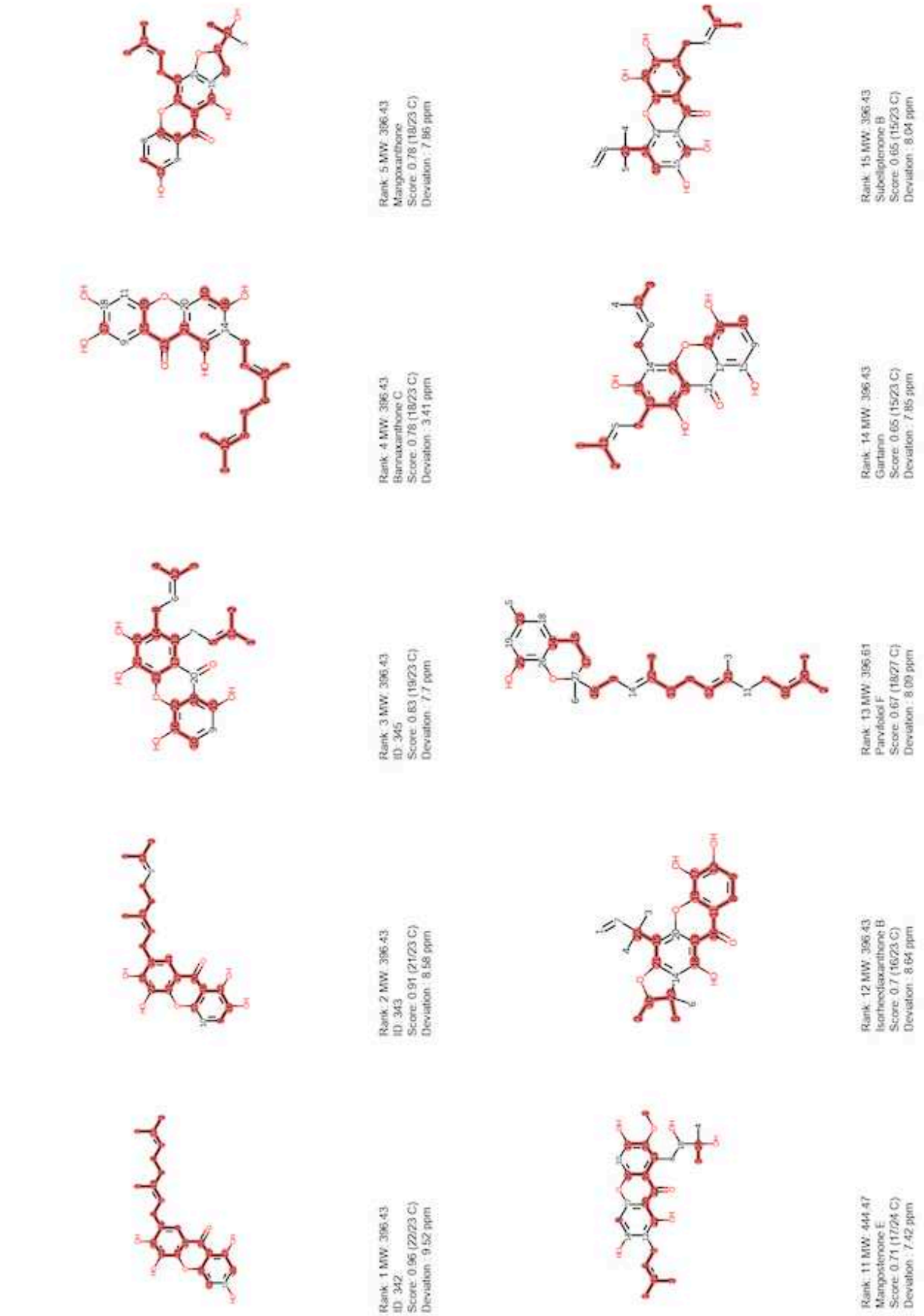


Figure S138. Part of displayed results for the ^{13}C-NMR dereplication (+ DEPT-90 and -135) of F12 using c-type *Garcinia* DB. Equivalent carbons were allowed, and a molecular weight filter was applied (396 and 444 g.mol^{-1}).

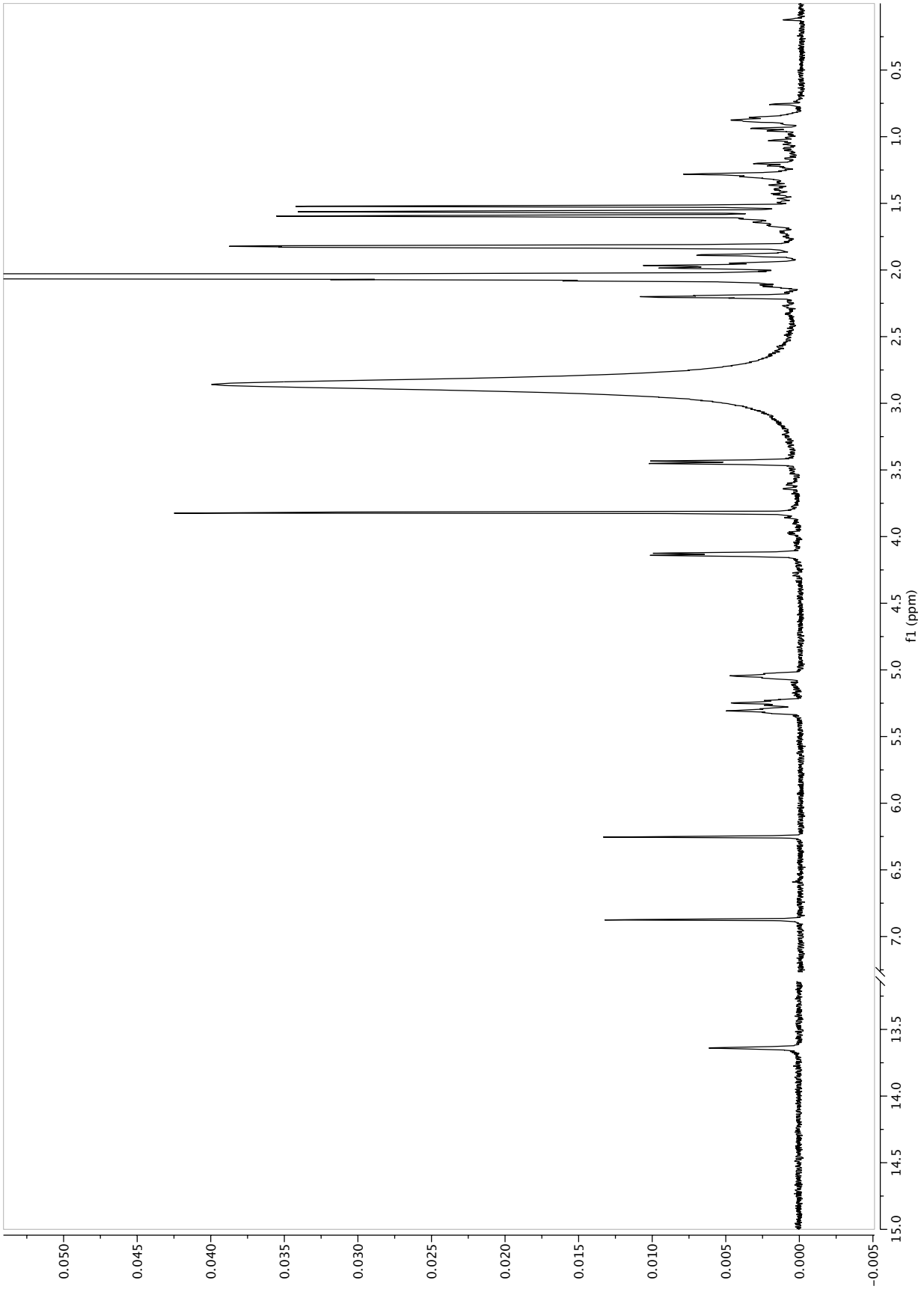


Figure S139. ^{1}H spectrum of isocowanin (11.5 mg) from F3 recorded in acetone-*d*6

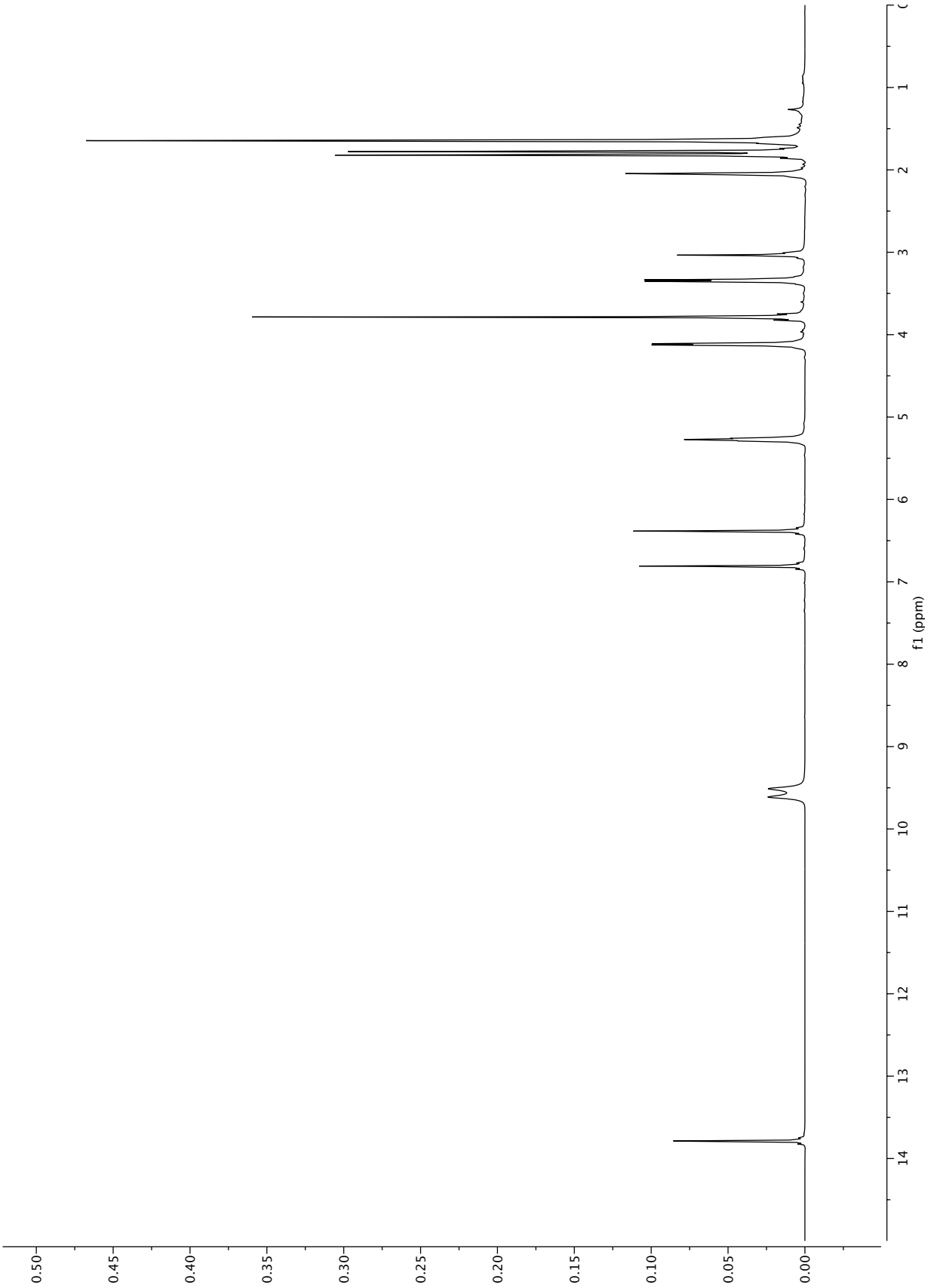


Figure S140. ^{1}H spectrum of α-mangostin (20.0 mg) from F4 recorded in acetone-*d*6

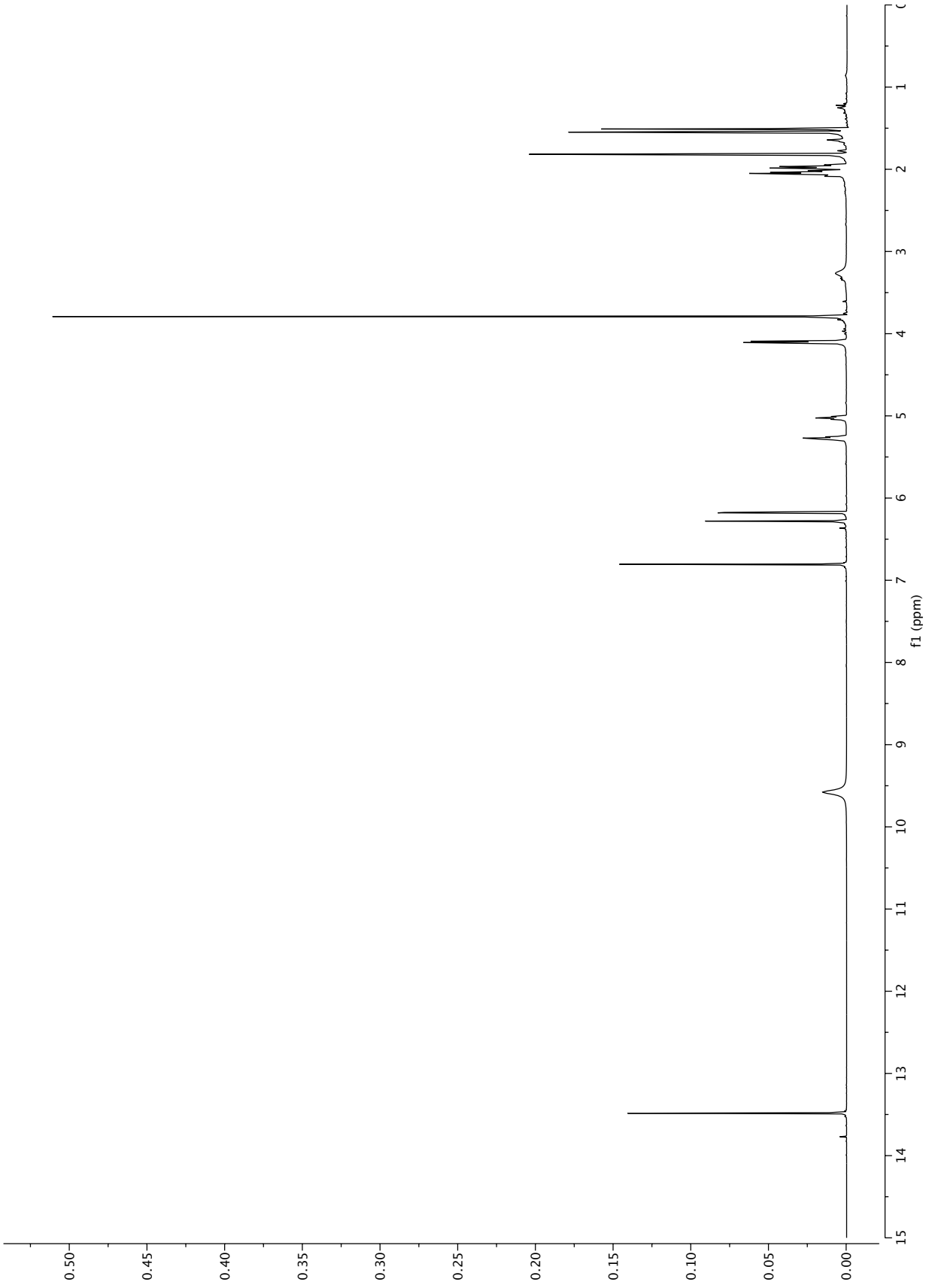


Figure S141. ^{1}H spectrum of rubraxanthone (50.0 mg) from F6 recorded in acetone-*d*6

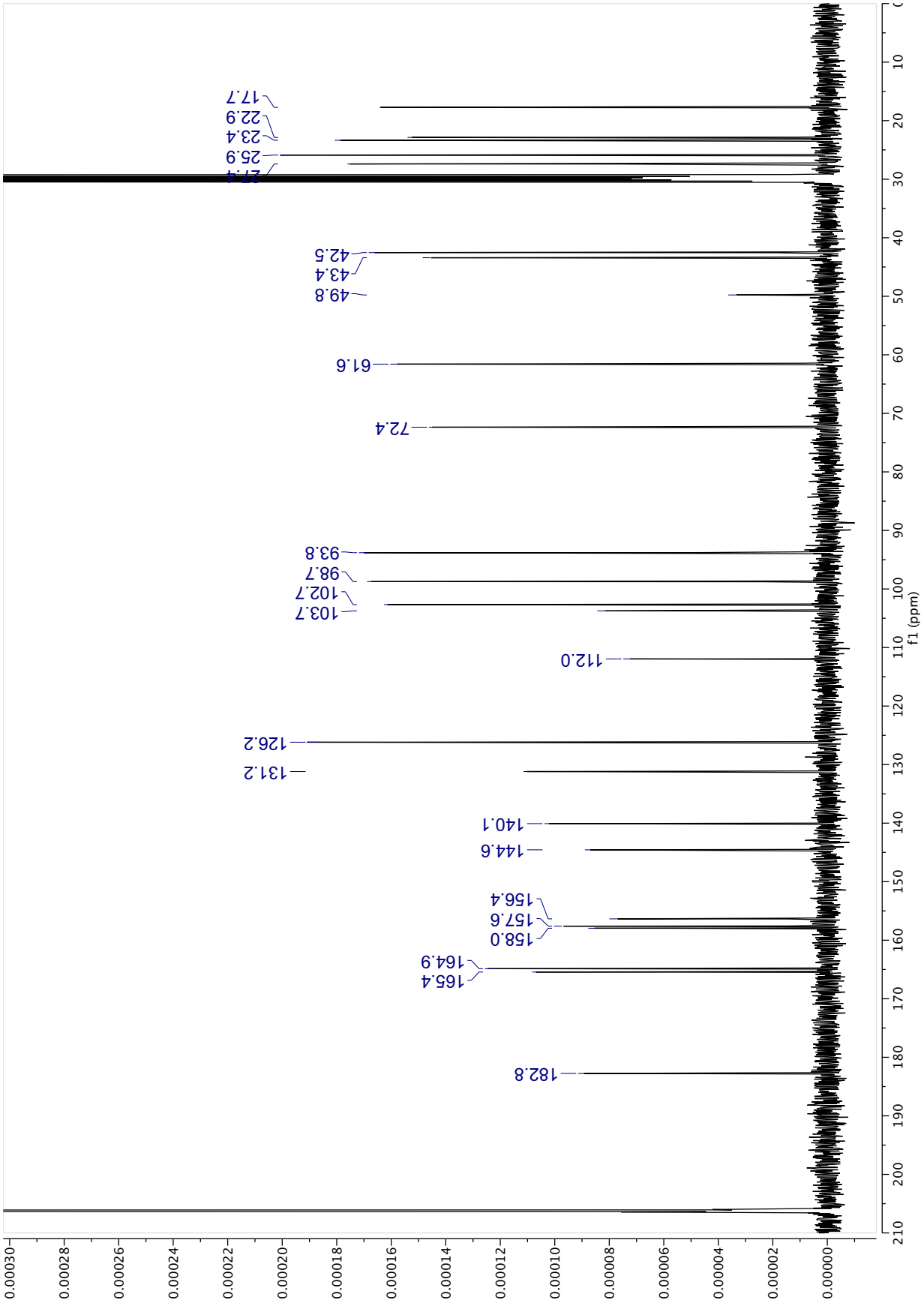


Figure S142. ^{13}C spectrum (2500 scans) of parvixanthone G from F10D (15.4 mg) recorded in acetone-*d6*.

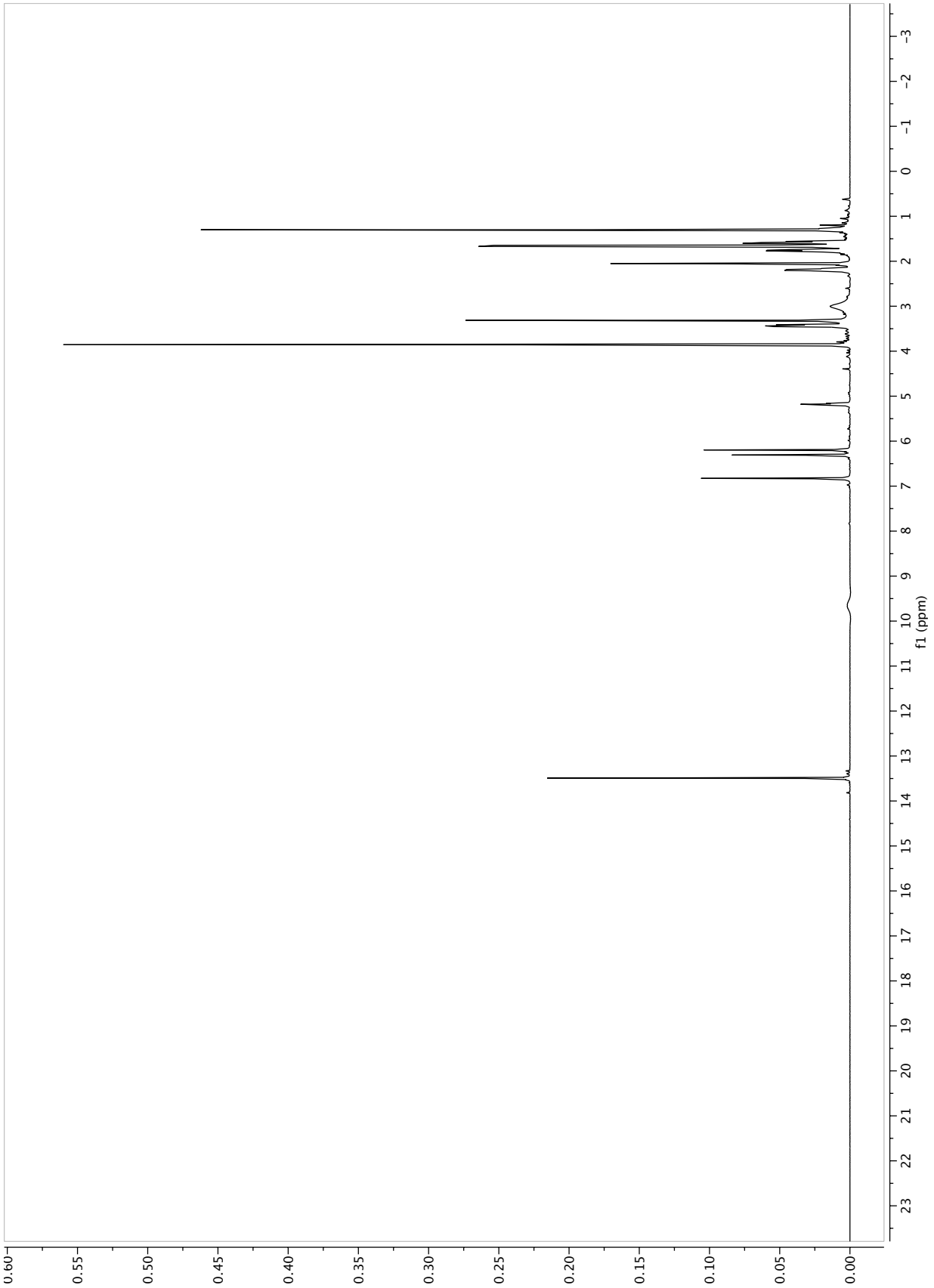


Figure S143. ^{1}H spectrum of parvixanthone G from F10D (15.4 mg) recorded in acetone-*d*6.

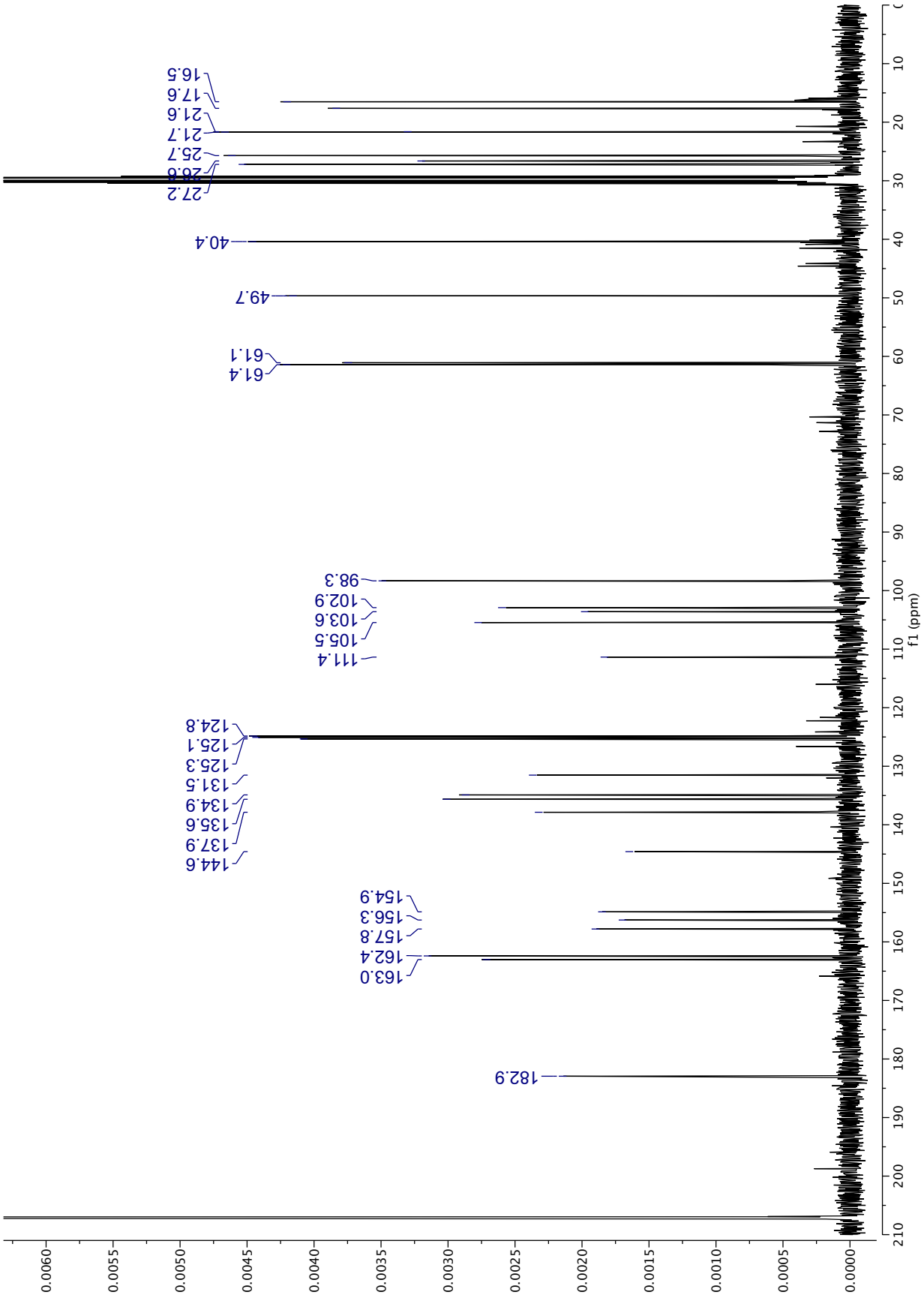


Figure S144. ^{13}C spectrum (900 scans) of isocowanol from F10H (40.0 mg) recorded in acetone-*d*6.

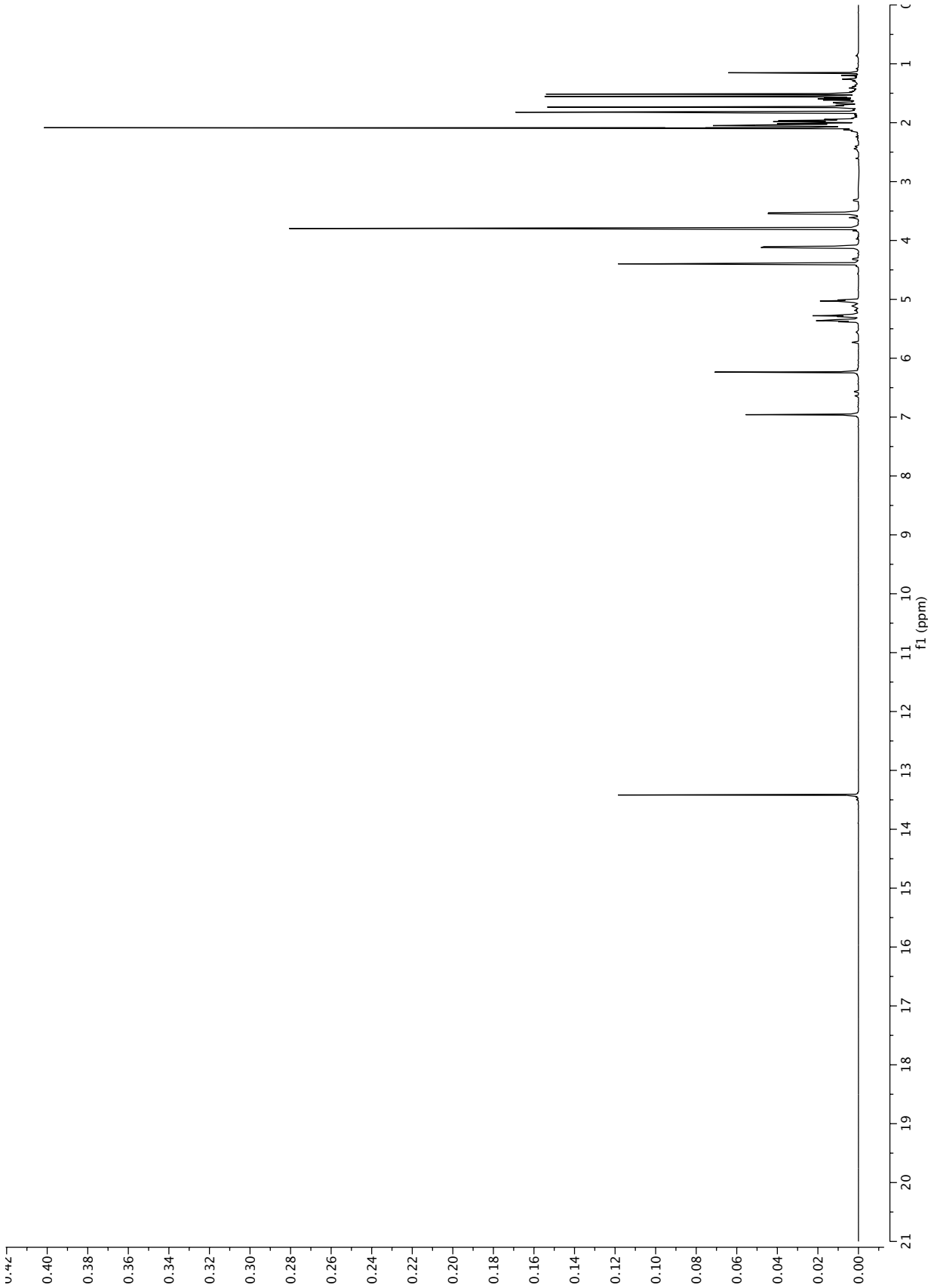


Figure S145. ^{1}H spectrum of isocowanol from F10H (40.0 mg) recorded in acetone-*d*6.

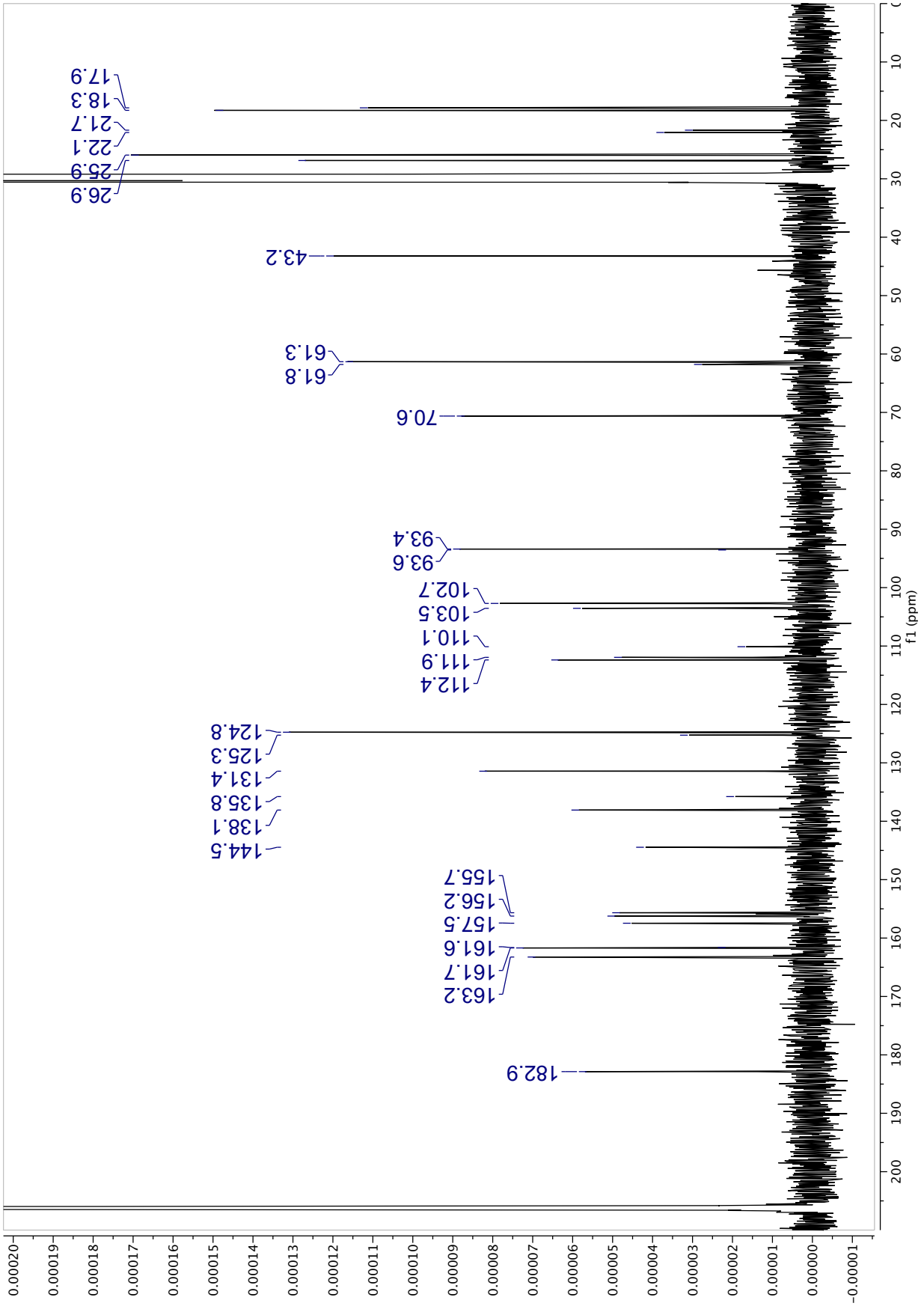


Figure S146. ^{13}C spectrum (44064 scans) of cratoxylone from F10F (40.0 mg) recorded in acetone-*d*6.

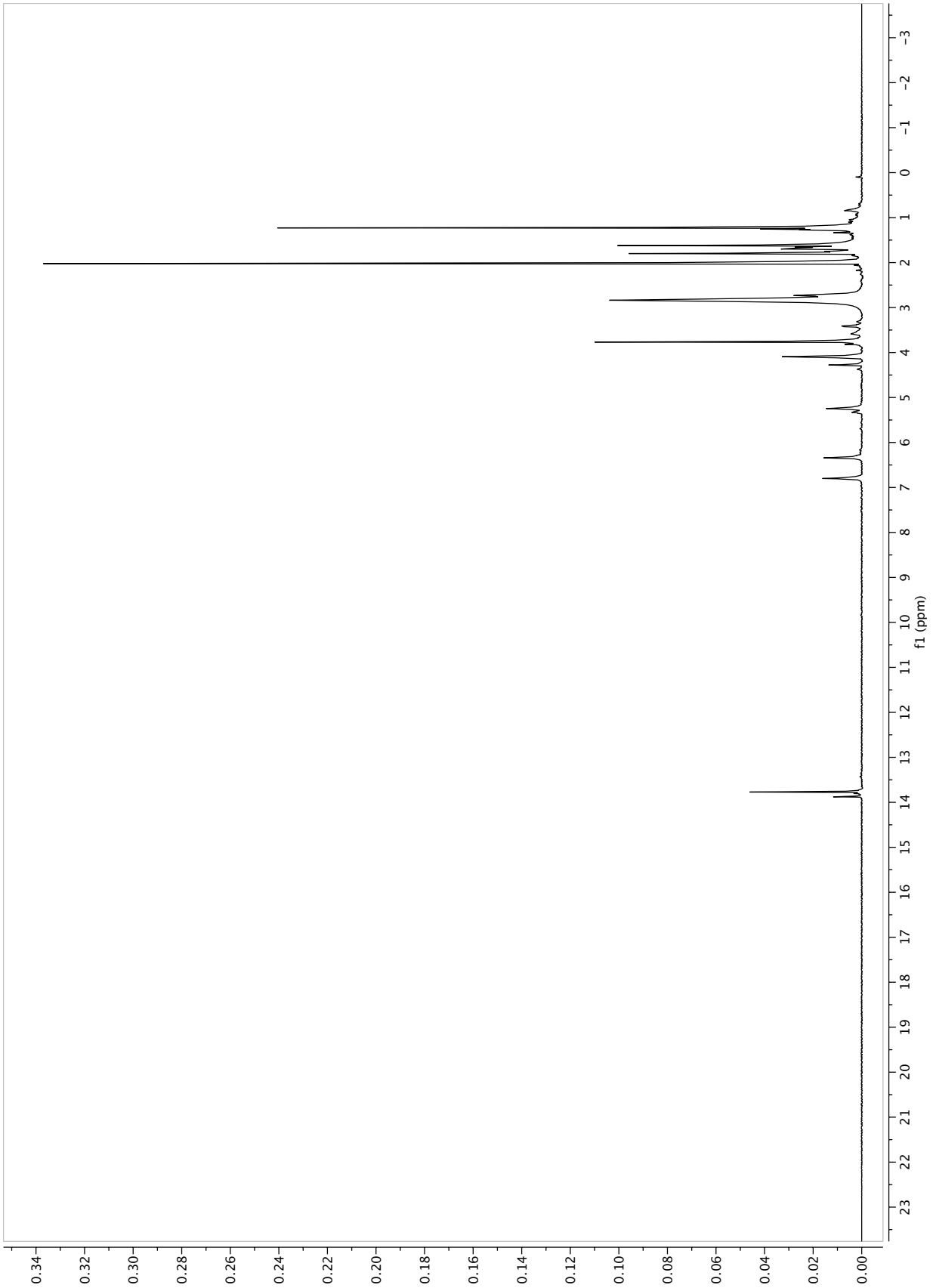


Figure S147. ^{1}H spectrum of cratoxylone from F10F (40.0 mg) recorded in acetone-*d*6.

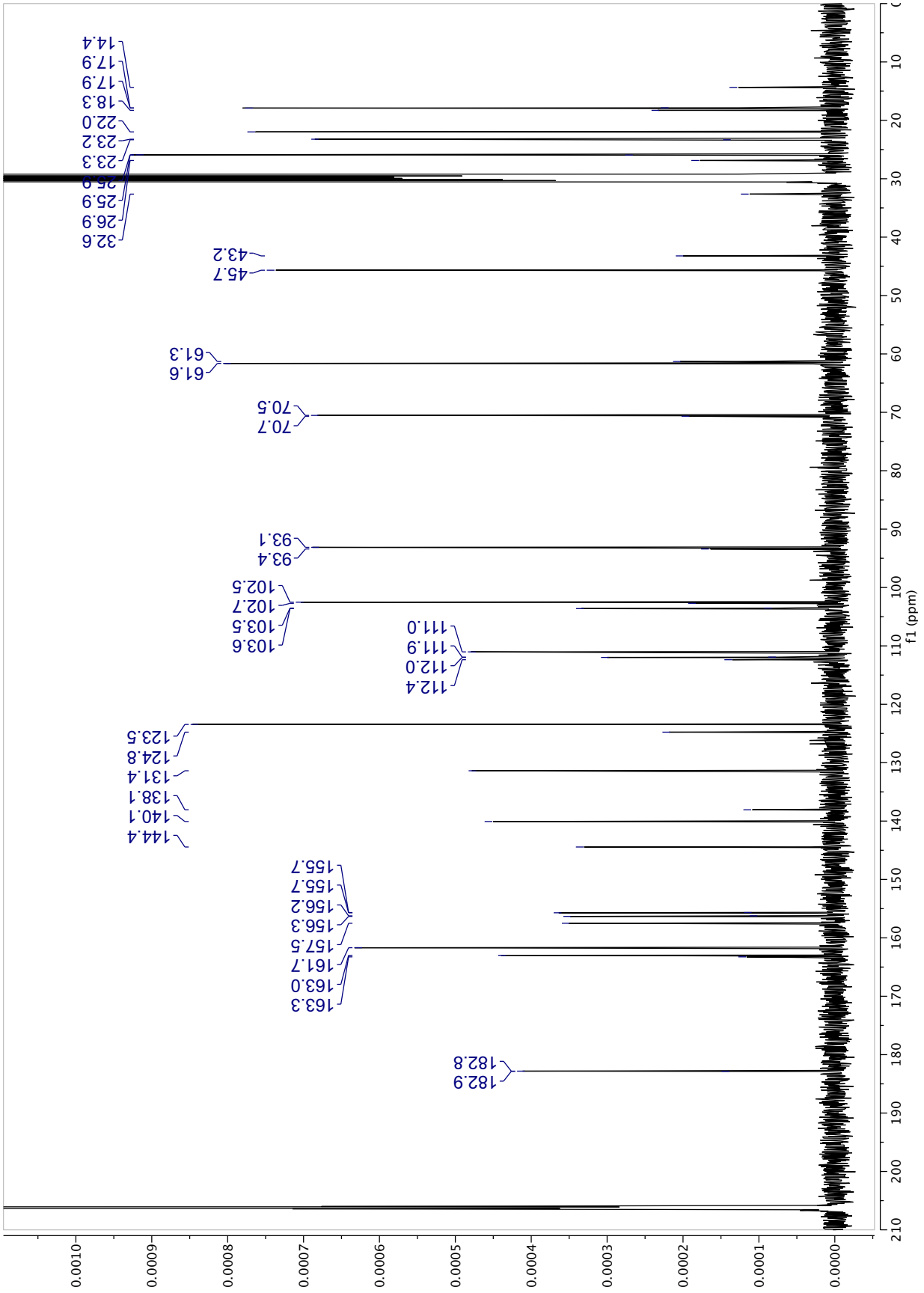


Figure S148. ^{13}C spectrum (17842 scans) of garcinone D from F10E (in mixture with cratoxylone) (8.1 mg) recorded in acetone-*d*6.

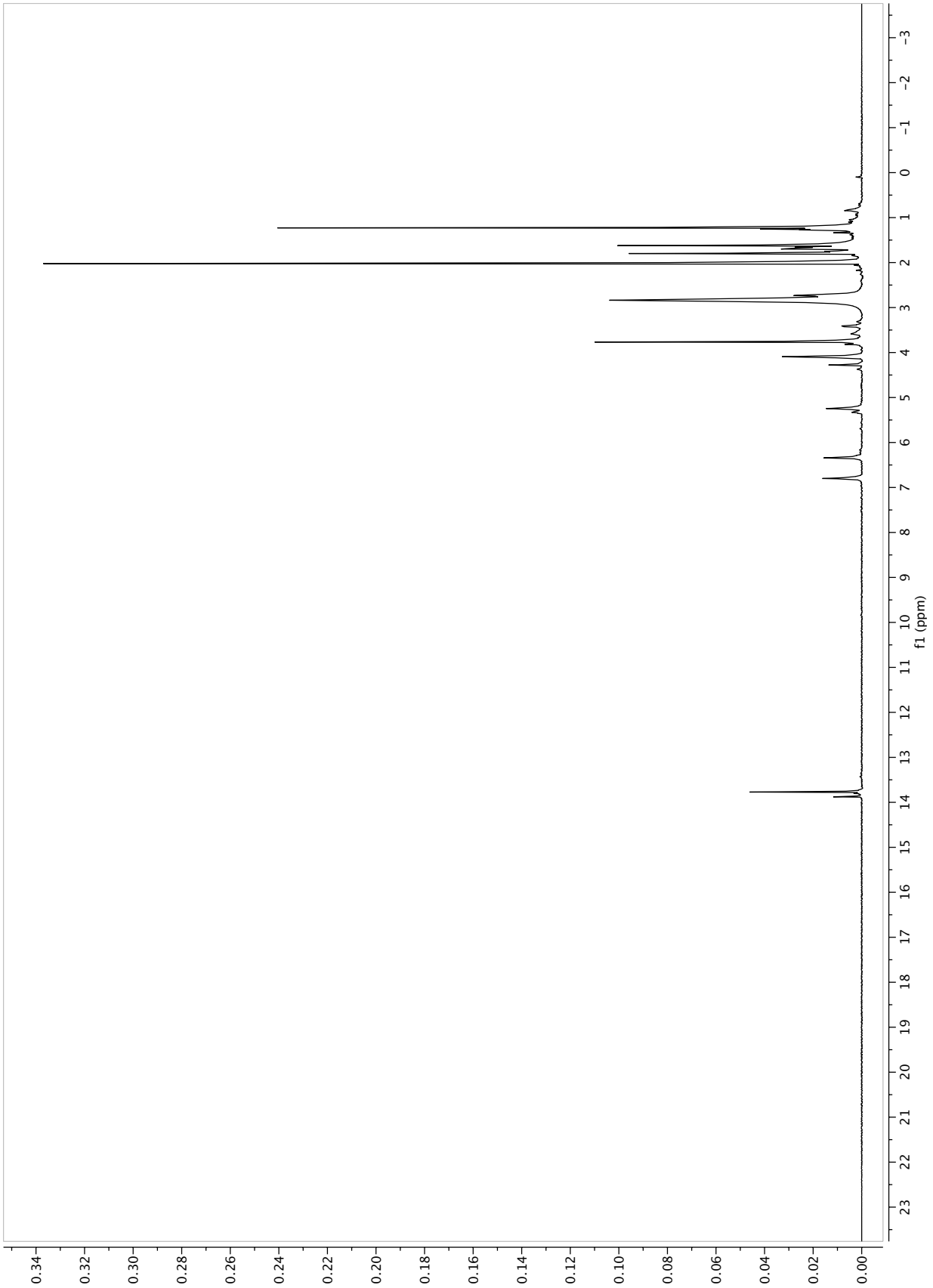


Figure S149. ^{1}H spectrum of garcinone D from F10E (in mixture with cratoxylone) (8.1 mg) recorded in acetone-*d*6.

Samples	IC_{50} results (µg/mL)	Average ± SD
*Aminoguanidine**	135.0 133.0 132.0 163.0 158.0 152.0	145.5 ± 12.6
F1	> 1000.0 745.0 > 1000.0	915.0 ± 120.2
F2	320.0 333.0 367.0	340.0 ± 19.8
F3	277.0 273.0 242.0	264.0 ± 15.6
F4	584.0 659.0 603.0	615.3 ± 31.8
F5	138.0 146.0 152.0	145.3 ± 5.7
F6	64.0 59.0 51.0	58.0 ± 5.4
F7	119.0 90.0 105.0	104.7 ± 11.8
F8	119.0 211.0 141.0	157.0 ± 39.2
F9	108.0 103.0 93.0	101.3 ± 6.2
F10	63.0 85.0 99.0	82.3 ± 14.8
F11	153.0 185.0 147.0	161.7 ± 16.7
F12	145.0 133.0 94.0	124.0 ± 21.8
F13	102.0 162.0 177.0	147.0 ± 32.4
F14	164.0 201.0 170.0	178.3 ± 16.2

Table S10. Anti-AGEs IC_{50} of KL5670 fractions tested in triplicate. Aminoguanidine corresponds to the reference compound.

Name	Retention time (min)	m/z	VIP score	RC
	0.95	428.1134	0.062	-1.630E-05
	1.04	573.1044	0.594	-1.699E-04
	1.36	425.1609	0.725	-3.296E-04
	1.36	570.0839	0.411	-1.836E-04
	1.71	329.2334	0.184	-8.969E-05
	1.71	425.1607	0.557	-1.917E-04
	1.75	417.2132	0.114	-1.514E-05
	1.85	318.0409	0.339	-9.312E-05
	2.00	315.0510	0.716	-2.727E-04
	2.00	431.2286	0.317	4.520E-05
	2.00	530.1434	0.244	4.520E-05
	2.12	431.2286	0.246	5.082E-05
	2.15	359.1135	0.244	-1.042E-04
	2.19	443.1698	0.673	-2.138E-04
	2.19	445.1866	0.952	-4.301E-04
	2.27	427.1759	0.317	-1.402E-04
	2.63	459.1655	0.424	-8.715E-05
1	2.66	443.1709	1.443	-2.511E-04
	2.73	570.0829	0.348	-1.216E-04
	2.98	443.1699	0.333	-1.377E-04
	3.15	425.1598	0.705	-9.946E-05
	3.23	443.1696	0.409	-1.646E-04
	3.23	457.1472	0.383	-1.739E-04
	3.23	509.2182	0.260	-1.085E-04
3	3.25	533.3487	1.160	-5.547E-04
	3.25	678.2709	0.834	-3.987E-04
	3.25	411.1443	0.637	-1.890E-04
2	3.25	487.3430	1.184	-5.615E-04
	3.33	441.1551	0.992	-4.195E-04
	3.40	587.3074	0.622	6.717E-05
	3.43	443.1705	0.430	-6.480E-05
	3.55	573.3285	0.380	-5.973E-05
	3.57	327.0873	0.623	-1.352E-04
	3.65	357.0982	0.505	-2.415E-04
	3.72	601.3238	0.982	1.004E-04
4	3.75	425.1605	1.482	-1.062E-04
	3.75	443.1705	0.722	-2.359E-04
	3.77	509.2181	0.540	-2.680E-04
5	3.92	427.1756	1.196	-1.170E-04
	3.92	479.3371	0.458	-1.786E-04
	3.92	509.2181	0.981	-3.708E-04
	3.94	413.1600	0.976	-1.907E-04
6	3.97	425.1605	1.950	-1.794E-04
7	3.97	570.0841	1.153	-1.600E-04

Name	Retention time (min)	m/z	VIP score	RC
8	4.09	425.1606	1.293	-2.103E-04
	4.09	616.0899	0.756	-2.846E-04
	4.12	525.2126	0.894	-1.817E-04
	4.14	409.1295	0.791	-2.350E-04
	4.14	441.1554	0.920	-3.251E-04
	4.22	411.1452	0.650	-1.858E-04
	4.29	425.1597	0.449	-5.092E-05
	4.34	315.2542	0.105	4.407E-05
	4.34	413.1602	0.972	-2.417E-04
	4.39	409.1647	0.812	-4.292E-04
9	4.41	341.1033	1.237	-4.034E-04
	4.44	891.3374	0.048	-2.994E-06
10	4.44	427.1758	2.301	-2.159E-04
	4.46	425.1600	0.654	-5.612E-05
	4.46	511.2339	0.818	-3.632E-04
	4.56	477.3219	0.648	-1.010E-04
	4.56	530.2337	0.282	-2.207E-05
	4.59	383.1152	0.557	-2.978E-05
	4.61	425.1596	0.608	-2.385E-04
	4.61	315.2541	0.186	-8.525E-05
	4.61	423.1452	0.801	-1.936E-04
11	4.66	427.1763	1.641	-2.681E-04
	4.71	651.1511	0.989	-1.506E-04
12	4.78	427.1764	1.010	-1.668E-04
	4.83	425.1598	0.657	-9.226E-05
	4.86	395.1503	0.963	-2.142E-04
	4.88	477.3219	0.568	-2.426E-04
	4.88	530.2335	0.297	-1.288E-04
	4.91	525.2127	0.405	-1.119E-04
	4.93	471.3473	0.727	-3.473E-04
	4.93	441.1550	0.333	-7.255E-05
	5.08	425.1599	0.687	-1.029E-04
	5.16	943.7034	0.360	-1.585E-04
	5.20	423.1452	0.625	-1.537E-04
	5.20	662.2763	0.949	-3.629E-04
	5.25	455.1705	0.326	-1.050E-04
	5.28	395.1498	0.711	-2.484E-04
	5.28	559.3846	0.050	-2.433E-05
13	5.33	425.1608	1.502	-4.539E-04
	5.33	479.2075	0.637	-1.484E-04
14	5.40	395.1497	1.085	-1.924E-04
	5.43	495.2350	0.276	-6.230E-05
15	5.43	425.1606	1.222	-4.216E-04
	5.53	383.1875	0.975	-9.445E-05

Name	Retention time (min)	m/z	VIP score	RC
	5.53	451.1757	0.351	-1.722E-04
16	5.55	395.1498	2.029	-4.166E-04
17	5.60	425.1605	1.292	-1.446E-04
	5.70	423.1455	0.800	-2.458E-04
	5.72	427.1753	0.872	-4.149E-04
	5.75	509.2181	0.570	-1.579E-04
	5.77	525.0855	0.343	-8.048E-05
18	5.77	425.1606	1.078	-1.166E-04
	5.79	427.2850	0.426	-1.964E-04
	5.89	379.1562	0.558	5.222E-05
	5.97	371.2233	0.782	-3.950E-04
	5.97	743.4530	0.667	-3.294E-04
	5.97	327.2331	0.509	-2.575E-04
	5.99	393.1354	0.338	-1.902E-05
	5.99	459.3117	0.525	-1.386E-05
19	6.11	395.1483	1.038	-2.235E-04
	6.18	439.1361	0.371	-1.118E-04
	6.21	409.1298	0.785	-5.615E-05
	6.21	459.3115	0.745	4.297E-05
	6.23	425.1603	0.969	-1.226E-04
20	6.23	493.2232	3.397	-4.487E-04
24	6.28	509.0914	1.305	-3.178E-04
	6.28	477.1535	0.722	-1.607E-04
21	6.28	475.3069	1.824	5.804E-04
	6.28	855.3151	0.124	-2.300E-05
23	6.28	409.1656	4.445	-1.062E-03
22	6.28	325.0721	1.519	-3.670E-04
	6.28	423.1477	0.302	-1.298E-04
	6.33	833.3160	0.804	-2.496E-04
	6.36	439.1361	0.442	-1.287E-04
	6.36	507.2025	0.458	-1.511E-04
	6.41	459.3119	0.675	1.680E-04
25	6.43	475.3067	3.303	1.060E-03
	6.45	421.1924	0.254	-9.111E-05
	6.55	461.3257	0.338	-5.974E-05
	6.57	379.1561	0.813	-3.190E-04
26	6.57	475.3065	1.860	5.957E-04
	6.60	495.2381	0.347	-4.518E-05
	6.67	369.1011	0.584	-3.004E-04
	6.72	365.1038	0.806	-3.244E-04
	6.77	561.2841	0.362	-1.921E-04
	6.79	577.2793	0.365	-1.926E-04
	6.98	509.0912	0.875	9.809E-05
27	6.98	409.1641	2.909	3.660E-04

Name	Retention time (min)	m/z	VIP score	RC
	6.98	445.1523	0.077	3.447E-06
	6.98	351.0878	0.670	7.441E-05
	6.98	391.1199	0.046	4.142E-06
	7.08	507.2027	0.420	-1.953E-04
	7.08	833.3162	0.380	-5.335E-05
28	7.17	475.3066	1.420	4.550E-04
	7.22	561.2848	0.361	-1.331E-04
	7.29	491.2074	0.707	-1.376E-04
	7.78	393.1379	0.396	-1.440E-04
29	7.78	477.2281	2.103	-7.857E-04
	7.87	397.2266	0.419	-9.877E-05
	8.15	833.3167	0.617	-5.164E-05
	8.19	423.1806	0.377	-1.711E-04
	8.26	441.2527	0.196	3.397E-05
30	8.33	407.1509	1.854	-5.587E-04
	8.33	507.0759	0.658	-2.008E-04
	8.44	833.3171	0.579	-4.393E-05
	8.67	433.1655	0.304	-1.406E-04
	8.78	477.2282	0.300	-8.783E-05
	9.30	355.1583	0.699	-2.729E-04
	9.49	545.2908	0.566	-2.542E-04
	9.55	381.1739	0.619	-1.382E-05

Table S11. The VIP scores and the regression coefficients of the PLS model applied to fractions of KL5670.

Rank	Compound (RT_MW)	Adducts	Formula	Score	Fragmentation Score	Mass Error (ppm)	Neutral mass (Da)	m/z	Structure
1	11.41_409.1658m/z	M-H	C24H26O6	49.8	62.7	0.44	/	409.1658	Hypothesis a
2	11.41_409.1658m/z	M-H	C24H26O6	48.5	55.9	0.44	/	409.1658	Hypothesis b
3	11.41_409.1658m/z	M-H	C24H26O6	48.4	55.7	0.44	/	409.1658	Hypothesis c
4	11.41_409.1658m/z	M-H	C24H26O6	48.4	55.5	0.44	/	409.1658	Hypothesis d
5	11.41_409.1658m/z	M-H	C24H26O6	48.3	55.1	0.44	/	409.1658	Hypothesis e
6	**11.41_409.1658m/z**	**M-H**	**C24H26O6**	**48.3**	**55.1**	**0.44**	**/**	**409.1658**	Rubraxanthone (**5**)
7	11.41_409.1658m/z	M-H	C24H26O6	48.3	55.0	0.44	/	409.1658	Hypothesis g
8	11.41_409.1658m/z	M-H2O-H	C24H28O7	47.5	50.8	0.43	/	409.1658	Hypothesis h

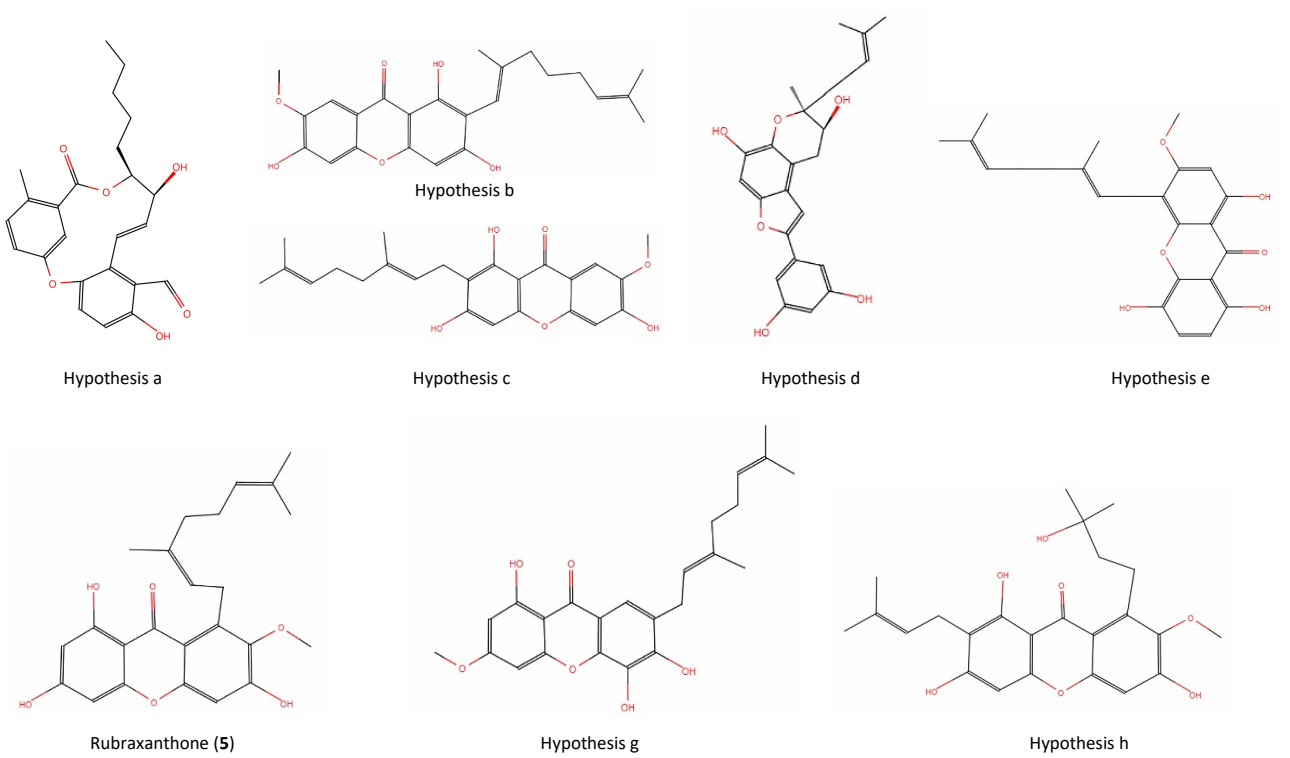
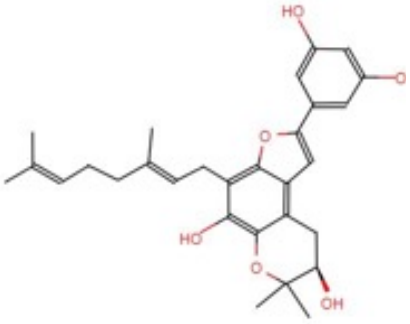
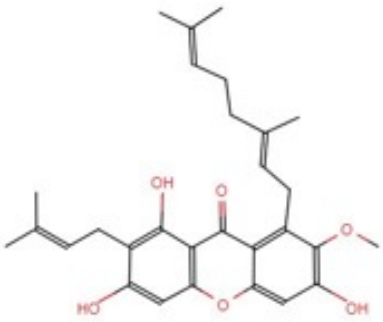


Table S12. Part of structures predicted for the ion at 409.1658 m/z (node **23**). The dereplication process is based on a ProgenesisQI in-silico fragmentation using the PNMRNP database (precursor fragmentation was set at 5 ppm and theoretical fragmentation was set at 10 ppm).

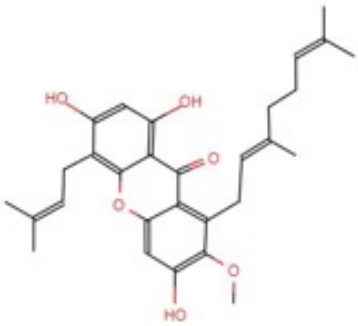
Rank	Compound (RT_MWn)	Adducts	Formula	Score	Fragmentation Score	Mass Error (ppm)	Neutral mass (Da)	m/z	Structure
1	7.78_478.2354n	M-H, 2M-H	C29H34O6	50.6	54.7	-0.34	478.2354	477.2281	Hypothesis i
2	7.78_478.2354n	M-H, 2M-H	C29H34O6	50.5	53.9	-0.34	478.2354	477.2281	Hypothesis j
3	**7.78_478.2354n**	**M-H, 2M-H**	**C29H34O6**	**50.5**	**54.1**	**-0.34**	**478.2354**	**477.2281**	isocowanin (**23**)
4	7.78_478.2354n	M-H, 2M-H	C29H34O6	50.5	53.9	-0.34	478.2354	477.2281	Hypothesis k
5	7.78_478.2354n	M-H, 2M-H	C29H34O6	50.5	54.1	-0.34	478.2354	477.2281	Hypothesis l



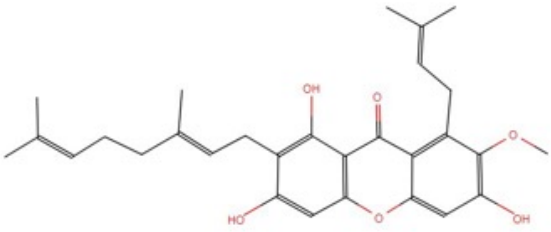
hypothesis i



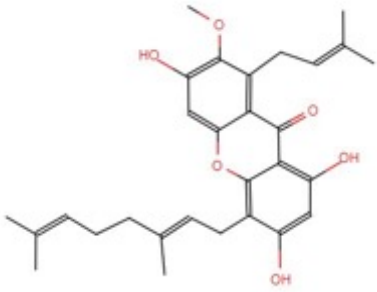
hypothesis j



Isocowanin (**29**)



hypothesis k



hypothesis l

Table S13. Part of structures proposition of the ion at 477.2281 m/z (node **29**). The proposition is based on a ProgenesisQI in-silico fragmentation using the PNMRNP database (precursor fragmentation was set at 5 ppm and theoretical fragmentation was set at 10 ppm).

Carbon numbering	Literature data	Experimental data	Dereplication by MixONat	
	δc (ppm)	δc in acetone-*d*6	δc (ppm) *Garcinia*DB	δc (ppm) matched found by MixONat
9	182.8	182.7	182.8	182.7
3	165.5	165.3	165.1	165.3
1	165.0	164.9	165.1	164.9
4a	158.1	157.9	157.9	157.9
6	157.6	157.4	154.1	/
10a	156.4	156.2	155.4	156.2
7	144.7	144.5	142.5	/
8	138.4	138.2	136.5	/
18	135.2	135.1	135.5	135.1
13	131.6	131.5	131.4	131.5
17	125.3	125.1	125.2	125.1
12	124.9	124.8	124.2	124.8
8a	112.2	111.9	112.3	111.9
9a	103.9	103.7	104.2	103.7
5	102.9	102.8	101.6	102.8
2	98.9	98.7	98.5	98.7
4	93.9	93.8	93.8	93.8
7-OMe	61.5	61.4	62.0	61.4
14	40.5	40.4	40.5	40.4
16	27.4	27.3	27.4	27.3
11	26.9	26.8	26.8	26.8
19	25.8	25.8	25.7	25.8
20	17.8	17.7	17.6	17.7
15	16.7	16.6	16.6	16.6

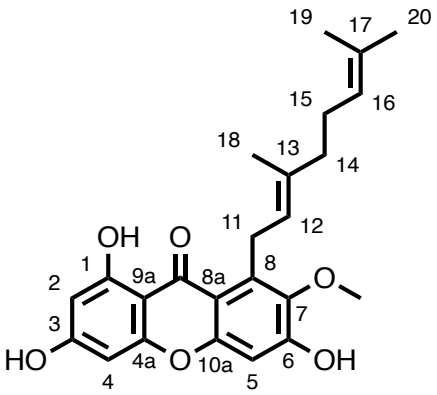


Table S14. Comparison of experimental δ_C of F6 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$). The comparison of experimental data with the confirmed the identification of **rubraxanthone** (**23**) in F6.[7]

[7]Munekazu, I. *et al.*, *Chem. Pharm. Bull.* **1995**, *44* (1), 232–234.

Carbon numbering	Literature data Isocowanin	Experimental data	Literature data Parvifolixanthone C	Experimental data	Dereplication by MixONat	
	δc (ppm)	δ_C matched in F3 (ppm) in acetone-*d*6 (Putative isocowanin)	δc (ppm)	δ_C matched in F3 (ppm) in $CDCl_3$ (Putative parvifolixanthone C)	δc (ppm) *Garcinia*DB	δc (ppm) matched found by MixONat when using δc in acetone-d6
9	183.1	183.0	182.4	182.4	184.5	183.0
1	162.7	162.7	161.5	161.5	162.7	162.7
3	162.5	162.4	160.9	161.2	162.3	162.4
4a	159.5	155.0	154.0	154.0	155.4	155.0
5a	156.4	156.4	154.6	154.8	153.0	/
6	155.1	155.0	155.6	155.9	154.1	154.1
7	144.6	144.5	142.1	142.8	142.7	/
8	138.2	138.2	137.0	137.3	136.6	136.4
3'	135.1	135.0	135.4	135.6	134.3	135.0
7'	131.5	131.6	124.3	124.3	131.9	132.1
3"	131.4	131.5	134.4	135.7	131.3	131.6
6'	125.2	125.2	124.1	124.0	123.6	123.4
2'	124.1	124.1	122.8	122.7	123.4	124.9
2"	123.5	123.4	121.1	120.8	123.3	122.3
8a	111.8	111.7	111.7	112.1	112.2	111.7
4	106.4	106.3	104.1	104.8	106.4	106.3
9a	104.0	103.8	103.4	104.0	104.7	103.8
5	102.9	102.8	101.5	101.8	102.5	102.8
2	98.3	98.3	98.5	98.4	97.9	98.3
7-OMe	61.5	61.4	61.8	62.1	62.0	/
4'	40.4	40.5	38.6	39.7	32.8	/
1'	27.3	27.3	25.8	25.8	26.2	25.9
5'	26.8	26.7	26.2	25.9	26.6	/
8'	25.8	25.8	25.7	25.7	25.7	27.1
4"	25.7	25.7	26.2	25.9	25.9	26.7
1"	22.1	22.1	26.2	25.9	22.3	/
3'-Me	18.1	18.0	17.3	17.8	23.9	23.2
9'	17.7	17.7	16.0	16.2	17.6	/
5"	16.6	16.6	16.0	16.1	18.0	/

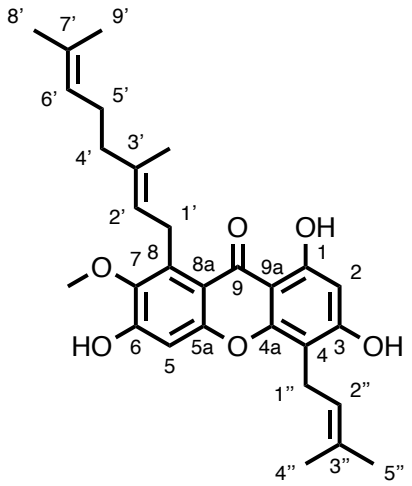


Table S15. Comparison of experimental δ_C of F3 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) (δC-SDF). The comparison of experimental data with the literature[8] confirmed the identification of **isocowanin (29)** in F3. The hypothesis of parvifolixanthone C can be excluded after comparison with previously reported δ_C in $CDCl_3$.[9,10]

[8] Ampofo, S. A. *et al.*, *Phytochemistry* **1986**, *25* (10), 2351-2355 in acetone-d6
[9] Rukachaisiriku, V. *et al. Tetrahedron* **2006**, *62*, 8578–8585 in $CDCl_3$: Errors in the chemical shift values are manifest for the compound described as parvifolixanthone C in this publication.
[10] Mohamed, G.A. *et al. Fitoterapia* **2014**, *98*, 215-221 in $CDCl_3$ (supporting information)

Carbon numbering	Literature data	Experimental data	Dereplication by MixONat	
	δc (ppm)	δ_C matched in F3 (ppm) in $CDCl_3$	δc (ppm) GarciniaDB	δc (ppm) in acetone-*d*6 matched found by MixONat
1	199.0	199.1	199.2	198.1
3	164.1	163.6	160.3	160.3
3'	138.2	138.7	138.2	138.2
7'	135.2	135.7	135.5	135.6
11'	134.8	135.1	134.6	135.0
15'	131.1	131.4	131.3	131.6
2	126.4	126.9	125.6	125.2
6'	124.3	124.3	125.0	125.1
10'	124.1	124.0	124.5	124.9
14'	123.8	123.9	124.5	123.4
2'	120.5	120.8	121.4	122.3
4	73.1	73.8	69.5	/
5	43.5	43.7	41.6	41.6
6	41.1	41.4	40.1	40.5
4'	39.8	39.8	39.9	40.2
8'	39.6	39.7	39.2	/
12'	39.6	39.7	39.2	39.2
1'	30.6	29.9	29.7	30.8
5'	26.6	26.7	27.2	27.3
9'	26.5	26.6	26.9	27.1
13'	26.4	26.5	26.3	26.2
15'Me	25.6	25.7	25.7	25.9
3Me	20.4	20.5	19.8	/
16'	17.6	17.8	17.7	17.7
3'Me	16.1	16.4	16.2	16.6
7'Me	15.9	16.2	16.1	16.1
11'Me	15.9	16.1	16.1	16.1

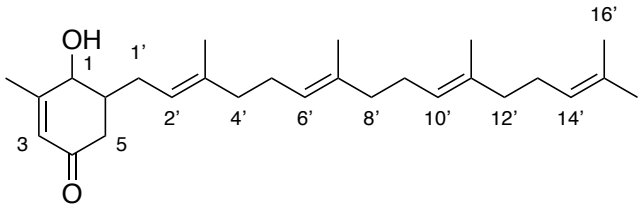


Table S16. Comparison of experimental δ_C of F3 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$). The comparison of experimental data with the literature[11] confirmed the identification of **tetraprenyltoluquinone** (**TPTQ**) in F3.

[11] Wahyuni, F. S. *et al.*, *Aust. J. Chem.* **2004**, *57* (3), 223.

Carbon numbering	Literature data	Experimental data	Dereplication by MixONat	
	δc (ppm)	δ_C matched in F2 (ppm) in $CDCl_3$	$\delta_{C\text{-}SDF}$ (ppm) *Garcinia*DB	δc (ppm) matched found by MixONat
9	181.8	182.1	181.7	182.1
3	159.8	159.9	160.6	159.9
1	157.8	158.0	157.9	158.0
4a	156.1	156.3	157.4	156.9
6	155.6	155.8	154.1	/
5a	154.5	155.1	154.8	155.4
7	142.7	142.9	142.8	143.9
8	136.9	137.2	136.8	137.3
17	131.8	132.2	131.9	130.7
11	126.9	127.2	127.2	127.5
16	123.1	123.3	123.4	123.8
10	115.6	115.8	115.6	115.2
8a	112.1	112.1	112.2	111.1
9a	104.4	104.5	104.6	104.3
2; 12	103.6	103.8	103.8	103.4
5	101.6	102.0	101.5	102.1
4	94.0	94.2	94.9	93.8
13	77.8	78.0	77.3	/
14	28.3	28.4	29.4	28.2
15	26.6	26.6	26.6	27.7
18	25.8	25.9	25.9	25.3
19	18.1	18.3	18.2	18.3
7-OMe	?	61.9	62.0	/

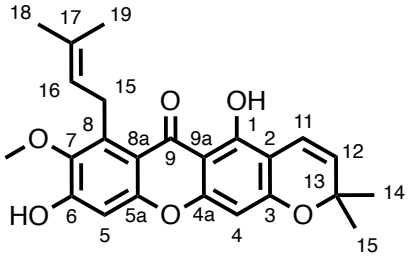


Table S17. Comparison of experimental δ_C of F2 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$). The comparison of experimental data with the literature[12] confirm the identification of **9-hydroxycalabaxanthone (30)** in F2.

[12] Al-Massarani, S. *et al.*, *Molecules* **2013**, *18* (9), 10599–10608.

Carbon numbering	Literature data	Experimental data	Dereplication by MixONat	
	δc (ppm)	δ_C matched in F9 (ppm) in acetone-*d*6	δc (ppm) GarciniaDB	δc (ppm) matched found by MixONat
9	183.2	182.9	183.7	182.9
3	165.8	165.5	165.1	164.8
1	165.0	164.8	163.3	164.1
4a	158.2	157.9	150.3	/
6	157.8	157.6	156.3	156.2
10a	156.3	156.2	154.2	/
7	144.8	144.6	141.6	/
8	138.4	137.6	131.6	131.2
17	136.2	136.2	132.1	132.1
13	131.7	131.6	150.3	131.5
16	125.4	125.3	123.6	124.8
8a	110.4	111.9	114.2	/
20	108.6	109.1	109.3	109.1
9a	103.9	103.6	102.2	102.8
5	103.1	102.9	102.6	102.9
2	98.7	98.7	97.9	98.7
4	93.9	93.8	93.8	93.8
12	75.5	75.6	75.8	75.8
MeO-7	61.0	61.1	62.3	61.7
11	34.7	34.4	35.9	34.4
14	32.7	32.6	31.1	32.5
15	27.8	27.8	26.5	27.0
19	27.8	25.6	25.7	25.7
18	25.8	25.6	17.7	17.7

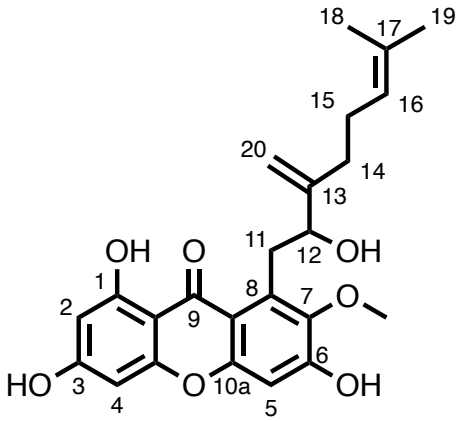
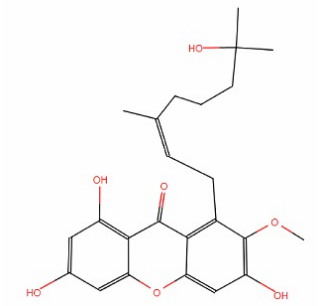


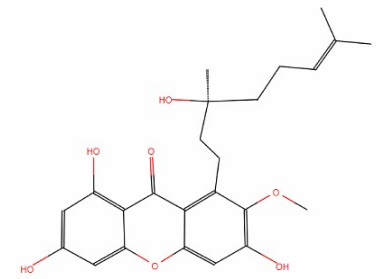
Table S18. Comparison of experimental δ_C of F9 with those of previously described NPs in *Garcinia* species and the predicted shits of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$). The comparison of experimental data with the literature[13] and chemical shifts found by MixONat confirm the identification of **parvixanthone H** (**8**) in F9.

[13] Xu, Y.-J. *et al.*,. *Nat. Prod.* **2001**, *64* (9), 1191–1195.

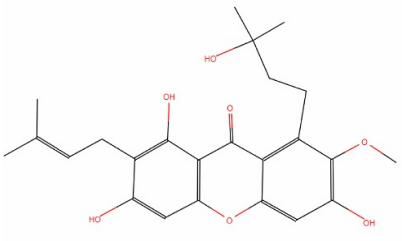
Rank	Compound (RT_MWn)	Adducts	Formula	Score	Fragmentation Score	Mass Error (ppm)	Neutral mass (Da)	m/z	Structure
1	8.42_428.1833n	M-H, 2M-H	C24H28O7	51.9	62.1	-0.47	428.1833	427.1760	Hypothesis m
2	**8.42_428.1833n**	**M-H, 2M-H**	**C24H28O7**	**51.2**	**58.6**	**-0.47**	**428.1833**	**427.1760**	Parvixanthone G (**10**)
3	8.42_428.1833n	M-H, 2M-H	C24H28O7	44.1	22.7	-0.47	428.1833	427.1760	Hypothesis n
4	8.42_428.1833n	M-H, 2M-H	C24H28O7	43.5	19.8	-0.47	428.1833	427.1760	Hypothesis o
5	8.42_428.1833n	M-H, 2M-H	C24H28O7	42.9	16.8	-0.47	428.1833	427.1760	Hypothesis p



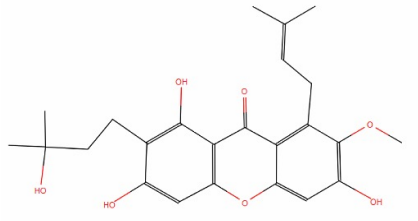
hypothesis m



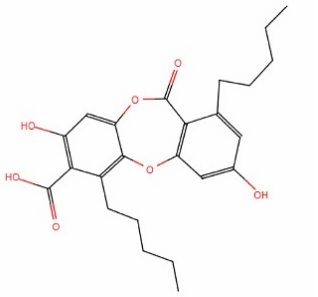
parvixanthone G (**10**)



hypothesis n



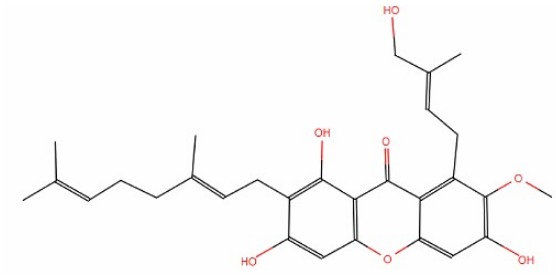
hypothesis o



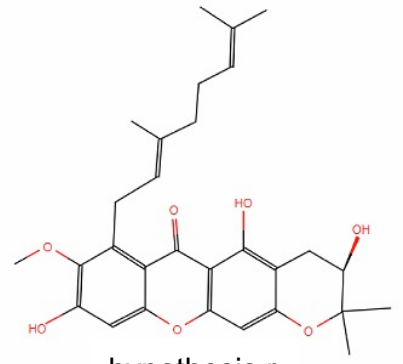
hypothesis p

Table S19. Part of structures predicted for the ion at 493.2228 m/z (node **10**). The proposition is based on a ProgenesisQI in-silico fragmentation using the PNMRNP database (precursor fragmentation was set at 5 ppm and theoretical fragmentation was set at 10 ppm).

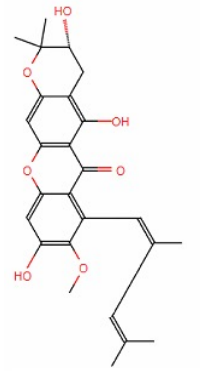
Rank	Compound (RT_MWn)	Adducts	Formula	Score	Fragmentation Score	Mass Error (ppm)	Neutral mass (Da)	m/z	Structure
1	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	50.1	58.4	-0.83	494.2300	493.2228	Hypothesis q
2	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	49.7	56.6	-0.83	494.2300	493.2228	Hypothesis r
3	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	49.7	56.6	-0.83	494.2300	493.2228	Hypothesis s
4	**11.22_494.2300n**	**M-H, 2M-H, M-H2O-H**	**C29H34O7**	**49.6**	**56.1**	**-0.83**	**494.2300**	**493.2228**	Isocowanol (**20**)
5	11.22_494.2300n	M-H, 2M-H, M-H2O-H	C29H34O7	49.6	56.1	-0.83	494.2300	493.2228	Hypothesis t



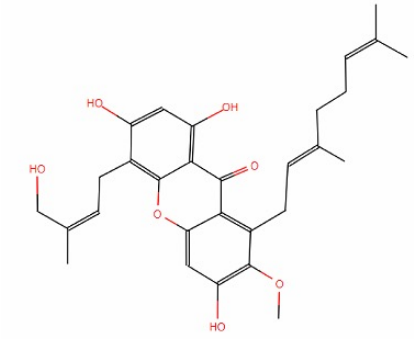
hypothesis q



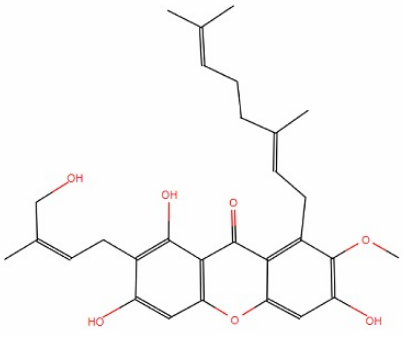
hypothesis r



hypothesis s



isocowanol (**20**)



hypothesis t

Table S20. Part of structures predicted for the ion at 493.2228 m/z (node **20**). The proposition is based on a ProgenesisQI in-silico fragmentation using the PNMRNP database (precursor fragmentation was set at 5 ppm and theoretical fragmentation was set at 10 ppm).

Carbon numbering	Literature data	Experimental data	Dereplication by MixONat	
	δc (ppm)	δ_C matched in F10 (ppm) in acetone-d6	δc (ppm) GarciniaDB	δc (ppm) matched found by MixONat
9	182.5	182.5	183.8	182.8
3	165.1	165.2	165.1	164.9
1	164.6	164.6	163.2	/
4a	157.8	157.7	158.0	158.0
6	157.3	157.2	156.7	156.4
10a	156.1	156.2	155.6	/
7	144.3	144.3	143.6	144.6
8	139.9	139.6	138.9	140.1
17	131.0	131.2	132.5	/
16	126.0	126.0	123.9	/
8a	111.8	111.8	113.2	112.0
9a	103.5	103.6	103.7	103.7
5	102.4	102.6	102.7	102.7
2	98.5	98.6	97.9	98.7
4	93.6	93.7	93.8	93.8
13	72.0	73.0 or 71.3	76.3	/
7-OMe	61.4	61.3	62.0	61.6
12	43.3	43.1	42.9	42.5
14	42.3	42.2	39.2	/
19	27.2	27.2	25.7	25.9
18	25.7	25.7	24.0	/
15	23.2	23.2	22.9	23.4
11	22.7	22.8	22.2	23.0
20	17.6	17.7	17.6	17.7

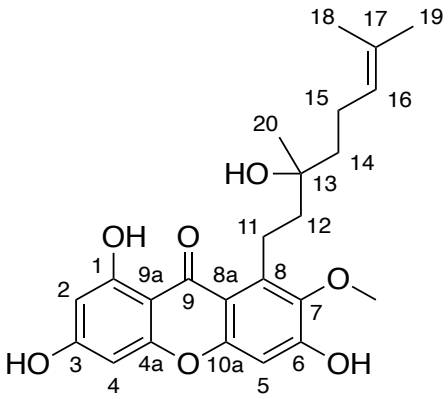


Table S21. Comparison of experimental δ_C of F10 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$).The comparison of experimental data with the literature[14] confirm the identification of **parvixanthone G** in F10.

[14] Xu, Y.-J. *et al.*, *J. Nat. Prod.* **2001**, *64* (9), 1191–1195.

Carbon numbering	δ_H (ppm)	δ_C (ppm)	Dereplication by MixONat with "GarciniaDB"
9		182.9	183.5
3		163.0	163.1
1		162.4	162.3
6		157.8	154.0
10a		156.3	154.6
4a		154.9	153.7
7		144.6	142.7
3'		137.9	137.6
8		135.6	136.5
3"		134.9	136.0
7'		131.5	132.0
2"	5.33 (1H, t, *J* = 7.5 Hz)	125.3	123.0
6'	5.03 (1H, t, *J* = 6.9 Hz)	125.1	125.7
2'	5.25 (1H, t, *J* = 6.5 Hz)	124.8	124.5
8a		111.4	111.1
4		105.5	106.8
9a		103.6	101.2
5	6.97 (1H, s)	102.9	104.3
2	6.20 (1H, s)	98.3	98.7
4"	4.40 (2H, s)	61.4	61.9
O-Me	3.78 (3H, s)	61.1	61.2
4'	1.96 (2H, t)	40.4	40.1
5'	1.72 (2H, s)	27.2	26.8
1'	4.09 (2H, d, *J* = 6.4 Hz)	26.6	25.8
7'Me	1.54* (3H, s)	25.7*	26.9
3"Me	1.72 (3H, s)	21.7	21.7
1"	3.49 (2H, d, *J* = 7.7 Hz)	21.6	22.3
7'Me	1.49* (3H, s)	17.6*	17.9
3'Me	1.80 (3H, s)	16.5	16.5
1-OH	13.44 (1H, s)		

Table S22. The δ_C and δ_H of the **isocowanol** (**20**) according to NMR experiments of F10 in acetone-*d*6. Signals with the same superscript (*) are interchangeable.

Carbon numbering	Literature data	Experimental data	Dereplication by MixONat	
	δc (ppm)	δc in acetone-d6	δc (ppm) GarciniaDB	δc (ppm) matched found by MixONat
9	182.5	182.5	183.8	182.8
3	165.1	165.2	165.1	164.9
1	164.6	164.6	163.2	/
4a	157.8	157.7	158.0	158.0
6	157.3	157.2	156.7	156.4
10a	156.1	156.2	155.6	/
7	144.3	144.3	143.6	144.6
8	139.9	139.6	138.9	140.1
17	131.0	131.2	132.5	/
16	126.0	126.0	123.9	/
8a	111.8	111.8	113.2	112.0
9a	103.5	103.6	103.7	103.7
5	102.4	102.6	102.7	102.7
2	98.5	98.6	97.9	98.7
4	93.6	93.7	93.8	93.8
13	72.0	73.0 or 71.3	76.3	/
7-OMe	61.4	61.3	62.0	61.6
12	43.3	43.1	42.9	42.5
14	42.3	42.2	39.2	/
19	27.2	27.2	25.7	25.9
18	25.7	25.7	24.0	/
15	23.2	23.2	22.9	23.4
11	22.7	22.8	22.2	23.0
20	17.6	17.7	17.6	17.7

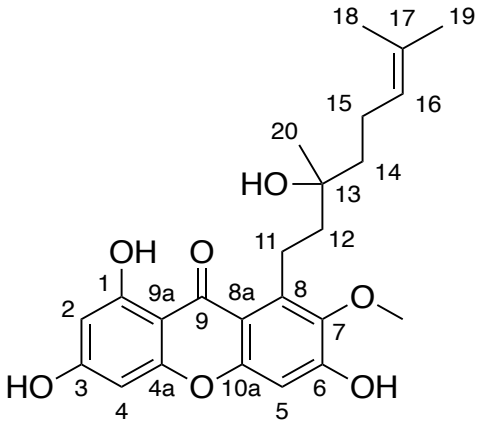
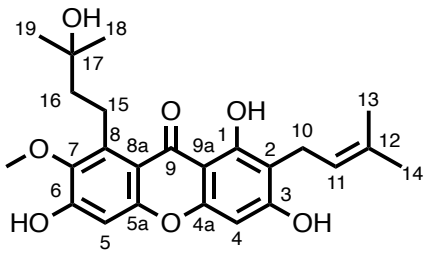


Table S23. Comparison of experimental δ_C of F10 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$). The comparison of experimental data with the literature[15] confirm the identification of **parvixanthone G** (**10**) in F10.

[15] Xu, Y.-J. *et al.*, *J. Nat. Prod.* **2001**, *64* (9), 1191–1195.

Experimental data		**Garcinone D (11)**				**Cratoxylone (12)**			
δc in acetone-*d*6	Intensity	Carbon numbering	δc found in literature	δc xanthoneDB	δc matched found by MixONat	Carbon numbering	δc found in literature	δc xanthoneDB	δc matched found by MixONat
182.9	0.00014					9	182.8	183.1	182.9
182.8	0.00041	9	182.8	182.9	182.8				
163.3	0.00012					3	163.3	163.5	163.3
163.0	0.00044	3	163.1	161.9	163.0				
161.7	0.00064	1	N.D.	161.4		1	161.6	160.4	161.7
157.5	0.00035	6	157.7	157.6	157.5	6	157.6	154.5	/
156.3	0.00036	5a	156.3	155.0	156.3				
156.2	0.00010					4b	156.2	155.7	156.2
155.7	0.00037	4a	155.7	155.5	155.7				
155.7	0.00011					4a	155.6	155.1	155.7
144.4	0.00034	7	144.5	144.5	144.4	7	144.5	142.7	/
140.1	0.00045	8	140.0	140.3	140.1				
138.1	0.00011					8	138.0	137.2	138.1
131.4	0.00049	12	131.3	130.5	131.4	17	131.3	132.2	131.4
124.8	0.00022					16	124.8	123.0	/
123.5	0.00085	11	123.5	122.2	123.5				
112.4	0.00014					8a	112.3		
112.0	0.00031					2	111.8	112.5	112.0
111.9	0.00008	8a	111.8	112.0	111.9				
111.0	0.00048	2	111.0	109.4	/				
103.6	0.00034	9a	103.6	105.0	/				
103.5	0.00008					9a	103.5	103.7	103.5
102.7	0.00018					5	102.7	101.6	102.7
102.5	0.00070	5	102.5	102.7	102.5				
93.4	0.00017					4	93.3	91.0	/
93.1	0.00069	4	93.1	93.4	93.1				
70.7	0.00019					12	70.6	70.1	70.7
70.5	0.00070	17	70.5	70.6	70.5				
61.6	0.00081	7-OMe	61.5	61.7	61.6				
61.3	0.00021					7-OMe	61.2	61.3	62.1
45.7	0.00074	16	45.7	45.8	45.7				
43.2	0.00020					11	43.2	42.9	43.2
29.6	0.00262		N.D.	29.1	/	13; 14	29.1	29.4	/
26.9	0.00018					15	26.8	26.6	26.9
25.9	0.00027					18	25.9	25.8	25.9
25.9	0.00094	13	25.9	25.5	25.9				
23.2	0.00069	15	23.2	23.3	23.2				
22.0	0.00076	10	22.0	20.9	22.0				
18.3	0.00023					19	18.3	18.0	18.3
17.9	0.00079	14	17.9	17.7	17.9				
17.9	0.00022					10	17.8	17.8	17.9



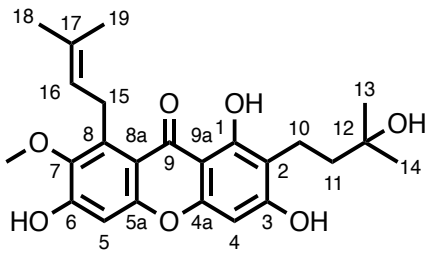


Table S24. Comparison of experimental δ_C of F10 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$). The comparison of experimental data with the literature[16] confirm the identification of **garcinone D** (**11**) and **cratoxylone** (**12**).

[16] Bennett, G. J. *et al.*, *Phytochemistry* **1993**, *32* (5), 1245–1251.

Carbon numbering	Literature data	Experimental data	Dereplication by MixONat	
	δc (ppm)	δc in acetone-d6	δc (ppm) GarciniaDB	δc (ppm) matched found by MixONat
9	182.9	182.8	182.1	182.8
3	163.1	162.9	160.7	/
1	161.7	161.7	161.6	161.7
6	157.6	157.5	154.5	/
10a	156.3	156.2	155.8	156.2
4a	155.8	155.7	155.1	155.7
7	144.6	144.4	142.6	/
8	138.2	138.1	137.1	138.1
18	131.4	131.5	135.8	135.1
13	131.2	131.4	132.1	131.5
17	124.5	124.8	121.4	/
12	123.6	123.5	123.2	123.5
8a	111.9	112.0	112.3	112.0
2	111.1	111.0	108.4	/
9a	103.7	103.6	103.7	103.7
5	102.8	102.7	101.6	102.7
4	93.2	93.1	93.3	93.1
7-OMe	61.4	61.4	62.1	61.4
16	26.1	25.9	26.6	26.9
15	25.9	25.9	25.8	/
20	25.8	25.8	25.8	/
11	22.1	22.0	21.5	22.0
14	18.4	18.3	18.2	18.3
19	18.0	17.9	18.0	17.9

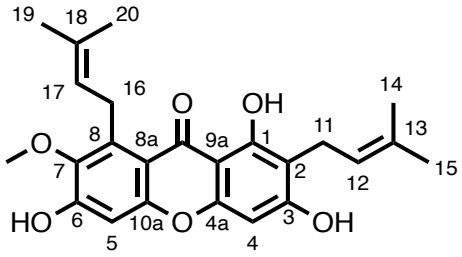


Table S25. Comparison of experimental δ_C of F4 with those of previously described NPs in *Garcinia* species and the predicted shifts of ACD/Labs NMR Predictors (C,H) ($\delta_{C\text{-}SDF}$).The comparison of experimental data with the literature[17] confirm the identification of **α-mangostin** (**27**).

[17] Ha, L. D. *et al.*, *Chem. Pharm. Bull.* **2009**, *57* (8), 5.

5.1.1. Additional information

Regarding inactive unidentified NPs with m/z 475 $[M+FA-H]^-$ ($C_{27}H_{42}O_4$?, nodes **21 (M1)**, **25 (M2)**, **26 (M3)**, **28 (M4)**), ProgenesisQI and Metascope failed to propose a relevant structure. In F4, MixONat suggested tetraprenyltoluquinone (TPTQ, Appendix 2 to Appendix 5) and α-mangostin (data not shown) as major NPs. To the best of our knowledge, no TPTQ derivatives were previously described. The molecular formula of these minor compounds did not match with any NPs from *Garcinia*DB. Thus, M1 to M4 were isolated from F4 for a complete structural determination (Appendix 6 to Appendix 21).

Spectral data confirm that they are TPTQ derivatives (Figure 21). A meticulous analysis of ^{1}H, ^{13}C, and 2D NMR data (Table 3) allowed us to suggest the following oxidized derivatives with the assistance of CMGNet (Yao *et al.*, 2023), MADByTE platform (Flores-Bocanegra *et al.*, 2022) (Appendix 22) and ACD/spectrus.

These results raise the problem of detection of TPTQ and its derivatives by MS. In addition, the molecules have been analyzed by several other MS systems (LDI, MALDI, ion trap), but none of them allowed the correct identification of the molecular mass. Recent analysis has shown that the molecules are degraded.

TPTQ				M1			M2			M3			M4		
Carbon numbering	**δ_C (ppm)**	**δ_H (ppm)**	**DEPT**	**δ_C (ppm)**	**δ_H (ppm)**	**DEPT**	**δ_C (ppm)**	**δ_H (ppm)**	**DEPTs**	**δ_C (ppm)**	**δ_H (ppm)**	**DEPTs**	**δ_C (ppm)**	**δ_H (ppm)**	**DEPTs**
1	197.9	/	C	197.9	/	C	198.1	/	C	198.0	/	C	197.9	/	C
3	165.0	/	C	165.1	/	C	165.2	/	C	165.1	/	C	165.0	/	C
3'	137.8	/	C	141.3	/	C	137.8	/	C	137.7	/	C	137.7	/	C
7'	135.5	/	C	137.8	/	C	137.6	/	C	135.5	/	C	135.2	/	C
11'	135.4	/	C	135.6	/	C	129.3	/	CH	135.0	/	C	132.0	/	C
15'	131.5	/	C	126.8	5.72 (1H, s)	CH	126.8	5.72 (1H, s)	CH	126.8	5.72 (1H, s)	CH	126.8	5.72 (1H, s)	CH
2	126.8	5.72 (1H, s)	CH	125.5	5.22-5.12* (4H, m)	CH	125.8	?	CH	125.4	5.14 - 5.22* (2H, m)	CH	125.3*	5.18-5.22* (2H, m)	CH
6'	125.2	5.05-5.10* (3H, m)	CH	125.1	5.22-5.12* (4H, m)	CH	125.5	?	CH	125.1	5.14 - 5.22* (2H, m)	CH	124.9*	5.18-5.22* (2H, m)	CH
10'	125.1	5.05-5.10* (3H, m)	CH	124.8	5.22-5.12* (4H, m)	CH	125.4	?	CH	122.3	5.13 (1H, d)	CH	122.4	5.11 (1H, t, J = 6.3 Hz)	CH
14'	125.0	5.05-5.10* (3H, m)	CH	122.4	5.22-5.12* (4H, m)	CH	122.3	5.19 (2H, s)	CH	72.9	4.11 (1H, d)	CH	72.9	4.11 (1H, s)	CH
2'	122.3	5.16 (1H, t, J = 7.4 Hz)	CH	109.1	/	C	113.6	4.91 (1H, s)	CH2	64.0	2.63 (1H, t)	CH	63.2	2.65 (1H, t, J = 6.2 Hz)	CH
4	72.9	4.07 (1H, d, J = 7.9 Hz)	CH	74.8	4.11 (2H, s)	CH	89.0	4.23 (1H, t)	CH2	44.7	2.0 **	CH2	60.6	/	C
6	44.6	2.0 - 2.1**	CH	72.9	4.11 (2H, s)	CH	72.9	4.11 (1H, s)	CH2	41.6	2.44 (2H, s)	CH2	44.7	2.28 (1H, m)	CH
5	41.6	2.0 - 2.1**	CH2	70.1	/	C	44.7	2.0 **	CH	40.5	2.0 **	CH2	41.6	2.09-2.15* (6H, m)	CH2
4'	40.6	2.0 - 2.1**	CH2	44.7	?	CH	41.6	2.44 (2H, s)	CH2	40.4	2.0 **	CH2	40.5	2.42 (2H, d)	CH2
8'	40.4	2.0 - 2.1**	CH2	43.1	2.42 (6H, d)	CH2	40.6	2.0 **	CH2	40.3	2.0 **	CH2	39.6	2.09-2.15* (6H, m)	CH2
12'	40.4	2.0 - 2.1**	CH2	41.6	2.42 (6H, d)	CH2	40.5	2.0 **	CH2	37.1	2.0 **	CH2	37.0	1.94-2.03* (6H, m)	CH2
1'	30.8	2.0 - 2.1**	CH2	40.4	2.42 (6H, d)	CH2	40.3	2.0 **	CH2	28.3	2.0 **	CH2	28.1	1.94-2.03* (6H, m)	CH2
5'	27.4	1.91-1.99* (6H, m)	CH2	36.7	2.0 - 2.1 **	CH2	38.9	2.0 **	CH2	27.2	2.0 **	CH2	27.4	1.94-2.03* (6H, m)	CH2
9'	27.3	1.91-1.99* (6H, m)	CH2	32.6	/	C	36.4	2.0 **	CH2	27.1	2.0 **	CH2	27.1	2.09-2.15* (6H, m)	CH2
13'	27.2	1.91-1.99* (6H, m)	CH2	27.4	2.0 - 2.1 **	CH2	27.2	2.0 **	CH2	25.8	1.65* (6H, s)	CH3	25.8	1.66 (3H, s)	CH3
15'Me	25.9	1.62 (3H, s)	CH3	27.1	2.0 - 2.1 **	CH2	27.1	2.0 **	CH2	25.1	1.65* (6H, s)	CH3	24.6	1.60 (2H, brs)	CH2
3Me	20.7	1.99 (3H, s)	CH3	25.8	1.63 (4H, s)	CH3	25.8	1.65 (6H, s)	CH3	20.7	2.0 **	CH3	20.7	2.02 (3H, s)	CH3
16'	17.8	1.56* (6H, s)	CH3	23.3	?	CH3	23.3	?	CH3	19.0	1.20 (3H, s)	CH3	17.7	1.60 (3H, s)	CH3
3'Me	16.3	1.56* (6H, s)	CH3	20.7	2.0 **	CH3	22.5	1.28 (5H, s)	CH3	17.7	?	CH3	16.8	1.21 (3H, s)	CH3
7'Me	16.2	1.57 (3H, s)	CH3	16.3	1.61* (6H, s)	CH3	20.7	2.0 **	CH3	16.3	1.62* (6H, s)	CH3	16.3	1.63* (6H, s)	CH3
11'Me	16.1	1.59 (3H, s)	CH3	16.1	1.61* (6H, s)	CH3	17.7	?	CH3	16.1	1.62* (6H, s)	CH3	16.1	1.63* (6H, s)	CH3
							17.1	1.69 (3H, s)	CH3						
							16.3	?	CH3						
							16.1	1.63 (6H, s)	CH3						

Table 3. NMR spectral data assignment of TPTQ compared to M1, M2, M3 and M4 in acetone-*d*6

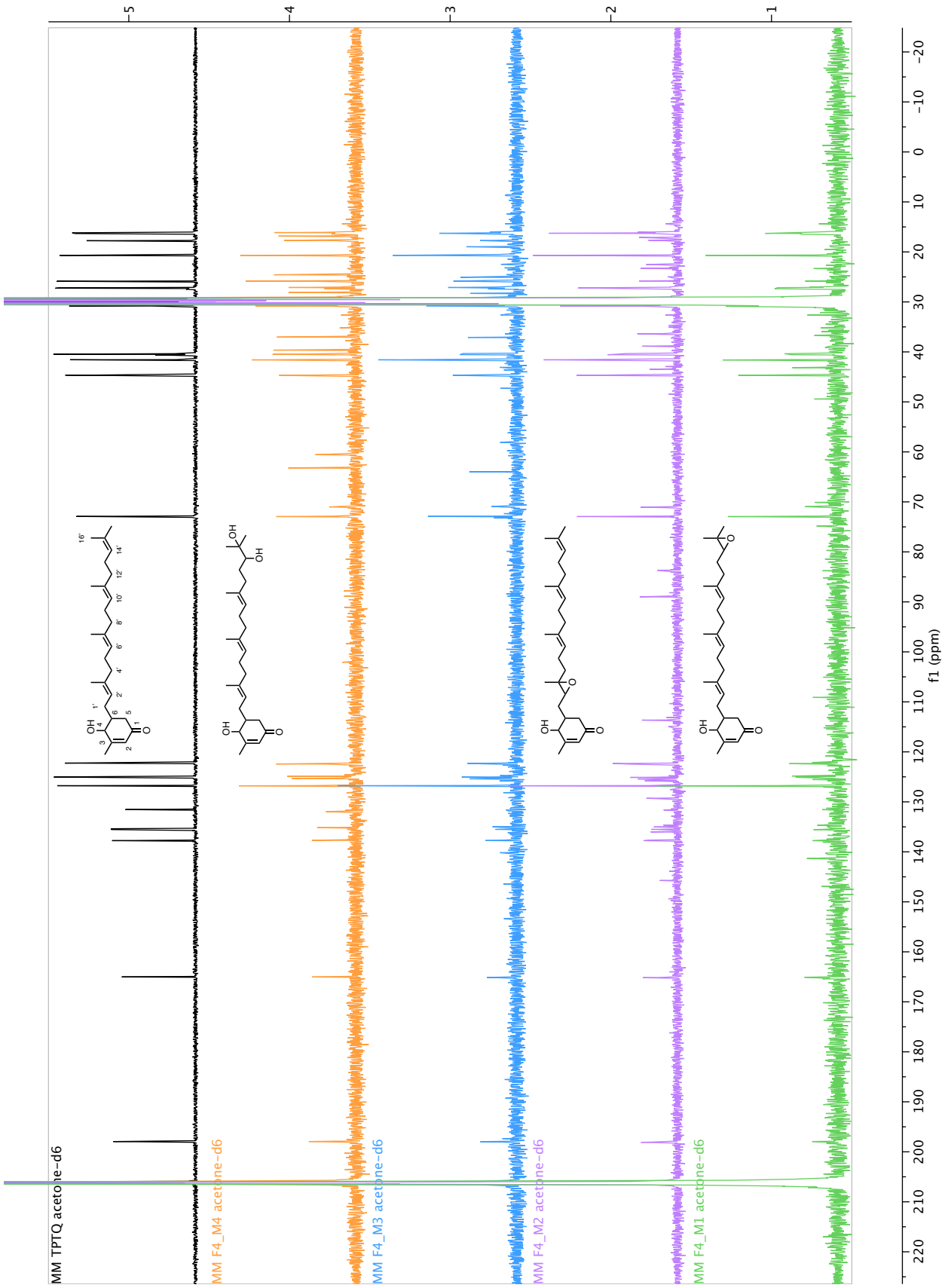


Figure 21. ^{13}C NMR spectra of TPTQ derivatives (M1, M2, M3 and M4) and TPTQ analyzed in acetone-*d*6.

5.2. Matrix free laser desorption ionization coupled to ion mobility spectrometry mass spectrometry: a powerful tool for the differentiation of natural isomers and molecular networking

This publication will be submitted in Analytical Chemistry journal.

Matrix free laser desorption ionization coupled to ion mobility spectrometry mass spectrometry: A powerful tool for the differentiation of natural isomers and molecular networking

Manon Meunier[1], Martina Haack[2], Dania Awad[2], Thomas Brück[2], Khalijah Awang[3], Marc Litaudon[4], Frédéric Saubion[5], Marc Legeay[5], Jules Leguy[5], David Guilet[1], Séverine Derbré[1*], Andreas Schinkovitz[1*]

[1] Univ Angers, SONAS, SFR QUASAV, Faculty of Health Sciences, Dpt Pharmacy, 16 Bd Daviers, 49045 Angers CEDEX 01, France

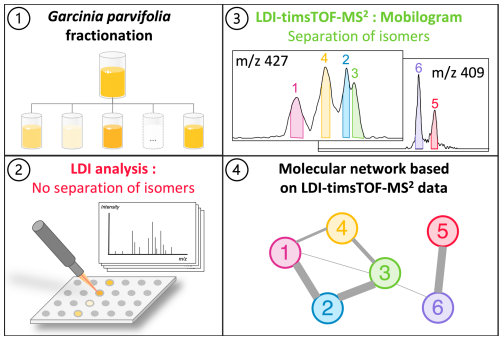
[2] Technical University of Munich, Department of Chemistry, Werner Siemens-Chair of Synthetic Biotechnology (WSSB), Lichtenbergstraße 4, 85748 Garching

[3] University of Malaya, Faculty of Sciences, Department of Chemistry, Malaysia

[4] Institut de Chimie des Substances Naturelles, CNRS-ICSN, UPR 2301, Université Paris-Saclay, 91198, Gif-sur-Yvette, France

[5] LERIA, EA2645, UNIV Angers, SFR MathSTIC, Faculty of Sciences, 2 boulevard Lavoisier, 49045 Angers cedex 01, France

ABSTRACT: The chemical profiling of complex mixtures comprising isomers is a major challenge in analytical chemistry and usually facilitated by liquid chromatography-mass spectrometry (LC-MS). In recent years matrix-free laser desorption ionization-mass spectrometry (LDI-MS) has become a powerful and versatile complement to LC-MS. Unlike LC-MS, the method permits the use of any volatile solvent, requires very little sample preparation, and provides immediate results. In addition, LDI-MS is particularly suitable for analyzing natural products (NPs). Despite these advantages, the absence of chromatographic separation remains a major limitation of the method, particularly for the analysis of isomers. With this in mind, the current work presents, a hyphenated LDI-ion mobility-MS^2 (LDI-IM-MS^2) approach. As a working example for the differentiation of constitutional xanthone isomers found in *Garcinia parvifolia* is being discussed. Finally, the so unexplored concept of LDI-IM-MS^2 based molecular networking is presented.



Liquid chromatography coupled to high-resolution tandem mass spectrometry (LC-MS^2) is widely considered as the gold standard for the analysis of complex mixtures of small molecules. Combining the assets of MS^2 and chromatographic separation, the method has become a cornerstone of contemporary "omics" research, including molecular networks (MNs).[1–4]

In recent year, matrix-free laser desorption ionization (LDI)-MS has been developed as a powerful complement to LC-MS. Benefiting from a time efficient experimental protocol, the method is particularly suitable for analyzing natural products (NPs) such as phenolics and alkaloids.[5–9] The latter show close structural similarities to matrices used in matrix-assisted laser desorption ionization (MALDI). Therefore, they can be easily ionized by simple laser irradiation and without any additional matrix support. This direct LDI-MS approach has been successfully applied to characterize highly complex mixtures of NPs such as plant powders[5] or crude extracts[7,10]. Moreover, LDI promotes the formation of both radical and quasi-molecular ions and may therefore detect compounds that are potentially missed by electrospray ionization (ESI)[6]. This effect was particularly observed for various phenolic compounds.[6]

The LDI-MS approach may further be easily integrated into a chemometric workflow. A recent study on crude extracts of *G. parvifolia* yielded comparable results for LDI-MS and UPLC-MS based partial least square-discriminant analysis (PLS-DA).[10] However, the same report also revealed a most fundamental problem of LDI-MS. Due to the absence of chromatographic separation constitutional isomers cannot be differentiated in LDI-MS.

One approach to bypass this limitation is ion mobility spectrometry (IMS). The method separates gas phase ions based on their interaction with a collision gas, which depends on the ions' conformation, spatial orientation, shape, charge, and mass.[11] By today IMS has been successfully applied to differentiate various isomers from different chemical groups such as sugars[12], lipids[13] or cannabinoids[14]. On the other hand, very little information about xanthones is available.

With this in mind, the present work evaluated the ion mobility (IM) of various xanthone isomers by a combined LDI-M-MS^2 approach (**I**). Next to single compounds, simple mixtures of previously isolated compounds from a bark extract (KL5670), a plant that is *G. parvifolia* particularly rich in xanthones, were analyzed (**II**). Extending the conventional use of LDI-MS, the feasibility of an LDI-IM-MS^2 based MN was further explored on fractions of one of the studied *Garcinia* extract (KL5073) (**III**).[1]

▪ Experimental section

Plant collection. *Garcinia parvifolia* bark samples were collected in Malaysia at two different locations: KL5670: Hutan Simpanan

Meranto Gua Musang, Kelantan (date of collection: 13.05.2009), KL5073: 1.2 km Kg. Toh Kah, Jerangau, Terengganu (date of collection: 20.08.2004). Plant material was verified by the botanist Mr. Teo Leong Eng and voucher specimens are kept at the Herbarium of the Department of Chemistry, University of Malaya, Kuala Lumpur, Malaysia.

Extraction procedure. Eighty gram of ground bark material (KL5670) was mixed with 100 mL of dichloromethane (DCM) (Carlo Erba reagents, Val-de-Reuil, France) and sonicated at room temperature for 30 min. After filtration and solvent evaporation, 4.48 g of dried extract were obtained. From this extract the following ten compounds were isolated: butyraxanthone D (**1**), cratoxylone (**2**), garcinone D (**3**), parvixanthone G (**4**), rubraxanthone (**5**), α-mangostin (**6**), parvixanthone H (**7**), isocowanol (**8**), isocowanin (**9**), and 9-hydroxycalabaxanthone (**10**) (Figure 1). A detailed experimental protocol is provided in a manuscript that has been submitted to the Journal of Natural Products.[15]

The same extraction protocol was applied for processing 10 g of KL5073, which yielded 578 mg of dried extract. The latter then underwent to chromatographic separation as outlined in below. Fractions were analyzed by LDI-IM-MS^2 in order to create the molecular network (MN) shown in **Figure 3**.

Flash chromatography conditions. Fractionation of KL5073 was conducted on a CombiFlash Rf flash chromatograph (Serlabo Technologies, Entraigues, France) using a Puriflash 50SIHC-JP-F0025 (25 g) silica gel column. The mobile phase consisted of petroleum ether (solvent A), an equivolumetric mixture of ethyl acetate and chloroform (solvent B), and methanol (solvent C). All solvents were purchased from Carlo Erba reagents (Val-de-Reuil, France). The flow rate was set to 20 mL/min and the following gradient was applied: 0 min, (A/B/C - 100/0/0); 0-60 min, (A/B/C - 0/100/0); 60-61 min, (A/B/C - 0/80/20); 61-68 min, (A/B/C - 0/80/20). Fractions were collected every 20 mL and analyzed by thin layer chromatography (data not shown). Overall, fifteen recombined fractions were obtained (Figure S1, supplementary information).

Preparation stock solutions and compound mixtures. Stock solutions (SS) of all compounds were prepared in acetone at a concentration of 5.0 mg/mL. Working solutions for mixtures were then prepared by mixing 200 µL of SS according to the following protocol: mix 1 compound **1**-**2**, mix 2 compound **2**-**3**, mix 3 compound **1**, **2** and **4**, mix 4 compound **1**-**4**, and mix 5 compound **5**-**6**.

Sample preparation. Mixtures of purified compounds and single compound experiments were dried and redissolved in DCM/MeOH (60/40) at a concentration of 5.0 mg/mL. Crude *G. parvifolia* extracts and fractions of KL5073 were dissolved in a mixture of DCM/MeOH (60/40) at a concentration of 10.0 mg/mL. All samples were deposited in triplicates (0.5 µL/each) on the MALDI sample carrier.

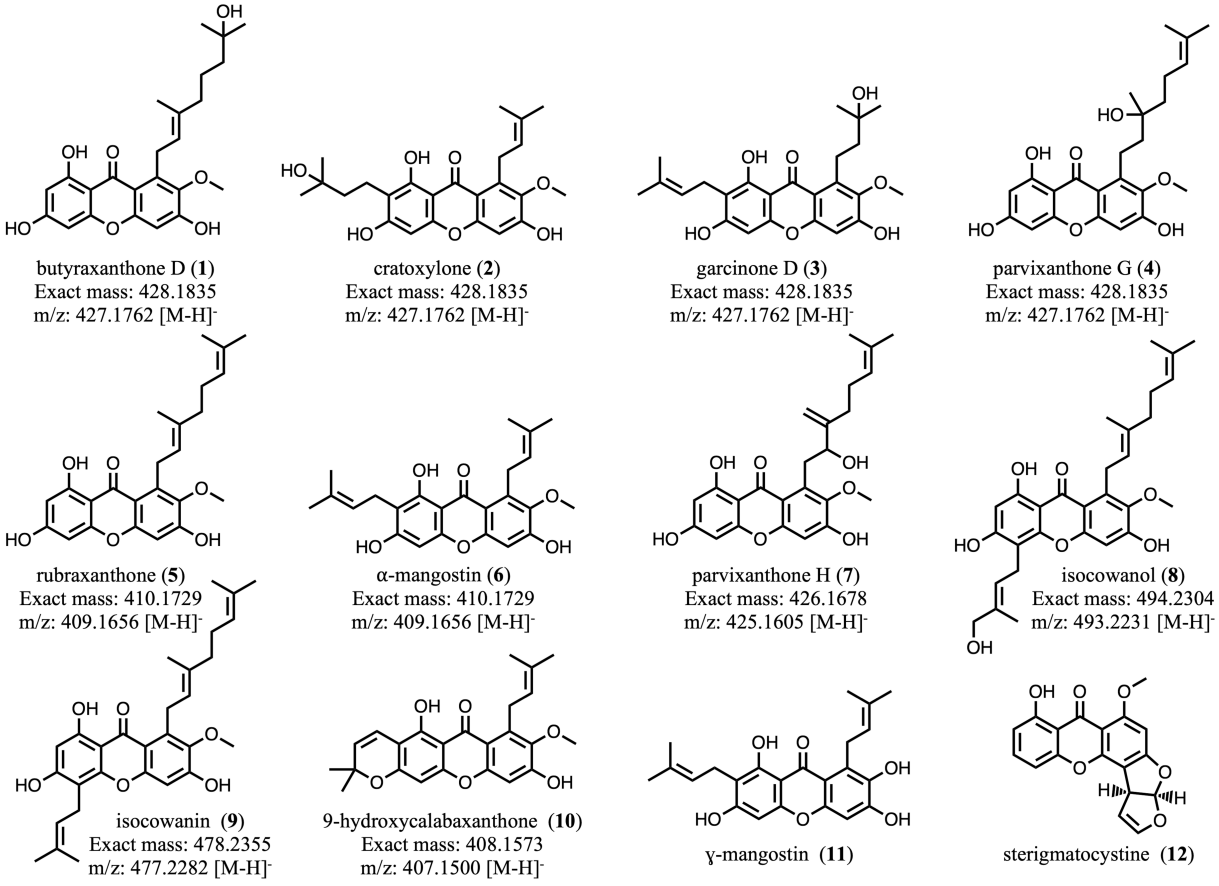


Figure 1. Isolated xanthones from the *Garcinia parvifolia* bark extract KL5670.

UPLC-MS experiments. Selected fractions of KL5670 were analyzed by UPLC-MS on an Acquity UPLC system (Waters, Manchester, UK) using the following conditions: the column temperature was set to 25°C. A HSS T3 C-18 (2.1 x 100 mm, 1.8 µm) UPLC column (Waters) was used and the mobile phase consisted of LC-MS grade water (solvent A) and acetonitrile

(solvent B, LC-MS grade, VWR chemicals), both supplemented with 0.1% of formic acid (LC-MS grade, VWR chemicals). The mobile phase gradient was as follows: 0min, 40% B; 5min, 80%B; 10min, 100%B; 12min, 100% B. The flow was set to 0.4 mL/min. Sample material of selected fractions were dissolved in LC-MS grade methanol (VWR chemicals) at a concentration of 0.02 mg/mL (injection volume 1 µL).

LDI-IM-MS² experiments. Experiments were conducted on a timsTOF fleX MALDI-2 (Bruker Daltonics GmbH & Co. KG, Bremen, Germany), equipped with a Bruker scanning SmartBeam 3D Laser and a Bruker MALDI-2 equipped with a dual laser system operating at 355 nm. Spectra were recorded applying the following settings: mass range, 50–1200 m/z; ion mobility ramp, 0.8 -1.3 Vs/cm^2. Laser frequency was set to 1000 Hz, and cumulative spectra of 500 acquisitions were collected for each sample spot. Laser intensity (70 – 100 %) and MS2 collision energy (28 - 55 V), were individually adapted to sample when analyzing fractions of KL5073. For single a collision energy of 35eV was applied for compounds **1**-**7** and **10**. For compounds **8**-**9**, 40 eV were used. Preceding experiments have shown that best results were obtained in negative ionization mode, so all spectra were recorded in high-resolution negative mode. Raw data were processed using the DataAnalysis 5.3 software.

For LDI-IM-MS² experiments on fractions of KL5073, only most intense MS¹ ions were selected for ion mobility as low intensity MS¹ ions did not produce satisfactory fragmentation.

The creation of a GNPS database. Individual MS2 spectra were first exported as .mgf format using the DataAnalysis 5.3. Then they were uploaded as metadata file to GNPS platform. The database was named Xanthones-DB and is available at https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=XANTHONES-DB.

Molecular networking. All MS² data were exported as text files (.mgf) using DataAnalysis 5.3 and manually merged into one file. A MN was then created by applying the following parameters: precursor ion and fragment ion mass tolerance (0.02 Da), cosine score (≥ 0.5), topK (5), minimum matched fragment ions (3) and maximum connected component size equals to 0. The library search options were set to a minimum of 4 matched peaks, and a score threshold of 0.7. The job record for the isolated compounds of with KL5670 is available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3a7d8e158cee45e2a8dcba86ce88f147. The job record for Xanthones-DB results is available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=dc53e87772d74e959148d7704a275f66. Visualization and treatment of the molecular network was performed by the Cytoscape 3.9.0 software.[16]

Results and discussion

Xanthones are a very interesting group of phenolic compounds that are known for a wide range of differentactivities.[17–20] Showing close structural similarity to dithranol, (a well-known MALDI matrix) they can be easily detected by matrix free LDI-MS even from complex mixtures.[10] However, various constitutional isomers of xanthones exist,[21,22] and their distinct differentiation is an important analytical challenge. Due to the absence of chromatographic separation, mixtures of isomers cannot be distinguished by classical LDI-MS.

With this in mind, the present study analyzed the IM of ten xanthones (figure 1) that were previously isolated from a bark extract of *G. parvifolia* (KL5670). Detailed information about the isolation and structure elucidation process is provided in a manuscript that has been submitted to the Journal of Natural Products.[15] These compounds comprised two groups of constitutional isomers, one with a monoisotopic mass of 428.1835 Da (compounds **1**-**4**, $C_{24}H_{28}O_7$), and another one with a monoisotopic mass of 410.1729 Da (compounds **5**-**6**, $C_{24}H_{26}O_6$).

Table 1. Experimental CCS values of isolated single compounds from KL5073 (1-10) and mixtures 1-5.*

	Compounds									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Single	197.7	206.3	207.3	205.0	194.0	206.5	204.4	220.4	218.7	202.3
Mix 1	198.7	207.4								
Mix 2		207.4	209.9							
Mix 3	199.1	207.7		205.3						
Mix 4	199.6	209.2	210.7	205.2						
Mix 5					198.7	208.4				
KL5073*	197.7			204.3	194.0					
Mean	198.8	207.6	209.3	205.2	196.4	207.5				
CV (%)	0.4	0.5	0.8	0.1	1.7	0.6				
AllCCS	207.2	208.0	208.0	206.7	200.5	200.0	203.9	216.5	214.3	201.6
CCSbase	205.6	206.5	206.5	205.6	201.2	201.9	199.1	217.8	218.0	206.5

First, all compounds were individually analyzed by LDI-IM-MS², before specific mixtures comprising between two and four isomers were studied. Results from these experiments are summarized in **Table 1**, including the CCS values determined by IMS measurements. This allows the direct comparison of CCS values obtained from differently designed instruments such as drift tube (DTIM) or traveling wave ion

* Predicted CCS values were obtained from the AllCCS and CCSbase database. Inter-experimental variations are expressed as coefficient of variation (CV). *Annotations for compounds detected in fractions of KL5073 without consecutive isolation and ultimate structural confirmation. These values were not used for the calculation of mean values.

mobility (TWIM) spectrometers.[23] As also shown in **Table 1**, very good inter-experimental repeatability for individual compounds and mixtures was observed. Coefficients of variation (CV) were in the range of 0.1 to 1.7% (Table 1), which is in line with previous reports (0.2-6.2%).[23–26]

By today many open access IM databases (DBs) such as CCSbase[26], AllCCS[25], ISiCLE[27], or DeepCCS[28] exist. They provide experimental and/or predicted CCS values for numerous compounds from various chemical classes. However, experimental data on xanthones are very rare. Solely α-mangostin (**6**) was previously studied and its reported CCS value is in line with current results (206.4 vs 207.5, respectively).[29] In order to still explore the rest of the compounds, predicted CCS values from the CCSbase and AllCCS databases added to **Table 1**. Both DBs showed very good accordance with experimental data for compounds **2-4** while largest differences were observed for compound **1**. In any case, the differences between predicted and experimental CCS values never exceeded 10%.

Extending the information provided in **Table 1**, **Figure 2** depicts representative mobilograms of mix 1 and 4 as well as the MS^2 spectra of compounds 1-4 acquired after LDI-IM-MS^2. (A complete summary of all mobilograms and MS^2 data is provided in Figure S3 (supporting information). Noteworthy, LDI-IM-MS facilitated the differentiation of all four isomers (compound 1-4) of mix 4 (Figure 2A).

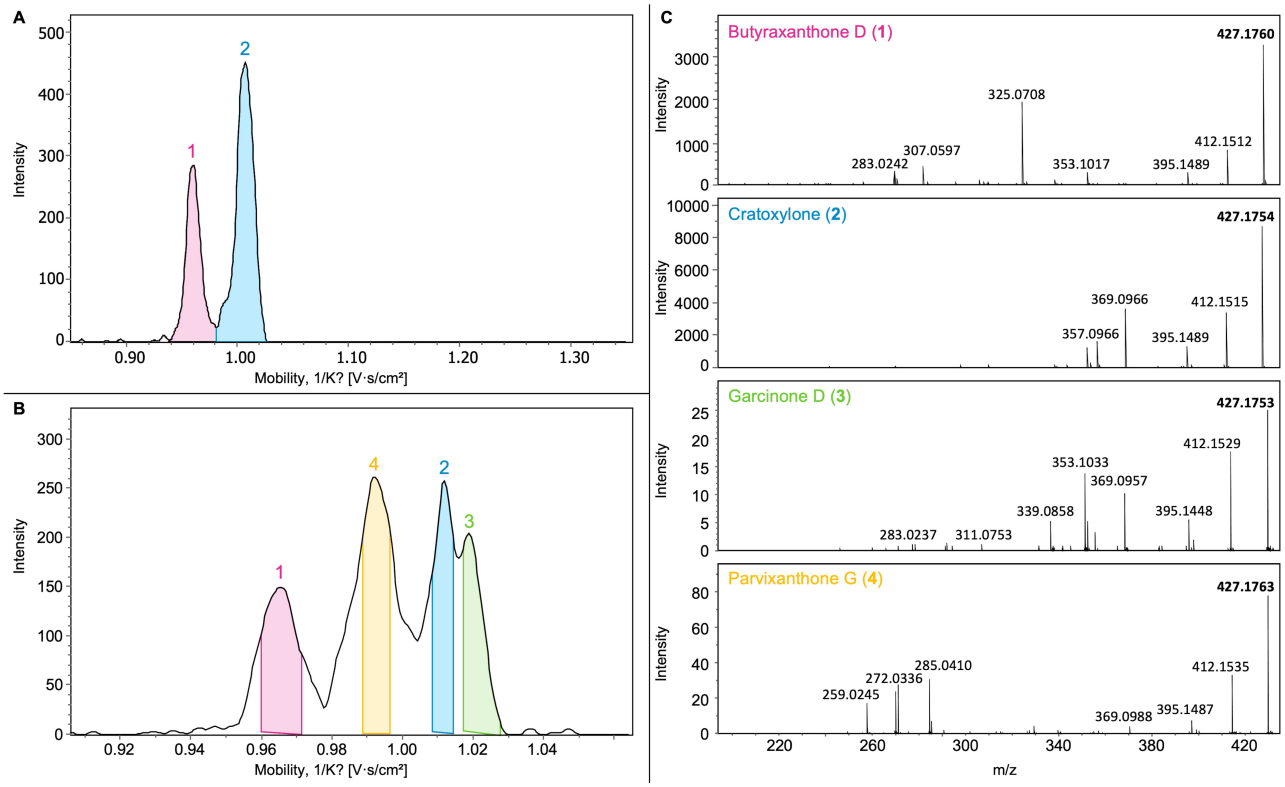


Figure 2. (A): Mobilogram of mix 1 (composed of butyraxanthone D (1) (pink) and cratoxylone (2) (blue)). (B): Mobilogram of mix 4 (composed of butyraxanthone D (1) (pink), parvixanthone G (4) (yellow), cratoxylone (2) (blue), and garcinone D (3) (green)). (C). LDI-IMS-MS^2 spectra of compounds 1-4.

A concurrently performed ultra-performance liquid chromatography (UPLC)-MS experiment failed to chromatographically separate compound 2 and 3 (Figure S4, supporting information). This underlines the outstanding potential of IMS and opens new perspectives for the application of LDI-IM-MS^2 such as molecular networking.

Exploring this possibility, the dry extract KL5073 was processed by flash chromatography and collected fractions were the analyzed by LDI-IM-MS^2. In parallel, LDI-IM-MS^2 spectra of compounds **1-10** were compiled in an *in-house* DB that was used to create a MN. As shown in figure 3A, MS^2 spectra allowed the putative annotation of compounds **1**, **4**, **5**, **7**. Additional proof to this prediction was added by observed CCS values, which were all in the range of reference compounds (isolated compounds from KL5073, **Table 1**). Finally, all LDI-IM-MS^2 spectra were uploaded to the Global Natural Products Social Molecular Networking platform (GNPS) and compiled in the Xanthones-DB (https://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=XANTHONES-DB).

The GNPS platform is open-access tool for sharing raw, processed, or annotated fragmentation MS^2 data. Comprising countless MS^2 data, the platform was used to improve current results and yielded the refined MN shown in **Figure 3B**. Compared to the initial MN (Figure 3A), it displays more nodes but also includes cosine scores, that may improve the precision of annotations. It should be noted that GNPS mostly comprises ESI-MS^2 data and that experimental fragmentation conditions are not uniform.

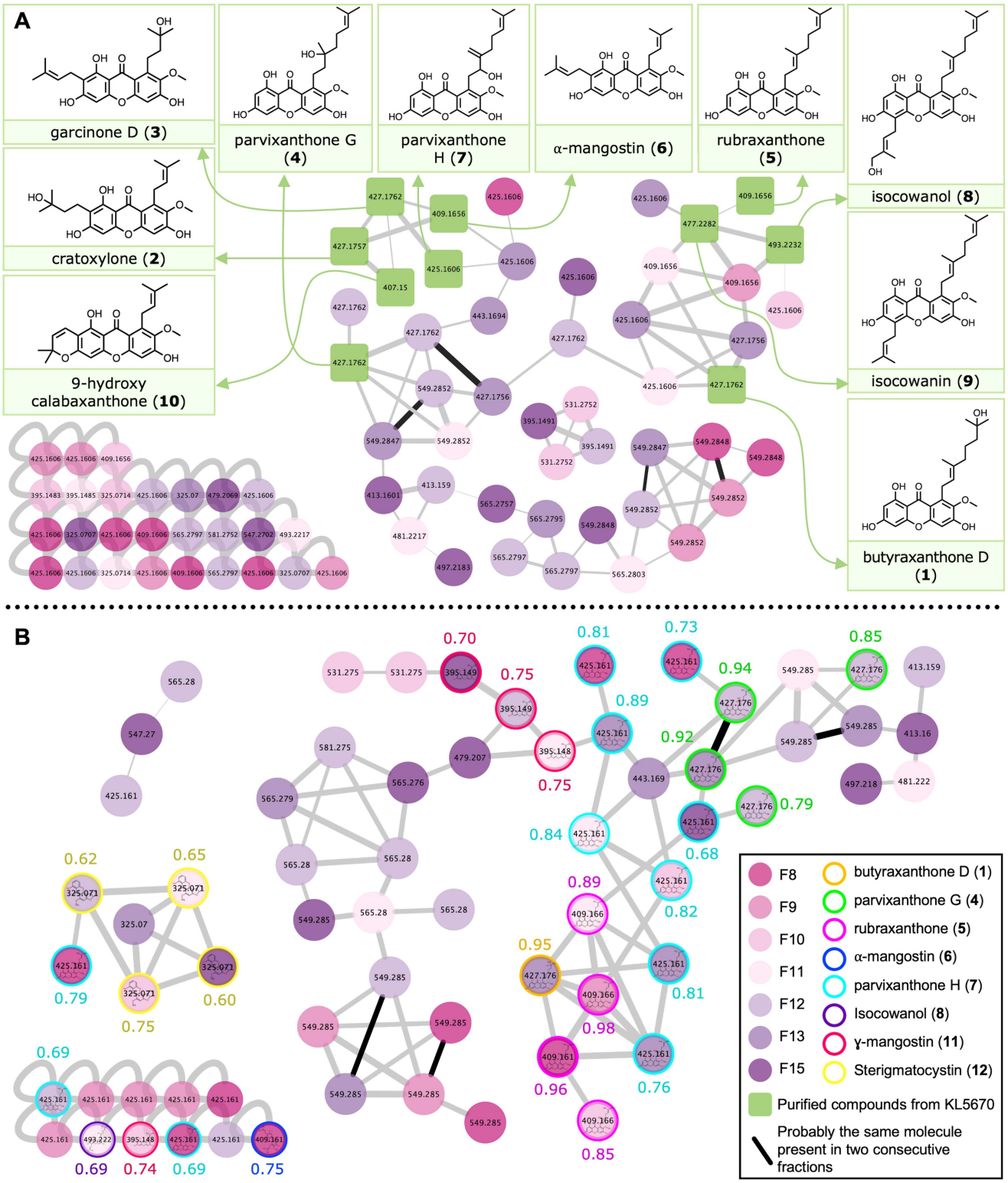


Figure 3. Molecular network based LDI-IMS-MS² data that were acquired from on fractions of KL5073 bark extracts. The color of the nodes represents the fraction in which the ions were found. Black colored edges potentially represent the same molecule detected in two consecutive fractions with the same CCS value. The green square nodes represent the isolated compounds from batch KL5670 (A). Circled nodes represent putative annotation by Xanthones-DB according to their cosine score (B).

This may impair the quality of annotations but is a general problem when comparing MS² spectra from different instruments. For the time being, most of the MS² annotations by the GNPS (**Figure 3B**) are based on our Xanthones-DB as only few data on xanthones were available on the platform. This should change in the future and may permit to further refine present results.

Conclusion

In recent years, IMS has gained significant importance for the characterization of complex mixtures of NPs.[30] While being a useful complement to LC-MS, IMS is a major methodological innovation for LDI-MS as it technically permits the separation of isomers without chromatography. Benefitting from this asset, the current work has provided a proof of concept for the feasibility of LDI-IM-MS^2 based molecular networking. This combined dereplication strategy can aid in the rapid detection and annotation of compounds in complex mixtures. Overall, the present works underlines most promising potential of LDI-IM-MS^2 as a powerful complement to standard LC-MS^2 in chemometrics.

AUTHOR INFORMATION

Corresponding Author

*AS: andreas.schinkovitz@univ-angers.fr; SD: severine.derbre@univ-angers.fr

Author Contributions

The manuscript was written through contributions of all authors. / All authors have given approval to the final version of the manuscript. / ‡These authors contributed equally. (Match statement to author names with a symbol)

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at:
Flash chromatogram of KL5073, LDI-IM-MS/MS of purified compounds and mixtures, UPLC-HRMS chromatogram of the coelution of garcinone D and cratoxylone.

ACKNOWLEDGMENT

This work was carried out in the framework of the International French Malaysian Natural Product Laboratory (IFM-NatPro-Lab) established between CNRS-ICSN and the University Malaya. The authors gratefully acknowledge the following institution for having provided financial support to the project: Le ministère de l'Europe et des Affaires Etrangères (MEAE), Campus France [PANASIA project, BioAsie, 2016- 2017, the DEMAiN project, PHC-Hibiscus, 2020-2022] and the Bavarian Ministry of State for Education, Culture, Science and Art (BayFrance grant no: FK-04-2022).

REFERENCES

(1) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016**, *34* (8), 828–837. https://doi.org/10.1038/nbt.3597.

(2) Aron, A. T.; Gentry, E. C.; McPhail, K. L.; Nothias, L.-F.; Nothias-Esposito, M.; Bouslimani, A.; Petras, D.; Gauglitz, J. M.; Sikora, N.; Vargas, F.; van der Hooft, J. J. J.; Ernst, M.; Kang, K. B.; Aceves, C. M.; Caraballo-Rodríguez, A. M.; Koester, I.; Weldon, K. C.; Bertrand, S.; Roullier, C.; Sun, K.; Tehan, R. M.; Boya P., C. A.; Christian, M. H.; Gutiérrez, M.; Ulloa, A. M.; Tejeda Mora, J. A.; Mojica-Flores, R.; Lakey-Beitia, J.; Vásquez-Chaves, V.; Zhang, Y.; Calderón, A. I.; Tayler, N.; Keyzers, R. A.; Tugizimana, F.; Ndlovu, N.; Aksenov, A. A.; Jarmusch, A. K.; Schmid, R.; Truman, A. W.; Bandeira, N.; Wang, M.; Dorrestein, P. C. Reproducible Molecular Networking of Untargeted Mass Spectrometry Data Using GNPS. *Nat. Protoc.* **2020**, *15* (6), 1954–1991. https://doi.org/10.1038/s41596-020-0317-5.

(3) Fox Ramos, A. E.; Evanno, L.; Poupon, E.; Champy, P.; Beniddir, M. A. Natural Products Targeting Strategies Involving Molecular Networking: Different Manners, One Goal. *Nat. Prod. Rep.* **2019**, *36* (7), 960–980. https://doi.org/10.1039/C9NP00006B.

(4) Qin, G.-F.; Zhang, X.; Zhu, F.; Huo, Z.-Q.; Yao, Q.-Q.; Feng, Q.; Liu, Z.; Zhang, G.-M.; Yao, J.-C.; Liang, H.-B. MS/MS-Based Molecular Networking: An Efficient Approach for Natural Products Dereplication. *Molecules* **2022**, *28* (1), 157. https://doi.org/10.3390/molecules28010157.

(5) Islam, S.; Alam, R.; Kim, S. Improved Coverage of Plant Metabolites Using Powder Laser Desorption/Ionization Coupled with Fourier-Transform Ion Cyclotron Mass Spectrometry. *Food Chem.* **2022**, *373*, 131541. https://doi.org/10.1016/j.foodchem.2021.131541.

(6) Le Pogam, P.; Schinkovitz, A.; Legouin, B.; Le Lamer, A.-C.; Boustie, J.; Richomme, P. Matrix-Free UV-Laser Desorption Ionization Mass Spectrometry as a Versatile Approach for Accelerating Dereplication Studies on Lichens. *Anal. Chem.* **2015**, *87* (20), 10421–10428. https://doi.org/10.1021/acs.analchem.5b02531.

(7) Schinkovitz, A.; Boisard, S.; Freuze, I.; Osuga, J.; Mehlmer, N.; Brück, T.; Richomme, P. Matrix-Free Laser Desorption Ionization Mass Spectrometry as a Functional Tool for the Analysis and Differentiation of Complex Phenolic Mixtures in Propolis: A New Approach to Quality Control. *Anal. Bioanal. Chem.* **2018**, *410* (24), 6187–6195. https://doi.org/10.1007/s00216-018-1225-1.

(8) Skopikova, M.; Hashimoto, M.; Richomme, P.; Schinkovitz, A. Matrix-Free Laser Desorption Ionization Mass Spectrometry as an Efficient Tool for the Rapid Detection of Opiates in Crude Extracts of Papaver Somniferum. *J. Agric. Food Chem.* **2020**, *68* (3), 884–891. https://doi.org/10.1021/acs.jafc.9b05153.

(9) Le Pogam, P.; Richomme, P.; Beniddir, M. A.; Duong, T.-H.; Bernadat, G.; Schinkovitz, A. A Thorough Evaluation of Matrix-Free Laser Desorption Ionization on Structurally Diverse Alkaloids and Their Direct Detection in Plant Extracts. *Anal. Bioanal. Chem.* **2020**, *412* (27), 7405–7416. https://doi.org/10.1007/s00216-020-02872-6.

(10) Meunier, M.; Bréard, D.; Awang, K.; Boisard, S.; Guilet, D.; Richomme, P.; Derbré, S.; Schinkovitz, A. Matrix Free Laser Desorption Ionization Assisted by ^{13}C NMR Dereplication: A Complementary Approach to LC-MS^2 Based Chemometrics. *Talanta* **2023**, *253*, 123998. https://doi.org/10.1016/j.talanta.2022.123998.

(11) Hines, K. M.; Ross, D. H.; Davidson, K. L.; Bush, M. F.; Xu, L. Large-Scale Structural Characterization of Drug and Drug-like Compounds

by High-Throughput Ion Mobility-Mass Spectrometry. *Anal. Chem.* **2017**, *89* (17), 9023–9030. https://doi.org/10.1021/acs.analchem.7b01709.

(12) Hofmann, J.; Hahm, H. S.; Seeberger, P. H.; Pagel, K. Identification of Carbohydrate Anomers Using Ion Mobility–Mass Spectrometry. *Nature* **2015**, *526* (7572), 241–244. https://doi.org/10.1038/nature15388.

(13) Zhang, X.; Quinn, K.; Cruickshank-Quinn, C.; Reisdorph, R.; Reisdorph, N. The Application of Ion Mobility Mass Spectrometry to Metabolomics. *Curr. Opin. Chem. Biol.* **2018**, *42*, 60–66. https://doi.org/10.1016/j.cbpa.2017.11.001.

(14) Hädener, M.; Kamrath, M. Z.; Weinmann, W.; Groessl, M. High-Resolution Ion Mobility Spectrometry for Rapid Cannabis Potency Testing. *Anal. Chem.* **2018**, *90* (15), 8764–8768. https://doi.org/10.1021/acs.analchem.8b02180.

(15) Meunier, M.; Bréard, D.; Boisard, S.; Blanchard, P.; Litaudon, M.; Awang, K.; Schinkovitz, A.; Derbré, S. Looking for Actives in the Haystack: Merging HRMS^2-Based Molecular Networking, Chemometrics and ^{13}C NMR-Based Dereplication Approaches. *Subtmitted in J. Nat. Prod. (unpublished work)* **2022**.

(16) Lopes, C. T.; Franz, M.; Kazi, F.; Donaldson, S. L.; Morris, Q.; Bader, G. D. Cytoscape Web: An Interactive Web-Based Network Browser. *Bioinformatics* **2010**, *26* (18), 2347–2348. https://doi.org/10.1093/bioinformatics/btq430.

(17) Failla, M. L.; Gutiérrez-Orozco, F. Mangosteen Xanthones: Bioavailability and Bioactivities. *Fruit and Veg. Phytochem.: Chem. and Hum. Health* **2017**, 165–182. https://doi.org/10.1002/9781119158042.ch9.

(18) Dharmaratne, H. R. W.; Napagoda, M. T.; Tennakoon, S. B. Xanthones from Roots of *Calophyllum Thwaitesii* and Their Bioactivity. *Nat. Prod. Res.* **2009**, *23* (6), 539–545. https://doi.org/10.1080/14786410600899118.

(19) Liu, Q.-Y.; Wang, Y.-T.; Lin, L.-G. New Insights into the Anti-Obesity Activity of Xanthones from *Garcinia Mangostana*. *Food Funct.* **2015**, *6* (2), 383–393. https://doi.org/10.1039/C4FO00758A.

(20) Permana, D.; Lajis, N. Hj.; Mackeen, M. M.; Ali, A. M.; Aimi, N.; Kitajima, M.; Takayama, H. Isolation and Bioactivities of Constitutents of the Roots of *Garcinia Atroviridis*. *J. Nat. Prod.* **2001**, *64* (7), 976–979. https://doi.org/10.1021/np000563o.

(21) Ishaque, M.; Bibi, Y.; Qayyum, A.; Iriti, M. Isolation and Structural Confirmation of Xanthone Isomers from *Dryopteris Ramosa* (Hope) C. Chr. and Their in Vitro Antioxidant Mechanism. *Arab J Sci Eng* **2021**, *46* (6), 5327–5337. https://doi.org/10.1007/s13369-020-05097-y.

(22) Li, H.; Chen, T.; Sun, J.; Wang, W.; Li, Y. Separation of Six Xanthones from *Swertia Franchetiana* by High-Speed Countercurrent Chroatography. *J. Sep. Sci.* **2017**, *40* (11), 2515–2521. https://doi.org/10.1002/jssc.201601134.

(23) Hinnenkamp, V.; Klein, J.; Meckelmann, S. W.; Balsaa, P.; Schmidt, T. C.; Schmitz, O. J. Comparison of CCS Values Determined by Traveling Wave Ion Mobility Mass Spectrometry and Drift Tube Ion Mobility Mass Spectrometry. *Anal. Chem.* **2018**, *90* (20), 12042–12050. https://doi.org/10.1021/acs.analchem.8b02711.

(24) Zhou, Z.; Xiong, X.; Zhu, Z.-J. MetCCS Predictor: A Web Server for Predicting Collision Cross-Section Values of Metabolites in Ion Mobility-Mass Spectrometry Based Metabolomics. *Bioinformatics* **2017**, *33* (14), 2235–2237. https://doi.org/10.1093/bioinformatics/btx140.

(25) Zhou, Z.; Luo, M.; Chen, X.; Yin, Y.; Xiong, X.; Wang, R.; Zhu, Z.-J. Ion Mobility Collision Cross-Section Atlas for Known and Unknown Metabolite Annotation in Untargeted Metabolomics. *Nat. Commun.* **2020**, *11* (1), 4334. https://doi.org/10.1038/s41467-020-18171-8.

(26) Ross, D. H.; Cho, J. H.; Xu, L. Breaking down Structural Diversity for Comprehensive Prediction of Ion-Neutral Collision Cross Sections. *Anal. Chem.* **2020**, *92* (6), 4548–4557. https://doi.org/10.1021/acs.analchem.9b05772.

(27) Colby, S. M.; Thomas, D. G.; Nuñez, J. R.; Baxter, D. J.; Glaesemann, K. R.; Brown, J. M.; Pirrung, M. A.; Govind, N.; Teeguarden, J. G.; Metz, T. O.; Renslow, R. S. ISiCLE: A Quantum Chemistry Pipeline for Establishing in Silico Collision Cross Section Libraries. *Anal. Chem.* **2019**, *91* (7), 4346–4356. https://doi.org/10.1021/acs.analchem.8b04567.

(28) Plante, P.-L.; Francovic-Fontaine, É.; May, J. C.; McLean, J. A.; Baker, E. S.; Laviolette, F.; Marchand, M.; Corbeil, J. Predicting Ion Mobility Collision Cross-Sections Using a Deep Neural Network: DeepCCS. *Anal. Chem.* **2019**, *91* (8), 5191–5199. https://doi.org/10.1021/acs.analchem.8b05821.

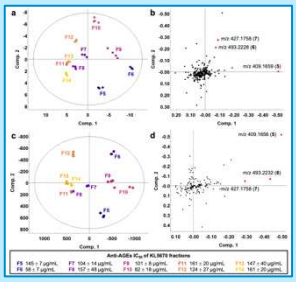
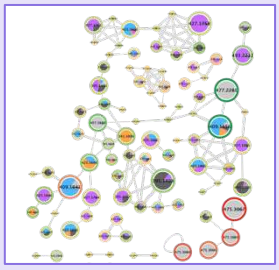
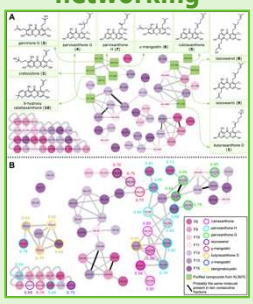
(29) Jariyasopit, N.; Limjiasahapong, S.; Kurilung, A.; Sartyoungkul, S.; Wisanpitayakorn, P.; Nuntasaen, N.; Kuhakarn, C.; Reutrakul, V.; Kittakoop, P.; Sirivatanauksorn, Y.; Khoomrung, S. Traveling Wave Ion Mobility-Derived Collision Cross Section Database for Plant Specialized Metabolites: An Application to *Ventilago Harmandiana* Pierre. *J. Proteome Res.* **2022**, *21* (10), 2481–2492. https://doi.org/10.1021/acs.jproteome.2c00413.

(30) Xia, J.; Xiao, W.; Lin, X.; Zhou, Y.; Qiu, P.; Si, H.; Wu, X.; Niu, S.; Luo, Z.; Yang, X. Ion Mobility-Derived Collision Cross-Sections Add Extra Capability in Distinguishing Isomers and Compounds with Similar Retention Times: The Case of Aphidicolanes. *Mar. Drugs* **2022**, *20* (9), 541. https://doi.org/10.3390/md20090541.

6. Conclusion

My PhD aimed to develop new approaches to identify (bio)markers in complex mixtures of NPs. Overall three different biochemometric and/or dereplication strategies based on LDI-MS or LDI-(IM)-MS/MS, UPLC-MS/MS and ^{13}C NMR were developed (Table 4).

Table 4. Advantages and disadvantages of analytical tools

	Chemometrics combined with ^{13}C NMR 		Chemometrics combined with molecular networks and ^{13}C NMR 	Ion mobility based molecular networking 
Analytical tool	**LDI-MS**	**^{13}C NMR**	**UPLC-MS/MS**	**LDI-IM-MS/MS**
Advantages	- Little to no sample preparation - Permits the use of any volatile solvent - Small sample quantities - Instant results (seconds)	- Identification of organic compounds (including constitutional isomers)	- Chromatographic separation - Separation of some isomers - Many data processing tools	- All advantages of LDI-MS + separation of isomers
Disadvantages	- No differentiation of isomers - Very few data processing tools (JEOL)	- Not sensitive (major compounds only)	- Some isomers not distinguished - Solvent limitation (ACN, MeOH, H_2O) - Solvent quantity - Long method development - Sample preparation	- MS/MS less sensitive (depending on MS^1 ionization)

LDI-MS based chemometrics offer many advantages, including an easy experimental protocol, small quantity of solvents and sample material as well as instant results. In addition, this strategy has demonstrated its efficiency for detecting anti-AGEs compounds from crude extracts of *G. parvifolia* (KL5670) prior to their isolation. However, due to the absence of chromatographic separation, the method cannot differentiate between isomers. Supplementary ^{13}C NMR experiments and the use of the MixONat ^{13}C NMR dereplication tool could bypass some of these issues. However, no existing dereplication tools or software that is adapted to process LDI-MS data. The development of such a software tool is an objective of a follow up project conducted in collaboration with the LERIA research team (University of Angers).

The second strategy proposed uses the combination of UPLC-MS/MS based MN, chemometrics and ^{13}C NMR allowed for early annotation of active NPs in complex mixtures, including isomers. In this context, section 4.1

highlighted the importance of confidence levels for correct compound annotations. To refine the process, level 1 (isolated compound) was divided into two subcategories: level 1a (convergent structure proposal using two orthogonal methods confirmed by literature-based matching) and level 1b (convergent structure proposal using two orthogonal methods). This way the precision and comprehensiveness of annotations could be significantly improved.

As a third strategy, the LDI-IM-MS/MS approach addressed the initial challenge of differentiating isomers in LDI-MS experiments. These efforts have resulted in the creation of a MN that is exclusively based on LDI-IM-MS/MS data (section 5.2, Figure 3). In addition, all MS/MS data were uploaded on the GNPS platform and compiled in the Xanthones-DB (in combination with the UPLC-MS/MS data at three different CE). This permitted the creation of refined MN shown in (section 5.2, Figure 3).

A significant part of the thesis focused on developing an optimized workflow based on the availability of instruments and (open-source) software. In this respect, many problems concerning the compatibility of original instrument files, with open-source post acquisition processing software were encountered. For example: LDI-MS data files (.tas, JEOL) were incompatible with the mMass software and had to be first converted into the csv. format. The mMass software did not permit spectra alignment, so a specific algorithm had to be developed by the LERIA team (University of Angers). Eventually most of these limitations could be resolved and a functional workflow from raw to processed data was established.

By today, LDI-MS and UPLC-MS based chemometrics are highly efficient tools for targeting active compounds in complex mixtures. However, the unequivocal identification of these compounds remains a significant challenge as data misinterpretation can easily occur. Therefore, the careful integration of multiple sources of information is required to increases the confidence level and the overall reliability of compound annotation in chemometrics.

As a future perspective, the interpretation of MS data could be assisted by artificial intelligence (AI), which would allow an optimized processing and interpretation of raw data. This approach could facilitate an efficient automatized strategy for the identification of known and unknown compounds from complex mixtures, with high level of confidence.

7. Perspectives

7.1. Additional discussion

Despite these applications, the challenges in accessing sufficient material for the isolation and characterization of novel bioactive metabolites remains a challenging task. Indeed, we face an alarmingly rate of species extinction due to factors such as habitat destruction and overharvesting that poses a significant threat to the availability of diverse plant species. This calls into question the ability to obtain large quantities of plants from natural populations and raises concerns about the conservation of biodiversity.

In addition to the challenges posed by declining plant populations, the rules governing the legal and ethical frameworks related to benefit sharing adds a layer of complexity. The Convention on biological diversity and the Nagoya protocol (Harrison, 2014; Convention on Biological Diversity, 2023) provide guidelines for the fair and

equitable sharing of benefits arising from the use of biological resources. These rules aim to ensure that countries of origin are appropriately recognized and compensated for the resources and knowledge they contribute.

In response to these challenges, there is a growing emphasis on implementing efficient and "sustainable" practices in the search for bioactive compounds. This involves the development of innovative techniques for the isolation and/or synthesis of target compounds. Collaboration between researchers, industry, and local communities is thus crucial to overcome the challenges associated with accessing plant material. This will not only contribute to the advancement of scientific knowledge but also promote the conservation of biodiversity and respect for the rights of communities involved.

7.2. Reanalysis of dataset from reusable plant extracts analysis

In this manuscript, the extracts were selected based on their chemical diversity and complexity for proof-of-concept purposes. As a result, they were chosen for their simplified chemistry and close similarity. Even though they have different activities. However, researchers interested in exploring extensive and diverse extracts must establish effective strategies to efficiently navigate this chemical complexity and gain access to novel (bio)active compounds. To address this challenge, P.-M. Allard *et al.* (Allard *et al.*, 2022), have proposed the creation of comprehensive spectral data as well as bioactivity results from extract collection to prioritize the time-consuming task of isolating bioactive compounds.

The authors report on their initial exploration of a collection of 1,600 plant extracts as part of a drug discovery program. They provide insights into the taxonomic coverage of the collection and present the results of HRMS profiling. In addition, the authors propose data reuse cases for other researchers interested in the computational exploration of NPs. With this in mind, the authors have submitted their dataset to the reanalysis of data user (ReDU) interface. As explained in section 1.2, ReDU is a community-based approach aimed at finding and reusing public data, including MS/MS data, on a repository scale.

In this context, the data obtained from the two batches of *G. parvifolia* (KL5670 and KL5073), together with the Pierre Fabre *Garcinia* extracts were merged to create a single MN (see section 2.5.5.3) (Figure 23 and Figure 24, Appendix 23 to Appendix 26). The objective is to use the spectral data obtained through this thesis (*i.e.*, MS^2 data from isolated or annotated xanthones) to annotate ions in the *Garcinia* extracts from Pierre Fabre.

As stated above (section 2.8.4), MS^2 data of isolated and characterized xanthones in this thesis were deposited in GNPS. Therefore, these xanthones were first spotted in the MN (Figures 22 and 23, green nodes) to **explore the annotation done by the GNPS**[7] with a given cosine score. The most interesting clusters is shown in Figure 23. No ions were annotated in Pierre Fabre's *Garcinia* extracts[18], suggesting the absence of the isolated xanthones in these *Garcinia* extracts. However, this work has made it possible to annotate other ions in the KL5073 extract

[18] According to GNPS dereplication results, only one compound from the Pierre Fabre *Garcinia* extract (highlighted in red, Figure 23) was putatively identified as γ-mangostin with a cosine score of 0.83.

of *Garcinia parvifolia* (*i.e.*, rubraxanthone, Appendix 23). As expected, ions comprised in Xanthones-DB were all annotated in KL5670.

Subsequently, annotated xanthones (section 4.1.1, Table 1) were also detected in KL5670 nodes in the MN (purple nodes) in an attempt to propagate the annotations in Pierre Fabre's *Garcinia* extracts (blue nodes) (Figure 23). Node 2 from KL5670 (Figure 23) was annotated as cheffouxanthone (level 1b, section 4.1.1, Table 1). Close to node 2, node 50 (highlighted in red) was annotated by GNPS as γ-mangostin (level 2)[8]. Regarding the level of annotation, node 50 should be an analogue of cheffouxanthone. So far, the limited number of xanthones or xanthone derivatives currently registered in GNPS does not allow a clear assignment of this compound.

In a second approach, the *in-house* Xanthones-DB[7] was explored for **annotation using ProgenesisQI's Metascope tool** (see section 2.8.4.3). It allowed the putative annotation (level of confidence 2) of the *Garcinia* extracts (Pierre Fabre) (Table 5,). The complete MN and annotation table is provided in Appendix 27. Nodes 29, 30, 33, 37, and 38 in Figure 24 displayed fragmentation scores ranging from 67.4% to 78.1% for compound WR3[19] (Figure 22). Notably, these nodes were also **annotated by the GNPS library** as compounds from Xanthones-DB as the same molecule, with similar cosine scores ranging from 0.67 to 0.73. Consequently, both annotation tools (*i.e.*, ProgenesisQI Metascope and GNPS) using the same MS^2 data suggest the same annotations. However, only one of these nodes could correspond to WR3 and the other nodes could be closely related structures. Since their monoisotopic masses are identical, these NPs are necessarily isomers.

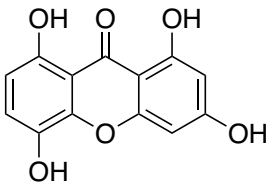


Figure 22. Structure of WR3 (1,3,5,8-tetrahydroxyxanthone) from *Canscora decussata*

In conclusion, the MS^2 data from this thesis have been compiled in a DB and made available to the community (see section 2.8). They were used to annotate some NPs of other *Garcinia* species from the Pierre Fabre collection. The cluster obtained using the MN, combining the extracts of this thesis (KL5670 and KL5073) and those of *Garcinia* species (Pierre Fabre), provides information on the structural similarity of the NPs (Figure 23). Many nodes remain to be annotated. This task is accessible through the sharing of experimental data by the entire scientific community. In the meantime, the work developed in this thesis proposes alternative workflows for annotating NPs more quickly and with a higher level of confidence.

[19] In parallel with this thesis work, various xanthones (previously isolated at SONAS Lab from *Canscora decussata* (Gentianaceae) were analyzed by UPLC-$HRMS^2$ in addition to the xanthones isolated from *Garcinia parvifolia* (KL5670). They were deposited on the GNPS. All these data also constitute an ***in-house* Xanthones-DB** that can be used directly with ProgenesisQI and its MetaScope tool.

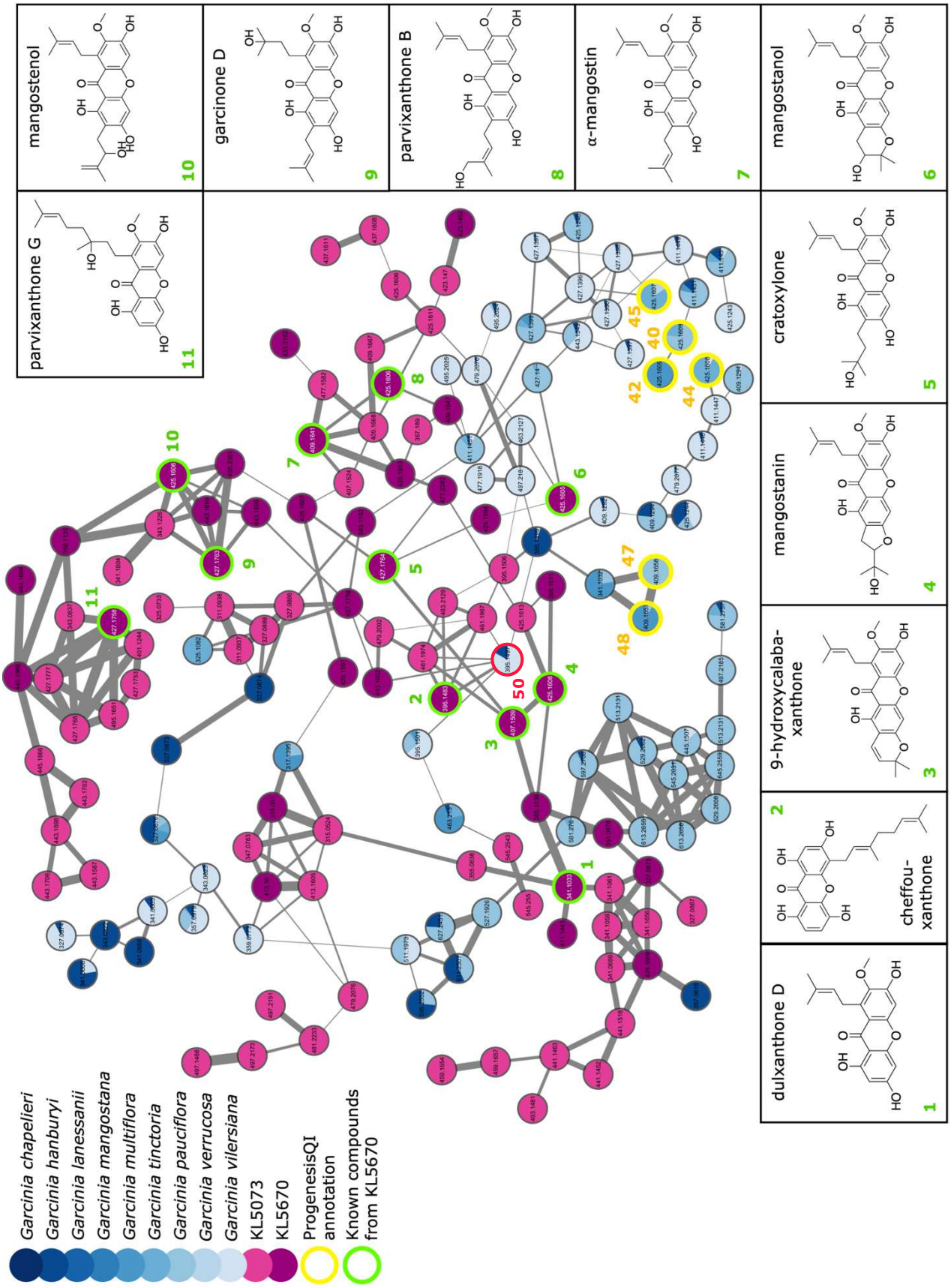


Figure 23. Molecular network (part 1) of *G. parvifolia* batches (KL5670 (dark pink) and KL5073 (light pink)) along with the previously identified molecules from KL5670 (green) and Pierre Fabre *Garcinia* extracts (blue gradient).

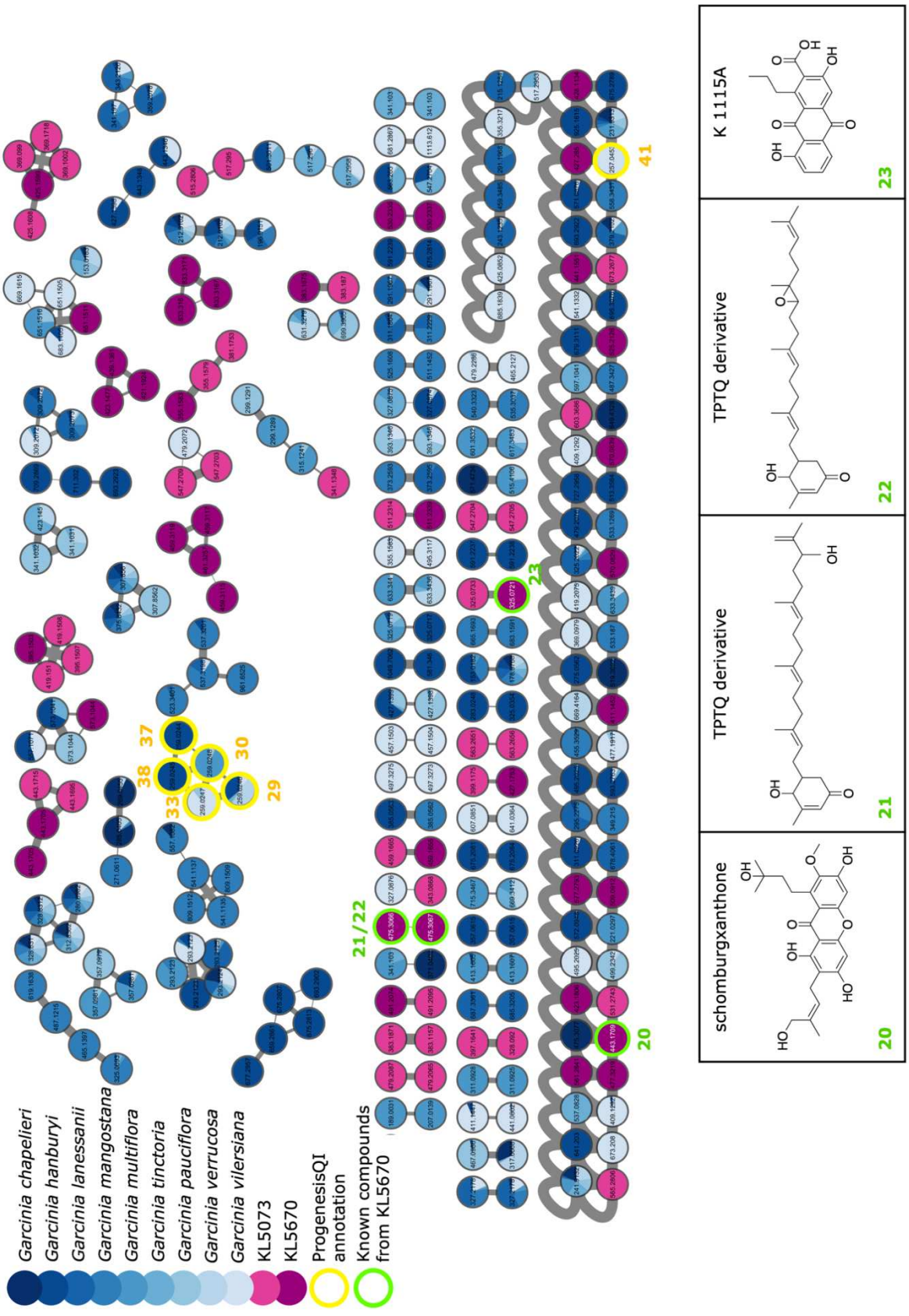


Figure 24. Molecular network (part 2) of *G. parvifolia* batches (KL5670 (dark pink) and KL5073 (light pink)) along with the previously identified molecules from KL5670 (green) and Pierre Fabre *Garcinia* extracts (blue gradient).

Table 5. Part of compound annotation of the MN based on KL5670, KL5073 and Pierre Fabre *Garcinia* extracts by Progenesis MetaScope and the *in-house* XanthonesDB

Network number	m/z	Retention time (min)	Score	Fragmentation Score	Compound ID
29	259.0246	2.04	33.3	70.1	WR3_NEG30
			31.4	60.8	WR3_NEG20
			21.1	6.9	WR3_NEG40
30	259.0248	2.23	34.6	76.1	WR3_NEG30
			31.6	61.5	WR3_NEG20
			23.1	16.5	WR3_NEG40
33	259.0247	2.62	35.0	78.1	WR3_NEG30
			32.4	65.4	WR3_NEG20
			21.6	9.4	WR3_NEG40
37	259.0244	3.09	32.6	67.4	WR3_NEG30
			31.1	60.6	WR3_NEG20
			23.2	18.8	WR3_NEG40
38	259.0245	3.18	33.3	70.6	WR3_NEG30
			31.7	62.7	WR3_NEG20
			23.6	20.2	WR3_NEG40
40	425.1609	4.02	29.5	51.7	parvixanthone H_NEG30
			27.9	44.0	parvixanthone H_NEG20
			23.9	23.0	parvixanthone H_NEG40
42	425.1609	4.16	21.3	10.1	parvixanthone H_NEG40
			20.8	8.1	parvixanthone H_NEG30
			19.4	1.6	parvixanthone H_NEG20
44	425.1608	4.43	32.1	64.4	parvixanthone H_NEG30
			31.8	63.0	parvixanthone H_NEG20
			21.3	9.3	parvixanthone H_NEG40
45	425.1607	4.71	29.2	49.5	parvixanthone H_NEG30
			27.5	41.3	parvixanthone H_NEG20
			20.5	5.2	parvixanthone H_NEG40
47	409.1658	5.50	30.2	53.7	alpha-mangostin_NEG20
			30.3	52.8	alpha-mangostin_NEG30
			29.1	47.3	alpha-mangostin_NEG40
			24.4	26.0	rubraxanthone_NEG20
			24.3	23.6	rubraxanthone_NEG30
			21.4	9.7	rubraxanthone_NEG40
48	409.1657	5.68	28.5	45.2	alpha-mangostin_NEG20
			28.6	44.4	alpha-mangostin_NEG30
			28.4	43.6	alpha-mangostin_NEG40
			24.0	23.7	rubraxanthone_NEG20
			23.7	20.7	rubraxanthone_NEG30
			21.2	8.9	rubraxanthone_NEG40
			39.1	0.0	garcinone D_NEG20

8. Résumé en français

8.1. Context

Les travaux présentés dans ce manuscrit ont été réalisés au sein du laboratoire "Substances d'Origine Naturelle et Analogues Structuraux" (SONAS). Cette unité de recherche se concentre sur l'analyse, la caractérisation et la valorisation de produits naturels (PNs) d'origine végétale et d'extraits d'intérêt pour l'Homme [par exemple, des composés anti-inflammatoires : dérivés de tocotriénols (Alsabil et al., 2016, 2017 ; Pein et al., 2018 ; Dinh et al, 2020), des agents anti-tartre écologiques : dérivés de l'acide cinnamique (Richomme et al., 2017), des inhibiteurs des produits finaux de glycation avancée (AGE) : flavonoïdes et acides phénoliques (Derbré et al., 2010)], ainsi que pour la santé animale (LabCom FeedIn Tech) et la santé végétale [alkaloïdes et xanthones potentialisant les défenses des plantes (Simoneau et al., 2014)].

Dans ce contexte, l'un des objectifs du laboratoire est de développer des stratégies permettant l'identification rapide de PNs dans des mélanges complexes. Dans un premier temps, les travaux d'A. Bruguière *et al.*, ont permis de développer un algorithme de déréplication basé sur la résonance magnétique nucléaire (RMN) du carbone 13, qui a donné naissance au logiciel MixONat (Bruguière et al., 2020). Ce dernier permet l'annotation rapide des PNs majoritaires dans des mélanges complexes (*i.e.*, extraits bruts, fractions).

Par ailleurs, le laboratoire a acquis en 2020 un système de chromatographie liquide ultra-performante couplé à un spectromètre de masse quadripolaire à temps de vol (UPLC-QTOF). Cet instrument permet de détecter des traces de métabolites dans des mélanges complexes et est équipé d'un détecteur de spectromètre de masse à haute résolution. Les données peuvent ensuite être traitées à l'aide du logiciel ProgenesisQI, acquis simultanément. Un avantage de ce flux de travail est la possibilité de générer un réseau moléculaire (MN). Cette technique de déréplication basée sur la spectrométrie de masse (SM) haute résolution et la fragmentation des ions (MS^2) permet également l'annotation rapide des PNs dans des mélanges complexes.

En tant que stratégies complémentaires, la SM et la RMN permettent aujourd'hui d'éviter les étapes longues et coûteuses de purification et analyses structurales des PNs afin de les identifier précocement grâce à la déréplication. Il est nécessaire de noter que, malgré ses performances analytiques puissantes et sa large application, la chromatographie liquide couplée à la spectrométrie de masse (LC-MS) présente encore certaines limites, telles que les restrictions liées aux solvants, le développement fastidieux des méthodes chromatographiques et la préparation des échantillons.

C'est pourquoi des études sur l'ionisation par désorption laser couplée à la spectrométrie de masse sans matrice (LDI-MS) ont été développées en laboratoire, car elles permettent l'analyse rapide et sensible d'un mélange complexe sans préparation fastidieuse de l'échantillon (Le Pogam et al., 2015 ; Skopikova et al., 2020). De nombreux PNs présentent de fortes similarités structurales avec les composés utilisés comme matrices dans l'ionisation par désorption laser assistée par matrice (MALDI) et peuvent donc être facilement ionisées par les lasers UV-LDI. Ainsi, cette méthode a été largement utilisée en laboratoire, car elle est applicable à un large éventail de PNs chimiquement divers, telles que les alcaloïdes (Schinkovitz et al., 2018 ; Le Pogam et al., 2020 ; Skopikova et al., 2020), les composés phénoliques ou les cannabinoïdes (dos Santos et al., 2019).

Dans ce contexte, l'objectif ultime de ma thèse est de discuter des différents outils disponibles afin d'identifier rapidement des composés (bio)actifs dans des mélanges complexes.

8.2. Introduction

L'étude des PNs présente un intérêt croissant dans de nombreux domaines parmi lesquels la santé et l'alimentation. Ces métabolites spécialisés sont de petites molécules produites par des organismes vivants (plantes, animaux ou microorganismes), provenant de sources terrestres ou marines. Contrairement aux métabolites primaires, ils sont impliqués dans diverses fonctions biologiques (par exemple, la survie ou la reproduction) qui ne sont pas liées à la croissance, au maintien des activités cellulaires et à la vie cellulaire. Ce manuscrit se concentre sur les PNs produits par les plantes. Dans le règne végétal, ils peuvent être divisés en trois grandes classes basées sur la voie de biosynthèse sous-jacente : les composés phénoliques (c'est-à-dire les polyphénols), les terpènes et les stéroïdes, et les composés contenant de l'azote (c'est-à-dire les alcaloïdes) (Twaij et Hasan, 2022). Par rapport aux composés synthétiques (CS), les PNs se caractérisent par une plus grande diversité chimique et leur grande complexité structurale, en particulier en ce qui concerne les aspects stéréochimiques (par exemple, les PNs ont tendance à avoir un plus grand nombre de centres chiraux). De plus, les PNs présentent un grand nombre d'atomes de carbone et d'oxygène, mais moins d'atomes d'azote et d'halogène par rapport aux CS (Clardy et Walsh, 2004).

Ainsi, les PNs ont joué un rôle important tout au long de l'histoire. Naturellement optimisés par l'évolution, ils sont connus pour exercer une large gamme d'effets biologiques très intéressants et ont fourni des pistes précieuses pour diverses applications, notamment en thérapeutique. Ils ont ont joué un rôle clé dans la découverte de médicaments, notamment contre les cancers ou les maladies infectieuses (Atanasov et al., 2015 ; Harvey, Edrada-Ebel et Quinn, 2015). Par exemple, la vancomycine, la quinine ou la morphine (Wilson et Danishefsky, 2006) tiennent toujours une place très importante dans les thérapeutiques médicamenteuses actuelles. Les PNs sont également largement utilisés pour des applications cosmétiques ou alimentaires (activités antioxydante (Lourenço, Moldão-Martins et Alves, 2019) ou antimicrobienne (Gyawali et Ibrahim, 2014)). De plus, les chercheurs ont identifié leur intérêt dans de nouvelles stratégies de biocontrôle ou biostimulation en agriculture (Jamiołkowska, 2020). Ces produits ont des propriétés bioactives qui protègent les plantes contre les pathogènes (Jamiołkowska, 2020 ; Efenberger-Szmechtyk, Nowak et Czyzowska, 2021) et peuvent également contrôler la croissance des plantes (gibbérellines (Hernández-García, Briones-Moreno et Blázquez, 2021 ; Hong, Jang et Ryu, 2021 ; Wu et al., 2021).

Encore aujourd'hui, le fractionnement bioguidé (FBG) est utilisé de manière conventionnelle pour la découverte de ces nouveaux métabolites (bio)actifs (Cunha et al., 2017). En général, une partie de la plante (fleurs, fruits, racines, tiges ou feuilles) est extraite et évaluée pour son activité biologique. Comme les extraits de plantes sont très complexes et contiennent différents types de métabolites, ceux qui présentent la meilleure (bio)activité subissent ensuite plusieurs étapes de fractionnement successives, chacune suivie d'un test biologique ou chimique. En conséquence, la complexité de l'échantillon initial est progressivement réduite, jusqu'à l'isolement d'un ou plusieurs PNs actifs. Une fois isolé(s), la caractérisation chimique complète d'un métabolite est généralement réalisée en utilisant une combinaison de techniques analytiques. Celles-ci incluent la spectrométrie de masse haute résolution (HRMS) et la RMN, ainsi que des méthodes pour déterminer la configuration absolue

du PN. Des techniques telles que la dichroïsme circulaire électronique (ECD) ou vibratoire (VCD), la cristallographie aux rayons X et des expériences de polarimètre (valeur [α]D) peuvent être utilisées à cette fin. Ce procédé a conduit à l'isolement et à l'élucidation de la structure de nombreux composés bioactifs (par exemple, la quinine, la morphine, la vinblastine, etc.) (Rates, 2001 ; Christensen, 2021).

Cependant, parce que le FBG nécessite un suivi continu de l'activité après chaque étape de séparation, il réduit considérablement la quantité d'échantillon. En outre, il comporte le risque d'isolement de composés précédemment décrits dans la littérature (Nahrstedt et Butterweck, 2010 ; Inui et al., 2012). Il peut également entraîner une perte d'activité due à une liaison irréversible des composants à la phase stationnaire du système chromatographique utilisé. Certaines phases stationnaires (réactives) (e.g. la silice) peuvent également induire une dégradation. De plus, le FBG est généralement biaisé par les métabolites majoritaires et peut donc omettre les métabolites présents en moins grande quantité dans les fractions obtenues (Kellogg et al., 2016).

Pour éviter cela, certains PNs (bio)actifs peuvent être obtenus par synthèse chimique totale ou par hémi-synthèse en utilisant un précurseur naturel. Cependant, la production de ces analogues structuraux est souvent très difficile et peut être très coûteuse, notamment lorsque les voies de synthèse sont difficiles à mettre en œuvre. Une alternative, l'ingénierie biosynthétique, consiste à utiliser les voies biosynthétiques d'un organisme pour produire des analogues de PNs (Atanasov et al., 2021).

Par ailleurs, une alternative intéressante développée ces dernières années, la chimiométrie, permet de faciliter la détection précoce de métabolites (bio)actifs dans des mélanges complexes. Ce terme est apparu pour la première fois au début des années 1970 et est défini comme "un domaine interdisciplinaire qui implique des statistiques multivariés, la modélisation mathématique, l'informatique et la chimie analytique" (Brown, 2017). Basée sur diverses méthodes analytiques, la chimiométrie permet la classification statistique de ces données. Par exemple, l'analyse en composantes principales (PCA) (Lever, Krzywinski et Altman, 2017) facilite le regroupement des extraits selon leur profil chimique. L'analyse de régression des moindres carrés partiels (PLS) (Nica et al., 2013) permet l'incorporation de données d'activité biologique ou chimique et donc la détection de marqueurs de (bio)activité des échantillons.

Quel que soit le modèle statistique utilisé, la chimiométrie nécessite des données analytiques solides pour produire des résultats exploitables. Dans ce contexte, la chromatographie liquide (ultra-performante) couplée à la spectrométrie de masse à haute résolution [(U)PLC-HRMS] est largement considérée comme une méthode de référence. En combinant les avantages de la séparation chromatographique avec la HRMS (généralement par des détecteurs QTOF), la méthode peut détecter une large gamme de PNs à faibles concentrations en raison de sa sensibilité et de sa résolution élevées (Aydoğan, 2020). Malgré ces avantages indéniables, l'(U)PLC-HRMS peut présenter certaines limites. Par exemple, la solubilisation des échantillons est généralement limitée aux phases mobiles, principalement des mélanges d'acétonitrile, de MeOH et d'eau, qui ne conviennent pas toujours aux composés les plus lipophiles. De plus, le développement de méthodes chromatographiques et spectrométriques est souvent laborieux et chronophage.

Comme de nombreux PNs présentent des similitudes structurales très proches avec les molécules de matrice utilisées dans les matrices en MALDI, ils peuvent être facilement ionisés sans support de matrice. À cet égard, l'ionisation par désorption/ionisation laser sans matrice couplée à la HRMS (LDI-HRMS) a montré des résultats

comparables à la LC-MS (Schinkovitz et al., 2018 ; Le Pogam et al., 2020 ; Skopikova et al., 2020) en un temps réduit. De plus, et contrairement aux sources d'ionisation par électrospray (ESI) principalement utilisées en LC-MS, la LDI favorise la formation d'ions radicalaires et quasi-moléculaires (Zenobi et Knochenmuss, 1998 ; Le Pogam et al., 2015). Cela peut élargir la gamme de composés détectables. Une étude comparative ESI/LDI a montré que certains composés phénoliques (par exemple, les quinones de lichen haemoventosin et parietin) étaient détectés par sous forme d'ion radicalaire en LDI-MS mais pas en ESI-MS (Le Pogam et al., 2015). Des observations similaires ont été faites pour certains isomères de vitamine E (Zhu et al., 2016). De plus, les expériences LDI-MS peuvent être réalisées en quelques secondes.

Sachant cela, la LDI-HRMS peut servir de méthode alternative potentielle à certaines applications de LC-HRMS. La méthode a été appliquée avec succès à divers extraits bruts de plantes, d'extraits de lichens et de propolis (Schinkovitz et al., 2018 ; Le Pogam et al., 2020 ; Skopikova et al., 2020) ainsi qu'à des matériaux végétaux non extraits (Islam, Alam and Kim, 2022).

Toutes les méthodes de spectrométrie de masse (y compris (U)PLC-HRMS et LDI-HRMS) sont des outils essentiels pour la chimiométrie car elles permettent l' « identification putative » de PNs au sein de mélanges complexes. Le terme fait référence à un système de confiance à cinq niveaux définis par la Société de métabolomique (Schymanski et al., 2014 ; Blaženović et al., 2018). Les critères de classification sont les suivants :

- Le niveau 0 correspond à l'attribution d'une structure tridimensionnelle incluant la stéréochimie (Blaženović et al., 2018). Elle peut être obtenue par des analyses par diffraction des rayons X, le dichroïsme circulaire et/ou la valeur alpha-D (à partir d'expériences de polarimètre) (Creek et al., 2014),
- Le niveau 1 correspond à une structure confirmée par un standard de référence ou deux méthodes orthogonales (e.g., temps de rétention (t_R) et MS ou MS et RMN),
- Le niveau 2 correspond à une structure probable correspondant à des données de littérature ou de base de données (DB),
- Le niveau 3 est un métabolite provisoire (correspondant à une classe de substance ou à une sous-structure),
- Le niveau 4 est un métabolite inconnu [parfois divisé en niveau 4 : formule moléculaire, et niveau 5 : masse exacte (Schymanski et al., 2014)].

La caractérisation complète (niveau 0 ou 1) permet une identification sans ambiguïté, mais le processus peut être assez long, surtout lorsque des composés connus sont caractérisés de manière redondante et que leurs données spectroscopiques ont été publiées plusieurs fois dans la littérature.

Depuis les années 1990, les scientifiques utilisent un processus appelé "déréplication" pour confirmer ces composés déjà rapportés (Hubert, Nuzillard et Renault, 2017). Cela implique de comparer les données expérimentales avec les données spectrales (MS, MS/MS ou RMN) provenant de la littérature ou d'une base de données. Certaines bases de données fournissent également des informations taxonomiques ou structurales (masse moléculaire, formule brute, SMILES, etc.), ce qui peut aider à limiter le nombre de candidats potentiels. Pour automatiser et accélérer la déréplication, les scientifiques ont développé des outils (bases de données, logiciels ou algorithmes).

Actuellement, il n'existe aucune base de données expérimentale complète comprenant tous les PNs rapportés. Pour les bases de données MS/MS, la recherche de spectres expérimentaux est en outre limitée par le manque de paramètres d'énergie de collision normalisés utilisés pour la fragmentation. Cette limitation ne s'applique pas aux composés volatils qui sont analysés par ionisation à impact électronique (EI) en chromatographie en phase gazeuse (GC) couplée à la spectrométrie de masse (GC-MS). Les paramètres standard de 70 eV produisent alors des schémas de fragmentation hautement reproductibles. De même, en RMN, bien que les informations spectrales soient beaucoup plus cohérentes entre les plateformes, elles peuvent varier en fonction de la nature du solvant et de la puissance du champ magnétique utilisé pour les mesures. Idéalement, les spectres d'un composé donné devraient être mesurés dans différentes conditions, et la préparation de l'échantillon et les paramètres d'acquisition devraient être bien documentés. Compte tenu de ces écueils, plusieurs programmes de simulation de spectres tels que MetFrag (Ruttkies et al., 2016), CFM-ID (Allen et al., 2014 ; Wang et al., 2021) ou MS-FINDER (Tsugawa et al., 2016) ont été développés pour produire des données MS/MS prédites spécifiques.

Une autre approche de la déréplication a été développée par M. Wang et al. (Wang et al., 2016) qui ont introduit le "global natural products social molecular networking" (GNPS). Le GNPS permet la génération de réseaux moléculaires qui regroupent les données de fragmentation MS/MS en fonction des similarités spectrales, permettant ainsi la visualisation des molécules apparentées. De plus, la plateforme facilite l'échange gratuit de données expérimentales MS/MS afin d'annoter rapidement les métabolites à travers les réseaux moléculaires et de propager l'annotation. Malgré ces avancées notables, un réseau moléculaire basé sur la HRMS ne peut fournir que des annotations putatives jusqu'au niveau de confiance 2 car il ne combine pas d'informations orthogonales provenant de différentes méthodes.

De plus, une limitation majeure de toutes les techniques de spectrométrie de masse (MS) est la différenciation des isomères. En l'absence de séparation chromatographique, cet inconvénient s'applique particulièrement à la LDI-HRMS et à l'(UP)LC-HRMS. Cependant, en (UP)LC-HRMS, certains isomères peuvent ne pas être différenciés. Mais la RMN du ^{13}C peut résoudre ce problème, car les stéréoisomères peuvent être différenciés par leurs déplacements chimiques du carbone-13 (δ_C). Cette méthode non destructive peut être utilisée pour annoter des composés connus et même inconnus avec un niveau de confiance similaire à celui de l'(UP)LC-HRMS. Cependant, le temps nécessaire pour acquérir des spectres RMN de haute qualité dépend de facteurs tels que la concentration de l'échantillon, l'intensité du champ magnétique et la complexité de la molécule étudiée. De plus, l'interprétation manuelle et l'analyse des spectres RMN pour déterminer la structure complète d'un composé sont des processus longs nécessitant une expertise significative dans le domaine. À cet égard, les stratégies de déréplication peuvent permettre de gagner du temps en annotant les composés déjà décrits avant l'isolement des composés. Ainsi, l'algorithme de déréplication MixONat basé sur la RMN du carbone-13 a été développé par A. Bruguière et al. (Bruguière et al., 2020, 2021). Cet algorithme permet une annotation fiable des métabolites dans les mélanges complexes, avec un niveau de confiance de 2. Lorsqu'il est combiné à des données HRMS, cette stratégie pourrait permettre l'élucidation de la structure des PNs dans les mélanges complexes avec un niveau de confiance de 1.

Dans le cadre de ma thèse de doctorat, nous avons visé à développer une approche hybride de la chimiométrie basée sur la spectrométrie de masse (MS) et de la déréplication par RMN du ^{13}C pour l'identification de PNs potentiellement actifs provenant de *Garcinia parvifolia* (*Clusiaceae*). Cette plante est originaire de Malaisie. Des études précédentes sur quatre extraits d'écorce de *G. parvifolia* (KL5670, KL5259, KL5248, KL5073) réalisées au

laboratoire SONAS ont révélé des effets inhibiteurs notables sur la formation de produits finaux de glycation avancée (AGEs) dans un essai interne (Derbré et al., 2010). *Garcinia parvifolia* est riche en xanthones, PNs qui présentent de fortes similitudes structurelles avec le dithranol, utilisée comme matrice en MALDI. Par conséquent, les écorces de cette plante constituent un candidat idéal pour développer des méthodes de chimiométrie basées sur LDI-MS, qui seront comparées aux méthodes conventionnelles basées sur LC-MS pour l'identification de métabolites actifs. À un stade ultérieur, cette approche sera étendue à la spectrométrie de masse à mobilité ionique (IMS) ainsi qu'à la création de réseaux moléculaires basés sur la LDI-IM-MS.

Dans le **chapitre 1** "État de l'art" de ma thèse, une revue présentera les "outils et stratégies actuels et émergents pour l'identification de produits naturels bioactifs dans les mélanges complexes". Ensuite, le **chapitre 2** "Méthodologie générale et conditions expérimentales" décrira toutes les techniques qui ont été appliquées en pratique pendant ma thèse. Cela comprendra des protocoles détaillés des essais anti-AGEs utilisés ainsi que toutes les expériences de (UP)LC-HRMS, LDI-HRMS et RMN.

Le **chapitre 3**, "Prédiction de l'activité des produits naturels dans les mélanges complexes en utilisant des analyses chimiométriques", se concentrera sur l'identification de marqueurs d'activité dans les mélanges complexes grâce à une approche combinée de LDI HRMS et de RMN du ^{13}C. Les résultats ont déjà été publiés dans le journal Talanta.

Le **chapitre 4**, "Stratégies de déréplication basées sur les réseaux moléculaires", analysera les améliorations, les intérêts et les limites des annotations basées sur les réseaux moléculaires des composés dans les mélanges complexes. Ce chapitre sera divisé en deux parties. La première partie portera sur un projet d'article intitulé "À la recherche d'actifs dans la botte de foin : fusion des approches de réseaux moléculaires basés sur $HRMS^2$, de la chimiométrie et de la déréplication par RMN du ^{13}C", tandis que la deuxième partie portera sur "La spectrométrie de masse par ionisation laser sans matrice couplée à la spectrométrie de masse à mobilité ionique: une approche complémentaire pour la différenciation des isomères naturels et la création de réseaux moléculaires".

Enfin, dans le **chapitre 5** "Conclusions et perspectives", une discussion conclusive sur les perspectives sera présentée.

8.3. Prédiction de l'activité des produits naturels dans des mélanges complexes à l'aide d'analyses chimiométriques

Les métabolites secondaires des plantes sont connus pour leur large éventail d'effets biologiques très intéressants et ont fourni des pistes précieuses pour diverses applications médicinales, cosmétiques et agricoles (Chemat *et al.*, 2019; Clemensen *et al.*, 2020; Newman and Cragg, 2020a). Toutefois, étant présents dans des mélanges complexes tels que les extraits bruts, les PNs (bio)actifs sont difficiles à identifier.

Une stratégie pour résoudre ce problème est le "fractionnement bioguidé" (FBG) (Cunha *et al.*, 2017). Cependant, comme expliqué en introduction, le FBG nécessite un contrôle continu de l'activité après chaque étape de séparation. Ainsi, cette méthode diminue considérablement la quantité de matériel de travail et augmente encore le risque d'isolement répétitif de composés déjà connus et décrits dans la littérature (Nahrstedt and Butterweck, 2010; Inui *et al.*, 2012).

Pour contourner cette limitation, la chimiométrie est devenue un outil important. Elle permet la caractérisation chimique des échantillons, ainsi que l'identification de marqueurs d'activité potentiel à partir de mélanges complexes par divers moyens statistiques. Alors que, par exemple, l'analyse en composantes principales (ACP) (Lever, Krzywinski and Altman, 2017) facilite le regroupement des échantillons en fonction de leur profil chimique, l'analyse par régression des moindres carrés partiels (PLS) peut inclure des données sur l'activité biologique ou chimique (Nica *et al.*, 2013) et permettre ainsi l'identification des marqueurs d'activité spécifiques à l'échantillon (Kellogg *et al.*, 2016a).

Indépendamment du modèle statistique appliqué, toutes les analyses chimiométries nécessitent des données analytiques solides pour produire des résultats exploitables. À cet égard, la chromatographie liquide couplée à la spectrométrie de masse en tandem à haute résolution (LC-MS^2) est considérée comme une méthode de référence. Combinant les avantages de la séparation chromatographique et la sensibilité de la MS, cette méthode permet de détecter une large gamme de PNs à de faibles concentrations (Aydoğan, 2020). Cependant, la solubilisation de l'échantillon est généralement limitée aux phases mobiles, comprenant principalement des mélanges d'acétonitrile, de méthanol et d'eau, qui ne sont pas bien adaptés aux composés lipophiles. De plus, même si l'ionisation par électrospray (ESI) couramment utilisée favorise généralement la formation d'ions quasi-moléculaires tels que $[M+H]^+$ ou $[M-H]^-$, certains constituants, par exemple les quinones du lichen, l'hémventosine et la pariétine, forment de préférence des ions radicaux et peuvent ne pas s'ioniser dans l'ESI (Le Pogam *et al.*, 2015). Des observations similaires ont également été faites pour certains isomères de la vitamine E (Zhu *et al.*, 2016).

Favorisant la formation d'ions radicaux et quasi-moléculaires (Zenobi and Knochenmuss, 1998; Le Pogam *et al.*, 2015; Zhu *et al.*, 2016), la MS par désorption/ionisation laser sans matrice (LDI-MS) peut donc constituer un complément précieux ou une alternative potentielle à la LC-ESI-MS (Chemat *et al.*, 2019). Comme de nombreux PNs présentent une bonne absorption à la longueur d'onde d'émission des lasers UV-LDI courants, la méthode couvre une large gamme de PNs chimiquement diverses telles que les alcaloïdes (Skopikova *et al.*, 2020), les substances phénoliques (Schinkovitz *et al.*, 2018; Le Pogam *et al.*, 2020) et les cannabinoïdes (dos Santos *et al.*, 2019). En outre, la LDI-MS peut être appliquée directement sur des extraits bruts de toute polarité (Le Pogam *et al.*, 2020; Skopikova *et al.*, 2020) sans étape de pré-purification de l'échantillon et fonctionne même sur du matériel végétal en poudre non extrait (Islam, Alam and Kim, 2022). Contrairement à la LC-MS^2, les expériences sont réalisées en quelques secondes. De plus, malgré ces résultats prometteurs et à la connaissance des auteurs, la LDI-MS n'a jamais été utilisée dans des analyses chimiométriques.

C'est pourquoi la présente publication décrit le développement étape par étape d'un modèle de régression PLS basé sur la LDI-MS pour l'identification de marqueurs d'activité à partir de mélanges complexes de PNs. Un essai fluorométrique mesurant l'inhibition de la formation de produits finaux de la glycation avancée (AGE) a été utilisé pour l'évaluation de l'activité (Derbré *et al.*, 2010; Séro *et al.*, 2013). Au cours de ce processus, un modèle statistique basé sur des mélanges bien définis de composés de référence actifs et inactifs a d'abord été établi (**I**). Ensuite, des échantillons réels comprenant quatre extraits d'écorce de *Garcinia parvifolia* avec différentes activités anti-AGEs ont été étudiés (**II**). Les résultats ont ensuite été comparés à la chimiométrie basée sur la LC-MS^2 réalisée simultanément (**III**). Enfin, des marqueurs d'activité sélectionnés, précédemment identifiés par LDI-MS et LC-MS^2, ont été analysés par MixONat, un outil de déréplication de la RMN ^{13}C (**IV**) qui a été développé dans

le laboratoire SONAS (https://sourceforge.net/projects/mixonat/). Ayant été appliqué avec succès pour l'identification des constituants majeurs de divers extraits de plantes tels que *Garcinia mangostana* (Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Suor-Cherer, *et al.*, 2020), *Calophyllum Brasiliense* (Silva-Castro *et al.*, 2021), *Rosmarinus officinalis* (Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Suor-Cherer, *et al.*, 2020) et l'huile essentielle de menthe poivrée (Bruguière *et al.*, 2018), la stratégie employée a permis d'obtenir des informations structurelles sur les marqueurs d'activité majeurs avant leur isolement.

8.4. Stratégies de déréplication à l'aide de réseaux moléculaires

8.4.1. Recherche d'actifs dans une botte de foin : combinaison des approches de déréplication basées sur le réseau moléculaire HRMS^2, la chimiométrie et la RMN ^{13}C

En raison de la propension des PNs à interagir avec des cibles d'intérêt thérapeutique ou agronomique, certains chimistes des PNs se concentrent sur l'identification de composés bioactifs dans des matrices complexes telles que des extraits de plantes médicinales ou de microorganismes (Cordell, 2000; Gwinn, 2018; Newman and Cragg, 2020b). Depuis longtemps, une approche classique, connue sous le nom de FBG, a été utilisée pour relever ce défi (Malviya and Malviya, 2017). Cependant, les étapes de fractionnement successives associées aux essais biologiques successifs sont coûteuses et présentent des limites. En particulier, l'isolement de PNs bien connus n'est pas rare et les efforts consentis ne sont pas à la hauteur des résultats (Inui *et al.*, 2012). Utilisée seule, la stratégie conduit aussi principalement à la caractérisation structurale des PNs bioactifs les plus abondants après leur isolement. Pour contourner ces problèmes, la méthodologie a ensuite été améliorée en utilisant simultanément la métabolomique non ciblée, la MS ou la RMN. Ces données sont ensuite liées à la bioactivité par des analyses biochimiques (Kellogg *et al.*, 2016b; Nothias *et al.*, 2018; Langeder *et al.*, 2023). Dans ce cas, les PNs bioactifs proposés sont finalement isolés pour une identification complète.

En effet, quelle que soit la méthode spectrale, l'annotation des composés sélectionnés reste un défi majeur dans la métabolomique non ciblée basée sur la MS ou la RMN. Cette dernière étape, appelée déréplication, repose sur des algorithmes qui comparent les données spectrales obtenues avec celles des métabolites d'une base de données donnée (DB) et attribuent finalement les meilleurs scores aux hypothèses structurelles les plus appropriées. Ces stratégies utilisent soit des spectres expérimentaux MS² [par exemple, global natural product social (GNPS)] / RMN (par exemple, NP-MRD), soit des BD *in silico* (Blaženović *et al.*, 2018b; Wishart *et al.*, 2022). En particulier, dans le cadre de la métabolomique non ciblée, la déréplication - c'est-à-dire l'identification sans ambiguïté des PN connus responsables de l'activité d'un extrait avant FBG (Appendino, Fontana and Pollastro, 2010; Hubert, Nuzillard and Renault, 2017b) - de matrices complexes issues d'organismes vivants est au centre des préoccupations des chercheurs impliqués dans l'identification de PNs bioactifs depuis près de vingt-cinq ans (Van Middlesworth and Cannell, 1998). Aujourd'hui, on estime à plus de 500 000 le nombre de PNs d'origine microbienne, marine et terrestre décrits dans la littérature (Ntie-Kang and Svozil, 2020). Par conséquent, la collecte de données expérimentales RMN (Wishart *et al.*, 2022; Kuhn *et al.*, 2021) et de MS (Fox

Ramos, Le Pogam, *et al.*, 2019; Olivier-Jimenez *et al.*, 2019; Vargas *et al.*, 2020; Agnès *et al.*, 2022) est une tâche qui nécessite un effort continu et collectif. Par ailleurs, les bases de données de spectres *in silico* deviennent de plus en plus importantes avec les progrès constants des algorithmes et logiciels de prédiction spectrale [par exemple CFM-ID4 (Wang *et al.*, 2021b, 2022), MetFrag (Ruttkies *et al.*, 2016b), ACD/NMR Predictors (C et H) (ACD Labs, no date), nmrshiftdb2 (Kuhn and Schlörer, 2015; Kuhn, no date)]. Par conséquent, l'utilisation de ces bases de données contenant les spectres de fragmentation prédits ou les déplacements chimiques ^{13}C-RMN (δ_C) des PNs permet aujourd'hui d'émettre des hypothèses structurelles pertinentes (Allard *et al.*, 2016; Bertrand, Guitton and Roullier, 2017; Kind and Fiehn, 2017; Abedini *et al.*, 2020; Bruguière *et al.*, 2021; Silva-Castro *et al.*, 2021; Azonwade *et al.*, 2023).

En ce qui concerne la déréplication de mélanges complexes, il convient de rappeler que, selon la MSI (Sumner *et al.*, 2007), les métabolites sont annotés avec un niveau de confiance. Si le niveau 0 se réfère à la structure 3D d'un PN pur isolé à l'aide de données spectrales appropriées, le niveau 1 nécessite soit la comparaison des données avec un standard, soit au moins deux techniques orthogonales pour proposer une structure 2D fiable. Le niveau 2 indique un métabolite probable grâce à une correspondance avec des données bibliographiques ou des bases de données. Le niveau 3 est attribué à une structure ou à une classe possible de PNs, tandis qu'une molécule inconnue reçoit le niveau 4 ou 5 (c.-à-d. 4 : formule moléculaire ; 5 : masse exacte) (Viant *et al.*, 2017; Blaženović *et al.*, 2018b). Par conséquent, il semble que seul le niveau 2 puisse être atteint en annotant les PNs à l'aide de bases de données spectrales prédites, tandis que le niveau 1 l'est lorsque deux méthodes d'identification sont utilisées (par exemple, MS et RMN).

Au cours de la dernière décennie, des outils et des logiciels de déréplication automatisés basés sur la MS² haute résolution (HRMS²) associée aux MN (Yang *et al.*, 2013) ou la RMN-^{13}C (Hubert *et al.*, 2014; Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Bréard, *et al.*, 2020) sont apparus pour identifier efficacement les PNs connues dans les mélanges. Alors que l'un des principaux avantages de la HRMS² est sa sensibilité, la RMN ^{13}C présente l'avantage de pouvoir distinguer les stéréo-isomères. Comme indiqué précédemment, les deux stratégies peuvent désormais utiliser des bases de données spectrales prédites plus larges pour explorer les correspondances avec des PNs connus, avec l'inconvénient d'un niveau de confiance plus faible (c'est-à-dire 2) dans l'annotation si elles sont utilisées indépendamment. Le travail présenté ici suppose l'intérêt de combiner les approches basées sur la MS et la RMN pour augmenter la confiance dans l'annotation vers le niveau 1.

Dans le but d'identifier de manière fiable les PNs actifs dans les mélanges bruts, l'efficacité d'une stratégie combinant les MN basée sur la HRMS², la chimiométrie, les approches de déréplication basées sur la RMN-^{13}C ainsi que les BD *in silico* a été étudiée dans le présent travail. Comme exemple de travail, un extrait d'écorce de *Garcinia parvifolia* (*Clusiaceae*) (KL5670) a été sélectionné parmi quatre chimiotypes sur la base de sa teneur en xanthones prénylées (c.-à-d. stéréoisomères) et de sa capacité à inhiber la formation de produits finaux de glycation avancée (AGE) (Meunier *et al.*, 2023a).

En 2018, L.-F. Nothias *et al.*, ont introduit le concept de MN basée sur la bioactivité (Nothias *et al.*, 2018). Après fractionnement de l'extrait brut pour obtenir un essai biologique pertinent, la bioactivité des fractions a été évaluée. Ensuite, l'approche a combiné le MN basé sur des données de MS/MS avec la bioactivité pour mettre en évidence les PNs actifs. Bien que dans cette publication, les auteurs aient calculé un score de bioactivité basé sur

la corrélation de Pearson entre l'intensité des ions détectés dans les échantillons et le niveau de bioactivité associé à chaque échantillon, il convient de noter qu'une autre méthode courante et efficace pour sélectionner les métabolites/PN d'intérêt est l'analyse des moindres carrés partiels (PLS), qui met en évidence la corrélation entre les variables indépendantes (données spectrales (X)) et une variable dépendante (données d'activité (Y)). En ce qui concerne le présent travail, la régression PLS a déjà permis d'identifier avec succès des inhibiteurs de la formation d'AGE directement à partir d'extraits bruts d'écorce de *G. parvifolia* (Derbré *et al.*, 2010). Dans la publication de Nothias *et al.*, le MN final basé sur la bioactivité a finalement été obtenu à l'aide de Cytoscape avant l'isolement et l'identification des PNs sélectionnées. Par ailleurs, le présent travail sélectionne tout d'abord les métabolites d'intérêt à l'aide de la HRMS² et un MN basé sur la bioactivité à l'aide des résultats chimiométriques [c'est-à-dire le VIP score et le coefficient de régression (RC) du modèle PLS] (Nothias *et al.*, 2018). Deuxièmement, l'annotation est effectuée à l'aide de l'outil *in silico* MetFrag inclus dans le logiciel ProgenesisQI (Ruttkies *et al.*, 2016b), la déréplication des fractions basée sur la RMN-^{13}C à l'aide du logiciel gratuit MixONat (Bruguière, Derbré, Dietsch, Leguy, Rahier, Pottier, Bréard, *et al.*, 2020) et les BD des PNs basées sur la chimiotaxonomie, y compris leur δ_C prédit (Figure 1) (Nuzillard, Leroy and Kuhn, 2021). Dans cette preuve de concept, la précision de l'annotation a finalement été évaluée par la purification et l'évaluation des PNs afin de permettre la discussion des résultats.

8.4.2. Désorption laser sans matrice couplée à la spectrométrie de mobilité ionique : un outil puissant pour la différenciation des isomères naturels et réseau moléculaire

La chromatographie liquide couplée à la spectrométrie de masse en tandem à haute résolution (LC-MS/MS) est largement considérée comme méthode de référence pour l'analyse des mélanges complexes. Combinant les avantages de la MS/MS et de la séparation chromatographique, cette méthode est devenue la pierre angulaire de la recherche "omique" contemporaine, y compris les MN (Wang *et al.*, 2016; Fox Ramos, Evanno, *et al.*, 2019; Aron *et al.*, 2020; Qin *et al.*, 2022).

Ces dernières années, l'ionisation par désorption/ionisation laser (LDI) sans matrice est une méthode complémentaire à la LC-MS. Bénéficiant d'un protocole expérimental rapide, la LDI-MS est particulièrement adaptée à l'analyse des PNs tels que les composés phénoliques et les alcaloïdes (Le Pogam *et al.*, 2015, 2020; Schinkovitz *et al.*, 2018; Skopikova *et al.*, 2020; Islam, Alam and Kim, 2022). Ces derniers présentent d'étroites similitudes structurelles avec les matrices utilisées dans l'ionisation par désorption/ionisation laser assistée par matrice (MALDI). Par conséquent, ils peuvent être facilement ionisés par simple irradiation laser et sans support matriciel supplémentaire. Cette approche LDI-MS directe a été appliquée avec succès pour caractériser des mélanges très complexes de PNs tels que des poudres de plantes ou des extraits bruts (Schinkovitz *et al.*, 2018; Meunier *et al.*, 2023b). De plus, la LDI favorise la formation d'ions radicaux et quasi-moléculaires et peut donc détecter des composés qui pourraient échapper à l'ionisation par électrospray (ESI). Cet effet a été particulièrement observé pour divers composés phénoliques (Le Pogam *et al.*, 2015).

L'approche LDI-MS peut également être facilement intégrée aux analyses chimiométriques. Une étude récente sur des extraits bruts de *G. parvifolia* a donné des résultats comparables pour la LDI-MS et l'UPLC-MS sur la base

d'une analyse discriminante par moindres carrés partiels (PLS-DA) [10]. Cependant, le même rapport a également révélé un problème fondamental de la LDI-MS. En raison de l'absence de séparation chromatographique, les isomères de constitutions ne peuvent pas être différenciés par cette méthode.

La spectrométrie de mobilité ionique (IMS) est une approche qui permet de contourner cette limitation. Les ions soumis à un champ électrique en présence d'un gaz porteur, sont séparé en fonction de leur conformation, de leur forme, de leur charge et de leur masse (Hines *et al.*, 2017). Aujourd'hui, l'IMS a été appliquée avec succès pour différencier divers isomères de différents groupes chimiques tels que les sucres (Hofmann *et al.*, 2015), les lipides (Zhang *et al.*, 2018) ou les cannabinoïdes (Hädener *et al.*, 2018). En revanche, il existe très peu d'informations sur les xanthones.

C'est pourquoi le présent travail a évalué la mobilité ionique (IM) de divers isomères de xanthones par une approche combinée LDI-IM-MS/MS (**I**). Outre les composés individuels, des mélanges simples de composés précédemment isolés à partir d'un extrait d'écorce de *G. parvifolia* (KL5670), une plante qui est particulièrement riche en xanthones, ont été analysés (**II**). En élargissant l'utilisation conventionnelle de la LDI-MS, la faisabilité d'un MN basé sur les données LDI-IM-MS/MS a été explorée sur des fractions de l'un des extraits de *G. parvifolia* étudiés (KL5073) (**III**) (Wang *et al.*, 2016).

8.5. Conclusion

Ces travaux visaient à développer de nouvelles approches pour identifier des (bio)marqueurs dans des mélanges complexes de PNs. Dans l'ensemble, trois stratégies biochimétriques et/ou de déréplication différentes basées sur la LDI-MS ou la LDI-(IM)-MS/MS, l'UPLC MS/MS et la RMN ^{13}C ont été développées (Table 3).

La chimiométrie basée sur la LDI-MS offre de nombreux avantages, notamment un protocole expérimental simple, une faible quantité de solvants et de matériel d'échantillonnage, ainsi que des résultats instantanés. En outre, cette stratégie a démontré son efficacité pour la détection de composés anti-AGEs à partir d'extraits bruts de *G. parvifolia* (KL5670) avant leur isolement. Cependant, en raison de l'absence de séparation chromatographique, la méthode ne permet pas de différencier les isomères. Des expériences ^{13}C supplémentaires et l'utilisation de l'outil de déréplication, MixONat, pourraient permettre de contourner certains de ces problèmes. Cependant, il n'existe pas d'outils ou de logiciels de déréplication adaptés au traitement des données LDI-MS. Le développement d'un tel outil est un objectif mené en collaboration avec l'équipe de recherche LERIA.

La seconde stratégie vise à combiner les MN basés sur des données de type UPLC-MS/MS avec la chimiométrie qui ont permis l'annotation précoce de molécules actives dans des mélanges complexes, y compris des isomères. Dans ce contexte, la section 4.1 a souligné l'importance des niveaux de confiance pour l'annotation correcte des composés. Pour affiner le processus, le niveau 1 (composé isolé) a été divisé en deux sous-catégories : le niveau 1a (correspondance basée sur la littérature) et le niveau 1b (proposition de structure convergente utilisant deux méthodes orthogonales). De cette manière, la précision et l'exhaustivité des annotations ont pu être améliorées de manière significative.

Dans le cadre d'une troisième stratégie, l'approche LDI-IM-MS/MS a permis de relever le défi initial de la différenciation des isomères dans les expériences LDI-MS. Ces expériences ont abouti à la création d'un MN exclusivement basé sur ce type de données (section 4.2, figure 3). De plus, toutes les données MS/MS ont été

déposées sur la plateforme GNPS sous le nom de Xanthones-DB (avec les données UPLC-MS/MS acquises à trois CE différents). Ceci a permis la création d'un MN illustré dans la section 4.2, Figure 3.

Une partie importante de la thèse a été consacrée au développement du traitement de données basé sur la disponibilité des instruments et des logiciels (open-source ou commerciaux). En effet, de nombreux problèmes concernant la compatibilité des fichiers originaux des instruments avec les logiciels libres ont été rencontrés. Par exemple, les fichiers de données LDI-MS (.tas, JEOL) étaient incompatibles avec le logiciel mMass et ont dû être convertis au préalable au format csv. De plus, le logiciel mMass ne permettait pas d'aligner les spectres, de sorte qu'un algorithme spécifique a dû être développé par l'équipe du LERIA.

Aujourd'hui, la chimiométrie basée sur la LDI-MS et l'UPLC-MS est un outil très efficace pour cibler les composés actifs dans des mélanges complexes. Cependant, l'identification certaine de ces composés reste un défi important, car les données peuvent facilement être mal interprétées. Par conséquent, l'intégration minutieuse de sources d'information multiples est nécessaire pour augmenter le niveau de confiance et la fiabilité globale de l'annotation des composés en chimiométrie.

Plus tard, l'interprétation des données de MS pourrait être assistée par l'intelligence artificielle (IA), qui permettrait d'optimiser le traitement et l'interprétation des données brutes. Cette approche pourrait faciliter une stratégie automatisée efficace pour l'identification de composés connus et inconnus à partir de mélanges complexes, avec un niveau de confiance élevé.

Table of appendices

Appendix 1. ProgenesisQI scoring and fragmentation scoring

Scoring

"Scores range from 0-100, with 100 being a perfect match, and are calculated using the mean of five similarity metrics:

- *Mass Similarity*: a mass difference of 0 gives a score of 100, which falls off rapidly to ~2 when the mass difference is ~1Da (for a 250Da compound).
- *Isotope Similarity*: this compares the intensities of each isotope between observed and theoretical distributions. A total intensity difference of 0 gives a score of 100, which falls linearly to 0 when the total intensity difference is equal to the maximum isotope intensity.
- *Retention Time Similarity*: a retention time difference of 0 gives a score of 100, which falls off rapidly to ~2 when the percentage error is ~20%. If the search method does not support searching by retention time or has been configured not to use it, this metric is 0.
- *CCS Similarity*: CCS difference of 0 gives a score of 100, which falls off rapidly to ~2 when the percentage error is ~20%. If the search method does not support searching by CCS or has been configured not to use it, this metric is 0.
- *Fragmentation Score*: this is a score from 0 to 100 representing how well your observed fragmentation data matches the search (which might be either theoretical fragmentation, or data from a fragment database). The exact scoring method depends on whether you carried out a theoretical or fragment database search. If fragment search is not available or has been turned off in the configuration of your chosen search method, this metric is 0."

Fragmentation Scoring

"The scoring method in the original MetFrag paper calculated a score based on weighted peak intensities, combined this with a score based on energy of the bonds broken, and normalized the results so that the best identification for a given compound has a score of 1.

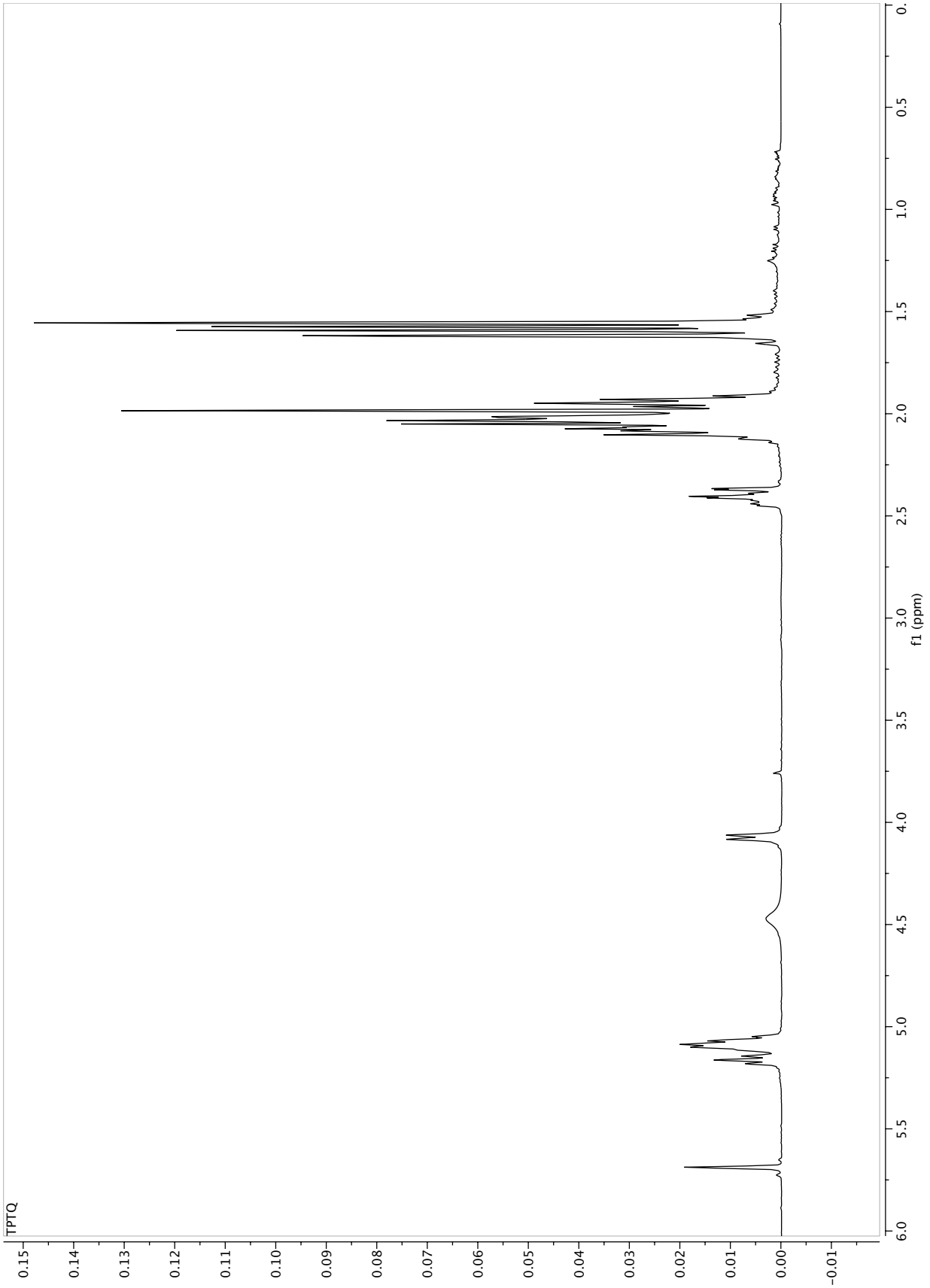
In Progenesis QI, you often have multiple compounds which you are identifying simultaneously, all of which may have many possible identifications. With the original MetFrag scoring algorithm, the top identification for each compound scored 1, so it was not possible to compare scores between compounds.

The algorithm in our modified version of MetFrag is based upon the same score calculated from weighted peak intensities, but we have modified it so that scores are comparable across compounds.

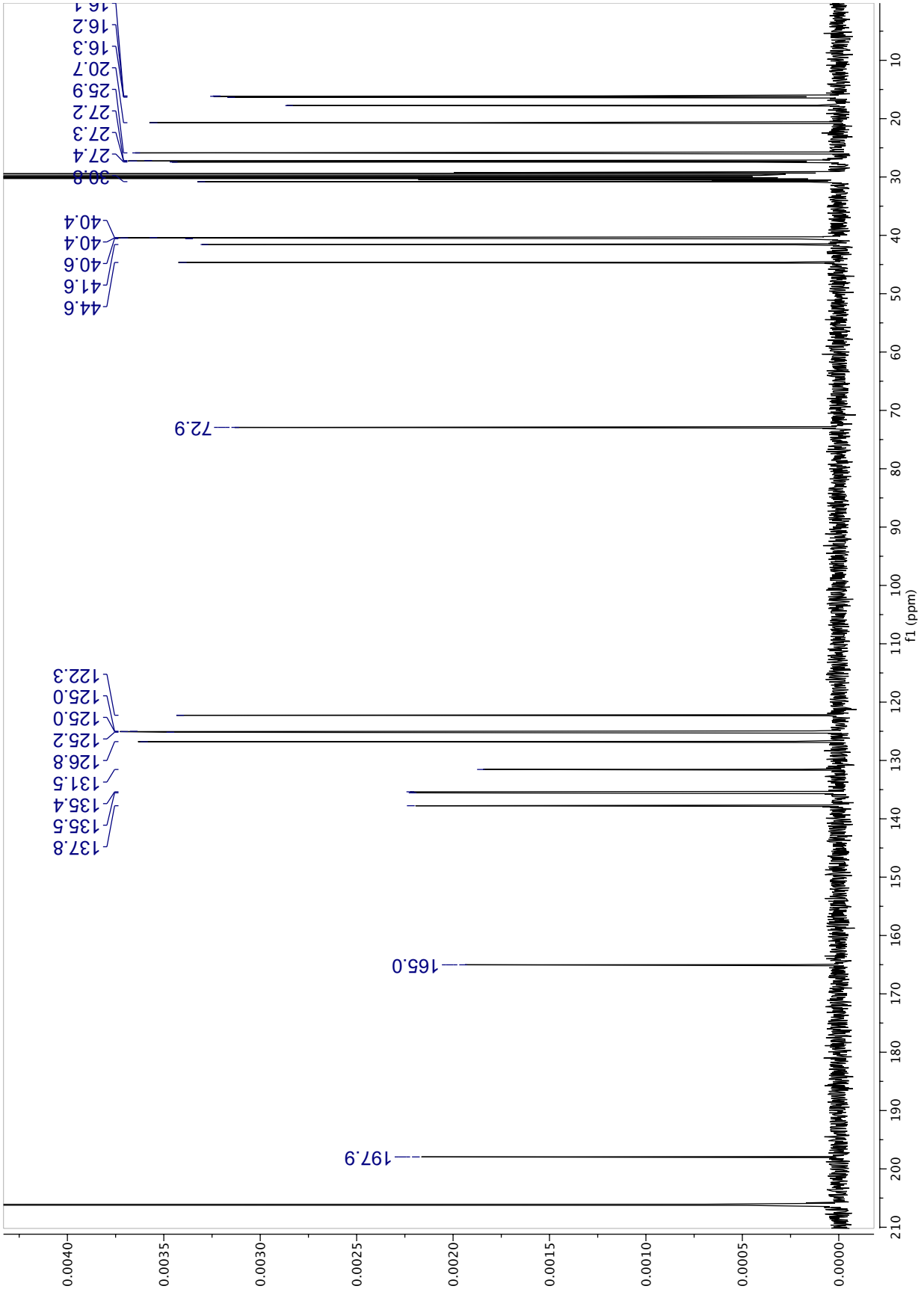
This is achieved by considering the sum of weighted peak intensities across all matched peaks and dividing it by the sum of weighted peak intensities across all peaks, matched or not. This gives us a relative score of how well the observed fragmentation trace can be explained by a given possible identification – a possible identification that can explain 90% of the peaks is preferred to one that can explain only 40% of the peaks (in practice this is not necessarily the case, because the peaks are weighted by intensity and mass, but it suffices to simplify the explanation)."

From ProgenesisQI website (*How does the theoretical fragmentation work? - Progenesis QI v3.0*, 2023; *How are the scores calculated for possible compound identifications?*, no date)

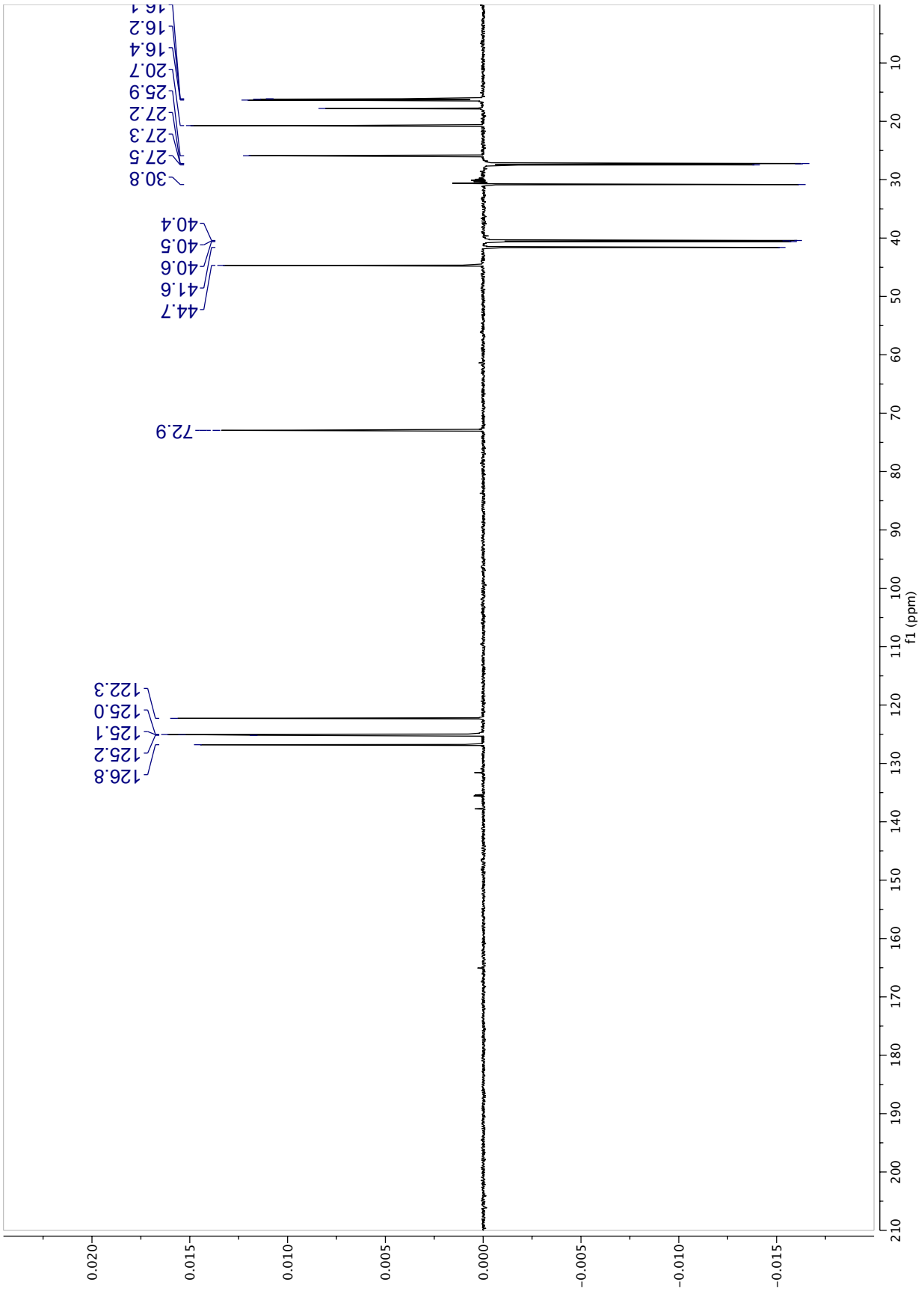
Appendix 2. ^{1}H spectrum of TPTQ in acetone-*d*6



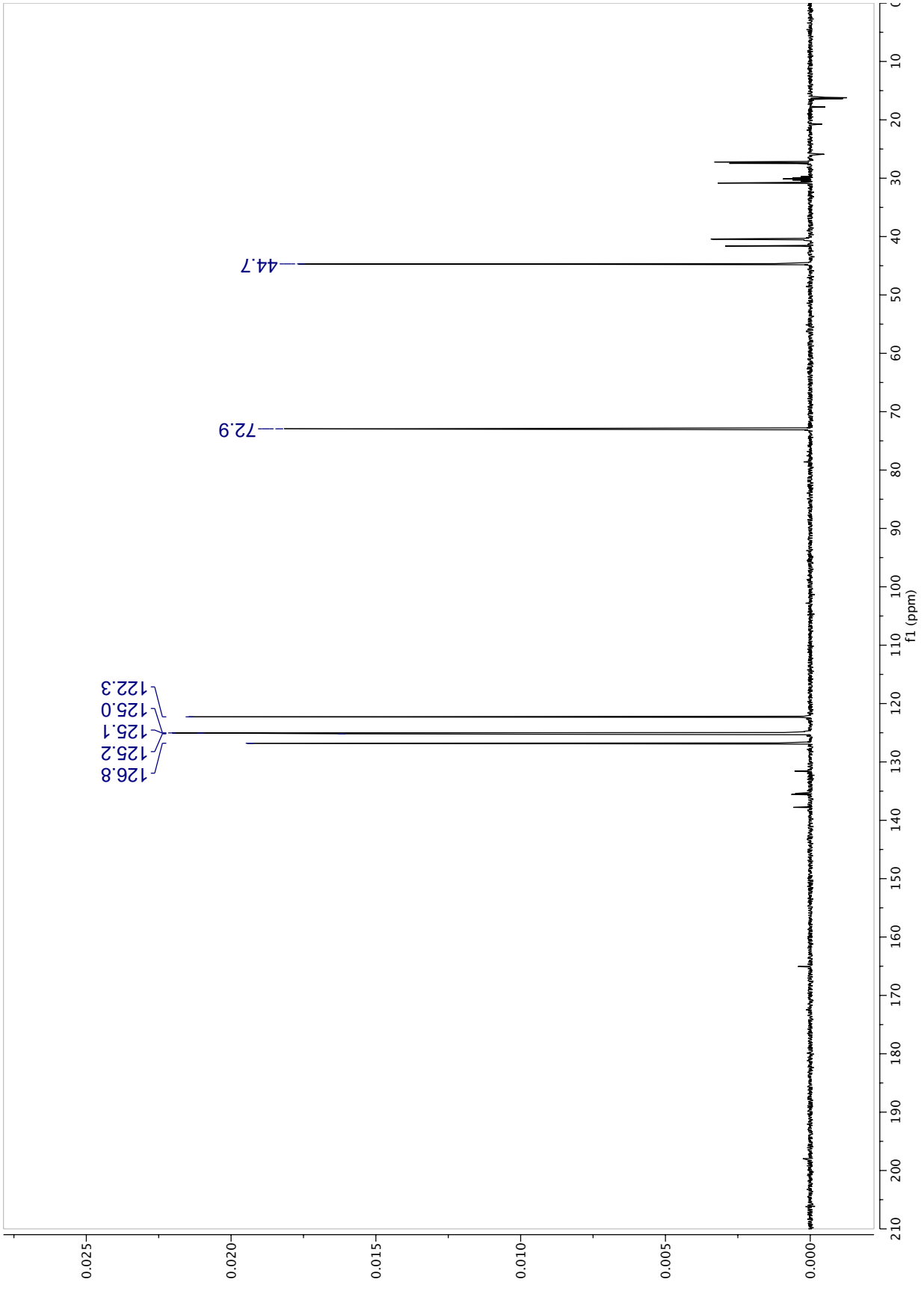
Appendix 3. ^{13}C spectrum of TPTQ in acetone-*d*6 (382 scans) (100.0 mg)



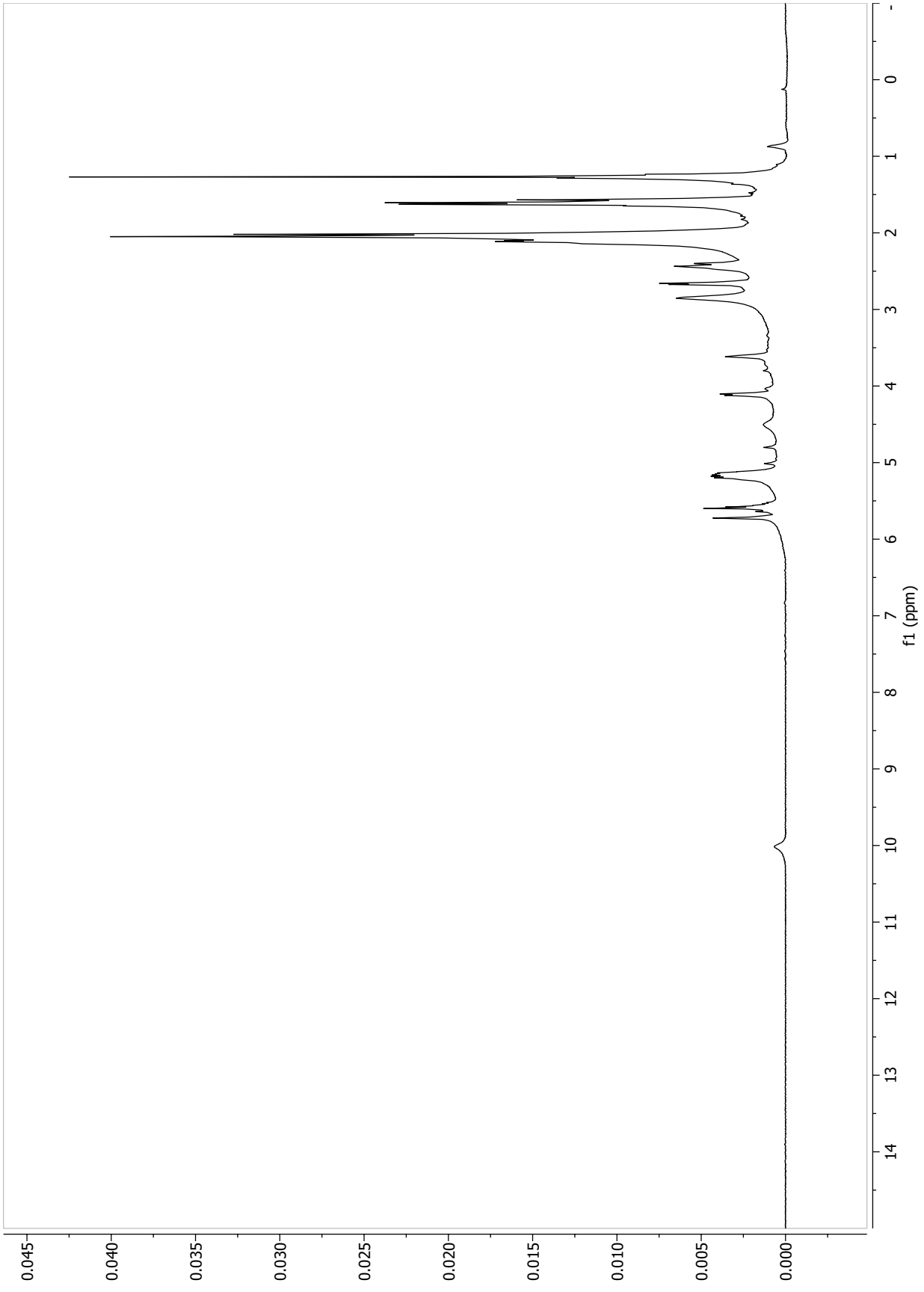
Appendix 4. DEPT-135 spectrum of TPTQ in acetone-*d*6 (500 scans) (100.0 mg)



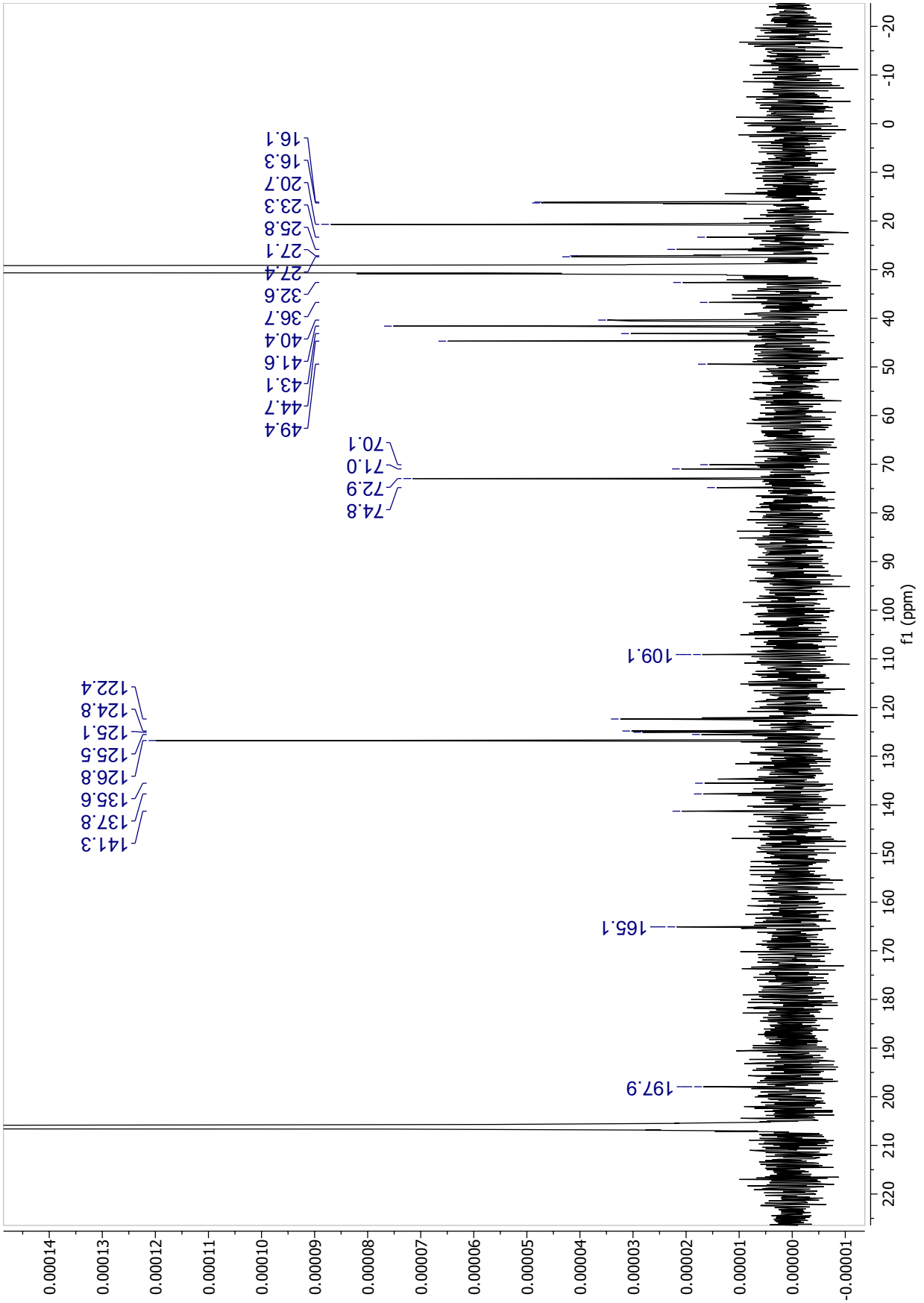
Appendix 5. DEPT-90 spectrum of TPTQ in acetone-*d*6 (500 scans) (100.0 mg)



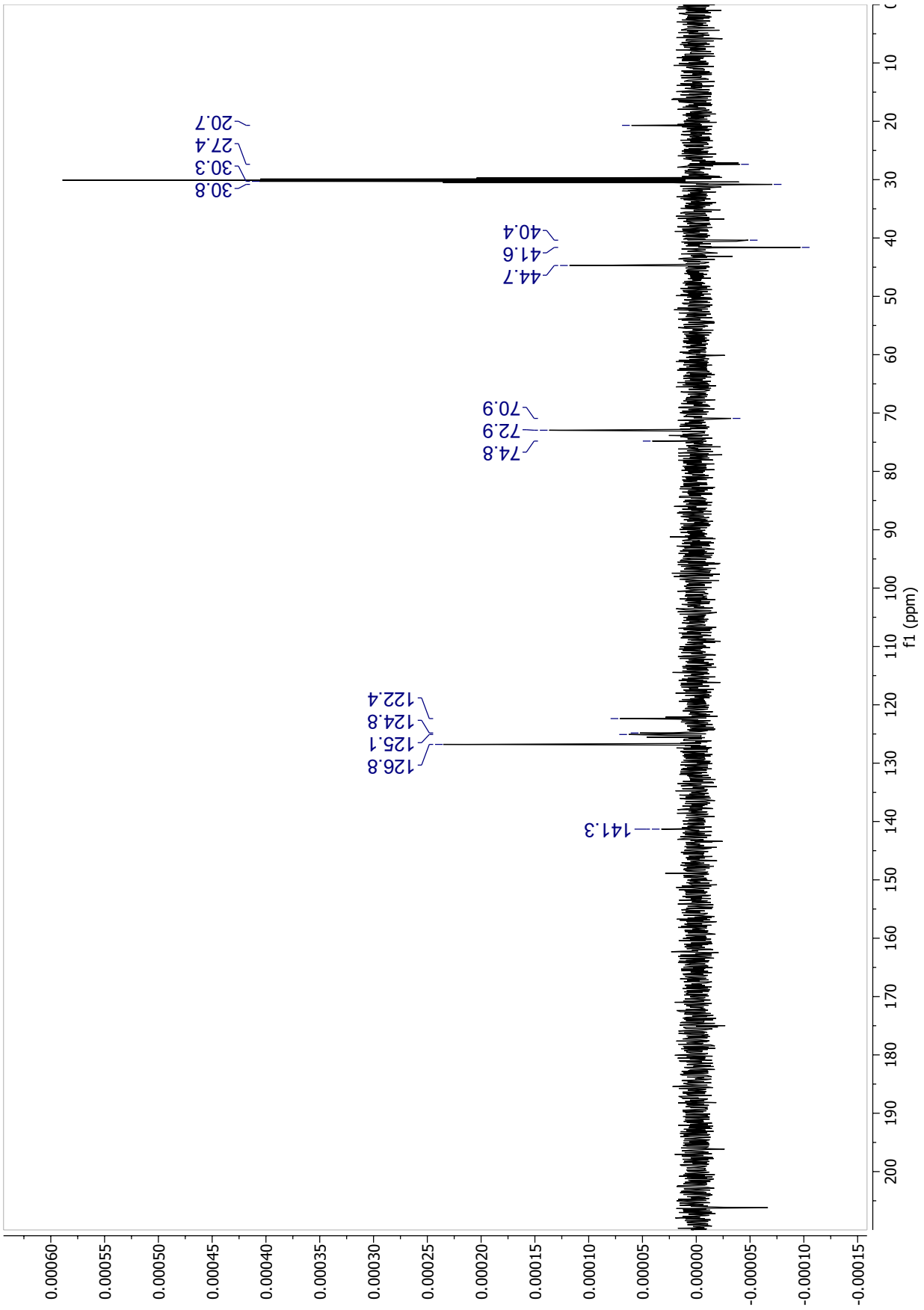
Appendix 6. ^{1}H spectrum of M1 in acetone-*d*6



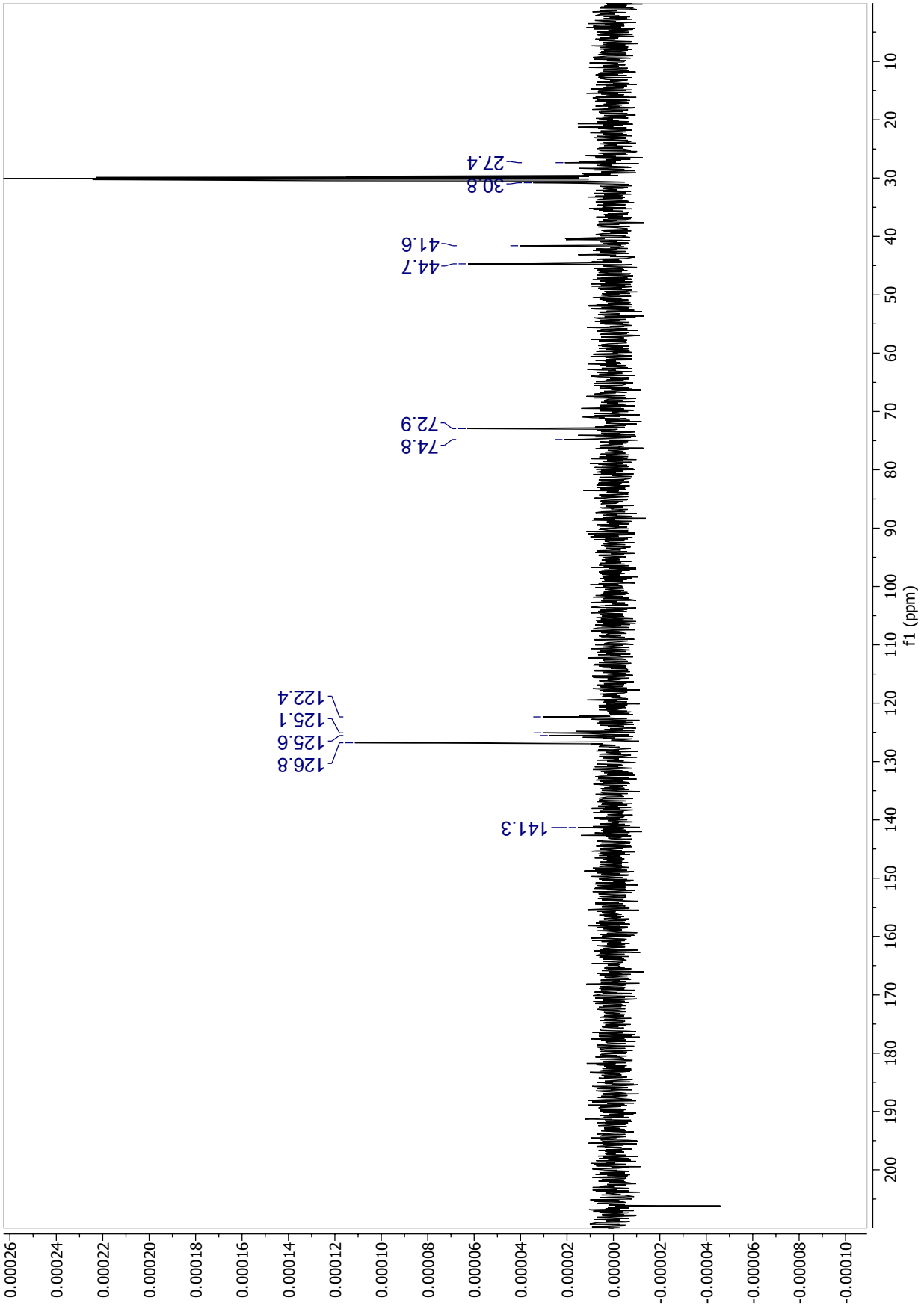
Appendix 7. ^{13}C spectrum of M1 in acetone-*d*6 (45798 scans) (4.9 mg)



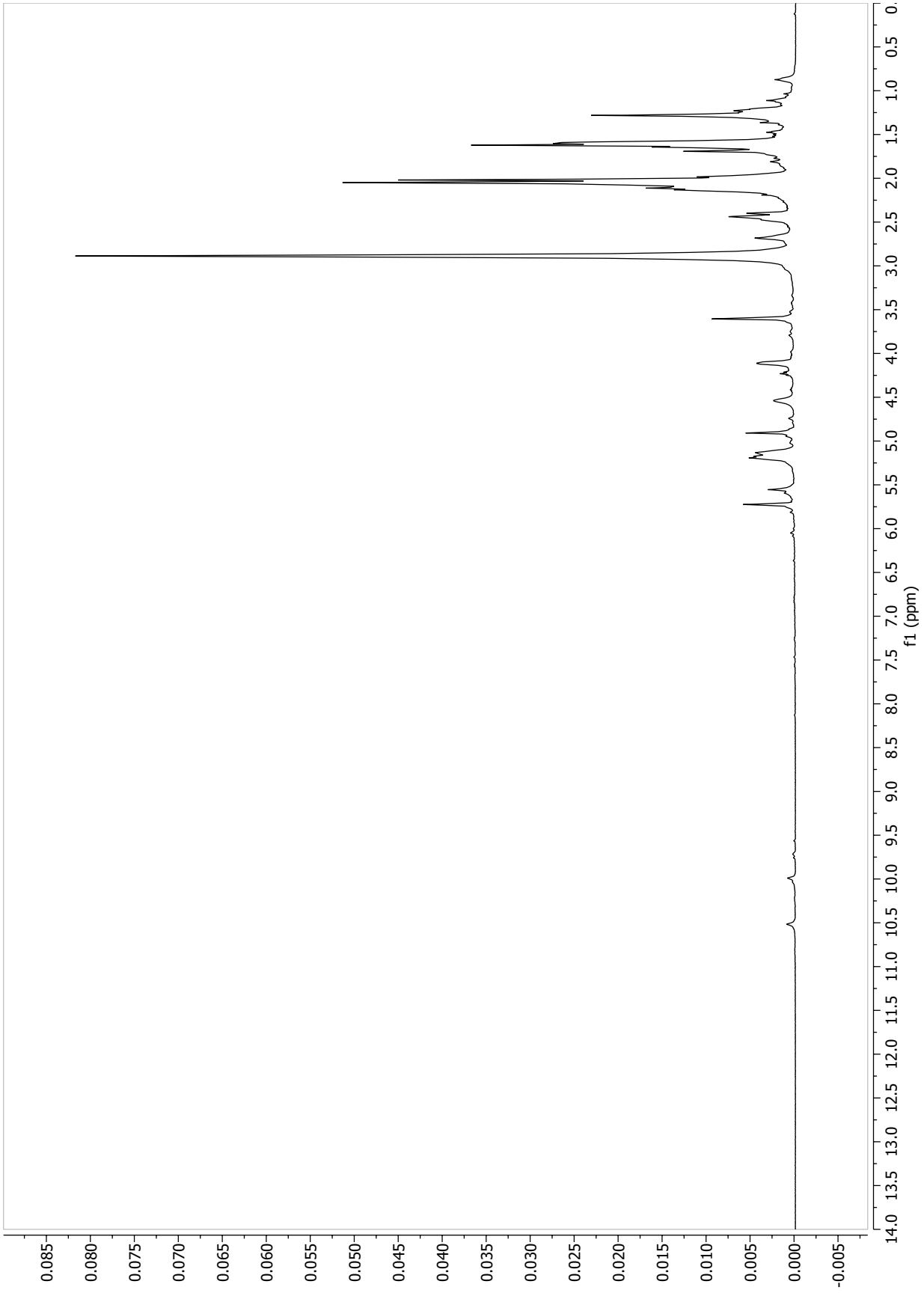
Appendix 8. DEPT-135 spectrum of M1 in acetone-*d*6 (17000 scans) (4.9 mg)



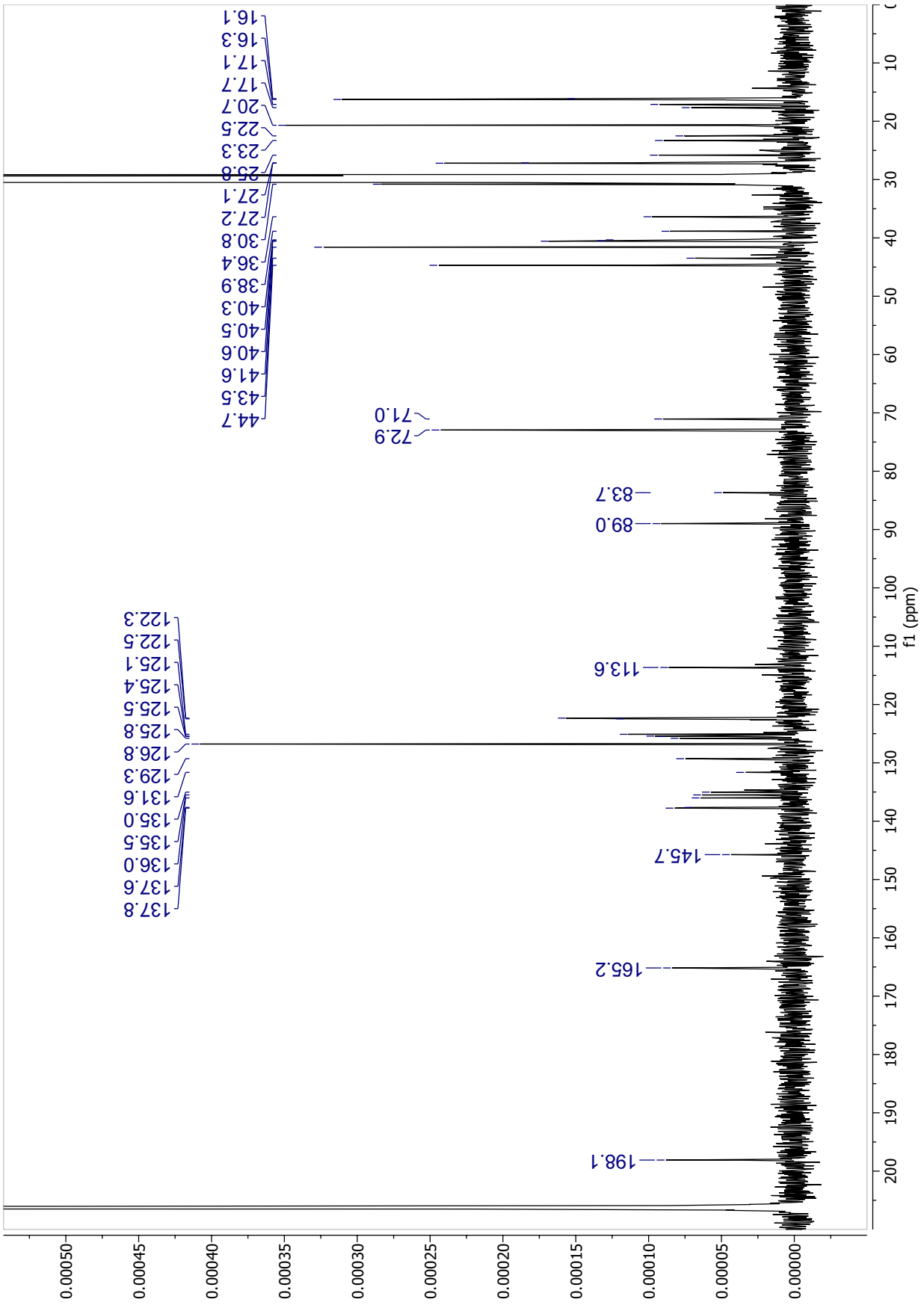
Appendix 9. DEPT-90 spectrum of M1 in acetone-*d*6 (11000 scans) (4.9 mg)



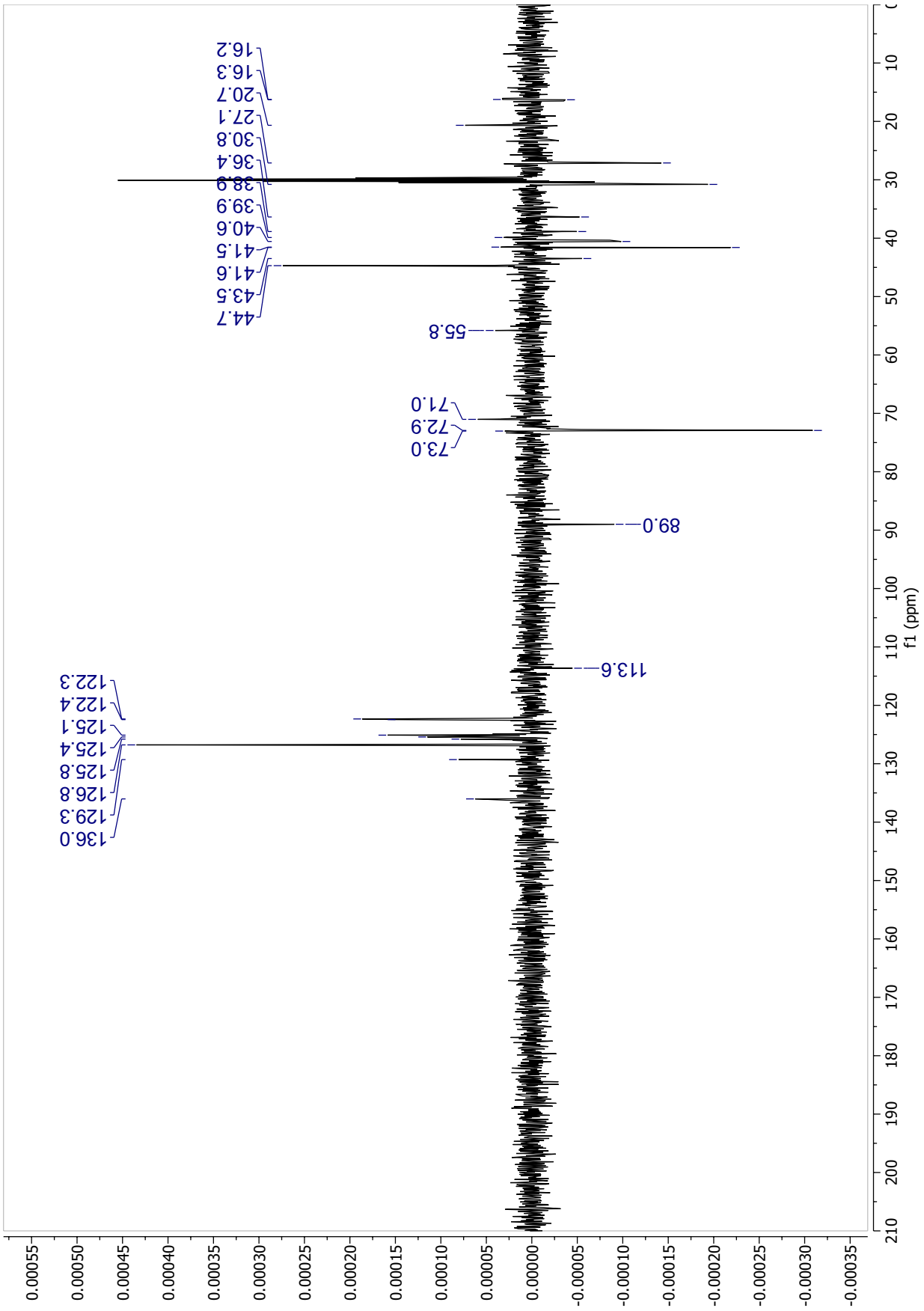
Appendix 10. ^{1}H spectrum of M2 in acetone-*d*6



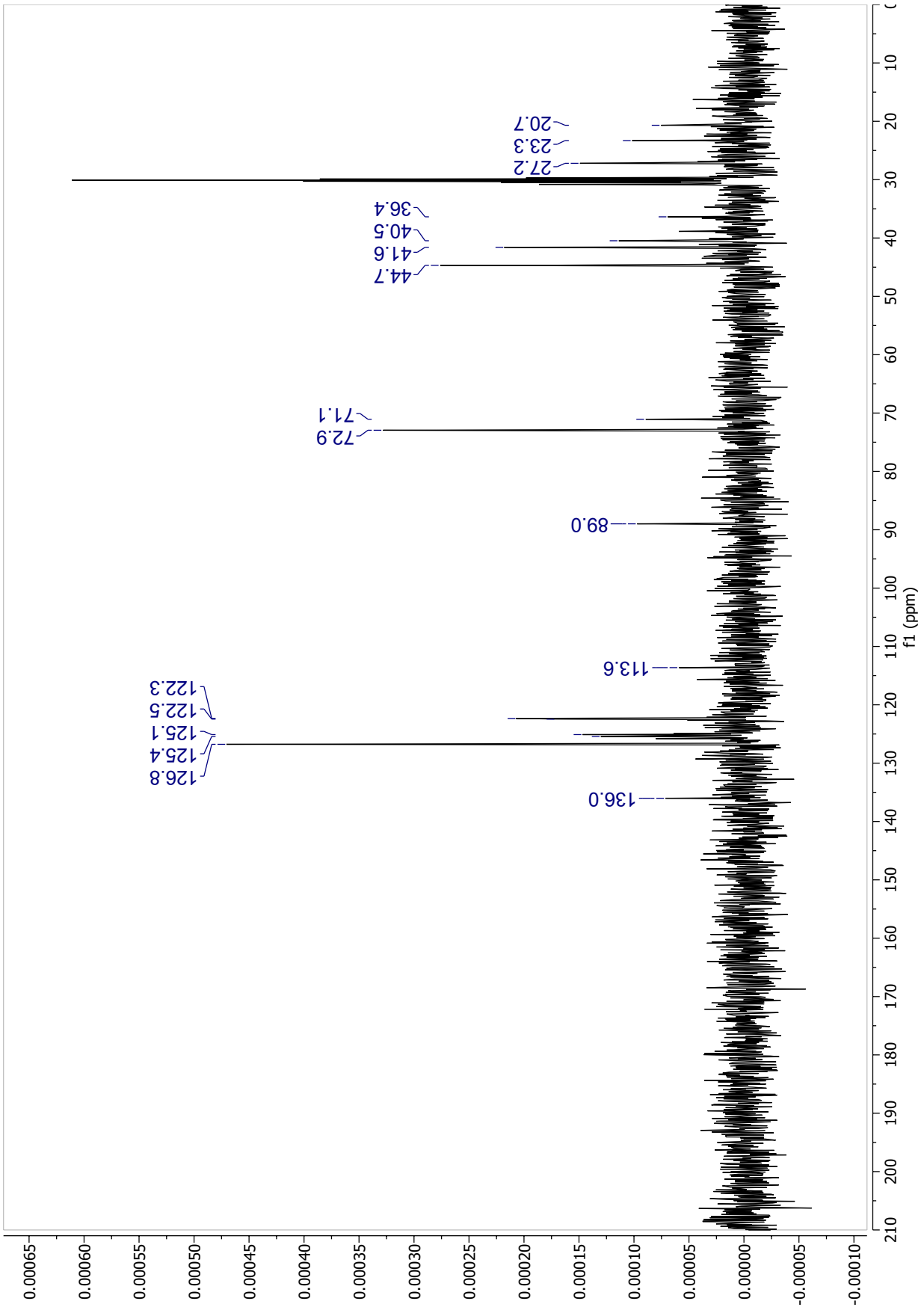
Appendix 11. ^{13}C spectrum of M2 in acetone-*d*6 (25000 scans) (5.8 mg)



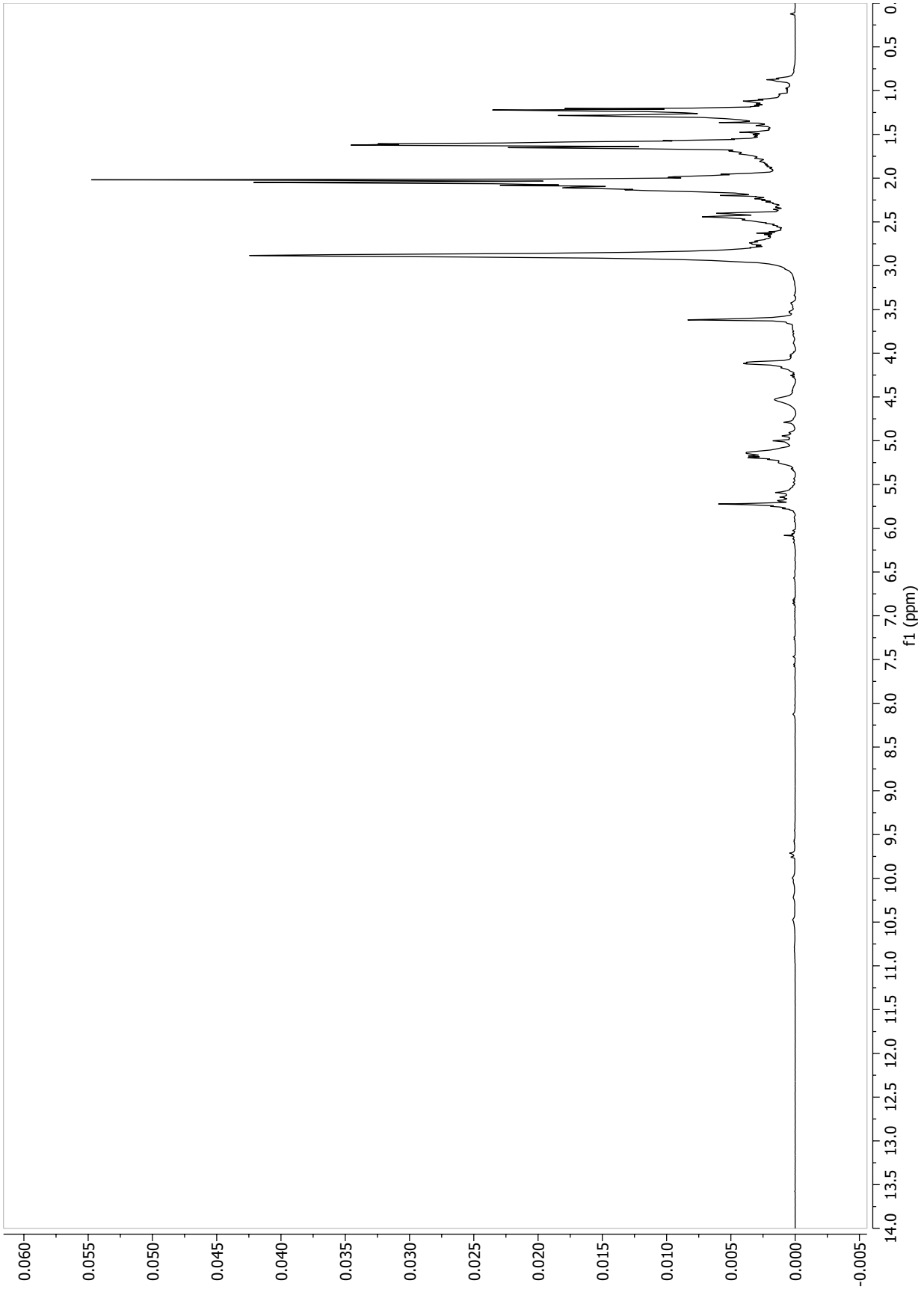
Appendix 12. DEPT-135 spectrum of M2 in acetone-*d*6 (6000 scans) (5.8 mg)



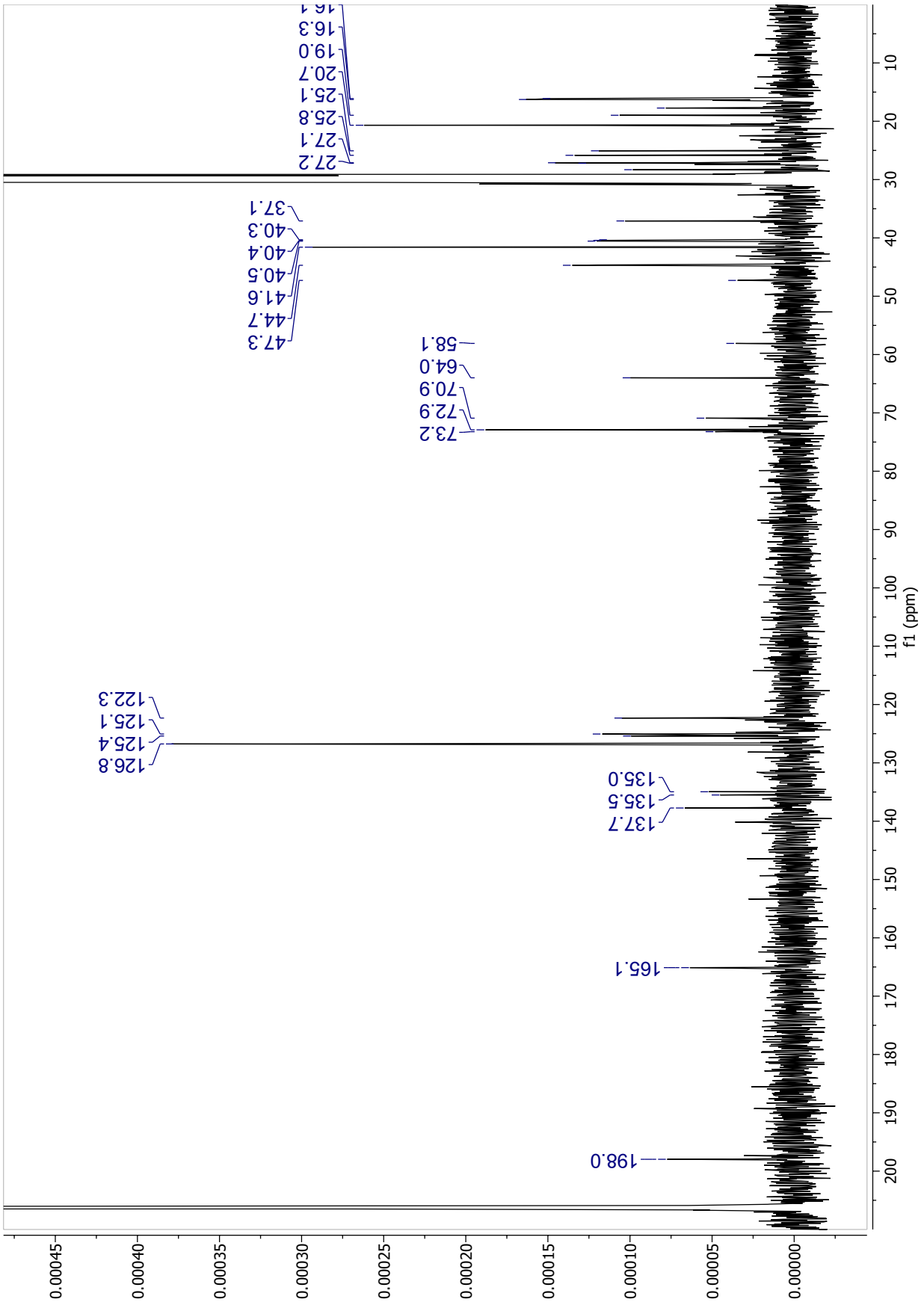
Appendix 13. DEPT-90 spectrum of M2 in acetone-*d*6 (3000 scans) (5.8 mg)



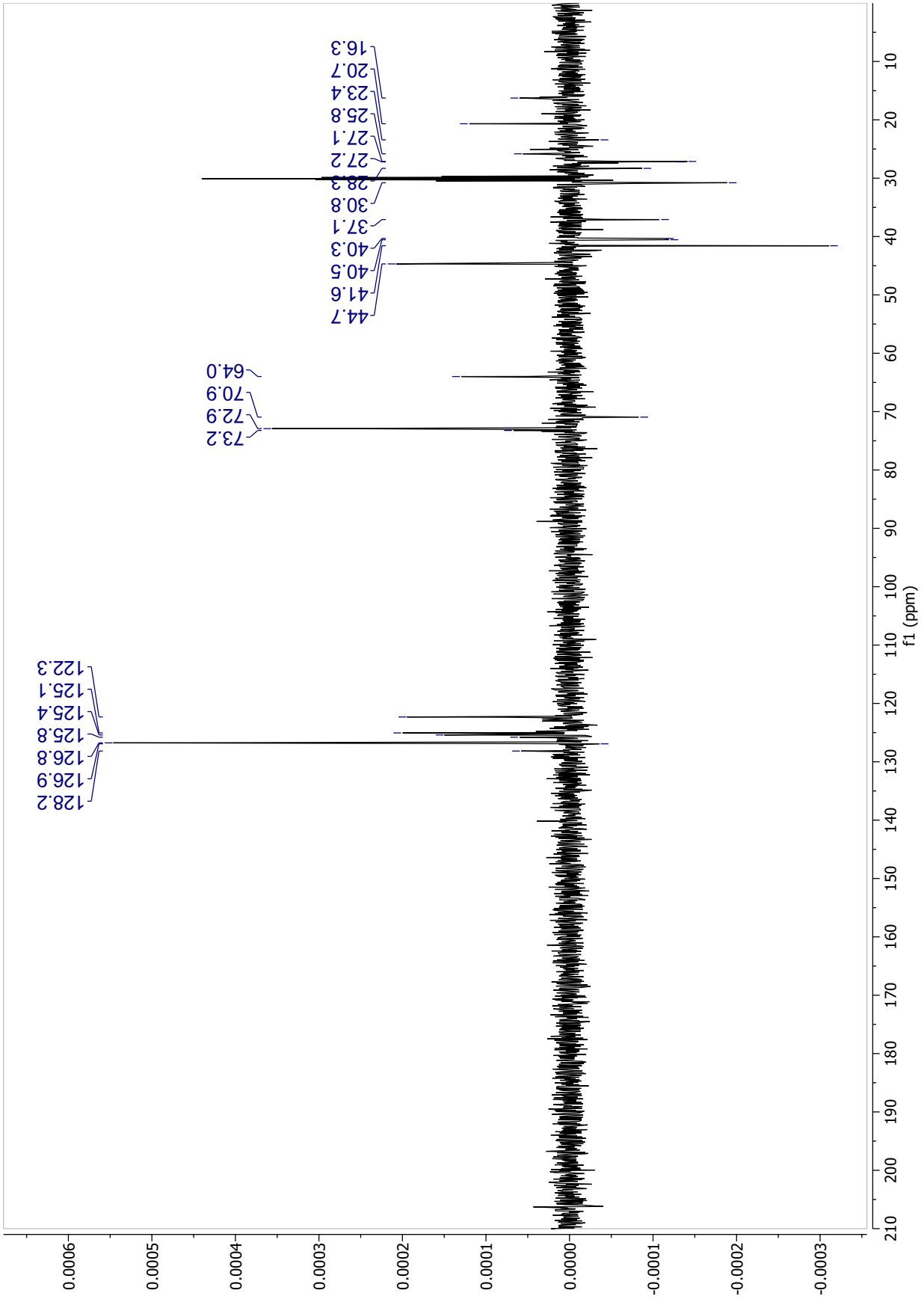
Appendix 14. ^{1}H spectrum of M3 in acetone-*d*6



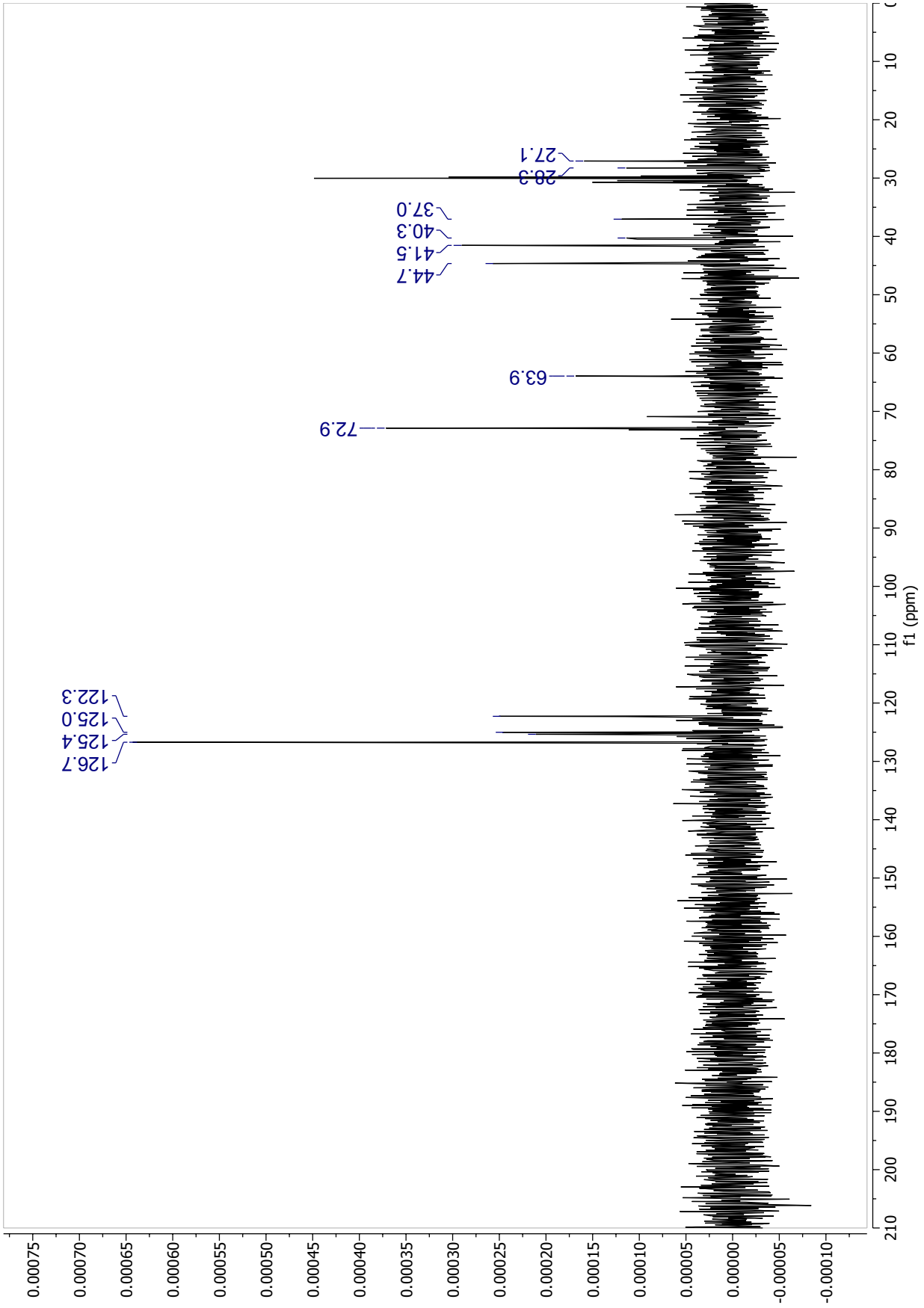
Appendix 15. ^{13}C spectrum of M3 in acetone-*d*6 (15401 scans) (6.1 mg)



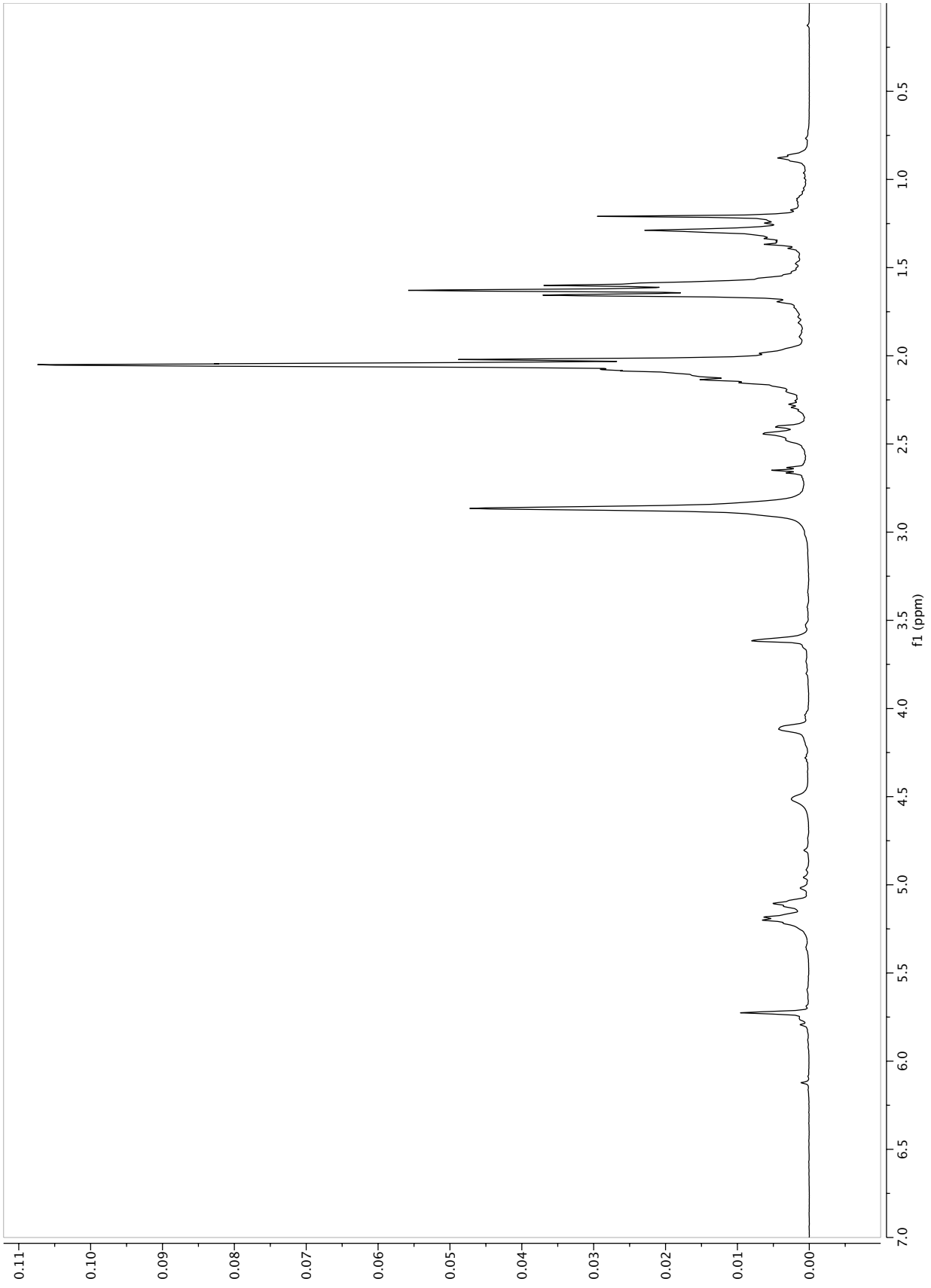
Appendix 16. DEPT-135 spectrum of M3 in acetone-*d*6 (10000 scans) (6.1 mg)



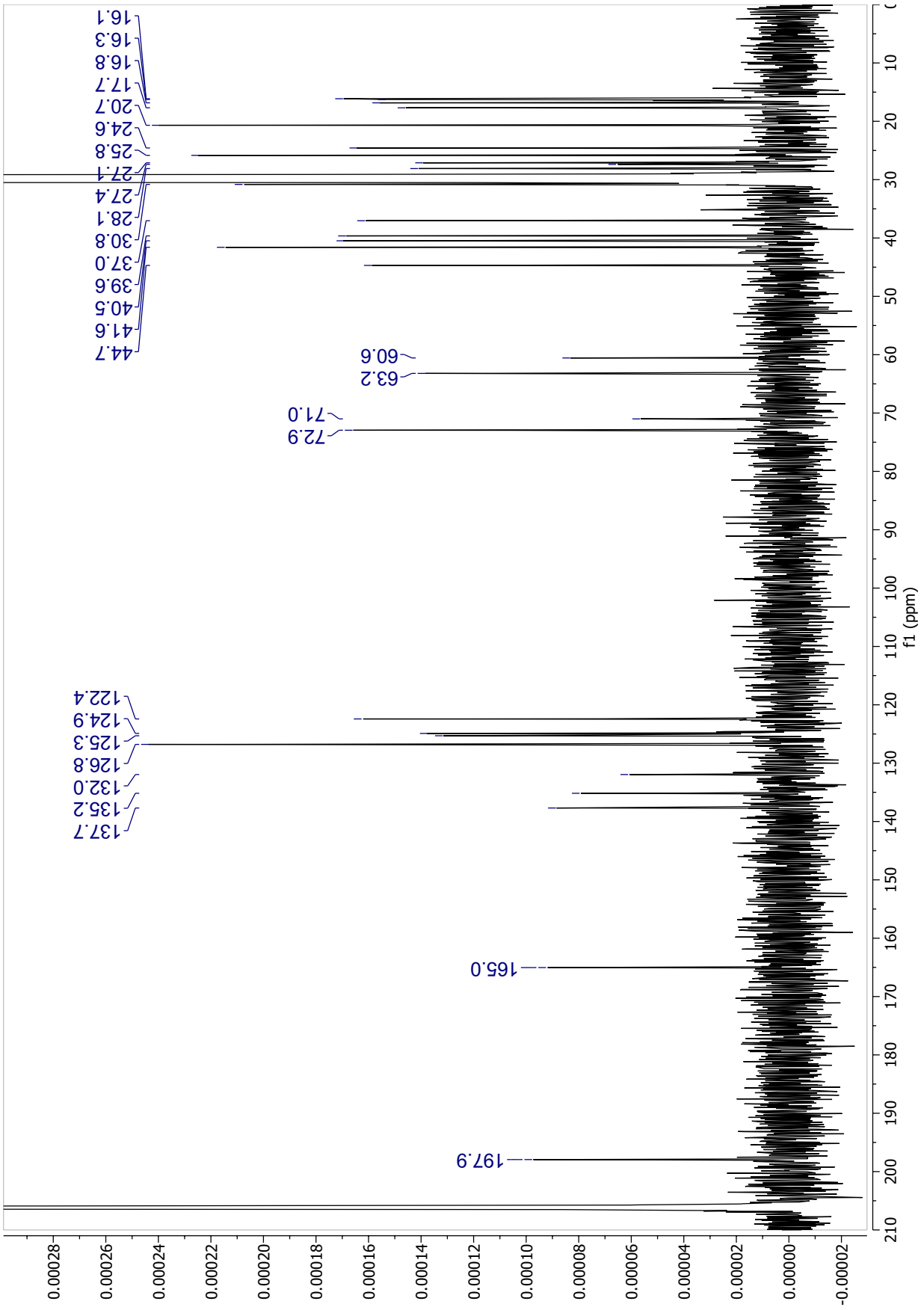
Appendix 17. DEPT-90 spectrum of M3 in acetone-*d*6 (2500 scans) (6.1 mg)



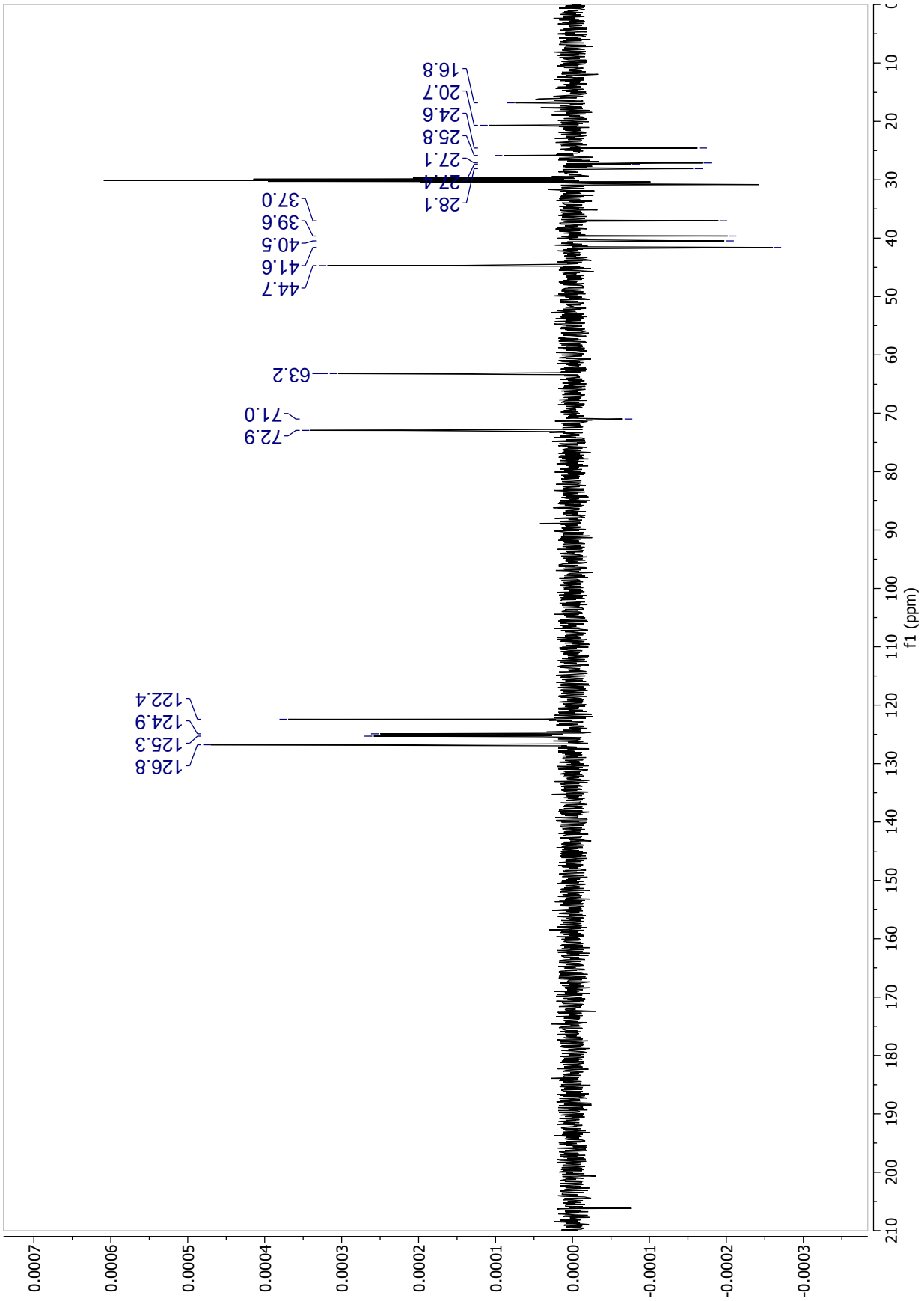
Appendix 18. ^{1}H spectrum of M4 in acetone-*d*6



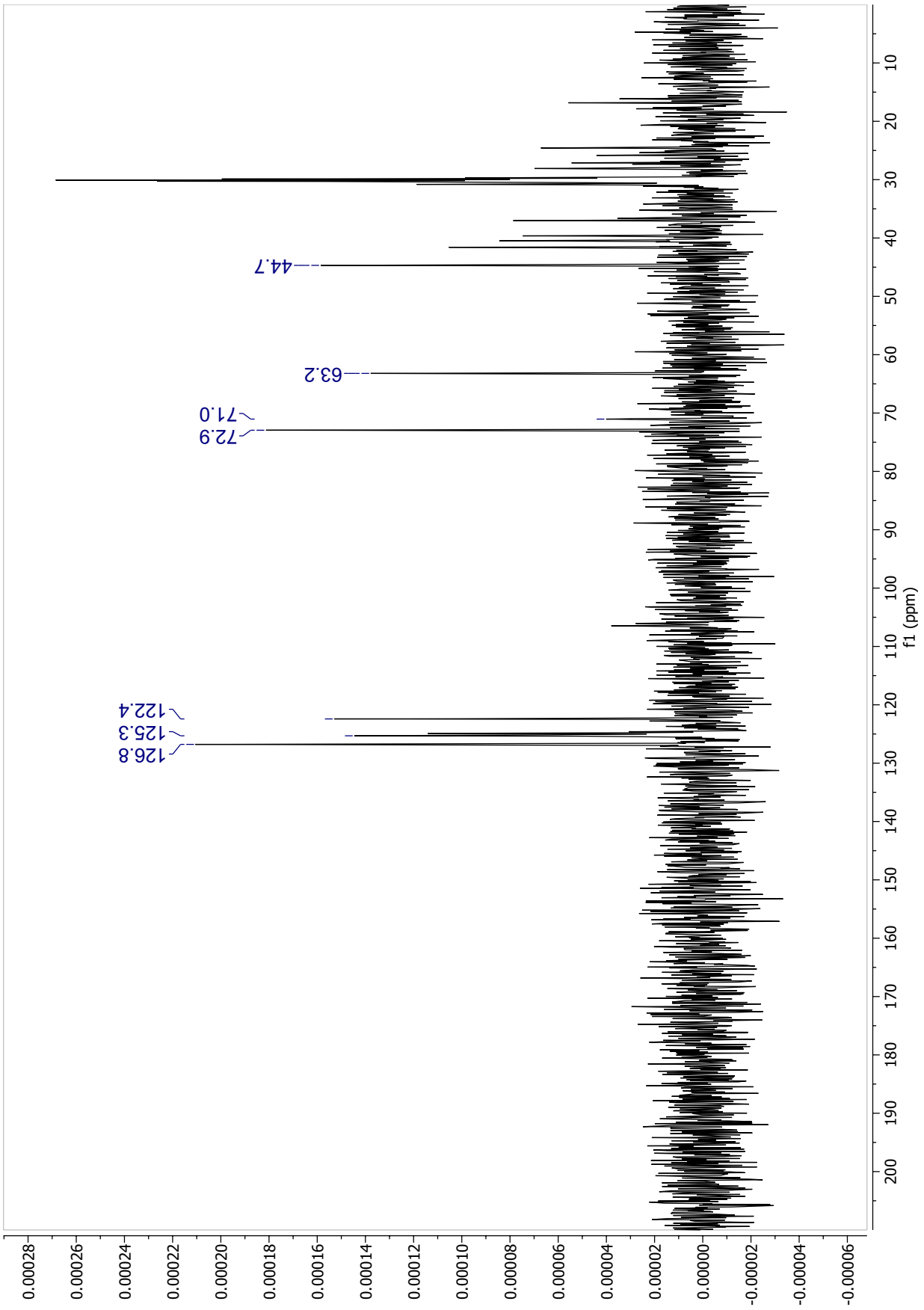
Appendix 19. ^{13}C spectrum of M4 in acetone-*d*6 (15000 scans) (5.2 mg)



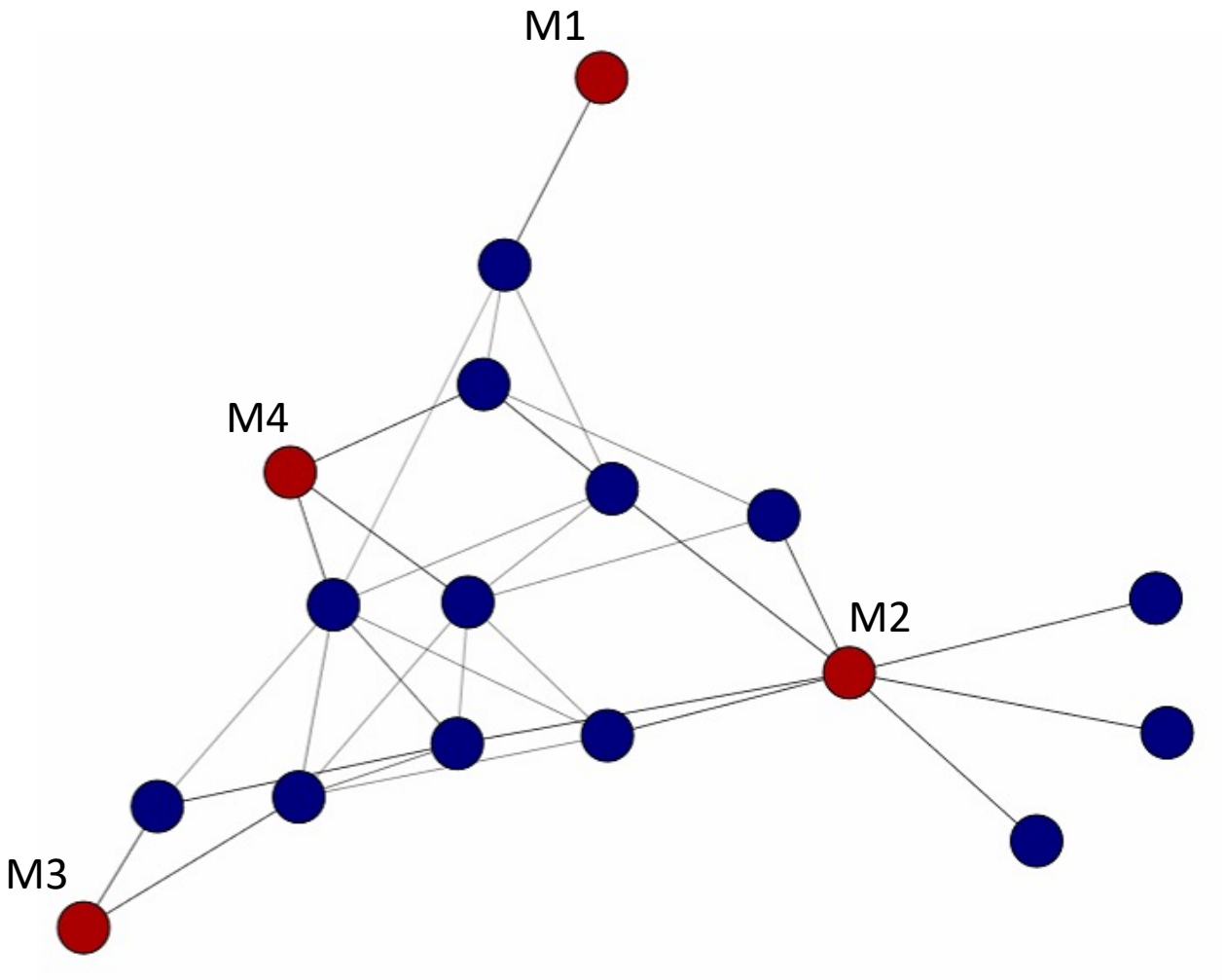
Appendix 20. DEPT-135 spectrum of M4 in acetone-*d*6 (10000 scans) (5.2 mg)



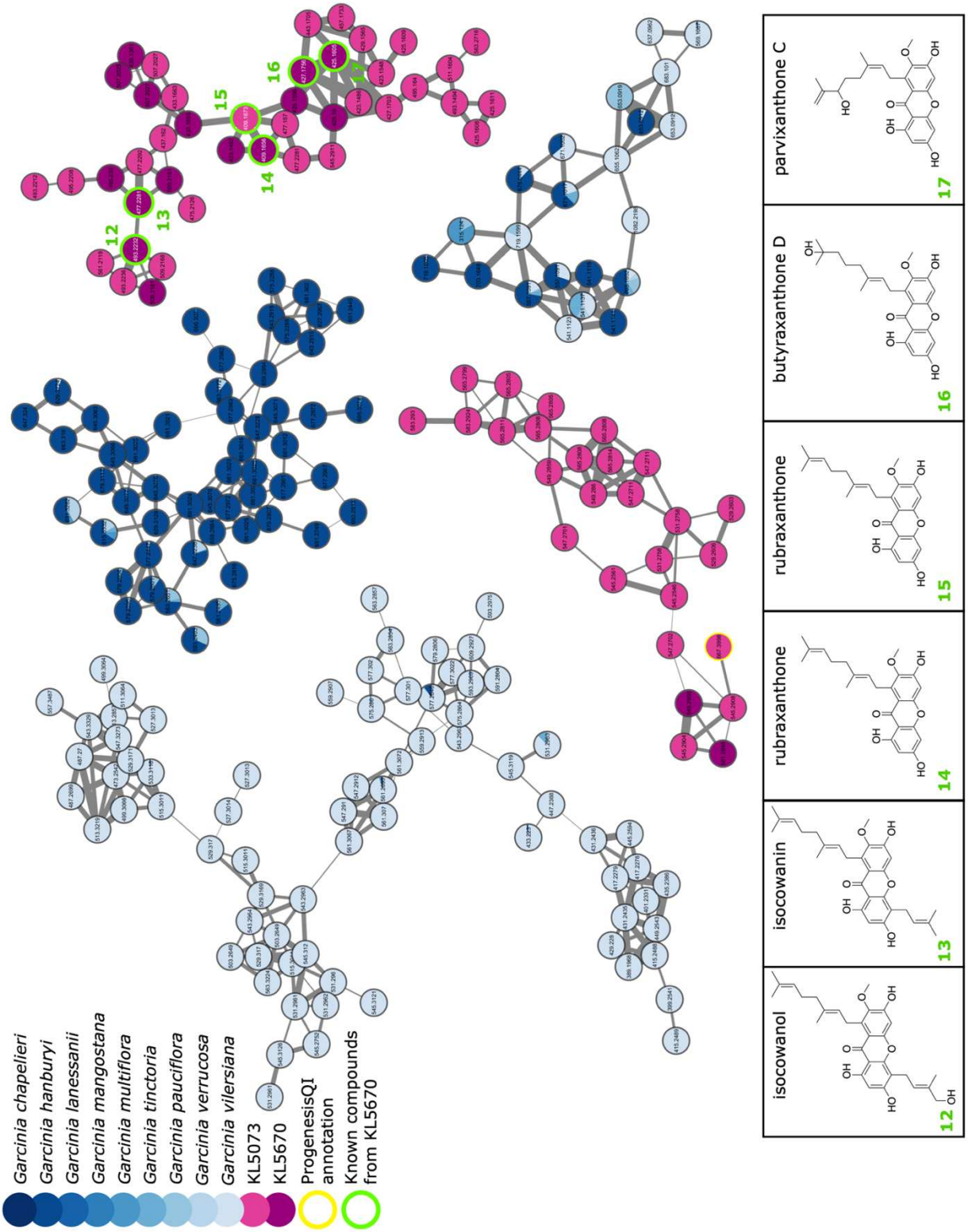
Appendix 21. DEPT-90 spectrum of M4 in acetone-*d*6 (2408 scans) (5.2 mg)



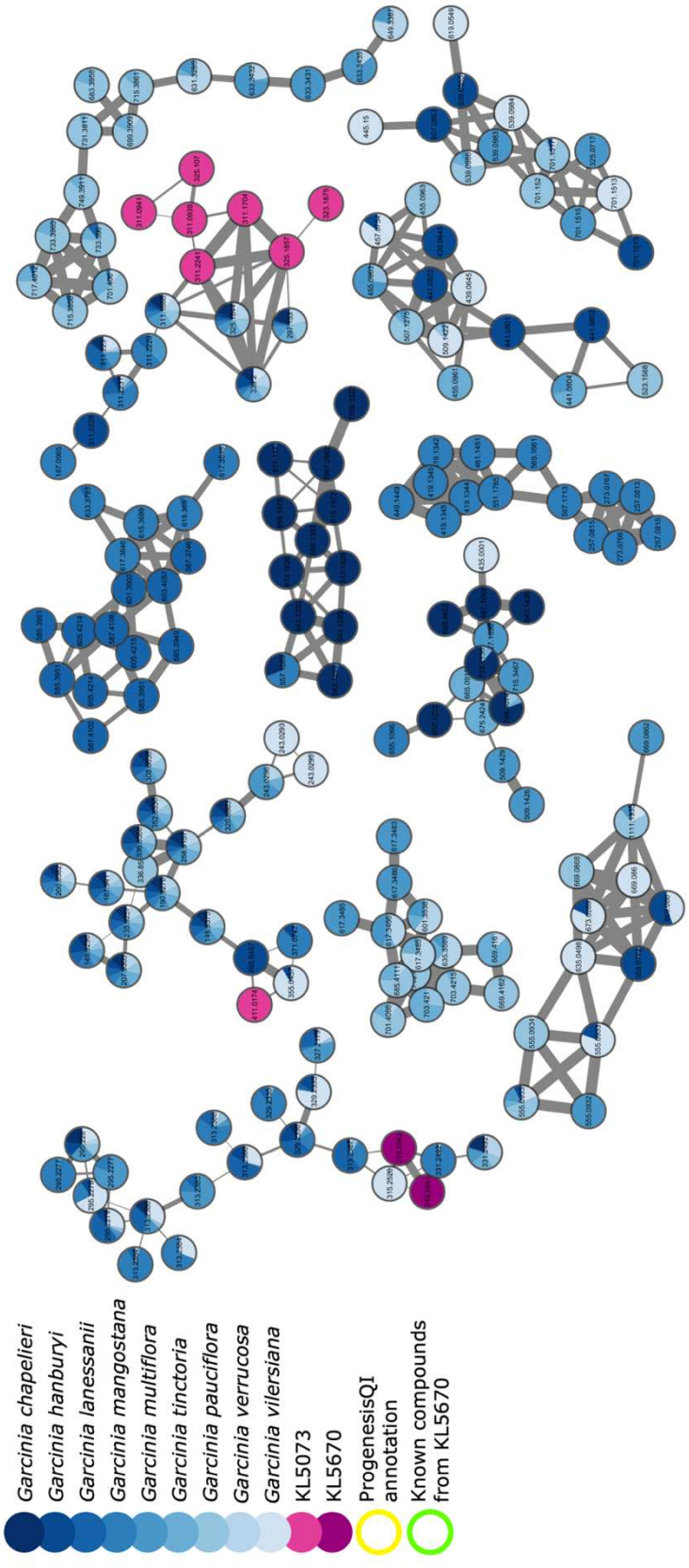
Appendix 22. MADByTE similarity network of TPTQ derivatives from F4



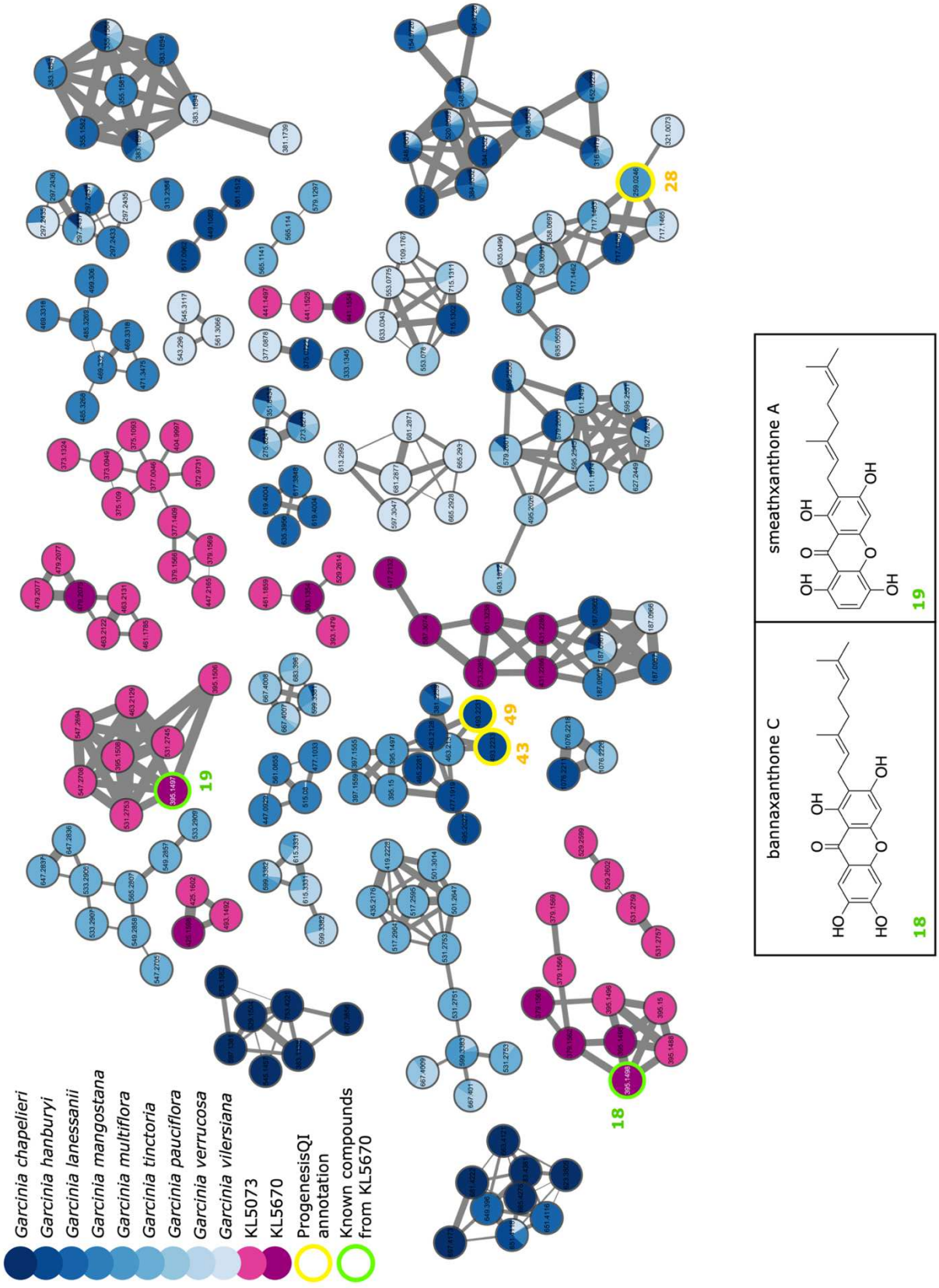
Appendix 23. Molecular network (part 2) of *G. parvifolia* batches (KL5670 (dark pink) and KL5073 (light pink)) along with the previously identified molecules from KL5670 (green) and Pierre Fabre *Garcinia* extracts (blue gradient).



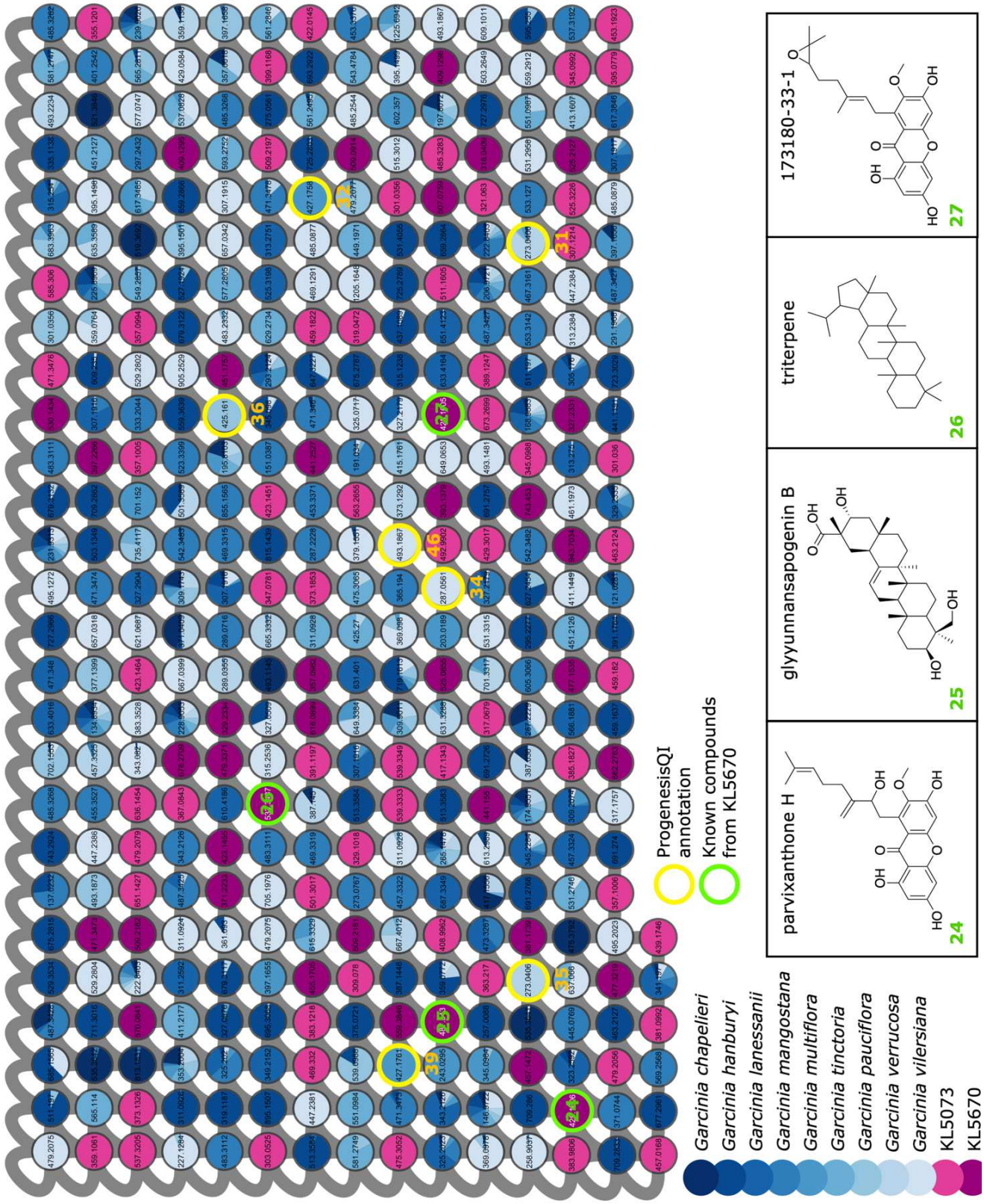
Appendix 24. Molecular network (part 3) of *G. parvifolia* batches (KL5670 (dark pink) and KL5073 (light pink)) along with the previously identified molecules from KL5670 (green) and Pierre Fabre *Garcinia* extracts (blue gradient).



Appendix 25. Molecular network (part 4) of *G. parvifolia* batches (KL5670 (dark pink) and KL5073 (light pink)) along with the previously identified molecules from KL5670 (green) and Pierre Fabre *Garcinia* extracts (blue gradient).



Appendix 26. Molecular network (part 6) of *G. parvifolia* batches (KL5670 (dark pink) and KL5073 (light pink)) along with the previously identified molecules from KL5670 (green) and Pierre Fabre *Garcinia* extracts (blue gradient).



Appendix 27. Compound annotation by Progenesis MetaScope and the *in-house* XanthonesDB

Network number	m/z	Retention time (min)	Score	Fragmentation Score	Compound ID
28	259.0246	0.62	33.8	72.8	WR3_NEG30
			30.8	58.1	WR3_NEG20
			21.1	7.4	WR3_NEG40
29	259.0246	2.04	33.3	70.1	WR3_NEG30
			31.4	60.8	WR3_NEG20
			21.1	6.9	WR3_NEG40
30	259.0248	2.23	34.6	76.1	WR3_NEG30
			31.6	61.5	WR3_NEG20
			23.1	16.5	WR3_NEG40
31	273.0406	2.31	37.9	97.1	WR2_NEG20
			37.6	91.8	WR2_NEG30
			21.3	10.5	WR2_NEG40
			21.3	10.5	WR4_NEG40
32	427.1758	2.58	20.2	2.3	parvixanthone G_NEG40
			19.9	0.9	parvixanthone G_NEG20
33	259.0247	2.62	35.0	78.1	WR3_NEG30
			32.4	65.4	WR3_NEG20
			21.6	9.4	WR3_NEG40
34	287.0561	2.63	37.6	92.6	WR7_NEG20
			38.0	91.0	WR7_NEG30
			36.9	87.8	WR5_NEG20
			29.0	50.5	WR7_NEG40
35	273.0406	2.78	38.3	98.9	WR2_NEG20
			37.9	93.0	WR2_NEG30
			20.6	6.9	WR2_NEG40
			20.6	6.9	WR4_NEG40
36	425.1610	3.02	29.0	49.5	parvixanthone H_NEG30
			23.5	22.4	parvixanthone H_NEG20
			23.0	18.7	parvixanthone H_NEG40
37	259.0244	3.09	32.6	67.4	WR3_NEG30
			31.1	60.6	WR3_NEG20
			23.2	18.8	WR3_NEG40
38	259.0245	3.18	33.3	70.6	WR3_NEG30
			31.7	62.7	WR3_NEG20
			23.6	20.2	WR3_NEG40
39	427.1761	4.00	26.4	36.1	garcinone D_NEG30
			46.1	35.9	garcinone D_NEG20
			26.7	35.8	cratoxylone_NEG20
			26.4	34.1	parvixanthone G_NEG30
			26.2	32.8	cratoxylone_NEG40
			25.9	32.0	cratoxylone_NEG30
			25.7	30.7	garcinone D_NEG40
			24.3	23.9	parvixanthone G_NEG20
			23.8	23.0	butyraxanthone D_NEG20
			23.0	19.8	butyraxanthone D_NEG30
			21.3	9.8	butyraxanthone D_NEG40
			20.1	2.5	parvixanthone G_NEG40
40	425.1609	4.02	29.5	51.7	parvixanthone H_NEG30
			27.9	44.0	parvixanthone H_NEG20
			23.9	23.0	parvixanthone H_NEG40
41	257.0453	4.13	36.6	87.9	WR6_NEG20
			35.7	85.9	WR6_NEG30
			23.7	28.6	WR6_NEG40
42	425.1609	4.16	21.3	10.1	parvixanthone H_NEG40
			20.8	8.1	parvixanthone H_NEG30

Network number	m/z	Retention time (min)	Score	Fragmentation Score	Compound ID
			19.4	1.6	parvixanthone H_NEG20
43	493.2233	4.38	21.9	11.2	isocowanol_NEG40
			19.8	0.8	isocowanol_NEG20
			19.5	0.6	isocowanol_NEG30
44	425.1608	4.43	32.1	64.4	parvixanthone H_NEG30
			31.8	63.0	parvixanthone H_NEG20
			21.3	9.3	parvixanthone H_NEG40
45	425.1607	4.71	29.2	49.5	parvixanthone H_NEG30
			27.5	41.3	parvixanthone H_NEG20
			20.5	5.2	parvixanthone H_NEG40
46	493.2234	5.17	21.7	10.5	isocowanol_NEG40
			21.2	7.8	isocowanol_NEG20
			20.2	4.2	isocowanol_NEG30
47	409.1658	5.50	30.2	53.7	alpha-mangostin_NEG20
			30.3	52.8	alpha-mangostin_NEG30
			29.1	47.3	alpha-mangostin_NEG40
			24.4	26.0	rubraxanthone_NEG20
			24.3	23.6	rubraxanthone_NEG30
			21.4	9.7	rubraxanthone_NEG40
48	409.1657	5.68	28.5	45.2	alpha-mangostin_NEG20
			28.6	44.4	alpha-mangostin_NEG30
			28.4	43.6	alpha-mangostin_NEG40
			24.0	23.7	rubraxanthone_NEG20
			23.7	20.7	rubraxanthone_NEG30
			21.2	8.9	rubraxanthone_NEG40
			39.1	0.0	garcinone D_NEG20
49	493.2231	6.03	22.0	11.2	isocowanol_NEG40
			20.0	1.0	isocowanol_NEG20
			19.7	0.9	isocowanol_NEG30

References

Allard, P.-M. *et al.* (2022) 'Open and reusable annotated mass spectrometry dataset of a chemodiverse collection of 1,600 plant extracts', *GigaScience*, 12, p. giac124. Available at: https://doi.org/10.1093/gigascience/giac124.

Allen, F. *et al.* (2014) 'CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra', *Nucleic Acids Research*, 42(W1), pp. W94–W99. Available at: https://doi.org/10.1093/nar/gku436.

Alsabil, K. *et al.* (2016) 'Semisynthetic and natural garcinoic acid isoforms as new mPGES-1 inhibitors', *Planta Medica*, 82(11/12), pp. 1110–1116. Available at: https://doi.org/10.1055/s-0042-108739.

Alsabil, K. *et al.* (2017) 'Efficient ortho-formylation in vitamin E series, application to the semi-synthesis of natural 5- and 7-formyl-δ-tocotrienols revealing an unprecedented 5-bromo-7-formyl exchange', *Tetrahedron*, 73(49), pp. 6863–6870. Available at: https://doi.org/10.1016/j.tet.2017.10.039.

Atanasov, A.G. *et al.* (2015) 'Discovery and resupply of pharmacologically active plant-derived natural products: A review', *Biotechnology Advances*, 33(8), pp. 1582–1614. Available at: https://doi.org/10.1016/j.biotechadv.2015.08.001.

Atanasov, A.G. *et al.* (2021) 'Natural products in drug discovery: advances and opportunities', *Nature Reviews Drug Discovery*, 20(3), pp. 200–216. Available at: https://doi.org/10.1038/s41573-020-00114-z.

Aydoğan, C. (2020) 'Recent advances and applications in LC-HRMS for food and plant natural products: a critical review', *Analytical and Bioanalytical Chemistry*, 412(9), pp. 1973–1991. Available at: https://doi.org/10.1007/s00216-019-02328-6.

Blaženović, I. *et al.* (2018) 'Software tools and approaches for compound identification of LC-MS/MS data in metabolomics', *Metabolites*, 8(2), p. 31. Available at: https://doi.org/10.3390/metabo8020031.

Brown, S.D. (2017) 'The chemometrics revolution re-examined: an assessment of the past, present and future of chemometrics', *Journal of Chemometrics*, 31(1), p. e2856. Available at: https://doi.org/10.1002/cem.2856.

Bruguière, A. *et al.* (2018) '^{13}C-NMR dereplication of Garcinia extracts: predicted chemical shifts as reliable databases', *Fitoterapia*, 131, pp. 59–64. Available at: https://doi.org/10.1016/j.fitote.2018.10.003.

Bruguière, A. *et al.* (2020) 'MixONat, a software for the dereplication of mixture based on ^{13}C NMR spectroscopy', *Analytical Chemistrysilva-cas*, 92(13), pp. 8793–8801. Available at: https://doi.org/10.1021/acs.analchem.0c00193.

Bruguière, A. *et al.* (2021) '^{13}C NMR dereplication using MixONat software: a practical guide to decipher natural products mixtures', *Planta Medica*, 87(12/13), pp. 1061–1068. Available at: https://doi.org/10.1055/a-1470-0446.

Bruguière, A., Derbré, S. and Richomme, P. (2020) *MixONat*, *SourceForge*. Available at: https://sourceforge.net/projects/mixonat/ (Accessed: 14 January 2022).

Castillo, S. *et al.* (2011) 'Algorithms and tools for the preprocessing of LC–MS metabolomics data', *Chemometrics and Intelligent Laboratory Systems*, 108(1), pp. 23–32. Available at: https://doi.org/10.1016/j.chemolab.2011.03.010.

ChemSpider Search plug-in - Progenesis QI (2023). Available at: https://www.nonlinear.com/progenesis/qi/v2.0/faq/compound-search-chemspider.aspx (Accessed: 12 April 2023).

Christensen, S.B. (2021) 'Natural products that changed society', *Biomedicines*, 9(5), p. 472. Available at: https://doi.org/10.3390/biomedicines9050472.

Clardy, J. and Walsh, C. (2004) 'Lessons from natural molecules', *Nature*, 432(7019), pp. 829–837. Available at: https://doi.org/10.1038/nature03194.

Convention on Biological Diversity (2023) *The Nagoya Protocol on access and benefit-sharing*. Secretariat of the Convention on Biological Diversity. Available at: https://www.cbd.int/abs/ (Accessed: 14 April 2023).

Creek, D.J. *et al.* (2014) 'Metabolite identification: are you sure? And how do your peers gauge your confidence?', *Metabolomics*, 10(3), pp. 350–353. Available at: https://doi.org/10.1007/s11306-014-0656-8.

Cunha, L.C.S. *et al.* (2017) 'Bioassay-guided fractionation and antimicrobial and cytotoxic activities of *Cassia bakeriana* extracts', *Revista Brasileira de Farmacognosia*, 27(1), pp. 91–98. Available at: https://doi.org/10.1016/j.bjp.2016.08.002.

Derbré, S. *et al.* (2010) 'Automating a 96-well microtiter plate assay for identification of AGEs inhibitors or inducers: application to the screening of a small natural compounds library', *Analytical and Bioanalytical Chemistry*, 398(4), pp. 1747–1758. Available at: https://doi.org/10.1007/s00216-010-4065-1.

Dictionary of natural products 31.2 (2015) *CHEMnetBASE*. Available at: https://dnp.chemnetbase.com/chemical/ChemicalSearch.xhtml?dswid=4743 (Accessed: 13 March 2023).

Dinh, C.P. *et al.* (2020) 'Structure-based design, semi-synthesis and anti-inflammatory activity of tocotrienolic amides as 5-lipoxygenase inhibitors', *European Journal of Medicinal Chemistry*, 202, p. 112518. Available at: https://doi.org/10.1016/j.ejmech.2020.112518.

Dodds, J.N. and Baker, E.S. (2019) 'Ion mobility spectrometry: fundamental concepts, instrumentation, applications, and the road ahead', *Journal of the American Society for Mass Spectrometry*, 30(11), pp. 2185–2195. Available at: https://doi.org/10.1007/s13361-019-02288-2.

Efenberger-Szmechtyk, M., Nowak, A. and Czyzowska, A. (2021) 'Plant extracts rich in polyphenols: antibacterial agents and natural preservatives for meat and meat products', *Critical Reviews in Food Science and Nutrition*, 61(1), pp. 149–178. Available at: https://doi.org/10.1080/10408398.2020.1722060.

Elemental composition plug-in - Progenesis QI (2023). Available at: https://www.nonlinear.com/progenesis/qi/v2.0/faq/compound-search-elemental-composition.aspx (Accessed: 12 April 2023).

Ernst, M. *et al.* (2014) 'Mass spectrometry in plant metabolomics strategies: from analytical platforms to data acquisition and processing', *Natural Product Reports*, 31(6), p. 784. Available at: https://doi.org/10.1039/c3np70086k.

Ge, G. and Wong, G.W. (2008) 'Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles', *BMC Bioinformatics*, 9(1), p. 275. Available at: https://doi.org/10.1186/1471-2105-9-275.

GNPS - Analyze, Connect, and Network with your Mass Spectrometry Data (2023). Available at: https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp (Accessed: 16 April 2023).

Gu, J. *et al.* (2013) 'Use of natural products as chemical library for drug discovery and network pharmacology', *PLoS one*. Edited by D. Cox, 8(4), p. e62839. Available at: https://doi.org/10.1371/journal.pone.0062839.

Gyawali, R. and Ibrahim, S.A. (2014) 'Natural products as antimicrobial agents', *Food Control*, 46, pp. 412–429. Available at: https://doi.org/10.1016/j.foodcont.2014.05.047.

Harrison, C. (2014) 'Patenting natural products just got harder', *Nature Biotechnology*, 32(5), pp. 403–404. Available at: https://doi.org/10.1038/nbt0514-403a.

Harvey, A.L., Edrada-Ebel, R. and Quinn, R.J. (2015) 'The re-emergence of natural products for drug discovery in the genomics era', *Nature Reviews Drug Discovery*, 14(2), pp. 111–129. Available at: https://doi.org/10.1038/nrd4510.

Hernández-García, J., Briones-Moreno, A. and Blázquez, M.A. (2021) 'Origin and evolution of gibberellin signaling and metabolism in plants', *Seminars in Cell & Developmental Biology*, 109, pp. 46–54. Available at: https://doi.org/10.1016/j.semcdb.2020.04.009.

Hines, K.M. *et al.* (2017) 'Large-scale structural characterization of drug and drug-like compounds by high-throughput ion mobility-mass spectrometry', *Analytical Chemistry*, 89(17), pp. 9023–9030. Available at: https://doi.org/10.1021/acs.analchem.7b01709.

Hinnenkamp, V. *et al.* (2018) 'Comparison of CCS values determined by traveling wave ion mobility mass spectrometry and drift tube ion mobility mass spectrometry', *Analytical Chemistry*, 90(20), pp. 12042–12050. Available at: https://doi.org/10.1021/acs.analchem.8b02711.

Hong, C.P., Jang, G.Y. and Ryu, H. (2021) 'Gibberellins enhance plant growth and ginsenoside content in *Panax ginseng*', *Journal of Plant Biotechnology*, 48(3), pp. 186–192. Available at: https://doi.org/10.5010/JPB.2021.48.3.186.

How are the compound ions grouped into compounds? (2023). Available at: https://www.nonlinear.com/progenesis/qi/v2.0/faq/how-does-deconvolution-work.aspx (Accessed: 12 April 2023).

How are the scores calculated for possible compound identifications? (no date). Available at: http://www.nonlinear.com/progenesis/qi/v2.0/faq/identifications-scoring-algorithm.aspx (Accessed: 16 February 2021).

How does the theoretical fragmentation work? - Progenesis QI v3.0 (2023). Available at: https://www.nonlinear.com/progenesis/qi/v3.0/faq/theoretical-fragmentation-algorithm.aspx (Accessed: 6 June 2023).

Hubert, J., Nuzillard, J.-M. and Renault, J.-H. (2017) 'Dereplication strategies in natural product research: how many tools and methodologies behind the same concept?', *Phytochemistry Reviews*, 16(1), pp. 55–95. Available at: https://doi.org/10.1007/s11101-015-9448-7.

Inui, T. *et al.* (2012) 'Unbiased evaluation of bioactive secondary metabolites in complex matrices', *Fitoterapia*, 83(7), pp. 1218–1225. Available at: https://doi.org/10.1016/j.fitote.2012.06.012.

Islam, S., Alam, R. and Kim, S. (2022) 'Improved coverage of plant metabolites using powder laser desorption/ionization coupled with Fourier-transform ion cyclotron mass spectrometry', *Food Chemistry*, 373, p. 131541. Available at: https://doi.org/10.1016/j.foodchem.2021.131541.

Jamiołkowska, A. (2020) 'Natural compounds as elicitors of plant resistance against diseases and new biocontrol strategies', *Agronomy*, 10(2), p. 173. Available at: https://doi.org/10.3390/agronomy10020173.

Kazmi, S.A. *et al.* (2006) 'Alignment of high resolution mass spectra: development of a heuristic approach for metabolomics', *Metabolomics*, 2(2), pp. 75–83. Available at: https://doi.org/10.1007/s11306-006-0021-7.

Kellogg, J.J. *et al.* (2016) 'Biochemometrics for natural products research: comparison of data analysis approaches and application to identification of bioactive compounds', *Journal of Natural Products*, 79(2), pp. 376–386. Available at: https://doi.org/10.1021/acs.jnatprod.5b01014.

Kind, T. *et al.* (2013) 'LipidBlast in silico tandem mass spectrometry database for lipid identification', *Nature Methods*, 10(8), pp. 755–758. Available at: https://doi.org/10.1038/nmeth.2551.

Kliman, M., May, J.C. and McLean, J.A. (2011) 'Lipid analysis and lipidomics by structurally selective ion mobility-mass spectrometry', *Biochimica et Biophysica Acta*, 1811(11), pp. 935–945. Available at: https://doi.org/10.1016/j.bbalip.2011.05.016.

Le Pogam, P. *et al.* (2015) 'Matrix-free UV-laser desorption ionization mass spectrometry as a versatile approach for accelerating dereplication studies on lichens', *Analytical Chemistry*, 87(20), pp. 10421–10428. Available at: https://doi.org/10.1021/acs.analchem.5b02531.

Le Pogam, P. *et al.* (2020) 'A thorough evaluation of matrix-free laser desorption ionization on structurally diverse alkaloids and their direct detection in plant extracts', *Analytical and Bioanalytical Chemistry*, 412(27), pp. 7405–7416. Available at: https://doi.org/10.1007/s00216-020-02872-6.

Lever, J., Krzywinski, M. and Altman, N. (2017) 'Principal component analysis', *Nature Methods*, 14(7), pp. 641–642. Available at: https://doi.org/10.1038/nmeth.4346.

Lianza, M. *et al.* (2021) 'The three pillars of natural product dereplication. Alkaloids from the bulbs of *Urceolina peruviana* (C. Presl) J.F. Macbr. as a preliminary test case', *Molecules*, 26(3), p. 637. Available at: https://doi.org/10.3390/molecules26030637.

LipidBlast Search plug-in - Progenesis QI (2023). Available at: https://www.nonlinear.com/progenesis/qi/v2.0/faq/compound-search-lipid-blast.aspx (Accessed: 12 April 2023).

Lopes, C.T. *et al.* (2010) 'Cytoscape web: an interactive web-based network browser', *Bioinformatics*, 26(18), pp. 2347–2348. Available at: https://doi.org/10.1093/bioinformatics/btq430.

López-Fernández, H. *et al.* (2015) 'Mass-up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery', *BMC Bioinformatics*, 16(1), p. 318. Available at: https://doi.org/10.1186/s12859-015-0752-4.

Lourenço, S.C., Moldão-Martins, M. and Alves, V.D. (2019) 'Antioxidants of natural plant origins: from sources to food industry applications', *Molecules*, 24(22), p. 4132. Available at: https://doi.org/10.3390/molecules24224132.

MetaScope search plug-in - Progenesis QI (no date). Available at: https://www.nonlinear.com/progenesis/qi/v2.0/faq/compound-search-metascope.aspx (Accessed: 9 March 2022).

Nahrstedt, A. and Butterweck, V. (2010) 'Lessons learned from herbal medicinal products: the example of St. John's Wort', *Journal of Natural Products*, 73(5), pp. 1015–1021. Available at: https://doi.org/10.1021/np1000329.

Nica, D.V. *et al.* (2013) 'A novel exploratory chemometric approach to environmental monitorring by combining block clustering with partial least square (PLS) analysis', *Chemistry Central Journal*, 7(1), p. 145. Available at: https://doi.org/10.1186/1752-153X-7-145.

Niedermeyer, T.H.J. and Strohalm, M. (2012) 'mMass as a Software Tool for the Annotation of Cyclic Peptide Tandem Mass Spectra', *PLoS ONE*. Edited by J.M. Koomen, 7(9), p. e44913. Available at: https://doi.org/10.1371/journal.pone.0044913.

NIST MS/MS Search plug-in - Progenesis QI (2023). Available at: https://www.nonlinear.com/progenesis/qi/v2.1/faq/compound-search-nist.aspx (Accessed: 12 April 2023).

Nonlinear Dynamics Progenesis software (no date). Available at: https://www.nonlinear.com/progenesis/ (Accessed: 9 March 2023).

Pein, H. *et al.* (2018) 'Endogenous metabolites of vitamin E limit inflammation by targeting 5-lipoxygenase', *Nature Communications*, 9(1), p. 3834. Available at: https://doi.org/10.1038/s41467-018-06158-5.

Pluskal, T. *et al.* (2010) 'MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data', *BMC Bioinformatics*, 11, p. 395. Available at: https://doi.org/10.1186/1471-2105-11-395.

Rates, S.M.K. (2001) 'Plants as source of drugs', *Toxicon*, 39(5), pp. 603–613. Available at: https://doi.org/10.1016/S0041-0101(00)00154-9.

Richomme, P. *et al.* (2017) 'New biodegradable anti-scaling agents'. Université d'Angers.

Ruttkies, C. *et al.* (2016) 'MetFrag relaunched: incorporating strategies beyond in silico fragmentation', *Journal of Cheminformatics*, 8(1), p. 3. Available at: https://doi.org/10.1186/s13321-016-0115-9.

Ruttkies, C., Neumann, S. and Posch, S. (2019) 'Improving MetFrag with statistical learning of fragment annotations', *BMC Bioinformatics*, 20(1), p. 376. Available at: https://doi.org/10.1186/s12859-019-2954-7.

dos Santos, N.A. *et al.* (2019) 'LDI and MALDI-FT-ICR imaging MS in *Cannabis* leaves: optimization and study of spatial distribution of cannabinoids', *Analytical Methods*, 11(13), pp. 1757–1764. Available at: https://doi.org/10.1039/C9AY00226J.

Schinkovitz, A. *et al.* (2018) 'Matrix-free laser desorption ionization mass spectrometry as a functional tool for the analysis and differentiation of complex phenolic mixtures in propolis: a new approach to quality control', *Analytical and Bioanalytical Chemistry*, 410(24), pp. 6187–6195. Available at: https://doi.org/10.1007/s00216-018-1225-1.

Schrimpe-Rutledge, A.C., Sherrod, S.D. and McLean, J.A. (2018) 'Improving the discovery of secondary metabolite natural products using ion mobility–mass spectrometry', *Current Opinion in Chemical Biology*, 42, pp. 160–166. Available at: https://doi.org/10.1016/j.cbpa.2017.12.004.

Schymanski, E.L. *et al.* (2014) 'Identifying small molecules via high resolution mass spectrometry: communicating confidence', *Environmental Science & Technology*, 48(4), pp. 2097–2098. Available at: https://doi.org/10.1021/es5002105.

Silva-Castro, L.F. *et al.* (2021) 'Using ^{13}C-NMR dereplication to aid in the identification of xanthones present in the stem bark extract of *Calophyllum brasiliense*', *Phytochemical Analysis*, 32(6), pp. 1102–1109. Available at: https://doi.org/10.1002/pca.3051.

Simoneau, P. *et al.* (2014) 'Potentiating agents for protecting plants from fungal infections'.

Skopikova, M. *et al.* (2020) 'Matrix-free laser desorption ionization mass spectrometry as an efficient tool for the rapid detection of opiates in crude extracts of Papaver somniferum', *Journal of Agricultural and Food Chemistry*, 68(3), pp. 884–891. Available at: https://doi.org/10.1021/acs.jafc.9b05153.

Smith, C.A. *et al.* (2006) 'XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification', *Analytical Chemistry*, 78(3), pp. 779–787. Available at: https://doi.org/10.1021/ac051437y.

Strohalm, M. *et al.* (2008) 'mMass data miner: an open source alternative for mass spectrometric data analysis', *Rapid Communications in Mass Spectrometry*, 22(6), pp. 905–908. Available at: https://doi.org/10.1002/rcm.3444.

Strohalm, M. *et al.* (2010) 'mMass 3: a cross-platform software environment for precise analysis of mass spectrometric data', *Analytical Chemistry*, 82(11), pp. 4648–4651. Available at: https://doi.org/10.1021/ac100818g.

Sturm, M. *et al.* (2008) 'OpenMS – An open-source software framework for mass spectrometry', *BMC Bioinformatics*, 9(1), p. 163. Available at: https://doi.org/10.1186/1471-2105-9-163.

Tsugawa, H. *et al.* (2015) 'MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis', *Nature Methods*, 12(6), pp. 523–526. Available at: https://doi.org/10.1038/nmeth.3393.

Tsugawa, H. *et al.* (2016) 'Hydrogen rearrangement rules: computational MS/MS Ffagmentation and structure elucidation using MS-FINDER software', *Analytical Chemistry* [Preprint]. Available at: https://doi.org/10.1021/acs.analchem.6b00770.

Twaij, B.M. and Hasan, Md.N. (2022) 'Bioactive secondary metabolites from plant sources: types, synthesis, and their therapeutic uses', *International Journal of Plant Biology*, 13(1), pp. 4–14. Available at: https://doi.org/10.3390/ijpb13010003.

Wang, F. *et al.* (2021) 'CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification', *Analytical Chemistry*, 93(34), pp. 11692–11700. Available at: https://doi.org/10.1021/acs.analchem.1c01465.

Wang, M. *et al.* (2016) 'Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking', *Nature Biotechnology*, 34(8), pp. 828–837. Available at: https://doi.org/10.1038/nbt.3597.

Wilson, R.M. and Danishefsky, S.J. (2006) 'Small molecule natural products in the discovery of therapeutic agents: the synthesis connection', *The Journal of Organic Chemistry*, 71(22), pp. 8329–8351. Available at: https://doi.org/10.1021/jo0610053.

Wolf, S. *et al.* (2010) 'In silico fragmentation for computer assisted identification of metabolite mass spectra', *BMC Bioinformatics*, 11, p. 148. Available at: https://doi.org/10.1186/1471-2105-11-148.

Wu, K. *et al.* (2021) 'New insights into gibberellin signaling in regulating plant growth–metabolic coordination', *Current Opinion in Plant Biology*, 63, p. 102074. Available at: https://doi.org/10.1016/j.pbi.2021.102074.

Zenobi, R. and Knochenmuss, R. (1998) 'Ion formation in MALDI mass spectrometry', *Mass Spectrometry Reviews*, 17(5), pp. 337–366. Available at: https://doi.org/10.1002/(SICI)1098-2787(1998)17:5<337::AID-MAS2>3.0.CO;2-S.

Zhu, L. *et al.* (2016) 'Simultaneous analysis of tocopherols, tocotrienols, phospholipids, γ-oryzanols and β-carotene in rice by ultra-high performance liquid chromatography coupled to a linear ion trap-orbitrap mass spectrometer', *Analytical Methods*, 8(28), pp. 5628–5637. Available at: https://doi.org/10.1039/C6AY00556J.

Nouvelles approches pour identifier les biomarqueurs présents dans des mélanges complexes de produits naturels

Mots clés: Produits naturels, biochimiométrie, HRMS^2, réseaux moléculaires, mobilité d'ions, RMN ^{13}C

Résumé: L'identification de produits naturels (PNs) (bio)actifs dans les mélanges complexes représente un défi important en raison de la présence de nombreux composés ayant des structures et des (bio)activités diverses. Cette thèse donne un aperçu des outils et des stratégies actuels et émergents pour l'identification de ces PNs (bio)actifs. Les approches traditionnelles, à savoir, le fractionnement bioguidé (FBG) suivi d'une analyse par résonance magnétique nucléaire (RMN) et spectrométrie de masse (SM) pour l'élucidation de la structure des composés, continuent de jouer un rôle important dans l'identification des molécules actives. Cependant, des progrès récents ont conduit au développement de nouvelles techniques telles que l'analyse (bio)chimiométrique, la déréplication basées sur la SM ou la RMN ou des approches combinées, qui permettent une hiérarchisation efficace pour l'élucidation des composés (bio)actifs. Cette thèse met ainsi en évidence les points forts et les limites de chaque technique et donne un aperçu du potentiel de leur utilisation combinée pour atteindre le plus haut niveau de confiance dans l'annotation de PNs (bio)actifs dans des matrices complexes.

New approaches to identify biomarkers in complex mixtures of natural products

Keywords: Natural products, biochemometrics, HRMS^2, molecular network, ion mobility, NMR ^{13}C

Abstract: The identification of (bio)active natural products (NPs) from complex mixtures is challenging due to the presence of numerous compounds with diverse structures and (bio)activities. This thesis provides an overview of current and emerging tools and strategies for the identification of (bio)active NPs in complex mixtures. Traditional approaches, bioassay-guided fractionation (BGF) followed by nuclear magnetic resonance (NMR) and mass spectrometry (MS) analyses for compound structure elucidation, continue to play an important role in active NPs identification. However, recent advances have led to the development of novel techniques such as (bio)chemometric analysis, dereplication based on MS or NMR or combined approaches, which allow efficient prioritization for the elucidation of (bio)active NPs. This thesis highlights the strengths and limitations of each technique and provides insights into their combined use to achieve the highest level of confidence in the identification of (bio)active natural products from complex matrices.