Discovery of new chemistry and enzymology from Cyanobacteria

Sara Raquel Ferreira Freitas
Doutoramento em Biologia
Departamento de Biologia
2023

Orientador
Doutor Pedro Nuno Leão, FCUP e CIIMAR

Coorientador
Professor Doutor Vítor Vasconcelos, FCUP e CIIMAR


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Nota: A presente página só deverá ser incluída no caso de o júri aprovar a tese com recomendação de correção, pelo estudante, dos erros, imprecisões ou incorreções formais identificados e expressamente referidos durante as provas.

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

A/O Orientadora(o),



Porto, __11__/__01__/__2024___

Para o meu eterno Diogo

Acknowledgements

The present work would not be possible to achieve without the help of a group of people that were a part of this journey, to whom I would like to express my gratitude.

First and foremost, I would like to thank my supervisor Dr. Pedro Leão for accepting me as his student, for helping and guiding me throughout all the project and for encouraging me to be a better researcher. For all the guidance, patience and discussion with the thesis writing, that pushed me to do my best. It was an honor to be your “sidekick” in establishing your lab, CNP, it was stressful, but it helped me understand that I really like to organize laboratories. So, thank you for all the trust. Also, for understanding when I needed time and space to return to the lab after my suspension. Your kind words and conversations really helped me to return to my normal life again. Thank you very much for everything.

To my co-supervisor, Professor Vitor Vasconcelos, I would also like to thank for the opportunity to do this project, for all the availability and support always present throughout the years. BBE (LEGE, still at the time) was my first laboratory where I started my master, and I always felt as a part of the family, thank you once again.

I would like to acknowledge Dr. Deniz Tasdmir and Arlette Wenzel-Storjohann, from GEOMAR Centre for Marine Biotechnology, research unit marine natural product chemistry, for kindly collaborate with this project regarding the biological testing of the isolated natural products.

I would also like to acknowledge all the help and support provided by my BBE colleagues and friends: Mica for always being there to help me, João Morais for all the knowledge of the cyanobacteria world and all the bad jokes that make me laugh every time, Flávio for all the support, all the playfulness and pranks that we did throughout these years and to Mariana Reis for all the advices and help in the lab, I’m very grateful. A special acknowledgement to Ralph, Tiago, Leonor, Sandra Pereira and Joana Almeida for helping me not only during these years, but also the bioassays performed for the thesis.

To Raquel Castelo Branco, I would like to express all my gratitude, we started together in LEGE and 11 years after we are still together, now in CNP. You are one of my closest friends, thank you for all the help and discussions about lab work, mainly about the biosynthesis area. For all the support throughout these rough years in my personal life, having you by my side helped me overcome a lot, I will always be grateful.

I would also like to acknowledge all my CNP colleagues and friends. In particular, Teresinha for being my teammate during the thesis writing and for all the long nights of work in the lab. To Sandra, Ligia, Ana, Adriana and Amaranta thank you for all the help and advices during these years, and to João Reis, for all the support and all the laughs, you are something. Also need to acknowledge Paulo with all the advices during my thesis writing, Eunice and Marco for helping me a lot during these last months with all the tasks from the lab work, and Mariana Girão for helping me with some bioassays. All of you, as well as other members of CNP, created an amazing work environment during these years, so thank you.

I would also like to acknowledge the Fundação para a Ciência e Tecnologia (FCT) for the PhD grant SFRH/BD/116009/2016, allowing me to do this project.

E por último,

Gostaria de agradecer a todos os meus amigos por todo o apoio que me deram sempre, principalmente nestes últimos anos, sem vocês nada disto era possível: ás minhas “irmãs” Sarinha e Susana, à Diana, Ana, Cátia, Zé e Vítor, à Mariana Carneiro, à Marta Ortigão, á Necas, Klai, Tiaguinho, Tiago e Leonara. Acompanhar e puder fazer parte das vossas vidas, faz-me uma pessoa feliz. Um agradecimento especial à minha Inês que me ajudou de uma forma incrível, nunca me vou esquecer o que fizeste por mim. Um agradecimento especial também ao Zeca por todo o carinho, apoio e incentivo em ser sempre melhor.

À minha família, quero agradecer aos meus pais por serem os meus pilares, são os dois uma força da natureza, obrigada por tudo o que fizeram e que fazem por mim, amo-vos muito. À minha irmã, o meu maior orgulho, obrigada por fazeres de mim a irmã mais feliz do mundo. A toda a minha família um grande obrigada por me aturarem e por estarem sempre do meu lado, o meu amor por vocês é incondicional.

Ao meu eterno Diogo, obrigada por todo o amor e apoio, pelos nossos bichinhos e por me ensinares a maior lição de sempre: aproveitar todos os momentos e todas as pessoas na nossa vida, e ser feliz. Estarás para sempre no meu coração.

This study was partially funded by Fundação para a Ciência e Tecnologia through grants UIDB/04423/2020 and UIDP/04423/2020 and the scholarship SFRH/BD/116009/2016. This project has also received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952374 and the project ATLANTIDA (red. NORTE-01-0145-FEDER-000040), supported by the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the European Regional Development Fund (ERDF).

Resumo

As indústrias farmacêuticas e biotecnológicas têm-se focado na pesquisa e aplicação de produtos naturais isolados de diversos organismos. Um grupo de organismos que marca cada vez mais a sua presença na descoberta de produtos naturais é o das cianobactérias, um grupo de bactérias fotossintéticas amplamente distribuídas por toda a Terra. A sua capacidade de adaptação a todos os tipos de ambientes e condições adversas faz com que a produção/isolamento dos seus metabolitos secundários seja uma área de investigação promissora. Muitos produtos naturais isolados de cianobactérias são utilizados em diversas aplicações biológicas e atualmente o foco tem-se direcionado para as linhas de construção e principais mecanismos responsáveis pela sua produção. Compreender a capacidade biossintética e a ecologia química responsável pela produção destes produtos naturais de cianobactérias poderá aumentar o seu uso em aplicações biotecnológicas novas. O objetivo deste projeto é o isolamento de novos produtos naturais de um grupo de estirpes de cianobactérias, isoladas de uma variedade de ecossistemas portugueses e que fazem parte da coleção de culturas LEGE (LEGEcc), estabelecida na sede do CIIMAR. Um teste inicial foi realizado baseado nas características metabolómicas e de bioactividade destas estirpes de cianobactérias, de forma a indicar as mais promissoras. A seleção de duas cianobactérias simbióticas presentes na LEGEcc resultou na descoberta de novos produtos naturais. Primeiro, a descoberta de uma nova classe de lipoglicopéptidos, desmamidas A-C da cianobactéria *Desmonostoc muscorum* LEGE 12446, uma cianobionte isolada de uma raiz coraloide de uma cicadácea (*Cycas revoluta*). As estruturas químicas dos compostos foram elucidadas usando a combinação de espectroscopia 1D e 2D de ressonância magnética nuclear (RMN) e espectrometria de massa. As desmamidas são decapéptidos apresentando uma tirosina *O*-glicosilada e um resíduo raro de ácido gordo, ácido 3,5-dihidroxi-2-metildecanóico. A biossíntese destes produtos naturais foi estudada usando uma combinação de ferramentas de bioinformática e de experiências de marcação, propondo o grupo de genes biossintéticos *dsm*, potencialmente responsável pela produção destes compostos. A atividade antibiótica da desmamida B contra o patógeno que causa a putrefação negra em plantas, *Xanthomonas campestris*, juntamente com a produção abundante desta classe de compostos pela cianobactéria (> 1,5% peso seco), abre o caminho para a investigação do seu potencial papel na endossimbiose radicular. A exploração da segunda cianobactéria *Nostoc* sp. LEGE 12450, isolada da raiz coraloide da cicadácea *Encephalartos horridus*, resultou no isolamento de três novos produtos

naturais. O processo de elucidação estrutural foi realizado com a combinação de RMN e espectrometria de massa. O nostopeptolideo L5 é um novo lipopéptido, pertencente à família de compostos nostopeptolideos, e é formado por oito aminoácidos e um resíduo de ácido gordo, ácido 2-metilhex-2-enóico. Embora a biossíntese dos nostopeptolideos já seja conhecida, a síntese do raro ácido gordo modificado foi sugerida aqui pela primeira vez. As suomilidas 1077 e 1007, que fazem parte da família das aeruginosinas, são novos tetrapéptidos glicosilados e sulfonados. O grupo de genes biossintéticos *suo* já foi descrito para a suomilida, e este também está presente no genoma da *Nostoc* sp. LEGE 12450. O perfil biológico do nostopeptolideo L5 demonstrou um EC_{50} de 19,4 μM na adesão das larvas de *Mytilus galoprovincialis*, indicando o potencial biotecnológico deste produto natural como um agente anti-incrustante. Os nostopeptolideos participam no processo endossimbiótico das cianobactérias e a descoberta das suomilidas pode fornecer uma nova camada de conhecimento dos estímulos químicos responsáveis por esta antiga parceria entre cianobactérias e cicadáceas. A investigação de cianobactérias portuguesas, específicas de um ecossistema, demonstrou a diversidade ampla e inexplorada deste grupo de organismos, no que diz respeito à descoberta de novos produtos naturais e das suas potenciais aplicações biológicas.

Palavras-chave: Cianobactérias, endossimbiontes, rede molecular GNPS, produtos naturais, péptidos, biossíntese.

Abstract

The pharmacological and biotechnological industries have been focusing on the application and research of natural products isolated from a variety of organisms. One phylum of organisms that increasingly gain traction in natural products research is cyanobacteria, a group of photosynthetic bacteria spread widely throughout Earth. The adaptation to varied environments and adverse conditions in this phylum makes the secondary metabolite production/isolation a promising area of research. Many natural products isolated from cyanobacteria are used in a diversity of biological applications and the focus is now directed to their assembly lines and leading mechanisms responsible for the production. Understanding the biosynthetic machinery and the chemical ecology behind the production of cyanobacterial natural products will enhance their biotechnological applications. So, the purpose of this project is the isolation of new natural products from a selected group of cyanobacterial strains, isolated from a wide range of Portuguese environments and that are part of the LEGE culture collection (LEGEcc), established in CIIMAR headquarters. An initial bioactivity and metabolomic screening highlighted the most promising cyanobacterial strains, and the selection of two symbiotic cyanobacteria led to the isolation of new natural products. First, the discovery of a new class of lipoglycopeptides, desmamides A-C from the nostoclean cyanobacterium *Desmonostoc muscorum* LEGE 12446, a cyanobiont isolated from a cycad (*Cycas revoluta*) coralloid root. The chemical structures of the compounds were elucidated using a combination of 1D and 2D NMR spectroscopy and mass spectrometry. Desmamides are decapeptides featuring a *O*-glycosylation of tyrosine and an unusual 3,5-dihydroxy-2-methyldecanoic acid residue. The biosynthesis of these natural products was studied using a combination of bioinformatics and incubation experiments, leading to the proposed *dsm* biosynthetic gene cluster. The antibiotic activity of desmamide B towards the black-rot disease-causing plant pathogen *Xanthomonas campestris,* together with abundant production of these class of compounds by the cyanobacterium (>1.5% d.w.), paves the way for exploration of their potential role in root endosymbiosis. Exploration of the second cyanobacterium *Nostoc* sp. LEGE 12450, isolated from an *Encephalartos horridus* cycad coralloid root, led to the isolation of three new natural products. The elucidation process was performed using a combination of NMR and mass spectrometry. Nostopeptolide L5 is a new lipopeptide, part of the nostopeptolide compound family, composed by eight amino acids and a 2-methylhex-2-enoic acid residue. Although the biosynthesis of nostopeptolides was already reported, the synthesis of the unusual fatty acyl moiety was proposed here for

the first time. Suomilides 1077 and 1007, belong to the aeruginosin compound family, and are new glycosylated and sulfonated tetrapetides. The *suo* biosynthetic gene cluster was already described for suomilide, being also present in the genome of *Nostoc* sp. LEGE 12450. Biological profiling of nostopeptolide L5 resulted in an EC_{50} of 19.4 μM against the settlement of *Mytilus galloprovincialis* larvae, indicating the biotechnological potential of this natural product as an antifouling agent. Nostopeptolides are known players of the endosymbiotic process of cyanobacteria, so the discovery of suomilides may provide better understanding of the chemical cues responsible for this ancient partnership between cycads and cyanobacteria. Exploration of the Portuguese niche-specific cyanobacteria showed the wide and unexplored diversity of this phylum in terms of natural products discovery and their potential biotechnological applications.

Keywords: Cyanobacteria, Endosymbionts, GNPS molecular networking, natural products, peptides, biosynthesis.

Table of Contents

Tables List xiii

Figures List xvii

Abbreviations List xxv

1. General Introduction 3

1.1. Cyanobacteria 3

1.1.1. Cyanobacterial symbiosis 7

1.2. Natural Products 11

1.2.1. Metabolome mining 13

1.2.2. Genome mining 15

1.2.3. Natural products biosynthesis 17

1.2.4. Other methodologies for natural products discovery 21

1.3. Cyanobacterial Natural Products 22

1.4. Objectives 29

1.5. References 30

2. Metabolomic and bioactivity screening of cyanobacteria from LEGE culture collection 39

2.1. Introduction 39

2.2. Results and discussion 40

2.2.1. Unexplored cyanobacterial LEGEcc strains (UC) 41

2.2.2. Explored cyanobacterial LEGEcc strains (EC) 57

2.3. Material and Methods 62

2.3.1. Cyanobacterial strain cultivation 62

2.3.2. Cyanobacterial extraction and fractionation 62

2.3.3. Anticancer assays 63

2.4. Conclusions 65

2.5. Supplementary data 66

2.6. References 82

3. Structure and biosynthesis of desmamides A-C, lipoglycopeptides from the endophytic cyanobacterium *Desmonostoc muscorum* LEGE 12446 87
3.1. Introduction 87
3.2. Results and discussion 89
3.2.1. MS-guided isolation of desmamides A-C (**1**-**3**) 89
3.2.2. Structure elucidation of **1**-**3** 90
3.2.3. Identification and analysis of the *dsm* biosynthetic gene cluster 96
3.2.4. Biological activity 99
3.3. Conclusion 99
3.4. Materials and methods 100
3.4.1. General experimental procedures 100
3.4.2. Cyanobacterial strain cultivation 101
3.4.3. Molecular Networking analysis 101
3.4.4. Extraction and isolation of desmamides 102
3.4.5. HRESIMS/MS analysis 104
3.4.6. Quantification of desmamides in *D. muscorum* LEGE 12446 104
3.4.7. Chiral-phase HPLC 105
3.4.8. Marfey's derivatization analysis 105
3.4.9. Incubation experiments with perdeuterated fatty acids 106
3.4.10. Monossaccharide derivatization analysis with PMP 106
3.4.11. Genome sequencing and bioinformatics analysis 107
3.4.12. Bioactivity profiling 108
3.5. Supplementary information 109
3.6. References 148
4. Secondary metabolite biosynthesis in the endophytic cyanobacterium *Nostoc* sp. LEGE 12450 155
4.1. Introduction 155
4.2. Results and discussion 157

4.2.1. MS-guided isolation of Nostopeptolide L5 (**1**) & Suomilides 1077 and 1007 (**2**,**3**) 158
4.2.2. Structure elucidation of **1** 158
4.2.3. Structure elucidation of **2** and **3** 163
4.2.4. Identification and analysis of the biosynthetic gene cluster of **1** and **2**-**3** ... 168
4.2.5. Biological activity 172
4.3. Material and Methods 173
4.3.1. General experimental procedures 173
4.3.2. Cyanobacterial strain cultivation 174
4.3.3. Molecular Networking analysis 174
4.3.4. Extraction and isolation of **1**-**3** 175
4.3.5. HRESIMS/MS analysis 176
4.3.6. Quantification of **1**-**3** in *Nostoc* sp. LEGE 12450 176
4.3.7. Marfey's derivatization analysis 177
4.3.8. Genome sequencing and bioinformatics analysis 178
4.3.9. Bioactivity profiling 178
4.4. Conclusions 180
4.5. Supplementary material 182
4.6. References 220
5. General Conclusion 229

Tables List

Chapter II – Metabolomic and bioactivity screening of cyanobacteria from LEGE culture collection

Table 1 – LEGEcc cyanobacterial strains selected for the screening assay. The order was updated according to the latest phylogenetic classification[23]. EC – explored cyanobacteria; UC – unexplored cyanobacteria; *metabolomic data. 41

Table 2 – Main information of the mass features present in the molecular networking analysis from the UC Nostocales strains. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and ***** corresponds to adducts not found in the experimental spectra. [a]mass annotated with lipid maps. 50

Table 3 - Main information of the mass features present in the molecular networking analysis from the UC Nodosilineales strains. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra. 55

Table 4 - Main information of the mass features present in the molecular networking analysis from the UC Pleurocapsales, Synechococcales and Leptolyngbyales strains. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra. 55

Table 5 - Main information of the mass features present in the molecular networking analysis from the EC Oculatellates strain *Calenema singularis* LEGE 06188. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra. 59

Table 6 - Main information of the mass features present in the molecular networking analysis from the EC Synechoccales strain unidentified filamentous Synechococcales LEGE 06118 and Chroococcidiopsales *Chroococcidiopsis* sp. LEGE 06174. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra. 60

Table 7 - Main information of the mass features present in the molecular networking analysis from the EC Oscillatoriales strain cf. *Oxynema acuminatum* sp. LEGE 06072. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+ (data not shown). In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra. 61

Table S1 – Cyanobacterial information of the LEGEcc strains selected for this project. *Order classification updated to the last phylogenetic publication[23]. 67

Table S2 – Cyanobacterial information regarding the extraction and fractionation of the strains present in the UC group. *after de-salting using C18 SPE cartridges; **might have presence of salts from the culture media. 68

Table S3 – Annotation of the dereplication hits of the UC fractions C and B, from *Desmonostoc muscorum* LEGE 12446. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 73

Table S4 - Annotation of the dereplication hits of the UC fraction C from *Nostoc* sp. LEGE 12450. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 74

Table S5 - Annotation of the dereplication hits of the UC fraction C from *Nostoc* sp. LEGE 12451. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 75

Table S6 - Annotation of the dereplication hits of the UC fractions C, B and A from *Nostoc* sp. LEGE 12454. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 75

Table S7 - Annotation of the dereplication hits of the UC fraction C from *Nostoc* sp. LEGE 12456. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 76

Table S8 - Annotation of the dereplication hits of the UC fractions B and C from unidentified Nostocales LEGE 14445. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 77

Table S9 - Annotation of the dereplication hits of the UC fractions A and B from *Nodosilinea* sp. LEGE 11468. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 77

Table S10 - Annotation of the dereplication hits of the UC fractions A and B from *Cymatolege* sp LEGE 11434. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 78

Table S11 - Annotation of the dereplication hits of the UC fractions A and B from *Myxosarcina* sp. LEGE 06146 (Pleurocapsales). Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 78

Table S12 - Annotation of the dereplication hits of the UC fractions A and B from *Romeriopsis navalis* LEGE 11480 (Lepotlyngbyales). Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 79

Table S13 - Annotation of the dereplication hits of the UC fractions A and B from cf. *Phormidesmis* sp. LEGE 11477 (Synechocalles). Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 79

Table S14 - Annotation of the dereplication hits of the EC fractions A and B from *Calenema singularis* LEGE 06188. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 81

Table S15 - Annotation of the dereplication hits of the EC fractions A and B from unidentified filamentous *Synechococcales* LEGE 06118. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's. 81

Chapter III - Structure and biosynthesis of desmamides A-C, lipoglycopeptides from the endophytic cyanobacterium *Desmonostoc muscorum* LEGE 12446

Table 1 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for desmamide A (**1**) in DMSO-d_6. 92

Table 2 – Desmamide analogues detected during isolation. 96

Table S1 - Mass features from the molecular networking analysis obtained using LC-HRESIMS/MS data for fractions A, B and C of the organic extracts of LEGE 12446. The more intense masses in fraction C are shown in bold. 109

Table S2 – LC-HRESIMS data for desmamides A-C (**1**-**3**) standards used for quantification of the compounds in the cyanobacterial extract/biomass. Also shown are the data in the organic extract. 111

Table S3 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for desmamide B (**2**) in DMSO-*d*6. 122

Table S4 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for desmamide C (**3**) in DMSO-d_6. 130

Table S5 - Annotation of the *dsm* biosynthetic gene cluster. 142

Table S6 - Results of tBlastn searches for homologs of cacaoidin BGC glycosyltransferases encoded by cao8, cao16 and cao24, against the *D. muscorum* LEGE 12446 genome. Shown are hits with an e-value < 1×10^{-20}. No hits were obtained for cao8 that satisfied this condition. 143

Table S7 - Biological activity of compounds **1** and **2**. IC_{50} values are in µM. 143

Chapter IV - Secondary metabolite biosynthesis in the endophytic cyanobacterium *Nostoc* sp. LEGE 12450

Table 1 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for nostopeptolide L5 (**1**) in DMSO-d_6. 160

Table 2 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for suomilide 1077 (**2**) in DMSO-d_6. 165

Table S1 - Mass features from the molecular networking analysis obtained using LC-HRESIMS/MS data for fractions C of the organic extracts of *Nostoc* sp. LEGE 12450. The more intense masses in the crude extract are shown in bold. 182

Table S2 – LC-HRESIMS data for Nostopeptolide L5 (**1**) and suomilides 1077/1007 (**2**,**3**) standards used for quantification of the compounds in the cyanobacterial extract/biomass. Also shown are the data in the organic extract. 184

Table S3 - ^{1}H and ^{13}C NMR Data comparison of nostopeptolide L5 (**1**) and nostopeptolide L1 in DMSO-*d*6. 194

Table S4 - Comparison of the NMR data of Suomilide and **2** and **3**, in DMSO-d_6. In bold are the signals assigned based on comparison with suomilide NMR data. Signals in red are only present in the suomilide structure. 204

Table S5 - Annotation of the *nos* biosynthetic gene cluster. 213

Table S6 - Annotation of the *suo* biosynthetic gene cluster. Proteins at bold are new, comparing to the *suo* BGC of *Nodularia sphaerocarpa* UHCC 0038 (Orf28-31). 214

Table S7 - Anticancer assays using the hepatocellular (HepG2) and colon (HCT 116) carcinoma cell lines in monolayer cultures (2D) and with HCT 116 cell line in 3D cultures. The compounds **1** and **2** were tested at a concentration of 10 *µ*M. 217

Figures List

Chapter I - General Introduction

Figure 1 – Examples of the cyanobacterial morphological diversity, represented by eight strains of the LEGE culture collection (LEGEcc). **(a)** unidentified filamentous cyanobacterium LEGE 11480; **(b)** unidentified Pleurocapsales LEGE 06147; **(c)** *Nostoc* cf. *edaphicum* LEGE 07299; **(d)** *Nostoc* sp. LEGE 12456; **(e)** unidentified Nostocales LEGE 06150; **(f)** unidentified Nostocales LEGE 11386; **(g)** cf. *Phormidesmis* sp. LEGE 11434; **(h)** *Chroococcidiopsis* sp. LEGE 06174.....4

Figure 2- Representation of the infection process of a cycad plant by a cyanobacterium to form the specialized coralloid roots. **(1)** release of hormogonium inducing factors (HIFs) by the plant; **(2)** cyanobacteria vegetative cells differentiate into hormogonia; **(3)** motile hormogonia filaments are directed to the cycad; **(4)** infection of the pre-coralloid root; **(5)** hormogonia filaments enter the root; **(6)** production of hormogonia repression factors (HRFs); **(7)** cyanobacteria differentiate to vegetative cells and heterocysts for nitrogen-fixation; **(8)** cyanobacteria form a layer around the coralloid root, cyanobacterial zone, being in permanent symbiosis with the cycad plant.....9

Figure 3 – Examples natural products biosynthetic pathways of polyketide origin. **(a)** Type I modular PKS biosynthetic pathway of erythromycin A: the starter unit is Propionyl-CoA, picked up by the loading module (LM) and followed by the incorporation of six extender units of (2S)-Methylmalonyl-CoA by modules 1-6; further biosynthetic steps of the cyclization and addition of the monosaccharide decorations are not represented. **(b)** Type I iterative PKS biosynthetic pathway of lovastatin: the starter unit is Acetyl-CoA (presumably it was a Malonyl-CoA that was further descarboxilated) and elongation is performed iteratively by eight units of Malonyl-CoA to form dehydromonacolin L (not shown) that is then modified to give lovastatin; LovC is a trans-acting enoyl reductase and besides C-C Claisen condensations, a Diels-Alder reaction also occurs. **(c)** Type II PKS biosynthetic pathway of doxorubicin: the starter unit is Propionyl-CoA, recognized specifically by DspC and DspD, and elongation is performed with nine units of Malonyl-CoA by DpsA and DpsB; dpsABCDG is considered the minimal PKS for this biosynthetic pathway, and several other enzymes involved are not represented here. **(d)** Type III PKS biosynthetic pathway of cylindrocyclophane A: the pathway starts with a modification of decanoinc acid (started unit) to form decanoyl-CylB thioester, as part of the fatty acid metabolism; the thioester is picked up by Type I mPKS enzymes and extended with two malonate units and the type III PKS CylI, using two malonyl-CoA unit's origins a monomeric precursor and other enzymatic reactions create cylindrocyclophane A; ACP* may participate in β-methyl formation and TE** may be inactive. AT – acyltransferase; KS – ketosynthase; ACP – acyl carrier protein; KR – keto reductase; DH – dehydratase; ER – enoyl reductase; ER^0 – inactive enoyl reductase; TE – thioesterase; cMT – carbon methyltransferase; CYC – cyclase; Ox – oxygenase; oMT – oxygen methyltransferase; ECH – enoyl-CoA hydratase. Each arrow is correspondent to one biosynthetic step. 19

Figure 4 - Non-ribosomal peptide biosynthetic pathway of Nostocyclopeptide A2. NcpA comprises three modules incorporating the amino acids tyrosine, glycine, and glutamine, followed by the enzyme NcpB with four modules in which the adenylation domain activates isoleucine, serine, 4-methyl-proline and phenylalanine. The peptide is cleaved from the PCP by domain R, a NADPH dependent reductase, and the resulting reactive aldehyde triggers the self-assembly of the macrocyclic imine. C – condensation domain; A – adenylation domain; PCP - peptidyl carrier protein; R – reductase domain. Each arrow is correspondent to one biosynthetic step. 20

Figure 5 - Selection of cyanobacterial natural products, representing the wide chemodiversity of the phylum. Characteristic structural features are highlighted in gray. 24

Chapter II - Metabolomic and bioactivity screening of cyanobacteria from LEGE culture collection

Figure 1 – Cell viability assays with the breast carcinoma cell line, T47D, showing the decrease in cell viability of >60% of the cyanobacterial fraction A from *Nodosilinea* sp. LEGE 11468, fractions A and B from unidentified Nostocales LEGE 11386, fractions A and B from unidentified Nostocales LEGE 14445 and fractions A and B from unidentified Pleurocapsales LEGE 06147. All fractions are at a concentration of 30 μg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 μg mL^{-1}). 42

Figure 2 - Cell viability assays with the neuroblastoma cell line, SYSH-S5, showing the decrease in cell viability of >60% of the cyanobacterial fraction A from unidentified Nostocales LEGE 14445, fractions A and B from unidentified Pleurocapsales LEGE 06147, fractions A and B from *Nostoc* sp. LEGE 12447 and the A

fraction of *Nostoc* sp. LEGE 12451, *Nostoc* sp. LEGE 12454, *Nostoc* sp. LEGE 12456 and *Nostoc* sp. LEGE 12450. All fractions are at a concentration of 30 µg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 µg mL^{-1}). 43

Figure 3 - Cell viability assays with the osteosarcoma cell line, MG-63, showing the decrease in cell viability of >60% of the cyanobacterial fraction A from unidentified Nostocales LEGE 14445 and fraction A from unidentified Pleurocapsales LEGE 06147. All fractions are at a concentration of 30 µg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 µg mL^{-1}). 43

Figure 4 – Antimicrobial assays performed with the disc diffusion method. **(A)** Inhibition halos of *S. aureus* with fractions B and C from unidentified Pleurocapsales LEGE 06147; **(B)** Inhibition halos of *S. aureus* with fractions B and C from *Nostoc* sp. LEGE 12447**; (C)** Inhibition halos of *B. subtilis* by fractions B and C from *Desmonostoc muscorum* LEGE 12446; **(D)** Inhibition halos of *S. typhimurium* by fractions A and B from *Nostoc* sp. LEGE 12450; **(E)** Inhibition halo of *E. coli* by fraction C of *Cymatolege* sp. LEGE 11434. Exposure of the bacterial strains with 15 µg of each fraction; Negative control corresponds to 100% DMSO. 44

Figure 5 – Main molecular clusters from *Desmonostoc muscorum* LEGE 12446, retrieved from the GNPS molecular networking analysis. Annotation of anabaenopeptin 891 and 905 (and respective adducts) with NPAtlas, of a glycerol lipid using Dictionary of Natural Products (DNP) and a spectral match to Pro-Ile from NIST14 (NIST standard reference database).....46

Figure 6 - Main molecular clusters from *Nostoc* sp. LEGE 12450, retrieved from the GNPS molecular networking analysis.....47

Figure 7 - Main molecular clusters from *Nostoc* sp. LEGE 12451, retrieved from the GNPS molecular networking analysis.....47

Figure 8 - Main molecular clusters from *Nostoc* sp. LEGE 12454, retrieved from the GNPS molecular networking analysis. The first cluster has a connection with masses also present in *Cymatolege* sp. LEGE 11434 (green), and with a mass feature present in several other UC strains (grey), which were not considered. 48

Figure 9 - Main molecular clusters from *Nostoc* sp. LEGE 12456, retrieved from the GNPS molecular networking analysis. Micropeptin C and D were identified with the dereplication tools (yellow) and Cyanopeptolin CP999 (blue). Related masses to micropeptins and/or cyanopeptolins probably indicate the presence of new analogues (green and orange).....49

Figure 10 – MS^2 spectra of the identified micropeptin C and cyanopeptolin CP999, with possible new analogues present in the UC *Nostoc* sp. LEGE 12456. 49

Figure 11 - Main molecular clusters from *Cymatolege* sp. LEGE 11434 retrieved from the GNPS molecular networking analysis.....52

Figure 12 – Main molecular clusters from *Romeriopsis navalis* LEGE 11480 retrieved from the GNPS molecular networking analysis. The first clusters (top left) is related with several other masses shared by other UC strains that were not considered for analysis. 53

Figure 13 - Main molecular clusters from cf. *Phormidesmis* sp. LEGE 11477 retrieved from the GNPS molecular networking analysis. 54

Figure 14 - Main molecular clusters from *Calenema singularis* LEGE 06188, retrieved from the GNPS molecular networking analysis. Annotation of oxooctacosanoic acid and glycerol 1,2-dialkanoates with Dictionary of Natural Products (DNP, yellow label). The grey mass features are shared by several EC strains. 58

Figure 15 - Main molecular clusters from unidentified filamentous Synechococcales LEGE 06118, and *Chroococcidiopsis* sp. LEGE 06174, retrieved from the GNPS molecular networking analysis. The considered masses (according to the defined intensity) are represented in green and blue, respectively. .58

Figure 16 - Main molecular clusters from cf. *Oxynema acuminatum* LEGE 06072, retrieved from the GNPS molecular networking analysis. The considered masses (according to the defined intensity) are represented in yellow. Only four of the five mass features are shown. 61

Figure S1 - Maximum-likelihood phylogenetic tree of the LEGEcc strains together with some reference strains, highlighting the 14 selected strains based on phylogenetic placement and uniqueness (red star). Collapsed clades are unlabeled and bootstrap values are omitted for clarity. Strain *Romeriopsis navalis* LEGE 11480 is not depicted in this phylogenetic tree, however it has only 91% identity to the cultured strain corresponding to the top BLAST hit (*Synechococcus* sp. UTEX 2973).....66

Figure S2 – Cytotoxic assays with the cell lines A549 (lung carcinoma), RKO (colon carcinoma) and T47D (breast carcinoma) with the fractions A-C from all the 14 cyanobacterial strains belonging to the UC group. All fractions are at a concentration of 30 μg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 μg mL^{-1}). 69

Figure S3 - Cytotoxic assays with the cell lines HepG2 (liver carcinoma), MG-63 (osteosarcoma) and SHSY-S5 (neuroblastoma) with the fractions A-C from all the 14 cyanobacterial strains belonging to the UC group. All fractions are at a concentration of 30 μg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 μg mL^{-1}). 70

Figure S4 - Antimicrobial assays performed with the disc diffusion method. **(A)** Inhibition halos of *S. aureus* with fractions B and C from unidentified Pleurocapsales LEGE 06147; **(B)** Inhibition halos of *S. aureus* with fractions B and C from *Nostoc* sp. LEGE 12447**; (C)** Inhibition halos of *B. subtilis* by fractions B and C from *Desmonostoc muscorum* LEGE 12446; **(D)** Inhibition halos of *S. typhimurium* by fractions A and B from *Nostoc* sp. LEGE 12450; **(E)** Inhibition halo of *E. coli* by fraction C of *Cymatolege* sp. LEGE 11434. Exposure of the bacterial strains with 15 μg of each fraction; Negative control corresponds to 100% DMSO. 71

Figure S5 – GNPS Molecular Networking analysis, full network visualized using Cytoscape (v. 3.8.0.). The UC strains were divided into five groups according to phylogenetic distribution: purple (Nostocales symbiotic strains), orange (Non-symbiotic Nostocales strains), lavender (shared mass features between both symbiotic and non-symbiotic Nostocales groups), green (Nodosilineales strains), pink (Pleurocapsales strains), blue (Synechococcales and Leptolyngbyales) and grey (blanks, plus an actinobacteria extract, *Streptomyces* sp. KENR25). Label information set for parent mass. Spectral library hits are triangle shaped and in case of known cyanobacterial metabolites (mainly pigments), diamond shape. Manual inspection of single-strain networks was conducted, focusing on networks composed of a single color. 72

Figure S6 - GNPS Molecular Networking analysis, full network visualized using Cytoscape (v. 3.8.0.). The EC strains were divided into five groups according to phylogenetic distribution: purple (Synechococcales), blue (Oscillatoriales_also including Coleofasciculales since the strains were collected from the same environment), green (Oculatellates), orange (Nostocales), pink (Chroococcidiopsidales and Leptolyngbyales) and grey (blanks, plus an actinobacteria extract, *Streptomyces* sp. KENR25). Label information set for parent mass. Spectral library hits are triangle shaped and in case of known cyanobacterial metabolites (mainly pigments), diamond shape. Manual inspection of single-strain networks was conducted, focusing on networks composed of a single color. 80

Chapter III - Structure and biosynthesis of desmamides A-C, lipoglycopeptides from the endophytic cyanobacterium *Desmonostoc muscorum* LEGE 12446

Figure 1 . Cyanobacterial lipoglycopeptides, hassalidin A and balticidin B. Sugar moieties – red; fatty acid chain – light blue; ring closure (ester bond) – green shade. 88

Figure 2 - Structure of desmamides A-C (**1**-**3**). Desmamide A is represented with both monosaccharide units, α-xylopyranosyl and β-arabinopyranosyl. 90

Figure 3 - Structure elucidation of **1**-**3**. A – Key HMBC, COSY and NOESY correlations and HRMS/MS fragmentations supporting the structure proposal for **1**. B – Key HMBC and COSY correlations supporting the presence of a 4-*O*-acetylxylopyranosyl-Tyr moiety. C – Comparison of the HRESIMS/MS spectra of the aglycone of **1** (isolated in-source) and of compound **3**, supporting that they are identical. 94

Figure 4 - Biosynthesis of the desmamides A-C (**1**-**3**). A – Schematic representation of the *dsm* BGC, including the proposed functions of the predicted open-reading frames, domain composition of detected PKS and NRPS (or NRPS-like) modules, and bioinformatically-predicted (antiSMASH) monomers activated by acyl-tranferase (AT) or adenylation (A) domains. A* indicates an adenylation domain within a region with homology to a fatty acyl-AMP ligase (FAAL), proposed to activate hexanoic acid. B – Proposed biosynthetic steps leading to **1**-**3**, highlighting the formation of the Dhmda residue from hexanoic acid, by PKS machinery. C – LC-HRESIMS analysis of methanolic extracts of *D. muscorum* LEGE 12446 cultures supplemented with perdeuterated fatty acids. Shown are EICs for *m/z* values corresponding to $[M+H]^+$ ions for compound **1** and for hypothesized isotopologues resulting from incorporation of each of the perdeuterated fatty acids. Also shown is Full MS data, highlighting the mass shift supporting incorporation of d_{11}-C_6 into **1**. 98

Figure S1 - GNPS molecular networking nodes present only in fraction C (eluting with MeOH), obtained from an organic extract of *D. muscorum* LEGE 12446 biomass. The mass features highlighted in red guided the isolation procedures. 109

Figure S2 - Fractionation scheme leading to the purification of **1**-**3.** 110

Figure S3 - LC-HRESIMS (positive mode) calibration curves (using standards of **1**-**3**) used for estimating the amount of desmamides in the organic extract used for isolation. 111

Figure S4 - HRESIMS spectrum of **1**. 112

Figure S5 - ^{1}H NMR (DMSO-d_6, 400 MHz) spectrum of compound **1**. 113

Figure S6 - ^{13}C NMR (APT, DMSO-d_6, 100 MHz) spectrum of compound **1**. 114

Figure S7 - HSQC (DMSO-d_6, 400 MHz) spectrum of compound **1**. 115

Figure S8 - HMBC (DMSO-d_6, 400 MHz) spectrum of compound **1**. 116

Figure S9 - COSY (DMSO-d_6, 400 MHz) spectrum of compound **1**. 117

Figure S10 - NOESY (DMSO-d_6, 400 MHz) spectrum of compound **1**. 118

Figure S11 - Band-selective HMBC (DMSO-d_6, 600 MHz) spectrum of compound **1**. 119

Figure S12 - LC-HRESIMS analysis of PMP-sugar derivatives to determine the sugar composition of **1**. Shown are EICs for each of the sugar standard derivatives and for the derivatized hydrolysate product of desmamide A (**1**). 120

Figure S13 - HRESIMS/MS spectrum of **1**, highlighting key fragmentations supporting the NMR-based residue sequence proposal. 120

Figure S14 - Annotated HRESIMS/MS spectrum for **3**, highlighting fragmentations leading to amino acid sequence confirmation. Three regions of the same spectrum are shown (A, B and C). 121

Figure S15 - HRESIMS spectrum of **2**. 124

Figure S16 - ^{1}H NMR (DMSO-d_6, 100 MHz) spectrum of compound **2**. 125

Figure S17 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **2**. 126

Figure S18 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **2**. 127

Figure S19 - HMBC (DMSO-d_6, 600 MHz) spectrum of compound **2**. 128

Figure S20 - COSY (DMSO-d_6, 600 MHz) spectrum of compound **2**. 129

Figure S21 - Coupling constant analysis (^{1}H and COSY data, DMSO-d_6, 600 MHz) underscoring the proposed sugar configuration of compound **2**. 130

Figure S22 - ^{1}H NMR (DMSO-d_6, 150 MHz) spectrum of compound **3**. 133

Figure S23 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **3**. 134

Figure S24 - Comparison of the HRESIMS/MS data for **3**, with those of the in-source-isolated aglycones of **1** and **2**, supporting that these are identical. 135

Figure S25 - Chiral-phase HPLC analysis of the amino acid residue composition of **1**. 136

Figure S26 - LC-HRESIMS analysis of the Marfey's derivatization products. EICs of $[M+H]^+$ ions for leucine (*m/z* 384.15), serine (*m/z* 358.09), glutamic acid (*m/z* 400.11), threonine and allo-threonine (*m/z* 372.11) are shown for the standards and hydrolysate. 137

Figure S27 - Examples of natural lipoglycopeptides of different structural classes. 138

Figure S28 - Comparison of the HRESIMS/MS data for *m/z* 1344.730, with those of the in-source aglycones of *m/z* 1212.684, supporting that these are identical. The structural difference is on the fatty acyl chain, having a hexanoic acid moiety in comparison to Dhmda of **1** and **3**. 139

Figure S29 - Comparison of the HRESIMS/MS data for *m/z* 1444.782, with those of the in-source aglycones of *m/z* 1402.774 and *m/z* 1270.729, supporting that these are identical. The structural difference is on amino acid substitution of Val for Leu/Ile (Ile is represented). 139

Figure S30 - HRESIMS/MS data for *m/z* 1360.723, showing the structural difference on the fatty acyl chain, having an 3,5-hydroxy-methylbutyric acid instead of the Dhmda of **1**. Fragmentation of the aglycone *m/z* 1228.683 was not possible due to the low intensity of the mass peak. 140

Figure S31 - HRESIMS/MS data for *m/z* 1416.788, showing the structural difference on the fatty acyl chain, having an 3,5-hydroxy-methyloctanoic acid instead of the Dhmda of **1**. Fragmentation of the aglycone *m/z* 1284.742 was not possible due to the low intensity of the mass peak. 140

Figure S32 - HRESIMS/MS data for *m/z* 1374.743, showing the structural difference on the fatty acyl chain, having an 3,5-hydroxy-methylheptanoic acid instead of the Dhmda of **1**. Fragmentation of the aglycone *m/z* 1242.699 was not possible due to the low intensity of the mass peak. 140

Figure S33 - HRESIMS/MS data for *m/z* 1370.748, showing the structural difference on the fatty acyl chain, without the hydroxyl group on position 5, in comparison with **1**. Fragmentation of the aglycone *m/z* 1238.664 was not possible due to the low intensity of the mass peak. 141

Figure S34 - HRESIMS/MS data for *m/z* 1042.580, showing the structural difference on the core of amino acids, with only seven amino acids, not having one unit of Leu and Thr, in comparison with **3**. 141

Figure S35 - HRESIMS/MS data for *m/z* 1198.675, showing the structural difference on the fatty acyl chain, having 3-hydroxy-2,5-methylhexanoic acid instead of Dhmda, in comparison with **3**. 141

Figure S36 - HRESIMS/MS spectrum of compound **2** ($[M+H]^+$). 143

Figure S37 - HRESIMS/MS spectrum of compound **3** ($[M+H]^+$). 144

Figure S38 - FT-IR spectrum of compound **1**. 144

Figure S39 - UV-vis spectrum of compound **1** 145

Figure S40 - FI-IR spectrum of compound **2**. 145

Figure S41 - UV-vis spectrum of compound **2** 146

Figure S42 - FT-IR spectrum of compound **3**. 146

Figure S43 - UV-vis spectrum of compound **3** 147

Chapter IV - Secondary metabolite biosynthesis in the endophytic cyanobacterium *Nostoc* sp. LEGE 12450

Figure 1 - Structure elucidation of **1**. (A) Key HMBC, COSY and NOESY correlations and HRMS/MS fragmentations supporting the structure proposal for **1**; (B) Compound **1** structure, evidencing the resolved stereochemistry. 160

Figure 2 - Structure elucidation of **2**. Key HMBC and COSY. Comparison of the suomilide with **2** and **3**, showing the main structural differences in the sugar unit. 167

Figure 3 - Biosynthesis of the nostopeptolide L5 (**1**). **A** – Schematic representation of the *nos* BGC, including the proposed functions of the predicted open-reading frames (Orfs), domain composition of detected PKS and NRPS (or NRPS-like) modules, and bioinformatically-predicted (antiSMASH) monomers activated by acyltranferase (AT) or adenylation (A) domains. **B** – Proposed biosynthetic steps leading to **1**, highlighting the formation of the MHEA residue from butyric acid, by PKS machinery; KS – ketosynthase, DH – dehydrogenase, MT – methyltransferase, KR – ketoreductase, T – PCP; FkbH – methyltransferase; E – epimerase; TE – thioesterase. 169

Figure 4 - Biosynthesis of the suomilides 1077 (**2**) and 1007 (**3**). A – Schematic representation of the *suo* BGC, including the proposed functions of the predicted open-reading frames (orf), domain composition of detected NRPS (or NRPS-like) modules, and the activated monomers activated by adenylation (A) domain, based on the putative *suo* BGC from *Nodularia sphaerocarpa* UHCC 0038. B – Proposed biosynthetic steps leading to **2** and **3**, following the putative *suo* BGC from *Nodularia sphaerocarpa* UHCC 0038. MT – methyltransferase; ST – sulfotransferase. Enzymes in bold are confirmed to be involved in the proposed biosynthetic steps. 172

Figure S1 - GNPS molecular networking nodes present on the methanolic fraction of *Nostoc* sp. LEGE 12450. **(A)** Adducts of nostopeptolide L5; (**B**) Mass features correspondent to the known anabaenopeptins; Sodium adduct of suomilides 1077 and 1007; The mass features highlighted in green were isolated in the present work. 182

Figure S2 - HRESIMS/MS of the anabaenopeptins found in the crude extract of *Nostoc* sp. LEGE 12450, showing the identical fragmentation pattern of three metabolites. 183

Figure S3 - Fractionation scheme leading to the purification of **1**-**3**. 183

Figure S4 - LC-HRESIMS (positive mode) calibration curves (using standards of **1**-**3**) used for estimating the amount of metabolites produced by *Nostoc* sp. LEGE 12450 in the organic extract used for isolation. 184

Figure S5 - HRESIMS spectrum of **1**. 185

Figure S6 - ^{1}H NMR (DMSO-d_6, 400 MHz) spectrum of compound **1**. 186

Figure S7 - ^{13}C NMR (APT, DMSO-d_6, 100 MHz) spectrum of compound **1**. 187

Figure S8 - HSQC (DMSO-d_6, 400 MHz) spectrum of compound **1**. 188

Figure S9 - HMBC (DMSO-d_6, 400 MHz) spectrum of compound **1** 189

Figure S10 - COSY (DMSO-d_6, 400 MHz) spectrum of compound **1**. 190

Figure S11 - NOESY (DMSO-d_6, 400 MHz) spectrum of compound **1** 191

Figure S12 - Band-selective HMBC (DMSO-d_6, 600 MHz) spectrum of compound **1**. 192

Figure S13 - Structures of nostopeptolide L1 isolated from *Nostoc* sp. UK2aImI, having an MHEA moiety connected to eight amino acids [Thr-*O*meSer-PhePr-4mePro-Dhb-Ile-Gln-Hyp]; and nostopeptolide A1, isolated from *Nostoc* sp. GSV224 and *Nostoc punctiforme* PCC 73102, showing the butyric acid unit and the eight amino acids [Thr-Ser-4mePro-LeuAc-Leu-Gly-Asn-Tyr-Pro]. 193

Figure S14 - HRESIMS/MS spectrum of **1**, highlighting key fragmentations supporting the NMR-based residue sequence proposal. 193

Figure S15 - LC-HRESIMS analysis of the Marfey's derivatization products. EICs of $[M+H]^+$ ions for Glu (*m/z* 400.11), Phe (*m/z* 418.14), Ser (*m/z* 358.10), Hyp (*m/z* 384.12), Pro (*m/z* 368.12), Tyr (*m/z* 434.13) Thr and allo-Thr (*m/z* 372.11) are shown for the standards and hydrolysate of nostopeptolide L5. 196

Figure S16 - Comparison of the EIC from the compounds **2** and **3**, showing the loss of 79.95 amu (atomic mass units) for both compounds, prior isolation. 197

Figure S17 - ^{1}H NMR (DMSO-d_6, 100 MHz) spectrum of compound **2**. 198

Figure S18 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **2**. 199

Figure S19 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **2**. 200

Figure S20 - HMBC (DMSO-d_6, 600 MHz) spectrum of compound **2** 201

Figure S21 - COSY (DMSO-d_6, 600 MHz) spectrum of compound **2**. 202

Figure S22 - NPAtlas search of cyanobacterial natural products, using the sulfate and glycosyl group as query. Four compound classes were retrieved and an example of each is represented. The sulfate group is highlighted by a red shade and the glycosyl unit drawn in blue. For aeruginosin 828A, the chlorine atom is highlighted in green as well as the Hpaa units of varlaxin 1022A. 203

Figure S23 - HRESIMS spectrum of **2**. 203

Figure S24 - HRESIMS/MS analysis of **2**. The main fragments are represented in the structure on the right side and are: **A** – loss of the sulphate group and sugar unit (*m/z* 592.309); **B** – loss of the oxazolidin-4-1 ring (*m/z* 59.00); **C** – loss of MgS-Ile moiety (*m/z* 215.12); **D** – loss of the Mgs, fragmenting from Ile (*m/z* 74.04); E – loss of the Aaep moiety (*m/z* 155.13) and $\mathbf{F_1}$ and $\mathbf{F_2}$ corresponding to the fragmentation of the C4 and C6 from the glycosyl unit, respectively. 206

Figure S25 - LC-HRESIMS analysis of the Marfey's derivatization products. EIC of $[M+H]^+$ ion for isoleucine (*m/z* 284.15) is shown for the standards and hydrolysate of **2**. 207

Figure S26 - Coupling constant analysis (^{1}H and COSY data, DMSO-d_6, 600 MHz) underscoring the proposed sugar configuration of compound **2**. 207

Figure S27 - ^{1}H NMR (DMSO-d_6, 150 MHz) spectrum of compound **3**. 208

Figure S28 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **3**. 209

Figure S29 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **3**. 210

Figure S30 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **3**, showing the proton annotation. **(A)** Exchangeable protons, highlighting the OH proton, δ^H 4.94 (d, 7.6 Hz) of position 2 of the sugar unit; **(B)** and **(C)** the remaining resonances of the structure. 211

Figure S31 - Comparison of the MS_2 from the suomilide 1077 (**2**) with soumilide 1007 (**3**), confirming the structural difference in the glycosyl unit connected to the Abn moiety. 212

Figure S32 - Coupling constant analysis (^{1}H and COSY data, DMSO-d_6, 600 MHz) underscoring the proposed sugar configuration of compound **3**. 212

Figure S33 - Alignment of the biosynthetic gene clusters (BGCs) from cyanobacterial strains known to produce nostopeptolides: *Nostoc* sp. GVS 224, *Nostoc punctiforme* PCC 73102 and *Nostoc* sp. LEGE 12450. *Nostoc* sp. XA010 genome was annotated by antiSMASH and aligned as well, since it was the closest homologue in the performed BGC annotation (Table S5). 213

Figure S34 - Alignment of the biosynthetic gene clusters (BGCs) from *Nostoc* sp. LEGE 12450 and with *Nodularia sphaerocarpa* UHCC 0038 from which the *suo* BGC was annotated first. 216

Figure S35 - Alignment of the BGCs with similarity retrieved by using ClusterBlast, the antiSMASH tool. 217

Figure S36 - Antifouling activity of **1**, with the anti-settlement assay using *M. galloprovincialis* larvae and assay performed for the EC_{50} calculations. 217

Figure S37 - FT-IR spectrum of compound **1**. 218

Figure S38 - FI-IR spectrum of compound **2**. 218

Figure S39 - UV-vis spectrum of compound **1** 219

Figure S40 - UV-vis spectrum of compound **2** 219

Abbreviations List

16S rRNA	Small subunit ribosomal RNA
1D	One dimensional
2D	Two dimensional
3D	Three dimensional
A	Adenylation domain
Abn	Azabiocyclononane
ACP	Acyl Carrier Protein
ACS	American Chemical Society
Aeap	1-amidino-3-(2-aminoethyl)-3-pyrroline
AGC	Automatic Gain Control
AMP	Adenosine monophosphate
amu	Atomic Mass Units
AntiSMASH	Antibiotics and secondary metabolite analysis shell
APT	Attatched-proton Test
AT	Acetyltransferase
ATCC	American Type Culture Collection
ATP	Adenosine Triphosphate
BGC	Biosynthetic Gene Cluster
BLAST	Basic Local Aligment Search Tool
BLASTp	Protein BLAST
BWA-MEM	Burrows-Wheeler Aligner for long sequence reads
C	Condensation domain
C18	Reverse-phase Silica
CAGECAT	CompArative GEne Cluster Analysis Toolbox
calcd	Calculated
CEMUP	Materials Center of the University of Porto
CH_2Cl_2	Dichloromethane
$CHCl_3$	Chlorophorm
Choi	2-carboxy-6-hydroxyoctahydroindole
CIIMAR	Centro Interdisciplinar de Investigação Ambiental e Marinha
cMT	Carbon Methyltransferase
CoA	Coenzime A
COSY	Correlated spetroscopy
CYC	Cyclase
d.w.	Dry weight
ddH2O	Deonized water
DH	Dehydratase
Dhma	3,5-dihydroxy-2-methyldecanoic acid
DMEM	Dulbecco's Modified Eagle Medium
DMIS	Direct Difussion Mode
DMSO	Dimethyl Sulfoxide
DMSO-*d*6	Deuterated Dimethyl Sulfoxide
DNA	Deoxyribonucleic acid
DNP	Dictionary of Natural Products
E	Epimerase
EC	Explored cyanobacteria

EC_{50}	Half Maximal Effective Concentration
ECH	Enoyl-coenzyme A hydratase
EIC	Extracted Ion Chromatogram
EPS	Exopolysaccharides
ER	Enoyl Reductase
EtOAc	Ethyl Acetate
FA	Fatty Acid
FAAL	Fatty acyl-AMP ligase
FCT	Fundação para a Ciência e Tecnologia
FCUP	Faculdade de Ciências da Universidade do Porto
FDR	False Discovery Rate
Fr	Fraction
Gb	Gigabase
gDNA	Genomic DNA
GEOMAR	Helmholtz Centre for Ocean Research Kiel
GNAT	Gcn5-Related N-Acetyltransferases
GNPS	Global Natural Products Social Molecular Networking
GOE	Great Oxidation Event
GT	Glycosyltransferase
Ham'F12	F-12 Nutrient medium
HCD	Higher-energy C-trap dissociation
HCl	Hydrochloric acid
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HESI	Heated Electrospray Ionization
HGT	Horizontal Gene Transfer
HIF	Hormogonia Inducing Factors
HMBC	Heteronuclear Multipe Bond Correlation
Hpaa	4-hydroxyphenylacetic acid
Hpg	4-hydroxyphenylglycine
Hpla	Hydroxy-phenyl lactic acid
HPLC	High Pressure Liquid Chromatography
HRF	Hormogonia Repressing Factors
HRMS	High Resolution Mass Spectrometry
HRMS/MS	High Resolution Tandem Mass Spectrometry
HSCoA	Coenzyme A
HSQC	Heteronuclear Single Quantum Coherence
IF	Infrared
IPA	Isopropanol
Kb	Kilo base
KR	Ketoreductase
KS	Ketosynthase
LC	Liquid Chromatography
LC-HRESIMS	Liquid Chromatography High Resolution Electrospray Ionization Mass Spectrometry
LC-HRESIMS/MS	Liquid Chromatography High Resolution Electrospray Ionization Tandem Mass Spectrometry
LC-MS	Liquid Chromatrography Mass Spectrometry
LEGEcc	LEGE culture collection

LM	Loading Module
m/z	Mass to charge ratio
M1-6	Module 1-6
MAAs	Mycosporin Amino Acids
Mal-Coa	Malonyl coenzyme A
MBOAT	Membrane Bound O-acyl Transferase
Mbp	Mega base pairs
MeCN	Acetonitrile
MEM	Minimum Essential Media
MeOH	Methanol
MF	Molecular Formula
MgS	2-O-methylglyceric acid 3-O-sulfate
MHEA	2-methylhex-2-enoic acid
MIBiG	Minimum Information about a Biosynthetic Gene Cluster
MS	Mass Spectrometry
MS1	Full mass spectra
MS2	Tandem Mass Spectrometry
MSI	Mass Spectrometry Imaging
MT	Methyltransferase
MTT	3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide
NGS	Next-Generation Sequencing
n-hex	Hexane
NIST	Mass spectral library
NMR	Nuclear Magnetic Ressonance
NOE	Nuclear Overhauser Effect
NOESY	Nuclear Overhauser Effect Spectroscopy
NP	Natural Products
NPAtlas	Natural Products Atlas
NRP	Non-ribosomal Peptides
NRPS	Non-ribosomal Peptide Syntethase
oMT	Oxygen Methyltransferase
Orf	Open Reading Frames
OSMAC	One Strain Many Compounds
Ox	Oxygenase
PBS	Phosphate-Buffered Saline
PCP	Peptidyl Carrier Protein
Pen-Strep	Penicillin-streptomycin
PK	Polyketides
PKS	Polyketide synthase
PMP	1-phenyl-3-methyl-5-pyrazolone
PP	4'-phosphopantetheine
PSII	Photosystem II
R	Reductase domain
RiPP	Ribosomal Synthesized and Post-Translational modified Peptides
RMN	Ressonância Magnética Nuclear
RNA	Ribonucleic acid
RPMI	Roswell Park Memorial Institute

RuBisCO	Ribulose-1,5-bisphosphate carboxylase/oxygenase
SAM	S-Adenosyl methionine
SPAdes	St. Petersburg genome assembler
SPE	Solid Phase Extraction
ST	Sulfotransferase
tblastn	Translated BLAST
TE	Thioesterase
TIC	Total Ion Chromatogram
t_R	Retention time
TSB	Tryptic Soy Broth
UC	Unexplored Cyanobacteria
UP	Universidade do Porto
UV	Ultra Violet
VLC	Vaccum Liquid Chromatography

Chapter I

General introduction

1. General Introduction

1.1. Cyanobacteria

Cyanobacteria are an ancient group of Gram-negative bacteria that have the outstanding capacity of performing oxygenic photosynthesis.[1] This bacterial phylum is widespread throughout the globe, being encountered in almost all types of environments. Due to their ability to adapt to such a range of environmental conditions, cyanobacteria show great metabolic, morphological and physiological diversity, making them an interesting group of organisms to study.[1,2]

The Earth's ability to harbor higher and more complex life forms was shaped by the photosynthetic activity of cyanobacteria, resulting in the Great Oxidation Event (GOE), which consisted of an oxygen increase in the atmosphere around 2.5 billion years ago.[1,3] Paleontological and geochemical data place cyanobacteria in the Archean eon, where the availability of abiogenic oxygen was very low, meaning that the first cyanobacterium, growing possibly in shallow warm waters, originated small hotspots of atmospheric oxygen.[1] The evolution of the photosynthetic system as well as protective mechanisms developed by cyanobacterial cells, transformed the Earth's geochemical scenario, changing the carbon, nitrogen, sulphur and phosphorous cycles.[1] The atmosphere changed by the decrease in the greenhouse effect of methane gases, cooling the Earth's surface and creating climate fluctuations, which increased cyanobacterial diversity and adaptability.[1] Cyanobacteria are considered to be the ancestors of chloroplasts, which equipped algae and plants with the photosynthetic tools for oxygenic photosynthesis, through an endosymbiotic event.[4] Both in cyanobacteria and chloroplasts, photosystem II (PSII) is the main responsible for providing oxygen and electrons after water photolysis.[1] After the first GOE, the presence of chloroplasts in more complex organisms (eukaryotes) further contributed to the biosphere development.[4] Nowadays, cyanobacteria are still a major oxygen producer and one of the main contributors to global primary production through fixation of carbon and nitrogen.[2,5]

Regarding morphology, unicellular cyanobacteria can be encountered as coccoid single cells or in colonial forms, while multicellular cyanobacteria are mainly found as filamentous strains.[6,7] These filamentous cyanobacteria can be further divided in true- or false-branching, presenting different specialized cells, such as heterocysts, akinetes, hormogonia and necridia.[7,8] Development of heterocysts is triggered by combined nitrogen limitation in the cyanobacterial environmental surroundings, and inside these specialized cells, atmospheric dinitrogen (N_2) is reduced to a useful source as ammonia

(NH_3), by nitrogenases enzymes, a process termed diazotrophy.[8,9] Nitrogenases are sensible to oxygen, so the thick cell wall of heterocysts protects these enzymes by blocking the entering of available oxygen in the cell, and creating an ideal microenvironment for nitrogen fixation. Morphologically, these cells are elliptical-shaped, distributed regularly along the filament, mainly in *Nostoc* strains, and are easily distinguished by microscopy due to its large size compared to the neighboring vegetative cells.[8,9] However, unicellular cyanobacteria (*e.g. Gloeothece*) can also fix nitrogen by using a different strategy to protect nitrogenases from oxygen, performing photosynthesis during the day (light period) and fixating nitrogen during the night (dark period).[1,9] Regarding the other cyanobacterial cell types, akinetes are spore-like cells that develop in adverse environmental conditions passing to a resting stage until favorable conditions are present again. Hormogonia are motile filaments, very important for cyanobacterial symbiosis since they disperse and can infect the host in the beginning of the symbiotic relationship. Necridia are deteriorating cells that can remove other cells along the filament.[8] The evolution of cyanobacteria has been a debatable subject over the past decades.[1,10] Phylogenetic analyses suggest that the evolution of multicellular cyanobacteria originated from an ancestral unicellular cyanobacterium and, furthermore, that cyanobacterial evolution from unicellular to multicellular taxa might not have been an unidirectional event.[1] This evolution improved cyanobacterial fitness and adaption by decreasing cellular burden and compartmentalization of specific tasks to specialized cells and expanding cyanobacterial morphological diversity (Figure 1).[10]

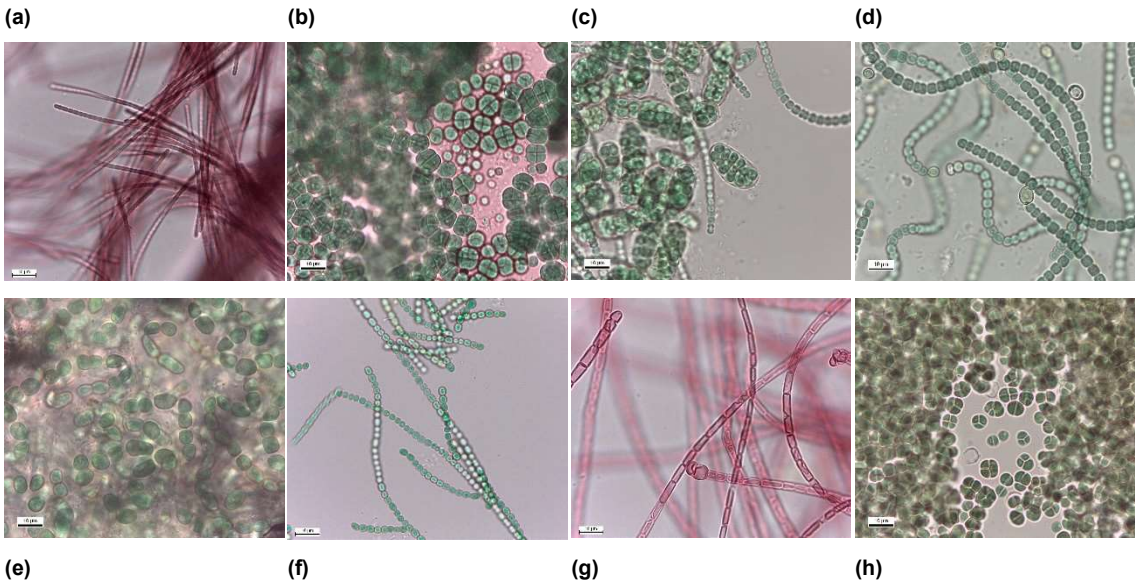


Figure 1 – Examples of the cyanobacterial morphological diversity, represented by eight strains of the LEGE culture collection (LEGEcc). **(a)** unidentified filamentous cyanobacterium LEGE 11480; **(b)** unidentified Pleurocapsales LEGE 06147; **(c)** *Nostoc* cf. *edaphicum* LEGE 07299; **(d)** *Nostoc* sp. LEGE 12456; **(e)** unidentified Nostocales LEGE 06150; **(f)** unidentified Nostocales LEGE 11386; **(g)** cf. *Phormidesmis* sp. LEGE 11434; **(h)** *Chroococcidiopsis* sp. LEGE 06174.

There are literature inconsistencies when it comes to the morphological and phylogenetic classification of cyanobacteria. This is due in part to their complex evolutionary history, and the same is transposed to the taxonomic classification of this phylum, which has been in constant change for over a century.[7] Initially, cyanobacteria were classified as blue-green algae, according to the botanical nomenclature, based mainly on morphological traits.[1] But with the knowledge of its prokaryotic nature, Rippka and co-workers[11] proposed the division of cyanobacteria in five sections, which became the primary basis for the nomenclatural classification in Bergey's Manual of Systematic Bacteriology: I (= Chroococcales), II (= Pleurocapsales), III (= Oscillatoriales), IV (= Nostocales) and V (= Stigonematales).[12] Regarding the botanical nomenclature, in the last decades, a polyphasic approach has been employed to revise the taxonomic classification of cyanobacteria, combining information not only from 16S rRNA gene sequence, but also from phylogenetic, morphological, physiological, ecological and biochemical data.[6] The usage of the two Codes of nomenclature still brings confusion among researchers in the field. Recently, an updated revision was published by Strunecký and co-workers[7] using a traditional polyphasic approach complemented with phylogenomic data. According to the authors, the updated classification of cyanobacteria above genus level is divided into 20 orders, namely: Gloeobacterales; Thermostichales; Aegeococcales; Pseudanabaenales; Gloeomargaritales; Acaryochloridales; Prochlorotrichales; Synechococcales; Nodosilineales; Oculatellales; Leptolyngbyales; Geitlerinematales; Desertifilales; Oscillatoriales; Coleofasciculales; Spirulinales; Chroococccales; Gomontiellales; Chroococcidiopsidales; Nostocales.[7] With the increase in available genomic data from a diverse array of cyanobacterial strains, it is possible to complement and improve the taxonomic classification of such a dynamic phylum.[13] Still, the occurrence of horizontal gene transfer (HGT), as well as vertical inherence and selective loss/gain of genes can further deepen the complexity of such taxonomically classification.[1] Nevertheless, such events can explain the morphological resemblance of phylogenetically distant taxa.[6,7]

Besides cyanobacterial diversity, adaptability and resilience are two other main characteristics of this phylum. These aspects stand out when the biogeographic distribution of cyanobacteria is examined, since they are present in all kinds of habitats across the Earth.[2] Cyanobacteria can be found in aquatic systems, freshwater and marine, in terrestrial environments from deserts, polar regions and temperate soils, in transitional zones between aquatic and terrestrial (mangroves, intertidal), in biological crusts, in subaerial structures and can even establish symbiotic associations with a wide range of organisms.[2,14-16] Freshwater cyanobacteria can inhabit waterbodies varying

both in salinity and temperature and can be divided in benthic and planktonic groups. The planktonic freshwater cyanobacteria are widely known by the bloom forming, with consequent toxin release, and scum-forming events due to a formation of large colonies of coccoid or filamentous cells. On the other hand, benthic freshwater cyanobacteria are present mainly in the sediments of water bodies and can develop into mats, that can later float in the water column. Diazotrophic strains are present in both planktonic and benthic systems.[2] Marine cyanobacteria are very important for the oceanic primary production by fixing carbon and nitrogen[17] and are found in a variety of niches, from coastal areas, open waters, corals reefs, mangroves and intertidal zones. They are also divided in planktonic, floating freely in the water column, and benthic cyanobacteria that adhere to the sediments, rocks, algae or aquatic plants.[2] These habitats are of great interest due to the high variability of abiotic factors associated, *e.g.* desiccation periods, which represent particular stressful conditions for cyanobacterial growth. Even so, cyanobacteria have evolved strategies to overcome these stress factors, such as the development of exopolysaccharides (EPS) that hydrate and protect the cells against desiccation and UV radiation, allowing the formation of cohesive biofilms.[18] The presence of terrestrial cyanobacteria in soils is dependent on factors such as humidity, temperature, pH, salinity, nutrients/minerals as well as interactions with other microorganisms.[2] Cyanobacteria are representative of the biome present in the soil where they are found; there are reports of cyanobacteria growing in arid regions, deserts, temperate forests, volcanic regions, mine spoils and polar areas. It is also expectable that in more heterogenous areas, the diversity of cyanobacteria will be higher, and that different regions will have their specific cyanobacterial terrestrial communities.[2] Cyanobacteria can also live in symbiosis with a range of organisms from lower to higher levels of organization, mainly due to their capacity of carbon/nitrogen fixation and photosynthetic apparatus.[16] Cyanobacterial symbiosis will be further addressed in subchapter 1.1.1.

Nowadays, cyanobacterial biodiversity is still underestimated, since it is difficult to analyze and/or collect samples from less explored geographic regions, or even maintaining the collected strains under laboratory conditions.[2,3] Recent advances have been made to try to overcome this limitation, such as the use of metagenomics analysis from environmental samples[19,20] or the use of new sequencing techniques to improve and expand the available cyanobacterial genomic data[21]. The pattern of cyanobacterial diversity also applies to its genomic content, that can vary from 1.44 megabase pairs (Mbp) of the cyanobacterium *Atelocyanobacterium thalassa* UCYN-A to 12.07 Mbp of *Scytonema hofmanni* PCC 7110, evidencing that some cyanobacterial strains

throughout evolution suffered from genomic “streamlining” modifications.[21,22] Gene content can diverge vastly from closely related cyanobacterial strains, being important to differentiate between core (primary metabolism) and flexible genes (all other pathways, ex: secondary metabolism) within the genome.[21] Cyanobacterial genomic diversity comes from having a high number of flexible genes within its genome, that can easily lead to new genome variants, through mechanisms such as HGT, virus-induced pressure (cyanophages), or even by symbiont-host interactions. These events can also alter the structure and dynamics of microbial communities, besides its genetic variability.[17,23] The adaptability and fitness of cyanobacteria to harsh environmental conditions arises from its genomic potential and gaining knowledge of that potential will help understand the interactions of cyanobacteria with other organisms. [13,21]

1.1.1. Cyanobacterial symbiosis

Symbiotic relationships between different organisms can be divided into parasitic, commensal or mutualistic and they are defined by the equilibrium between the costs/benefits of the relationship itself.[22,24] Mutualism in particular, is based on a peaceful transaction in which the symbiont provides a service to the host in exchange of another, both benefiting from the association. Cyanobacteria can maintain mutualistic symbiosis with a wide range of organisms.[16,22] The interdependency between the cyanobacterial symbiont (cyanobiont) and its host will define if the symbiosis is obligate or facultative.[22,25] In obligate symbiosis the partner cannot survive without the other, as is the case of *Nostoc azollae*, an obligate symbiont of *Azolla filiculoides*. In obligate symbiosis, genome streamlining can occur, characterized by a genomic content reduction of the symbiont that is further compensated by the host. In the *Nostoc-Azolla* association, a high number of pseudogenes is present in the cyanobiont.[17,25] Facultative cyanobionts establish symbiosis through environmental or host-derived chemical cues, for example the interaction between the cyanobiont *Richelia* and the diatom *Rhizosolenia* present in oligotrophic regions, due to the diazotrophic life style of this cyanobacterium.[26] Regarding genome alterations, genomic amplification can occur in facultative symbiosis, in which there is an increase of the cyanobiont genomic content, allowing the expansion of its metabolic repertoire, and thus meet the requirements of the host organism and adapt to the new created environment.[25,27] Cyanobacteria are excellent symbiont microorganisms due to their ability to fix nitrogen- and/or carbon, to provide defense through the production of bioactive secondary metabolites (photo-protective, anti-grazing, antimicrobial, etc.), and above all, to their remarkable adaptability to the host

conditions.[24] In return, the host can be a carbon source (in cases where the cyanobiont is not able to perform this task), can provide shelter against predation, provide oxygen and allow the cyanobiont to grow in a less competitive environment.[24] Even so, the main reason for cyanobacteria to change from a free-living state to a symbiotic relationship seems to be nitrogen-fixation, since cyanobacteria are a major player in the global nitrogen cycle.[8,14] In the marine environment, cyanobacteria are known to enter symbiosis with both planktonic (diatoms, dinoflagellates and macroalgae) and benthonic species (sponges, ascidians and echuroid worms) mainly in oligotrophic but also in nutrient rich areas.[16] A recent study showed that *Synechococcus*, *Prochlorococcus* and *Lyngbya* are the most encountered strains in marine symbiosis.[24] In respect to terrestrial symbiosis, there are examples scattered in all plant divisions: with algae, cyanolichens are a good representative of symbiosis, having also associated fungi; with bryophytes (non-vascular plants), cyanobionts are found in liverworts, hornworts and mosses; with pteridophytes (vascular plants), the well-known water fern *Azolla*; with gymnosperms (naked seed vascular plants), one of the most ancient symbiosis, with cycads; and finally, with angiosperms (flowering plants) the only known symbiosis is with the genus *Gunnera*.[9,16] In plant symbioses the diversity encountered will be dependent on the local microbiome in close proximity to the host, since symbiosis is established with those microorganisms, with the exception of *Azolla*, which is vertically inherited.[8] In some cases, there is the development of specific structures to harbor these cyanobionts: for example, the coralloid roots in cycads and the red stem glands of *Gunnera*.[8] The most common cyanobacterial genus in terrestrial symbiosis is *Nostoc*.[16] Firstly, because terrestrial environments are dominated by these geographically-widespread genus; secondly, they have a facultative life cycle (free-living or facultative symbionts); and lastly, they can be multicellular, developing heterocysts crucial for nitrogen fixation, and it was already shown that an increase of 25-35% of these specialized cells occurs in cyanobionts.[8]

Considering the evolutionary leap of cyanobacterial symbiosis in global transformative events (chloroplasts and photosynthesis) and trying to comprehend the evolution of cyanobacterial symbiosis with more complex organisms, its symbiosis with the “living fossil” gymnosperm cycads can be considered a good model. There are records of cycads growing on Earth for over 300 million years, co-existing with dinosaurs in the Mesozoic era, where this group of plants was scattered all over the planet.[8,16] Nowadays, cycads are mainly distributed in tropical and sub-tropical regions and several specimens are present in other climates, but growing in botanical gardens/green houses.[8] Cycadales are currently the only gymnosperm order with the ability to form new

symbiosis with cyanobacteria and they are divided into three families: *Cycadacea*, *Strangeriaceae* and *Zamiaceae*.[8,16] The ancient association with cyanobacteria probably arose to overcome the poor-nutrient soils of that era that led to the development of a specialized root system to facilitate symbiosis, called coralloid roots.[8] Besides these engorged, dichotomously-branched roots, cycads have a tap-root, consistent with the primary root system of plants, lateral roots and an aerial stem covered with persistent leafs.[8,16] Coralloid roots have been documented in all cycads specimens studied, showing a negative geotropism growth, growing laterally and sometimes reaching the soil surface.[16] Before reaching this stage, a pre-coralloid root is formed where cyanobacteria are not present. After the maturation process, the sheath protecting the top of the pre-coralloid root is replaced by an external thin layer with dermal breaks through which cyanobacteria enter and stablish symbiosis.[8] The infection process of cyanobacteria within coralloid roots is represented in Figure 2, adapted from Chang, et. al.[8]

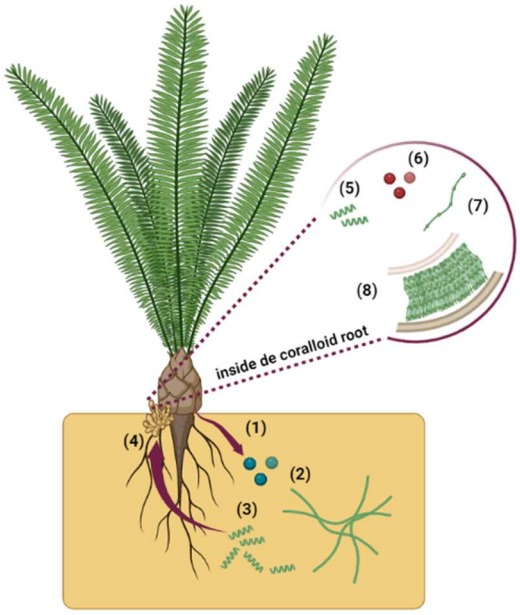


Figure 2- Representation of the infection process of a cycad plant by a cyanobacterium to form the specialized coralloid roots. **(1)** release of hormogonium inducing factors (HIFs) by the plant; **(2)** cyanobacteria vegetative cells differentiate into hormogonia; **(3)** motile hormogonia filaments are directed to the cycad; **(4)** infection of the pre-coralloid root; **(5)** hormogonia filaments enter the root; **(6)** production of hormogonia repression factors (HRFs); **(7)** cyanobacteria differentiate to vegetative cells and heterocysts for nitrogen-fixation; **(8)** cyanobacteria form a layer around the coralloid root, cyanobacterial zone, being in permanent symbiosis with the cycad plant.

Briefly, the process starts with the production of hormogonium inducing factors (HIF) by the host, mainly under nitrogen depleted conditions (1), to elicit the cyanobiont to differentiate from vegetative cells to hormogonia (2). These motile filaments separate

from the vegetative cells (3), and then migrate to the root, infecting the host organism (4).[8,9] A study using methanolic extracts of *Cycas revoluta* coralloid roots led to the isolation of the diacylglycerol 1-palmitoyl-2-linoleoyl-sn-glycerol, proven to be a HIF of *Nostoc*.[28] Also, the production of HIFs alone seems to be insufficient for a successful infection of the host, so other chemoattractants must have a role, such as sugar-based molecules.[9,24,28] After hormogonia entry (5), the filaments become sessile and the production of hormogonia repressing factors (HRF) occurs (6) for cyanobacteria to return to their vegetative state and differentiate into heterocysts, for nitrogen-fixation (7).[8,9] Nostopeptolides (specialized metabolites) are thought to have hormogonia repressing activity, since the mutant of the polyketide synthase (PKS) biosynthetic gene cluster *pks2*- used to compare the metabolite production in *Nostoc punctiforme*, led to a decrease in nostopeptolide production and to an increase in hormogonia differentiation. However, the production of these metabolites seems to be downregulated in symbiosis. Thus, nostopeptolides might be responsible for the autogenic regulation of hormogonia, specifically during the immunity period, an interval before hormogonium development can happen again, even in the presence of HIF.[29,30] The heterocyst-containing cyanobacterial filaments form a blue-green ring inside the coralloid root, called cyanobacterial zone, separating the cortical layer of the root in two (8), being in permanent symbiosis with the cycad.[8] The frequency of heterocysts within the coralloid root can increase up to 80%, compared to the free-living state.[8] Fixed-nitrogen is then transferred to the cycad in the form of amino acids (glutamine or glutamine/citrulline), unlike all other symbiotic relationships that supply ammonia.[16] Another metabolic difference presented by the cyanobiont is related to carbon uptake: it can either assume heterotrophic behavior with the carbon being provided by the host or, perform CO_2 fixation using RuBisCO enzyme.[8,16] Light-dependent photosynthetic activity is also altered inside the coralloid root, probably inactive with the exception of the occurrence of dermal breaks where cyanobacteria can access light.[8] The consortium cyanobacteria-cycad shows low specificity, since a diverse range of cyanobacterial genotypes can be found in only one cycad plant and/or one specific cyanobacterial strain can be found in multiple cycad hosts.[8,16] The majority of cyanobacterial strains isolated from cycad coralloid roots are *Nostoc* strains, but records of *Calothrix*, *Scytonema* and *Richelia* are also present in the literature.[8] Besides cyanobacteria, a range of other heterotrophic bacteria were found inside the coralloid roots, for example, *Caulobacter* was isolated from natural-growing coralloid roots collected in Mexico, and its association with the *Nostoc* species may be responsible for the production of specialized metabolites.[31]

There is some knowledge regarding genes that are essential for symbiosis, not specific to cycads, but found throughout the plant-cyanobacterial symbiotic relationships. Briefly, the *hrm* operon acts on hormogonia suppression and differentiation of cells into heterocysts; *hetR* and *hetF* are responsible for heterocyst development; *ntcA* is responsible for the transcriptional regulation of several genes encoding proteins responsible for nitrogen regulation and others, including *hetR*, which has also been shown to be required for host infection; the *nif* operon contains the genetic information for ensuring nitrogenase activity and it is found in both free-living and symbiotic cyanobacteria; *ctpH* is controlled by HIF and is related with PSII; *tprN* is also related to heterocysts, mainly in the maturation of these specialized cells, but its role in symbiosis is still not known.[8,9] Moreover, additional genes involved in phototaxis and motility were identified in cyanobacteria stimulated with plant HIFs and can be important for the colonization of the host, *ptx-hmp*, *hps-pil-gvp*, respectively.[27,32] Information regarding the genetic determinants involved in cyanobacterial-cycads symbiosis, and cyanobacterial symbiosis in general, is still very limited, but new genomic tools, such as metatranscriptomic analysis using next-generation sequencing (NGS) of the host-cyanobiont association, can provide better insights on how these interactions occur.[20]

Some limitations are in play when talking about cyanobacterial-cycad associations, mainly the laboratory cultivation of isolated symbiont strains, experimental designs to access the interactions between both organisms, working with cycads recovered from their natural habitat (with the natural occurring cyanobionts) and not from botanical gardens, and finally the separation of the cyanobiont from its host, that can lead to morphological, physiological and genetic changes.[8]

All symbiotic relationships are dependent on the exchange of infochemicals between host and cyanobiont, and so exploring these metabolites will provide knowledge in one of the most ancient associations on the planet, mainly on signal-receptor evolution.[24,30] Symbiotic interactions of cyanobacteria with several hosts can trigger activation of specific biosynthetic pathways, leading sometimes to the production of specific secondary metabolites. Therefore, symbiosis may represent a unique niche for the discovery of new natural products and enzymatic entities.[8,24]

1.2. Natural Products

All living organisms have the ability of producing specialized metabolites to cope with environmental and/or biotic pressures during their existence, and these metabolites are

widely termed natural products (NP). In that sense, natural products encompass all small molecules produced by organisms, not essential for growth, development or reproduction.[33-36] Through the course of evolution, these specialized metabolites were refined to fulfil a specific task, improving the fitness of the producing organism by aiding communication, defense and feeding strategies, as well as protection against stress factors. [34,36,37] Natural products represent an incredible source to explore chemical diversity and biological activities, and have attracted the interest of the pharmaceutical industry for the development of therapeutic agents to treat human health conditions.[37-39] Although, over the centuries, traditional medicine used organic extracts for the treatment of several human diseases, only with the advances in organic chemistry applied to pharmacology, these NP-containing extracts from plants, fungi and/or bacteria were purified and subsequently commercialized as drugs or NP-derived drugs.[33,40] When comparing the use of synthetic molecules with natural products for biomedical resources, the latter provide advantages such as: more complex and diverse structures (*e.g.* ring systems, functional groups); higher molecular mass values; relatively higher numbers of sp_3 carbons, hydrogen and oxygen atoms; more stereogenic centers; higher hydrophilicity and molecular rigidity.[37,40,41] These features are responsible for natural products' bioactivity in a wide range of molecular targets.

The traditional NP-drug discovery workflow is bioactivity-guided, where organisms are cultured, extracted and screened against a molecular target. The resulting bioactive fraction is purified through several rounds of isolation steps, until reaching a pure metabolite, whose structure is further characterized.[34,37,38,42] From the early 19th century on, this discovery approach enriched the pharmaceutic industry with a pool of potent molecules with antibacterial and anticancer activities, for example the isolation of penicillin and paclitaxel, from fungal and plant origin, respectively.[42,43] Also, with the discovery of the first bacterial antibiotic (streptomycin), it was noticeable the potential of bacteria for new natural products isolation, with mainly actinobacteria but also myxobacteria and cyanobacteria taking center stage.[38,42] Nowadays, some of the discovered natural products or NP-derived drugs are still commercialized, so this research line is still important for the discovery of new and potent therapeutic agents.[44] Besides cancer and infectious diseases, natural products and NP-derived molecules are also used for other therapeutic areas (cardiovascular, multiple sclerosis, diabetes, etc.), and other applications such as agriculture, nutraceuticals, cosmetics and dyes industries.[33,34,44,45] Even though the use of the this classical NP-drug discovery rendered several active molecules, several disadvantages led to the abandonment of NP-drug discovery programs by the pharmaceutical industry. The main cause was the re-

discovery rate of known compounds, that led to an abrupt decrease of new active molecules retrieval, accompanied by the need to perform dereplication prior to compound purification.[34,37,38,46] Other limitations that arose from the bioactivity-guided isolation were: time consuming workflows; difficulty in growing the organisms under laboratory conditions; harvesting enough biological material from nature; following-up with the isolation of all the high-throughputs screening hits; abundance of the isolated active compound required for biologic testing; total synthesis of intricate chemical scaffolds; and financial costs that led a promising research area into a stagnation point. Also, for a bioactive molecule to enter and pass all clinical trials phases, and subsequently reach the pharmaceutical market a stretched timeline was needed, and some industries could not sustain such costs. [35,37,38,46] Fortunately, from the middle to the end of the 19th century, two major methodologies were developed, becoming crucial for the evolution of NP-research, namely metabolomics (1960's) and genomics (1990's). While the first allowed the dereplication of the molecules present in the complex organic extracts, genome mining opened a new door to the unforeseen chemical and biosynthetic diversity of prolific organisms, mainly bacteria.

1.2.1. Metabolome mining

The high rediscovery rates of known molecules was one of the major problems with the bioactivity-guided NP-research, a problem somehow expected, since the same organisms were explored during the drug discovery "golden age", and maintained in the same conditions.[40] To avoid rediscovery of known metabolites dereplication tools were needed, and the use of metabolomics fulfilled that need, since it retrieves data from complex metabolite mixtures found in biological extracts. Besides dereplication, metabolomics or metabolome mining, can be used to prioritize a NP-containing extract for isolation based on its novelty, or to find new analogues of already known compounds.[34,37] The dereplication process consists of matching specific features of a new compound to an already known metabolite, through the use of mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR), and it crucially depends on the use of databases and spectral libraries.[47] Both techniques are still widely used in the NP-research area.

NMR can be used for natural product structure elucidation, showing the connections between the atoms in the structure, but it can also be used for metabolome mining, where complex mixtures are analyzed, and specific resonances are matched to a spectral library.[47] Several spectroscopy analysis can be performed with NMR, such as one

dimensional (1D) carbon (C_{13}) and proton (^{1}H) NMR analysis and two dimensional (2D) spectral analysis. With the 2D experiments is possible to: assign a proton to its carbon atom (HSQC – Heteronuclear Single Quantum Coherence); check the neighboring carbons of a specific proton (HMBC – Heteronuclear Multiple Bond Correlation); and even see the neighboring protons of a specific proton (COSY – Correlated Spectroscopy).[48] Even though NMR provides a structural fingerprint of molecules, acquisition of some 2D NMR experiments can be time-consuming, since higher number of scans are needed to deconvolute complex structures resonances. The assigning process itself can also take long periods of time, due to the presence of artefacts that can deepen the complexity of such assignment. These artefacts and/or impurities are not suitable for similarity recognition NMR tools used for dereplication.[49] In the last decades, efforts in improving NMR instruments and probes to use less material for high quality analysis, as well as the development of algorithms to dereplicate and prioritize extracts for isolation were achieved.[37,49] Even with such fidelity in good quality NMR spectra, assigning a NP-structure solely using this technique can still be a challenge, mainly due to presence of complex chemical scaffolds, so a second layer of information is needed, for example by using mass spectrometry.

Mass spectrometry analysis became a crucial step of NP-research since it was possible to obtain a metabolic profile from complex organic extracts with a single analysis. Advances in regard to ionization techniques, accurate and high resolution mass analyzers, made possible to calculate molecular formulas from spectral data and see clear isotopic patterns, *e.g.* from halogenated compounds.[47] In that sense, high resolution mass spectrometry (HRMS) is essential to analyze multiple metabolite mixtures, but using only full scan mode (MS1) acquisition can impose a challenge for data annotation, since several natural products can share the same molecular mass/formula, while having completely different structures.[34] Therefore, the use of high resolution tandem mass spectrometry (HRMS/MS – MS2) provides a second layer of structural data, being able to show the fragmentation pattern of a specific parent ion from a complex extract. Both HRMS and HRMS/MS can be coupled with liquid chromatography (LC), a chromatographic step helping with the separation of the complex extracts, direct diffusion mode (DMIS) only using spectral analysis, and imaging (MSI) that allows to trace compound distribution inside and/or surrounding the organism.[37,47] The daunting amount of HRMS/MS data acquired in a single run of a complex extract, powered NP-researchers to join bioinformaticians to develop numerous *in silico* tools, for spectral analysis and annotation.[38,50] A recent review by Sorokina and Steinbeck[33],

identified around 90 open-access databases of natural products since the beginning of the millennium.

GNPS (Global Natural Products Social Molecular Networking) is a web-based software of mass spectrometry and it became an essential tool for dereplication and to help prioritize extracts and/or organisms for new natural products isolation.[34,37,51] Molecular networking is one of the tools in GNPS, and it operates based on untargeted MS2 data of one (or more) extract(s), creating a network with several clusters that are grouped according to the MS2 similarity of all the metabolites in the analyzed extract(s). Structurally similar compounds will share similar fragmentation patterns, so with this tool it is also possible to annotate analogues of known compounds, by matching experimental data with MS2 from spectral libraries and databases. With GNPS's molecular networking it is also possible to analyze extracts from different organisms, making it possible to map which molecules are organism-specific, facilitating the selection for new NP-discovery, and revealing early on the uniqueness of such metabolite.[34,37] Nevertheless, this MS-based metabolomics approach also shows some limitations, starting with the simplest problem, metabolite ionization. A metabolite needs to ionize to be further analyzed using MS tools. Other simple and crucial points are how easily metabolites fragment under the selected experimental conditions, since some bonds are more difficult to break then others, and the quality of the uploaded spectra that will reflect on the quality of the network. Annotation of experimental spectra is also a major challenge, so most of the spectral libraries only represent a minor part of all NP-metabolome, which in turn leads to low NP-chemical diversity in such libraries/databases, hampering the dereplication process.[34,37,50] To overcome this issue, some tools were developed to recognized only substructures of compounds or were even developed for specific type of natural products, which is the case of Dereplicator and Dereplicator+ tools present in GNPS.[34]

Although metabolomics greatly enhanced NP-research, overcoming the rediscovery of known compounds, the unforeseen chemical scaffolds of some natural products cannot be assigned and understood solely using this approach. Knowledge on the biosynthetic machinery and building blocks behind these scaffolds will give a glimpse behind nature's curtain in regard to natural products development and evolution.

1.2.2. Genome mining

Near the end of 20th century, DNA sequencing of a NP-producing microorganism, *Streptomyces coelicolor*, showed higher biosynthetic potential when compared to the

compounds that were known to be produced by that particular strain. These new insights into the rich biosynthetic genomic regions from NP-producing organisms, gave rise to a new approach for NP-discovery, genome mining.[36,42,46] By definition, genome mining consists on finding genes that will participate on the biosynthesis of a secondary metabolite, being sometimes possible to predict its structure.[37,42] Due to the ongoing decrease of sequencing costs and advances in sequencing technology, genome mining became a quick available tool for the NP-research community to gain more insights into chemical diversity, perform *in silico* dereplication a prioritize specific organisms for new NP-isolation.[34,38,45] Besides *Streptomyces*, other NP-prolific organisms, such as cyanobacteria, fungi and myxobacteria, revealed more genomic richness in comparison with the known metabolome.[38] There are two key points regarding genome data in such microorganisms: first is the presence of biosynthetic gene clusters (BGC), defined by a group of clustered genes that encode for specific enzymes, which in turn are responsible for the synthesis of a specific molecule; and second is the conservation of such core biosynthetic enzymes throughout gene evolution.[37,41,42,45]

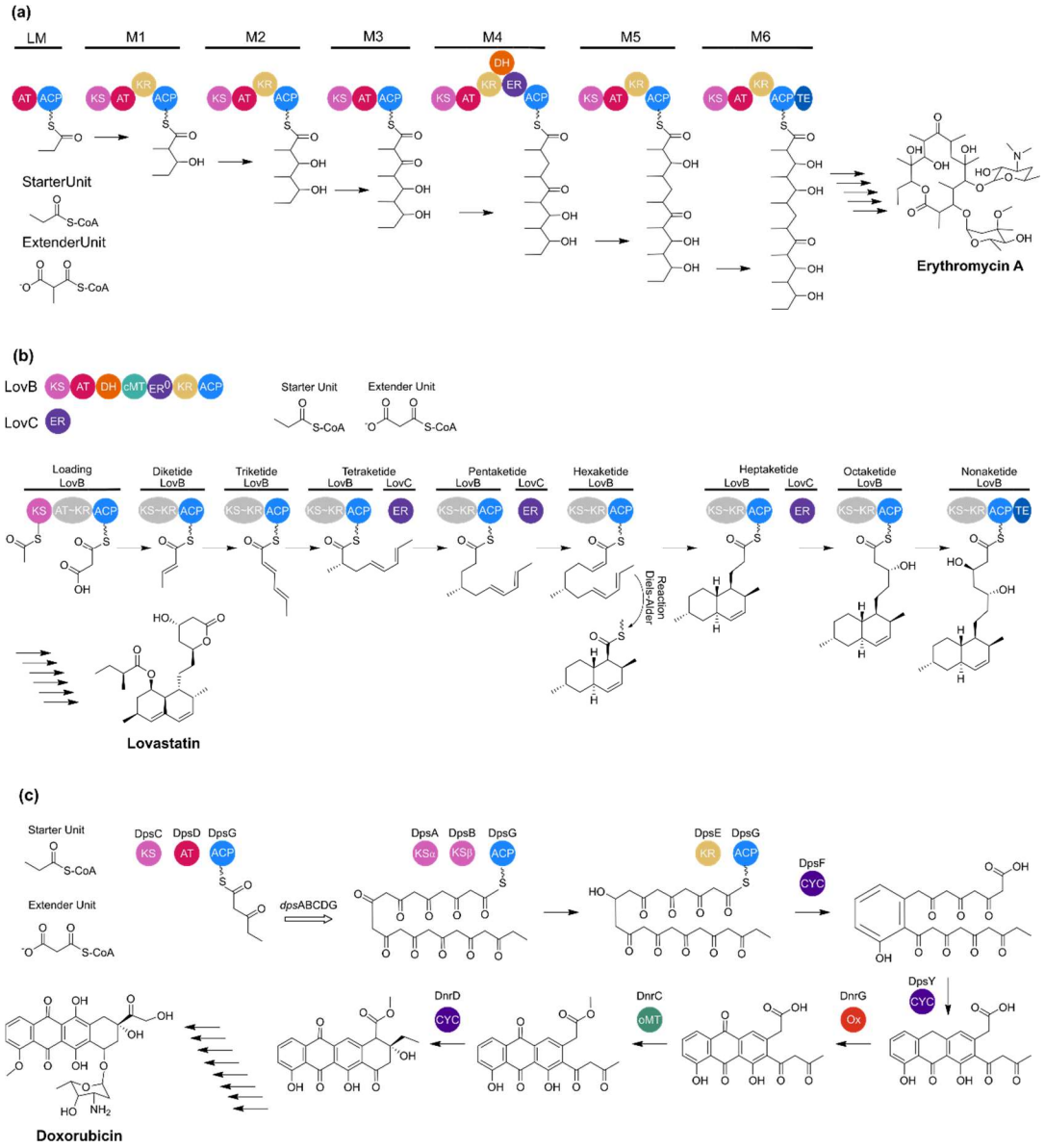
To access such genomic information, it is important to understand the genome mining workflow. Briefly, microorganism(s) are submitted to (meta)genomic extraction, and the resulting DNA is sequenced using short/long-reads technology to recover as much good-quality information as possible; followed by assembly of individual sequence reads to form contigs; in case of metagenome data, similar contigs are linked to one or more microorganisms in a process termed binning; next, genes in contigs are annotated; identification of the BGC of interest and finally natural product structural prediction.[34,38] Prior to sequencing, all of these steps are performed using several bioinformatic tools, for example, SPAdes, MaxBin and AntiSMASH, for assembly, binning and annotation, respectively.[34] AntiSMASH has proven to be a crucial tool for BGC identification according to homology searches of the biosynthetic gene encoding enzymes. Some of the secondary metabolite biosynthetic enzyme classes in AntiSMASH are polyketides synthases (PKS), non-ribosomal peptides synthetases (NRPS), ribosomal synthesized and post-translational modified peptides (RiPP), terpenes, carbohydrates, alkaloids and hybrid forms.[52,53] An additional bioinformatic tool, essential for genome mining NP-research and dereplication is the MIBiG repository (Minimum Information about a Biosynthetic Gene Cluster), joining the information of annotated biosynthetic pathways to its characterized natural products.[54]

1.2.3. Natural products biosynthesis

The main biosynthetic pathways present in natural products research are polyketides (PKs), non-ribosomal peptides (NRPs), ribosomal peptides (RiPPs), hybrid forms of these pathways and terpenes. PKS are a group of multi-domain enzymes or enzyme-complexes composed mainly by the core domains, acetyltransferase (AT), ketosynthase (KS) and thioesterase (TE) and other facultative domains. Its biosynthesis starts with the connection of an acyl-CoA starter unit to an acyl carrier protein (ACP) catalyzed by the AT-domain, followed by carbon chain elongation catalyzed by the KS, through decarboxylative Claisen condensations, this process can occur multiple times and ends with hydrolysis or cyclization of the elongated polyketide from the ACP, by the TE-domain.[36,55,56] Besides these domains, ketoreducatse (KR), dehydratase (DH) and enoyl reductase (ER) domains lead to altered reduction states of the keto groups incorporated in the polyketide. PKS can be divided into three types: type I, II and III PKS.[55,56] Type I PKS are multifunctional enzymes that have linear and covalent bound domains, that can be iterative (fungal), in which the same set of domains is used repeatedly for polyketide elongation, or non-iterative/modular (bacterial), in which one or more different domain-containing modules, arranged in an assembly-line manner, are responsible for the carbon chain elongation.[55,56] Still within the Type I PKS, these can be sub-divided into *cis*-AT and *trans*-AT PKS, in which the *trans*-AT PKS while working in a modular fashion, lacks the AT domain, using a free-standing enzyme that acts in *trans*. These type of PKSs are responsible for a huge chemical and enzymatic diversity in natural products biosynthesis, diverging from the collinearity PKS biosynthetic principles.[36,41,42] Type II PKS are monofunctional multi-enzyme complexes, that catalyze several rounds of aromatic polyketide elongation. These enzymes are iterative and characterized by the presence of two KS domains (*α* and *β*) and a cyclase.[36,55,56] Finally, the type III PKS are homodimeric enzymes, acting directly on acyl-CoA thioesters, catalyzing consecutive Claisen condensations, cyclizations and aromatizations, originating poly-*β*-keto chains. These enzymes are also iterative, mainly found in plants but with recent records of bacterial origin, and can have different catalytic domains such as chalcone, stilbene or acridone synthases.[36,55-57]

In general, PKS natural products show an high structural and enzymatic diversity, as well as, several distinct biological functions, such as the antibacterial erythromycin (Type I modular PKS)[55], the anti-cholesterol lovastatin (Type I iterative PKS)[58], the antitumor doxorubicin (Type II PKS)[59] and the cytotoxic cylindrocyclophanes (Type III PKS)[60].

The biosynthetic pathways of these polyketide natural products are represented in figure 3.[55-60] From all the discussed domains, KS-domains are the most conserved in polyketide biosynthesis, functioning as a fingerprint of these biosynthetic pathways.[36]



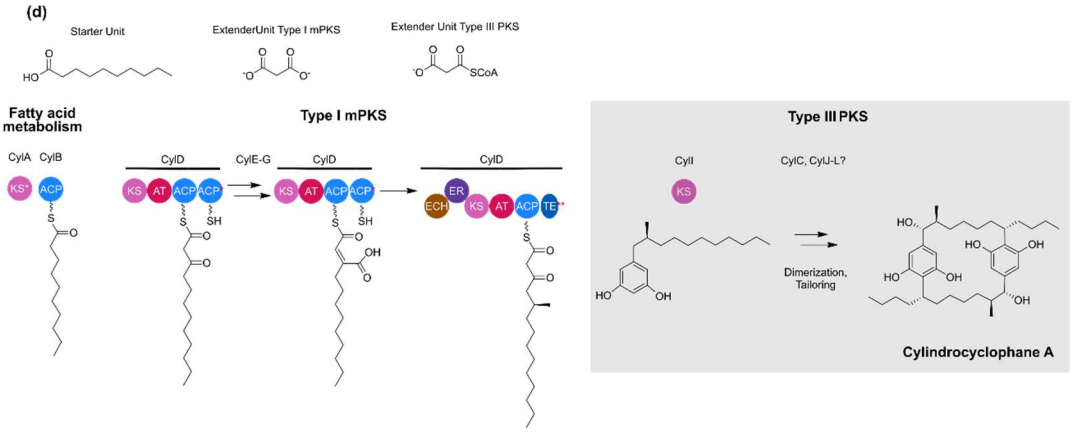


Figure 3 – Examples natural products biosynthetic pathways of polyketide origin. **(a)** Type I modular PKS biosynthetic pathway of erythromycin A: the starter unit is Propionyl-CoA, picked up by the loading module (LM) and followed by the incorporation of six extender units of (2S)-Methylmalonyl-CoA by modules 1-6; further biosynthetic steps of the cyclization and addition of the monosaccharide decorations are not represented. **(b)** Type I iterative PKS biosynthetic pathway of lovastatin: the starter unit is Acetyl-CoA (presumably it was a Malonyl-CoA that was further descarboxilated) and elongation is performed iteratively by eight units of Malonyl-CoA to form dehydromonacolin L (not shown) that is then modified to give lovastatin; LovC is a trans-acting enoyl reductase and besides C-C Claisen condensations, a Diels-Alder reaction also occurs. **(c)** Type II PKS biosynthetic pathway of doxorubicin: the starter unit is Propionyl-CoA, recognized specifically by DspC and DspD, and elongation is performed with nine units of Malonyl-CoA by DpsA and DpsB; dpsABCDG is considered the minimal PKS for this biosynthetic pathway, and several other enzymes involved are not represented here. **(d)** Type III PKS biosynthetic pathway of cylindrocyclophane A: the pathway starts with a modification of decanoinc acid (started unit) to form decanoyl-CylB thioester, as part of the fatty acid metabolism; the thioester is picked up by Type I mPKS enzymes and extended with two malonate units and the type III PKS CylI, using two malonyl-CoA unit's origins a monomeric precursor and other enzymatic reactions create cylindrocyclophane A; ACP* may participate in β-methyl formation and TE** may be inactive. AT – acyltransferase; KS – ketosynthase; ACP – acyl carrier protein; KR – keto reductase; DH – dehydratase; ER – enoyl reductase; ER^0 – inactive enoyl reductase; TE – thioesterase; cMT – carbon methyltransferase; CYC – cyclase; Ox – oxygenase; oMT – oxygen methyltransferase; ECH – enoyl-CoA hydratase. Each arrow is correspondent to one biosynthetic step.

Another important family of biosynthetic multi-modular enzyme complexes are the non-ribosomal peptides synthetases (NRPS), following an assembly-line activation and condensation reactions to join amino acids, originating small peptides.[36,41,61,62] The core domains present in the NRPS modules are the adenylation (A) domain, the peptidyl carrier protein (PCP) domain and the condensation (C) domain.[36,61] NRPS biosynthesis starts with the recognition and activation of an amino acid by the A-domain, followed by PCP-domain binding the active amino acid to its cofactor 4'-phosphopantetheine arm (PP), through a thioester bond, and the C-domain catalyzes the peptide bond (from N- to C-terminal) to form the peptide chain. This process continues through different modules until a thioesterase cleaves the peptide either by hydrolysis originating a linear peptide, or by internal cyclization giving a cyclic peptide.[61] In some cases, an epimerase domain might be present, showing the switched chirality of the incorporated amino acid (mainly from L- to D-form).[36,41] Each module is responsible for a specific amino acid

activation and incorporation, and its collinearity to form the resulting peptide, makes these biosynthetic enzymes a good candidate for natural product structural prediction.[42] The A-domains can recognize the twenty proteinogenic amino acids (both L- or D-configurations) as well as non-proteinogenic amino acids, for example ornithine or hydroxy amino acids.[61] Tailoring enzymes can posteriorly modify the peptide to encompass methylations, halogenations, hydroxylations, glycosylations, acylations or even heterocyclic ring formations, increasing the structural diversity of such natural products.[36,61] Several NRPS-derived compounds are widely known as commercialize drugs, as penicillin and vancomycin.[36] But also other bioactive compounds, such as the cyanobacterial toxin Nostocyclopepetide A2, are constructed according to the NRPS assembly line, as represented in figure 4.[63] Besides NRPS and PKS biosynthetic machineries, hybrid forms containing biosynthetic enzymes of both natures, are responsible for well-known bioactivities, such as the cyanobacterial toxin microcystin[64], anticancer natural products aplidine (dehydrodidemnin B)[65] and Curacin A[66].

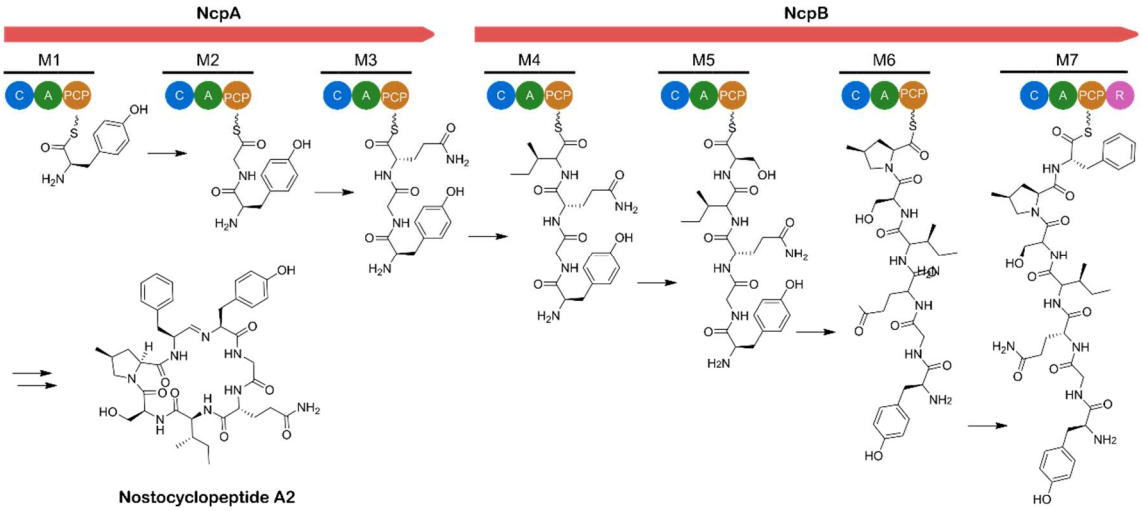


Figure 4 - Non-ribosomal peptide biosynthetic pathway of Nostocyclopeptide A2. NcpA comprises three modules incorporating the amino acids tyrosine, glycine, and glutamine, followed by the enzyme NcpB with four modules in which the adenylation domain activates isoleucine, serine, 4-methyl-proline and phenylalanine. The peptide is cleaved from the PCP by domain R, a NADPH dependent reductase, and the resulting reactive aldehyde triggers the self-assembly of the macrocyclic imine. C – condensation domain; A – adenylation domain; PCP - peptidyl carrier protein; R – reductase domain. Each arrow is correspondent to one biosynthetic step.

RiPPs are also a large reservoir for chemical and enzymatic natural product diversity, with recent research efforts directed to this group of ribosomal synthetize peptidic natural products. The biosynthetic pathways of RiPPs are mainly composed of genes encoding for a precursor peptide, tailoring enzymes, and a protease. The precursor peptide has usually a N-terminal leader peptide and a C-terminal core peptide, in which the first

encodes for all post translation modifications recognition sequences applied to the latest, culminating with a proteolytic cleavage of the mature core peptide from the leader, once the tailoring reactions occur.[62,67]

Lastly, terpenoids are one of the largest NP-bioactive classes, also comprising an high structural diversity with metabolites isolated from bacteria, fungi and plants.[42] In respect to the biosynthetic machinery and assembly, terpenes are composed of simple isoprenoid units, cyclized by terpene cyclases, enzymes catalyzing independent biosynthetic reactions, followed by downstream modifications with several tailoring enzymes.[41]

Even though the use of the described genomic tools greatly aided the genome mining NP-workflow, its rate of success is based on gene annotation of already know NP-biosynthetic classes. For example, if metabolites are derived from primary metabolites suffering tailoring reactions, with no specific enzyme responsible for its assembly, or if the enzymes responsible for a NP-assembly are scattered in the genome and not in a BGC, it constitutes a huge challenge for biosynthetic pathway identification.[41] When directed to the discovery of new biosynthetic routes or enzymes, other mining strategies are more suited to access such cryptic information.[37,38,45] Following that line of thought, it is important the differentiate between three categories of BGC expression and/or association to the natural product itself: when a BGC is not active under specific conditions (*e.g.* laboratorial growth) is termed silent; if a BGC is identified as a NP-producer, but the metabolite resulting from that BGC is not known or identified, it is termed orphan; and when, neither the natural product or the BGC are identified using the available tools, is termed cryptic.[41] All the sequenced data that belong the category of these cryptic pathways, has been termed the “microbial dark matter”, the unknown unknowns, where the true NP-diversity is hidden.

1.2.4. Other methodologies for natural products discovery

Currently, efforts are being made in the NP-research field to explore the cryptic biosynthetic pathways and metabolites, to unravel new chemical diversity and unforeseen enzymology. For that, two major approaches are usually followed, taxonomy and ecosystem-guided approaches.[37,38,42] In the first methodology, phylogenomic studies can point out less explored organisms for NP-discovery, since some NP-families can be linked to specific taxa, and a high potential for chemical diversity is presented in deep-branched less explored taxa.[38] The ecosystem-guided NP-discovery is directed towards

exclusive ecological niches both geographically and/or symbiotic interactions.[38] Ecological conditioning is a major driving force for species evolution, encompassing also its secondary metabolism, as the example given previously, of a symbiont producing a specialized metabolite improving its host fitness. Using this approach to gain insights on the cryptic biosynthetic and chemical NP-potential can be performed using culture dependent and independent methodologies. Culturomics (culture-dependent) is an already known tactic to improve natural products expression and production, for example with the OSMAC workflow (one strain many compounds), altering culture conditions and performing co-culturing of strains. Nevertheless, recent strategies coupling culturomics techniques with genome mining have proven to access new chemical diversity and biosynthetic pathways: modifying genetic promoters/regulators/resistance genes in the native host or by heterologous expression, with the help of synthetic biology techniques [39,42]; *in situ* cultivation of uncultivable microorganisms using iChip technology[68]; improving high-throughput microbial culturing with machine learning techniques[69], are some examples of culture-based NP-discovery. On the other hand, culture-independent strategies have already been applied for NP-discovery, by recovering the microorganisms directly from the environment, *e.g.* cyanobacterial mats in tropical regions.[70] However, with improved metagenome sequencing techniques, it became possible to sequence entire microbial communities, using environmental DNA to create libraries used posteriorly for cloning strategies.[38]

Nowadays, natural products research is shifting towards the integration of several mining strategies, applying several “omics” technologies to have a better picture of the cues that lead to the unforeseen chemical and biosynthetic diversity. The focus on ecology-based NP-discovery is not only improving the discovery on new chemical scaffolds and enzymatic reactions, but also on understanding the complex interactions between microbial communities, and both will improve the researchers toolbox for all NP-research lines, from new drug discovery to more intricate synthetic organism manipulation.

1.3. Cyanobacterial Natural Products

Cyanobacteria are well-known producers of toxins, such as microcystin and cylindrospermopsin, occurring in freshwater habitats, by the formation of toxic blooms.[13,71] This phenomenon called the attention of several researchers, and until the present day, cyanotoxins are the focus of monitoring studies to control the levels in water

bodies in many countries, since it poses a risk to the surrounding microbiota and human health.[72] Nevertheless, with the emergence of bacterial NP-research the focus drifted towards the discovery of bioactive molecules, that rendered potent cyanobacterial metabolites, using the bioactivity-guided isolation workflow.[71,73,74] The perceived diversity of both chemical scaffolds and inhibitory effects of the isolated metabolites, classified cyanobacteria as a promising source for new natural products with pharmaceutical applications.[75] Half a century later, cyanobacteria proved its worth providing the NP-research community with fascinating chemical diversity, wide range of biological applications and an unusual set of enzymes, which reactions are hard to translate for synthetic chemists.[71,74,76]

Cyanobacterial natural products can be divided into ten classes, according to structural characteristics: alkaloids, depsipeptides, lipopeptides, lipids, macrolide/lactones, peptides, polyketides, polysaccharides, terpenes and others.[75,77] Unique chemical scaffolds can be pointed out for each of these chemical classes, as is demonstrated in Figure 5. The indole alkaloids are isoprenoid-indole compounds such as hapalindoles and hapalindole-like compounds like for example ambiguine isonitrile **1** and scytonemin **2** showcasing nitrile and imine bonds, respectively;[71] depsipeptides are peptides that have one or more ester bonds adding to the amide bonds, *e.g.* microviridin B **3** and apratoxin **4** that also has a *tert*-butyl side chain;[71,76] lipopeptides are composed by a core peptide decorated with a variety of acyl moieties, for example puwainaphycin A **5** and barbamide **6** a chlorinated lipopeptide;[74,78] examples of cyanobacterial lipids are glycolipids used for heterocysts formation and γ-linoleic acid;[13,79] macrolide compounds are composed of lactone rings, such as cryptophycin-1 **7** and scytophycin **8**, both containing an epoxide group;[80] cyanobacteria peptides are very diverse, seen for example in pattelamides **9** with thiazole and oxazole rings or aeruginoside **10** decorated with a monosaccharide;[74] polyketides are also very diverse group of compounds, for example jamaicamides **11** with a alkynyl bromide and a vinyl chloride, and anatoxin-a **12** with a azabicyclo ring;[71] an example from cyanobacterial polysaccharides is nostoflan;[76] a terpenoid isolated from cyanobacteria is, for example, the odorous compound geosmin;[71] finally in the “others” class, diverse structures are included like the alkylresorcinols bartolosides **13** with two chlorinated alkyl chains, and nocuolin A **14** having a 1,2,3-oxadiazine moiety.[81,82] Dittman and co-workers[71], accounted in 2015 around 1100 cyanobacterial unique metabolites, and in 2019 Demay and co-workers[75] mentioned 1630 cyanobacterial natural products, divided into 260 compound families.

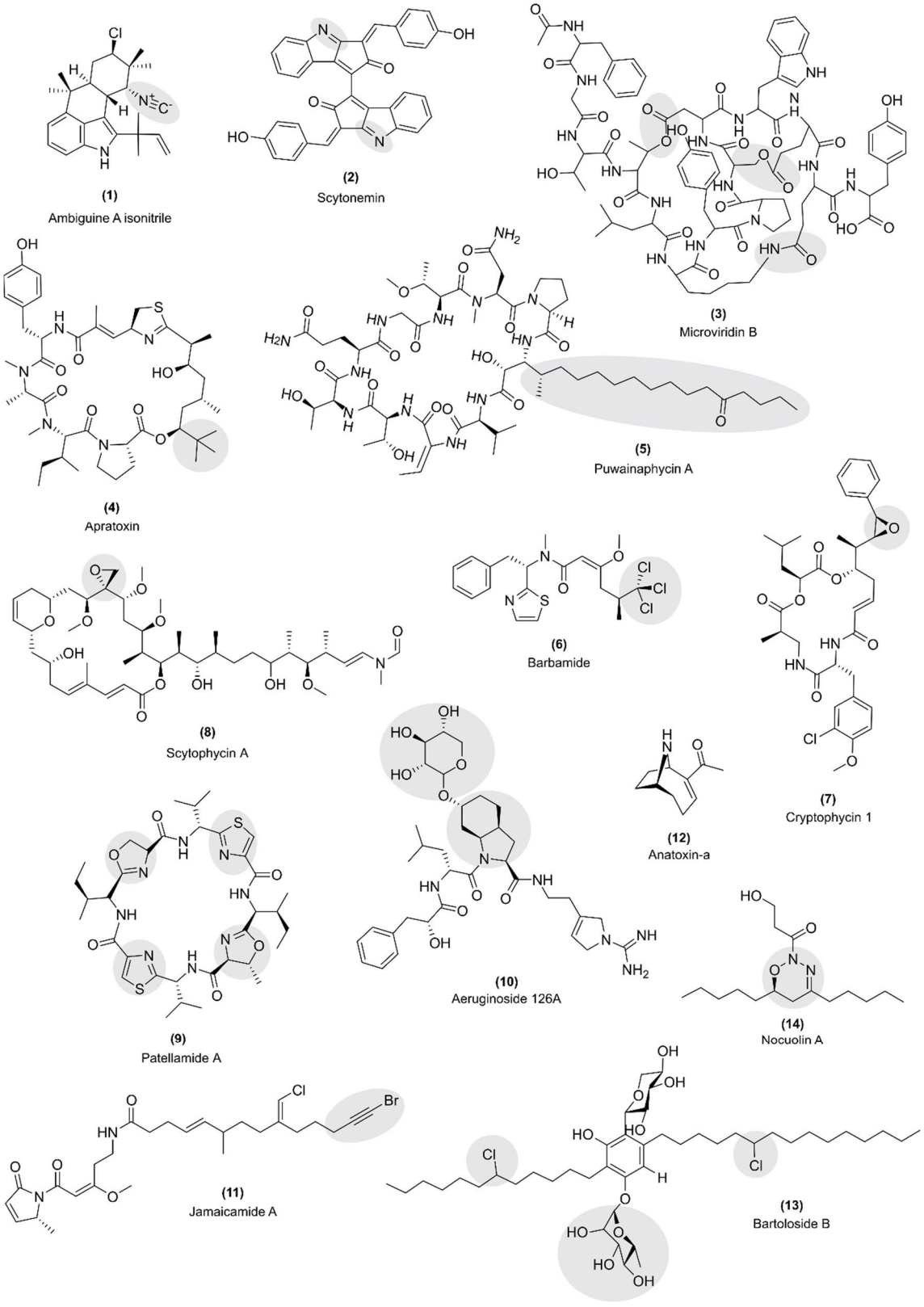


Figure 5 - Selection of cyanobacterial natural products, representing the wide chemodiversity of the phylum. Characteristic structural features are highlighted in gray.

Since an increase of 500 compounds happened in four years, it is only expected that in current times around 2000 cyanobacterial natural products were discovered, evidencing that cyanobacteria are still an endless source to find new chemical scaffolds. The majority of retrieved metabolites from cyanobacteria are from peptidic nature (peptides, depsipeptides and lipopeptides), accounting for 50% of the cyanobacterial-NPs. Besides peptides, cyanobacterial produces several lipids and alkaloids.[75] In comparison with other organisms, cyanobacterial-NPs show a huge structural diversity based on high value molecular weights, complex three-dimensional (3D) structures and stereochemistries, having an unique profile of molecules as highlighted in figure 5.[83]

The genome mining revolution helped to understand the taxonomic distribution of cyanobacterial-NPs by analyzing the biosynthetic genes encoding enzymes from available cyanobacterial genomes.[13,84] The first reports of phylum comparative genomic analysis revealed that late-branches of the cyanobacterial tree of life had a higher frequency of biosynthetic gene clusters, namely the Oscillatorialles, Nostocales and Chroococcales.[5,13] By the possibility of correlating compound families to specific taxa, the orders Oscillatoriales and Nostocales are the largest producers of cyanobacterial-NPs with 47% and 30%, respectively.[75] For Oscillatoriales the major classes of metabolites were peptides (including lipo- and depsipeptides) and macrolides, as exemplified by the natural products **4**, **6** and **10**, together with the very well-known lyngbyatoxin and hectochlorin. For Nostocales, peptides, alkaloids and terpenes were the major classes of natural products, exemplified by **1**, **2**, **3** and **5** but also geosmin and anabaenopeptins. From Chroococcales, compound **3** was isolated, as well as microcystins and aeruginosamides.[75] At genus level, *Lyngbya-Moorena* is responsible for most of metabolites isolated from cyanobacteria, followed by the genus *Nostoc*.[75]

In terms of the biosynthetic potential, it was noted that 70% of the biosynthetic gene clusters found in late-branching strains were assigned to hybrid NRPS/PKS, dedicating 5% of its genome for the production of these pathways.[13] Terpenes and RiPPs biosynthetic pathways were widely distributed among the phylum, and the early-branching cyanobacteria were richer in RiPPs and PKS biosynthetic routes.[84] These results might reflect a bias towards the most accessible and studied strains, since several taxa still remain to be investigated, and bioinformatic tools might not recognize different biosynthetic machinery.[41,85] Nevertheless, it was perceivable that a high percentage of these biosynthetic gene clusters found in cyanobacterial genomes were orphan (~80%), highlighting the potential for the discovery of novel compounds.[13] There is still a lot to discover, but cyanobacteria already demonstrated a huge biosynthetic potential, by

using some intricate and unique enzymes to generate diverse natural products. Some examples are: BrtB, a *β*-propeller enzyme from the bartosolides BGC that performs O-C bond formation using a carboxylate nucleophile and alkyl halide electrophile to form esterified bartolosides (homolog of CylK from cylindrocyclophanes)[86]; the nosperin biosynthetic gene cluster that encodes a *trans*-AT PKS, from a lichen cyanobiont[87]; NocG from nocuolin biosynthetic pathway that has new family of KS domains performing C-C with high selectivity, using as substrate acyl-ACP thioesters[82]; BarE from barbamide biosynthetic pathway that through an *α*-ketothioester decarbonylation reaction shortens the thioester carbon chain by CO removal[88]; nonheme iron-dependent halogenase WelO5 of the hapalindole-like welwitindolinone pathway that acts on free-standing substrates as well as the dimetal-carboxylate halogenases CylC (cylindrocyclophanes) and the homologs ColE (columbamides) and BrtJ (bartolosides), that performs a carbon center activation with cryptic halogenation in the case of cylindrocyclophanes[89]; NosE and NosF from nostopeptolide L1 biosynthetic route, responsible for the nonproteinogenic amino acid 4-methyl-proline[90]; enzymes of the lyngbypeptin B biosynthetic pathway, LynB2 (split module) a MT-ACP didomain that generates (*E*)-3-methoxybut-2-enoate as a unique polyketide starter and, the LynB7 a nonheme diiron oxygenase, that catalyzes the formation of a 2-thiazole ring[91]; finally the PatG macrocyclization enzyme from the cyanobactin patellamides biosynthetic gene cluster that shows high substrate promiscuity[92]. Complementing the examples depicted above, several other cyanobacterial BGCs show unique enzymatic activities such as aeruginosins, lyngbyatoxin, apratoxin and prochlorosins.[74] Genome mining efforts helped with the discovery of such exceptional enzymes which in turn, improves the tool box for synthetic biology approaches.[13] Many of these enzymes are excellent catalysts and serve as inspiration to modify known metabolites or create libraries of semi-synthetic natural products.[74] The huge chemical diversity of cyanobacterial-NPs is attributed mainly to the action of unique tailoring enzymes present in the biosynthetic pathways, as well as the unusual precursors loaded into the biosynthetic machinery.[13,71]

Another interesting fact concerning biosynthetic pathways in cyanobacteria is the presence of mobile elements within or surrounding the gene clusters.[5] Popin and co-workers[84] saw that many pathways of known natural products, such as anabaenopeptins, geosmin and microcystins, were found inside plasmids in late-branched cyanobacteria. Most of the BGCs found in plasmids, were from hybrid assembly lines, and the taxonomic distribution of these was in accordance the overall biosynthetic pathways of cyanobacterial-NPs in late-branching cyanobacteria.[5,84] So, horizontal gene transfer events can be partially responsible for the increased number of BGCs in these

cyanobacterial strains.[13,71] Together with HGT, other genetic events can be enumerated to explain the genetic diversity found in the cyanobacterial biosynthetic pathways throughout the phylum, such as recombination, gene deletion, (in)action of tailoring enzymes and substrate specificity.[71] These events can also be responsible for the synthesis of multiple compound variants within one strain, which is frequently observed in cyanobacterial-NPs. The reason behind this peculiarity is not yet well defined but it can be attributed to synergistic effects of the combined variants, conferring adaptive advantages to the cyanobacterium itself.[71]

Cyanobacterial natural products are known to confer ecological advantages for the producer strain, to thrive and adapt in a wide range of environments.[77,80] While there is not a clear distinction in cyanobacterial-NPs structures and biosynthetic enzymes between the several environments, some notes can be made: the two most explored habitats are the marine and freshwater ones; the marine cyanobacteria strains are more prolific in producing natural products (53%); on the other hand, freshwater metabolites are bigger, more polar and hydrophilic than marine metabolites.[73,75,83] Overall, the scaffolds from marine and freshwater cyanobacterial natural products are different, showing very low overlap between them, but all of them serve the purpose of assisting cyanobacteria survival.[83] Many of these metabolites display a biological activity/function that can be assigned to a specific ecological task, *e.g.* metal chelators siderophores or UV-protecting mycosporin amino acids (MAAs), but in other cases understanding its role is not straightforward.[13,77,93]

Cyanobacteria produce a variety of bioactive metabolites that can have biotechnological of medicinal applications. A wide range of bioactivities are reported from cyanobacterial-NPs, such as antimicrobial (**1**, **5**, **9**), algicidal (cyanobacterin), molluscicidal (**6**), antioxidant (**2**, MAAs), anti-inflammatory (**2,** phycocyanin), anticancer (**4**, **5**, **8**, **12**), multidrug resistance reversion (**10**), antidiabetic (columbamides), anti-viral (**14**), anti-parasitical (carbamin A), immunosuppressive (**13**), protease inhibitor (**3**, **11**) and ion channel (in)activator (**12**).[75-77] The most reported bioactivity from cyanobacterial-NPs is cytotoxicity (anticancer) and lethality. Analysing the chemical scaffolds of the bioactive metabolites, three types of structures stand out, alkaloids, lipopeptides and polyketides.[75] All of these bioactivities are applied in numerous research and technological fields, such as pharmaceutical, cosmetic, food and bioprocess industries.[77] A good example of a commercialized product is the anticancer drug Adcetris, derived from dolastatin 10, a highly potent cyanobacterial cytotoxic compound.[76,94] Also, the powerhouse *Spirulina* can be consumed orally, possessing a high nutritional value, it can

enhance the immune system (antioxidant, antiviral, anticancer activities) and it can even regulate hormones and be used to treat obesity.[76] In agriculture, cyanobacterial-NPs can be used for soil fertilization or pesticides production, and in aquaculture in the production of animal feed.[75,76] Recently, cyanobacteria have also been used in the biofuel industry because of its hydrocarbon production, in bioplastics production using biodegradable polymers, and on colorant dyes leading to environmentally-friendly paints.[75,76]

Although the potential of working with cyanobacteria for the discovery of new active metabolites and enzymes is tremendous, some limitations can occur when working with this diverse phylum. Starting with accessing strains from more inhospitable habitats and transposing them to laboratorial growth, that is being currently overcome by the use of metagenomic sequencing of environmental samples, using the genome guided strategy.[38,73] Another strategy to access taxonomical diversity, is by exploring culture collections that usually comprise strains from several environments and have already optimal growth conditions for each strain.[95] CIIMAR, has an in-house culture collection of cyanobacteria, the LEGE culture collection (LEGEcc), that holds around 400 cyanobacterial strains, isolated from diverse geographies, from both aquatic and terrestrial habitats. In terms of taxonomy, LEGEcc strains are distributed in six orders, including from the late-branched strains with higher potential for new natural products discovery.[96] In the case in which cyanobacterial strains grow well under laboratory conditions, a metabolomic and/or bioassay-guided approach can be followed to access new chemical diversity. Since the yield of isolated natural products from cyanobacteria is usually low, together with the amount of biomass needed, integrating genome and metabolome information with synthetic biology, is a promising strategy to overcome time and yield constrains.[34,71] Nevertheless, expression of cyanobacterial metabolites in heterologous hosts can be challenging because of the genetic differences of the host as well as the difficulty of expressing large-size biosynthetic pathways, as is the case of NRPS and PKS assembly lines.[71,77]

In conclusion, and as discussed above, the current best approach to unravel novel chemistry and enzymology from cyanobacterial natural products is by combining several methodologies and “omics” techniques, to gain knowledge on the ecological relevance of such cryptic information.[13,34,77]

1.4. Objectives

The overall goal of this project is the chemical and biosynthetic exploration of the LEGEcc, which includes cyanobacterial strains phylogenetically predicted to contain a wealth of natural products and also previous bioactivity and analytical data. The purpose is to discover new natural products, their putative biosynthetic gene clusters, looking also to possible pharmaceutical and/or biotechnological applications. The exploration of the LEGEcc had successfully led to the isolation of several unique metabolites, subsequent biosynthetic pathways and demonstrating a range of bioactivities.[96-100] The specific objectives of this project were:

(i) Perform an expeditious, small-scale screening to a group of cyanobacterial strains, to highlight the most promising one's for new natural products isolation;

(ii) Isolate the natural products from strains selected in objective (i);

(iii) Elucidate the structures of isolated compounds using state-of-the-art NMR and MS methodologies;

(iv) Elucidate the biosynthesis of isolated natural products;

(v) Characterize the bioactivity and potency of the isolated natural products, using a variety of assays to reveal important pharmaceutical or biotechnological applications.

1.5. References

1. Shestakov, S. V.; Karbysheva, E. A., The origin and evolution of cyanobacteria. *Biology Bulletin Reviews* **2017,** *7* (4), 259-272.

2. Gaysina, L. A.; Saraf, A.; Singh, P., Chapter 1 - Cyanobacteria in Diverse Habitats. In *Cyanobacteria*, Mishra, A. K.; Tiwari, D. N.; Rai, A. N., Eds. Academic Press: 2019; pp 1-28.

3. Selão, T. T., Exploring cyanobacterial diversity for sustainable biotechnology. *Journal of Experimental Botany* **2022,** *73* (10), 3057-3071.

4. Bergman, B.; Zheng, W.-W.; Klint, J.; Ran, L., On the origin of plants and relations to contemporary cyanobacterial-plant symbioses. *Plant Biotechnology* **2008,** *25* (3), 213-220.

5. Shih, P. M.; Wu, D.; Latifi, A.; Axen, S. D.; Fewer, D. P.; Talla, E.; Calteau, A.; Cai, F.; Tandeau de Marsac, N.; Rippka, R.; Herdman, M.; Sivonen, K.; Coursin, T.; Laurent, T.; Goodwin, L.; Nolan, M.; Davenport, K. W.; Han, C. S.; Rubin, E. M.; Eisen, J. A.; Woyke, T.; Gugger, M.; Kerfeld, C. A., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences* **2013,** *110* (3), 1053-1058.

6. Komárek, J. K., J.; Mares, J.; Johnsen, J., Taxonomic Classification of Cyanoprokaryotes (Cyanobacterial Genera) 2014, Using a Polyphasic Approach. *Preslia* **2014,** *86*, 295-335.

7. Strunecký, O.; Ivanova, A. P.; Mareš, J., An updated classification of cyanobacterial orders and families based on phylogenomic and polyphasic analysis(1). *J Phycol* **2022**.

8. Chang, A. C. G.; Chen, T.; Li, N.; Duan, J., Perspectives on Endosymbiosis in Coralloid Roots: Association of Cycads and Cyanobacteria. *Frontiers in Microbiology* **2019,** *10*.

9. Meeks, J. C., Physiological Adaptations in Nitrogen-fixing Nostoc–PlantSymbiotic Associations. In *Prokaryotic Symbionts in Plants*, Pawlowski, K., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp 181-205.

10. Schirrmeister, B. E.; de Vos, J. M.; Antonelli, A.; Bagheri, H. C., Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proceedings of the National Academy of Sciences* **2013,** *110* (5), 1791-1796.

11. Rippka, R.; Deruelles, J.; Waterbury, J. B.; Herdman, M.; Stanier, R. Y., Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. *Microbiology* **1979,** *111* (1), 1-61.

12. Castenholz, R. W.; Wilmotte, A.; Herdman, M.; Rippka, R.; Waterbury, J. B.; Iteman, I.; Hoffmann, L., Phylum BX. Cyanobacteria. In *Bergey's Manual® of Systematic Bacteriology: Volume One : The Archaea and the Deeply Branching and Phototrophic Bacteria*, Boone, D. R.; Castenholz, R. W.; Garrity, G. M., Eds. Springer New York: New York, NY, 2001; pp 473-599.

13. Calteau, A.; Fewer, D. P.; Latifi, A.; Coursin, T.; Laurent, T.; Jokela, J.; Kerfeld, C. A.; Sivonen, K.; Piel, J.; Gugger, M., Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* **2014,** *15* (1), 977.

14. Brito, A.; Ramos, V.; Seabra, R.; Santos, A.; Santos, C. L.; Lopo, M.; Ferreira, S.; Martins, A.; Mota, R.; Frazão, B.; Martins, R.; Vasconcelos, V.; Tamagnini, P., Culture-dependent characterization of cyanobacterial diversity in the intertidal zones of the Portuguese coast: a polyphasic study. *Syst Appl Microbiol* **2012,** *35* (2), 110-9.

15. Pentecost, A.; Whitton, B. A., Subaerial Cyanobacteria. In *Ecology of Cyanobacteria II: Their Diversity in Space and Time*, Whitton, B. A., Ed. Springer Netherlands: Dordrecht, 2012; pp 291-316.

16. Amar N. Rai, B. B., Ulla Rasmussen, *Cyanobacteria in Symbiosis*. 1 ed.; Springer Dordrecht: New York, Boston, Dordrecht, London, Moscow, 2002; p XI, 355.

17. Hess, W. R., Cyanobacterial genomics for ecology and biotechnology. *Current Opinion in Microbiology* **2011,** *14* (5), 608-614.

18. Brito, Â.; Ramos, V.; Mota, R.; Lima, S.; Santos, A.; Vieira, J.; Vieira, C. P.; Kaštovský, J.; Vasconcelos, V. M.; Tamagnini, P., Description of new genera and species of marine cyanobacteria from the Portuguese Atlantic coast. *Molecular Phylogenetics and Evolution* **2017,** *111*, 18-34.

19. Rego, A.; Fernandez-Guerra, A.; Duarte, P.; Assmy, P.; Leão, P. N.; Magalhães, C., Secondary metabolite biosynthetic diversity in Arctic Ocean metagenomes. *Microb Genom* **2021,** *7* (12).

20. Schorn, M. A.; Jordan, P. A.; Podell, S.; Blanton, J. M.; Agarwal, V.; Biggs, J. S.; Allen, E. E.; Moore, B. S., Comparative Genomics of Cyanobacterial Symbionts Reveals Distinct, Specialized Metabolism in Tropical Dysideidae Sponges. *mBio* **2019,** *10* (3).

21. Teikari, J.; Baunach, M.; Dittmann, E., Cyanobacterial Genome Sequencing, Annotation, and Bioinformatics. In *Engineering Natural Product Biosynthesis: Methods and Protocols*, Skellam, E., Ed. Springer US: New York, NY, 2022; pp 269-287.

22. Estrela, S.; Kerr, B.; Morris, J. J., Transitions in individuality through symbiosis. *Current Opinion in Microbiology* **2016,** *31*, 191-198.

23. Zhang, D.; He, Y.; Gin, K. Y.-H., Novel Freshwater Cyanophages Provide New Insights into Evolutionary Relationships between Freshwater and Marine Cyanophages. *Microbiology Spectrum* **2021,** *9* (2), e00593-21.

24. Mutalipassi, M.; Riccio, G.; Mazzella, V.; Galasso, C.; Somma, E.; Chiarore, A.; de Pascale, D.; Zupo, V., Symbioses of Cyanobacteria in Marine Environments: Ecological Insights and Biotechnological Perspectives. *Mar Drugs* **2021,** *19* (4).

25. Huo, D.; Li, H.; Cai, F.; Guo, X.; Qiao, Z.; Wang, W.; Yu, G.; Li, R., Genome Evolution of Filamentous Cyanobacterium Nostoc Species: From Facultative Symbiosis to Free Living. *Microorganisms* **2021,** *9* (10), 2015.

26. Harke, M. J.; Frischkorn, K. R.; Haley, S. T.; Aylward, F. O.; Zehr, J. P.; Dyhrman, S. T., Periodic and coordinated gene expression between a diazotroph and its diatom host. *Isme j* **2019,** *13* (1), 118-131.

27. Warshan, D.; Espinoza, J. L.; Stuart, R. K.; Richter, R. A.; Kim, S.-Y.; Shapiro, N.; Woyke, T.; C Kyrpides, N.; Barry, K.; Singan, V.; Lindquist, E.; Ansong, C.; Purvine, S. O.; M Brewer, H.; Weyman, P. D.; Dupont, C. L.; Rasmussen, U., Feathermoss and epiphytic Nostoc cooperate differently: expanding the spectrum of plant–cyanobacteria symbiosis. *The ISME Journal* **2017,** *11* (12), 2821-2833.

28. Hashidoko, Y.; Nishizuka, H.; Tanaka, M.; Murata, K.; Murai, Y.; Hashimoto, M., Isolation and characterization of 1-palmitoyl-2-linoleoyl-sn-glycerol as a hormogonium-inducing factor (HIF) from the coralloid roots of Cycas revoluta (Cycadaceae). *Scientific Reports* **2019,** *9* (1), 4751.

29. Liaimer, A.; Jenke-Kodama, H.; Ishida, K.; Hinrichs, K.; Stangeland, J.; Hertweck, C.; Dittmann, E., A polyketide interferes with cellular differentiation in the symbiotic cyanobacterium Nostoc punctiforme. *Environ Microbiol Rep* **2011,** *3* (5), 550-8.

30. Liaimer, A.; Helfrich, E. J. N.; Hinrichs, K.; Guljamow, A.; Ishida, K.; Hertweck, C.; Dittmann, E., Nostopeptolide plays a governing role during cellular differentiation of the symbiotic cyanobacterium Nostoc punctiforme. *Proceedings of the National Academy of Sciences* **2015,** *112* (6), 1862-1867.

31. Gutiérrez-García, K.; Bustos-Díaz, E. D.; Corona-Gómez, J. A.; Ramos-Aboites, H. E.; Sélem-Mojica, N.; Cruz-Morales, P.; Pérez-Farrera, M. A.; Barona-Gómez, F.; Cibrián-Jaramillo, A., Cycad Coralloid Roots Contain Bacterial Communities Including Cyanobacteria and Caulobacter spp. That Encode Niche-Specific Biosynthetic Gene Clusters. *Genome Biology and Evolution* **2018,** *11* (1), 319-334.

32. Harwood, T. V.; Zuniga, E. G.; Kweon, H.; Risser, D. D., The cyanobacterial taxis protein HmpF regulates type IV pilus activity in response to light. *Proc Natl Acad Sci U S A* **2021,** *118* (12).

33. Sorokina, M.; Steinbeck, C., Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics* **2020,** *12* (1), 20.

34. van der Hooft, J. J. J.; Mohimani, H.; Bauermeister, A.; Dorrestein, P. C.; Duncan, K. R.; Medema, M. H., Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society Reviews* **2020,** *49* (11), 3297-3314.

35. Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G., Retrospective analysis of natural products provides insights for future discovery trends. *Proceedings of the National Academy of Sciences* **2017,** *114* (22), 5601-5606.

36. Jensen, P. R., Natural Products and the Gene Cluster Revolution. *Trends Microbiol* **2016,** *24* (12), 968-977.

37. Atanasov, A. G.; Zotchev, S. B.; Dirsch, V. M.; Orhan, I. E.; Banach, M.; Rollinger, J. M.; Barreca, D.; Weckwerth, W.; Bauer, R.; Bayer, E. A.; Majeed, M.; Bishayee, A.; Bochkov, V.; Bonn, G. K.; Braidy, N.; Bucar, F.; Cifuentes, A.; D'Onofrio, G.; Bodkin, M.; Diederich, M.; Dinkova-Kostova, A. T.; Efferth, T.; El Bairi, K.; Arkells, N.; Fan, T.-P.; Fiebich, B. L.; Freissmuth, M.; Georgiev, M. I.; Gibbons, S.; Godfrey, K. M.; Gruber, C. W.; Heer, J.; Huber, L. A.; Ibanez, E.; Kijjoa, A.; Kiss, A. K.; Lu, A.; Macias,

F. A.; Miller, M. J. S.; Mocan, A.; Müller, R.; Nicoletti, F.; Perry, G.; Pittalà, V.; Rastrelli, L.; Ristow, M.; Russo, G. L.; Silva, A. S.; Schuster, D.; Sheridan, H.; Skalicka-Woźniak, K.; Skaltsounis, L.; Sobarzo-Sánchez, E.; Bredt, D. S.; Stuppner, H.; Sureda, A.; Tzvetkov, N. T.; Vacca, R. A.; Aggarwal, B. B.; Battino, M.; Giampieri, F.; Wink, M.; Wolfender, J.-L.; Xiao, J.; Yeung, A. W. K.; Lizard, G.; Popp, M. A.; Heinrich, M.; Berindan-Neagoe, I.; Stadler, M.; Daglia, M.; Verpoorte, R.; Supuran, C. T.; the International Natural Product Sciences, T., Natural products in drug discovery: advances and opportunities. *Nature Reviews Drug Discovery* **2021,** *20* (3), 200-216.

38. Hemmerling, F.; Piel, J., Strategies to access biosynthetic novelty in bacterial genomes for drug discovery. *Nature Reviews Drug Discovery* **2022,** *21* (5), 359-378.

39. Zhao, F.; Sun, C.; Liu, Z.; Cabrera, A.; Escobar, M.; Huang, S.; Yuan, Q.; Nie, Q.; Luo, K. L.; Lin, A.; Vanegas, J. A.; Zhu, T.; Hilton, I. B.; Gao, X., Multiplex Base-Editing Enables Combinatorial Epigenetic Regulation for Genome Mining of Fungal Natural Products. *Journal of the American Chemical Society* **2023,** *145* (1), 413-421.

40. Chen, Y.; Garcia de Lomana, M.; Friedrich, N.-O.; Kirchmair, J., Characterization of the Chemical Space of Known and Readily Obtainable Natural Products. *Journal of Chemical Information and Modeling* **2018,** *58* (8), 1518-1532.

41. Scott, T. A.; Piel, J., The hidden enzymology of bacterial natural product biosynthesis. *Nat Rev Chem* **2019,** *3* (7), 404-425.

42. Ziemert, N.; Alanjary, M.; Weber, T., The evolution of genome mining in microbes - a review. *Nat Prod Rep* **2016,** *33* (8), 988-1005.

43. Balandrin, M. F.; Kinghorn, A. D.; Farnsworth, N. R., Plant-Derived Natural Products in Drug Discovery and Development. In *Human Medicinal Agents from Plants*, American Chemical Society: 1993; Vol. 534, pp 2-12.

44. Newman, D. J.; Cragg, G. M., Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products* **2020,** *83* (3), 770-803.

45. Biermann, F.; Helfrich, E. J. N., Hidden Treasures: Microbial Natural Product Biosynthesis off the Beaten Path. *mSystems* **2021,** *6* (4), e00846-21.

46. Baltz, R. H., Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J Ind Microbiol Biotechnol* **2019,** *46* (3-4), 281-299.

47. Přichystal, J.; Schug, K. A.; Lemr, K.; Novák, J.; Havlíček, V., Structural Analysis of Natural Products. *Analytical Chemistry* **2016,** *88* (21), 10338-10346.

48. Kwan, E. E.; Huang, S. G., Structural Elucidation with NMR Spectroscopy: Practical Strategies for Organic Chemists. *European Journal of Organic Chemistry* **2008,** *2008* (16), 2671-2688.

49. Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y.; Nannapaneni, Y.; Duggan, B. M.; Min, J.; Lin, E. C.; Gerwick, E. C.; Cottrell, G. W.; Gerwick, W. H., Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Scientific Reports* **2017,** *7* (1), 14243.

50. Dührkop, K.; Nothias, L. F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S., Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol* **2021,** *39* (4), 462-471.

51. 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. *Nature Biotechnology* **2016,** *34* (8), 828-837.

52. Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W., NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *Journal of Natural Products* **2021,** *84* (11), 2795-2807.

53. Blin, K.; Shaw, S.; Kloosterman, A. M.; Charlop-Powers, Z.; van Wezel, G. P.; Medema, Marnix H.; Weber, T., antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **2021,** *49* (W1), W29-W35.

54. Terlouw, B. R.; Blin, K.; Navarro-Muñoz, J. C.; Avalon, N. E.; Chevrette, M. G.; Egbert, S.; Lee, S.; Meijer, D.; Recchia, Michael J. J.; Reitz, Zachary L.; van Santen, Jeffrey A.; Selem-Mojica, N.; Tørring, T.; Zaroubi, L.; Alanjary, M.; Aleti, G.; Aguilar, C.; Al-Salihi, Suhad A. A.; Augustijn, Hannah E.; Avelar-Rivas, J A.; Avitia-Domínguez, Luis A.; Barona-Gómez, F.; Bernaldo-Agüero, J.; Bielinski, V. A.; Biermann, F.; Booth, Thomas J.; Carrion Bravo, Victor J.; Castelo-Branco, R.; Chagas, Fernanda O.; Cruz-Morales, P.; Du, C.; Duncan, Katherine R.; Gavriilidou, A.; Gayrard, D.; Gutiérrez-García, K.; Haslinger, K.; Helfrich, Eric J. N.; van der Hooft, Justin J. J.; Jati, Afif P.; Kalkreuter, E.; Kalyvas, N.; Kang, Kyo B.; Kautsar, S.; Kim, W.; Kunjapur, Aditya M.; Li, Y.-X.; Lin, G.-M.; Loureiro, C.; Louwen, Joris J. R.; Louwen, Nico L. L.; Lund, G.; Parra, J.; Philmus, B.; Pourmohsenin, B.; Pronk, Lotte J. U.; Rego, A.; Rex, Devasahayam Arokia B.; Robinson, S.; Rosas-Becerra, L R.; Roxborough, Eve T.; Schorn, Michelle A.; Scobie, Darren J.; Singh, Kumar S.; Sokolova, N.; Tang, X.; Udwary, D.; Vigneshwari, A.; Vind, K.; Vromans, Sophie P. J. M.; Waschulin, V.; Williams, Sam E.; Winter, Jaclyn M.; Witte, Thomas E.; Xie, H.; Yang, D.; Yu, J.; Zdouc, M.; Zhong, Z.; Collemare, J.; Linington, Roger G.; Weber, T.; Medema, Marnix H., MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research* **2022,** *51* (D1), D603-D610.

55. Chan, Y. A.; Podevels, A. M.; Kevany, B. M.; Thomas, M. G., Biosynthesis of polyketide synthase extender units. *Nat Prod Rep* **2009,** *26* (1), 90-114.

56. Wang, J.; Zhang, R.; Chen, X.; Sun, X.; Yan, Y.; Shen, X.; Yuan, Q., Biosynthesis of aromatic polyketides in microorganisms using type II polyketide synthases. *Microbial Cell Factories* **2020,** *19* (1), 110.

57. Hou, L.; Huang, H.; Li, H.; Wang, S.; Ju, J.; Li, W., Overexpression of a type III PKS gene affording novel violapyrones with enhanced anti-influenza A virus activity. *Microbial Cell Factories* **2018,** *17* (1), 61.

58. Wang, J.; Liang, J.; Chen, L.; Zhang, W.; Kong, L.; Peng, C.; Su, C.; Tang, Y.; Deng, Z.; Wang, Z., Structural basis for the biosynthesis of lovastatin. *Nature Communications* **2021,** *12* (1), 867.

59. Malla, S.; Prasad Niraula, N.; Singh, B.; Liou, K.; Kyung Sohng, J., Limitations in doxorubicin production from Streptomyces peucetius. *Microbiological Research* **2010,** *165* (5), 427-435.

60. Nakamura, H.; Hamer, H. A.; Sirasani, G.; Balskus, E. P., Cylindrocyclophane biosynthesis involves functionalization of an unactivated carbon center. *J Am Chem Soc* **2012,** *134* (45), 18518-21.

61. Martínez-Núñez, M. A.; López, V. E. L. y., Nonribosomal peptides synthetases and their applications in industry. *Sustainable Chemical Processes* **2016,** *4* (1), 13.

62. Wenski, S. L.; Thiengmag, S.; Helfrich, E. J. N., Complex peptide natural products: Biosynthetic principles, challenges and opportunities for pathway engineering. *Synthetic and Systems Biotechnology* **2022,** *7* (1), 631-647.

63. Kopp, F.; Mahlert, C.; Grünewald, J.; Marahiel, M. A., Peptide Macrocyclization: The Reductase of the Nostocyclopeptide Synthetase Triggers the Self-Assembly of a Macrocyclic Imine. *Journal of the American Chemical Society* **2006,** *128* (51), 16478-16479.

64. Zhou, C.; Chen, H.; Zhao, H.; Wang, Q., Microcystin biosynthesis and toxic effects. *Algal Research* **2021,** *55*, 102277.

65. Xu, Y.; Kersten, R. D.; Nam, S. J.; Lu, L.; Al-Suwailem, A. M.; Zheng, H.; Fenical, W.; Dorrestein, P. C.; Moore, B. S.; Qian, P. Y., Bacterial biosynthesis and maturation of the didemnin anti-cancer agents. *J Am Chem Soc* **2012,** *134* (20), 8625-32.

66. Chang, Z.; Sitachitta, N.; Rossi, J. V.; Roberts, M. A.; Flatt, P. M.; Jia, J.; Sherman, D. H.; Gerwick, W. H., Biosynthetic Pathway and Gene Cluster Analysis of Curacin A, an Antitubulin Natural Product from the Tropical Marine Cyanobacterium Lyngbya majuscula. *Journal of Natural Products* **2004,** *67* (8), 1356-1367.

67. Montalbán-López, M.; Scott, T. A.; Ramesh, S.; Rahman, I. R.; van Heel, A. J.; Viel, J. H.; Bandarian, V.; Dittmann, E.; Genilloud, O.; Goto, Y.; Grande Burgos, M. J.; Hill, C.; Kim, S.; Koehnke, J.; Latham,

J. A.; Link, A. J.; Martínez, B.; Nair, S. K.; Nicolet, Y.; Rebuffat, S.; Sahl, H.-G.; Sareen, D.; Schmidt, E. W.; Schmitt, L.; Severinov, K.; Süssmuth, R. D.; Truman, A. W.; Wang, H.; Weng, J.-K.; van Wezel, G. P.; Zhang, Q.; Zhong, J.; Piel, J.; Mitchell, D. A.; Kuipers, O. P.; van der Donk, W. A., New developments in RiPP discovery, enzymology and engineering. *Natural Product Reports* **2021,** *38* (1), 130-239.

68. Berdy, B.; Spoering, A. L.; Ling, L. L.; Epstein, S. S., In situ cultivation of previously uncultivable microorganisms using the ichip. *Nature Protocols* **2017,** *12* (10), 2232-2242.

69. Huang, Y.; Sheth, R. U.; Zhao, S.; Cohen, L. A.; Dabaghi, K.; Moody, T.; Sun, Y.; Ricaurte, D.; Richardson, M.; Velez-Cortes, F.; Blazejewski, T.; Kaufman, A.; Ronda, C.; Wang, H. H., High-throughput microbial culturomics using automation and machine learning. *Nature Biotechnology* **2023**.

70. Salvador, L. A.; Biggs, J. S.; Paul, V. J.; Luesch, H., Veraguamides A-G, cyclic hexadepsipeptides from a dolastatin 16-producing cyanobacterium Symploca cf. hydnoides from Guam. *J Nat Prod* **2011,** *74* (5), 917-27.

71. Dittmann, E.; Gugger, M.; Sivonen, K.; Fewer, D. P., Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol* **2015,** *23* (10), 642-652.

72. Pearson, L. A.; Dittmann, E.; Mazmouz, R.; Ongley, S. E.; D'Agostino, P. M.; Neilan, B. A., The genetics, biosynthesis and regulation of toxic specialized metabolites of cyanobacteria. *Harmful Algae* **2016,** *54*, 98-111.

73. Welker, M.; Dittmann, E.; von Döhren, H., Cyanobacteria as a source of natural products. *Methods Enzymol* **2012,** *517*, 23-46.

74. Kehr, J. C.; Gatte Picchi, D.; Dittmann, E., Natural product biosyntheses in cyanobacteria: A treasure trove of unique enzymes. *Beilstein J Org Chem* **2011,** *7*, 1622-35.

75. Demay, J.; Bernard, C.; Reinhardt, A.; Marie, B., Natural Products from Cyanobacteria: Focus on Beneficial Activities. *Mar Drugs* **2019,** *17* (6).

76. Khalifa, S. A. M.; Shedid, E. S.; Saied, E. M.; Jassbi, A. R.; Jamebozorgi, F. H.; Rateb, M. E.; Du, M.; Abdel-Daim, M. M.; Kai, G. Y.; Al-Hammady, M. A. M.; Xiao, J.; Guo, Z.; El-Seedi, H. R., Cyanobacteria-From the Oceans to the Potential Biotechnological and Biomedical Applications. *Mar Drugs* **2021,** *19* (5).

77. Verma, S.; Thapa, S.; Siddiqui, N.; Chakdar, H., Cyanobacterial secondary metabolites towards improved commercial significance through multiomics approaches. *World Journal of Microbiology and Biotechnology* **2022,** *38* (6), 100.

78. Mareš, J.; Hájek, J.; Urajová, P.; Kopecký, J.; Hrouzek, P., A hybrid non-ribosomal peptide/polyketide synthetase containing fatty-acyl ligase (FAAL) synthesizes the β-amino fatty acid lipopeptides puwainaphycins in the Cyanobacterium Cylindrospermum alatosporum. *PLoS One* **2014,** *9* (11), e111904.

79. Micallef, M. L.; D'Agostino, P. M.; Sharma, D.; Viswanathan, R.; Moffitt, M. C., Genome mining for natural product biosynthetic gene clusters in the Subsection V cyanobacteria. *BMC Genomics* **2015,** *16* (1), 669.

80. Chlipala, G. E.; Mo, S.; Orjala, J., Chemodiversity in freshwater and terrestrial cyanobacteria - a source for drug discovery. *Curr Drug Targets* **2011,** *12* (11), 1654-73.

81. Leão, P. N.; Nakamura, H.; Costa, M.; Pereira, A. R.; Martins, R.; Vasconcelos, V.; Gerwick, W. H.; Balskus, E. P., Biosynthesis-Assisted Structural Elucidation of the Bartolosides, Chlorinated Aromatic Glycolipids from Cyanobacteria. *Angewandte Chemie International Edition* **2015,** *54* (38), 11063-11067.

82. Martins, T. P.; Glasser, N. R.; Kountz, D. J.; Oliveira, P.; Balskus, E. P.; Leão, P. N., Biosynthesis of the Unusual Carbon Skeleton of Nocuolin A. *ACS Chemical Biology* **2022,** *17* (9), 2528-2537.

83. Saldívar-González, F. I.; Valli, M.; Andricopulo, A. D.; da Silva Bolzani, V.; Medina-Franco, J. L., Chemical Space and Diversity of the NuBBE Database: A Chemoinformatic Characterization. *J Chem Inf Model* **2019,** *59* (1), 74-85.

84. Popin, R. V.; Alvarenga, D. O.; Castelo-Branco, R.; Fewer, D. P.; Sivonen, K., Mining of Cyanobacterial Genomes Indicates Natural Product Biosynthetic Gene Clusters Located in Conjugative Plasmids. *Frontiers in Microbiology* **2021,** *12*.

85. Welker, M.; von Döhren, H., Cyanobacterial peptides - nature's own combinatorial biosynthesis. *FEMS Microbiol Rev* **2006,** *30* (4), 530-63.

86. Reis, J. P. A.; Figueiredo, S. A. C.; Sousa, M. L.; Leão, P. N., BrtB is an O-alkylating enzyme that generates fatty acid-bartoloside esters. *Nature Communications* **2020,** *11* (1), 1458.

87. Kampa, A.; Gagunashvili, A. N.; Gulder, T. A. M.; Morinaka, B. I.; Daolio, C.; Godejohann, M.; Miao, V. P. W.; Piel, J.; Andrésson, Ó. S., Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. *Proceedings of the National Academy of Sciences* **2013,** *110* (33), E3129-E3137.

88. Guo, S.; Sang, Y.; Zheng, C.; Xue, X.-S.; Tang, Z.; Liu, W., Enzymatic α-Ketothioester Decarbonylation Occurs in the Assembly Line of Barbamide for Skeleton Editing. *Journal of the American Chemical Society* **2023,** *145* (9), 5017-5028.

89. Eusebio, N.; Rego, A.; Glasser, N. R.; Castelo-Branco, R.; Balskus, E. P.; Leão, P. N., Distribution and diversity of dimetal-carboxylate halogenases in cyanobacteria. *BMC Genomics* **2021,** *22* (1), 633.

90. Liu, L.; Jokela, J.; Herfindal, L.; Wahlsten, M.; Sinkkonen, J.; Permi, P.; Fewer, D. P.; Døskeland, S. O.; Sivonen, K., 4-Methylproline Guided Natural Product Discovery: Co-Occurrence of 4-Hydroxy- and 4-Methylprolines in Nostoweipeptins and Nostopeptolides. *ACS Chemical Biology* **2014,** *9* (11), 2646-2655.

91. Kudo, F.; Chikuma, T.; Nambu, M.; Chisuga, T.; Sumimoto, S.; Iwasaki, A.; Suenaga, K.; Miyanaga, A.; Eguchi, T., Unique Initiation and Termination Mechanisms Involved in the Biosynthesis of a Hybrid Polyketide-Nonribosomal Peptide Lyngbyapeptin B Produced by the Marine Cyanobacterium Moorena bouillonii. *ACS Chemical Biology* **2023**.

92. Sarkar, S.; Gu, W.; Schmidt, E. W., Expanding the chemical space of synthetic cyclic peptides using a promiscuous macrocyclase from prenylagaramide biosynthesis. *ACS Catal* **2020,** *10* (13), 7146-7153.

93. Nandagopal, P.; Steven, A. N.; Chan, L.-W.; Rahmat, Z.; Jamaluddin, H.; Mohd Noh, N. I., Bioactive Metabolites Produced by Cyanobacteria for Growth Adaptation and Their Pharmacological Properties. *Biology* **2021,** *10* (10), 1061.

94. Tan, L. T., Pharmaceutical agents from filamentous marine cyanobacteria. *Drug Discov Today* **2013,** *18* (17-18), 863-71.

95. Brito, Â.; Vieira, J.; Vieira, C. P.; Zhu, T.; Leão, P. N.; Ramos, V.; Lu, X.; Vasconcelos, V. M.; Gugger, M.; Tamagnini, P., Comparative Genomics Discloses the Uniqueness and the Biosynthetic Potential of the Marine Cyanobacterium Hyella patelloides. *Frontiers in Microbiology* **2020,** *11*.

96. Ramos, V.; Morais, J.; Castelo-Branco, R.; Pinheiro, Â.; Martins, J.; Regueiras, A.; Pereira, A. L.; Lopes, V. R.; Frazão, B.; Gomes, D.; Moreira, C.; Costa, M. S.; Brûle, S.; Faustino, S.; Martins, R.; Saker, M.; Osswald, J.; Leão, P. N.; Vasconcelos, V. M., Cyanobacterial diversity held in microbial biological resource centers as a biotechnological asset: the case study of the newly established LEGE culture collection. *Journal of Applied Phycology* **2018,** *30* (3), 1437-1451.

97. Gutiérrez-del-Río, I.; Brugerolle de Fraissinette, N.; Castelo-Branco, R.; Oliveira, F.; Morais, J.; Redondo-Blanco, S.; Villar, C. J.; Iglesias, M. J.; Soengas, R.; Cepas, V.; Cubillos, Y. L.; Sampietro, G.; Rodolfi, L.; Lombó, F.; González, S. M. S.; López Ortiz, F.; Vasconcelos, V.; Reis, M. A., Chlorosphaerolactylates A–D: Natural Lactylates of Chlorinated Fatty Acids Isolated from the Cyanobacterium Sphaerospermopsis sp. LEGE 00249. *Journal of Natural Products* **2020,** *83* (6), 1885-1890.

98. Figueiredo, S. A. C.; Preto, M.; Moreira, G.; Martins, T. P.; Abt, K.; Melo, A.; Vasconcelos, V. M.; Leão, P. N., Discovery of Cyanobacterial Natural Products Containing Fatty Acid Residues**. *Angewandte Chemie* **2021,** *133* (18), 10152-10160.

99. Freitas, S.; Silva, N. G.; Sousa, M. L.; Ribeiro, T.; Rosa, F.; Leão, P. N.; Vasconcelos, V.; Reis, M. A.; Urbatzka, R., Chlorophyll Derivatives from Marine Cyanobacteria with Lipid-Reducing Activities. *Mar Drugs* **2019,** *17* (4).

100. Antunes, J.; Pereira, S.; Ribeiro, T.; Plowman, J. E.; Thomas, A.; Clerens, S.; Campos, A.; Vasconcelos, V.; Almeida, J. R., A Multi-Bioassay Integrated Approach to Assess the Antifouling Potential of the Cyanobacterial Metabolites Portoamides. *Marine Drugs* **2019,** *17* (2), 111.

Chapter II

Metabolomic and bioactivity screening of cyanobacteria from LEGE culture collection

2. Metabolomic and bioactivity screening of cyanobacteria from LEGE culture collection

2.1. Introduction

Cyanobacteria are a group of photosynthetic bacteria widespread throughout Earth, being able to adapt to grow in a variety of conditions and colonizing multiple habitats.[1] Their plasticity is also observed in their ability to produce a large diversity of natural products.[2] Many of these compounds show potent bioactivities, as illustrated by the cyanobacterial metabolite dolastatin 10,[3] which served as inspiration for the development of powerfull cancer drugs, available in the market.[4] The unprecedent chemical diversity present in the metabolites discovered from cyanobacteria is the result of the biosynthetic machinery responsible for their production.[5] The biosynthetic pathways leading to known bioactive cyanobacterial natural products are mainly of the ribosomal synthetized and post-translational modified (RiPPs), non-ribosomal peptide (NRPs), polyketide (PK), and hybrids of the latter.[6] In fact, most of the known biosynthetic gene clusters (BGCs) – i.e. group of clustered genes responsible for metabolite production, resistance and transport – of cyanobacterial origin are PKS/NRPS hybrids, which result in a considerable diversity of chemical scaffolds.[7]

However, to access such diverse cyanobacteria is something difficult depending on the environments on which they grow, and culturing these organisms under laboratory conditions might be also a challenge.[8] Consequently, most of the studied cyanobacterial strains are those which grow abundantly in the natural environment and/or easily accessed, which in turn leads to a rediscovery rate of the same natural products and biosynthetic pathways.[9] One way to tackle this problem is the use of sequencing techniques directly on environmental samples to recover (meta)genomic information, and then use synthetic biology to express promising biosynthetic pathways in heterologous hosts in order to discover new natural products.[9] Although some successful cases led to the discovery of new chemical diversity from unexplored cyanobacteria, heterologous expression of BGCs is sometimes difficult and ineffective.[6] So, an alternative route to explore diverse and unprecedent cyanobacteria is using the resources of culture collections, since usually they comprise a large diversity of strains from different ecosystems, adapted to laboratorial growth conditions.[10]

The Interdisciplinary Center of Marine and Environmental research (CIIMAR) has an in-house cyanobacteria culture collection, LEGEcc[11], with ca. 400 strains from aquatic

(marine, freshwater and estuarine) and terrestrial (symbiotic or non-symbiotic) environments. Most of these cyanobacterial strains were collected from Portuguese habitats (84%), comprising a huge morphological and phylogenetic diversity, which is unexplored at many levels, including their chemistry. Nevertheless, this collection also has specimens from other continents (American, Africa, Europe, Oceania, Antarctica, and Asia). In terms of habitats, most of the cyanobacterial strains are from aquatic habitats (marine, freshwater, brackish and hypersaline) but it also contains strains isolated from terrestrial habitats (3%). The phylogenetic distribution of the cyanobacterial strains are mainly on the Synechococcales, Chroococcales, Nostocales, Oscillatoriales, Pleurocapsales and Croococcidiopsidales orders, which provides a huge opportunity for new natural products discovery.[11] According to a phylogenomic analysis linking cyanobacteria phylogenetic placement with the biosynthetic potential for natural products production, the late branched groups are the most biosynthetically diverse and a great part of the BGCs on those cyanobacterial orders were not linked to any known metabolite, being considered orphan[7]. This provides a huge opportunity to discover new metabolites and enzymes from these cyanobacterial orders.[12] Also, cyanobacterial strains from the LEGEcc have led to the discovery of bioactive secondary metabolites, as well as unique enzymes.[11,13-15] For example, the discovery of portoamides[13] and bartolosides[16,17] from LEGEcc cyanobacterial strains proved not only the chemical diversity of this collection but also its untapped enzymology. In order to discover new natural products and the respective biosynthetic pathways, a selection of unexplored LEGEcc strains was performed taking into account their phylogenetic placement in natural products-rich clades[11], as well as selection of cyanobacterial strains with previously reported bioactivity and/or metabolomic data but with no isolated metabolite and/or BGC.[15,18-21] To help prioritize the strains for large-scale cultivation, and downstream metabolite isolation, a metabolomic screening was performed for all selected strains. In addition, the group of strains selected based on phylogenetic placement underwent a biological activity screening. Using this metabolomics approach makes it possible to performed dereplication prior to compound isolation[10], and merging this information with bioactivity screening provides a framework of prioritized strains for bioactive natural products discovery.

2.2. Results and discussion

The cyanobacterial strains selected from the LEGEcc[11] to be explored for the discovery of new natural products and biosynthetic pathways were divided in two groups,

unexplored cyanobacteria (UC) and explored cyanobacteria (EC), as shown in Table 1. The UC group corresponded to phylogenetically unique cyanobacterial strains found in clades known to be prolific in natural products; in particular, from Oscillatoriales and Nostocales orders, encompassing 14 cyanobacterial strains (Table 1, Figure S1). The EC group comprises 12 cyanobacterial strains with already demonstrated bioactivity in past studies, and/or with metabolomic information suggesting the presence of novel natural products, but for which no follow-up studies were performed (Table 1). Both groups were screened to prioritize strains to explore. For UC strains a bioactivity and metabolomic screening was applied using anticancer, antimicrobial bioassays and GNPS molecular networking[22], respectively, while for EC only a metabolomic screening was performed using also GNPS molecular networking. Detailed information of all the cyanobacterial strains used in this screening effort is shown in the supplementary information (Table S1).

Table 1 – LEGEcc cyanobacterial strains selected for the screening assay. The order was updated according to the latest phylogenetic classification[23]. EC – explored cyanobacteria; UC – unexplored cyanobacteria; *metabolomic data.

Cyanobacterial strain	Order	Group
Unidentified filamentous Synechococcales LEGE 06070	Synechococcales	EC[18-20]
cf. *Oxynema acuminatum* LEGE 06072	Oscillatoriales	EC[18-20]
Toxifilum mysidocida LEGE 06108	Oculatellales	EC[21]
Unidentified filamentous Synechococcales LEGE 06118	Synechococcales	EC*
Leptolyngbya Saxicola LEGE 06131	Leptolyngbyales	EC*
Myxosarcina sp. LEGE 06146	Pleurocapsales	UC
Unidentified Pleurocapsales LEGE 06147	Pleurocapsales	UC
Unidentified Nostocales LEGE 06150	Nostocales	EC*
Chroococcidiopsis sp. LEGE 06174	Chroococcidiopsidales	EC*
Calenema singularis LEGE 06188	Oculatellales	EC*
Microcoleus sp. LEGE 07076	Oscillatoriales	EC[18-20]
Coleofasciculus sp. LEGE 07092	Coleofasciculales	EC[18-20]
Scytonema sp. LEGE 07182	Nostocales	EC*
Nostoc cf. *edaphicum* LEGE 07299	Nostocales	EC*
Unidentified Nostocales LEGE 11386	Nostocales	UC
Cymatolege sp. LEGE 11434	Nodosilineales	UC
Nodosilinea sp. LEGE 11468	Nodosilineales	UC
cf. *Phormidesmis* sp. LEGE 11477	Leptolyngbyales	UC
Romeriopsis navalis LEGE 11480	Synechococcales	UC
Desmonostoc muscorum LEGE 12446	Nostocales	UC
Nostoc sp. LEGE 12447	Nostocales	UC
Nostoc sp. LEGE 12450	Nostocales	UC
Nostoc sp. LEGE 12451	Nostocales	UC
Nostoc sp. LEGE 12454	Nostocales	UC
Nostoc sp. LEGE 12456	Nostocales	UC
Unidentified Nostocales LEGE 14445	Nostocales	UC

2.2.1. Unexplored cyanobacterial LEGEcc strains (UC)

Since no information was available for the cyanobacterial strains present in this group, a bioactivity and metabolomic screening was performed. The biological screening was carried out with anticancer and antimicrobial assays and the metabolomic characterization was performed using GNPS molecular networking. For that, cyanobacterial strains were grown in small scale (4 L) and the freeze-dried biomass

(Table S2) was extracted with CH_2Cl_2/MeOH (2:1, v/v) by repeated percolation. Each organic extract was further fractionated in three fractions of increasing polarity (*n*-hexane [A], ethyl acetate [B] and methanol [C]), using a pre-packed solid-phase extract (SPE) silica cartridge, to decrease the complexity of the crude extract. These fractions were prepared at the desired concentrations and exposed to the selected cell lines and microbial strains. For the anticancer cytotoxic assays, a fraction (tested at 30 μg mL^{-1} concentration) was considered bioactive if a reduction in cell viability of >60% was observed (Figure S2, S3); this criterion yielded bioactive fractions in three target cell lines, namely breast carcinoma (T47D) in Figure 1, neuroblastoma (SHSY-S5) in Figure 2 and osteosarcoma (MG-63) in Figure 3. The cyanobacterial strains whose fractions demonstrated bioactivity were: *Nodosilinea* sp. LEGE 11468 (fr. A); unidentified Nostocales LEGE 11386 (fr. A and B); unidentified Nostocales LEGE 14445 (fr. A and B), unidentified Pleurocapsales LEGE 06147 (fr. A and B); *Nostoc* sp. LEGE 12447 (fr. A and B); *Nostoc* sp. LEGE 12451 (fr. A); *Nostoc* sp. LEGE 12454 (fr. A); *Nostoc* sp. 12456 (fr. A) and *Nostoc* sp. LEGE 12450 (fr. A).

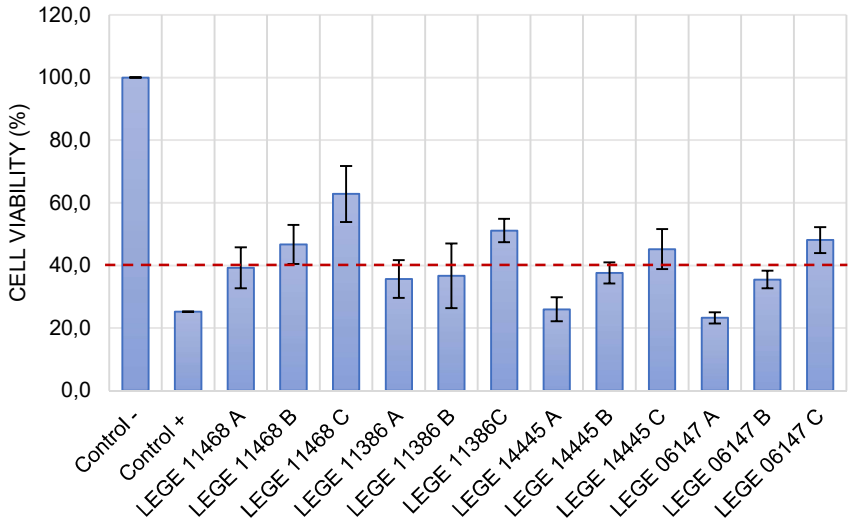


Figure 1 – Cell viability assays with the breast carcinoma cell line, T47D, showing the decrease in cell viability of >60% of the cyanobacterial fraction A from *Nodosilinea* sp. LEGE 11468, fractions A and B from unidentified Nostocales LEGE 11386, fractions A and B from unidentified Nostocales LEGE 14445 and fractions A and B from unidentified Pleurocapsales LEGE 06147. All fractions are at a concentration of 30 µg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 µg mL^{-1}).

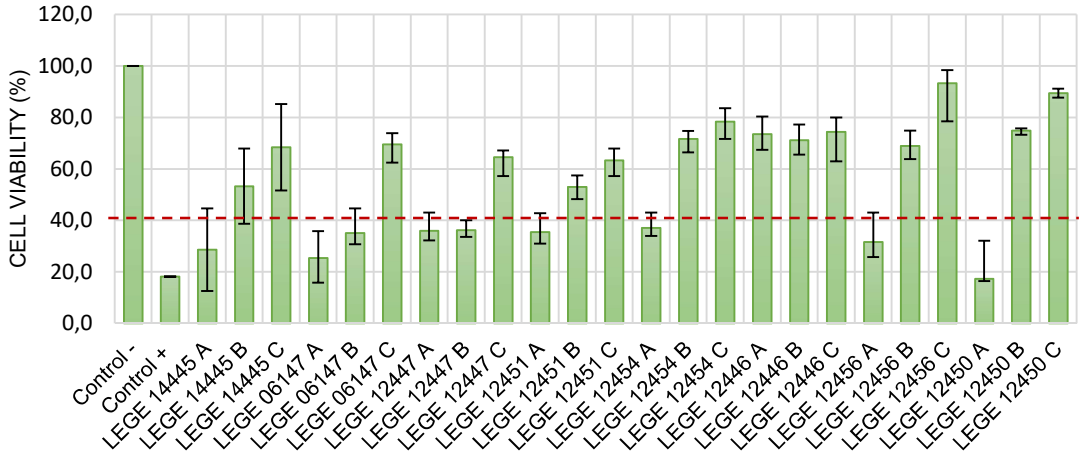


Figure 2 - Cell viability assays with the neuroblastoma cell line, SYSH-S5, showing the decrease in cell viability of >60% of the cyanobacterial fraction A from unidentified Nostocales LEGE 14445, fractions A and B from unidentified Pleurocapsales LEGE 06147, fractions A and B from *Nostoc* sp. LEGE 12447 and the A fraction of *Nostoc* sp. LEGE 12451, *Nostoc* sp. LEGE 12454, *Nostoc* sp. LEGE 12456 and *Nostoc* sp. LEGE 12450. All fractions are at a concentration of 30 µg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 µg mL^{-1}).

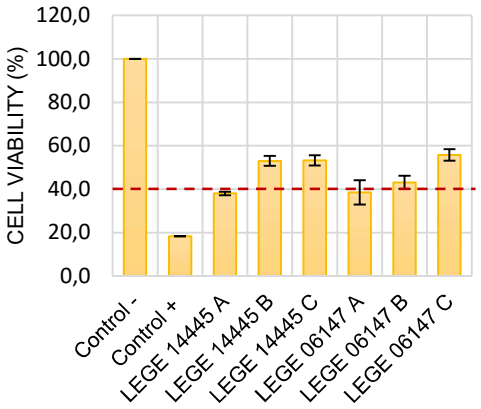


Figure 3 - Cell viability assays with the osteosarcoma cell line, MG-63, showing the decrease in cell viability of >60% of the cyanobacterial fraction A from unidentified Nostocales LEGE 14445 and fraction A from unidentified Pleurocapsales LEGE 06147. All fractions are at a concentration of 30 µg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 µg mL^{-1}).

Considering the cytotoxicity screening results (Figures 1-3), it is clear that fraction A (*n*-hex) tended to show more cytotoxic activity, when compared to the others. Also, marine and symbiotic cyanobacteria were found to be the most active strains. The fraction with the strongest activity was fraction A from the symbiotic *Nostoc* sp. LEGE 12450. Both marine and symbiotic cyanobacteria are known to produce natural products with UV protecting properties, such as scytonemin[24], and these compounds are of

alkaloid nature (less polar), eluting better with the non-polar solvents, such as that used for fraction A (*n*-hex). Usually, these antioxidant natural products, as well as other bacterial pigments, also demonstrate antiproliferative effects towards cancer cell lines[25,26], and these could justify the observed activities.

The antimicrobial assays were performed at a concentration of 15 μg mL^{-1} with two Gram-negative, two Gram-positive bacteria and one fungal strain (Figure S4). Growth inhibition halos (± 1 mm) were observed for the four bacterial strains, the Gram-positives *Staphylococcus aureus* (*S. aureus*) and *Bacillus subtilis* (*B. subtilis*) and the Gram-negatives *Salmonella typhimurium* (*S. typhimurium*) and *Escherichia coli* (*E. coli*). The cyanobacterial strains showing bioactivity in these assays were: unidentified Pleurocapsales LEGE 06147 fraction B and C (Figure 4A); *Nostoc* sp. LEGE 12447 fractions B and C (Figure 4B); *Desmonostoc muscorum* LEGE 12446 fraction B and C (Figure 4C); *Nostoc* sp. LEGE 12450 fractions A and B (Figure 4D) and *Cymatolege* sp. LEGE 11434 fraction C (Figure 4E). Although the inhibition of bacterial cell growth was mild, the cyanobacterial strains *Desmonostoc muscorum* LEGE 12446, *Nostoc* sp. LEGE 12450 and *Cymatolege* sp. LEGE 11434 show a clear inhibition halo.

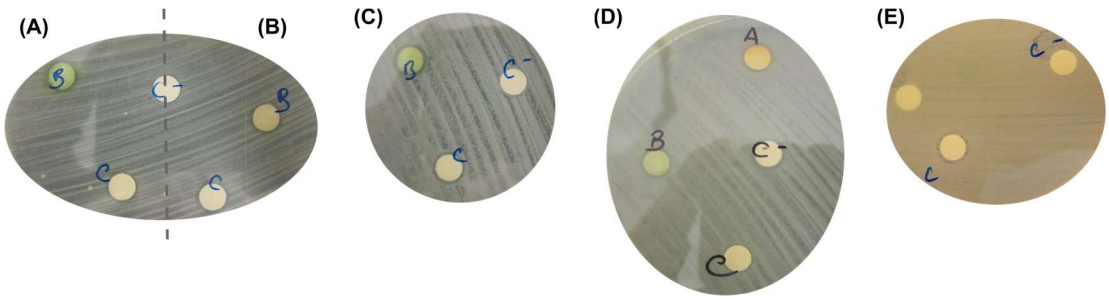


Figure 4 – Antimicrobial assays performed with the disc diffusion method. **(A)** Inhibition halos of *S. aureus* with fractions B and C from unidentified Pleurocapsales LEGE 06147; **(B)** Inhibition halos of *S. aureus* with fractions B and C from *Nostoc* sp. LEGE 12447**; (C)** Inhibition halos of *B. subtilis* by fractions B and C from *Desmonostoc muscorum* LEGE 12446; **(D)** Inhibition halos of *S. typhimurium* by fractions A and B from *Nostoc* sp. LEGE 12450; **(E)** Inhibition halo of *E. coli* by fraction C of *Cymatolege* sp. LEGE 11434. Exposure of the bacterial strains with 15 µg of each fraction; Negative control corresponds to 100% DMSO.

Unlike what was observed in the cytotoxicity assays, in the antimicrobial assays, methanolic fractions (fraction C) proved to be the most active. Bacteria are known to produce antibiotic compounds of peptidic nature (*e.g.* vancomycin) being more polar, such as those eluting with the solvents used in fraction C. Cyanobacteria also have some

peptidic natural products with antibacterial (*e.g.* kawaguchipeptin A), as well as antifungal activities (hassalidins)[27].

To gain some insights into the metabolomic profile of the UC strains an untargeted metabolomics analysis using GNPS molecular networking[22] was performed. All the cyanobacterial fractions were analyzed by LC-HRESIMS/MS and submitted to the analysis pipeline. The entire network was examined with Cytoscape[28] (Figure S5) to map individual clusters specific to a single strain, in order to discover new unique metabolites. Strains were also grouped by taxonomic Order in the analysis. Dereplication efforts were performed in order to identify known natural products in the tested fractions, using the tools GNPS Dereplicator[29], Dereplicator VarQuest[30] and Dereplicator+[31], as well as manual searches in the databases NPAtlas[32] and Dictionary of Natural Products (DNP)[33]. Only the most abundant mass features (≥ $1x10^7$) were selected for analysis, since those would be the ones eventually followed-up for natural product isolation (i.e., such intensity values would roughly indicate sufficient amounts of the target compound for isolation and structure elucidation). All the GNPS dereplication hits were manually inspected in the LC-HRESIMS/MS data, and the cyanobacterial hits were filtered according to the score value (≥ 10), the *p*-value (≤ $1.0E^{-10}$) and the FDR % (false discovery rate, = 0) (i.e., one of these conditions had to be met for the dereplication hit to be considered). Strain-specific molecular clusters (i.e. present only in a particular strain) were found for the UC Nostocales strains *Desmonostoc muscorum* LEGE 12446; *Nostoc* sp. LEGE 12450; *Nostoc* sp. LEGE 12451; *Nostoc* sp. LEGE 12454; *Nostoc* sp. LEGE 12456; and unidentified Nostocales LEGE 14445, as shown in Table 2, the Nodosilineales strains *Nodosilinea* sp. LEGE 11486; *Cymatolege* sp. LEGE 11434 (Table 3), and one strain of Pleurocapsales (*Myxosarcina* sp. LEGE 06146), Synechococcales (*Romeriopsis navalis* LEGE 11480) and Leptolyngbyales (cf. *Phormidesmis* sp. LEGE 11477), as shown in Table 4.

For *Desmonostoc muscorum* LEGE 12446, 11 unique mass features were present in the molecular clusters, from fractions B and C, as shown in Figure 5 and Table 2. Two features were dereplicated as anabaenopeptins with NPAtlas and two other were dereplicated as primary lipids with DNP. Regarding the GNPS dereplication tools, no hit was observed for these 11 mass features with Dereplicator and Dereplicator+. On the other hand, Dereplicator VarQuest retrieved 26 hits for cyanobacterial natural products, most of all based on the $[M+2H]^{2+}$ ions. Inspection of those ions on the LC-HRESIMS/MS data showed that most of the hits were related to the mass features present in the molecular clusters, and not to the known compounds annotated with dereplicator

VarQuest (Table S3). Further analysis of the results from dereplicator VarQuest led to the discovery of two new abundant masses (not present in the molecular clusters for this cyanobacterium), that were dereplicated using NPAtlas as the known metabolites anabaenopeptin 863a (or SA8) and anabaenopeptin 848 (or SA7) (Table S3).

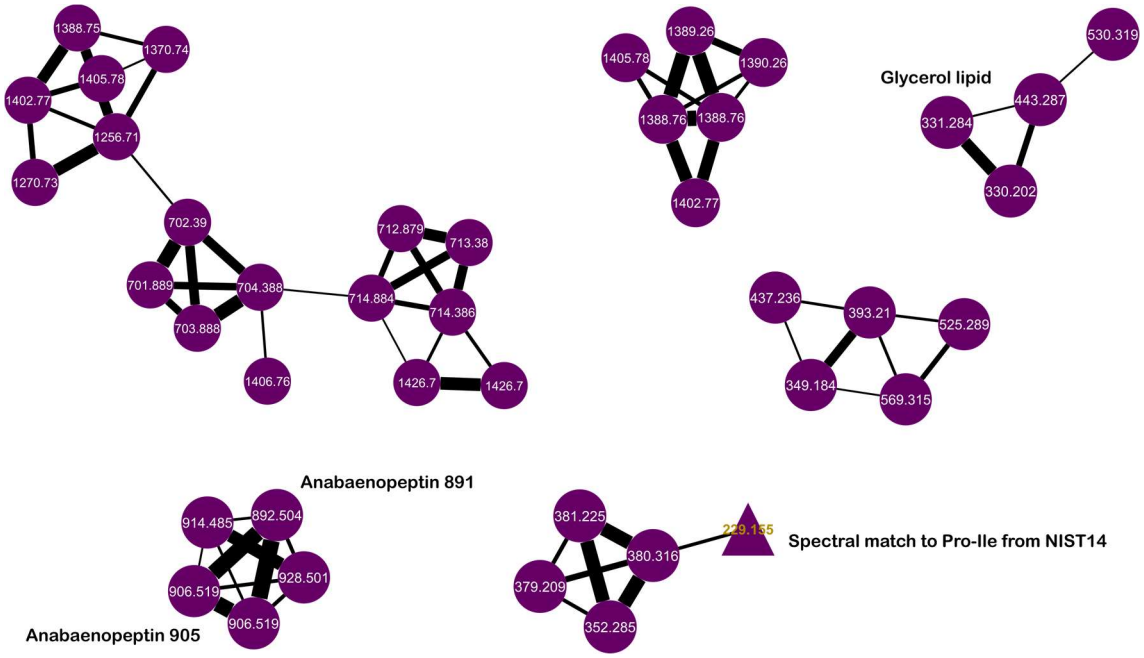


Figure 5 – Main molecular clusters from *Desmonostoc muscorum* LEGE 12446, retrieved from the GNPS molecular networking analysis. Annotation of anabaenopeptin 891 and 905 (and respective adducts) with NPAtlas, of a glycerol lipid using Dictionary of Natural Products (DNP) and a spectral match to Pro-Ile from NIST14 (NIST standard reference database).

For *Nostoc* sp. LEGE 12450, five intense mass features were present in the molecular clusters, from fraction C, and did not show any cyanobacterial hit on the dereplication using NPAtlas and DNP (Table 2, Figure 6). Dereplicator VarQuest listed seven hits, again all based on the $[M+2H]^{2+}$ ions and were not considered (Table S4). Finally, with Dereplicator+, diglycolipid 14 was annotated in the LC-HRESIMS/MS spectrum of this cyanobacterium (Table S4). The mass features of 1007 Da and 1077 Da show a clear loss of 79.96 amu (anatomic mass units), that can correspond to a presence of a phosphate or sulfate group in the structure.

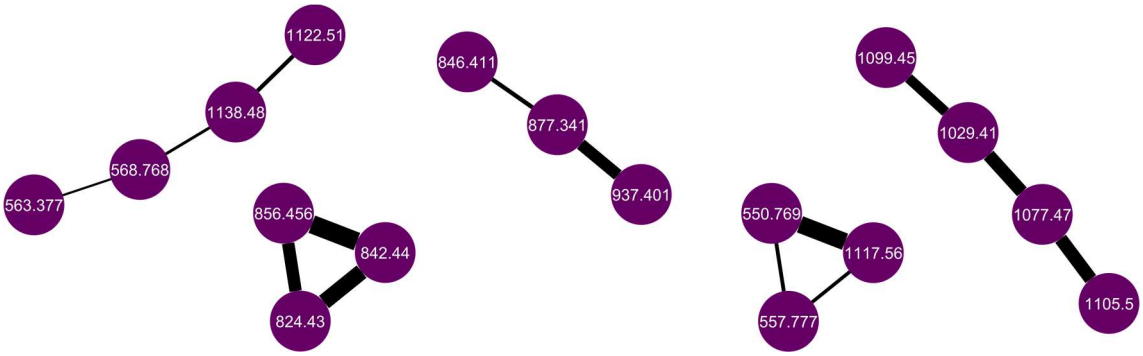


Figure 6 - Main molecular clusters from *Nostoc* sp. LEGE 12450, retrieved from the GNPS molecular networking analysis.

The *Nostoc* sp. LEGE 12451 strain showed three intense mass features in the molecular clusters, all in fraction C, and the only dereplication hits were with NPAtlas and Dereplicator VarQuest (Table S5). NPAtlas dereplicated microcystin-HphHph and with Dereplicator VarQuest all four hits were based on the $[M+2H]^{2+}$ adducts but an inspection of the annotated MS2 spectra showed that these were not correspondent to the masses retrieved from the molecular clusters. The compound with mass 915 Da has a neutral loss of 79.96 amu, as observed for the compounds of *Nostoc* sp. LEGE 12450 highlighted above.

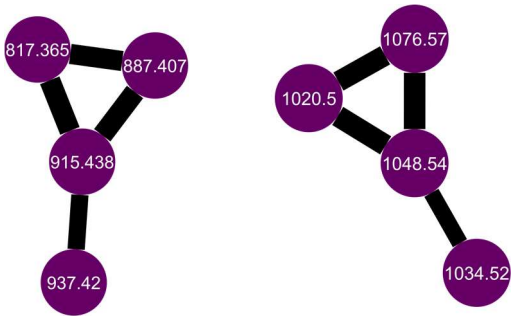


Figure 7 - Main molecular clusters from *Nostoc* sp. LEGE 12451, retrieved from the GNPS molecular networking analysis.

In contrast to the other symbiotic strains, *Nostoc* sp. LEGE 12454 showed more abundant mass features from fraction A than B and C, having in total 12 abundant masses (Table 2). In terms of dereplication, in fraction C two hits were retrieved using Dereplicator VarQuest, related with microcystins, but a quick search on the MS2 data confirmed that these compounds were not the same present in this cyanobacterium, as fragmentations were not matching (Table S6). The calculated molecular formula (M.F.) (Δppm 0.189) of *m/z* 391.2844 $[M+H]^+$ was searched on Lipid Maps[34], and resulted in the

identification of the sterol lipid 24:2. Using DNP, vitamin K1 5' S-hydroxy was annotated for this cyanobacterium and lastly, Dereplicator+ retrieved two hits both in fraction A, trihydroxy-7',8'-didehydro-§-carotene (*m/z* 583.410 $[M+H]^+$) and 7,8-dihydrodiadinoxanthin (*m/z* 585.430 $[M+H]^+$). Both these masses were not present in the molecular clusters retrieved from the molecular networking analysis for this cyanobacterium (Figure 8).

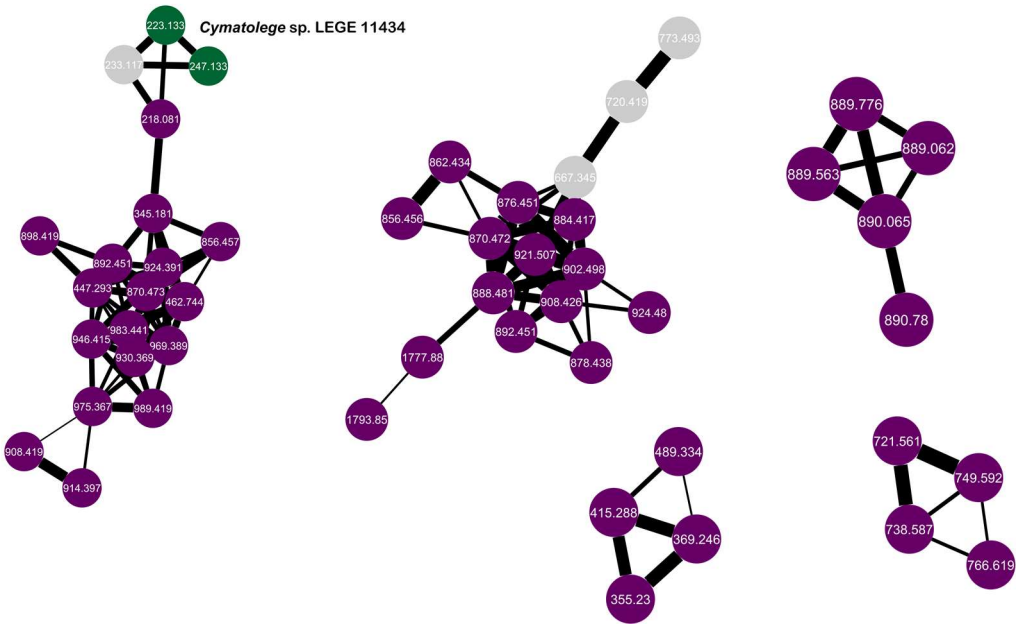


Figure 8 - Main molecular clusters from *Nostoc* sp. LEGE 12454, retrieved from the GNPS molecular networking analysis. The first cluster has a connection with masses also present in *Cymatolege* sp. LEGE 11434 (green), and with a mass feature present in several other UC strains (grey), which were not considered.

Micropeptin C was annotated in the LC-HRESIMS data of *Nostoc* sp. LEGE 12456, using NPAtlas, DNP, Dereplicator VarQuest and Dereplicator+, being one of the mass values present in a strain-specific cluster (Table 2, Table S7). Cyanopeptolin CP999 was also dereplicated using NPAtlas and corresponded to one of the mass features present in the molecular cluster (Figure 9). Micropeptin D, was also found in the LC-HRESIMS of this cyanobacterium, being one of the 15 hits of Dereplicator VarQuest, but the intensity of this metabolite was low (Table S7). With the use of this dereplication tool, four new intense masses were discovered, in which two are shared with other symbiotic UC strains (*m/z* 622.2724 $[M+H]^+$ and *m/z* 513.3286 $[M+H]^+$).

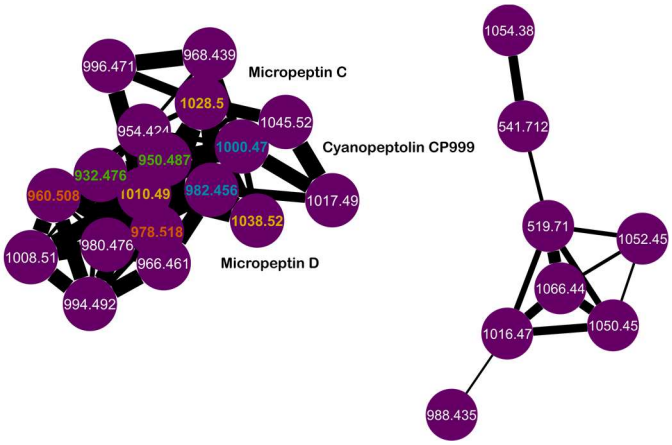


Figure 9 - Main molecular clusters from *Nostoc* sp. LEGE 12456, retrieved from the GNPS molecular networking analysis. Micropeptin C and D were identified with the dereplication tools (yellow) and Cyanopeptolin CP999 (blue). Related masses to micropeptins and/or cyanopeptolins probably indicate the presence of new analogues (green and orange).

The inspection of the fragmentation pattern of micropeptin C and cyanopeptolin CP999 was performed. Since the MS2 data of these two compounds is similar to the other two mass features present in the molecular cluster (*m/z* 950.4859 [M+H]+ and *m/z* 978.5173 $[M+H]^+$), as shown in Figure 10, the latter are possibly new analogues of these families of cyanobacterial compounds.

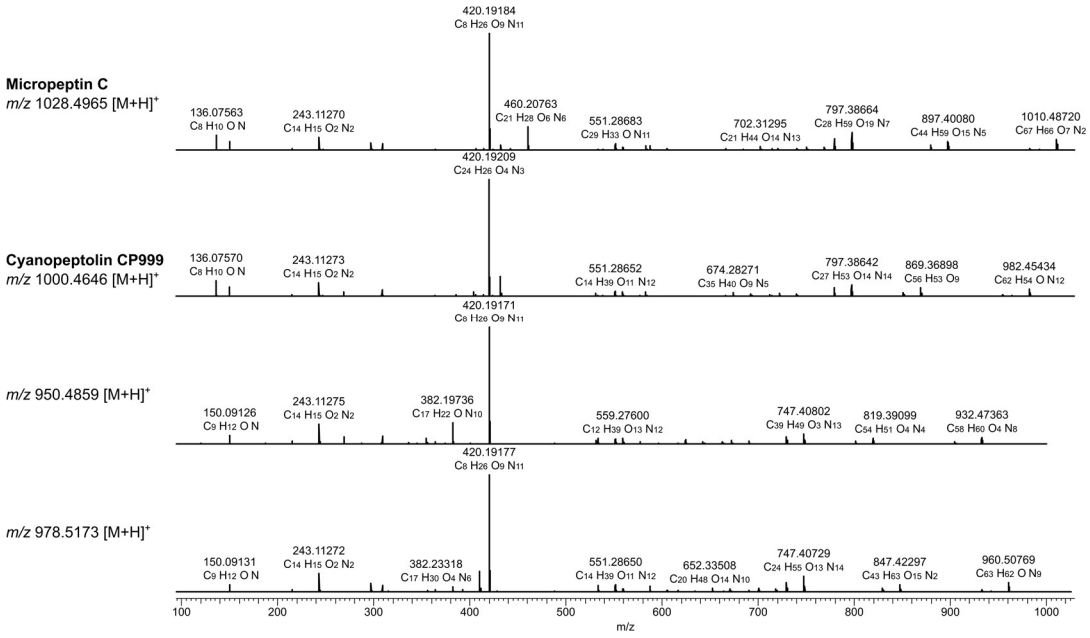


Figure 60 – MS^2 spectra of the identified micropeptin C and cyanopeptolin CP999, with possible new analogues present in the UC *Nostoc* sp. LEGE 12456.

Table 2 – Main information of the mass features present in the molecular networking analysis from the UC Nostocales strains. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra. [a]mass annotated with lipid maps.

Strain	Fr	t_R (min)	Monoisotopic mass (m/z)	Adducts (calcd)						NPAtlas	DNP	Dereplicator	Dereplicator VarQuest	Dereplicator +
				$[M+H]^+$	$[M+Na]^+$	$[M+2H]^{2+}$	$[M+2Na]^{2+}$	$[2M+H]^+$	$[2M+Na]^+$					
Desmonostoc muscorum LEGE 12446	C	7.2	1387.7491	**1388.7564**	1410.7383	694.8818	716.8638	2776.5055*	2798.4875*			0 hits	26 hits	0 hits
		7.9	1255.7067	1256.7139	1278.6959	**628.8606**	650.8425	2512.4206*	2534.4025*					
		7.4	1401.7650	1402.7722	1424.7542	**701.8898**	723.8717	2804.5372*	2826.5191*					
		6.2	905.5111	**906.5184**	928.5003	453.7628	475.7448	1812.0295*	1834.0115*	Anabenopeptin 905				
		5.7	891.4956	**892.5028**	914.4848	446.7551	468.7370	1783.9984*	1805.9804*	Anabaenopeptin 891; Anabaenopeptin SA1				
		10.8	330.2764	**331.2837**	353.1656	166.1455*	188.1274*	661.5600*	683.5420*		9,10-Dihydroxyoctadecanoic acid			
		8.9	492.3659	493.3732	**515.3551**	247.1902*	269.1722*	985.7391*	1007.7210*					
		9.6	471.3556	472.3629	**494.3448**	236.6851*	258.6670*	943.7185*	965.7004					
		15.6	816.5549	**817.5622**	839.5441	409.2847*	431.2667*	1634.1170*	1656.0990*		3-O-(9Z-Octadecenoyl), 5-hexadecanoyl			
	B	9.6	648.4568	649.4641	671.4460	**325.2357**	347.2176*	1297.9209*	1319.9028*			0 hits	0 hits	0 hits
		12.5	602.4524	603.4597	**625.4416**	302.2335*	324.2154*	1205.9121*	1227.8940*					
Nostoc sp. LEGE 12450	C	6.9	1099.5226	**1100.5299**	1122.5118	550.7686	572.7505	2200.0525*	2222.0344*			0 hits	7 hits	1 hit
		4.9	823.4222	**824.4295**	846.4114	412.7184	434.7003	1647.8517*	1669.8337*					
		5.3	1006.4160	**1007.4232**	1029.4052	504.2175*	526.1972	2013.8392*	2035.8211*					
		6.1	1076.4584	**1077.4657**	1099.4476	539.2365*	561.2184	2153.9241*	2175.9061*					
		14.6	1001.6225	**1002.6298**	1024.6118	501.8185*	523.8005*	2004.2524*	2026.2350*					
Nostoc sp. LEGE 12451	C	5.8	914.4311	**915.4383**	937.4203	458.2228*	480.2047*	1829.8694*	1851.8513*			0 hits	4 hits	0 hits
		7.1	1047.5276	**1048.5348**	1070.5168	524.7711	546.7530	2096.0624*	2118.0443*	Microcystin-HphHph				
		5.9	1019.4950	**1020.5101**	1042.4920	510.7587	532.7406	2040.0129*	2061.9948*					

Nostoc sp. LEGE 12454	C	6.1	869.4648	**870.4721**	892.4540	435.7397	457.7216	1739.9369	1761.9188			0 hits	2 hits	0 hits
		6.1	344.1732	**345.1805**	367.1624*	173.0939*	195.0758*	689.3537*	711.3356*					
	B	7.0	875.4434	**876.4507**	898.4326	438.7290*	460.7109	1751.8940	1773.8760			0 hits	0 hits	0 hits
		10.9	986.7110	987.7182	1009.7008	494.3628	**516.3447**	1974.4292*	1996.4112*					
	A	14.2	734.5695	735.5768	757.5587	368.2920*	390.2740*	1470.1463*	**1492.1283**			0 hits	0 hits	2 hits
		13.5	664.4907	665.4979	**687.4799**	333.2526*	355.2345*	1329.9886*	1351.9705					
		14.4	748.5845	749.5918	**771.5737**	375.2995*	397.2815*	1498.1763*	1520.1582					
		13.8	692.5215	693.5287	**715.5107**	347.2680*	369.2500*	1386.0502*	1408.0322					
		12.0	462.3907	463.3980	**485.3799**	232.2026*	254.1846*	925.7887*	947.7706*					
		14.0	720.5530	721.5603	**743.5422**	361.2838*	383.2657*	1442.1133*	1464.0953					
		12.8	466.3439	467.3512	**489.3331**	234.1792*	265.1612*	933.6950*	955.6770*		Vitamin K1, 5' S-Hydroxy			
		12.6	390.2765	391.2837[a]	**413.2657**	196.1455*	218.1274*	781.5602*	803.5421*					
Nostoc sp. LEGE 12456	C	7.7	949.4786	950.4859	**972.4679**	475.7466*	497.7285*	1899.9646*	1921.9465*			0 hits	15 hits	2 hit
		8.7	977.5101	978.5173	**1000.4993**	489.7623*	511.7442*	1956.0274*	1978.0093*					
		7.2	999.4573	1000.4646	**1022.4465**	500.7359*	522.7179*	1999.9218*	2021.9038*	Cyanopeptolin CP999				
		8.2	1027.4892	1028.4965	**1050.4784**	514.7519*	536.7338*	2055.9856*	2077.9676*	Micropeptin C; Cyanopeptolin CP1027	Micropeptin C			
Nostoc sp. LEGE 14445	B	12.8	940.9375	941.9448	963.9267	471.4760	**493.4580**	1882.8823*	1904.8642*			0 hits	0 hits	0 hits
	C	11.9	422.4113	**423.4186**	445.4005*	212.2129*	234.1949*	845.8299*	867.8118*		5-Octacosenoic acid	0 hits	1 hit	1 hit

The remaining Nostocales UC cyanobacterium with a strain-specific molecular clusters was “unidentified Nostocales LEGE 14445”, showing only two unique intense mass features, each from a different fraction (B and C). The one in fraction C was identified as 5-octacosenoic acid (using DNP). Other two dereplication results were obtained using Dereplicator+ and Dereplicator VarQuest, but none of the mass features were present the molecular clusters of this cyanobacterium (Table S8). Dereplicator+ identified another fatty acid produced by this cyanobacterium, while the hit from VarQuest was present throughout the entire LC-HRESIMS run, possibly belonging to the background noise of the run (Table S8).

The Nodosilineales UC strains did not show many single-strain molecular networks from the molecular networking analysis. Even so, both strains from this order exhibited abundant masses not found in the other UC strains, as represented in Table 3. The *Nodosilinea* sp. LEGE 11468 showed three intense mass features in the molecular clusters. Regarding dereplication, Dereplicator+ retrieved two hits, found in the LC-HRESIMS/MS data, but none are correspondent to the masses present in the molecular clusters (Table S9). Dereplicator VarQuest retrieved four hits of cyanobacterial origin but once again, not the mass features of the molecular clusters. From *Cymatolege* sp. LEGE 11434, the five mass features retrieved from the molecular clusters were not identified with the dereplication tools (Figure 11). Therefore, the results from DNP, Dereplicator VarQuest and Dereplicator+ did not correspond to any of the mass features present in Figure 11 (Table S10).

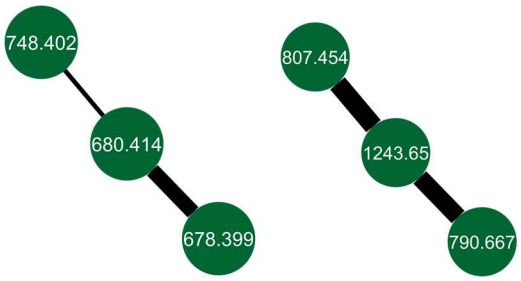


Figure 11 - Main molecular clusters from *Cymatolege* sp. LEGE 11434 retrieved from the GNPS molecular networking analysis.

The remaining UC strains with single-strain molecular clusters are grouped in Table 4, and belong to the Pleurocapsales, Synechococcales and Leptolyngbyales cyanobacterial orders. For *Myxosarcina* sp. LEGE 06146, only two mass features were

abundant (≥ 10^7) in the molecular clusters of this cyanobacterium, and none were dereplicated using the tools mentioned above (Table 4). The two dereplication hits of this cyanobacterium were obtained with Dereplicator VarQuest but did not correspond to masses present in the molecular clusters (Table S11). On the other hand, from *Romeriopsis navalis* LEGE 11480, 20 unique and intense mass features were present in the molecular clusters of this cyanobacterium, mainly present in fraction C (Figure 12, Table 4). One of these 20 mass features was dereplicated as a glycerol lipid, using DNP (Table S12). The other dereplication hit, obtained with dereplicator VarQuest, corresponded to a mass feature not found in the molecular network of this strain.

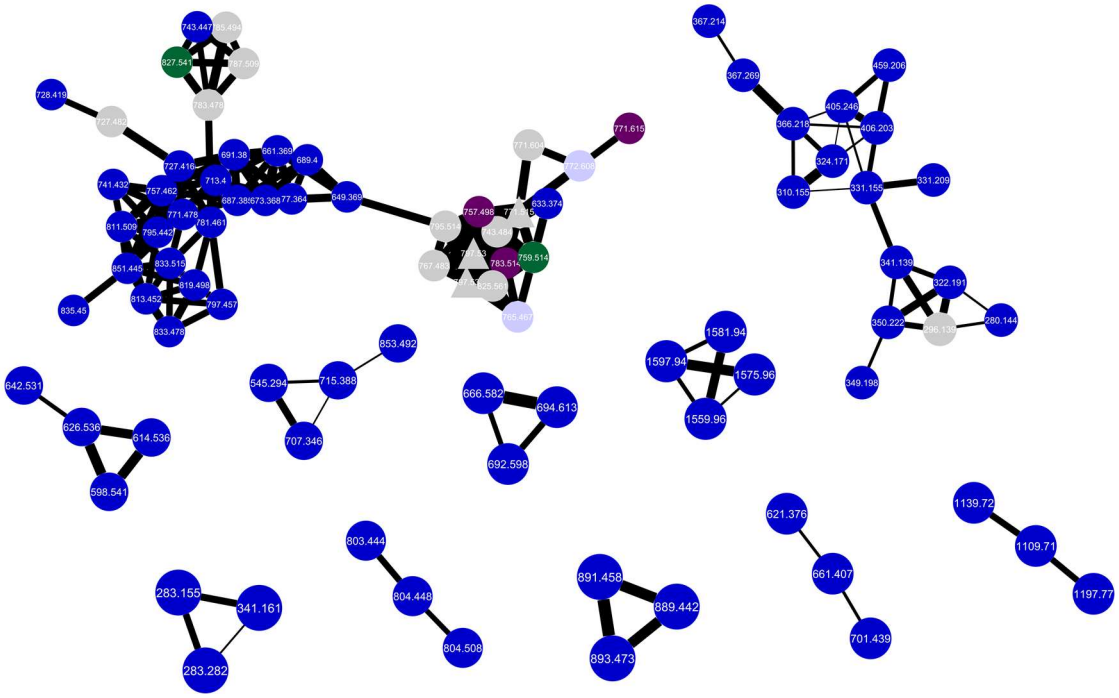


Figure 7 – Main molecular clusters from *Romeriopsis navalis* LEGE 11480 retrieved from the GNPS molecular networking analysis. The first clusters (top left) is related with several other masses shared by other UC strains that were not considered for analysis.

The last strain from the UC group is cf. *Phormidesmis* sp. LEGE 11477 with five mass features selected from the molecular networking analysis (Figure 13). The only dereplication hits were from the GNPS tools, which identified anabaenopeptins A and D in fraction A of this cyanobacterium (Table S13). The remaining hits did not correspond to the masses retrieved from the molecular clusters of this cyanobacterium (Table S13).

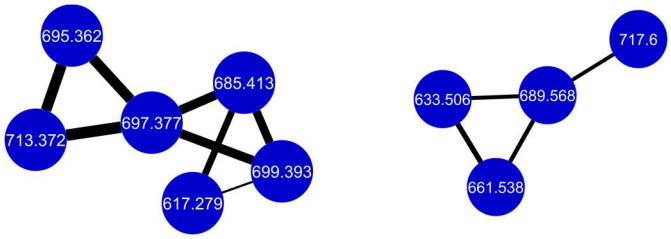


Figure 8 - Main molecular clusters from cf. *Phormidesmis* sp. LEGE 11477 retrieved from the GNPS molecular networking analysis.

Although the bioactivity screening showed only mild bioactivity, combining that information with the one retrieved from the GNPS molecular networking analysis, led to the selection of the following strains for large-scale growth aiming at new natural product isolation: *Desmonostoc muscorum* LEGE 12446, *Nostoc* sp. LEGE 12450, *Nostoc* sp. LEGE 12451, *Nostoc* sp. LEGE 12454, *Nostoc* sp. LEGE 12456 and *Cymatolege* sp. LEGE 11434. The selected cyanobacterial strains above were listed according to the priority for isolation, based on the number of mass features retrieved from the GNPS molecular networking analysis, interesting chemical groups (sulfate/phosphate groups) and the perceived growth rate of the strain (higher being better).

Table 3 - Main information of the mass features present in the molecular networking analysis from the UC Nodosilineales strains. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra.

Strain	Fr	t_R (min)	Monoisotopic mass (m/z)	Adducts (calcd)						NPAtlas	DNP	Dereplicator	Dereplicator Varquest	Dereplicator +
				$[M+H]^+$	$[M+Na]^+$	$[M+2H]^{2+}$	$[M+2Na]^{2+}$	$[2M+H]^+$	$[2M+Na]^+$					
Nodosilinea sp. LEGE 11468	A	9.5	526.4706	**527.4778**	549.4598	264.2426*	286.2245*	1053.9484*	1075.9303*			0 hits	4 hits	2 hits
	B	14.0	1445.0641	1446.0714	1468.0533	723.5393	745.5213	2891.1255	2913.1174*					
	B	14.0	938.5379	**939.5451**	961.5271	470.2762*	492.2581*	1878.9927*	1900.0649*					
Cymatolege sp. LEGE 11434	A	14.3	794.5207	795.5280	**817.5099**	398.2676*	420.2496*	1590.0487	1612.0306		Ribocyclophane E	0 hits	1 hit	2 hits
	A	16.4	819.7673	820.7745	**842.7565**	410.8909*	432.8728*	1640.5418	1662.5237					
	B	17.8	875.8290	876.8362	**898.8182**	438.9218*	460.9037*	1752.6652*	1774.6472*			0 hits	1 hit	0 hits
	C	14.8	784.4642	785.4715	807.4534	**393.2394**	404.2304	1569.3957	1591.9177			0 hits	0 hits	4 hits
	C	17.3	641.4297	642.4370	**664.4189**	321.7221*	343.7041*	1283.8667	1305.8486					

Table 4 - Main information of the mass features present in the molecular networking analysis from the UC Pleurocapsales, Synechococcales and Leptolyngbyales strains. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra.

Strain	Fr	t_R (min)	Monoisotopic mass (m/z)	Adducts (calcd)						NPAtlas	DNP	Dereplicator	Dereplicator Varquest	Dereplicator +
				$[M+H]^+$	$[M+Na]^+$	$[M+2H]^{2+}$	$[M+2Na]^{2+}$	$[2M+H]^+$	$[2M+Na]^+$					
Myxosarcina sp. LEGE 06146	B	9.8	380.2343	**381.2412**	403.2235	191.1244*	213.1063*	761.4758*	783.4577*			0 hits	2 hits	0 hits
	B	13.0	717.5383	718.5456	**740.5275**	359.7764*	381.7584*	1436.0838	1458.0658*					
Romeriopsis navalis LEGE 11480	C	13.8	638.3779	639.3852	**661.3671**	320.1962*	342.1782*	1277.7631*	1299.7450*			0 hits	1 hit	0 hits
	C	13.2	626.3790	**627.3863**	649.3682	314.1968*	336.1787*	1253.7653	1275.7473					
	C	13.4	654.3739	655.3812	**677.3631**	328.1942*	350.1762*	1309.7551	1331.7370					
	C	13.9	650.3786	651.3859	**673.3678**	326.1966*	348.1785*	1301.7645*	1323.7464*					

	C	14.8	704.4924	705.4997	**727.4817**	353.2535*	375.2354*	1409.9922*	1431.9705*		35-O-β– 3,5-anhydrogalacturonopyranosylbacteriohopanetetrol			
	C	16.2	758.4725	759.4797	**781.4617**	380.2435*	402.2254*	1517.9522*	1539.9341*					
	C	14.2	772.4511	773.4583	**795.4403**	387.2328*	409.2148*	1545.9094*	1567.8914*					
	C	5.7	366.2614	367.2686	**389.2506**	184.1380*	206.1199*	733.5300*	755.5119*					
	C	0.7	287.2818	**288.2890**	310.2710	144.6482*	166.6301*	575.5708*	597.5527*					
	C	0.8	436.2280	**437.2353**	459.2172*	219.1213*	241.1032*	873.4632*	895.4452*					
	C	9.0	522.3036	523.3109	**545.2928**	262.1591*	284.1410*	1045.6145*	1067.5694*					
	C	8.6	666.4104	667.4178	**689.3996**	334.2125*	356.1944*	1333.8280*	1555.8000*					
	C	15.9	798.4652	**799.4725**	821.4544	400.2399*	422.2218*	1597.9377	1619.9196					
	C	12.9	657.5181	**658.5254**	680.5073	329.7663*	351.7483*	1316.0435	1338.0254					
	C	13.3	654.3738	655.3811	**677.3630**	328.1942*	350.1761*	1309.7549	1331.7368					
	C	13.2	626.3791	**627.3863**	649.3683	314.1968*	336.1787*	1253.7654	1275.7473					
	C	9.0	362.1568	**363.1641**	385.1460*	182.0857*	204.0676*	725.3209	747.3028*					
	C	13.0	750.5275	751.5348	**773.5168**	376.2710*	398.2530*	1502.0624	1524.0443		Glycerol 1,2-dialkanoates			
	B	5.4	330.1476	**331.1549**	353.1369*	166.0811*	188.0630*	661.3026*	683.2845*			0 hits	0 hits	0 hits
	B	4.0	282.1475	**283.1548**	305.1367	142.0810*	164.0630*	565.3023*	587.28442*					
cf. *Phormidesmis* sp. LEGE 14777	C	11.6	343.3081	**344.3154**	366.2973	172.6613*	194.6433*	687.6235	709.6054			0 hits	2 hits	0 hits
	A	14.3	632.4983	**633.5056**	655.4876*	317.2564*	339.2384*	1266.0040*	1287.9859*			2 hits	5 hits	2 hits
	A	14.6	660.5297	**661.5370**	683.5189*	331.2721*	353.2541*	1322.0667*	1344.0486*					
	A	15.0	688.5607	**689.5680**	711.5499*	345.2876*	367.2696*	1378.1287*	1400.1107*					
	A	15.4	716.5912	**717.5984**	739.5804*	359.3029*	381.2848*	1434.1896*	1456.1715*					

2.2.2. Explored cyanobacterial LEGEcc strains (EC)

The 12 cyanobacterial LEGEcc strains present in this group were characterized as bioactive against a panel of bioassays or demonstrated metabolomic information suggesting the presence of novel natural products (NP-like masses) (Table 1, Table S1), although no isolation protocol was applied to discover the potential new natural products. To help with the selection of the most promising strains for large-scale culture and subsequent natural products isolation, an untargeted metabolomic analysis was performed using GNPS molecular networking[22]. Dereplication was also performed using the same tools for the UC group, NPAtlas[32], DNP[33], Dereplicator[29], Dereplicator VarQuest[30] and Dereplicator+[31]. The cyanobacterial strains were grouped according to order classification and a search of single-strain clusters was conducted. Contrary to the UC group, fewer single-strain molecular clusters were detected for this group of cyanobacteria (Figure S6). The results were from strains of Synechoccalles, Oscillatoriales, Oculatellales, Nostocales and Chroococcidiopsidales orders, as shown in table 5. The EC strains with intense mass features for isolation from the molecular networking analysis, were mainly from the EC strains with NP-like masses (unidentified filamentous Synechococcales LEGE 06118, *Calenema singularis* LEGE 06188, and *Chroococcidiopsis* sp. LEGE 06174). However, some mass features were also retrieved from the EC strain with reported bioactivity, cf. *Oxynema acuminatum* LEGE 06072. The dereplicator+ showed many hits for all the EC strains (data not shown), but most of them were not from cyanobacterial origin and/or belong to the primary metabolism, so they were not considered as hits for dereplication purposes.

The most prolific EC strain was *Calenema singularis* LEGE 06188, showing 34 abundant mass features, distributed among the three used fractions (Table 5). Regarding dereplication, with DNP two hits were identified as fatty acid related compounds (Figure 14). The two hits from Dereplicator VarQuest, as well as the other hit from Dereplicator+, were not correspondent to the mass features present in the molecular clusters (Table S14).

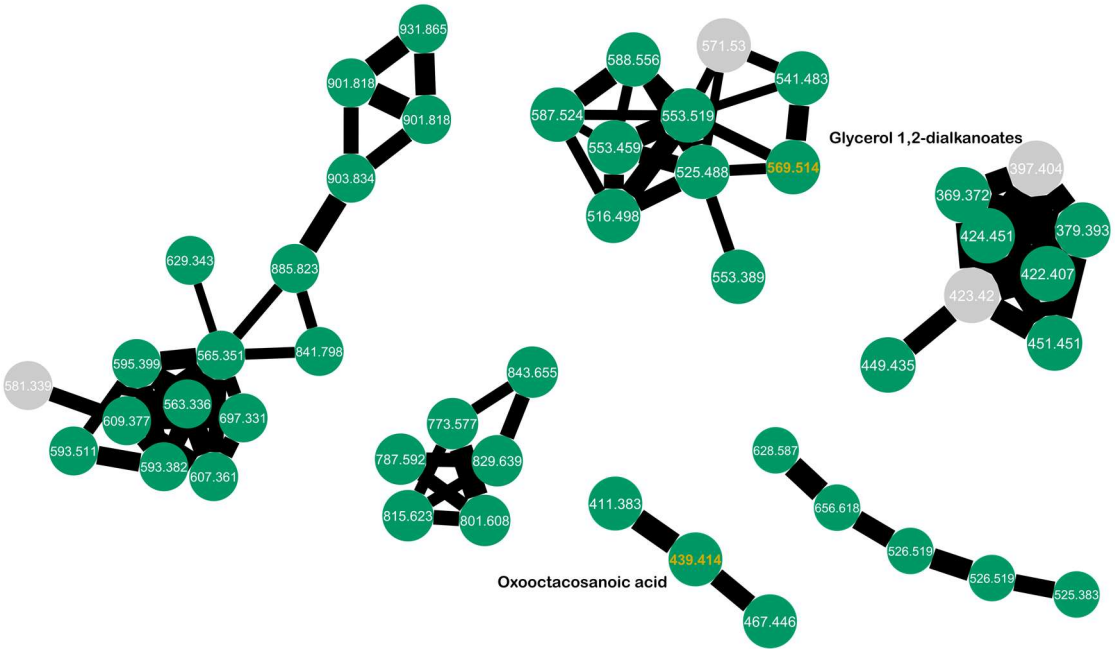


Figure 9 - Main molecular clusters from *Calenema singularis* LEGE 06188, retrieved from the GNPS molecular networking analysis. Annotation of oxooctacosanoic acid and glycerol 1,2-dialkanoates with Dictionary of Natural Products (DNP, yellow label). The grey mass features are shared by several EC strains.

Other two EC strains with abundant features, showed three masses for isolation as shown in figure 15, two from unidentified filamentous Synechococcales LEGE 06118, represented in green, and one from *Chroococcidiopsis* sp. LEGE 06174, represented in blue, the latter corresponding to a halogenated compound (Table 6). In terms of dereplication, only two hits were shown for the first cyanobacterium, using Dereplicator+. Nevertheless, these hits were not correspondent to the masses shown in the molecular clusters, shown in Figure 15 (Table S15).

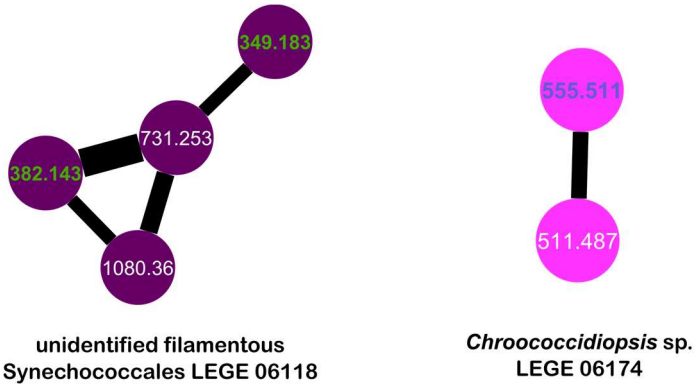


Figure 15 - Main molecular clusters from unidentified filamentous Synechococcales LEGE 06118, and *Chroococcidiopsis* sp. LEGE 06174, retrieved from the GNPS molecular networking analysis. The considered masses (according to the defined intensity) are represented in green and blue, respectively.

Table 5 - Main information of the mass features present in the molecular networking analysis from the EC Oculatellates strain *Calenema singularis* LEGE 06188. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra.

Strain	Fr	t_R (min)	Monoisotopic mass (m/z)	Adducts (calcd)						NPAtlas	DNP	Dereplicator	Dereplicator Varquest	Dereplicator +
				$[M+H]^+$	$[M+Na]^+$	$[M+2H]^{2+}$	$[M+2Na]^{2+}$	$[2M+H]^+$	$[2M+Na]^+$					
Calenema singularis LEGE 06188	C	11.2	788.5638	789.5171	**811.5530**	395.2892*	406.2801*	1578.1348*	1600.1168*			0 hits	0 hits	0 hits
		11.6	410.3751	411.3824	**433.3643**	206.1948*	217.1858*	821.7575*	843.7394					
		12.0	858.6417	859.6490	**881.6309**	430.3281*	441.3191*	1718.2907	1740.2726					
		12.1	525.5115	**526.5188**	548.5007*	263.7630*	274.7540*	1052.0303*	1074.0122*					
		12.3	466.4382	467.4454	**489.4274**	234.2264*	245.2173*	933.8836*	955.8656*					
		12.6	570.5219	571.5292	**593.5111**	286.2682*	297.2592*	1142.0511*	1164.0330*					
		12.7	655.6103	**656.6176**	678.5995*	328.8124*	339.8034*	1312.2279*	1334.2099*					
		14.7	884.8158	**885.8231**	907.8050*	443.4152*	454.4062*	1770.6389*	1792.6208*					
		14.9	552.4519	**553.4592**	575.4411*	277.2332*	288.2242*	1105.9111*	1127.8930*					
		15.8	542.3617	**543.3690**	656.3509*	272.1881*	283.1791*	1085.7306*	1107.7126					
		16.1	586.3876	587.3948	**609.3768**	294.2011*	605.1920*	1173.7824	1195.7643					
		16.7	562.3899	563.3971	**585.3791**	282.2022*	293.1932*	1125.7870*	1147.7690*					
		16.9	570.3924	571.3997	**593.3816**	286.2035*	297.1845*	1141.7921	1163.7741					
		17.0	638.3670	**639.3743**	661.3562*	320.1808*	331.1818*	1277.7413*	1299.7233*					
	B	12.0	438.4063	439.4136	**461.3955**	220.2104*	231.2014*	877.8199*	899.8018		Oxooctacosanoic acid	0 hits	1 hit	1 hit
		12.1	524.3745	**525.3818**	547.3637*	263.1945*	274.1855*	1049.7563*	1071.7382*					
		12.2	585.5324	**586.5397**	608.5216*	293.7735*	304.7645*	1172.0721*	1194.0540*					
		12.3	440.4224	**441.4297**	463.4116*	221.2185*	232.2094*	881.8520*	903.8340*					
		12.3	450.4433	**451.4503**	473.4322*	226.2288*	237.2197*	901.8933*	923.8752*					
		12.4	540.4750	541.4823	**563.4642**	271.2448	282.2358	1081.9573*	1103.9392*					
		12.4	368.3647	**369.3720**	391.3539	185.1896*	196.1806*	737.7367*	759.7186					
		12.4	627.5794	**628.5866**	650.5686*	314.7970*	325.7879*	1256.1660*	1278.1479*					

		12.5	570.5217	**571.5289**	593.5109*	286.2681*	297.2591*	1142.0506*	1164.0326*					
		12.6	396.3963	**397.4036**	419.3855	199.2054*	210.1964*	793.7999	815.7819					
		12.7	586.5165	587.5238	**609.5057**	294.2655*	305.2565*	1174.0403*	1196.0223*					
		12.7	568.5060	569.5132	**591.4952**	285.2603*	296.2512	1138.0192*	1160.0012*		glycerol 1,2 dialkanoates			
		13.2	423.4429	**424.4502**	446.4321	212.7287*	223.7197*	847.8931*	869.8751*					
		13.5	426.4432	427.4505	**449.4324**	214.2289*	225.2198*	853.8937*	875.8756					
		15.9	840.7899	**841.7972**	863.7791*	421.4022*	432.3932*	1682.5870*	1704.5690*					
		17.5	875.8287	876.8360	**898.8179**	438.9216*	449.9126*	1752.6647	1774.6466					
	A	11.2	519.4647	**520.4720**	542.4539	260.7396*	271.7306	1039.9367*	1061.9186*			0 hits	1 hit	0 hits
		11.8	562.5434	563.5507	585.5327	**282.2790**	293.2700*	1126.0942*	1148.0761*					
		12.9	530.3998	**531.4071**	553.9892*	266.2072*	277.1982*	1061.8069*	1083.7889					
		13	498.4639	499.4712	**521.4531**	250.2392*	261.2302*	997.9351*	1019.9170*					

Table 6 - Main information of the mass features present in the molecular networking analysis from the EC Synechoccales strain unidentified filamentous Synechococcales LEGE 06118 and Chroococcidiopsales *Chroococcidiopsis* sp. LEGE 06174. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+. In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra.

Strain	Fr	t_R (min)	Monoisotopic mass (m/z)	Adducts (calcd)						NPAtlas	DNP	Dereplicator	Dereplicator Varquest	Dereplicator +
				$[M+H]^+$	$[M+Na]^+$	$[M+2H]^{2+}$	$[M+2Na]^{2+}$	$[2M+H]^+$	$[2M+Na]^+$					
unidentified filamentous Synechococcales LEGE 06118	B	0.6	381.1351	**382.1424**	404.1243	191.5748*	202.5658*	763.2774*	785.2594*			0 hits	0 hits	2 hits
		0.9	326.1934	327.2006	**349.1826**	164.1040*	175.0949*	653.3940*	675.3759*					
Chroococcidiopsis sp. LEGE 06174	A	15.2	552.5266	**553.5339**	575.5159	277.2706	288.2616	1106.0606	1128.0425			0 hits	0 hits	0 hits

Regarding the EC strain cf. *Oxynema acuminatum* LEGE 06072, with reported bioactivity from previous studies, it showed five intense mass features in the GNPS molecular networking analysis, from fractions C and B (Figure 16, Table 7).

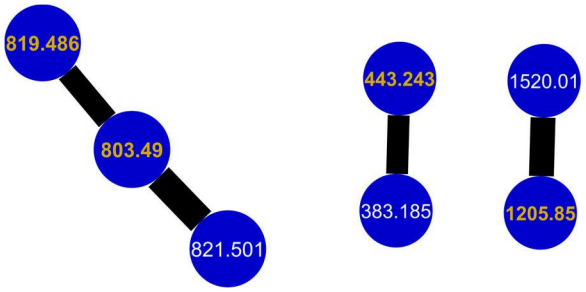


Figure 16 - Main molecular clusters from cf. *Oxynema acuminatum* LEGE 06072, retrieved from the GNPS molecular networking analysis. The considered masses (according to the defined intensity) are represented in yellow. Only four of the five mass features are shown.

In terms of dereplication none of the masses from the molecular clusters had a hit using the dereplication tools described above. This cyanobacterium demonstrated bioactivity mainly the reduction of in vivo anti-settlement of mussel plantigrade larvae[18], acute effects on *Artemia salina* and a growth inhibition of *Nannochloropsis* sp. and *Synechocystis salina*.[19,20]

Table 7 - Main information of the mass features present in the molecular networking analysis from the EC Oscillatoriales strain cf. *Oxynema acuminatum* sp. LEGE 06072. The mass features were selected from a single cyanobacterial strain and with an intensity on the total ion chromatogram (TIC) above or equal to 1.0E+07. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator VarQuest, Dereplicator+ (data not shown). In bold are the most abundant adduct, and * corresponds to adducts not found in the experimental spectra.

Strain	Fr	t_R (min)	Monoisotopic mass (m/z)	Adducts (calcd)					
				$[M+H]^+$	$[M+Na]^+$	$[M+2H]^{2+}$	$[M+2Na]^{2+}$	$[2M+H]^+$	$[2M+Na]^+$
cf. *Oxynema acuminatum* LEGE 06072	C	11.6	796.4956	797.5029	**819.4848**	399.2551*	410.2460*	1593.9985*	1615.9804*
		11.7	780.5005	781.5077	**803.4897**	391.2575*	402.2485*	1562.0082*	1583.9901*
		12.8	1182.8643	1183.8716	**1205.8535**	592.4394*	603.4304*	2366.7359*	2388.7178*
		12.8	764.5057	765.5130	**787.4949**	383.2601*	394.2511*	1530.0187*	1552.0007
	B	9.3	442.2347	**443.2420**	465.2239*	222.1246*	233.1156*	885.4767*	907.4586*

Other EC strains showed smaller molecular clusters (≤ 2 mass features) from the GNPS molecular networking analysis, specifically *Microcoleus* sp. LEGE 07076 and *Nostoc* cf. *edaphicum* LEGE 07299, however the masses were less abundant (<1.0E^{07}),

so they were not considered for metabolite annotation and dereplication. For that reason, the selected EC strains for large-scale growth are *Calenema singularis* LEGE 06188 due to the higher number of mass features, *Chroococcidiopsis* sp. LEGE 06174 since it has a halogenated abundant mass, and cf. *Oxynema acuminatum* LEGE 06072 since it was the only strain with reported bioactivity from previous studies, with molecular networking results.

2.3. Material and Methods

2.3.1. Cyanobacterial strain cultivation

For the screening assays, freshwater strains were cultured in Z8 medium[35] and for marine strains the medium was supplemented with marine tropical salt (25 g L^{-1}, TropicMarin) and vitamin B12 (10 μg L^{-1}). Growth conditions were maintained at a temperature of 25º C and with a photoperiod of 16h/8h light and dark respectively (light intensity of 10−30 μmol photons s^{-1} m^{-2}) and a progressive upscaling was performed from 40 mL flasks up to 6L flasks with 4L medium and in this final stage with constant aeration. At the exponential phase of growth (ca. 30 days), cells were harvested through centrifugation or with a plankton net, frozen and freeze-dried. The only exception were the cyanobacterial strains *Cymatolege* sp. LEGE 11434, *Scytonema* sp. LEGE 07182, unidentified filamentous Synechococcales LEGE 06118, unidentified Nostocales LEGE 06150 and *Toxifilum mysidocida* LEGE 06108 that were upscaled only to 1 L volume culture since their growth was slower. The marine strains had an extra washing step after biomass collection with ddH_2O, to remove the excess of salts present in the media. However, some cyanobacterial strains were very sensitive to this washing step and some cells burst during the process. Cyanobacterial strains identified as growing with light restrictions were maintained at and intensity of 3.5-4.0 μmol photons s^{-1} m^{-2}. Specific growth conditions for some cyanobacterial strains are described in the supplementary material, Table S1.

2.3.2. Cyanobacterial extraction and fractionation

The cyanobacterial freeze-dried biomass was extracted with a mixture of CH_2Cl_2/MeOH (2:1) by repeated percolation at room temperature (3x), with a heating plate at 40 ºC (3x) and a final extraction in the ultrasonic bath (10 min) to ensure cell disruption and extraction. The crude extracts were further fractionated with pre-packed

SiOH normal phase SPE cartridges (Strata SI-1, Phenomenex), from 2 g to 20 g according to the mass of the crude extract. A gradient elution from 100% hexane (*n*-hex) to 100% ethyl acetate (EtOAc) to 100% methanol (MeOH) was used to separate the crude extracts according to different solvent polarities, yielding 3 fractions, A, B and C respectively. For fraction A, a gradient from 100% *n*-hex to 60% n-hex in EtOAc, for fraction B 60%EtOAc in *n*-hex to 25% MeOH in EtOAc and for fraction C, 75% MeOH in EtOAc to 100% MeOH. As explained above, some marine cyanobacterial strains were susceptible to the washing step, so for some strains, fraction C had salts from the media. To those fractions an extra SPE cleaning step was performed using reversed-phase silica cartridges (Strata C18-E, 100 mg, 55 *μ*m, 70 Å, Phenomenex), according to the manufacturers protocol. The mass of the resulting fractions, as well as those of the crude extracts are presented in the supplementary material, Table S2.

2.3.3. Anticancer assays

The anticancer cell lines selected for the bioactivity screening were the colon carcinoma (RKO), breast carcinoma (T47D), lung carcinoma (A459), liver carcinoma (HepG2), osteosarcoma (MG-63) and neuroblastoma (SHSY-S5) all acquired from the American Type Culture Collection (ATCC). Cells were cultured in DMEM Glutamax medium (Dulbecco's Modified Eagle Medium DMEM GlutaMAX™), supplemented with 10% (v/v) fetal bovine serum, 1% of penicillin-streptomycin (Pen-Strep 100 IU mL^{-1} and 10 mg mL^{-1}, respectively) and 0.1% fungicide Amphotericin B (250 μg mL^{-1}) and were maintained in a humidified incubator at 37 °C, with 5% CO_2. For the neuroblastoma cell line SHSY-S5 MEM:Ham's F12 medium (1:1) supplemented with 10% (v/v) fetal bovine serum, 1% of penicillin-streptomycin (Pen-Strep 100 IU mL^{-1} and 10 mg mL^{-1}, respectively) and 0.1% fungicide Amphotericin B (250 μg mL^{-1}). The cell lines were transferred every 5 or 7 days. For experimental procedure, after reaching 80% confluency in the culture flask, cells were washed with PBS (Phosphate-Buffered Saline) and harvested with Tripllex (trysin substituent) and seeded in 96-well plates at a concentration of $3.3x10^4$ cells mL^{-1} for exposure to the cyanobacterial fractions. The cell lines were exposed to the fractions at a concentration of 30 μg mL^{-1} (dissolved in dimethyl sulfoxide, DMSO) during 48h. After that period, cells were incubated with 0.5 mg MTT for 4h and posterior the formazan salts were dissolved with DMSO, and the absorbance read at 550 nm in a GEN5TM-Multi-detection Microplate Reader. The negative control for the anticancer assays was 1% DMSO, while the positive control was the membrane disruptor digitonin (25 *μ*g mL^{-1}).

2.3.4. Antimicrobial assays

Antibacterial assays were performed using the disc diffusion method and using *Staphylococcus aureus*, *Bacilus subtilis*, *Escherichia coli*, *Salmonela typhimurium* and *Candida albicans*. Mueller-Hinton agar medium (Oxoid) was used for bacterial inoculation and Sabouraud agar medium (Oxoid) was used for fungi inoculation. Petri dishes with the media and inocula were incubated for 24h at 37 °C and the turbidity adjusted to 0.5 McFarland standard (OD_{600} = 0.095-0.110). The plates were seeded with this inoculum, the 6 mm blank discs (Oxoid) distributed in the inoculated plates and impregnated with 15 µL of the 1 mg mL^{-1} stock solution of each fraction. Blank disks impregnated with 15 µL DMSO served as negative controls. After 30 min at room temperature, the plates were incubated at 37 ºC for 24h and the antimicrobial activity was estimated by the development of an inhibition halo around the disk.

2.3.5. GNPS molecular networking analysis and dereplication

The cyanobacterial fractions were analyzed using GNPS Molecular Networking[22] analysis. For LC-HRESIMS analysis, fractions were prepared at 1.0 mg mL^{-1} in LC-MS grade MeOH, filtered through a 0.2 *µ*m syringe regenerated cellulose filter, and 5 *µ*L were injected in an ACE Ultracore 2.5 SuperC18 column (75 × 2.1 mm, Avantor). Samples were eluted at 0.35 mL min^{-1} over a linear gradient from 99.5% solution A (95% H_2O, 5% MeOH, 0.1% v/v HCOOH) and 0.5% solution B (95% isopropanol, 5% MeOH, 0.1% v/v HCOOH) to reach 10% solution B over 0.5 min, followed by an increase to 60% solution B in 8 min and then to 90% in 1 min; these conditions were kept for 6 min before returning to the initial conditions. The column oven was set to 40 °C, UV monitoring was carried out at 254.0 nm. The capillary voltage of the Heated Electrospray Ionization (HESI) was set to 3.8 kV and its temperature to 300 ºC. The sheath gas and auxiliary gas flow rate were at 35 and 10 (arbitrary units as provided by the software settings). Full MS scans were acquired at a resolution of 70,000 fwhm (range of 150–2000 *m/z*), and data dependent MS/MS (ddMS2, Discovery mode) at a resolution of 17,500 fwhm (isolation window used was 3.0 amu, and normalized collision energy was 35). For the molecular network analysis, the protocol available in GNPS documentation was followed, the precursor ion mass tolerance and the fragment ion mass tolerance were set for 0.02 Da, and the minimum of matched fragment ions was set to 3. The resulting network file was visualized using Cytoscape v3.8.0[28].

For dereplication purposes, cyanobacterial fractions were analyzed using Dereplicator[29], Dereplicator VarQuest[30] and Dereplicator+[31] from the GNPS platform. All the dereplication results were filtered according to the score value (≥ 10), *p*-value (≤ $1.0E^{-10}$) and FDR % (= 0), so one of these criteria had to met to be considered a hit. A manual search of the monoisotopic masses from the molecular networking analysis was also performed using the NPAtlas[32] and the Dictionary of Natural Products (DNP)[33] with 10 Δppm mass range. Additional information, regarding the hits from the dereplication tools, was also confirmed using PubChem[36].

2.4. Conclusions

All the cyanobacterial strains selected for this project, 14 unexplored cyanobacteria (UC group) and 12 explored cyanobacteria (EC group) with already reported bioactivity and/or NP-like mass features, were submitted to screening assays to help prioritize the cyanobacterial strains for large-scale growth and subsequent metabolite isolation. To the UC group a bioactivity screening performed with anticancer and antimicrobial assays was performed, as well as an untargeted metabolomics analysis to map metabolites to a specific and unique cyanobacterial strain. Since the EC group had already some published information regarding biological activity, the screening efforts were only applied to the metabolome analysis of these strains.

The selected strains from the UC group for large-scale growth and new natural products isolation are *Desmonostoc muscorum* LEGE 12446, *Nostoc* sp. LEGE 12450, *Nostoc* sp. LEGE 12451, *Nostoc* sp. 12454, *Nostoc* sp. 12456 and *Cymatolege* sp. LEGE 11434. From the EC group, the selected strains based on the number and type of masses features retrieved from the GNPS molecular networking analysis are *Calenema singularis* LEGE 06188 and *Chroococcidiopsis* sp. LEGE 06174. From the EC strains with already reported bioactivity, only cf. *Oxynema acuminatum* LEGE 06072 showed strain-specific molecular clusters in the untargeted metabolomics analysis, so it is also selected for large-scale growth.

The results of this screening effort evidence the natural products richness of cyanobacteria from late-branching orders, mainly Nostocales, which was already well-recognized. Inside the Nostocales strains, it was also visible the uniqueness of the symbiotic cyanobacterial strains since all demonstrated at least one type of bioactivity and showed higher chemical diversity in the GNPS molecular networking analysis.

2.5. Supplementary data

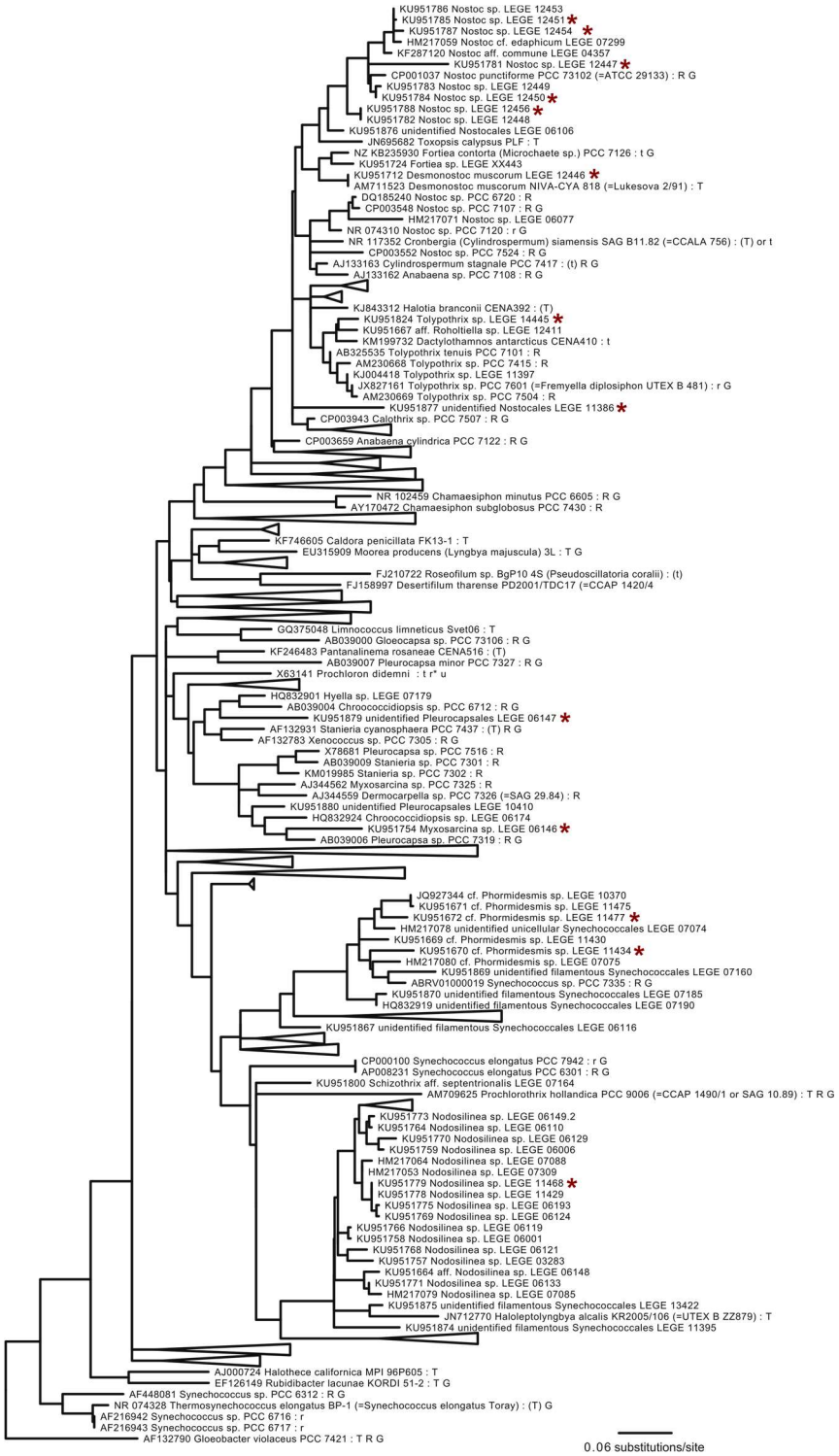


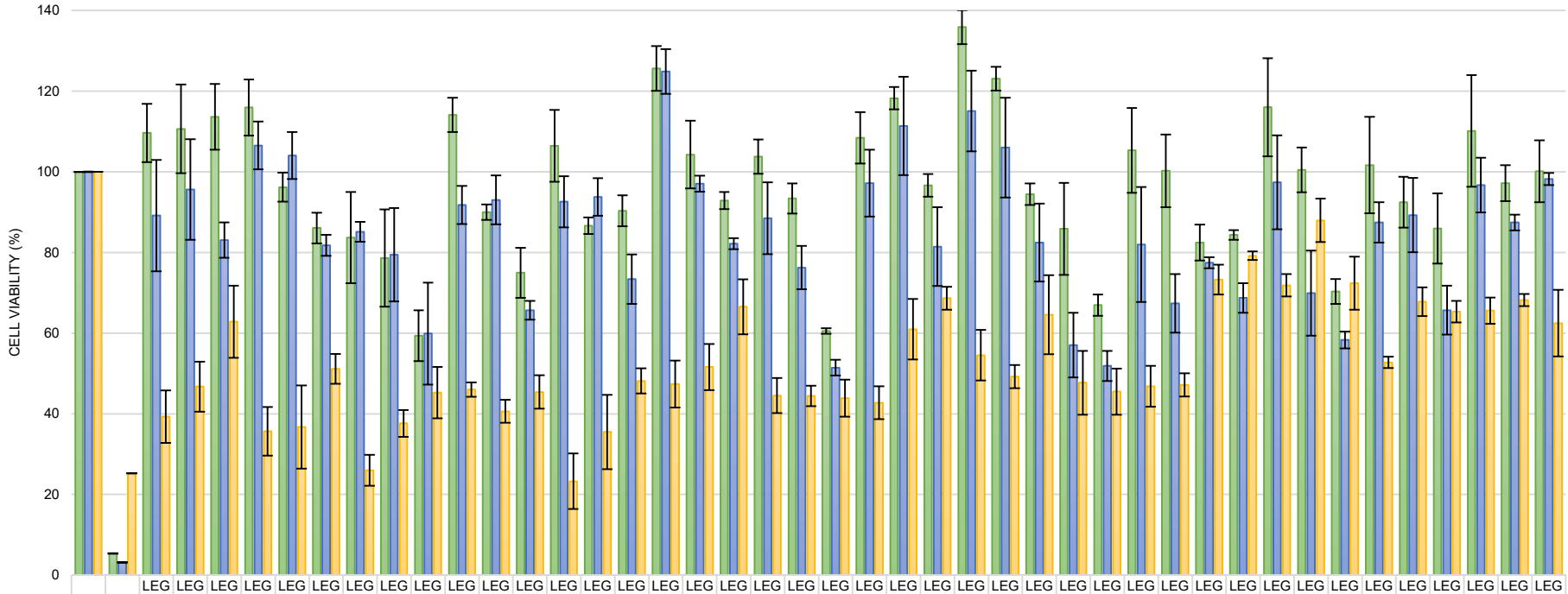
Figure S1 - Maximum-likelihood phylogenetic tree of the LEGEcc strains together with some reference strains, highlighting the 14 selected strains based on phylogenetic placement and uniqueness (red star). Collapsed clades are unlabeled and bootstrap values are omitted for clarity. Strain *Romeriopsis navalis* LEGE 11480 is not depicted in this phylogenetic tree, however it has only 91% identity to the cultured strain corresponding to the top BLAST hit (*Synechococcus* sp. UTEX 2973).

Table S1 – Cyanobacterial information of the LEGEcc strains selected for this project. *Order classification updated to the last phylogenetic publication[23].

	Strain Code	Cyanobacterial strain	Order*	Environment	Location	Media	Growth conditions	Information
UC	LEGE 12446	*Desmonostoc muscorum*	Nostocales	Terrestrial	Portugal, Coimbra	Z8	Standard	Cycad symbiont
	LEGE 12447	*Nostoc* sp.	Nostocales	Terrestrial	Portugal, Lisbon	Z8	Standard	Cycad symbiont
	LEGE 12450	*Nostoc* sp.	Nostocales	Terrestrial	Portugal, Lisbon	Z8	Standard	Cycad symbiont
	LEGE 12451	*Nostoc* sp.	Nostocales	Terrestrial	Portugal, Lisbon	Z8	Standard	Cycad symbiont
	LEGE 12454	*Nostoc* sp.	Nostocales	Terrestrial	Portugal, Lisbon	Z8	Standard	Cycad symbiont
	LEGE 12456	*Nostoc* sp.	Nostocales	Terrestrial	Portugal, Lisbon	Z8	Standard	Cycad symbiont
	LEGE 14445	Unidentified Nostocales	Nostocales	Freshwater	Colombia, Bogotá	Z8	Standard	
	LEGE 11386	Unidentified Nostocales	Nostocales	Marine	Portugal	Z8/25‰ + B12	Standard	Marine sponge
	LEGE 11468	*Nodosilinea* sp.	Nodosilineales	Marine	Portugal, Porto	Z8/25‰ + B12	Standard	
	LEGE 11477	cf. *Phormidesmis* sp.	Leptolyngbyales	Marine	Portugal, Matosinhos	Z8/25‰ + B12	Low light, slow growth	
	LEGE 11480	*Romeriopsis navalis*	Synechococcales	Marine	Portugal, Porto	Z8/25‰ + B12	Low light, slow growth	
	LEGE 06147	Unidentified Pleurocapsales	Pleurocapsales	Marine	Portugal, Caminha	Z8/25‰ + B12	Standard	
	LEGE 11434	*Cymatolege* sp.	Nodosilineales	Marine	Portugal, Porto	Z8/25‰ + B12	Low light, 14 °C, slow growth	
	LEGE 06146	*Myxosarcina* sp.	Pleurocapsales	Marine	Portugal, Caminha	Z8/25‰ + B12	Standard	
EC	LEGE 06070	unidentified filamentous Synechococcales	Synechococcales	Brackish	Portugal, Porto	Z8	Standard	Allelopathic activity
	LEGE 06072	cf. *Oxynema acuminatum*	Oscillatoriales	Brackish	Portugal, Aveiro	Z8	Standard	Cytotoxic, antifouling, antimicrobial activity
	LEGE 07076	*Microcoleus* sp.	Oscillatoriales	Brackish	Portugal, Seixas	Z8	Standard	Cytotoxic, antifouling, allelopathic activity
	LEGE 07092	*Coleofasciculus* sp.	Coleofasciculales	Brackish	Portugal, Aveiro	Z8		Cytotoxic, antifouling, allelopathic activity
	LEGE 07299	*Nostoc* cf. *edaphicum*	Nostocales	Brackish	Portugal, Seixas	Z8	Standard	Antimicrobial activity
	LEGE 06108	*Toxifilum mysidocida*	Oculatellales	Marine	Portugal, Lagos	Z8/25‰ + B12	Standard	Cytotoxic, allelopathic activity; 740-790 Da
	LEGE 07182	*Scytonema* sp.	Nostocales	Marine	Portugal, Caminha	Z8/25‰ + B12	Standard	762 Da
	LEGE 06150	unidentified Nostocales	Nostocales	Marine	Portugal, Esposende	Z8/25‰ + B12	Standard, slow growth	703 Da
	LEGE 06188	*Calenema singularis*	Oculatellales	Marine	Portugal, Vila Nova de Gaia	Z8/25‰ + B12	Standard	700-770 Da
	LEGE 06174	*Chroococcidiopsis* sp.	Chroococcidiopsidales	Marine	Portugal, Vila Nova de Gaia	Z8/25‰ + B12	Standard	750 Da
	LEGE 06131	*Leptolyngbya saxicola*	Leptolyngbyales	Marine	Portugal, Lagos	Z8/25‰ + B12	Standard	1004 Da
	LEGE 06118	unidentified filamentous Synechococcales	Synechococcales	Marine	Portugal, Lagos	Z8/25‰ + B12	Standard	740-820 Da

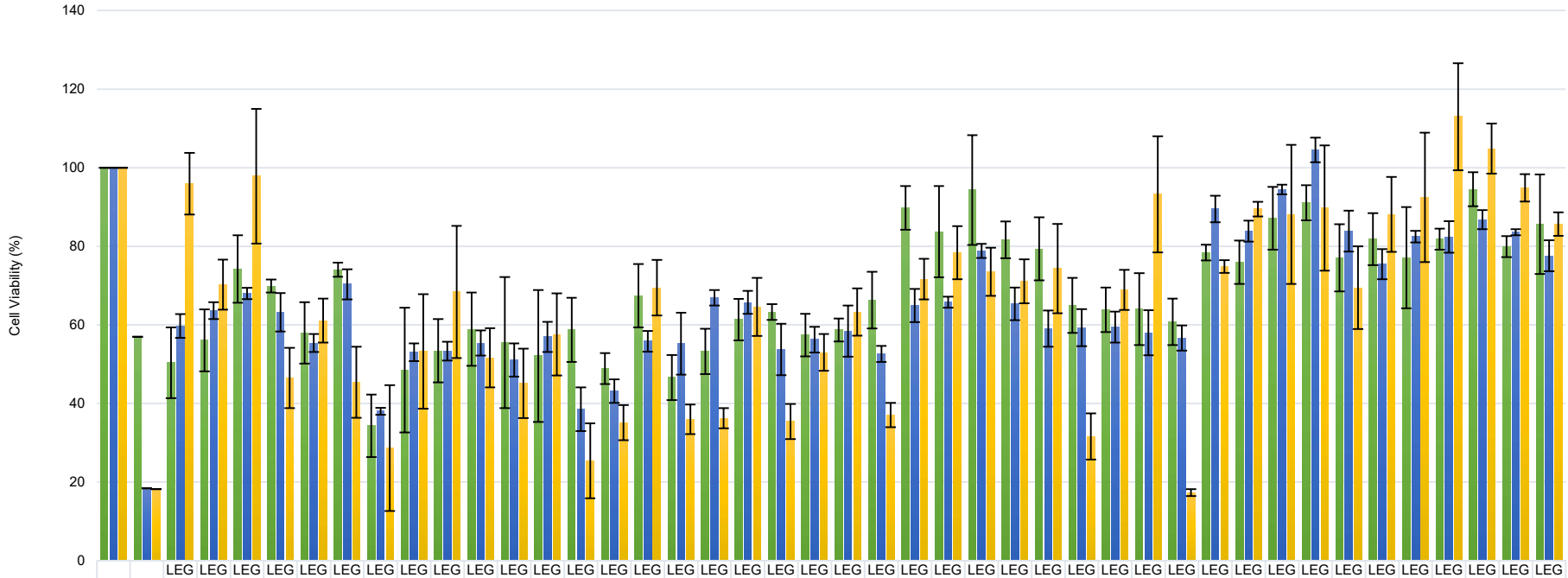
Table S2 – Cyanobacterial information regarding the extraction and fractionation of the strains present in the UC group. *after de-salting using C18 SPE cartridges; **might have presence of salts from the culture media.

Strain code	f.d. biomass (g)	Crude extract (mg)	Fractions (mg)		
			A	B	C
LEGE 14445	1.5	72.3	6.1	19.0	43.3
LEGE 11386	1.5	292.1	18.3	37.9	206.6**
LEGE 11468	5.3	996.5	15.8	24.9	204.0**
LEGE 06146	2.3	58.7	6.6	3.6	46.9
LEGE 06147	2.2	166.3	19.6	27.0	98.8
LEGE 12446	1.0	134.2	12.1	7.4	70.1
LEGE 12447	0.3	35.7	4.5	7.6	22.9
LEGE 12450	2.0	253.9	6.0	11.6	57.5
LEGE 12451	0.6	61.3	5.2	40.0	15.4
LEGE 12454	1.1	90.0	8.0	25.8	46.4
LEGE 12456	0.8	124.0	6.0	15.4	89.7
LEGE 11480	0.6	107.5	12.2	16.2	6.6*
LEGE 11477	6.0	689.7	2.9	3.8	3.2*
LEGE 11434	1.0	235.3	1.5	1.6	1.0*



	Control -	Control +	LEG E 1146 8 A	LEG E 1146 8 B	LEG E 1146 8 C	LEG E 1138 6 A	LEG E 1138 6 B	LEG E 1138 6C	LEG E 1444 5 A	LEG E 1444 5 B	LEG E 1444 5 C	LEG E 0614 6 A	LEG E 0614 6 B	LEG E 0614 6 C	LEG E 0614 7 A	LEG E 0614 7 B	LEG E 0614 7 C	LEG E 1244 7 A	LEG E 1244 7 B	LEG E 1244 7 C	LEG E 1245 1 A	LEG E 1245 1 B	LEG E 1245 1 C	LEG E 1245 4 A	LEG E 1245 4 B	LEG E 1245 4 C	LEG E 1244 6 A	LEG E 1244 6 B	LEG E 1244 6 C	LEG E 1245 6 A	LEG E 1245 6 B	LEG E 1245 6 C	LEG E 1245 0 A	LEG E 1245 0 B	LEG E 1245 0 C	LEG E 1148 0 A	LEG E 1148 0 B	LEG E 1148 0 C	LEG E 1147 7 A	LEG E 1147 7 B	LEG E 1147 7 C	LEG E 1143 4 A	LEG E 1143 4 B	LEG E 1143 4 C
A549	100,	5,4	109,	110,	113,	115,	96,2	86,1	83,7	78,6	59,4	114,	90,0	75,0	106,	86,6	90,3	125,	104,	92,9	103,	93,4	60,5	108,	118,	96,6	135,	123,	94,5	85,9	67,0	105,	100,	82,4	84,3	116,	100,	70,3	101,	92,4	86,0	110,	97,2	100,
RKO	100,	3,1	89,2	95,6	83,1	106,	104,	81,8	85,1	79,4	59,9	91,8	93,0	65,7	92,6	93,8	73,4	124,	97,0	82,2	88,5	76,3	51,4	97,2	111,	81,5	115,	106,	82,5	57,0	51,9	82,0	67,4	77,4	68,7	97,4	69,9	58,3	87,4	89,3	65,7	96,7	87,4	98,2
T47D	100,	25,3	39,3	46,7	62,8	35,7	36,7	51,2	26,0	37,6	45,2	46,0	40,6	45,4	23,3	35,5	48,1	47,3	51,6	66,5	44,5	44,4	43,9	42,7	61,0	68,7	54,5	49,2	64,6	47,7	45,5	46,8	47,2	73,3	79,2	71,9	88,0	72,4	52,7	67,8	65,3	65,6	68,2	62,5

Figure S2 – Cytotoxic assays with the cell lines A549 (lung carcinoma), RKO (colon carcinoma) and T47D (breast carcinoma) with the fractions A-C from all the 14 cyanobacterial strains belonging to the UC group. All fractions are at a concentration of 30 μg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 μg mL^{-1}).



	Control -	Control +	LEGE 11468A	LEGE 11468B	LEGE 11468C	LEGE 11386A	LEGE 11386B	LEGE 11386C	LEGE 14445A	LEGE 14445B	LEGE 14445C	LEGE 06146A	LEGE 06146B	LEGE 06146C	LEGE 06147A	LEGE 06147B	LEGE 06147C	LEGE 12447A	LEGE 12447B	LEGE 12447C	LEGE 12451A	LEGE 12451B	LEGE 12451C	LEGE 12454A	LEGE 12454B	LEGE 12454C	LEGE 12446A	LEGE 12446B	LEGE 12446C	LEGE 12456A	LEGE 12456B	LEGE 12456C	LEGE 12450A	LEGE 12450B	LEGE 12450C	LEGE 11480A	LEGE 11480B	LEGE 11480C	LEGE 11477A	LEGE 11477B	LEGE 11477C	LEGE 11434A	LEGE 11434B	LEGE 11434C
HepG2	100,	56,9	50,4	56,1	74,2	69,9	58,0	74,1	34,3	48,5	53,4	58,9	55,5	52,1	58,7	48,9	67,4	46,6	53,2	61,4	63,3	57,4	58,7	66,3	89,8	83,7	94,3	81,6	79,4	65,0	63,8	64,0	60,8	78,4	75,9	87,1	91,1	77,1	81,8	77,1	81,8	94,5	79,9	85,6
MG-63	100,	18,4	59,7	63,6	68,0	63,2	55,4	70,3	38,0	53,0	53,3	55,4	51,0	56,9	38,5	43,1	55,8	55,2	66,9	65,7	53,7	56,3	58,4	52,6	64,9	65,8	78,8	65,4	59,0	59,3	59,5	58,0	56,7	89,5	83,9	94,4	104,	83,9	75,4	82,5	82,4	86,8	83,6	77,6
SHSY-S5	100,	18,2	95,9	70,3	97,8	46,5	61,1	45,4	28,6	53,2	68,4	51,6	45,1	57,6	25,4	35,1	69,5	35,9	36,2	64,6	35,4	53,0	63,2	37,0	71,6	78,4	73,5	71,1	74,3	31,6	68,9	93,2	17,3	74,9	89,5	88,1	89,7	69,5	88,1	92,4	113,	104,	94,9	85,7

Figure S3 - Cytotoxic assays with the cell lines HepG2 (liver carcinoma), MG-63 (osteosarcoma) and SHSY-S5 (neuroblastoma) with the fractions A-C from all the 14 cyanobacterial strains belonging to the UC group. All fractions are at a concentration of 30 µg mL^{-1}, the negative control corresponds to 1% DMSO and the positive control to digitonin (25 µg mL^{-1}).

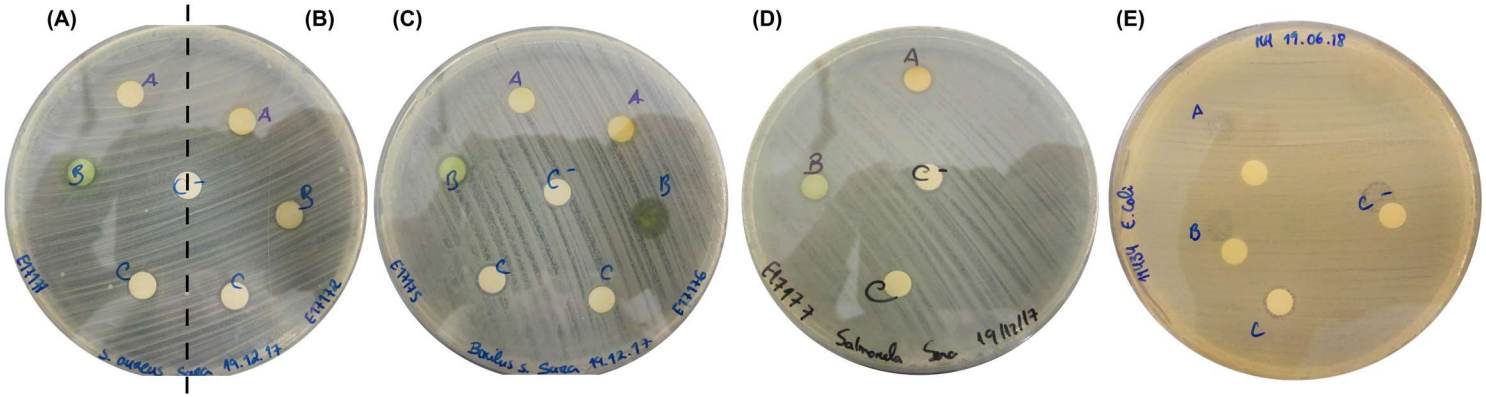


Figure S4 - Antimicrobial assays performed with the disc diffusion method. **(A)** Inhibition halos of *S. aureus* with fractions B and C from unidentified Pleurocapsales LEGE 06147; **(B)** Inhibition halos of *S. aureus* with fractions B and C from *Nostoc* sp. LEGE 12447**; (C)** Inhibition halos of *B. subtilis* by fractions B and C from *Desmonostoc muscorum* LEGE 12446; **(D)** Inhibition halos of *S. typhimurium* by fractions A and B from *Nostoc* sp. LEGE 12450; **(E)** Inhibition halo of *E. coli* by fraction C of *Cymatolege* sp. LEGE 11434. Exposure of the bacterial strains with 15 µg of each fraction; Negative control corresponds to 100% DMSO.

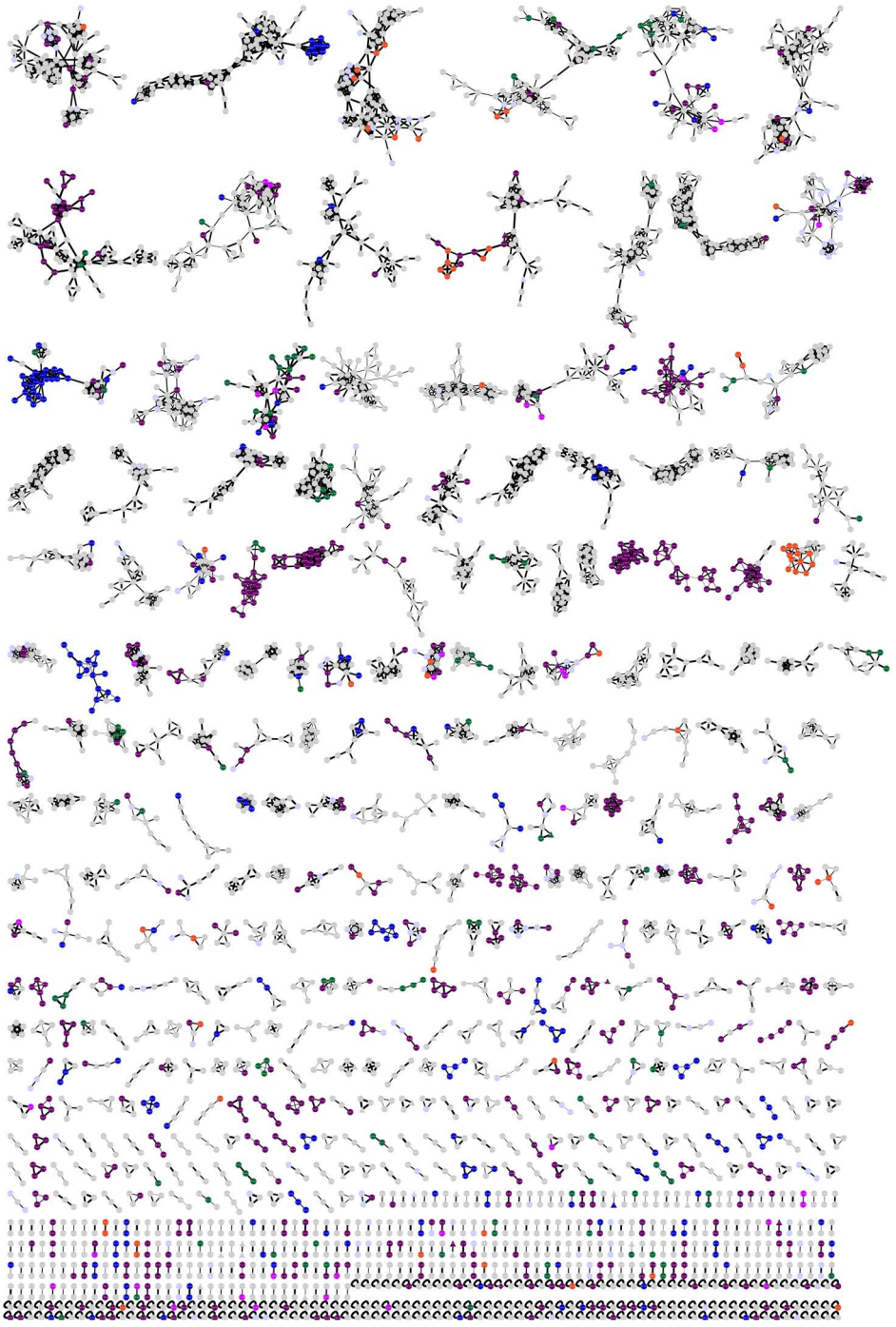


Figure S5 – GNPS Molecular Networking analysis, full network visualized using Cytoscape (v. 3.8.0.). The UC strains were divided into five groups according to phylogenetic distribution: purple (Nostocales symbiotic strains), orange (Non-symbiotic Nostocales strains), lavender (shared mass features between both symbiotic and non-symbiotic Nostocales groups), green (Nodosilineales strains), pink (Pleurocapsales strains), blue (Synechococcales and Leptolyngbyales) and grey (blanks, plus an actinobacteria extract, *Streptomyces* sp. KENR25). Label information set for parent mass. Spectral library hits are triangle shaped and in case of known cyanobacterial metabolites (mainly pigments), diamond shape. Manual inspection of single-strain networks was conducted, focusing on networks composed of a single color.

Table S3 – Annotation of the dereplication hits of the UC fractions C and B, from *Desmonostoc muscorum* LEGE 12446. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Desmonostoc muscorum LEGE 12446	NPAtlas	Anabaenopeptin 905	905.5123	$C_{45}H_{67}N_{11}O_9$		$M+H^+$	906.5190	$C_{45}H_{68}N_{11}O_9$	-0.705	Confirmed (Fr.C 6.2 min t_R)
		Anabaenopeptin 891	891.4967	$C_{44}H_{65}N_{11}O_9$		$M+H^+$	892.5034	$C_{44}H_{66}N_{11}O_9$	-0.604	Confirmed (Fr. C_5.7 min t_R)
	DNP	9,10-Dihydroxyoctadecanoic acid		$C_{19}H_{38}O_4$		$M+H^+$	331.2842	$C_{19}H_{39}O_4$	-0.290	Glyceryl palmitate (glycerol lipids)
		3-O-(9Z-Octadecenoyl), 5-hexadecanoyl		$C_{51}H_{76}O_8$		$M+H^+$	817.5627	$C_{51}H_{77}O_8$	1.730	
	Dereplicator VarQuest	Tolybyssidin A	1464.870	$C_{71}H_{116}N_{16}O_{17}$	694.881	$M+2H^{2+}$	1388.7570	$C_{68}H_{108}N_8O_{22}$	-0.229	Adduct of *m/z* 1388.757 $[M+H]^+$
		Lyngbyacyclamide A	1442.840	$C_{69}H_{114}N_{14}O_{19}$	702.390	$M+2H^{2+}$	1402.7730	$C_{69}H_{110}N_8O_{22}$	0.044	Adduct of *m/z* 1402.773 $[M+H]^+$
		Pahayokolide B	1302.650	$C_{63}H_{90}N_{12}O_{18}$	705.871	$M+2H^{2+}$	1388.7570	$C_{68}H_{108}N_8O_{22}$	-0.229	
		Lobocyclamide C	1369.807*	$C_{63}H_{111}N_{13}O_{20}$	680.866	$M+2H^{2+}$	1360.7263	-	-	Adduct of *m/z* 1360.726 $[M+H]^+$ Fr.C 6.3 min t_R
		Tolybyssidin B	1490.830	$C_{72}H_{114}N_{16}O_{16}S$	703.886	$M+2H^{2+}$	1406.7659	-	-	Adduct of *m/z* 1406.766 $[M+H]^+$ Fr.C 6.2 min t_R
		Lyngbyazothrin C_4'''-Demethoxy_5''-deacyl[1]	1282.680	-	703.394	$M+2H^{2+}$	1402.7730	$C_{69}H_{110}N_8O_{22}$	0.044	[1]Lyngbyazothrin B
		Tychonamide A_13-Demethoxy[2]	1455.760	-	691.856	$M+2H^{2+}$	1360.7263	-	-	[2]Tychonamide B
		Laxaphycin B2	1378.840	$C_{65}H_{114}N_{14}O_{18}$	637.865	$M+2H^{2+}$	1274.7248	-	-	Adduct of *m/z* 1274.725 $[M+H]^+$ Fr.C 6.6 min t_R
		Lobocyclamide B	1397.840	$C_{65}H_{115}N_{13}O_{20}$	703.886	$M+2H^{2+}$	1406.7659	-	-	
		Laxaphycin B	1394.840	$C_{65}H_{114}N_{14}O_{19}$	703.394	$M+2H^{2+}$	1402.7730	$C_{69}H_{110}N_8O_{22}$	0.044	
		Laxaphycin B_104R-Hydroxy	1410.830	-	703.394	$M+2H^{2+}$	1402.7730	$C_{69}H_{110}N_8O_{22}$	0.044	
		Lyngbyazothrin C_5''-deacyl[3]	1312.690	-	703.394	$M+2H^{2+}$	1402.7730	$C_{69}H_{110}N_8O_{22}$	0.044	[3]Lyngbyazothrin A
		Oscillamide C	956.512	$C_{49}H_{68}N_{10}O_{10}$	454.753	$M+2H^{2+}$	906.5196	-	-	Adduct of *m/z* 906.5196 $[M+H]^+$ Fr.C 5.0 min t_R; possibly halogenated
		Oscillatorin	1239.640	$C_{60}H_{85}N_{15}O_{14}$	647.834	$M+2H^{2+}$	1256.7145	$C_{63}H_{100}N_8O_{18}$	-0.397	Adduct of *m/z* 1256.715 $[M+H]^+$
		Laxaphycin D	1366.810	$C_{63}H_{110}N_{14}O_{19}$	705.872	$M+2H^{2+}$	1388.7570	$C_{68}H_{108}N_8O_{22}$	-0.229	
		Aeruginazole DA1497	1497.580	$C_{68}H_{91}N_{17}O_{14}S_4$	721.338	$M+2H^{2+}$	-	-	-	Fr.C 6.8-7.6 min t_R
		Tychonamide A	1485.780	$C_{73}H_{107}N_{13}O_{20}$	714.385	$M+2H^{2+}$	1388.7570	$C_{68}H_{108}N_8O_{22}$	-0.229	

Oscillamide B	868.427	$C_{41}H_{60}N_{10}O_9S$	446.756	$M+2H^{2+}$	892.5034	$C_{44}H_{66}N_{11}O_9$	-0.604	Anabenopeptin 891
Cyanothecamide A	829.379	$C_{38}H_{56}O_{10}N_9S$	849.487	$M+H^+$	849.4866	$C_{44}H_{65}N_8O_9$	-0.367	Anabaenopeptin 848 (or SA7); Fr.C 8.6 min t_R
Lobocyclamide C	1369.810	$C_{63}H_{111}N_{13}O_{20}$	637.373	$M+2H^{2+}$	1256.7145	$C_{63}H_{100}N_8O_{18}$	-0.397	
Patellamide C	762.298	$C_{37}H_{46}N_8O_6S_2$	406.249	$M+2H^{2+}$	-	-	-	Low abundant
Microcystin_2-[2-Amino-3-(4-hydroxy-2-cyclohexen-1-yl)propanoic_acid]_analogue	1048.560	-	451.228	$M+2H^{2+}$	863.5027	$C_{45}H_{67}N_8O_9$	0.171	Anabaenopeptin 863a (or SA8); Fr.C 8.9 min t_R
Planktocyclin	800.425	$C_{39}H_{60}N_8O_8S$	763.423	$M+H^+$	763.4231	-	-	Fr.C 6.6 min t_R
Malevamide C	1511.940	$C_{79}H_{125}N_{13}O_{16}$	706.428	$M+2H^{2+}$	1406.7659	-	-	
Aeruginazole DA1304	1304.460	$C_{61}H_{72}N_{14}O_{13}S_3$	720.863	$M+2H^{2+}$	1360.7263	-	-	
Patellamide F	762.298	$C_{37}H_{46}N_8O_6S_2$	451.228	$M+2H^{2+}$	863.5027	$C_{45}H_{67}N_8O_9$	0.171	

Table S4 - Annotation of the dereplication hits of the UC fraction C from *Nostoc* sp. LEGE 12450. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Nostoc sp. LEGE 12450		Wewakpeptin A	983.631	$C_{52}H_{85}N_7O_{11}$	561.760	$M+2H^{2+}$	1100.5304	$C_{55}H_{74}N_9O_{15}$	0.464	
		Largamide H	1198.660	$C_{57}H_{90}N_{12}O_{16}$	559.282					
		Grassypeptolide A	1101.540	$C_{56}H_{79}N_9O_{10}S_2$	569.742					
	Dereplicator VarQuest	Prenylagaramide A	1080.490	$C_{54}H_{68}N_{100}O_{14}$	569.743					
		[D-Asp3]-Microcystin-FR	1014.520	-	561.760					
		Nostofungicidine	1080.530	$C_{48}H_{76}N_{10}O_{18}$	569.742					
		Aeruginazole DA1274	1274.450	$C_{60}H_{70}N_{14}O_{21}S_3$	566.196	$M+2H^{2+}$	-	-	-	
	Dereplicator+	Diglycolipid 14	750.530	$C_{43}H_{74}O_{10}$	782.340	$M+H^+$	751.5350	$C_{43}H_{75}O_{10}$	-0.539	Fr.C 12.5-13.2 min t_R

Table S5 - Annotation of the dereplication hits of the UC fraction C from *Nostoc* sp. LEGE 12451. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Nostoc sp. LEGE 12451	NPAtlas	Microcystin-HphHph	1047.532	$C_{57}H_{73}N_7O_{12}$		$M+H^+$	1048.5354	$C_{52}H_{74}N_7O_{12}$	-3.411	
	Dereplicator VarQuest	Puwainaphycin A	1234.720	$C_{58}H_{98}N_{12}O_{17}$	625.335	$M+2H^{2+}$	625.3348	-	-	Low abundant
		Prepatellamide-B-formate	822.319	-	485.181	$M+2H^{2+}$	915.4388	$C_{56}H_{61}N_5O_5S$	0.020	Adduct of *m/z* 915.4388 $[M+H]^+$
		Puwainaphycin E	1192.710	$C_{56}H_{96}N_{12}O_{16}$	543.750	$M+2H^{2+}$	543.7458	-	-	Low abundance
		Dolastatin 16	878.515	$C_{47}H_{70}N_6O_{10}$	428.159	$M+2H^{2+}$	428.1595	-	-	Low abundance

Table S6 - Annotation of the dereplication hits of the UC fractions C, B and A from *Nostoc* sp. LEGE 12454. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Nostoc sp. LEGE 12454	DNP	Vitamin K1, 5' S-Hydroxy	466.345	$C_{31}H_{46}O_3$		$M+H^+$	467.3517	$C_{31}H_{47}O_3$	-0.624	Confirmed
	Dereplicator VarQuest	Microcystin LL	951.532	$C_{49}H_{73}N_7O_{12}$	856.456	$M+H^+$	856.4558	$C_{40}H_{66}N_5O_{15}$	1.001	Fr.C 5.6 min t_R; *m/z* 951.532 $[M+H]^+$ at 15.8 min t_R
		[D-Asp3,(E)-Dhb7]Microcystin_HtyHty_[D-Asp3,(E)-Dhb7]Microcystin_HtyHty	1065.510	-	465.688	$M+2H^{2+}$	876.4512	$C_{46}H_{58}N_{11}O_7$	-0.387	Fr.C 7.0 min t_R;
	Dereplicator+	trihydroxy-7',8'-didehydro-§-carotene	582.410	$C_{40}H_{54}O_3$	583.410	$M+H^+$	583.4149	$C_{40}H_{55}O_3$	0.494	Pigment; Fr.A 14.2 min t_R
		7,8-dihydrodiadinoxanthin	584.420	$C_{40}H_{56}O_3$	585.430	$M+H^+$	585.4306	$C_{40}H_{57}O_3$	-0.325	Pigment: Fr.A 16.7 min t_R

Table S7 - Annotation of the dereplication hits of the UC fraction C from *Nostoc* sp. LEGE 12456. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Nostoc sp. LEGE 12456	NPAtlas	Cyanopeptolin CP999	999.4589	$C_{51}H_{65}N_7O_{14}$		$M+H^+$	1000.4651	$C_{51}H_{66}N_7O_{14}$	-1.006	Confirmed;
		Cyanopeptolin CP1027	1027.4902	$C_{53}H_{69}N_7O_{14}$		$M+H^+$	1028.4972	$C_{53}H_{70}N_7O_{14}$	-0.288	Confirmed; $[M+H-H_2O]^+$ *m/z* 1010.4872
		Micropeptin C								
	DNP	Micropeptin C		$C_{53}H_{69}N_7O_{14}$		$M+H^+$	1028.4972	$C_{53}H_{70}N_7O_{14}$	-0.288	
	Dereplicator VarQuest	Micropeptin EI 992	992.497	$C_{48}H_{68}N_{10}O_{13}$	982.456	$M+H^+$	1000.4651	$C_{49}H_{64}N_{10}O_{13}$	0.247	Mass of micropeptin EI992 not found
		Micropeptin C	1027.490	$C_{53}H_{69}N_7O_{14}$	996.471	$M+H^+$	984.4698	$C_{47}H_{62}N_{13}O_{11}$	0.615	$[M+H-H_2O]^+$ *m/z* 966.4601; Fr.C 7.9 min t_R
		Prenylagaramide B	928.506	$C_{49}H_{68}N_8O_{10}$	960.508	$M+H^+$	978.5179	$C_{50}H_{72}N_7O_{13}$	-0.889	$[M+H-H_2O]^+$ *m/z* 960.5077
		Nostopeptolide A	1080.590	$C_{53}H_{80}N_{10}O_{14}$	1086.610	$M+H^+$	1086.6097	$C_{54}H_{80}N_{13}O_{11}$	-0.313	**New intense mass to isolate;** Fr.C 8 min t_R
		Micropeptin D	1055.520	$C_{55}H_{73}N_7O_{14}$	1038.520	$M+H^+$	1056.5291	$C_{55}H_{74}N_7O_{14}$	-0.26	Confirmed; $[M+H-H_2O]^+$ *m/z* 1038.5191; Fr.C 9.1 min t_R
		Anacyclamide A10	1052.530	$C_{49}H_{72}N_{12}O_{17}$	932.476	$M+H^+$	950.4864	$C_{46}H_{66}N_{10}O_{12}$	0.245	$[M+H-H_2O]^+$ *m/z* 932.4758
		Kawaguchipeptin B	1284.550	$C_{58}H_{76}N_{16}O_{18}$	622.272	$M+2H^{2+}$	622.2724	$C_{28}H_{39}N_5O_{11}$	-1.327	The adduct in the spectrum is $[M+H]^+$; **New intense mass to isolate**; Fr.C 5.5 min t_R
		Minutissamide C[1]	1189.670	$C_{55}H_{91}N_{13}O_{16}$	657.344	$M+2H^{2+}$	1313.6814	-	-	[1]Minutissamide D
		Puwainaphycin C	1226.670	$C_{56}H_{95}ClN_{12}O_{16}$	672.350	$M+2H^{2+}$	1311.6652	-	-	Low abundance
		Pahayokolide A	1471.760	$C_{72}H_{105}N_{13}O_{20}$	672.350	$M+2H^{2+}$	1311.6652	-	-	Low abundance
		Aeruginopeptin 95A	1145.530	$C_{56}H_{75}N_9O_{17}$	1010.490	$M+H^+$	1028.4972	$C_{53}H_{70}N_7O_{14}$	-0.288	
		Oscillacyclin	994.476	$C_{47}H_{66}N_{10}O_{14}$	994.492	$M+H^+$	1012.5008	$C_{50}H_{72}N_6O_{16}$	0.287	$[M+H-H_2O]^+$ *m/z* 994.4922; **new mass for isolation**; Fr.C 8.7 min t_R
		Dolastatin 1 Demethoxy,11-N-Me[2]	968.595	$C_{50}H_{80}N_8O_{11}$	541.712	$M+2H^{2+}$	541.7116	-	-	[2]Dolastatin 12; Exact mass not found
		Majusculamide C	984.590	$C_{50}H_{80}N_8O_{12}$	513.329	$M+2H^{2+}$	513.3286	$C_{25}H_{45}N_4O_7$	-0.36	The adduct in the sepctrum is $[M+H]^+$ **New intense mass to isolate**; Fr.C 7.2 min t_R
		Trichamide	1098.450	$C_{46}H_{66}N_{16}O_{12}S_2$	565.79	$M+2H^{2+}$	1086.6097	$C_{54}H_{80}N_{13}O_{11}$	-0.313	

Dereplicator+	Deca-4,5-diene-7,9-diyne-1,3-diol	162.070	$C_{10}H_{10}O_2$	163.070	$M+H^+$	163.0277	$C_8H_5NO_3$	4.828	Fr.C 0.6 min t_R; (Z)-2-cyano-3-(furan-2-yl)prop-2-enoate
	Micropeptin C	1027.490	$C_{53}H_{69}N_7O_{14}$	1028.500	$M+H^+$	1028.4972	$C_{53}H_{70}N_7O_{14}$	-0.288	

Table S8 - Annotation of the dereplication hits of the UC fractions B and C from unidentified Nostocales LEGE 14445. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
unidentified Nostocales LEGE 14445	DNP	5-Octacosenoic acid	422.4123	$C_{28}H_{54}O_2$		$M+H^+$	423.4186	$C_{28}H_{54}O_2$	-1.246	Confirmed
	Dereplicator VarQuest	Dolastatin I	516.215	$C_{24}H_{32}N_6O_5S$	214.089	$M+2H^{2+}$	-	-	-	Fr.C 0.6-19.5 min t_R
	Dereplicator+	Monoacylglyceryl-N,N,N-trimethylhomoseri	497.370	$C_{28}H_{51}NO_6$	498.380	$M+H^+$	498.3785	$C_{28}H_{52}NO_6$	-0.892	Confirmed

Table S9 - Annotation of the dereplication hits of the UC fractions A and B from *Nodosilinea* sp. LEGE 11468. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Nodosilinea sp. LEGE 11468	Dereplicator+	"Diadinochrome;5,8-Epoxy-7',8'-didehydr"	582.410	$C_{40}H_{54}O_3$	583.410	$M+H^+$	583.4149	$C_{40}H_{55}O_3$	0.562	Confirmed; Fr.A 12.5 min t_R; Pigment
		(2S,3S,4R)-2-[(2'R)-2'-hydroxydocosanoyl	655.610	$C_{40}H_{81}NO_5$	656.620	$M+H^+$	656.6184	$C_{40}H_{82}NO_5$	-0.504	Confirmed; Fr.A 13.9 min t_R; Spingolipd (ceraminde)
	Dereplicator VarQuest	Puwainaphycin D	1224.690	$C_{57}H_{97}ClN_{12}O_{15}$	623.870	$M+2H^{2+}$	623.8707	-	-	Fr.A 4.7 min t_R; Low abundance
		Dolastatin 12	968.595	$C_{50}H_{80}N_8O_{11}$	532.406	$M+2H^{2+}$	532.4065	-	-	Fr.A 17.2 min t_R
		Dolastatin_1_31-Methyl	982.610	-	435.286	$M+2H^{2+}$	435.2890	-	-	$[M+H]^+$ in the spectrum; Fr. 9.3 min t_R
		Dolastatin 14	1088.69	$C_{59}H_{92}N_8O_{11}$	552.432	$M+2H^{2+}$	552.4324	-	-	$[M+H]^+$ in the spectrum; Fr.A 14.9 min t_R

Table S10 - Annotation of the dereplication hits of the UC fractions A and B from *Cymatolege* sp LEGE 11434. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Cymatolege sp. LEGE 11434	DNP	Ribocyclophane E	794.4372	$C_{43}H_{67}ClO_{11}$		$M+H^+$	795.5285	-	-	
	Dereplicator VarQuest	Wewakpeptin C	1003.600	$C_{54}H_{81}N_7O_{11}$	550.417	$M+2H^{2+}$	550.4172	-	-	$[M+H]^+$ in the spectrum; Fr.A 16.1 min t_R
		Minutissamide A	1117.610	$C_{51}H_{83}N_{13}O_{15}$	573.400	$M+2H^{2+}$	573.3999	-	-	$[M+H]^+$ in the spectrum; Fr.B 10.2 min t_R
	Dereplicator+	Spheroidenone	582.440	$C_{41}H_{58}O_2$	583.450	$M+H^+$	583.4503	$C_{41}H_{59}O_2$	-1.093	Carotenoid; Fr.A 14.1 min t_R
		Thiothece-474	582.440	$C_{41}H_{58}O_2$	583.450	$M+H^+$	583.4503	$C_{41}H_{59}O_2$	-1.093	Carotenoid; Fr.A 14.1 min t_R
		N-[(1S,2S,3R)-2,3-Dihydroxy-1-(hydroxyme	655.610	$C_{40}H_{81}NO_5$	656.620	$M+H^+$	656.6172	$C_{40}H_{82}NO_5$	-2.438	Spingolipd (ceraminde); Fr.C 13.9 min t_R
		WB_3559B	652.500	$C_{37}H_{68}N_2O_7$	653.510	$M+H^+$	653.5097	$C_{36}H_{63}N_9O_2$	-0.388	Fr.C 13.2 min t_R
		GSL-1b	729.540	$C_{40}H_{75}NO_{10}$	730.550	$M+H^+$	730.5463	$C_{40}H_{76}NO_{10}$	-0.101	Fatty acid glycoside; Fr.C 13.0 min t_R
		DKxanthene_574	574.280	$C_{32}H_{38}N_4O_6$	575.280	$M+H^+$	575.4663	$C_{32}H_{59}N_6O_3$	3.43	Alkaloid; Fr.C 12.4 min t_R

Table S11 - Annotation of the dereplication hits of the UC fractions A and B from *Myxosarcina* sp. LEGE 06146 (Pleurocapsales). Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Myxosarcina sp. LEGE 06146	Dereplicator VarQuest	Palau'amide	851.504	$C_{46}H_{69}N_5O_{10}$	435.289	$M+2H^{2+}$	435.2890	-	-	$[M+H]^+$ in the spectrum; Fr.B 9.4 min t_R
		Malevamide B	1428.940	$C_{76}H_{124}N_{14}O_{14}$	362.987	$M+2H^{2+}$	362.9864	-	-	Charge 4 in the spectrum; Fr.B 1.0 min t_R

Table S12 - Annotation of the dereplication hits of the UC fractions A and B from *Romeriopsis navalis* LEGE 11480 (Lepotlyngbyales). Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental Mass	calcd. M.F	Δppm	Observations
Romeriopsis navalis LEGE 11480	DNP	Glycerol 1,2-dialkanoates	750.5282	$C_{43}H_{74}O_{10}$		$M+H^+$	751.5335	$C_{43}H_{75}O_{10}$	-2.588	Confirmed
	Dereplicator VarQuest	Lobocyclamide B	1383.820	-	699.392	$M+2H^{2+}$	699.3928	-	-	[M+H]+ in the spectrum

Table S13 - Annotation of the dereplication hits of the UC fractions A and B from cf. *Phormidesmis* sp. LEGE 11477 (Synechocalles). Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental Mass	calcd. M.F	Δppm	Observations
cf. *Phormidesmis* sp. LEGE 14777	DEREPLICATOR	Anabaenopeptin D	827.420	$C_{44}H_{57}N_7O_9$	828.430	$M+H^+$	828.4287	$C_{44}H_{58}N_7O_9$	-0.426	Confirmed (Fr.A 7.7 min t_R)
		Anabaenopeptin A	843.420	$C_{44}H_{57}N_7O_{10}$	844.420	$M+H^+$	844.4240	$C_{44}H_{58}N_7O_9$	0.051	Confirmed (Fr.A 6.7 min t_R)
	DEREPLICATOR Varquest	Anabaenopeptin NZ825	825.443	$C_{45}H_{59}N_7O_8$	812.433	$M+H^+$	812.4334	$C_{44}H_{58}N_7O_8$	-0.933	Loss of CH_2; Fr.A 8.8 min t_R
		Anabaenopeptin NZ825 4'-Hydroxy	841.437	$C_{45}H_{59}N_7O_8$	828.429	$M+H^+$	828.4287	$C_{44}H_{58}N_7O_9$	-0.426	Anabaenopeptin D
		Anabaenopeptin D	827.422	$C_{44}H_{57}N_7O_9$	844.424	$M+H^+$	844.4240	$C_{44}H_{58}N_7O_9$	0.051	Anabaenopeptin A
		Majusculamide C	970.574	$C_{49}H_{78}N_8O_{12}$	425.775	$M+2H^{2+}$	425.7751	-	-	Fr.B 0.7 min t_R
		Sch 378167 5',5"-Diamide	1135.660	$C_{54}H_{87}N_9O_{15}$	579.845	$M+2H^{2+}$	579.8452	-	-	Fr.C 4.7 min t_R
		Sch 378167 5'-Amide	1136.650	$C_{54}H_{87}N_9O_{15}$	601.858	$M+2H^{2+}$	601.8582	-	-	Fr.C 4.7 min t_R
		Dolastatin G 37-O-De-Me	1058.690	$C_{59}H_{94}N_6O_{13}$	599.410	$M+2H^{2+}$	599.4098	-	-	Fr.B 11.8 min t_R; $[M+H]^+$ in the spectrum;
	DEREPLICATOR+	Anabaenopeptin A	843.420	$C_{44}H_{57}N_7O_{10}$	844.420	$M+H^+$	844.4240	$C_{44}H_{58}N_7O_9$	0.051	Confirmed (Fr.A 6.7 min t_R)
		Zeaxanthin_monomethyl_ether; 3-Hydroxy-3	582.440	$C_{41}H_{58}O_2$	583.450	$M+H^+$	583.4508	$C_{41}H_{59}O_2$	-0.27	Confirmed; Fr.A_14.0 min t_R)

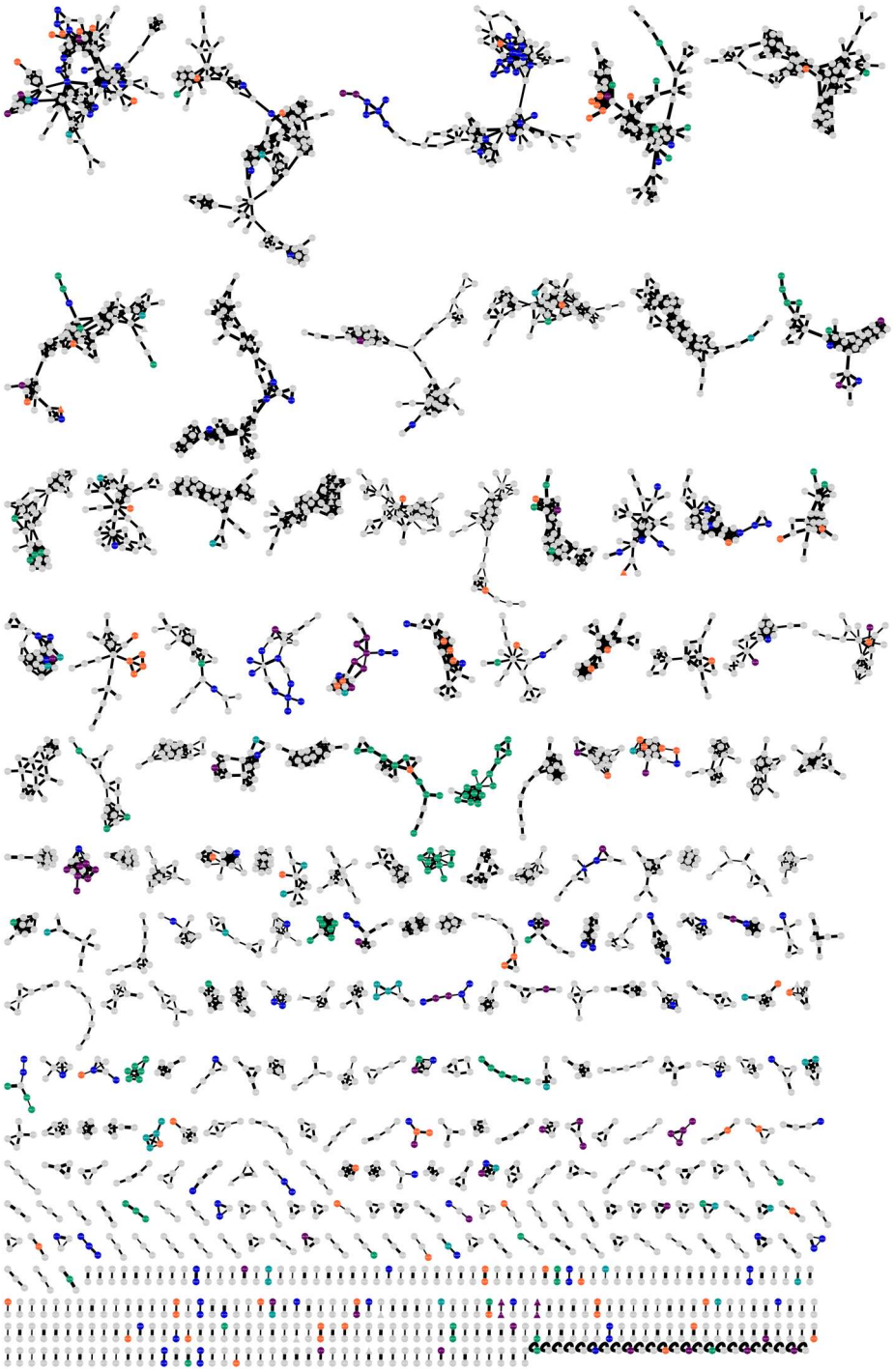


Figure S6 - GNPS Molecular Networking analysis, full network visualized using Cytoscape (v. 3.8.0.). The EC strains were divided into five groups according to phylogenetic distribution: purple (Synechococcales), blue (Oscillatoriales_also including Coleofasciculales since the strains were collected from the same environment), green (Oculatellates), orange (Nostocales), pink (Chroococcidiopsidales and Leptolyngbyales) and grey (blanks, plus an actinobacteria extract, *Streptomyces* sp. KENR25). Label information set for parent mass. Spectral library hits are triangle shaped and in case of known cyanobacterial metabolites (mainly pigments), diamond shape. Manual inspection of single-strain networks was conducted, focusing on networks composed of a single color.

Table S14 - Annotation of the dereplication hits of the EC fractions A and B from *Calenema singularis* LEGE 06188. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Calenema singularis LEGE 06188	DNP	Oxooctacosanoic acid	438.4073	$C_{28}H_{54}O_3$		$M+H^+$	439.4142	$C_{28}H_{55}O_3$	-0.778	Confirmed
		glycerol 1,2 dialkanoates	568.5067	$C_{35}H_{68}O_5$		$M+H^+$	569.5138	$C_{35}H_{69}O_6$	-0.214	Confirmed
	Dereplicator VarQuest	Majusculamide C	970.574	$C_{49}H_{78}N_8O_{12}$	498.803	$M+2H^{2+}$	996.5998	$C_{50}H_{80}N_{10}O_{11}$	-0.486	Fr.B 0.7 min t_R
		Dolastatin 1_11-Nme	998.605	-	551.424	$M+2H^{2+}$	551.4246	-	-	Fr.A 14.0 min t_R; $[M+H]^+$ adduct in the spectrum
	Dereplicator +	Porpoisamide B	598.373	$C_{33}H_{50}N_4O_6$	599.390	$M+H^+$	599.3902	$C_{31}H_{55}N_2O_9$	-0.096	Fr.B 0.6 min t_R

Table S15 - Annotation of the dereplication hits of the EC fractions A and B from unidentified filamentous *Synechococcales* LEGE 06118. Dereplication tools used were NPAtlas, Dictionary of Natural Products (DNP), Dereplicator, Dereplicator+ and Dereplicator VarQuest. The hits were searched using NPAtlas and Pubchem to retrieve the known compound information's.

Strain	Platform	Hit	Exact Mass	M.F.	Spectrum Mass	Adduct	Experimental mass	calcd. M.F	Δppm	Observations
Unidentified filamentous Synechococcales LEGE 06118	Dereplicator+	Mycosporine-gly_val	344.158	$C_{15}H_{24}N_2O_7$	345.167	$M+H^+$	345.1670	$C_{16}H_{21}N_6O_3$	0.159	Fr.B 6.9-8.3 min t_R
		Cryptophycin 175	672.237	$C_{35}H_{42}Cl_2N_2O_7$	673.247	$M+H^+$	673.2469	$C_{30}H_{37}N_6O_{12}$	0.762	Is not halogenated

2.6. References

1. Gaysina, L. A.; Saraf, A.; Singh, P., Chapter 1 - Cyanobacteria in Diverse Habitats. In *Cyanobacteria*, Mishra, A. K.; Tiwari, D. N.; Rai, A. N., Eds. Academic Press: 2019; pp 1-28.

2. Popin, R. V.; Alvarenga, D. O.; Castelo-Branco, R.; Fewer, D. P.; Sivonen, K., Mining of Cyanobacterial Genomes Indicates Natural Product Biosynthetic Gene Clusters Located in Conjugative Plasmids. *Frontiers in Microbiology* **2021,** *12*.

3. Tan, L. T., Pharmaceutical agents from filamentous marine cyanobacteria. *Drug Discov Today* **2013,** *18* (17-18), 863-71.

4. Khalifa, S. A. M.; Shedid, E. S.; Saied, E. M.; Jassbi, A. R.; Jamebozorgi, F. H.; Rateb, M. E.; Du, M.; Abdel-Daim, M. M.; Kai, G. Y.; Al-Hammady, M. A. M.; Xiao, J.; Guo, Z.; El-Seedi, H. R., Cyanobacteria-From the Oceans to the Potential Biotechnological and Biomedical Applications. *Mar Drugs* **2021,** *19* (5).

5. Kehr, J. C.; Gatte Picchi, D.; Dittmann, E., Natural product biosyntheses in cyanobacteria: A treasure trove of unique enzymes. *Beilstein J Org Chem* **2011,** *7*, 1622-35.

6. Dittmann, E.; Gugger, M.; Sivonen, K.; Fewer, D. P., Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol* **2015,** *23* (10), 642-652.

7. Calteau, A.; Fewer, D. P.; Latifi, A.; Coursin, T.; Laurent, T.; Jokela, J.; Kerfeld, C. A.; Sivonen, K.; Piel, J.; Gugger, M., Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* **2014,** *15* (1), 977.

8. Selão, T. T., Exploring cyanobacterial diversity for sustainable biotechnology. *Journal of Experimental Botany* **2022,** *73* (10), 3057-3071.

9. Hemmerling, F.; Piel, J., Strategies to access biosynthetic novelty in bacterial genomes for drug discovery. *Nature Reviews Drug Discovery* **2022,** *21* (5), 359-378.

10. van der Hooft, J. J. J.; Mohimani, H.; Bauermeister, A.; Dorrestein, P. C.; Duncan, K. R.; Medema, M. H., Linking genomics and metabolomics to chart specialized metabolic diversity. *Chemical Society Reviews* **2020,** *49* (11), 3297-3314.

11. Ramos, V.; Morais, J.; Castelo-Branco, R.; Pinheiro, Â.; Martins, J.; Regueiras, A.; Pereira, A. L.; Lopes, V. R.; Frazão, B.; Gomes, D.; Moreira, C.; Costa, M. S.; Brûle, S.; Faustino, S.; Martins, R.; Saker, M.; Osswald, J.; Leão, P. N.; Vasconcelos, V. M., Cyanobacterial diversity held in microbial biological resource centers as a biotechnological asset: the case study of the newly established LEGE culture collection. *Journal of Applied Phycology* **2018,** *30* (3), 1437-1451.

12. Shih, P. M.; Wu, D.; Latifi, A.; Axen, S. D.; Fewer, D. P.; Talla, E.; Calteau, A.; Cai, F.; Tandeau de Marsac, N.; Rippka, R.; Herdman, M.; Sivonen, K.; Coursin, T.; Laurent, T.; Goodwin, L.; Nolan, M.; Davenport, K. W.; Han, C. S.; Rubin, E. M.; Eisen, J. A.; Woyke, T.; Gugger, M.; Kerfeld, C. A., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences* **2013,** *110* (3), 1053-1058.

13. Leão, P. N.; Pereira, A. R.; Liu, W. T.; Ng, J.; Pevzner, P. A.; Dorrestein, P. C.; König, G. M.; Vasconcelos, V. M.; Gerwick, W. H., Synergistic allelochemicals from a freshwater cyanobacterium. *Proc Natl Acad Sci U S A* **2010,** *107* (25), 11183-8.

14. Gutiérrez-del-Río, I.; Brugerolle de Fraissinette, N.; Castelo-Branco, R.; Oliveira, F.; Morais, J.; Redondo-Blanco, S.; Villar, C. J.; Iglesias, M. J.; Soengas, R.; Cepas, V.; Cubillos, Y. L.; Sampietro, G.; Rodolfi, L.; Lombó, F.; González, S. M. S.; López Ortiz, F.; Vasconcelos, V.; Reis, M. A., Chlorosphaerolactylates A–D: Natural Lactylates of Chlorinated Fatty Acids Isolated from the Cyanobacterium Sphaerospermopsis sp. LEGE 00249. *Journal of Natural Products* **2020,** *83* (6), 1885-1890.

15. Brito, Â.; Gaifem, J.; Ramos, V.; Glukhov, E.; Dorrestein, P. C.; Gerwick, W. H.; Vasconcelos, V. M.; Mendes, M. V.; Tamagnini, P., Bioprospecting Portuguese Atlantic coast cyanobacteria for bioactive secondary metabolites reveals untapped chemodiversity. *Algal Research* **2015,** *9*, 218-226.

16. Leão, P. N.; Nakamura, H.; Costa, M.; Pereira, A. R.; Martins, R.; Vasconcelos, V.; Gerwick, W. H.; Balskus, E. P., Biosynthesis-Assisted Structural Elucidation of the Bartolosides, Chlorinated Aromatic Glycolipids from Cyanobacteria. *Angewandte Chemie International Edition* **2015,** *54* (38), 11063-11067.

17. Reis, J. P. A.; Figueiredo, S. A. C.; Sousa, M. L.; Leão, P. N., BrtB is an O-alkylating enzyme that generates fatty acid-bartoloside esters. *Nature Communications* **2020,** *11* (1), 1458.

18. Almeida, J. R.; Freitas, M.; Cruz, S.; Leão, P. N.; Vasconcelos, V.; Cunha, I., Acetylcholinesterase in Biofouling Species: Characterization and Mode of Action of Cyanobacteria-Derived Antifouling Agents. *Toxins (Basel)* **2015,** *7* (8), 2739-56.

19. Lopes, V. R.; Fernández, N.; Martins, R. F.; Vasconcelos, V., Primary screening of the bioactivity of brackishwater cyanobacteria: toxicity of crude extracts to Artemia salina larvae and Paracentrotus lividus embryos. *Mar Drugs* **2010,** *8* (3), 471-82.

20. Lopes, V. R.; Vasconcelos, V. M., Bioactivity of benthic and picoplanktonic estuarine cyanobacteria on growth of photoautotrophs: inhibition versus stimulation. *Mar Drugs* **2011,** *9* (5), 790-802.

21. Costa, M.; Garcia, M.; Costa-Rodrigues, J.; Costa, M. S.; Ribeiro, M. J.; Fernandes, M. H.; Barros, P.; Barreiro, A.; Vasconcelos, V.; Martins, R., Exploring bioactive properties of marine cyanobacteria isolated from the Portuguese coast: high potential as a source of anticancer compounds. *Mar Drugs* **2013,** *12* (1), 98-114.

22. 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. *Nature Biotechnology* **2016,** *34* (8), 828-837.

23. Strunecký, O.; Ivanova, A. P.; Mareš, J., An updated classification of cyanobacterial orders and families based on phylogenomic and polyphasic analysis(1). *J Phycol* **2022**.

24. Nandagopal, P.; Steven, A. N.; Chan, L.-W.; Rahmat, Z.; Jamaluddin, H.; Mohd Noh, N. I., Bioactive Metabolites Produced by Cyanobacteria for Growth Adaptation and Their Pharmacological Properties. *Biology* **2021,** *10* (10), 1061.

25. Stevenson, C. S.; Capper, E. A.; Roshak, A. K.; Marquez, B.; Eichman, C.; Jackson, J. R.; Mattern, M.; Gerwick, W. H.; Jacobs, R. S.; Marshall, L. A., The identification and characterization of the marine natural product scytonemin as a novel antiproliferative pharmacophore. *J Pharmacol Exp Ther* **2002,** *303* (2), 858-66.

26. Numan, M.; Bashir, S.; Mumtaz, R.; Tayyab, S.; Rehman, N. U.; Khan, A. L.; Shinwari, Z. K.; Al-Harrasi, A., Therapeutic applications of bacterial pigments: a review of current status and future opportunities. *3 Biotech* **2018,** *8* (4), 207.

27. Verma, S.; Thapa, S.; Siddiqui, N.; Chakdar, H., Cyanobacterial secondary metabolites towards improved commercial significance through multiomics approaches. *World Journal of Microbiology and Biotechnology* **2022,** *38* (6), 100.

28. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003,** *13* (11), 2498-504.

29. Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L.-F.; Ninomiya, A.; Takada, K.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of peptidic natural products through database search of mass spectra. *Nat Chem Biol* **2017,** *13* (1), 30-37.

30. Gurevich, A.; Mikheenko, A.; Shlemov, A.; Korobeynikov, A.; Mohimani, H.; Pevzner, P. A., Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat Microbiol* **2018,** *3* (3), 319-327.

31. Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L. F.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of microbial metabolites through database search of mass spectra. *Nat Commun* **2018,** *9* (1), 4035.

32. van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castaño-Espriu, L.; Chang, C.; Clark, T. N.; Cleary Little, J. L.; Delgadillo, D. A.; Dorrestein, P. C.; Duncan, K. R.; Egan, J. M.; Galey, M. M.; Haeckl, F. P. J.; Hua, A.; Hughes, A. H.; Iskakova, D.; Khadilkar, A.; Lee, J.-H.; Lee, S.; LeGrow, N.; Liu, D. Y.; Macho, J. M.; McCaughey, C. S.; Medema, M. H.; Neupane, R. P.; O'Donnell, T. J.; Paula, J. S.; Sanchez, L. M.; Shaikh, A. F.; Soldatou, S.; Terlouw, B. R.; Tran, T. A.; Valentine, M.; van der Hooft, J. J. J.; Vo, D. A.; Wang, M.; Wilson, D.; Zink, K. E.; Linington, R. G., The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Central Science* **2019,** *5* (11), 1824-1833.

33. Dictionary of Natural Products 32.0. https://dnp.chemnetbase.com (accessed 20 Dec 2021).

34. LIPID MAPS. https://www.lipidmaps.org (accessed March).

35. Kotai, J., Instructions for preparation of modified nutrient solution Z8 for algae. *Oslo: Norwegian Institute for Water Research* **1972,** *5*.

36. PubChem. https://pubchem.ncbi.nlm.nih.gov (accessed March).

Chapter III

Structure and biosynthesis of desmamides A-C, lipoglycopeptides from the endophytic cyanobacterium *Desmonostoc muscorum* LEGE 12446

3. Structure and biosynthesis of desmamides A-C, lipoglycopeptides from the endophytic cyanobacterium *Desmonostoc muscorum* LEGE 12446

3.1. Introduction

Cyanobacteria are photoautotrophic bacteria found in a wide range of environments and exist in single-celled, colonial, filamentous and branched filamentous forms.[1] These organisms can also differentiate specialized cells[1], such as heterocysts, which carry out nitrogen fixation[2] and are the hallmark of the Nostocales, a late-branching order within the Cyanobacteria phylum. Nostocalean cyanobacteria typically have large genomes, complex morphological features and can either be free-living, obligate symbionts of the water fern *Azolla* or facultative symbionts of other organisms, such as cycads, mosses, liverworts, and fungi.[1,3] In cycads, the cyanobacterial holobiont forms a conspicuous circular layer in the coralloid roots.[4] Free-living cyanobacterial holobionts are recruited by the plant and can invade their host at different stages of root development.[3,4] Once inside the roots, the cyanobacteria increase their heterocyst differentiation frequency to fix nitrogen for the host, and establish a permanent symbiosis under complete darkness.[3,4] Cyanobacteria within the order Nostocales present relatively large genomes (average 7.6 Gb)[5], which typically contain between 10 and 30 biosynthetic gene clusters (BGCs), encoding the production of several classes of natural products.[6] These include terpenes, alkaloids, peptides of ribosomal (RiPPs) or non-ribosomal (NRPs) origin, polyketides and hybrids thereof.[6,7] The chemical talent of members of the Nostocales was well-recognized before the genomic era: early discoveries of natural products from this order include remarkable structures such as the cylindrocyclophanes,[8] hapalindoles and related alkaloids,[9] cryptophycins,[10] and scytonemin.[11] Symbiotic Nostocales are known to be rich producers of secondary metabolites,[12,13] some of which have been shown to mediate interactions with the host.[14,15] For example, the facultative cycad symbiont cyanobacterium *Nostoc punctiforme* PCC 73102 produces nostopeptolides, cyclic lipopeptides whose production is regulated by the plant host and are involved in cross-talk between the two organisms.[15] A large variety of cyanobacteria and, in particular, members of the order Nostocales, produce cyclic lipopeptides.[16] These compounds are usually synthesized by hybrid polyketide synthase/non-ribosomal

peptide synthetase (PKS/NRPS) pathways, which yields a fatty acyl moiety (that can be modified to varying degrees) connected to a peptidic component.[16] Cyanobacterial cyclic lipopeptides are structurally diverse and typically exhibit antifungal activity; however, their precise ecological roles are mostly unknown.[16]

A very small fraction of natural cyclic lipopeptides – from cyanobacteria or other organisms – are glycosylated and referred to as lipoglycopeptides. This designation has been used to encompass both glycosylated lipopeptides and acylated glycopeptides.[17,18] Natural lipoglycopeptides are rare and are grouped in a small number of structural classes: the cyanobacterial hassallidins/balticidins,[19,20] the actinobacterial teicoplanin/teichomycins,[21] mannopeptimycins (γ, δ and ε),[22] ramoplanins,[23] and gausemycins,[18] as well as occidiofungins produced by a Gram-negative *Burkholderia* strain.[24] Potent antibacterial or antifungal activities are associated with each of these compound classes; teicoplanin and related semi-synthetic lipoglycopeptides are used in the clinic to treat serious infections caused by Gram-positive bacteria.[25] Looking with more detail into lipoglycopeptides produced by cyanobacteria, both hassalidins and balticidins share a peptide core ring of eight amino acids, with a exocyclic amino acid linked to a fatty acyl chain, as shown on figure 1.[19,20] In the case of hassalidins (A-E), the variants show modifications on the length of the fatty acid chain (C14-C18) and on the degree of glycosylation (1-3), which occurs on the N-Me-thr residue (1 glycosylation), as well as on the β-hydroxyl group of the fatty acid chain (1-2 glycosylations). Also, the monosaccharide units vary from mannose, rhamnose, pentose, N-acetylhexosamine and diacetylhexose.[19] The balticidins (A-D) variants share the same fatty acid chain and glycosylation pattern, varying on the chlorination of the fatty acyl chain (A and B) and on the structure of the lipopeptide (linear or cyclic). The monosaccharides present in the structure are mannose and the disaccharide arabinose-galacturonic acid.[20] Glycolysations occur on the same positions of hassallidins.

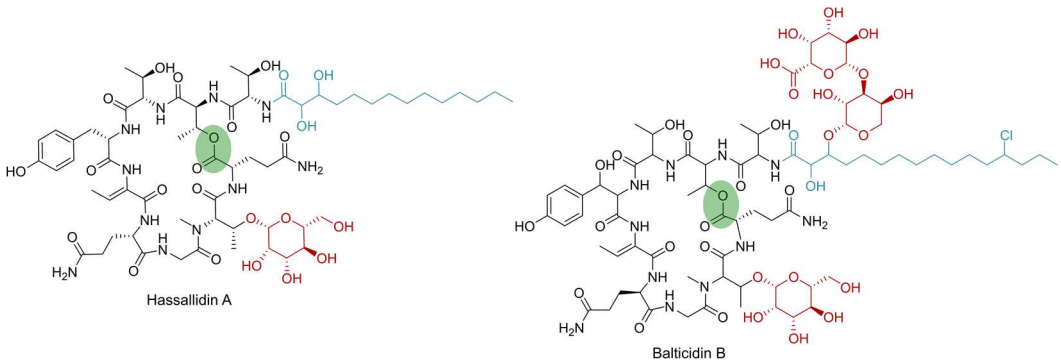


Figure 1 . Cyanobacterial lipoglycopeptides, hassalidin A and balticidin B. Sugar moieties – red; fatty acid chain – light blue; ring closure (ester bond) – green shade.

The chemical exploration of the nostocalean cyanobacterium *Desmonostoc muscorum* LEGE 12446, a plant facultative symbiont isolated from a cycad (*Cycas revoluta*) coralloid root, was undertaken. Metabolomic analysis of a pre-fractionated organic extract of this cyanobacterium led to the detection and eventual isolation of two abundant metabolites, desmamides A (**1**) and B (**2**), which represent a new structural class of lipoglycopeptides. Their aglycone – desmamide C (**3**) – was also isolated in this effort. The biosynthesis of **1**-**3** was investigated and a putative biosynthetic gene cluster (BGC) associated with desmamide production – *dsm* – was identified. Compounds **1** and **2** showed moderate cytotoxic activity towards cancer cell lines and inhibited the growth of the plant pathogenic bacterium *Xanthomonas campestris* with moderate potency.

3.2. Results and discussion

3.2.1. MS-guided isolation of desmamides A-C (**1**-**3**)

As a result of the screening assay performed in chapter II, the exploration of the xenic unicyanobacterial culture of *D. muscorum* LEGE 12446 was performed. Global Natural Products Social Molecular Networking (GNPS)[26] analysis of the resulting fractions revealed the presence of a cluster comprised of five mass features in the fraction eluting with 100% MeOH, two of which had close retention time values and $[M+H]^+$ ions with *m/z* 1388.750, and 1256.714 (Figure S1; Table S1). Dereplication using GNPS tools Dereplicator,[27] Dereplicator VarQuest,[28] and Dereplicator+[29] as well as manual searches on the Dictionary of Natural Products database[30] and Natural Products Atlas[31] returned no hits, confirming the novelty of the compounds. To obtain sufficient amount of the natural products for NMR-based structure elucidation and biological activity testing, the strain was grown in large-scale (160 L), which yielded 70.3 g of dry biomass. Repeated percolation with CH_2Cl_2/MeOH (2:1, v/v) afforded an organic extract (14.7 g) that contained the target mass features, as revealed by Liquid Chromatography coupled to High Resolution Electrospray Ionization Mass Spectrometry (LC-HRESIMS) analysis, enabling a MS-guided isolation approach (Figure S2). Following extraction, a normal-phase Vacuum Liquid Chromatography (VLC) was performed to obtain 13 fractions of increasing polarity. LC-HRESIMS analysis indicated that the compounds were abundant in fractions eluting with 1:2 and 3:7 EtOAc:MeOH (v/v). These fractions were combined (2.90 g) and LC-HRESIMS analysis was used to follow the targeted mass features over a series of chromatographic procedures. Normal phase flash chromatography, reversed-phase chromatography on a pre-packed SPE cartridge, semipreparative and analytical

HPLC were necessary to obtain pure desmamides A (**1**, 26.3 mg) and B (**2**, 20.4 mg), together with their aglycone, desmamide C (**3**, 4.0 mg), as shown in Figure 2. Compounds **1** and **3** correspond to the two GNPS-detected mass features targeted, while compound **2** was obtained during isolation of **1**. A considerable amount of **1**-**3** remained in adjacent chromatography fractions – an estimate based on LC-HRESIMS quantification (Figure S3; Table S2), revealed that the combined mass of these metabolites reaches 235 mg in the initial extract (1.6% of dry cell material).

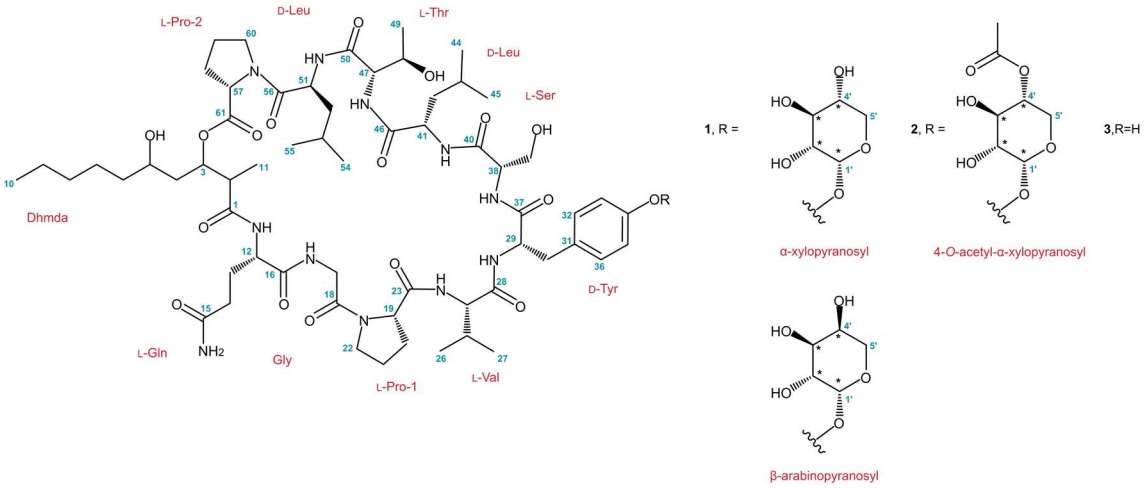


Figure 2 - Structure of desmamides A-C (**1**-**3**). Desmamide A is represented with both monosaccharide units, α-xylopyranosyl and β-arabinopyranosyl.

3.2.2. Structure elucidation of **1**-**3**

Structure elucidation efforts were focused on the most abundant metabolite – **1**. High Resolution Electrospray Ionization Mass Spectrometry (HRESIMS) analysis of **1** showed a *m/z* 1388.7577 for the $[M+H]^+$ ion (Figure S4). This was consistent with a molecular formula of $C_{66}H_{105}N_{11}O_{21}$ and the presence of 20 degrees of unsaturation. Inspection of the ^{1}H and ^{13}C APT NMR data (400 MHz, DMSO-d_6) of **1**, present in table 1 (Figures S5-S6), quickly revealed the presence of the anomeric proton (H-1', δ_C 97.6, δ_H 5.31 $J_{1'-2'}$ 3.5 Hz) of a sugar unit. Multiple resonances consistent with the exchangeable NH protons (δ_H 8.5-7.50), amide carbons (δ_C 175-165), and alpha positions of amino acid residues (δ_C 60-40, δ_H 4.5-3.5) were observed. Furthermore, an aliphatic terminal methyl group triplet (CH_3-10, δ_C 14.0, δ_H 0.86) and a small methylene envelope (δ_H 1.28-1.25) were clearly identified. These observations collectively suggested that **1** was a lipoglycopeptide. Combined analysis of 1D and 2D (HSQC, HMBC and COSY) NMR

data for **1**, presented in figure 3 (Figures S7-S9), indicated that a major conformer and at least one minor conformer existed in solution, complicating the interpretation of the NMR spectra. Still, its analysis enabled the assembly of several structural fragments, indicated in Figure 3A, including those of a number of proteinogenic amino acid residues (Thr, Ser, Gln, two Pro, Val, Gly and two Leu). Furthermore, it was possible to establish the presence of a Tyr residue featuring *O*-glycosylation with an aldopentose pyranose ring, as the anomeric proton showed an HMBC correlation to C-34 and NOE correlation to H-33/35 (Figures S8, S10). Determination of the nature of the sugar residue connected to Tyr in **1** was hampered by extensive overlap of the sugar proton multiplet signals, limiting the measure of the associated coupling constants. Still, the clear $J_{1'\text{-}2'}$ value of 3.5 Hz for the anomeric proton (H-1') indicated *α*-glycosylation. To identify the chemical nature of the sugar unit in **1**, an acid hydrolysis to free the sugar unit was carried out, followed by LC-HRESIMS analysis of 1-phenyl-3-methyl-5-pyrazolone (PMP) derivatives of ribose, xylose, and arabinose and of the hydrolyzed sugar residue of **1**. Interestingly, this analysis indicated that xylose was the most abundant sugar in **1** (~80%), but arabinose was also found to a smaller extent (~20%, from LC-HRESIMS (Figure S12). Hence, it is possible that some of the predicted minor conformer resonances observed in the NMR data for **1** could actually correspond to this arabinose-bearing, co-eluting minor compound. The presence of an α-methylated and β- and δ-hydroxylated acyl moiety could also be deduced from the NMR data. As shown in Figure 3A, this spin system was supported by HMBC and COSY correlations (Figures S8, S9) and degenerated into aliphatic methylenes, which could be connected by HMBC correlations to a terminal linear alkane moiety. Hence, this residue was found to correspond to 3,5-dihydroxy-2-methyldecanoic acid (Dhmda) residue, *O*-substituted at C-3 (Figure 3A). Due to poor resolution of the multiple amide carbon resonances, inter-residue connectivity could not be unambiguously established from the initial HMBC data acquired for **1**. To circumvent this, a band-selective constant-time HMBC[32] experiment was conducted (Figure S11). The increased resolution in the amide δ_C resonance region enabled the determination of the connectivity between the NMR-derived residues for **1**, with the exception of the (Pro-1)-Val connectivity (Figure 3A). Nevertheless, a cyclic peptide would be required to satisfy the degrees of unsaturation, and the connectivity between Pro and Val was supported by NOE data (Figure 3A; Figure S10). From HRESIMS/MS analysis of **1**, the residue sequences [Dhmda-Gln-Gly-Pro-Val-Tyr] and [Tyr-Ser-Leu-Thr-Leu-Pro-Dhmda-Gln] could be established (Figure 3A, Figures S13-14), which validated the NMR-derived findings.

Table 1 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for desmamide A (**1**) in DMSO-d_6.

unit	no.	δC, type	δH[a] (*J* in Hz)	HMBC[b]	COSY	NOESY
Dhmda	1	172.5, C				
	2	43.7, CH	2.55, t (6.7)	1, 3, 4, 11	3, 11	4, Gln-NHa
	3	72.9, CH	5.06, m	1, 2, 4, 5, 11, 61	2, 4	11, Gln-NHa
	4	38.4, CH_2	1.46, m	3, 5	3, 5	2, Gln-NHa
	5	66.3, CH	3.33, m	7	4, 6, Dhmda-OH	
	6	37.7, CH_2	1.28, m	7	5	
	7	24.8, CH_2	1.28, m	6, 8		
	8	24.8, CH_2	1.28, m	6, 8, 9		
	9	22.1, CH_2	1.25, m	6, 8, 10	10	
	10	14.0, CH_3	0.86, m	9	9	
	11	13.6, CH_3	0.98, d (7.0)	1, 2, 3	2	3, Gln-NHa
	OH		4.30, d (5.7)	4, 5, 6	5	
Glu	12	53.1, CH	4.15, m	1, 13, 14, 16	13a, 13b, Gln-NHa	Gly-NH
	13a	27.7, CH_2	1.86, m	12, 14, 15	12	Gln-NHa
	13b		1.74, m	12, 14, 15	12, 14a	Gln-NHa
	14a	31.7, CH_2	2.12, m	12, 13, 15	13b	Gln-NHa
	14b		1.88, m	12, 15		Gln-NHa
	15	173.8, C				
	16	171.7, C				
	NHa		7.93, m	1, 12	12	2, 3, 4, 11, 13a, 13b, 14a, 14b
	NHb		6.72, d (11.5)	14, 15	Gln-NHb	
	NHb		7.20, d (18.8)	15	Gln-NHb	
Gly	17[a]	41.3, CH_2	3.96, m	16, 18	17b, Gly-NH	
	17b		3.80, m	16, 18	17a	Gly-NH
	18	167.1, C				
	NH		7.76, t (5.3)	16	17a	12, 17b
Pro_1	19	59.2, CH	4.38, m	21, 23	20	Val-NH
	20	28.9, CH_2	2.0, m	21, 22, 23	19	
	21	24.4, CH_2	1.97, m		22a	22b
	22a	46.6, CH_2	3.51, m		21, 22b	
	22b		3.75, m		22a	21
	23	172.4, C				
Val	24	59.8, CH	3.76, m	25, 27, 28	25, Val-NH	26, Tyr-NH
	25	29.5, CH	1.72, m	24, 26, 27, 28	24, 26, 27	Val-NH
	26	18.6, CH_3	0.46, d (6.6)	24, 25, 27	25	24, Val-NH
	27	18.7, CH_3	0.66, d (6.6)	24, 25, 26	25	Val-NH
	28	170.4, C				
	NH		8.11, d (5.8)		24	19, 25, 26, 27
Tyr	29	54.1, CH	4.53, m	28, 30, 37	30a, 30b, Tyr-NH	32/36, Tyr-NH, Ser-NH
	30a	36.6, CH_2	2.64, m	32/36, 29	29, 30b	32/36, Tyr-NH
	30b		3.13, m	32/36	29, 30a	32/36, Tyr-NH
	31	131.2, C				
	32	129.9, CH	7.13, d (8.5)	30, 33, 34	33/35	29, 30a, 30b
	33	116.5, CH	6.91, d (8.5)	31, 34	32/36	1'

	34	155.4, C				
	35	116.5, CH	6.91, d (8.5)	30, 33, 34	32/36	1'
	36	129.9, CH	7.13, d (8.5)	31, 34	33/35	29, 30a, 30b
	37	171.5, C				
	NH		8.21, d (8.4)	28	29	24, 29, 30a, 30b
	38	56.5, CH	4.16, m	37, 39, 40	Ser-NH	39a
	39a	61.5, CH_2	3.65, m	40	Ser-OH	38, Ser-NH
Ser	39b		3.75, m	40	Ser-OH	
	40	169.8, C				
			8.08, d (6.8)	37, 38, 39	38	29, 39a, Ser-OH, Leu_1-NH
			5.13, t (6.1)	39	39a, 39b	Ser-NH
	41	48.8, CH	4.32, m		Leu_1-NH	42, Leu_1-NH
	42	40.3, CH_2	1.42-1.50, m	41, 43, 44/45		41, Leu_1-NH
	43	24.0, CH	1.64, m	42, 44/45	44/45	
Leu_1	44	23.2, CH_3	0.87, m	43	43	
	45	23.2, CH_3	0.87, m	43	43	
	46	172.1, C				
			7.55, d (8.0)	40	41	41, 42, Ser-NH, Thr-NH
	47	58.3, CH	4.19, m	46, 48, 49, 50	Thr-NH	48, 49
	48	66.8, CH	3.91, m	50	49, Thr-OH	47, Thr-NH, Leu_2-NH
Thr	49	19.4, CH_3	1.0, d (6.4)	47, 48	48	47, Thr-OH
	50	169.5, C				
	NH		7.87, d (7.5)	46, 47, 48	47	48, Leu_1-NH, Thr-OH
	OH		4.90, d (4.6)	47, 48, 49	48	49, Thr-NH
	51	48.8, CH	4.65, m	50, 56	52, Leu_2-NH	
	52	41.1, CH_2	1.48, m	51, 53, 54/55	51	Leu_2-NH
	53	24.0, CH	1.64, m	52, 54/55	54/55	
Leu_2	54	21.4, CH_3	0.83, m	52, 53	53	
	55	21.4, CH_3	0.83, m	52, 53	53	
	56	170.6, C				
	NH		7.57, d (8.0)	50	51	48, 52
	57	59.0, CH	4.27, m	58, 59, 61	58	59, 3'
	58	28.5, CH_2	2.13, m		57, 59	
Pro_2	59	24.7, CH_2	1.85, m	58, 60	58, 60	57
	60	45.9, CH_2	3.46, m		59	
	61	171.4, C				
	1'	97.6, CH	5.31, d (3.5)	34, 63, 64, 66	2'	33/35
	2'	71.6, CH	3.36, m	64, 66	1', OHa	
	3'	73.2, CH	3.55, m	63, 65	OHb	57
	4'	69.7, CH	3.34, m	64, 66		
Xylosyl	5'a	62.5, CH_2	3.45, m	62, 64, 65	5'b	
	5'b		3.33, m	62	5'a	OHb, OHc
	OHa		5.01, d (6.2)	-	2'	
	OHb		4.93, d (4.9)	63, 64, 65	3'	5'b
	OHc		4.99, d (4.8)	65, 66		5'b

[a]from HSQC, [b]from proton to indicated carbon.

The 1D and 2D NMR data of **2** in DMSO-d_6 (Table S3, Figures S16-S20) was found to be highly similar to those of **1**. In HRESIMS analysis, **2** showed a *m/z* 1430.7666 corresponding to a molecular formula $C_{68}H_{107}N_{11}O_{22}$ and 21 degrees of unsaturation (Figure S15). The 42.0089 amu mass difference to **1** is consistent with the presence of an acetyl substitution (C_2H_2O, calcd 42.0106 amu). Comparison of the ^{1}H and ^{13}C NMR data of **1** and **2** revealed the differences to be associated with the sugar resonances, including those of C-3'/4'/5' (Table S3). Also, a prominent sharp singlet (δ_H 2.02, s, 3H) and an additional carbonyl resonance (δ_C 170.0) were observed in **2**. Analysis of 2D NMR data of this compound (Table S3) clarified that the latter signals corresponded to the suspected acetyl group, in particular through the ^{2}J HMBC correlation between the acetyl methyl to the carbonyl, represented in Figure 3B. An HMBC correlation between H-4' and this same carbonyl confirmed the acetylation of the OH group at C-4' position of the sugar moiety (Figure 3B). The acetylated sugar residue is proposed to be α-xylopyranosyl, as in **1**, on the basis of the small $J_{1'\text{-}2'}$ value of 3.5 Hz (H-1', δ_H 5.40), and from coupling constant analysis of ^{1}H and COSY data (Figure S21).

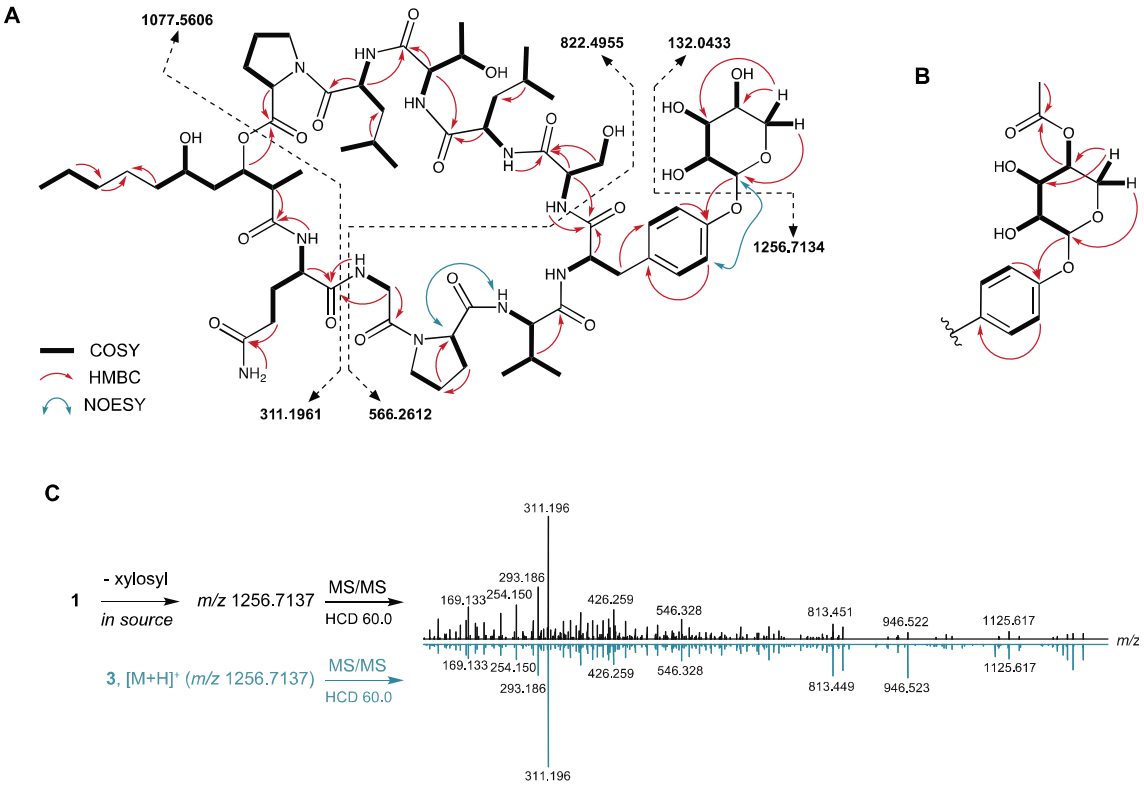


Figure 3 - Structure elucidation of **1**-**3**. A – Key HMBC, COSY and NOESY correlations and HRMS/MS fragmentations supporting the structure proposal for **1**. B – Key HMBC and COSY correlations supporting the presence of a 4-*O*-acetylxylopyranosyl-Tyr moiety. C – Comparison of the HRESIMS/MS spectra of the aglycone of **1** (isolated in-source) and of compound **3**, supporting that they are identical.

Compound **3** showed ^{1}H and ^{13}C NMR data in DMSO-d_6 (Table S4; Figures S22-S23) similar to those of **1** and **2** but lacking the sugar signals, while featuring a broad singlet resonating at δ_H 9.13, characteristic of a phenol group exchangeable proton. As suggested by the NMR data, this metabolite was confirmed to be the aglycone of **1** and **2** on the basis of identical HRESIMS/MS spectra for **3** and for the in-source isolated aglycone ions of **1** and **2**, shown in Figure 3C (Figure S24).

The absolute configurations of the chiral amino acid residues in **1** (and therefore, also in **2** and **3**) was determined using chiral HPLC and Marfey's analysis. The Tyr residue and the two Leu residues were found to have D-configuration, while the remaining chiral amino acids were of L-configuration (Figures S25, S26). Determination of the absolute configuration of the α-xylopyranosyl residue was not possible, thus only the relative configuration is reported. Nevertheless, given the fact that L-xylose is very rare in nature[33], we propose a D-configuration for this sugar residue.

Compounds **1** and **2** are cyclic decapeptides that feature Tyr *O*-glycosylation and contain a modified fatty acyl residue as part of the macrocycle. This residue is connected through an amide bond to the N-terminal amino acid and through an ester bond to the C-terminus of the peptide chain. Desmamides A and B thus represent a novel lipoglycopeptide class, the second to be described in cyanobacteria after the hassallidins/balticidins.[19,20] Unlike **1** and **2**, these compounds are nonapeptides with an exocyclic fatty acyl residue connected to the N-terminal amino acid, which is also exocyclic. Several *O*-glycosylation patterns have been described for hassallidins/balticidins, occurring on the 3-hydroxy group of the fatty acyl moiety and/or on an endocyclic N-Me-Thr residue. Compounds **1** and **2** are also structurally distinct from the other previously reported natural lipoglycopeptide scaffolds (Figure S27). Nevertheless and remarkably, *O*-glycosylation of Tyr is observed in several of these, namely in mannopeptimycins γ/δ/ε[22] and in the gausemycins.[18] This decoration is also present in the fungal glycopeptide cycloaspeptide F[34] and in cacaoidin[35], a recently described ribosomally synthesized and post-translationally modified (RiPP) glycopeptide. In the case of ramoplanin, *O*-glycosylation occurs in the non-proteinogenic amino acid Hpg (4-hydroxyphenylglycine)[23], which is structurally similar to Tyr. Protection of the 4-OH position in sugar residues, as observed for **2**, is also a feature of mannopeptimycins γ/δ/ε,[22] while several hassallidins contain acetylated sugar moieties.[19] No other examples of lipopeptides bearing a Dhmda moiety were found, nor shorter- or longer-chain versions thereof.

During the isolation efforts, several minor analogues of desmamides were detected in adjacent fractions, as in the case of hassallidins.[19] The mass features found showed minor structural differences, mainly on the fatty acyl chain, as shown in table 2 (Figures S28-S35). This is a common feature of cyanobacterial peptidic compounds and might be involved in synergistic interactions in the producer organism.[36]

Table 2 – Desmamide analogues detected during isolation.

Analogue *(m/z)*	Intensity	Calc. M.F.	Mass difference	M.F.	Structural diff.	Main Fragments
1344.7297	4.29E+08	$C_{64}H_{103}O_{20}N_{11}$	-44.03	$2(CH_2)O$	FA (C6)	1212-L-T-L-S-Y-V-P-G
1212.6871	1.72E+07	$C_{59}H_{94}O_{16}N_{11}$	aglycone			382-L-T-L-S-Y
1402.7742	2.22E+08	$C_{67}H_{108}N_{11}O_{21}$	14.02			1270-H_2O-L-T-L-S-Y
1444.7816	4.87E+07	$C_{69}H_{110}N_{11}O_{22}$	acetylated sugar	CH_2	Ile/Leu instead of Val	311-G-P-L/I-T
1270.7293	2.93E+08	$C_{62}H_{99}N_{11}O_{17}$	aglycone			311-P-L-T-L
1360.7234	4.69E+06	$C_{64}H_{102}N_{11}O_{21}$	-28.03	$2(CH_2)$	3,5-hydroxy-2-methylbutyric acid	1228-H_2O-L-T-L-S-Y
1228.6789	1.48E+06	$C_{59}H_{94}N_{11}O_{17}$	aglycone			283-P-L-V-Y
1416.7839	6.76E+07	$C_{68}H_{110}N_{11}O_{21}$	28.03	$2(CH_2)$	3,5-hydroxy-2-methyloctanoic acid	1284-H_2O-L-T-L-S-Y-V-P-G
1284.74702	-	$C_{63}H_{102}N_{11}O_{17}$	aglycone			339-G-P-V-T-S
1374.7392	4.15E+07	$C_{65}H_{104}N_{11}O_{21}$	-14.02	CH_2	3,5-hydroxy-2-methylhepatnoic acid	1242-H_2O-L-T-L-S-Y
1242.6971	2.23E+07	$C_{60}H_{96}N_{11}O_{17}$	aglycone			297-G-P-V-Y
1370.7478	1.91E+07	$C_{66}H_{104}N_{11}O_{20}$	-18.01	$-H_2O$	3-hydroxy-2-methyldecanoic acid	293-G-V-Y-S (-)xylosyl-L-T-L
1042.5788	9.20E+07	$C_{51}H_{80}N_9O_{14}$	-214.14	$C_{10}H_{18}N_2O_3$	without Leu-Thr	1042-H_2O-L-S-Y-P-V-G-P 1042-L-S-Y 311-G-P-V; 311-P-L-S
1198.6748	1.99E+06	$C_{58}H_{92}O_{16}N_{11}$	-58.05	$3(CH_2)O$	3-hydroxy-2,5-methylhexanoic acid	1198-H_2O-L-T-L 368-G-P-V-Y-S-L-T-L

3.2.3. Identification and analysis of the *dsm* biosynthetic gene cluster

The structures of **1-3** could, in principle, be reconciled with canonical, hybrid PKS and NRPS biosynthetic logic. This working hypothesis was used to search the genome data of *D. muscorum* LEGE 12446 (Genbank: JADEXS020000000). Annotation of the genome data using antiSMASH v6.0[37] revealed several hybrid PKS-NRPS BGCs. One such BGC, represented in Figure 4A, featured two PKS modules upstream of 10 NRPS modules, for which the corresponding adenylation (A) domains are predicted select and activate the amino acids that make up the desmamides. Notably, these are colinear with the N-to-C terminal amino acid sequence of the peptide and contained epimerase (E) domains in the modules predicted to load the experimentally determined D-amino acids. This BGC was further considered as the putative desmamide gene cluster (*dsm*) and an detailed annotation was carried out (Table S5) leading to a biosynthetic proposal, shown

in Figure 4B. According to this proposal, the biosynthesis of **1**-**3** begins with activation of hexanoic acid by an adenylation domain in DsmG, which is related to fatty acyl-AMP ligases (Table S5). This intermediate would then be elongated using malonyl-CoA by the PKS module in DsmG, which includes a ketoreductase domain (KR), to generate a linear eight-carbon beta-hydroxy acyl-ACP intermediate. An additional malonyl-CoA unit would then be used by the PKS module in DsmH to elongate this intermediate to a 10-carbon unit, reduce its beta-keto group (KR domain) and methylate (methyltransferase domain, MT) its alpha position to generate the 11-carbon Dhmda residue. In both PKS modules, the KR domain do not present residue signatures that enable inferring the configuration of C-2, C-3, and C-5, as evidenced by antiSMASH v6.0 analysis, which did not result in a classification of the KR type for either domain. To test the proposed biosynthesis of Dhmda experimentally, in particular whether hexanoic acid (C_6) is the starter unit for the biosynthesis of **1**-**3**, *D. muscorum* LEGE 12446 cultures were supplemented with perdeuterated fatty acids, namely hexanoic (d_{11}-C_6), octanoic (d_{15}-C_8) and decanoic (d_{19}-C_{10}) acids. LC-HRESIMS analysis showed that only d_{11}-C_6 was incorporated into **1**-**3**, present in Figure 4C, indicating that hexanoic acid is a substrate in the biosynthesis of the desmamides, as hypothesized. However, no ACP domain was found in between the adenylation domain proposed to load hexanoic acid and the first PKS module, which is unusual. A KR domain downstream of the A domain in DsmG is of the C-1 type, and therefore inactive.[38] The PKS module of DsmG includes a DH domain, containing all catalytical residues (as per antiSMASH analysis), which does not seem necessary to generate the Dhmda residue of **1**-**3**. Following the proposed biosynthesis pathway and canonical hybrid PKS/NRPS biosynthesis,[39] the resulting C_{11} acyl intermediate is passed to the first NRPS module in DsmI which loads Gln, followed by incorporation of the additional amino acids into the growing chain by the NRPS machinery encoded in DsmI and DsmJ. Finally, the thioesterase in the C-terminus of DsmJ would catalyze the concomitant cyclization and chain release of the desmamides (Figure 4A, B).

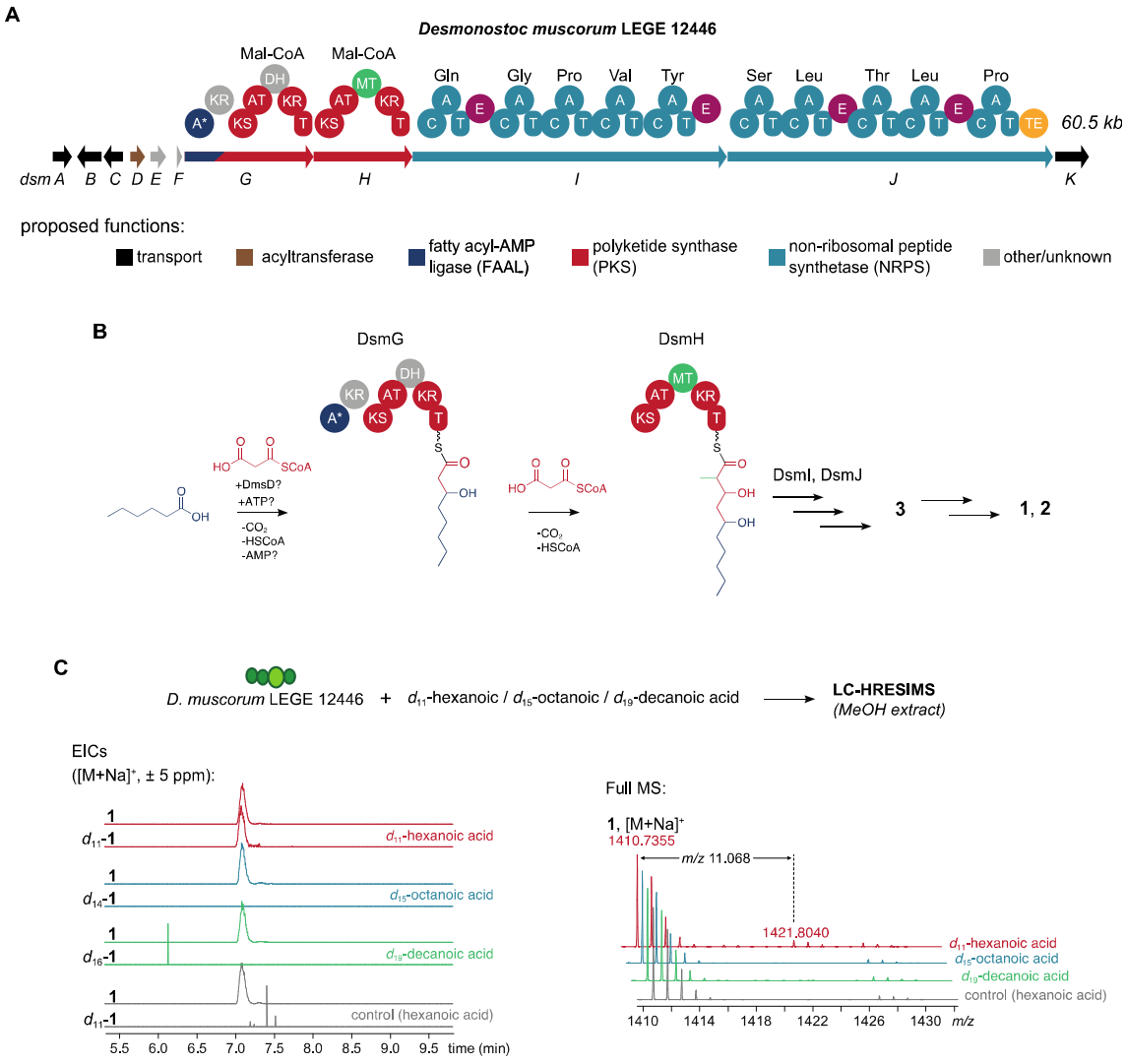


Figure 4 - Biosynthesis of the desmamides A-C (**1**-**3**). A – Schematic representation of the *dsm* BGC, including the proposed functions of the predicted open-reading frames, domain composition of detected PKS and NRPS (or NRPS-like) modules, and bioinformatically-predicted (antiSMASH) monomers activated by acyl-tranferase (AT) or adenylation (A) domains. A* indicates an adenylation domain within a region with homology to a fatty acyl-AMP ligase (FAAL), proposed to activate hexanoic acid. B – Proposed biosynthetic steps leading to **1**-**3**, highlighting the formation of the Dhmda residue from hexanoic acid, by PKS machinery. C – LC-HRESIMS analysis of methanolic extracts of *D. muscorum* LEGE 12446 cultures supplemented with perdeuterated fatty acids. Shown are EICs for *m/z* values corresponding to $[M+H]^+$ ions for compound **1** and for hypothesized isotopologues resulting from incorporation of each of the perdeuterated fatty acids. Also shown is Full MS data, highlighting the mass shift supporting incorporation of d_{11}-C_6 into **1**.

Intriguingly, despite **1** and **2** being glycosylated, no glycosyltransferase-encoding gene in the *dsm* locus, or its genomic neighborhood, was found. A genome-wide search for glycosyltransferases (GTs), in particular those known to glycosylate tyrosine residues in the biosynthesis of natural products, indicated multiple GTs, some of which were low-identity homologs of the cacaoidin BGC GTs (Table S6); however, none was near the *dsm* locus. Homologs of the mannopeptimycin glycosyltransferases[40] were not detected

in the searches. It is also unclear if any of the *dsm* BGC-encoded enzymes are responsible for the 4-*O*-acetylation of the sugar residue. One possibility would be the predicted acyltransferase DsmD. Enzymes containing a GNAT-like fold such as DsmD (which contains two GNAT-like domains) are known to *N*-acylate several substrates[41] but have also been reported to catalyze *O*-acetylation of sugar residues.[42,43] Interestingly, hassallidins also feature *O*-acetylation in sugar residues decorating the peptide. In this case, the authors suggested that HasR, an acyltransferase encoded in the hassallidin (*has*) BGC could be responsible for these decorations.[19] No homologs of HasR were detected in the genome of *D. muscorum* LEGE 12446.

Hassallidins are widely distributed among the Nostocales.[19] To check if that was the case for desmamides, and whether *dsm*-like BGCs could be found in genomes of cultured cyanobacteria or in metagenome data, ClusterBlast was used. This antiSMASH could not retrieve any similar BGCs and BlastP searches (nr/env_nr databases) using DsmG/DsmH/DsmI/DsmJ amino acid sequences as query did not retrieve hits with a genomic context that would configure a full *dsm* BGC. Curiously, several Dsm proteins have a very high identity to homologs from *Nostoc* sp. FACHB-973 and *Desmonostoc muscorum* CCALA 125 (Table S5). Because such homologs are present in more than one contig in each of the genome assemblies, it is unclear if these are part of *dsm* BGCs.

3.2.4. Biological activity

Compounds **1** and **2** were tested against a panel of human cancer cell lines, namely liver cancer cell line HepG2, colorectal adenocarcinoma cell line HT-29, malignant melanoma cell line A-375, colon cancer cell line HCT-116, lung carcinoma cell line A-549, breast cancer line MDA-MB-231 as well as the non-cancerous keratinocyte line HaCaT. Neither compound demonstrated cytotoxicity in the tested cancer cell lines. Both **1** and **2** were tested against a wide variety of Gram positive and Gram negative bacterial and fungal pathogens. They showed no inhibitory activity at the standard concentration of 100 μg mL^{-1} except a moderate activity against the plant pathogenic bacterium *Xanthomonas campestris* with IC_{50} values of 48 and 334 μM, respectively.

3.3. Conclusion

In this study focused on exploring a cycad coralloid root symbiotic cyanobacterium for its novel secondary metabolites. This resulted in isolation and structure elucidation of

the new lipoglycopeptides **1** and **2** and their aglycone **3**. The proposed *dsm* BGC is possibly involved in the biosynthesis of these compounds. Desmamides feature an unprecedented type of fatty acyl moiety among lipopeptides, which is derived from a hexanoic acid starter unit. While the enzyme functions predicted to be encoded in the *dsm* BGC are congruent with the formation of **3** and, therefore, the aglycones of **1** and **2**, it was not possible to rationalize the glycosylating enzyme(s) through bioinformatics predictions. The desmamides expand the diversity of natural lipoglycopeptides, a small group of natural products with remarkable antibacterial and antifungal properties. The biological activity of the desmamides was explored, and found that these were inactive toward several different human cancer cell lines. However, compound **2** showed antibiotic activity towards the black rot disease-causing plant pathogen *Xanthomonas campestris*. Lipopeptides and lipoglycopeptides often have important antibacterial activities. Given their abundance in *D. muscorum* LEGE 12446, high concentrations of desmamides are likely to occur in a natural setting and could potentially be related to a defensive ecological role in its host plant. This will necessarily involve studying the potential production of **1**-**3** inside the coralloid roots. For example, two *Nostoc* sp. coralloid root endosymbionts, isolated from *Macrozamia* spp. cycads, were found to produce the toxin nodularin *in planta*.

Overall, the discovery of the desmamides and of the *dsm* BGC sets the foundation exploring the role of these very abundant (1.6% d.w.) metabolites, in particular in the context of cycad coralloid root symbiosis. Despite all the recent progress in understanding the secondary metabolism of the facultative endosymbiont *Nostoc punctiforme* PCC 73102, these systems are still poorly understood in terms of chemical ecology.

3.4. Materials and methods

3.4.1. General experimental procedures

Optical rotations were measured using a JASCO P-2000 polarimeter (JASCO) with SpectraManager 2.14.02 software. Infrared spectra were collected on a Nicolet iS5 FTIR spectrometer (ThermoScientific) with OMNIC 9.8.372 software. The UV spectra were acquired on a UV-1600PC spectrometer (VWR) controlled by MWAVE 1.0.20 software. 1D and 2D NMR data were obtained in the Materials Center of the University of Porto (CEMUP) on a Bruker Avance III, 400 MHz controlled by TopSpin 3.2. Compounds were analyzed in deuterated DMSO (DMSO-d_6, Sigma). Chemical shifts (^{1}H and ^{13}C) are

expressed in δ (ppm), referenced to the residual non-deuterated solvent used (^{1}H, 2.500 ppm; ^{13}C, 39.52 ppm). NMR data were analyzed in MNova 12.0.4. LC-HRESIMS and LC-HRESIMS/MS analyses were performed on an UltiMate 3000 UHPLC (Thermo Fisher Scientific) system composed of a LPG-3400SD pump, WPS-3000SL autosampler and VWD-3100 UV/VIS detector coupled to a Q Exactive Focus Hybrid Quadrupole-Orbitrap Mass Spectrometer controlled by Q Exactive Focus Tune 2.9 and Xcalibur 4.1 (Thermo Fisher Scientific). For LC-HRESIMS data, full Scan mode was used with capillary voltage set to -3.8kV, capillary temperature to 300°C and sheath gas flow rate to 35 units. HPLC separations were performed in a JASCO PU-4180 HPLC pump with MD-4010 Photo Diode Array Detector, and separations were monitored at 254 nm wavelength. All solvents used were ACS grade, except for HPLC solvents (HPLC gradient grade) and LC-MS solvents (MS-grade). Stable-isotope labeled fatty acids (d_{11}-hexanoic, d_{15}-octanoic, d_{19}-decanoic and d_{23}-dodecanoic were obtained from CDN Isotopes Inc.; amino acids were purchased from Fluorochem (D-allo-Thr; L-allo-Thr; D-Ser; L-Ser; L-Val; D-Val; L-Pro; D-Pro; D-Tyr; L-Tyr; D-Leu; L-Leu; D-Thr; L-Glu; D-Glu), and Glenthan Life Sciences (L-Thr). Monosaccharide standards were acquired from Sigma Aldrich (D-xylose; L-arabinose), Alfa Aesar (D-ribose; D-arabinose) and Acros Organics (L-ribose). The reagent PMP was obtained from Alfa Aesar, and *Nα*-(2,4-Dinitro-5-fluorophenyl)-L-alaninamide (Marfey's reagent) were purchased from Sigma Aldrich.

3.4.2. Cyanobacterial strain cultivation

The cyanobacterial strain *D. muscorum* LEGE 12446, kindly provided by LEGEcc,[44] was cultured in Z8 media[45], at 25 °C, with constant aeration, under a 14h/10h light (10−30 μmol photons s^{-1} m^{-2})/dark cycle. Large-scale culturing of this strain was carried out in 80 L low-density polyethylene sleeves for a total of 160 L and cells were harvested during late exponential/early stationary phase (1.5-2 months of growth) through centrifugation (7178 x *g*), immediately frozen, and freeze-dried until further usage.

3.4.3. Molecular Networking analysis

The cyanobacterial strain *D. muscorum* LEGE 12446 was analyzed using GNPS Molecular Networking analysis[26], as reported in chapter II. Briefly, 4 L of culture were harvested at the exponential phase of growth, frozen and freeze-dried. A crude extraction of the dry cell material using a mixture of CH_2Cl_2/MeOH (2:1, v/v) was performed at room

temperature and the resulting extract was fractionated in a Strata SI-1 Silica cartridge (5 g, 55 *μ*m, 70A, Phenomenex) using a gradient of increasing polarity from *n*-hexane (*n*-hex) to EtOAc to MeOH, to yield three fractions (A-C). For LC-HRESIMS analysis, fractions were prepared at 1.0 mg mL^{-1} in LC-MS grade MeOH, filtered through a 0.2 *μ*m syringe regenerated cellulose filter, and 5 *μ*L were injected in an ACE Ultracore 2.5 SuperC18 column (75 × 2.1 mm). Samples were eluted at 0.35 mL min^{-1} over a linear gradient from 99.5% solution A (95% H_2O, 5% MeOH, 0.1% v/v HCOOH) and 0.5% solution B (95% isopropanol, 5% MeOH, 0.1% v/v HCOOH) to reach 10% solution B over 0.5 min, followed by an increase to 60% solution B in 8 min and then to 90% in 1 min; these conditions were kept for 6 min before returning to the initial conditions. The column oven was set to 40 °C, UV monitoring was carried out at 254.0 nm. The capillary voltage of the Heated Electrospray Ionization (HESI) was set to 3.8 kV and its temperature to 300 ºC. The sheath gas and auxiliary gas flow rate were at 35 and 10 (arbitrary units as provided by the software settings). Full MS scans were acquired at a resolution of 70,000 fwhm (range of 150–2000 *m/z*), and data dependent MS/MS ($ddMS^2$, Discovery mode) at a resolution of 17,500 fwhm (isolation window used was 3.0 amu, and normalized collision energy was 35). For the molecular network analysis, the protocol available in GNPS documentation was followed and the resulting network file was visualized using Cytoscape v3.8.0[46].

3.4.4. Extraction and isolation of desmamides

Freeze-dried biomass from large-scale culturing (70.3 g, d.w.) was extracted by repeated percolation using a mixture of CH_2Cl_2/MeOH (2:1, v/v) at room temperature and at 40 ºC. A total extract mass of 14.7 g was obtained. Further fractionation using vacuum liquid chromatography (VLC) was performed, using silica gel 60 (0.015-0.040 mm, Merck). The extract was dry loaded onto the column using 15 g of silica. A mixture of 1:1 (v/v) EtOAc/*n*-hex was used to start the separation, and stepwise gradient to 100% EtOAc and then to 100% MeOH was used, yielding 13 fractions. Fractions containing the mass features of interest (eluting with 2:3 (v/v) to 3:7 (v/v) EtOAc/MeOH) were further fractionated using flash chromatography (silica gel 60, 0.040-0.063 mm, Merck), with a stepwise gradient from 1:1 (v/v) EtOAc/MeOH to 100% MeOH, yielding seven fractions. The mass features of interest were present in fractions 2-4, which were pooled and further separated by reversed-phase solid phase extraction (SPE) in a Strata C18-E cartridge (50 g, 55 *μ*m, 70A, Phenomenex) using a stepwise gradient from 5% MeOH (aq) (solution A) to 95% IPA with 5% MeOH (solution B), yielding a total of 11 fractions.

The first fraction (A, 565 mg), eluting with 5% MeOH (aq), was selected for reversed phase semi-preparative HPLC separation using a C18 column (100 Å, 250 × 10 mm, 5 *μ*m, ACE). A gradient elution program was employed, starting with 35% MeCN (aq), increasing to 60% MeCN (aq) over 20 min, and then to 100% MeCN in 5 min, before returning to the initial conditions. Fraction A2 (33.5 mg, t_R 8-9 min) was collected and subjected to an additional purification step in the same system but with a gradient starting in 35% MeCN (aq) and increasing to 46% MeCN (aq) in 15 min, and then to 100% MeCN in 5 min before returning to the initial conditions. This purified fraction A2 corresponded to spectroscopically pure desmamide A (compound **1**, t_R = 10.0 min, 26.3 mg). HPLC fraction A5 (5.8 mg, t_R =11-13 min) corresponded to partially purified desmamide C and was further separated by semi-preparative reversed phase HPLC with an Aeris Peptide XB-C18 column (100 Å, 250 × 10 mm, 5 *μ*m, Phenomenex). A gradient from 35%MeCN (aq) to 45% MeCN (aq) in 22 min was applied to yield spectroscopically pure desmamide C (**3**, t_R =15.5 min, 4.0 mg). Desmamide B (**2**) was purified from SPE fraction E (977 mg), eluting with 3:2 (v/v) of a mixture of solution A (5% MeOH (aq)) and solution B (IPA with 5% MeOH), following semi-preparative HPLC using an Aeris Peptide XB-C18 column (100 Å, 250 × 10 mm, 5 *μ*m, Phenomenex). The separation was carried out with isocratic 35% MeCN (aq) for 16 min, and then a linear gradient to 50% MeCN (aq) in 3 min, before returning to the initial conditions. Only 205 mg of sample (fraction E) were processed. Fraction E6 corresponded to spectroscopically pure desmamide B (**2**, t_R =18.0 min, 20.4 mg). The physical chemistry analysis optical rotation, infrared (IF) and UV absorption were performed for all three compounds and are reported as follows.

Desmamide A (**1**). Amorphous white powder; $[\alpha]_D^{23}$ +41 (*c* 0.01, MeOH); IR ν_{max} 3271.3, 2954.5, 2929.6, 2929.6, 2871.3, 2350.9, 2317.8, 1738.7, 1733.1, 1667.3, 1660.5, 1651.6, 1645.3, 1634.5, 1574.4, 1557.3, 1548.6, 1539.0, 1532.4, 1511.6, 1455.34, 1417.9 cm^{-1} (Figure S38); UV (MeOH) λ_{max} (log ε) 219.0 (3.2), 272.0 (2.9), 278.5 (2.8) (Figure S39); ^1H and ^{13}C NMR spectroscopic data (DMSO-d_6) Table 1; HRESIMS *m/z* 1388.7577 $[M+H]^+$ (calcd. for $C_{66}H_{106}N_{11}O_{21}$ 1388.7559); HRESIMS/MS (Figure S13).

Desmamide B (**2**). Amorphous white powder; $[\alpha]_D^{23}$ +37 (*c* 0.01, MeOH); IR ν_{max} 3271.2, 2953.9, 2928.9, 2870.3, 2351.0, 2327.9, 1737.9, 1732.6, 1667.0, 1660.1, 1651.4, 1645.0, 1634.3, 1574.1, 1556.9, 1548.2, 1538.5, 1532.2, 1511.5, 1462.8, 1417.6 cm^{-1} (Figure S40); UV (MeOH) λ_{max} (log ε) 222.0 (3.2), 272.0 (2.9), 278.5 (2.8) (Figure S41); ^1H and ^{13}C NMR spectroscopic data (DMSO-d_6) Table S3; HRESIMS *m/z* 1430.7666 $[M+H]^+$ (calcd. for $C_{68}H_{108}N_{11}O_{22}$ 1430.7665); HRESIMS/MS (Figure S34).

Desmamide C (**3**). Amorphous white powder; $[\alpha]_D^{23}$ +10 (*c* 0.01, MeOH); IR ν_{max} 3270.6, 2955.9, 2929.2, 2871.2, 2350.8, 1660.2, 1651.3, 1538.6, 1515.9, 1447.3 cm^{-1} (Figure S42); UV (MeOH) λ_{max} (log ε) 231.0 (3.2), 277.5 (3.1) (Figure S43); ^{1}H and ^{13}C NMR spectroscopic data (DMSO-d_6) Figure S22-S23; HRESIMS *m/z* 1256.7149 $[M+H]^+$ (calcd. for $C_{61}H_{98}N_{11}O_{17}$ 1256.7142); HRESIMS/MS (Figure S34).

3.4.5. HRESIMS/MS analysis

HRESIMS/MS analysis of pure **1**-**3** was performed by direct injection of each compound (1.0 mg mL^{-1}, flow 0.005 mL min^{-1}) into the spectrometer, with a 35 000 fwhm resolution, using an isolation window of 1 *m/z*, loop count of 3, and AGC target of 5 × 10^4, and a collision energy of 60 (arbitrary units). Analysis of the minor analogues was performed by LC-HRESIMS/MS following the same conditions described in the molecular networking section (3.4.3.), but with $ddMS^2$ in confirmation mode (with each of the mass values) and with normalized collision energy between 50-75. In the case of glycosylated metabolites, additional *in-source* fragmentation followed by MS^2 in confirmation mode was performed for some metabolites (indicated in the Figures S28-S35).

3.4.6. Quantification of desmamides in *D. muscorum* LEGE 12446

To quantify the production of desmamides in the cyanobacterial *D. muscorum* LEGE 12446, a calibration curve for each compound was obtained from LC-HRESIMS analysis. Briefly, desmamide A-C standards were prepared at 0.5, 0.1, 0.05, 0.01, 0.005 and 0.001 mg mL^{-1}. Five-microliters (5 *μ*L) of each standard solution were injected in an LC-HRESIMS system fitted with an ACE Ultracore 2.5 SuperC18 column (75 × 2.1 mm) following the same chromatographic and spectrometric conditions described in the molecular networking analysis method. The area under the curve from the extracted ion chromatograms (EIC) for each compound were plotted (0.1 to 0.001 mg mL^{-1}), and a linear regression was fitted. This was then used to extrapolate the concentration of each compound in the initial crude extract used for isolation and in the biomass from which it was obtained (Table S2, Figure S3).

3.4.7. Chiral-phase HPLC

For the acid hydrolysis of **1**, 1.5 mg of compound was placed into 1.0 mL of 6 N HCl for 24h at 110°C under continuous stirring. After the reaction, HCl was removed under a N_2 stream. Amino acids standards (proline, valine, leucine, tyrosine), L- and D- isomers were prepared at 0.5 mg mL^{-1} with MS-grade H_2O. The hydrolysate of **1** was prepared 1.0 mg mL^{-1} with MS-grade H_2O. A volume of 10 μL of each sample was injected onto the HPLC system and separated in a Chirex 3126 (D)-penicillamine column (50 × 4.6 mm, Phenomenex), using a 1 mM $CuSO_4/H_2O$ solution as mobile phase, at a 1.0 mL min^{-1} flow rate. To confirm the absolute configuration of the amino acids, co-injection of the hydrolysate with both versions of each amino acid standard (15 μL+ 5 μL, respectively) was performed (Figure S5). Retention time (t_R, min) for standards: L-Pro 4.7, D-Pro 9.7, L-Val 6.4, D-Val 11.1, L-Leu 21.3, D-Leu 34.5, L-Tyr 24.0, and D-Tyr 43.8. Retention time (t_R, min) for hydrolysate: L-Pro 4.4, L-Val 5.9, D-Leu 32.5, and D-Tyr 42.3.

3.4.8. Marfey's derivatization analysis

The Marfey's derivatization method was used to assign the configuration of Ser, Gln, Thr, and *allo*-Thr residues. L- and D- isomer standards of each amino acid were prepared at 50 mM in 50 μL of LC-MS grade H_2O. For the Gln residue, standards where prepared using L-Glu and D-Glu. A Marfey's reagent (1-fluoro-2,4-dinitrophenyl-5-L-alanine amide) solution was prepared at a 1% (w/v) concentration in 100 μL acetone. For hydrolysate (0.6 mg, obtained as for the chiral HPLC analysis) derivatization, 100 μL of 1 M $NaHCO_3$ and 100 μL of the Marfey's reagent acetone solution were added to a glass vial, which was placed on an oil bath at 40 ºC for 1 h with a magnetic stirrer. The reaction was quenched with 100 μL of 1 M HCl, diluted with 200 μL MeCN (LC-MS grade) and filtered (0.2 µm). The same protocol was applied to the amino acid standards: to 50 μL of each standard were added 20 μL of 1 M NaHCO3 and 100 μL of the Marfey's reagent solution. The reaction (40 ºC, 1h) was quenched with 20 μL 1 M HCl and diluted with 810 µL MeCN. After derivatization all samples were analyzed by LC-HRESIMS, using an ACE UltraCore 2.5 SuperC18 column (75 × 2.1 mm, Avantor). Eluents were solutions A (95% H_2O, 5% MeOH, 0.1% v/v HCOOH) and B (95% IPA, 5% MeOH, 0.1% v/v HCOOH). The separation was carried out as follows: 100% A for 1 min, then a linear gradient for 7 min to 60% B and then a steeper gradient to 100% B in 1 min and held at 100% B for 2 min before returning to the initial conditions. Injection volume was 10 μL. Retention time (t_R, min) for standards: L-Ser 7.29, D-Ser 7.35, L-Leu 12.04, D-Leu 13.61, L-Thr 10.29, D-Thr

16.84, L-*allo*-Thr 10.50, D-allo-Thr 12.87, L-Glu 5.13, D-Glu 5.47. Retention time (t_R, min) for hydrolysate: L-Ser 7.29, D-Leu 13.62, L-Thr 10.23, D-Glu 5.54 (Figure S26).

3.4.9. Incubation experiments with perdeuterated fatty acids

To study the putative incorporation of a fatty acid unit in desmamide A, supplementation of *D. muscorum* LEGE 12446 with a series of deuterium-labelled fatty acids – d_{11}-hexanoic acid (d_{11}-C6), d_{15}-octanoic acid (d_{15}-C8), d_{19}-decanoic acid (d_{19}-C10) and d_{23}-dodecanoic acid (d_{23}-C12) – was carried out. Stock solutions of the labelled fatty acids were prepared in DMSO (500 mM). Small-scale cultures (25 mL) growing in Z8 medium were supplemented by pulse feeding (three times, two-day interval) with 8.33 μL of deuterated fatty acid stock solution, to a final concentration of 0.5 mM. A control experiment with non-deuterated hexanoic acid (C6) was performed. Experiments were carried out in duplicate. Cultures are placed in the same growth conditions described above with shaking in an orbital shaker at 200 rpm. After 7 days, cells were harvested by centrifugation, the pellet rinsed with ddH_2O and extracted with 15 mL of a mixture of CH_2Cl_2/MeOH (2:1, v/v) in an orbital shaker for 30 min. The extracts were then filtered, dried under vacuum, and resuspended in MeOH for LC-HRESIMS analysis (2 mg mL^{-1}). Samples were separated in an ACE UltraCore SuperC18 (2.5 µm, 95 Å, 75 × 2.1 mm) column. Elution was performed with a 0.4 mL min^{-1} flow using mixtures of eluents A (H_2O/MeOH 1:1, 0.1% v/v HCOOH) and B (IPA, 0.1% v/v HCOOH). The following program was applied: isocratic 10% B for 1 min, then a linear gradient to 65% B over 5 min, holding at 65% B for 12 min before increasing to 85% B over 2 min and holding again for 9 min at 85% B before returning to the initial conditions. The column oven was set to 40 °C, data was obtained in Full Scan (switching mode), with a capillary voltage of HESI +3.5 kV/-3.8 kV and a capillary temperature of 300 ºC.

3.4.10. Monossaccharide derivatization analysis with PMP

To determine the identity of the sugar residue in **1**, we used a sugar derivatization method using 1-phenyl-3-methyl-5-pyrazolone (PMP) to allow for LC-MS-based determination of monosaccharides.[47] Some adaptations to the published method were made. Briefly, monosaccharide standards (D-arabinose, D-xylose, and D-ribose) were prepared at 0.5 mg mL^{-1} with LC-MS grade H_2O. To 100 μL of each sugar solution, were added 100 μL of 0.3 M NaOH and 100 μL of PMP reagent (0.5 M in MeOH). Reaction mixtures were placed in an oil bath at 70 °C for 30 min. The mixtures were then quenched

with 100 μL 0.3 M HCl and dried in a rotatory evaporator. Equal volumes of H_2O and $CHCl_3$ (1 mL) were added to extract the PMP residues. The organic layer was discarded, and this process was repeated twice by adding $CHCl_3$ to the aqueous layer. Finally, the aqueous layer was dried and dissolved using 1 mL LC-MS grade H_2O. The same protocol was applied to 0.5 mg of the hydrolysate of **1**, which was dissolved in 0.5 mL of LC-MS grade H_2O. For monosaccharide derivatives analysis, an LC-HRESIMS run was performed, using an ACE UltraCore SuperC18 (2.5 μm, 95 Å, 75×2.1 mm) column and the elution was carried out isocratically in a mixture of (83:17, v/v) of H_2O with 0.1% v/v HCOOH and MeCN with 0.1% v/v HCOOH, respectively. The flow rate used was 0.4 mL min^{-1}, 5 μL of each derivative monosaccharide and hydrolysate solution were injected for analysis, and the separation was monitored using a wavelength of 245 nm. The column oven was set to 30 °C. Retention time (t_R, min) for standards: ribose 5.1, arabinose 10.7, xylose 11.5. Retention time (t_R, min) for hydrolysate: arabinose 10.7 (minor), xylose 11.4 (major) (Figure S12).

3.4.11. Genome sequencing and bioinformatics analysis

The cyanobacterium *D. muscorum* LEGE 12446 was sequenced through a specialized sequencing service (enhanced genome service) provided by the company MicrobesNG, using a combination of short-read Illumina and long-read Nanopore sequencing technologies. Fresh biomass, harvested after 15 days' cultivation period, were sent to the service provider. Following gDNA extraction and sequencing, raw data was submitted to a bioinformatics pipeline. Briefly, identification of the closest reference genomes for reading mapping was done using Kraken 2,[48] BWA-MEM was used to check the quality of the reads and *de novo* assembly was performed using SPAdes.[49] Because the culture was not axenic, the genomic data was treated as a metagenome: the retrieved contigs were analyzed using the binning tool MaxBin 2.0.[50] This procedure yielded a nearly complete genome (Genbank: JADEXS020000000) with an estimated size of 9.64 Mb, distributed among 12 contigs. AntiSMASH v6.0[37] was used to annotate BGCs in the genome data using relaxed detection settings and all extra features selected. Manual annotation and searches for enzyme homologs in the genome of *D. muscorum* LEGE 12446 were performed using BlastP or tBlastn.

3.4.12. Bioactivity profiling

The cell viability and antimicrobial assays were performed in collaboration with Dr. Deniz Tasdemir and Arlette Wenzel-Storjohann, from GEOMAR Centre for Marine Biotechnology, research unit marine natural product chemistry.

The sensitivity of malignant melanoma cell line A-375 (CLS), colon cancer cell line HCT-116 (DSMZ), lung carcinoma cell line A-549 (CLS), human breast cancer line MDA-MB-231 (CLS) and the non-cancerous human keratinocyte line HaCaT (CLS) to the samples was evaluated by monitoring the metabolic activity using the CellTiterBlue Cell Viability Assay (Promega). HaCaT cells were cultivated in RPMI medium, A549 and MDA-MB-231 cells in DMEM:Ham's F12 medium (1:1) supplemented with 15 mM HEPES and A-375, and HCT-116 cells in DMEM medium supplemented with 4.5 g L^{-1} D-Glucose and 110 mg L^{-1} Sodium pyruvate. All media were supplemented with L-glutamine, 10% fetal bovine serum, 100 U mL^{-1} penicillin, and 100 mg mL^{-1} streptomycin. The cultures were maintained at 37 °C under a humidified atmosphere and 5% CO_2. The cell lines were transferred every 3 or 4 days. For experimental procedure, cells were seeded in 96-well plates at concentration of 10 000 cells per well in RPMI. After 24 h incubation, the medium was removed, and 100 μL fresh medium including the test sample was added to the cells. A 100 μg mL^{-1} doxorubicin, as a standard therapeutic drug, as standard therapeutic drug, was used as a positive control. Following compound addition, plates were cultured for 24 h at 37 °C. Afterwards, the assay was performed according to the manufacturer's instructions of *CellTiterBlue* Viability Assay and measured using the microplate reader Tecan Infinite M200 at excitation 560 nm and emission of 590 nm. IC_{50} value was calculated as the concentration that shows 50% inhibition of the viability. Antimicrobial assay was performed using *Xanthomonas campestris* DSM 2405. The cultivation took place in TSB medium (1.2% tryptic soy broth; 0.5% NaCl). An overnight culture of the test organism was prepared and diluted to an optical density (600 nm) of 0.03. To prepare the assay, the samples (20 mg mL^{-1} stock solution) were dissolved in medium and transferred into a 96-well microtiter plate and 200 μL of the diluted culture was added to each well. The inoculated microplate was incubated for 7 h at 28 °C and 200 rpm. To detect the inhibitory effect of the substances, 10 μL of a resazurin solution (0.3 mg mL^{-1} phosphate-buffered saline) was added to the microplate and incubated again for 1 h, and the fluorescence signal (560 nm/590 nm) was measured using the microplate reader Tecan Infinite M200. Chloramphenicol was used as the positive control. IC_{50} value was calculated as the concentration that shows 50% inhibition of the viability.

3.5. Supplementary information

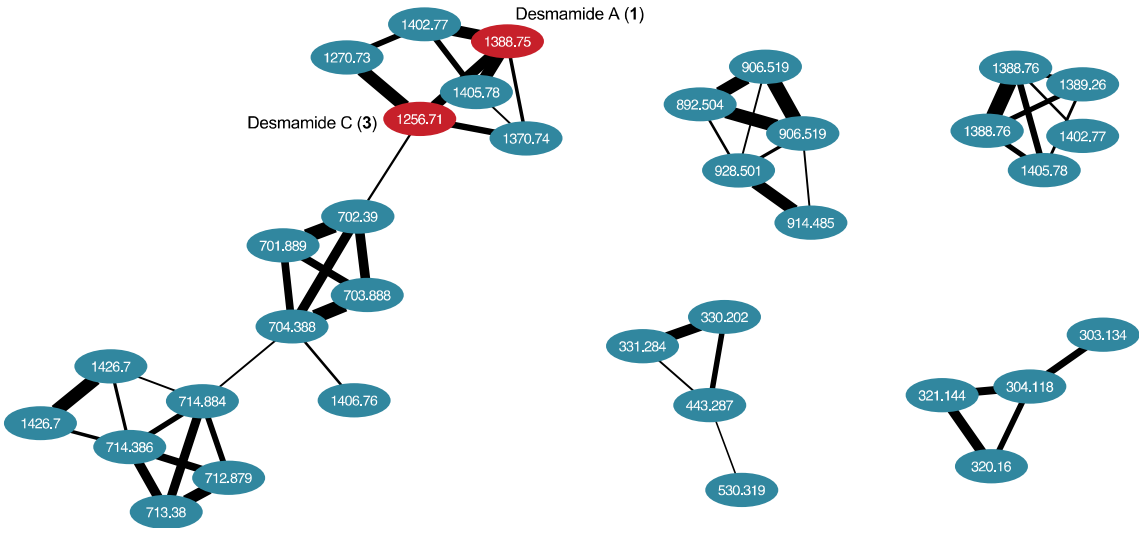


Figure S1 - GNPS molecular networking nodes present only in fraction C (eluting with MeOH), obtained from an organic extract of *D. muscorum* LEGE 12446 biomass. The mass features highlighted in red guided the isolation procedures.

Table S1 - Mass features from the molecular networking analysis obtained using LC-HRESIMS/MS data for fractions A, B and C of the organic extracts of LEGE 12446. The more intense masses in fraction C are shown in bold.

Mass value	Intensity		Adducts (calc.)					
m/z	Fraction C	Crude	$[M+NH_4]^+$	$[M+Na]^+$	$[M+2H]^{2+}$	$[M+H+NH4]^{2+}$	$[M+H+Na]^{2+}$	$[M+2Na]^{2+}$
1270.730	1.60E+06	-	1287.756	1292.711	635.869	643.878*	646.356*	657.851
1402.770	2.60E+07	3.40E+05	1419.801	1424.756	701.891	709.900*	712.378*	723.873
1256.710	2.30E+07	7.21E+05	1273.741	1278.696	628.861	636.870	639.348*	650.843
1370.740	1.30E+06	-	1387.772*	1392.728*	685.877	693.886*	696.364*	707.859
1388.750	4.30E+08	1.15E+07	1405.783	1410.738	694.882	702.891	705.369*	716.864
906.519	6.70E+07	1.92E+07	923.545*	928.501	453.763	461.773*	464.250*	475.745
892.504	3.60E+07	2.77E+07	909.530*	914.485	446.756	454.765*	457.243*	468.737
530.319	1.70E+06	1.97E+05	547.345	552.301	265.663*	273.673*	276.150*	287.645*
443.286	1.46E+06	1.05E+05	460.313	465.268	222.147*	230.157*	232.634*	244.129*
330.202	2.90E+06	1.26E+05	347.229*	352.184	165.605*	173.614*	176.092*	187.587*
331.284	1.32E+08	3.31E+05	348.311*	353.266*	166.146*	174.155*	176.633*	188.128*
320.16	1.43E+06	-	337.187*	342.142	160.584*	168.594*	171.071*	182.566*
321.144	5.30E+06	2.26E+05	338.171*	343.126	161.076*	169.085*	171.563*	183.058*
303.134	4.20E+06	3.08E+05	320.160*	325.116	152.071*	160.080*	162.558*	174.053*

*Masses not found in the MS spectra.

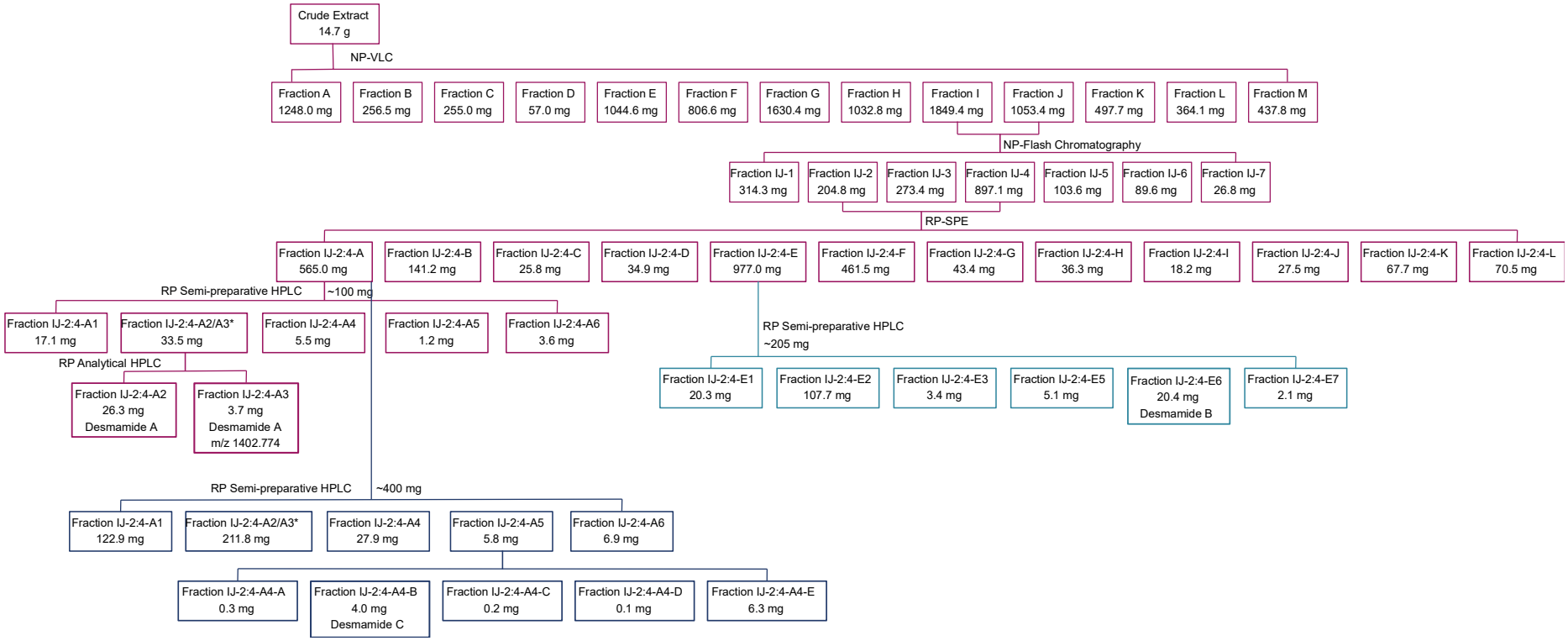


Figure S2 - Fractionation scheme leading to the purification of **1-3.**

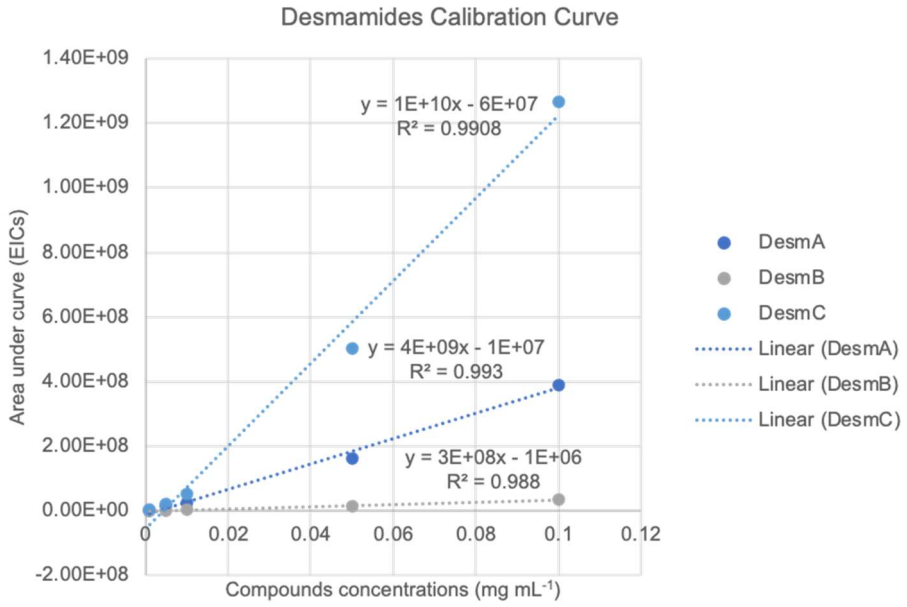


Figure S3 - LC-HRESIMS (positive mode) calibration curves (using standards of **1**-**3**) used for estimating the amount of desmamides in the organic extract used for isolation.

Table S2 – LC-HRESIMS data for desmamides A-C (1-3) standards used for quantification of the compounds in the cyanobacterial extract/biomass. Also shown are the data in the organic extract.

Concentration	**1 (EIC)**	**2 (EIC)**	**3 (EIC)**
mg mL^{-1}	**Area under curve**	**Area under curve**	**Area under curve**
0.5	1.95E+09	3.64E+08	4.59E+09
0.1	3.90E+08	3.25E+07	1.27E+09
0.05	1.60E+08	1.25E+07	5.03E+08
0.01	2.48E+07	1.82E+06	5.24E+07
0.005	1.29E+07	8.63E+05	1.98E+07
0.001	1.03E+06	1.71E+05	2.24E+06
Crude Extract (1 mg mL^{-1})	7.73E+07	1.29E+07	3.07E+06
g of compound per gram of extract	0.022	0.046	0.006
g of compound per gram of biomass	0.456	0.967	0.132

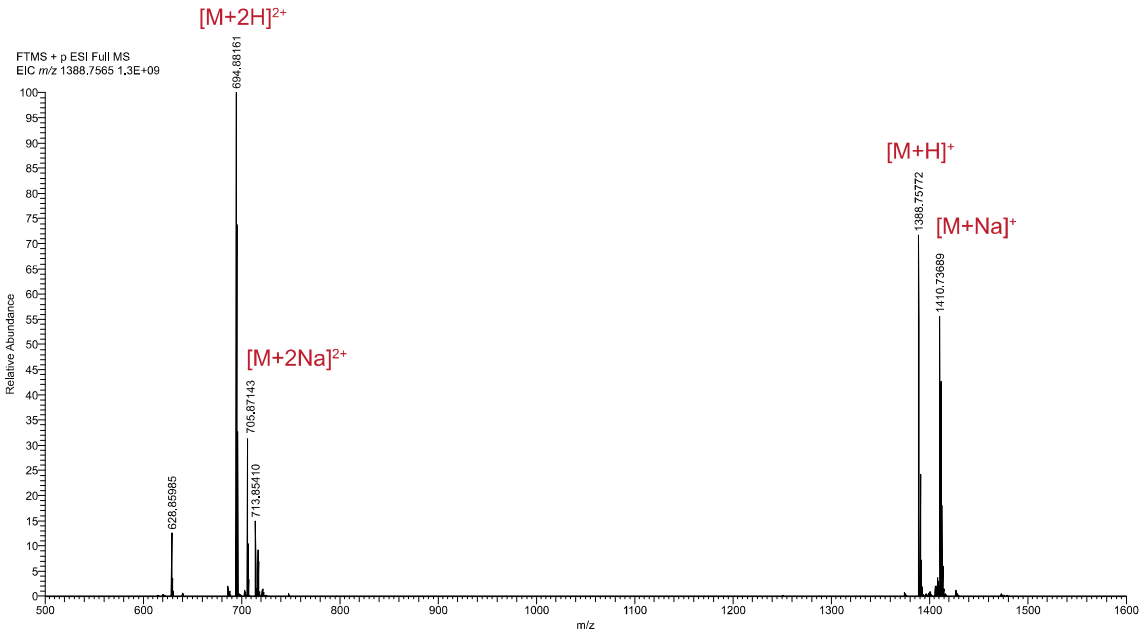


Figure S4 - HRESIMS spectrum of **1**.

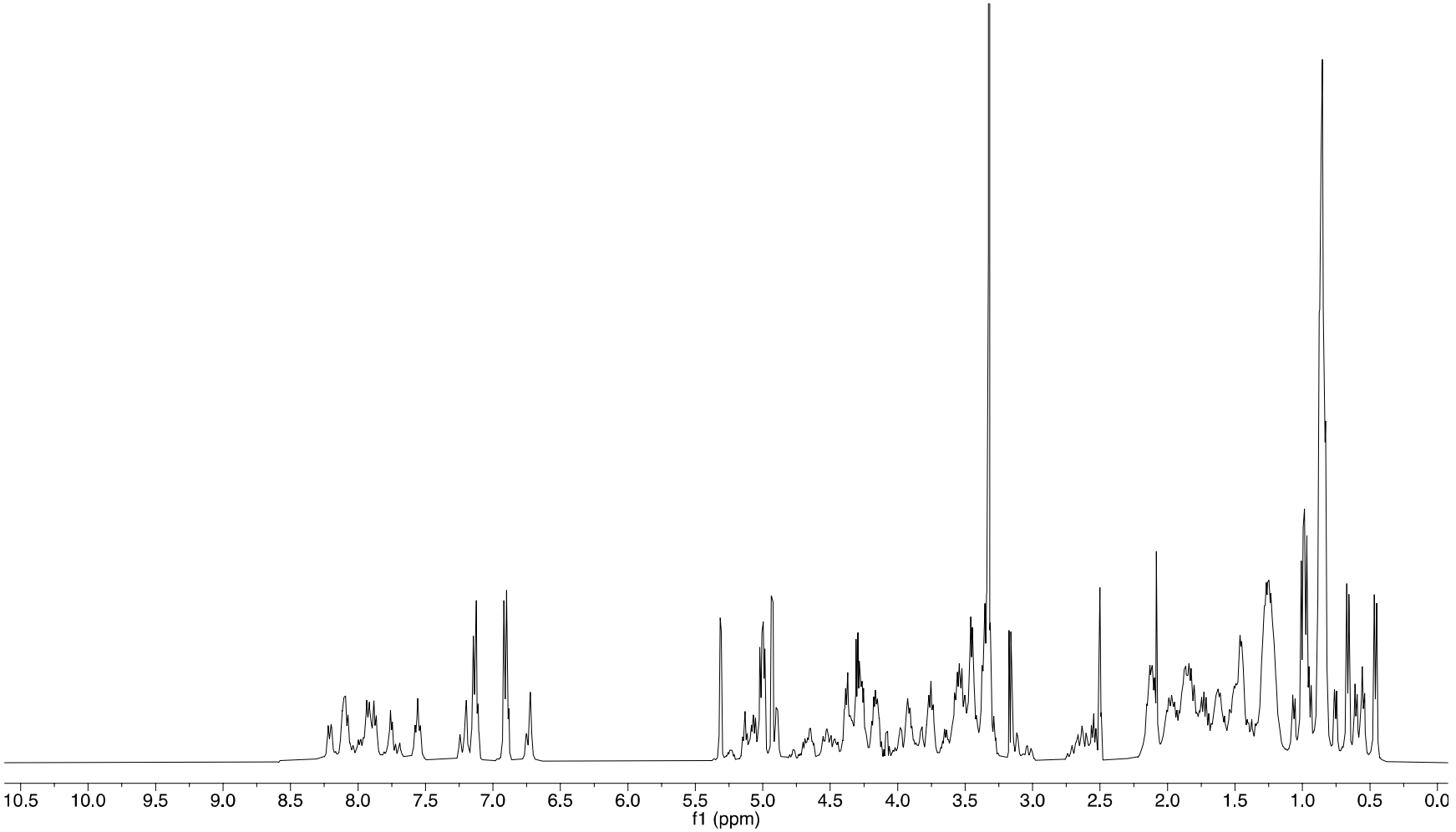


Figure S5 - ^{1}H NMR (DMSO-d_6, 400 MHz) spectrum of compound **1**.

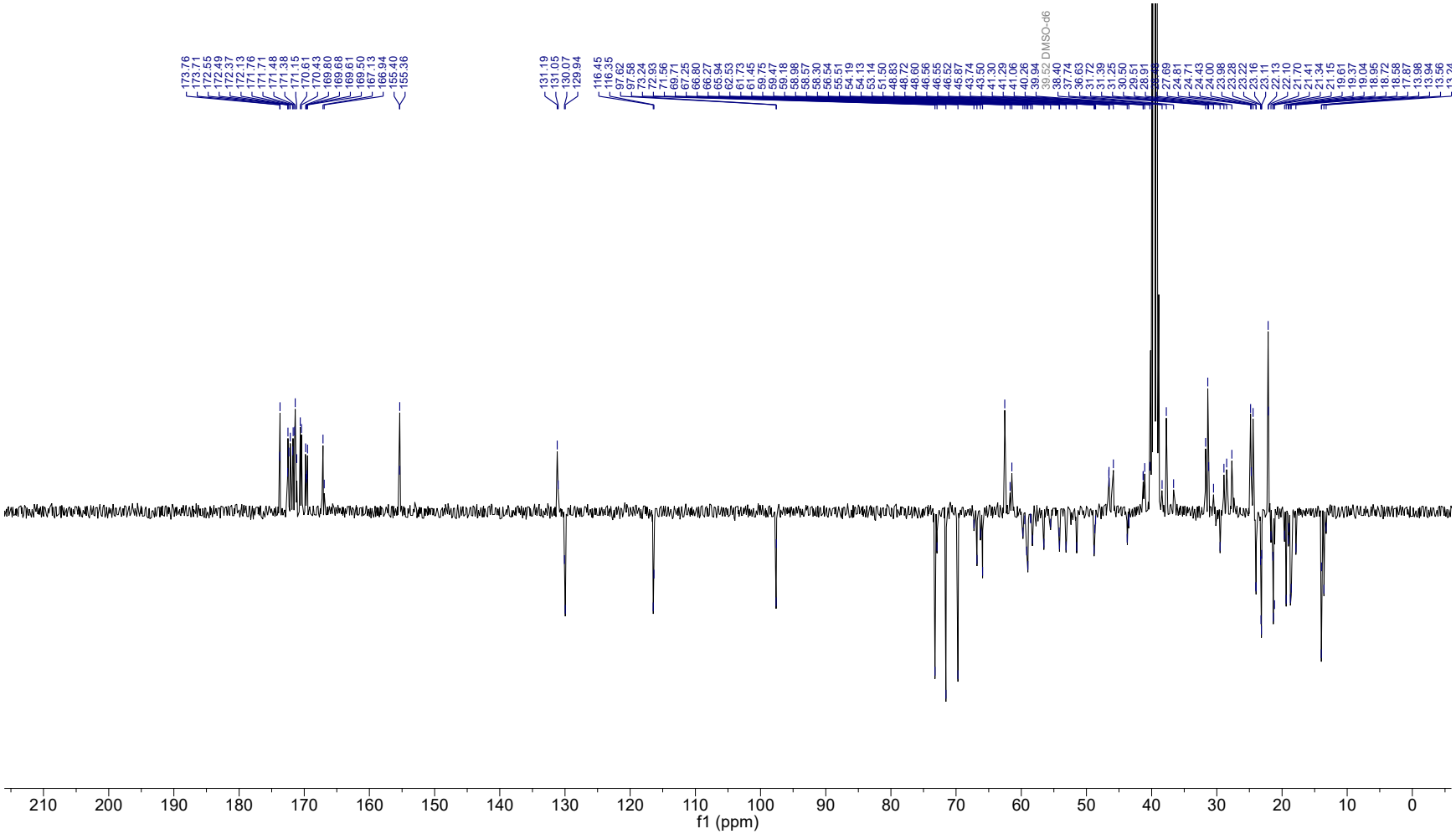


Figure S6 - ^{13}C NMR (APT, DMSO-d_6, 100 MHz) spectrum of compound **1**.

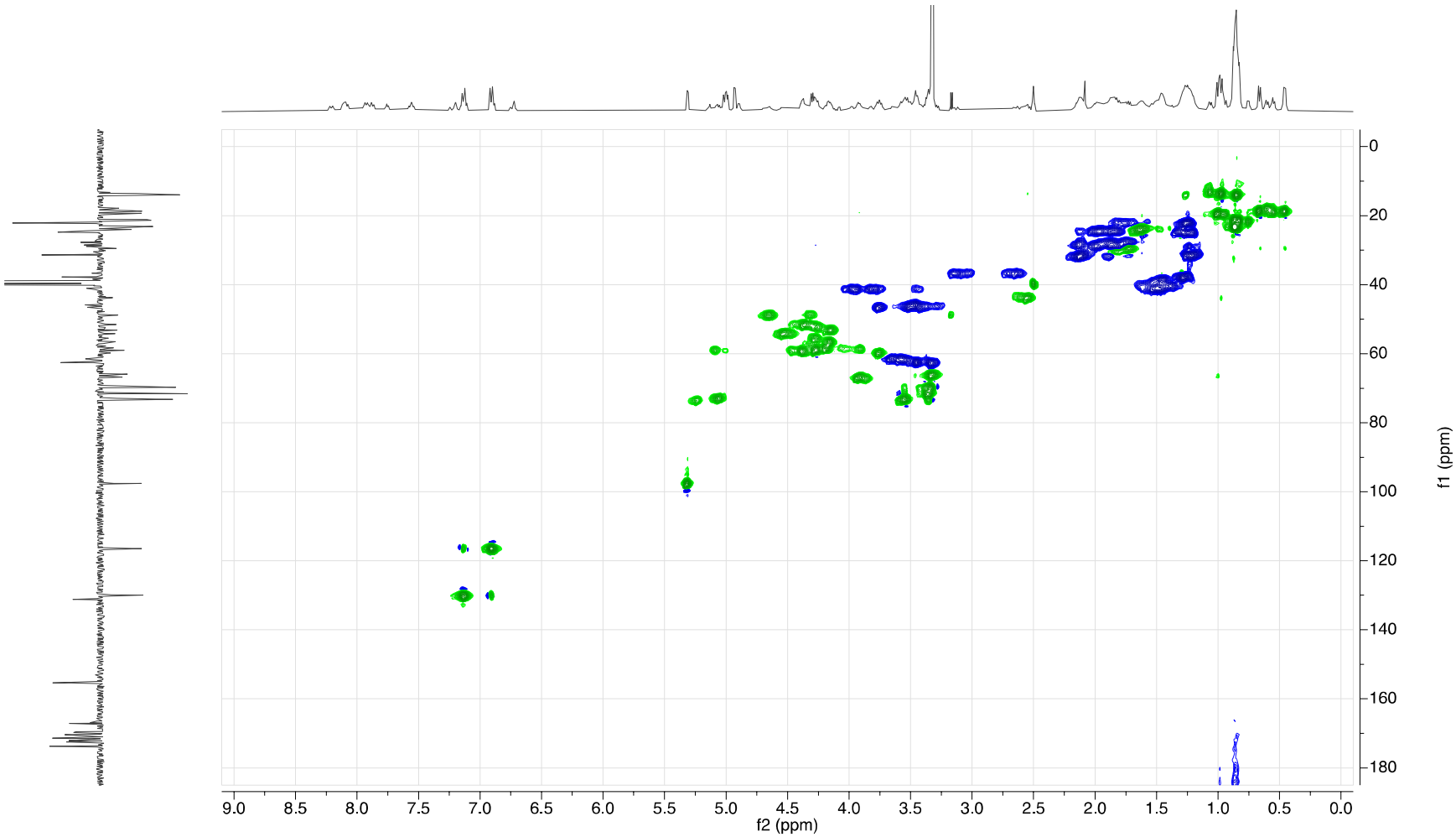


Figure S7 - HSQC (DMSO-d_6, 400 MHz) spectrum of compound **1**.

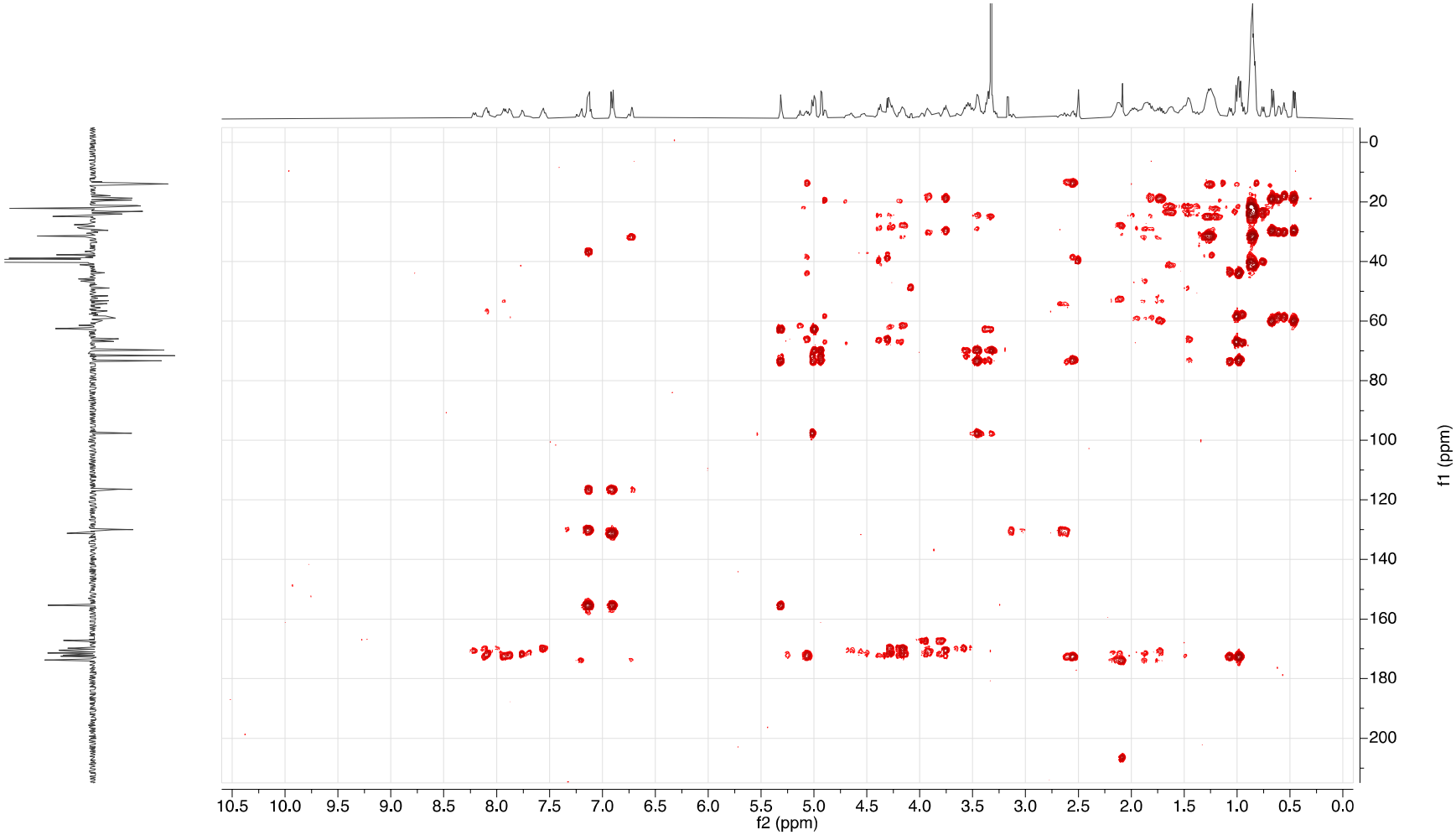


Figure S8 - HMBC (DMSO-d_6, 400 MHz) spectrum of compound **1**.

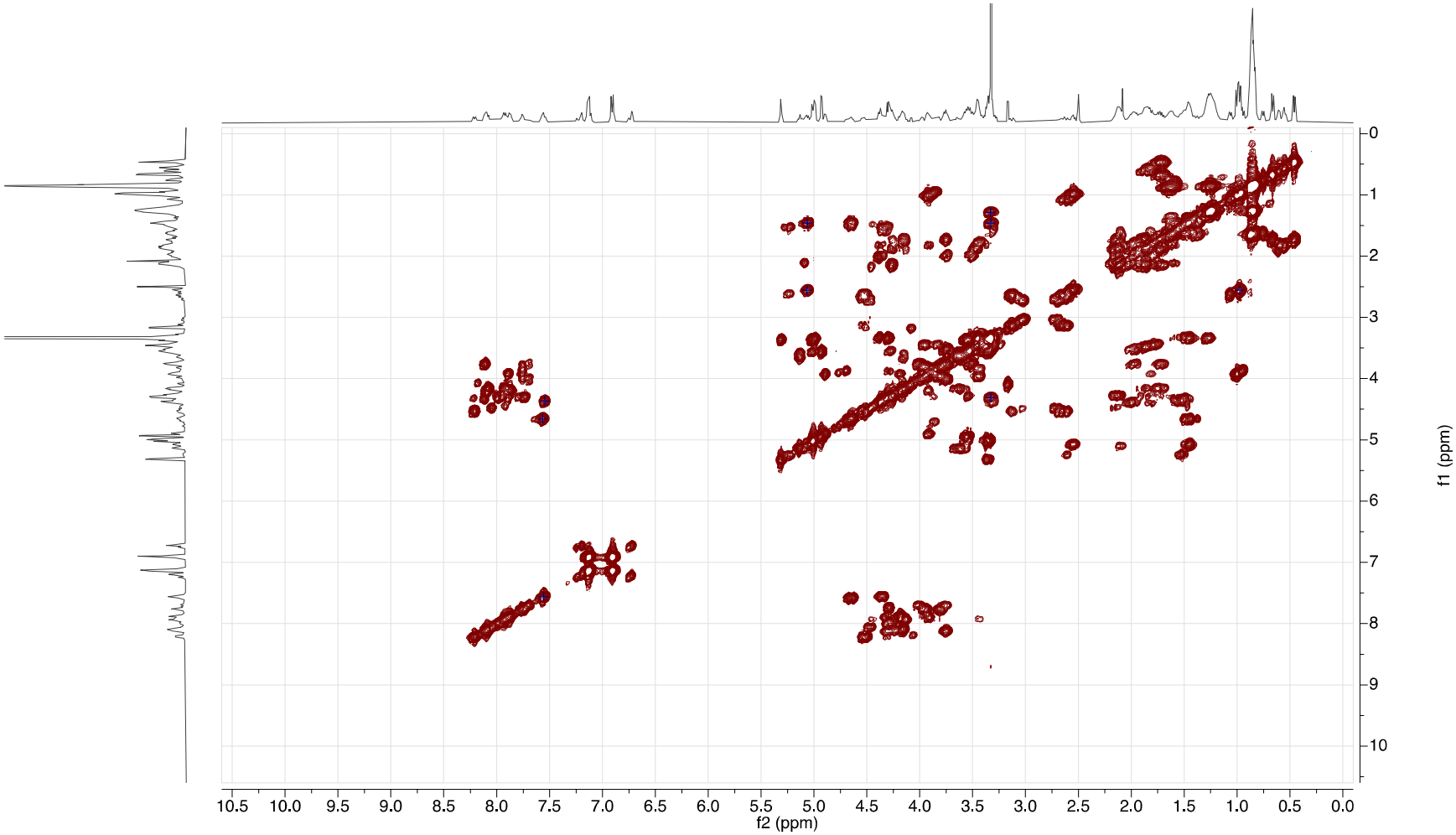


Figure S9 - COSY (DMSO-d_6, 400 MHz) spectrum of compound **1**.

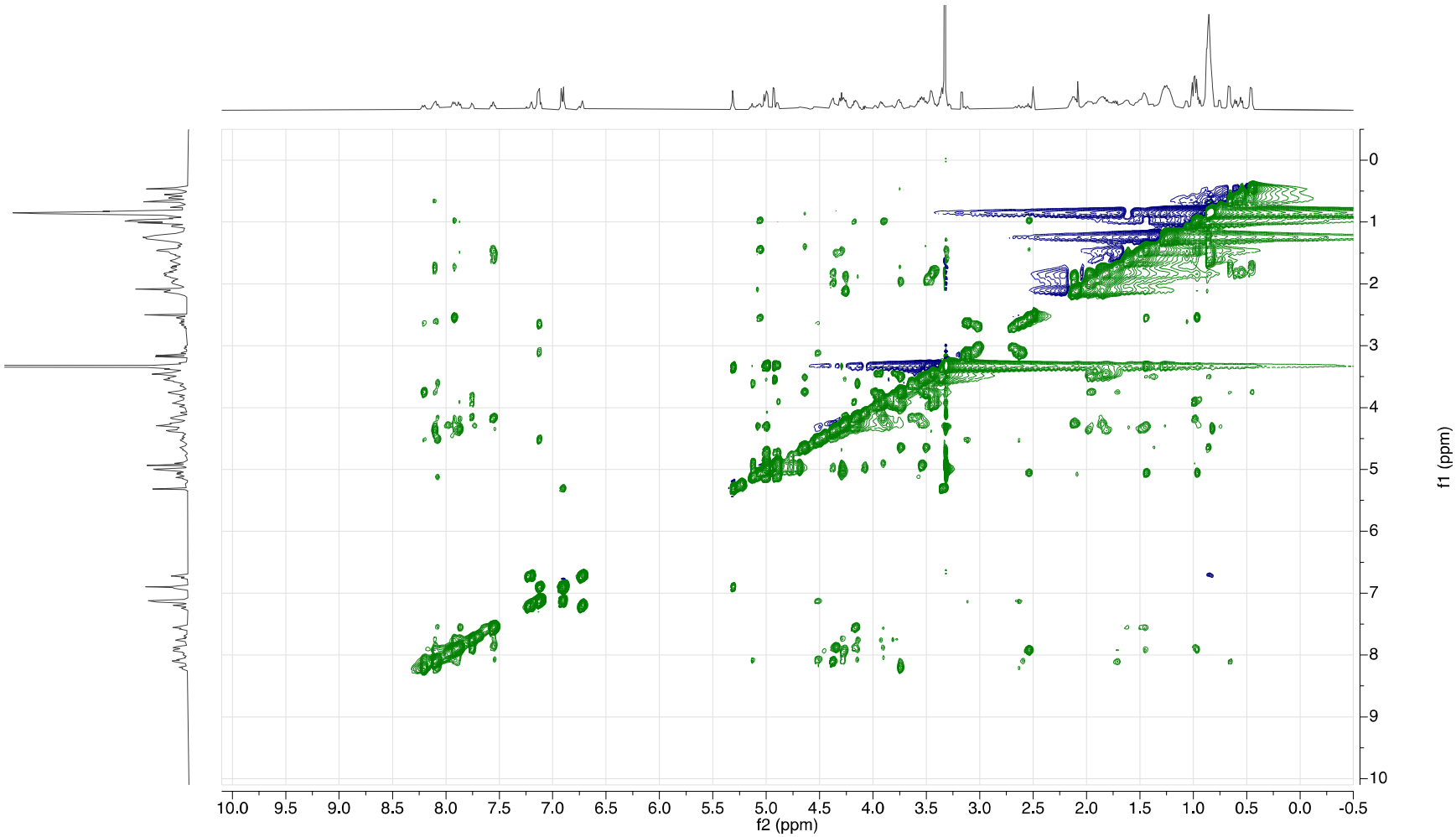


Figure S7 - NOESY (DMSO-d_6, 400 MHz) spectrum of compound **1**.

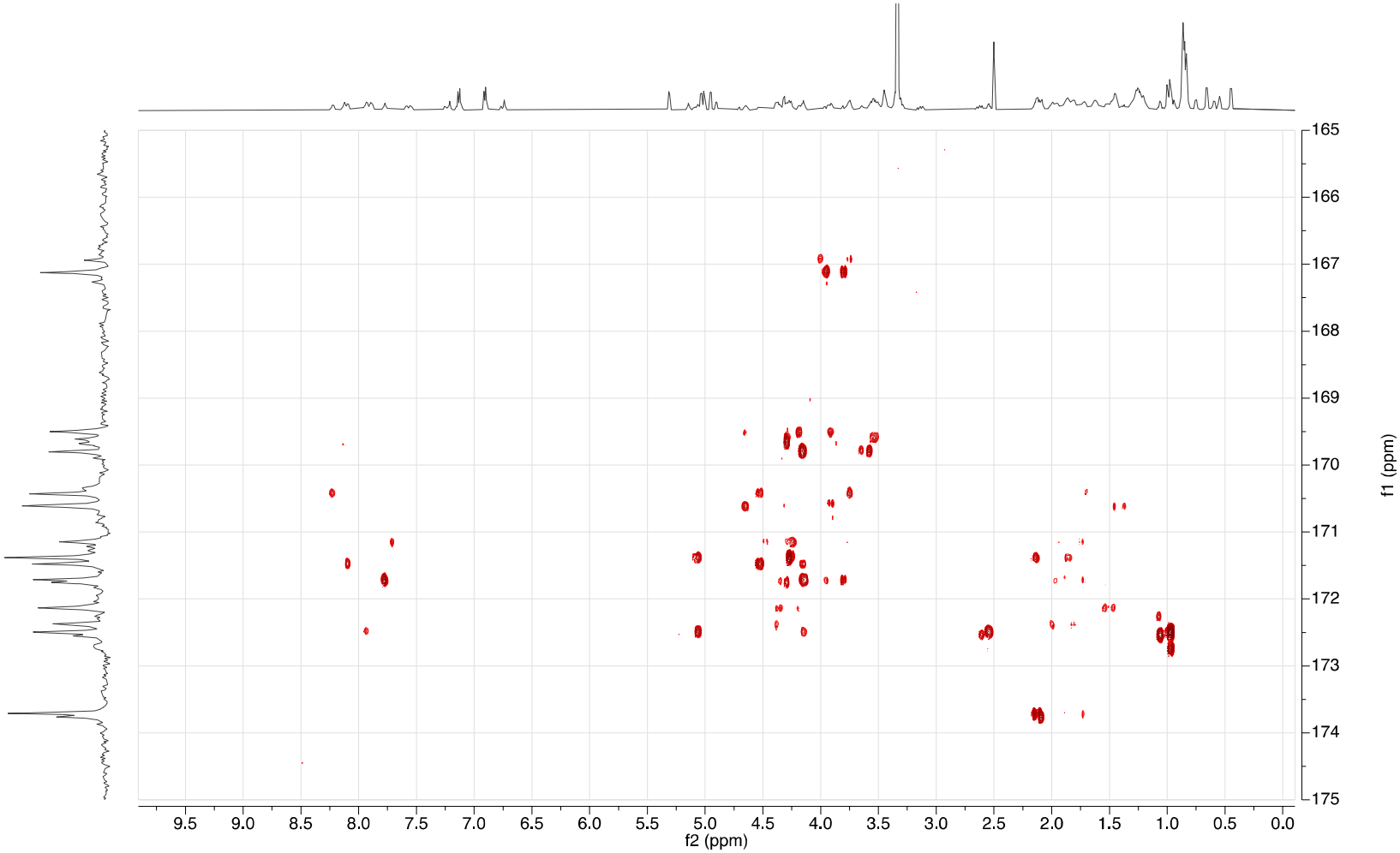


Figure S8 - Band-selective HMBC (DMSO-d_6, 600 MHz) spectrum of compound **1**.

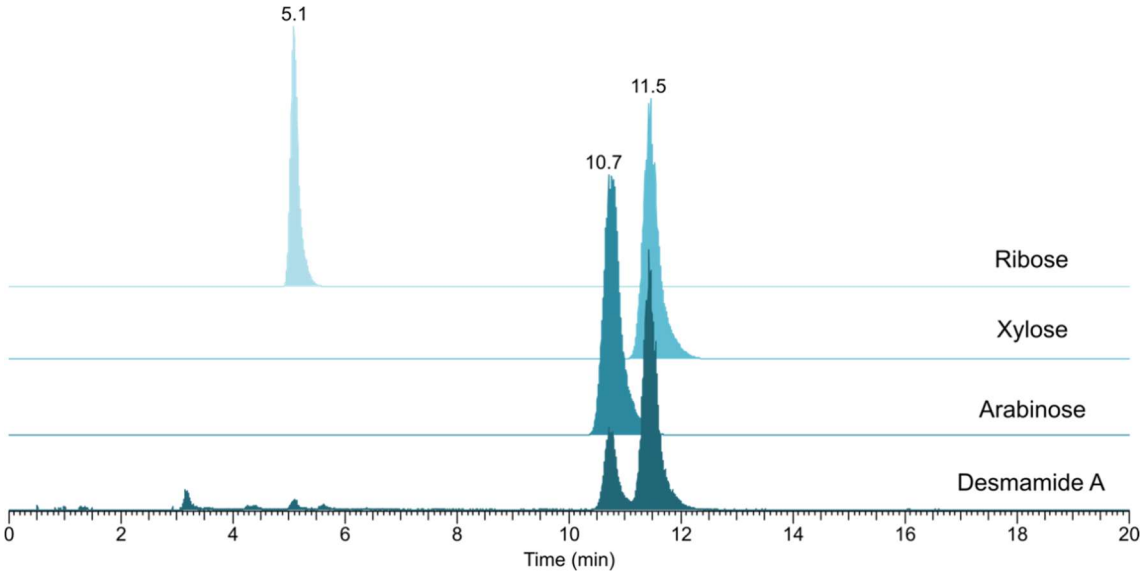


Figure S9 - LC-HRESIMS analysis of PMP-sugar derivatives to determine the sugar composition of **1**. Shown are EICs for each of the sugar standard derivatives and for the derivatized hydrolysate product of desmamide A (**1**).

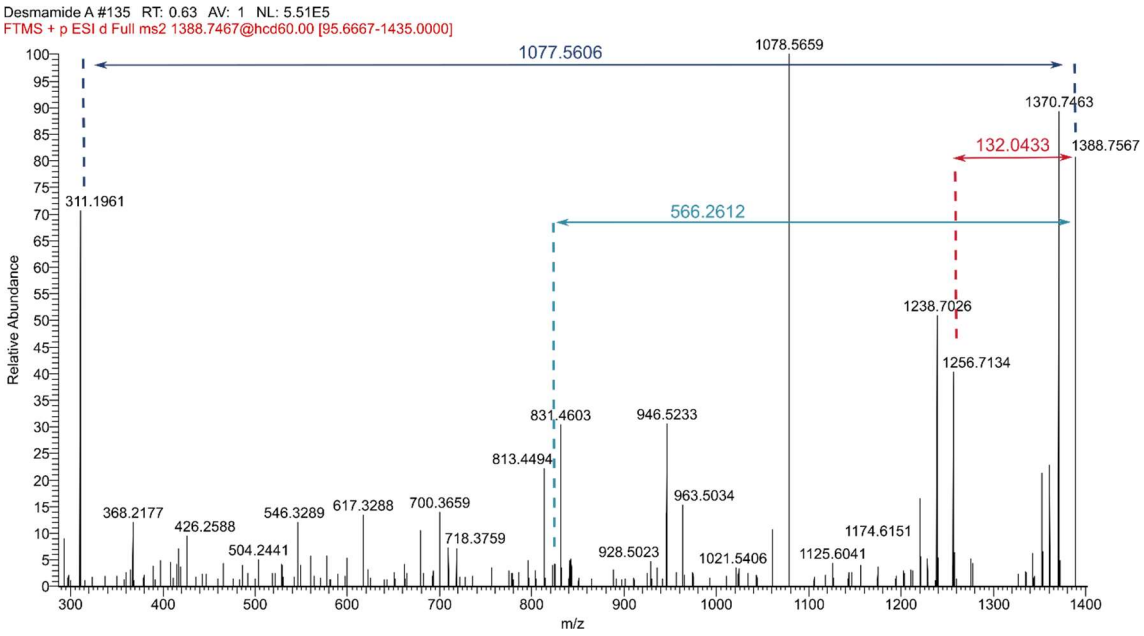


Figure S10 - HRESIMS/MS spectrum of **1**, highlighting key fragmentations supporting the NMR-based residue sequence proposal.

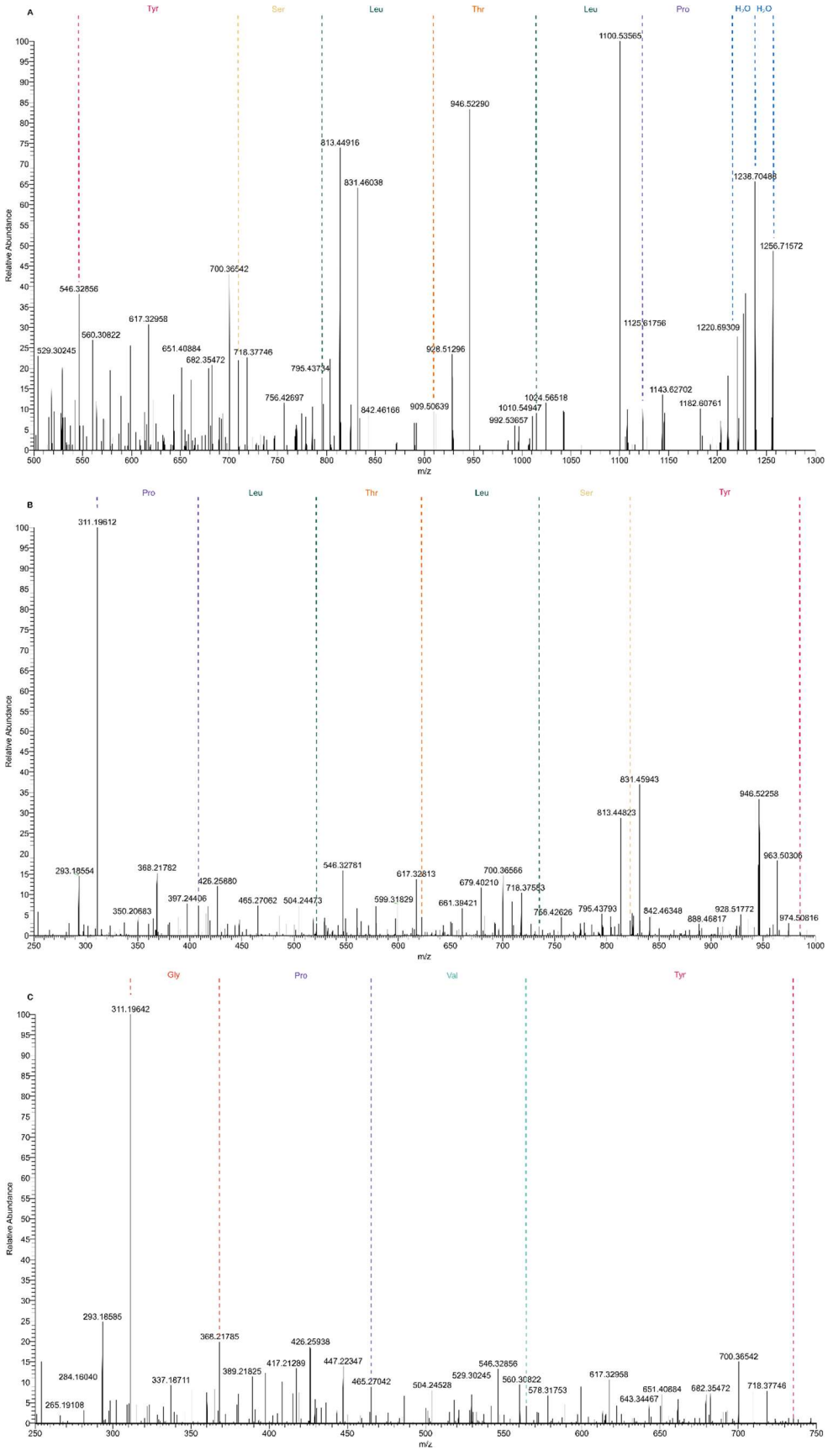


Figure S14 - Annotated HRESIMS/MS spectrum for **3**, highlighting fragmentations leading to amino acid sequence confirmation. Three regions of the same spectrum are shown (A, B and C).

Table S3 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for desmamide B (**2**) in DMSO-*d*6.

unit	no.	δC, type	δH[a] (*J* in Hz)	HMBC[b]	COSY
Akyl chain	1	172.5, C			
	2	43.7, CH	2.55, m	13.6, 72.9, 172.5	0.97, 5.06
	3	72.9, CH	5.06, m	171.7-172.5	1.46, 2.55
	4	38.4,CH_2	1.46, m		3.33, 5.06
	5	65.9, CH	3.33, m		1.28, 1.46, 4.31
	6	37.8, CH_2	1.28, m	24.8, 31.4	3.33
	7	31.4, CH_2	1.24, m	24.8	
	8	24.8, CH_2	1.29, m	31.4	0.85
	9	22.1, CH_2	1.26, m	14.0, 24.8, 31.4	0.85
	10	14, CH_3	0.85, m	22.1, 31.4	1.24-1.29
	11	13.6, CH_3	0.97, d (7.0)	43.7, 72.9, 172.5	2.55
	OH		4.31, d (5.8)	38.4, 65.9	3.33
Glu	12	53.2, CH	4.14, m		7.93
	13a	27.7, CH_2	1.88, m		1.75, 2.12
	13b		1.75, m		1.88, 2.12
	14a	31.7, CH_2	2.12, m	53.2, 173.7	1.74-1.89
	14b		1.89, m		
	15	173.7, C			
	16	171.7, C			
	Nha		7.93, d (7.7)		4.14
	NHb		6.74, d (11.5)		7.23
	NHb		7.23, d (18.8)		6.74
Gly	17a	41.3, CH_2	3.96, m	167.1	3.80, 7.76
	17b		3.80, m	167.1	3.96, 7.76
	18	167.1, C			
	NH		7.76, m	171.7	3.96, 3.80
Pro_1	19	59.2, CH	4.38, m	28.9	2.0
	20	28.9, CH_2	2.0, m		4.38
	21	24.4, CH_2	1.98, m		3.76/3.51
	22a	46.6, CH_2	3.51, m		1.98, 3.76
	22b		3.76, m		1.98, 3.51
	23	172.3, C			
Val	24	59.7, CH	3.75, m	18.6, 29.5, 170.4	1.72, 8.11
	25	29.5, CH	1.72, m	18.6, 59.7	0.46, 0.66, 3.75
	26	18.6, CH_3	0.46, d (6.7)	18.7, 29.5, 59.7	1.72
	27	18.7, CH_3	0.66, d (6.6)	18.6, 29.5, 59.7	1.72
	28	170.4, C			
	NH		8.11, d (5.7)		3.75

Tyr	29	54.2, CH	4.53, m		2.65, 3.13, 8.25
	30a	36.7, CH_2	2.65, m	130.2	3.13, 4.53
	30b		3.13, m		2.65, 4.53
	31	131.5, C			
	32	130.0/130.2, CH	7.15, d (8.5)	36.7, 116.5, 155.0	6.93
	33	116.5, CH	6.93, d (8.6)	131.5, 155.0	7.15
	34	155, C			
	35	116.5, CH	6.93, d (8.6)	131.5, 155.0	7.15
	36	130.0/130.2, CH	7.15, d (8.5)	36.7, 116.5, 155.0	6.93
	37	171.5, C			
	NH		8.25, d (6.6)		4.53
Ser	38	56.6, CH	4.16, m	61.4, 169.8	3.64/3.57, 8.10
	39a	61.4, CH_2	3.64, m		4.16, 5.13
	39b		3.57, m		4.16
	40	169.8, C			
	NH		8.10, d (5.7)		4.16
	OH		5.13, t (5.0)		3.64
Leu_1	41	48.9, CH	4.32, m		1.50-1.55, 7.57
	42	40, CH_2	1.50-1.55, m	23.2	4.32
	43	24, CH	1.65, m		0.87
	44	23.2, CH_3	0.87, m	24.0, 40.0	1.65
	45	23.2, CH3	0.87, m	24.0, 40.0	1.65
	46	172.2, C			
	NH		7.57, m		4.32
Thr	47	58.3, CH	4.19, m	19.6	3.91, 7.89
	48	66.8, CH	3.91, m		1.0, 4.19, 4.89, 7.89
	49	19.6, CH_3	1.0, m	58.3, 66.8	3.91
	50	169.5, C			
	NH		7.89, d (7.8)		3.91, 4.19
	OH		4.89, d (4.5)	19.6, 58.3, 66.8	3.91
Leu_2	51	48.9, CH	4.65, m		1.40-1.50, 7.60
	52	41, CH_2	1.40-1.50, m		4.65
	53	24, CH	1.65, m		0.83
	54	21.7, CH_3	0.83, m	24.0, 41.0	1.65
	55	21.7, CH3	0.83, m	24.0, 41.0	1.65
	56	170.6, C			
	NH		7.60, m		4.65
Pro_2	57	59, CH	4.27, dd (8.7)	45.9	2.13
	58	28.5, CH_2	2.13, m		4.27, 1.87
	59	24.4, CH2	1.87, m		2.13, 3.46
	60	45.9, CH_2	3.46, m	24.4	1.87

	61	171.4, C			
Xylosyl	1'	97.3, CH	5.40, d (3.5)	58.8, 69.9, 71.5, 155.0	3.48, 3.58
	2'	71.5, CH	3.48, dd (3.5, 13.0)	69.9	3.78, 5.31, 5.40
	3'	69.9, CH	3.78, m	71.4	3.48, 4.62, 5.29
	4'	71.4, CH	4.62, dd (9.6, 10.7)	58.8, 69.9, 170.0	3.78, 3.37/3.58
	5'a	58.8, CH_2	3.37, dd (5.8)	71.4, 97.3	3.58, 4.62
	5'b		3.58, m	71.4, 97.3	3.37, 4.62, 5.40, 5.29
	OHa		5.31, d (5.6)	69.9, 71.5, 97.3	3.48
	OHb		5.29, d (5.0)	71.4, 97.3	3.78, 3.58
4'-O-Ac	6'	170, C			
	7'	21.2, CH_3	2.02, m	71.4	

[a]from HSQC, [b]from proton to indicated carbon.

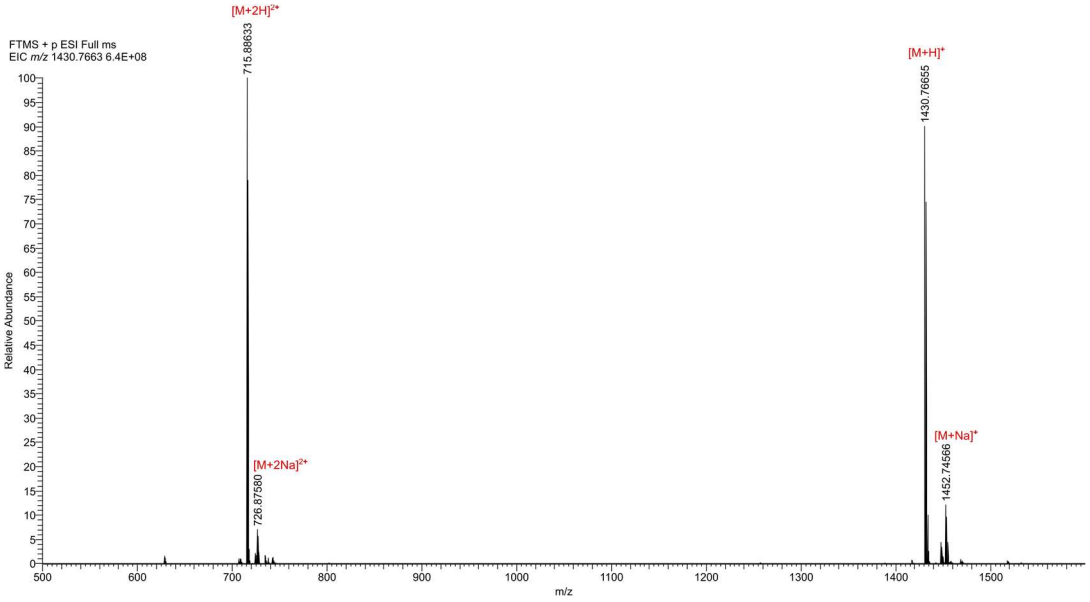


Figure S15 - HRESIMS spectrum of **2**.

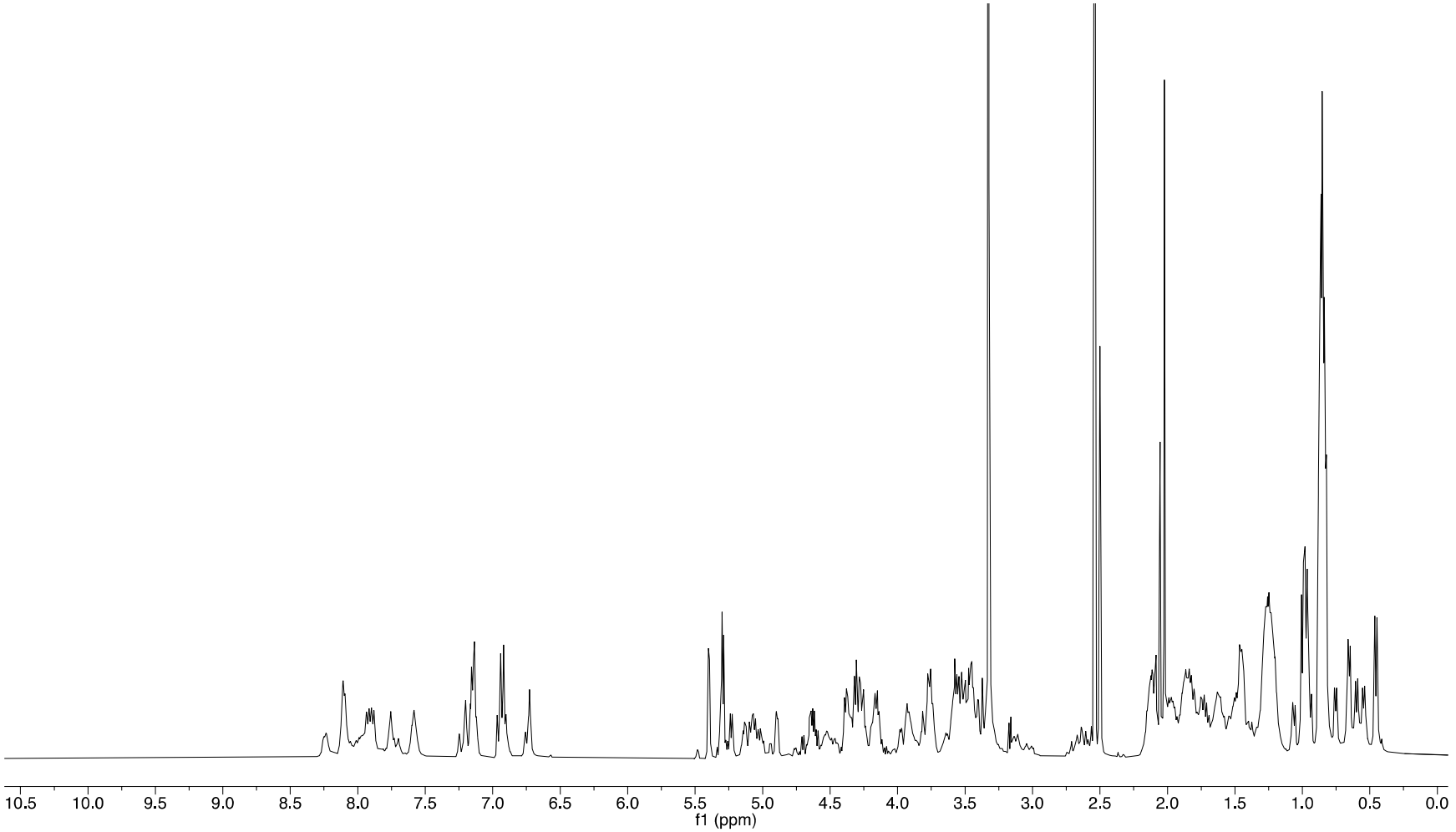


Figure S16 - ^{1}H NMR (DMSO-d_6, 100 MHz) spectrum of compound **2**.

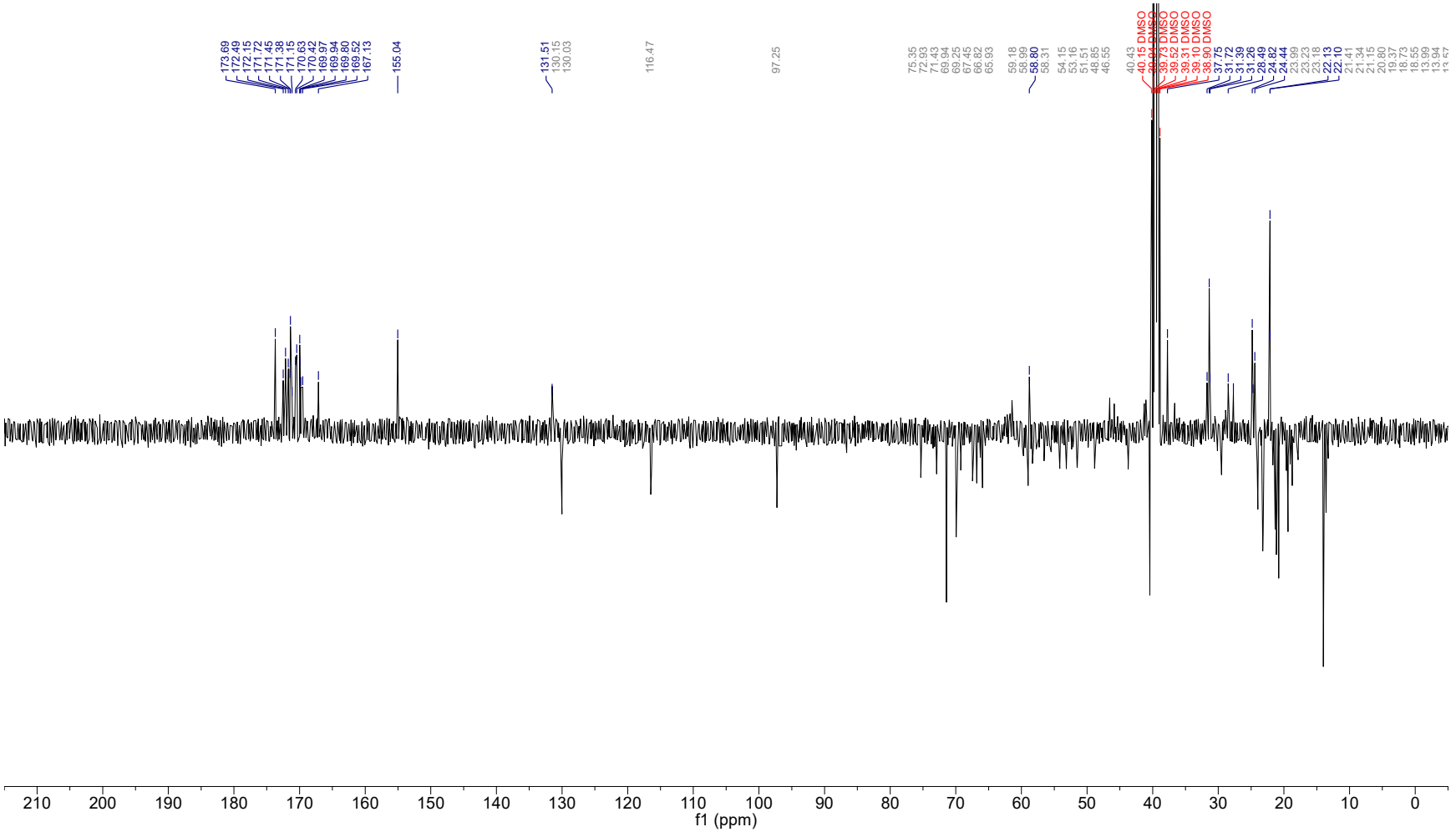


Figure S17 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **2**.

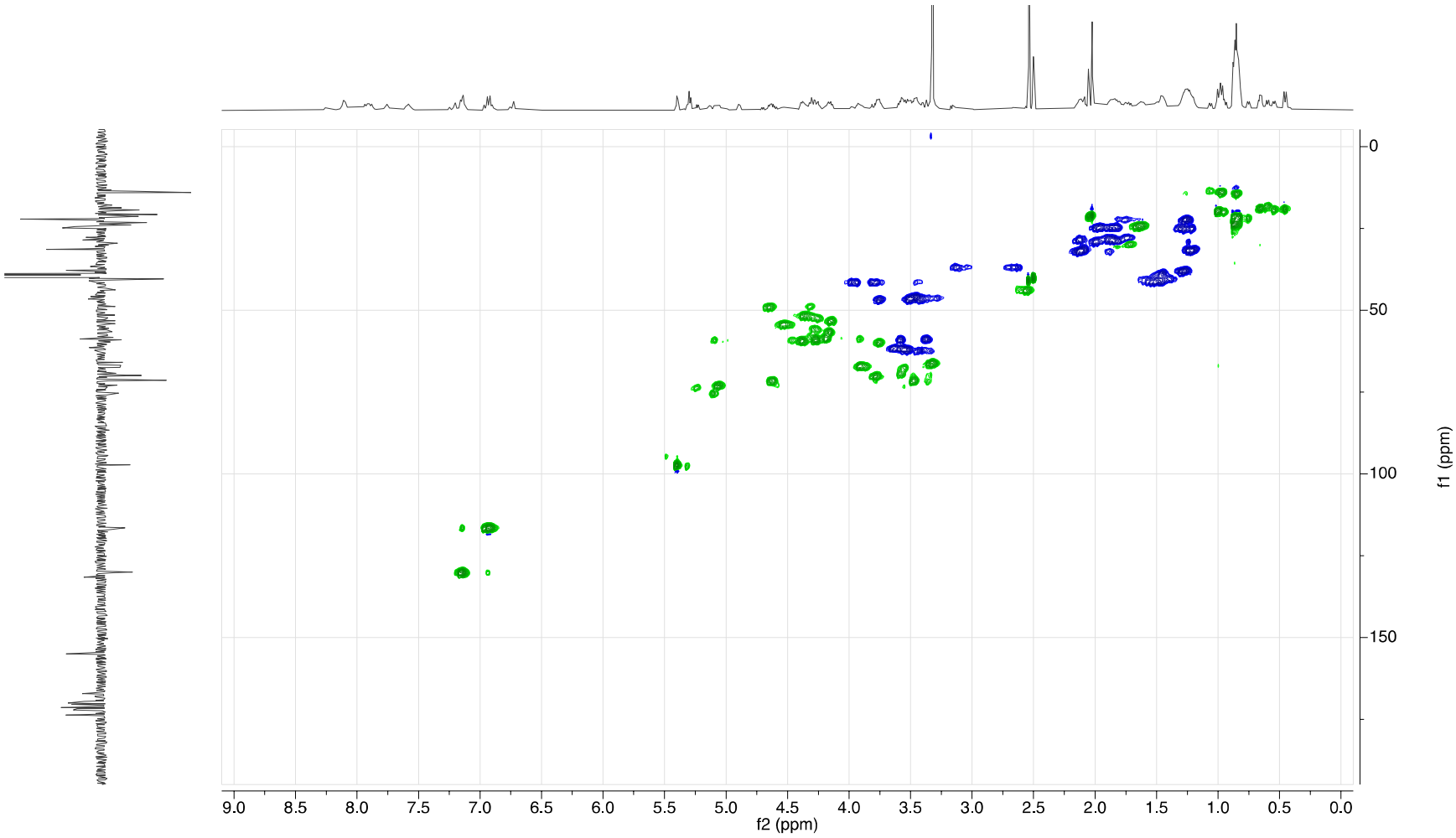


Figure S18 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **2**.

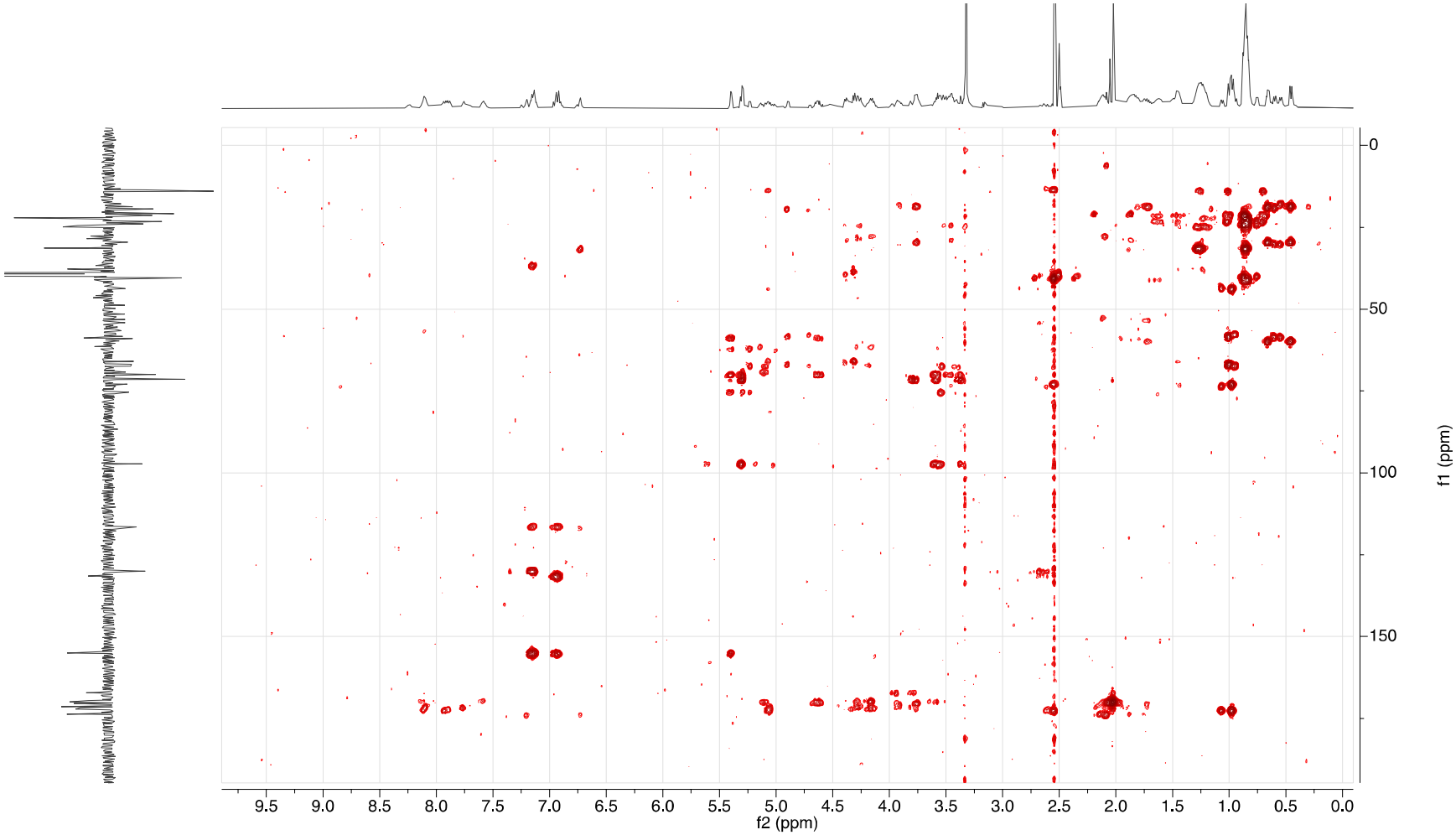


Figure S19 - HMBC (DMSO-d_6, 600 MHz) spectrum of compound **2**.

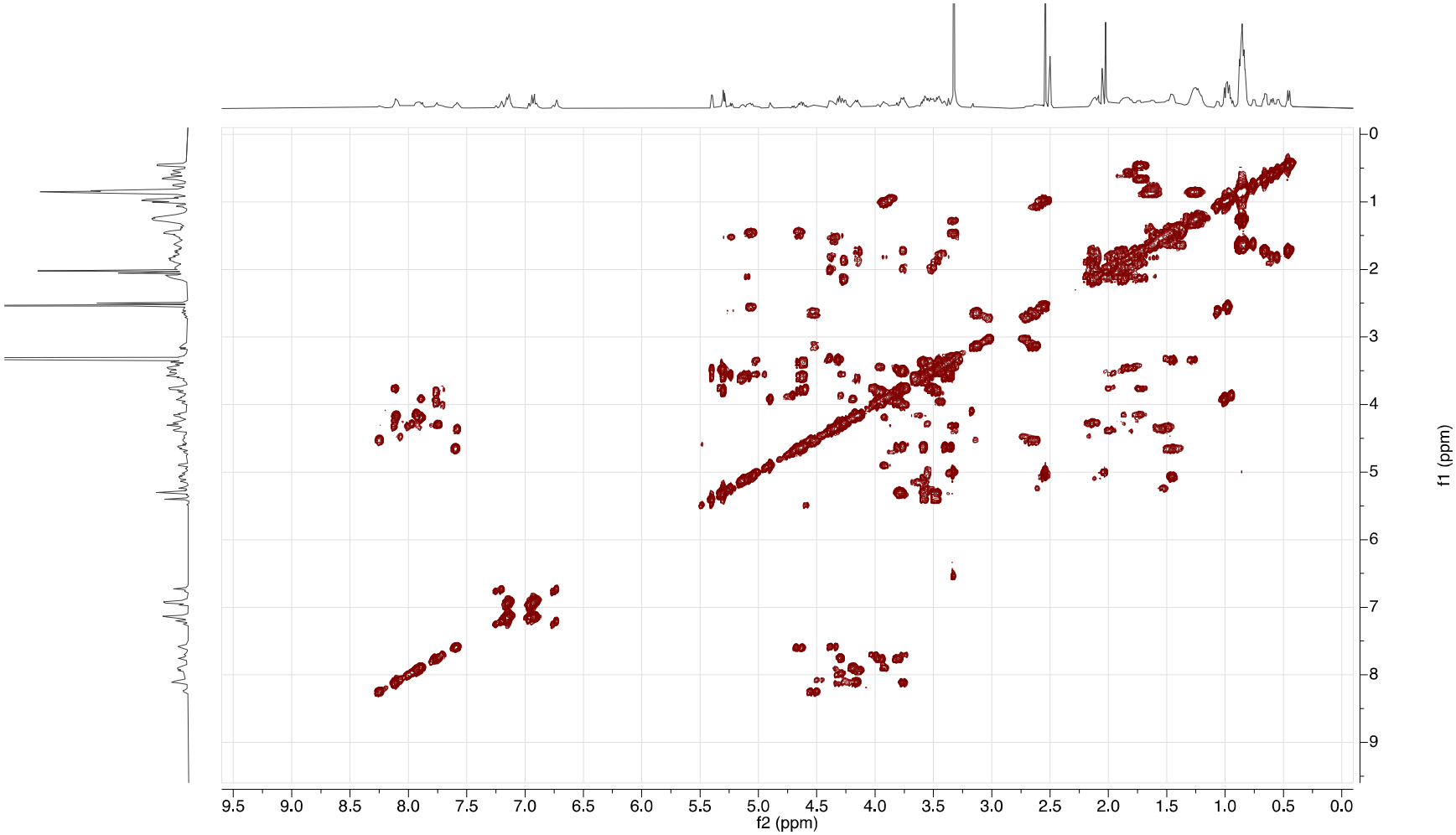
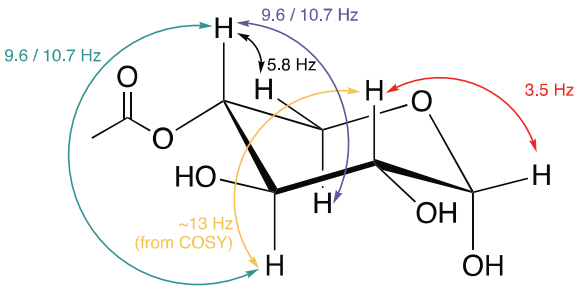


Figure S20 - COSY (DMSO-d_6, 600 MHz) spectrum of compound **2**.



4-*O*-acetyl-α-xylose
(depicted is the assumed D-configuration)

Figure S11 - Coupling constant analysis (^{1}H and COSY data, DMSO-d_6, 600 MHz) underscoring the proposed sugar configuration of compound **2**.

Table S4 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for desmamide C (**3**) in DMSO-d_6.

unit	no.	δ_C, type	δ_H (*J* in Hz)
Dhmda	1	172.5, C	
	2	43.6, CH	2.55, m
	3	72.9, CH	5.05, m
	4	38.4, CH_2	1.45, m
	5	65.9, CH	3.33, m
	6	37.7, CH_2	1.28, m
	7	31.4, CH_2	1.28, m
	8	24.8, CH_2	1.28, m
	9	22.1, CH_2	1.25, m
	10	14.0, CH_3	0.86, m
	11	13.6, CH_3	0.97, m
	OH		4.31, d (5.8)
Gln	12	53.2, CH	4.14, m
	13a	27.7, CH_2	1.86, m
	13b		1.75, m
	14a	31.7, CH_2	2.13, m
	14b		1.88, m
	15	173.7, C	
	16	171.7, C	
	NHa		7.95, d (7.7)
	NHb		6.71, d (13.2)
	NHb		7.18, d (24.8)
Gly	17a	41.3, CH_2	3.96, m
	17b		3.79, m
	18	167.0, C	
	NH		7.73, t (5.3)

Pro_1	19	59.1, CH	4.38, m
	20	28.8, CH_2	2.0, m
	21	24.4, CH_2	1.97, m
	22a	46.5, CH_2	3.51, m
	22b		3.75, m
	23	172.4, C	
Val	24	59.8, CH	3.76, m
	25	29.4, CH	1.74, m
	26	18.7, CH_3	0.50, d (6.7)
	27	18.8, CH_3	0.70, d (6.6)
	28	170.5, C	
	NH		8.14, d (6.7)
Tyr	29	54.4, CH	4.45, m
	30a	36.6, CH_2	2.62, m
	30b		3.05, m
	31	128.0, C	
	32	129.8, CH	7.00, d (8.5)
	33	114.9, CH	6.61, d (8.6)
	34	155.8, C	
	35	114.9, CH	6.61, d (8.6)
	36	129.8, CH	7.00, d (8.5)
	37	171.6, C	
	NH		8.33, br
	OH		9.13, s
Ser	38	56.6, CH	4.14, m
	39a	61.5, CH_2	3.65, m
	39b		3.76, m
	40	169.9, C	
	NH		8.11, d (5.7)
	OH		5.11, m
Leu_1	41	49.0, CH	4.33, m
	42	40.1, CH_2	1.43-1.51, m
	43	24.0, CH	1.64, m
	44	23.1, CH_3	0.87, m
	45	23.1, CH_3	0.87, m
	46	172.2, C	
	NH		7.68, m
Thr	47	58.3, CH	4.16, m
	48	66.8, CH	3.91, m
	49	19.4, CH_3	1.0, d (6.3)
	50	169.6, C	
	NH		7.92, d (7.5)
	OH		4.89, d (4.8)
Leu_2	51	49.0, CH	4.62, m

	52	41.1, CH_2	1.48, m
	53	24.0, CH	1.64, m
	54	21.3, CH_3	0.83, m
	55	21.3, CH_3	0.83, m
	56	170.7, C	
	NH		7.70, m
Pro_2	57	59.0, CH	4.27, dd (3.5, 8.7)
	58	28.8, CH_2	2.13, m
	59	24.4, CH_2	1.86, m
	60	45.8, CH_2	3.46, m
	61	171.4, C	

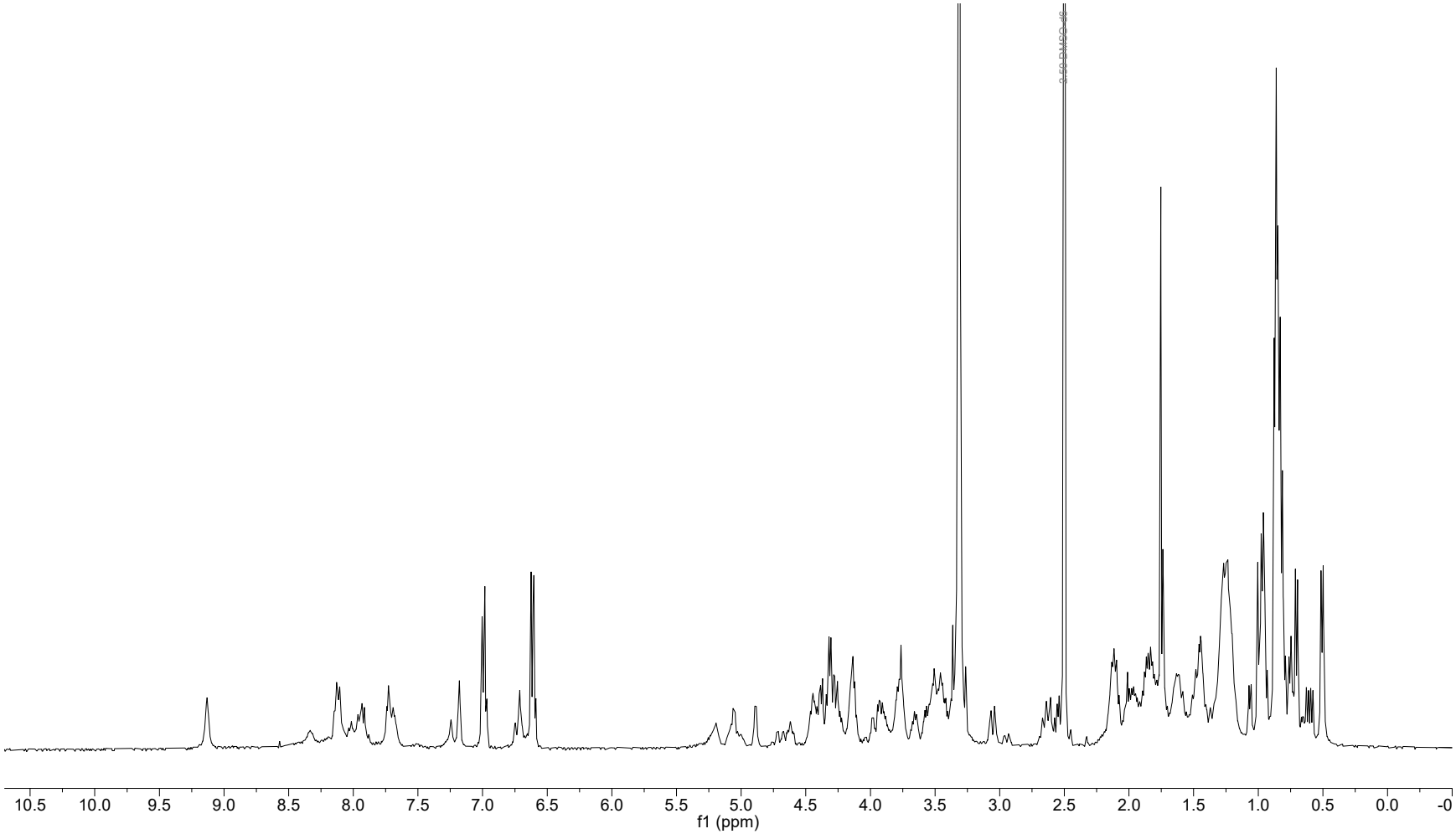


Figure S12 - ^{1}H NMR (DMSO-d_6, 150 MHz) spectrum of compound **3**.

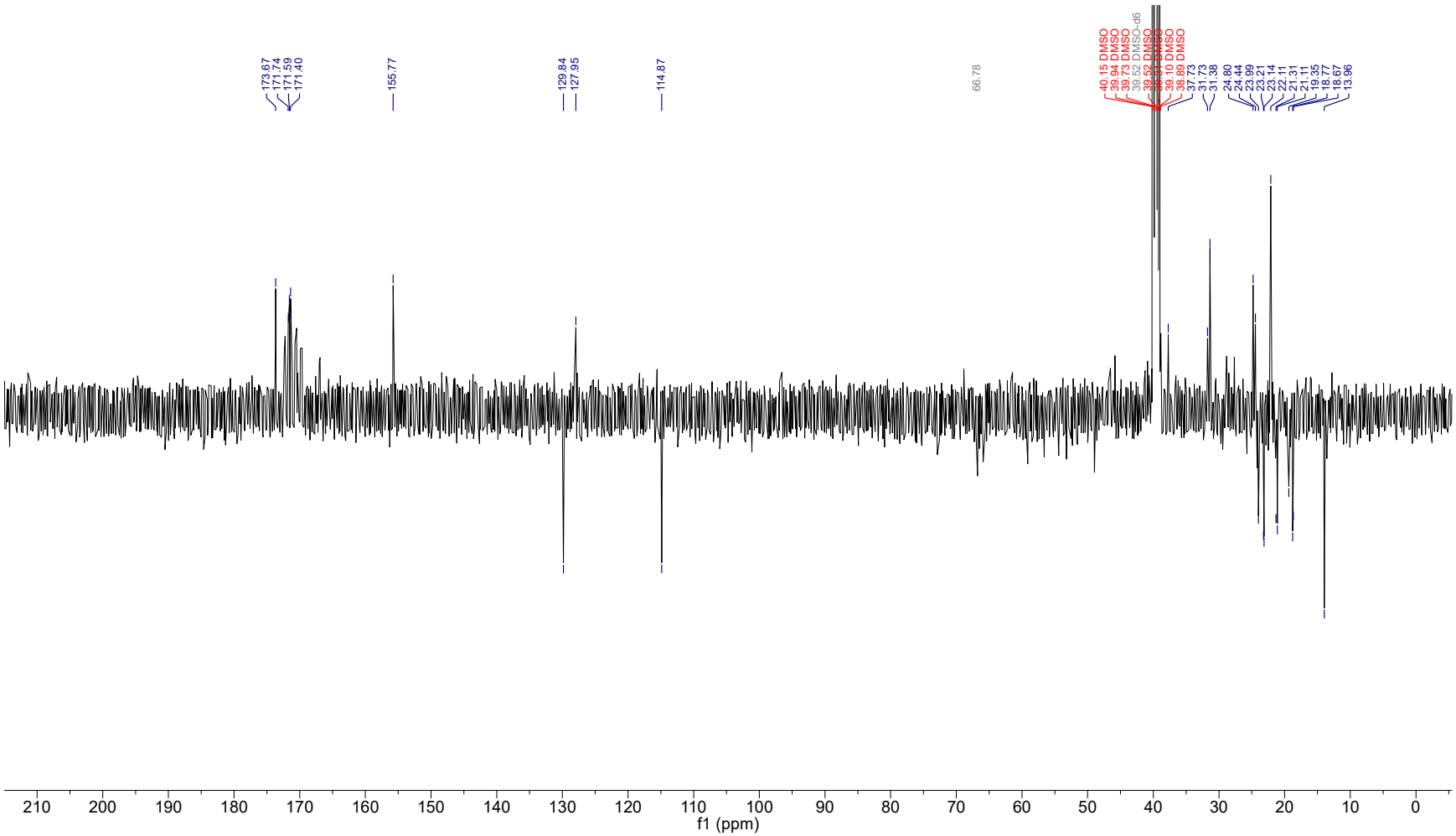


Figure S13 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **3**.

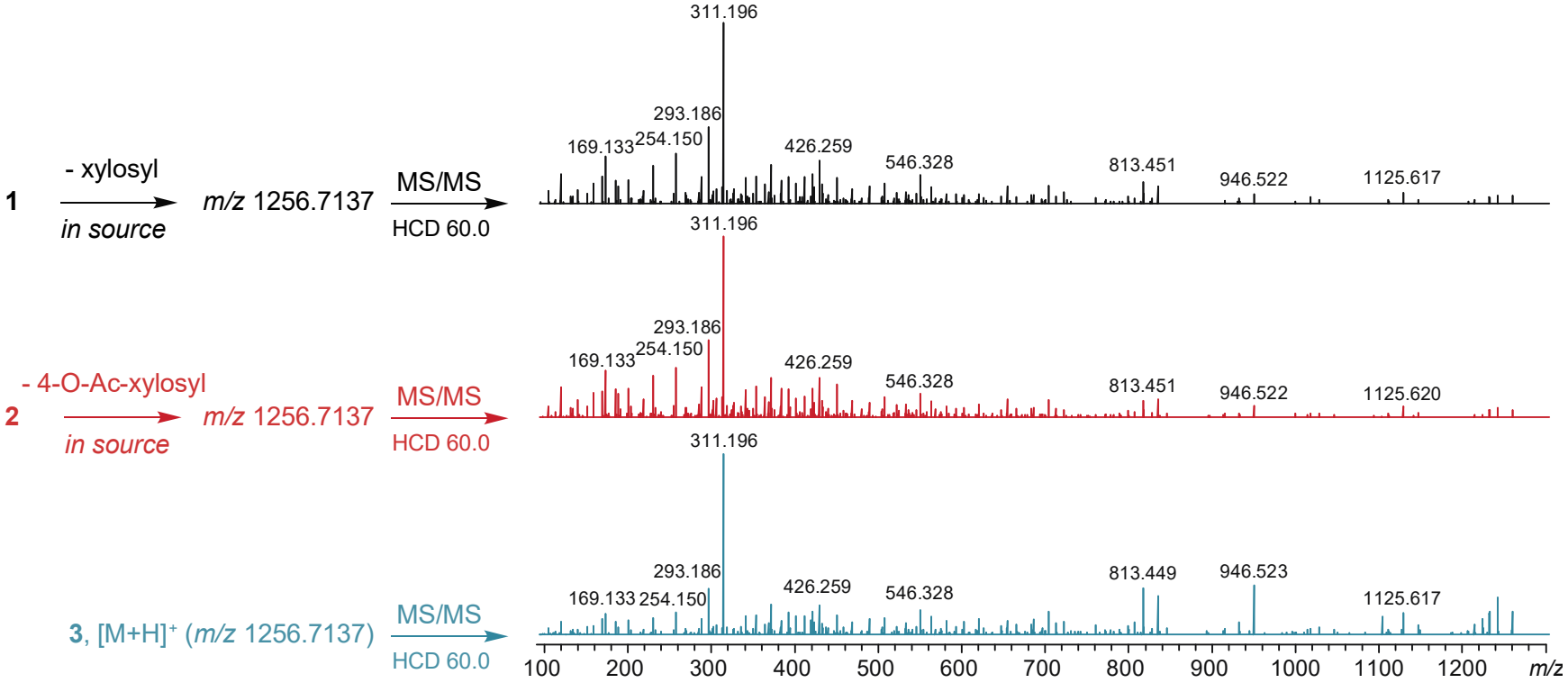


Figure S24 - Comparison of the HRESIMS/MS data for **3**, with those of the in-source-isolated aglycones of **1** and **2**, supporting that these are identical.

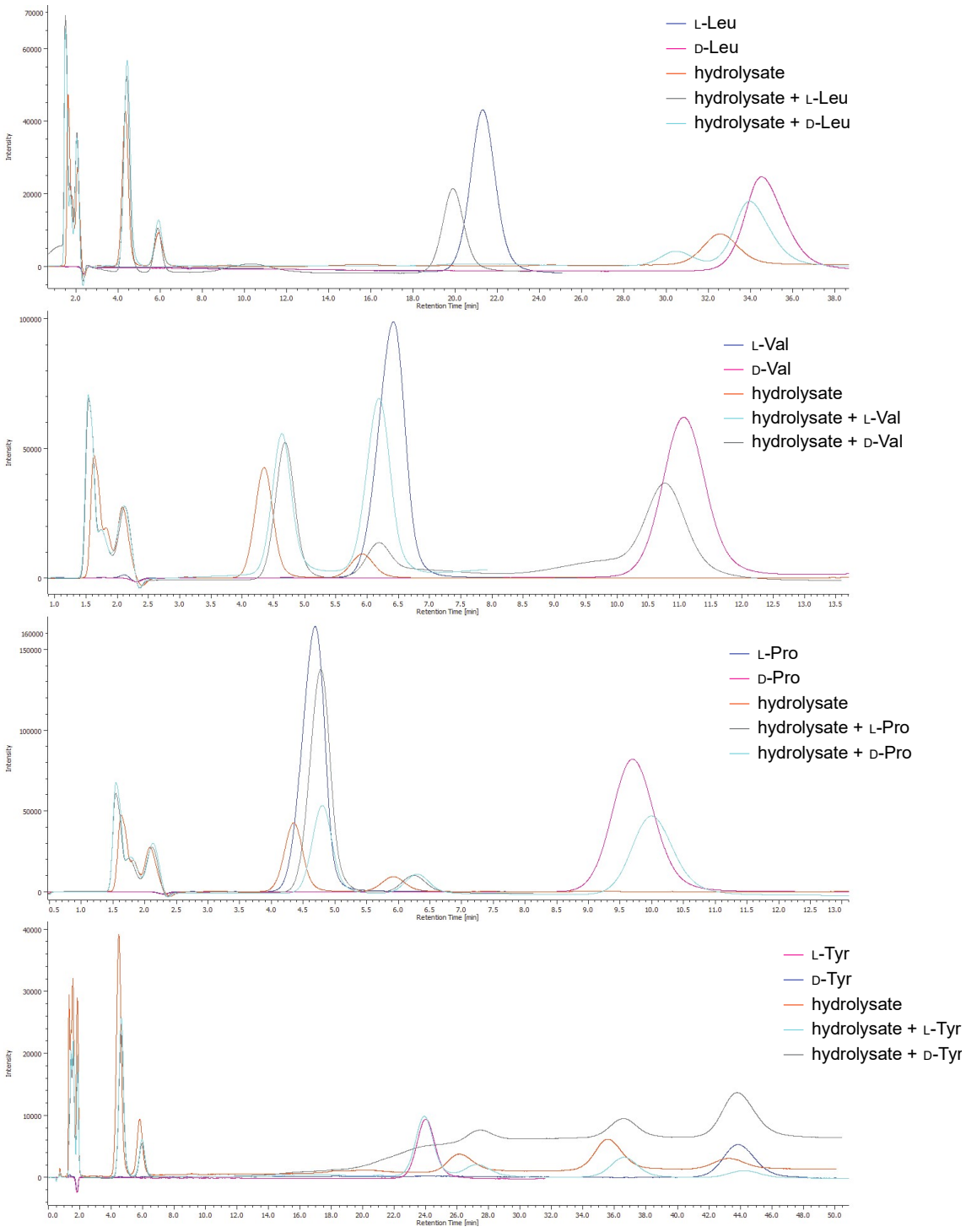


Figure S25 - Chiral-phase HPLC analysis of the amino acid residue composition of **1**.

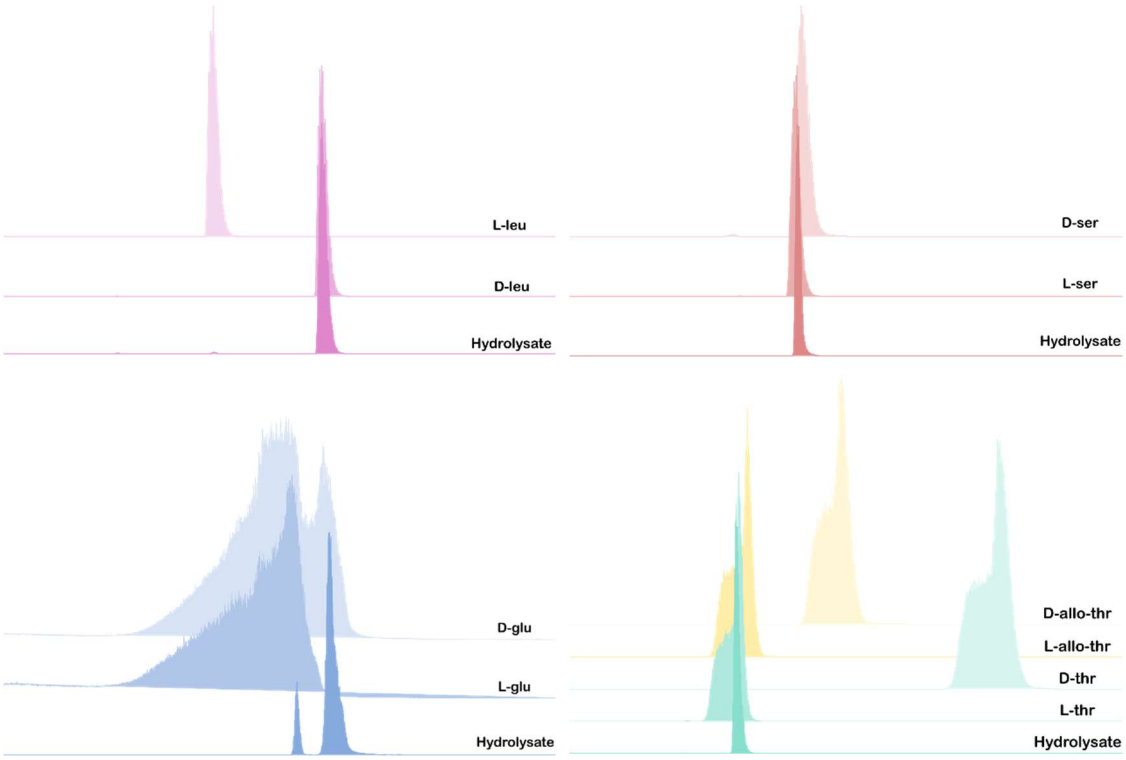


Figure S26 - LC-HRESIMS analysis of the Marfey's derivatization products. EICs of $[M+H]^+$ ions for leucine (*m/z* 384.15), serine (*m/z* 358.09), glutamic acid (*m/z* 400.11), threonine and allo-threonine (*m/z* 372.11) are shown for the standards and hydrolysate.

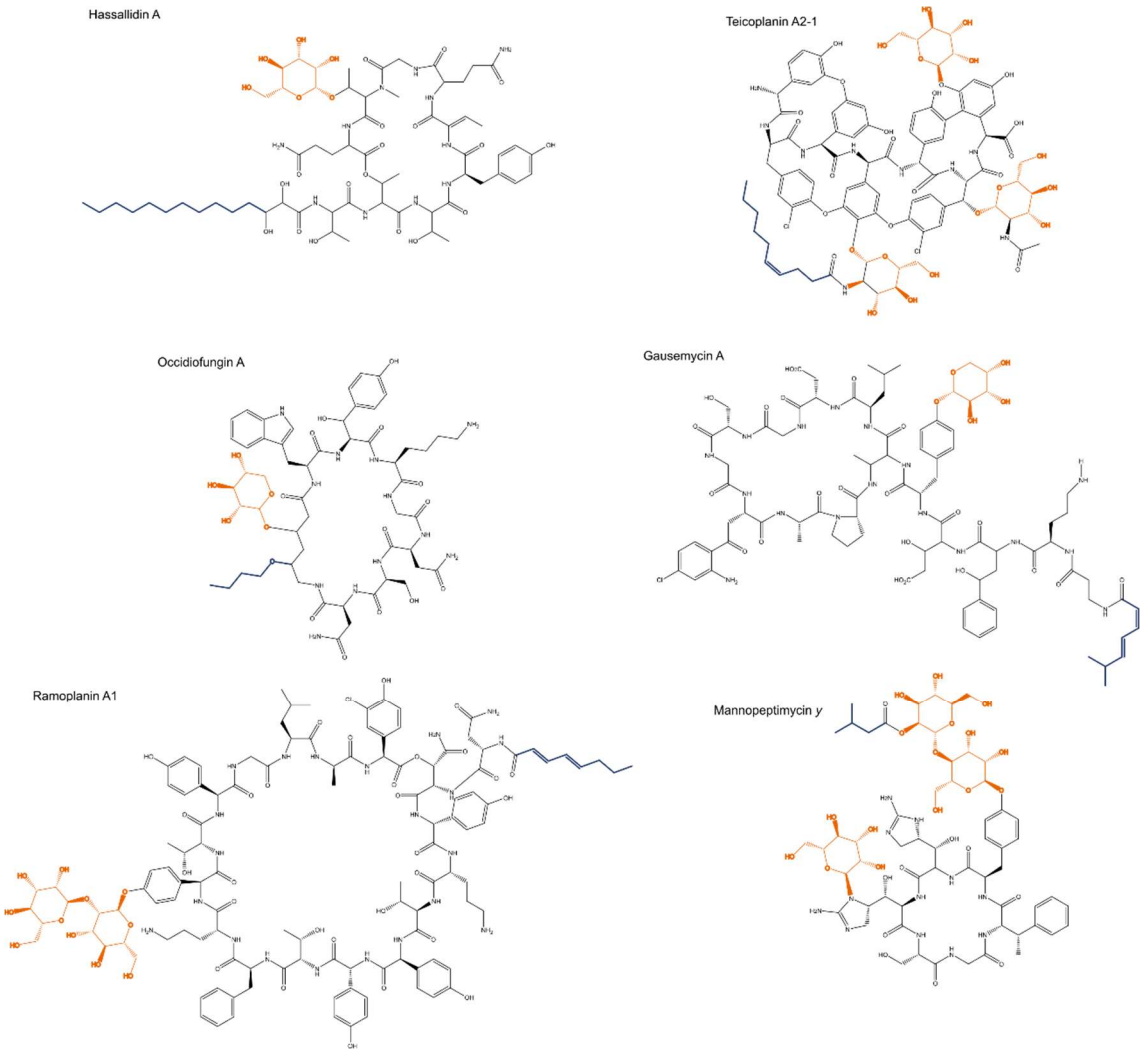


Figure S27 - Examples of natural lipoglycopeptides of different structural classes.

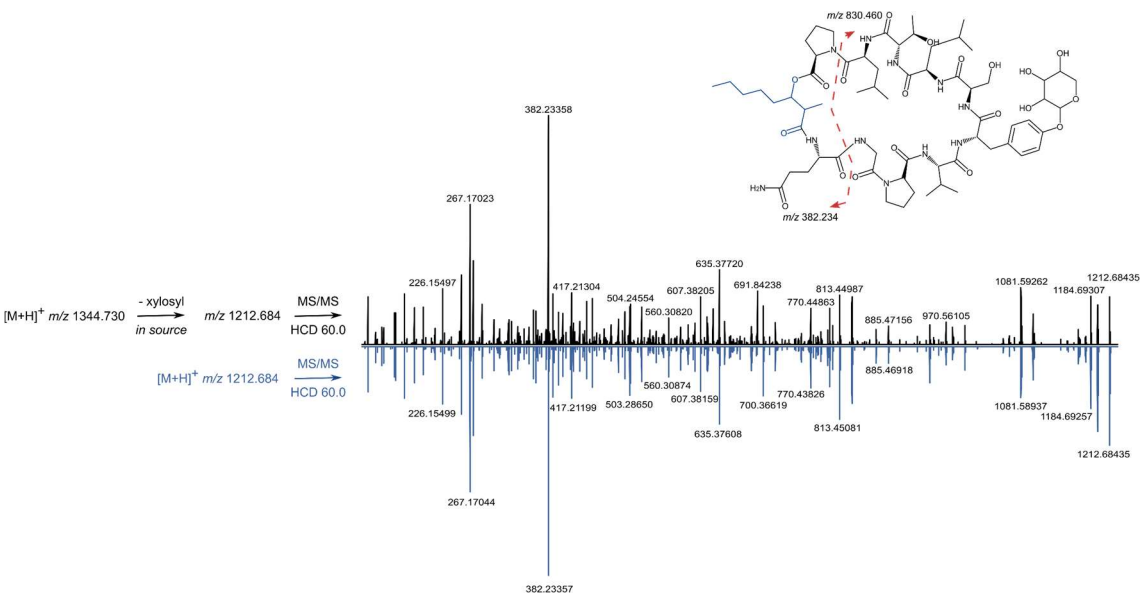


Figure S28 - Comparison of the HRESIMS/MS data for *m/z* 1344.730, with those of the in-source aglycones of *m/z* 1212.684, supporting that these are identical. The structural difference is on the fatty acyl chain, having a hexanoic acid moiety in comparison to Dhmda of **1** and **3**.

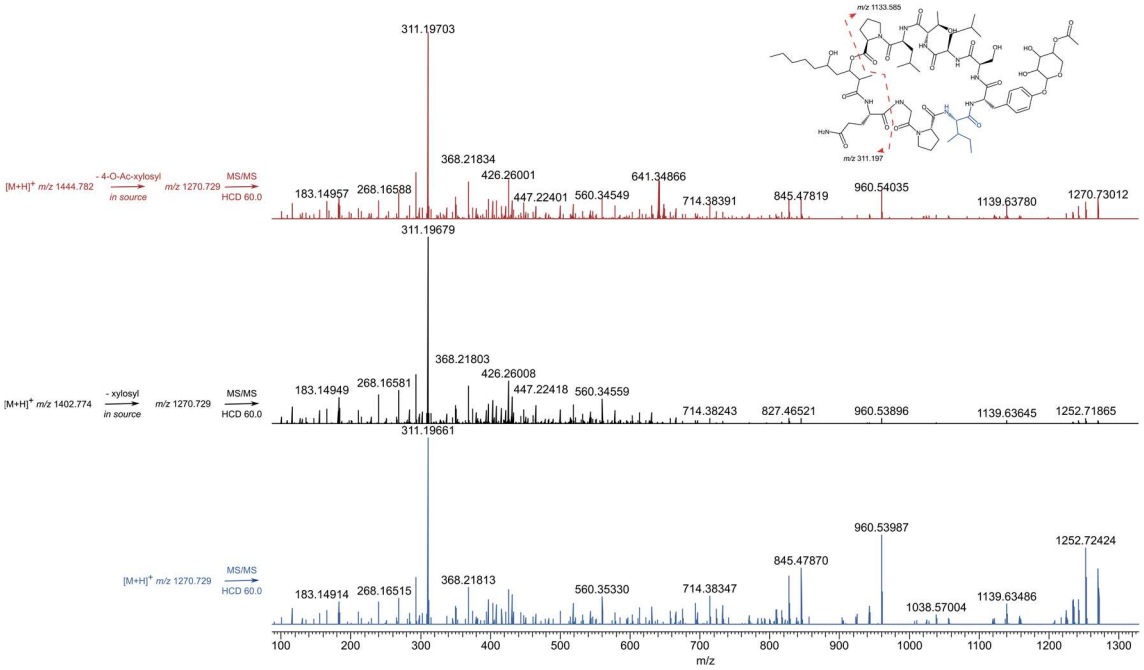


Figure S29 - Comparison of the HRESIMS/MS data for *m/z* 1444.782, with those of the in-source aglycones of *m/z* 1402.774 and *m/z* 1270.729, supporting that these are identical. The structural difference is on amino acid substitution of Val for Leu/Ile (Ile is represented).

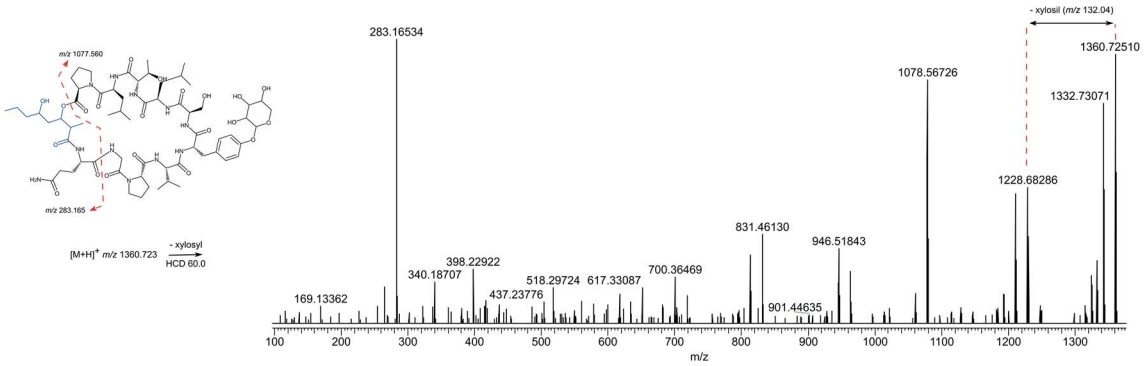


Figure S14 - HRESIMS/MS data for *m/z* 1360.723, showing the structural difference on the fatty acyl chain, having an 3,5-hydroxy-methylbutyric acid instead of the Dhmda of **1**. Fragmentation of the aglycone *m/z* 1228.683 was not possible due to the low intensity of the mass peak.

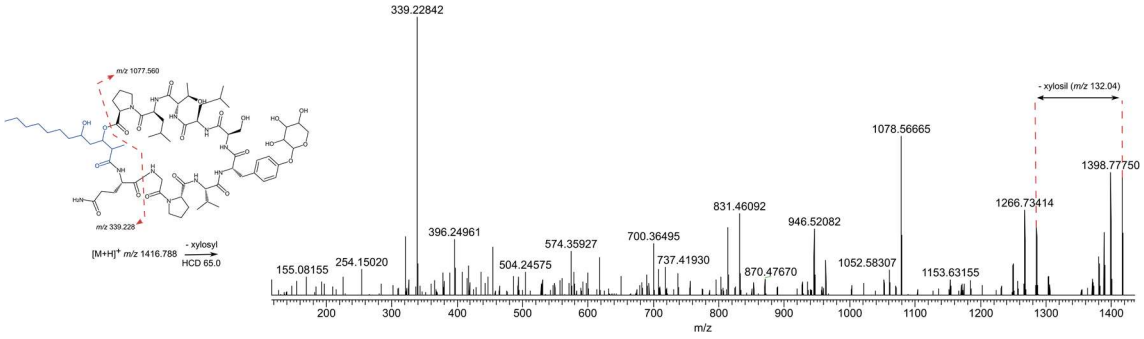


Figure S15 - HRESIMS/MS data for *m/z* 1416.788, showing the structural difference on the fatty acyl chain, having an 3,5-hydroxy-methyloctanoic acid instead of the Dhmda of **1**. Fragmentation of the aglycone *m/z* 1284.742 was not possible due to the low intensity of the mass peak.

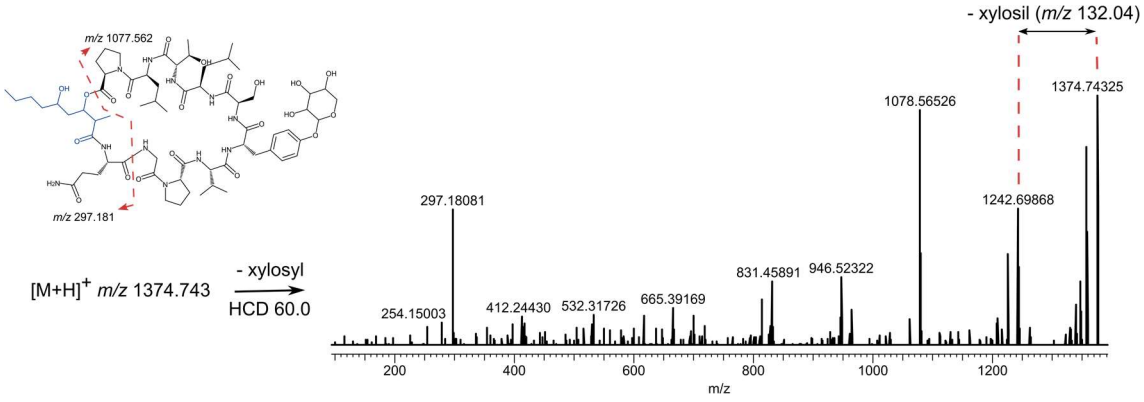


Figure S16 - HRESIMS/MS data for *m/z* 1374.743, showing the structural difference on the fatty acyl chain, having an 3,5-hydroxy-methylheptanoic acid instead of the Dhmda of **1**. Fragmentation of the aglycone *m/z* 1242.699 was not possible due to the low intensity of the mass peak.

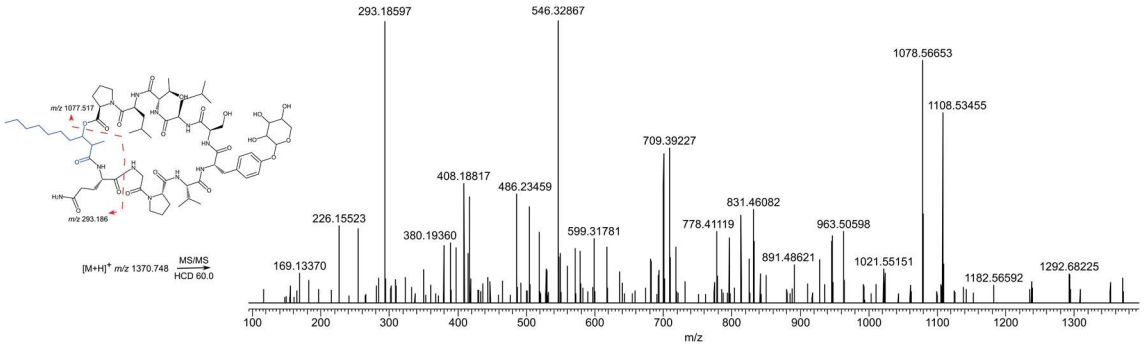


Figure S17 - HRESIMS/MS data for *m/z* 1370.748, showing the structural difference on the fatty acyl chain, without the hydroxyl group on position 5, in comparison with **1**. Fragmentation of the aglycone *m/z* 1238.664 was not possible due to the low intensity of the mass peak.

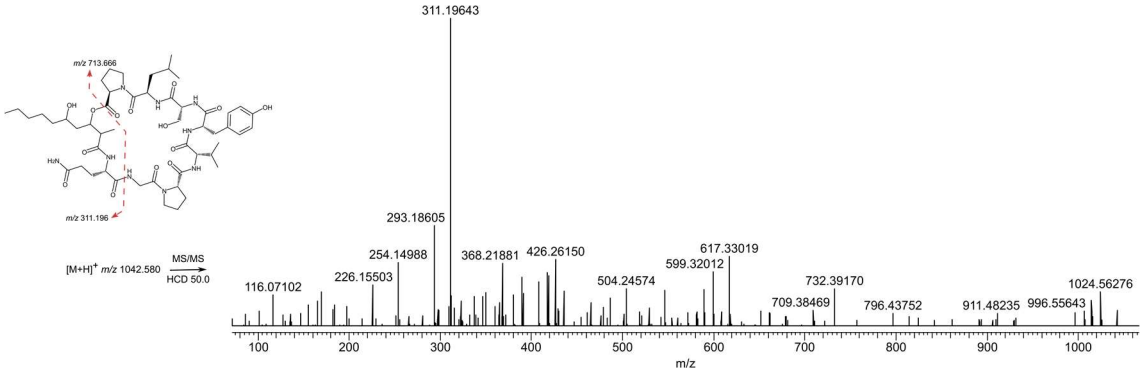


Figure S34 - HRESIMS/MS data for *m/z* 1042.580, showing the structural difference on the core of amino acids, with only seven amino acids, not having one unit of Leu and Thr, in comparison with **3**.

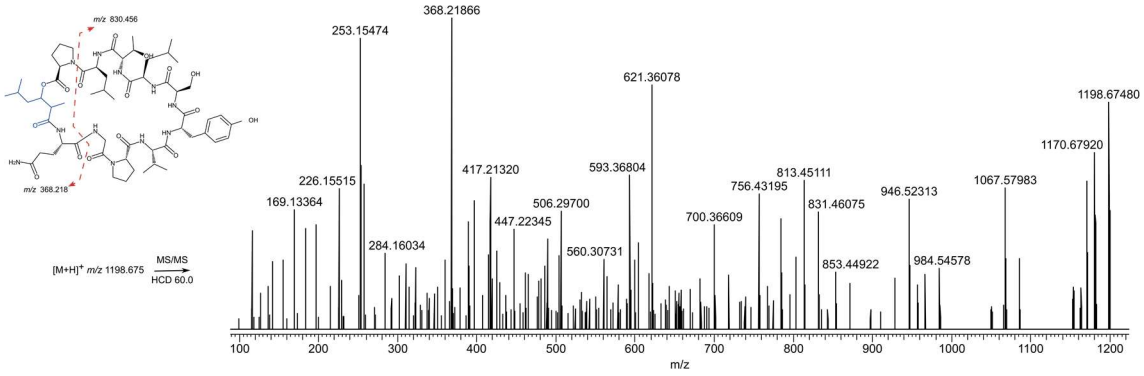


Figure S35 - HRESIMS/MS data for *m/z* 1198.675, showing the structural difference on the fatty acyl chain, having 3-hydroxy-2,5-methylhexanoic acid instead of Dhmda, in comparison with **3**.

Table S5 - Annotation of the *dsm* biosynthetic gene cluster.

Protein	Length (aa)	Accession Number	Locus ID	predicted function	closest homolog (refseq)	organism	identity/ similarity (%)	accession no
-1	210	MCF2150407.1	IQ276_029055	hypothetical protein	hypothetical protein [Anabaena azotica]	Anabaena azotica FACHB-119	86/91	WP_190476212.1
DsmA	384	MCF2150408.1	IQ276_029060	DevB-like, part of ABC transporter	HlyD family efflux transporter periplasmic adaptor subunit [Nostoc sp. FACHB-973]	Nostoc sp. FACHB-973	99/99	WP_190882283.1
DsmB	491	MCF2150409.1	IQ276_029065	DevA-like, part of ABC transporter / cylcic nucleotide binding domain protein	cyclic nucleotide-binding domain-containing protein [Nostoc sp. FACHB-973]	Nostoc sp. FACHB-973	99/99	WP_190882284.1
DsmC	388	MCF2150410.1	IQ276_029075	DevC-like, part of ABC transporter	ABC transporter permease DevC	Nostoc sp. FACHB-973	99/100	WP_190882285.1
DsmD	282	MCF2150411.1	IQ276_029080	GNAT-family acyltransferase	GNAT family N-acetyltransferase [Nostoc sp. FACHB-973]	Nostoc sp. FACHB-973	99/99	WP_190882286.1
DsmE	305	MCF2150412.1	IQ276_029085	hypothetical protein	DUF928 domain-containing protein [Nostocaceae]	Nostoc sp. FACHB-973	100/100	WP_190882287.1
DsmF	90	MCF2150413.1	IQ276_029090	hypothetical protein	hypothetical protein [Nostocaceae]	Nostoc sp. FACHB-973	82/87	WP_190882288.1
DsmG	2587	MCF2150414.1	IQ276_029095	Fatty acid activation (putative Fatty acyl-AMP ligase) / PKS	SDR family NAD(P)-dependent oxidoreductase [Nostoc sp. FACHB-973]	Nostoc sp. FACHB-973	97/98	WP_190882289.1
DsmH	1996	MCF2150415.1	IQ276_029100	PKS	SDR family NAD(P)-dependent oxidoreductase [Nostoc sp. FACHB-973]	Nostoc sp. FACHB-973	99/99	WP_190882290.1
DsmI	6355	MCF2150416.1	IQ276_029105	NRPS	non-ribosomal peptide synthase/polyketide synthase [Scytonema sp. UIC 10036]	Scytonema sp. UIC 10036	69/79	WP_155746086.1
DsmJ	6575	MCF2151708.1	IQ276_029115	NRPS	non-ribosomal peptide synthetase [Scytonema sp. UIC 10036]	Scytonema sp. UIC 10036	71/81	WP_155746081.1
DsmK	675	MCF2150417.1	IQ276_029120	ABC-type transport system	ATP-binding cassette domain-containing protein [Nostoc sp. FACHB-973]	Nostoc sp. FACHB-973	99/99	WP_190881694.1
+1	428	MCF2150418.1	IQ276_029125	transaminase	valine--pyruvate transaminase [Desmonostoc muscorum]	Desmonostoc muscorum CCALA 125	99/99	WP_221707008.1

Table S6 - Results of tBlastn searches for homologs of cacaoidin BGC glycosyltransferases encoded by cao8, cao16 and cao24, against the *D. muscorum* LEGE 12446 genome. Shown are hits with an e-value < 1×10^{-20}. No hits were obtained for cao8 that satisfied this condition.

	% Identical Sites	% Pairwise Identity	E Value	Hit end	Hit start	Query coverage	Sequence Length
Cao24 homolog	**30.50%**	30.50%	4.45E-35	5824034	5823126	73.14%	315
Cao16 homolog1	**29.00%**	29.00%	1.35E-22	5736691	5737326	58.77%	217
Cao16 homolog2	**31.90%**	31.90%	3.46E-21	4687977	4687414	50.97%	191

Table S7 - Biological activity of compounds **1** and **2**. IC_{50} values are in µM.

	Cytotoxicity, IC_{50}							Antibacterial, IC_{50}
compound	HaCaT	A-375	A-549	HCT-116	MDA-MB-231	HepG2	HT-29	*X. campestris*
1	>100	>100	52 ± 1	>100	45 ± 0.3	>100	>100	47 ± 1
2	44 ± 8	44 ± 0.3	43 ± 0.3	57 ± 10	28 ± 2	41 ± 1	42 ± 0.7	34 ± 4
control[a]	19 ± 0.4	9 ± 4	6 ± 0.7	63 ± 9	13 ± 0.7	5 ± 1	2 ± 0.2	2 ± 0.1

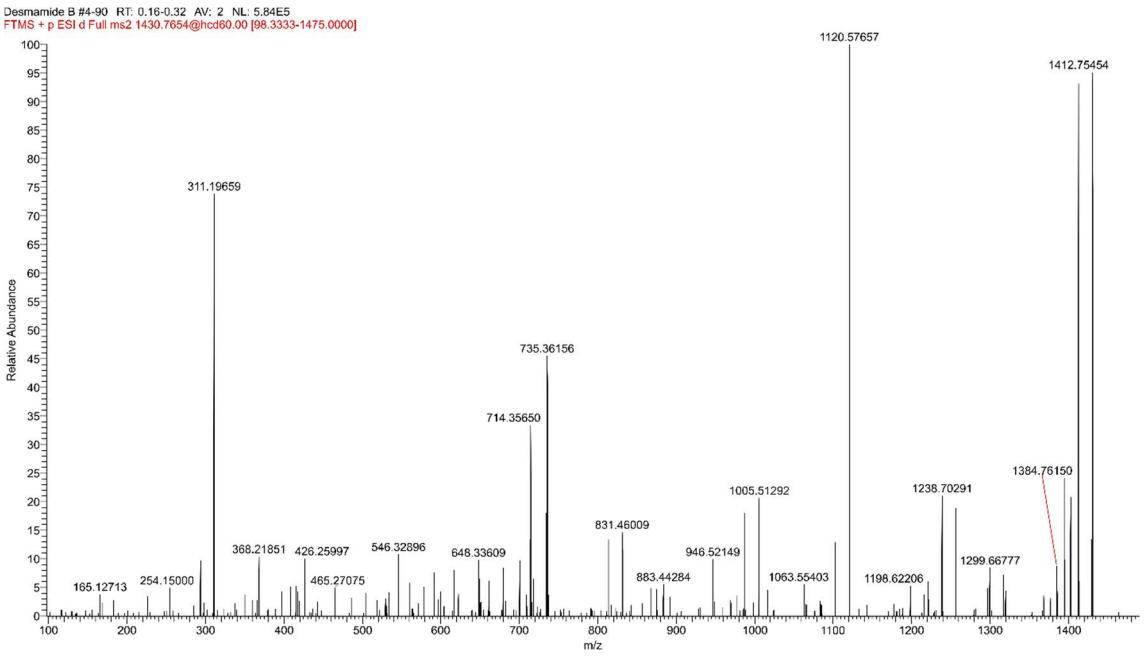


Figure S36 - HRESIMS/MS spectrum of compound **2** ($[M+H]^+$).

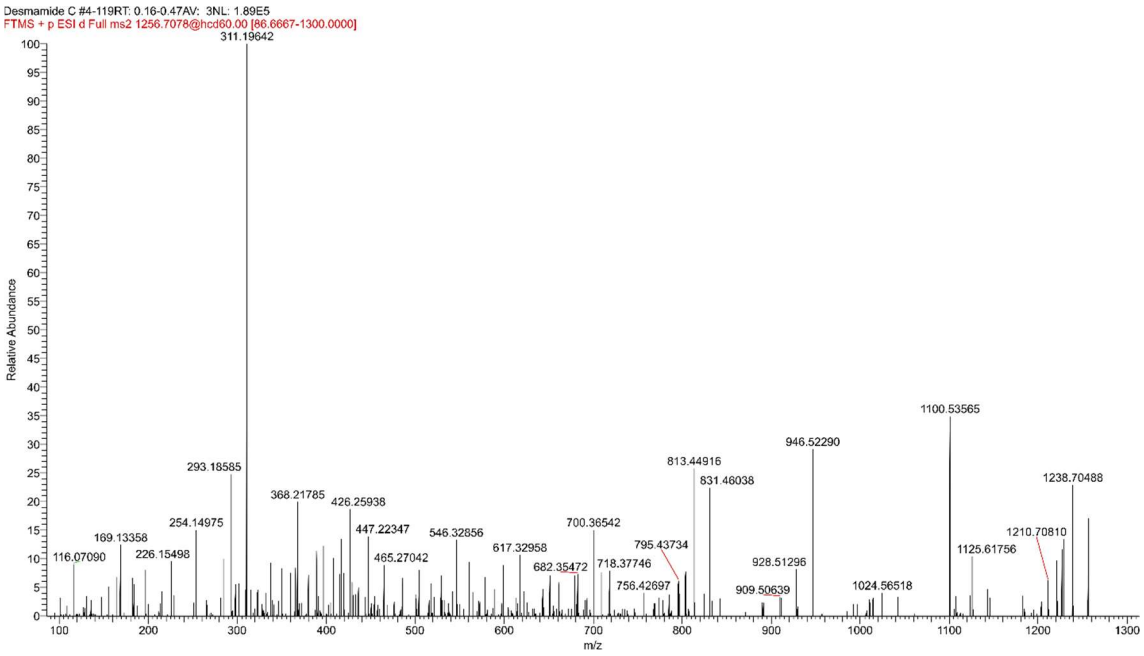


Figure S37 - HRESIMS/MS spectrum of compound **3** ([M+H]$^+$).

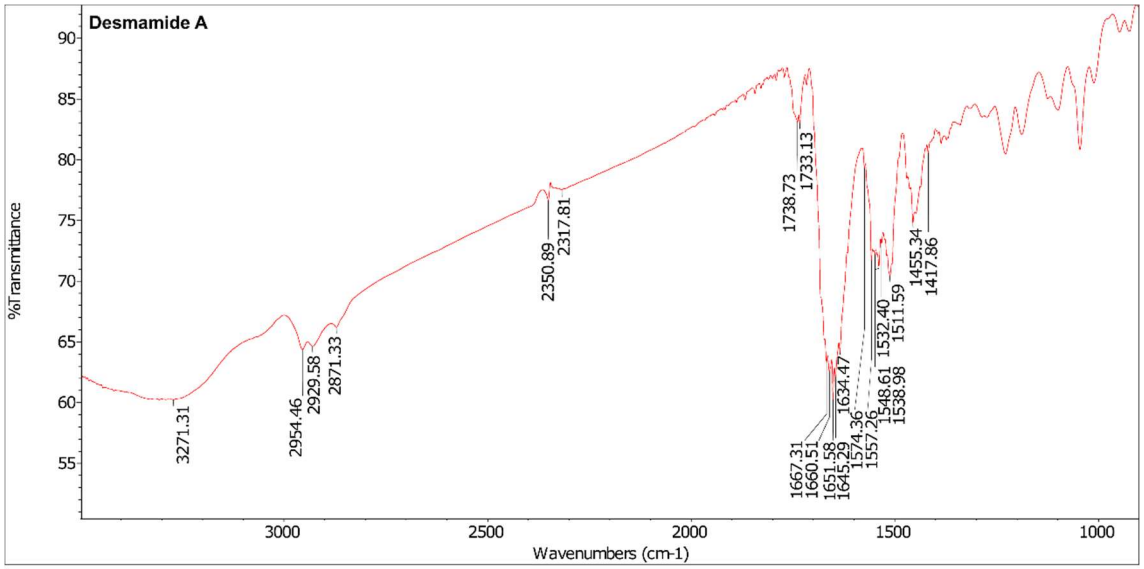


Figure S38 - FT-IR spectrum of compound **1**.

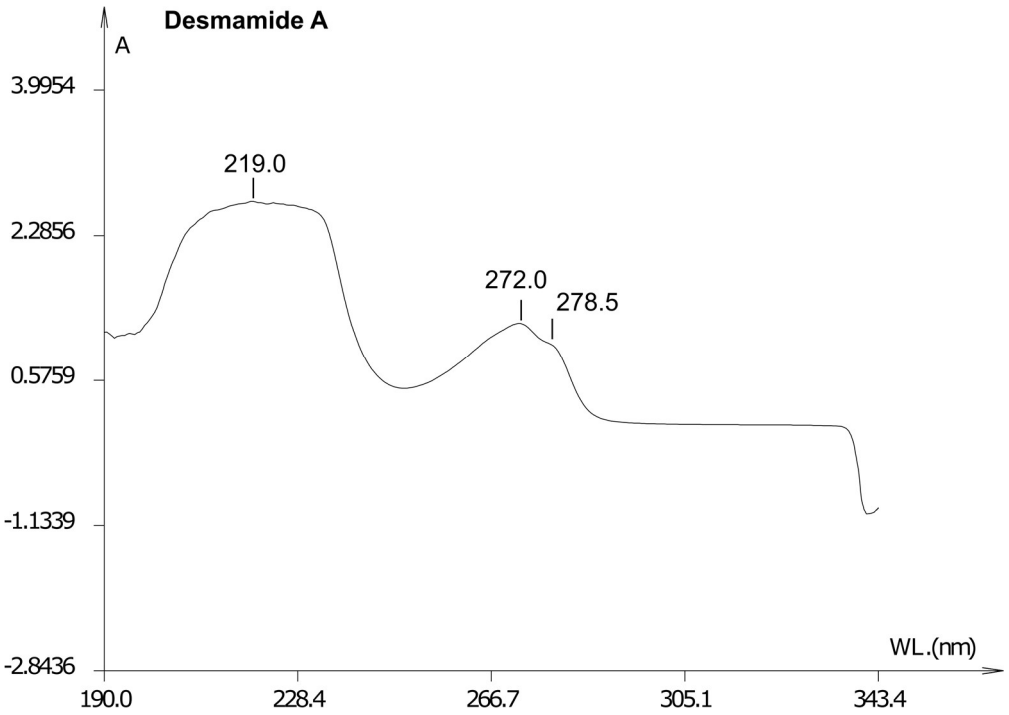


Figure S39 - UV-vis spectrum of compound **1**.

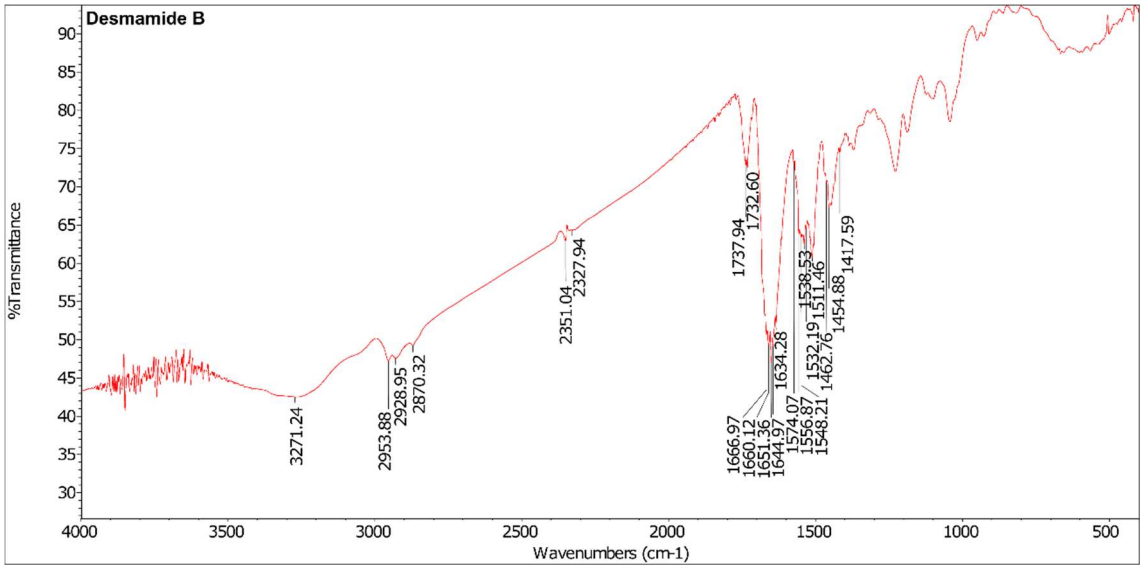


Figure S40 - FI-IR spectrum of compound **2**.

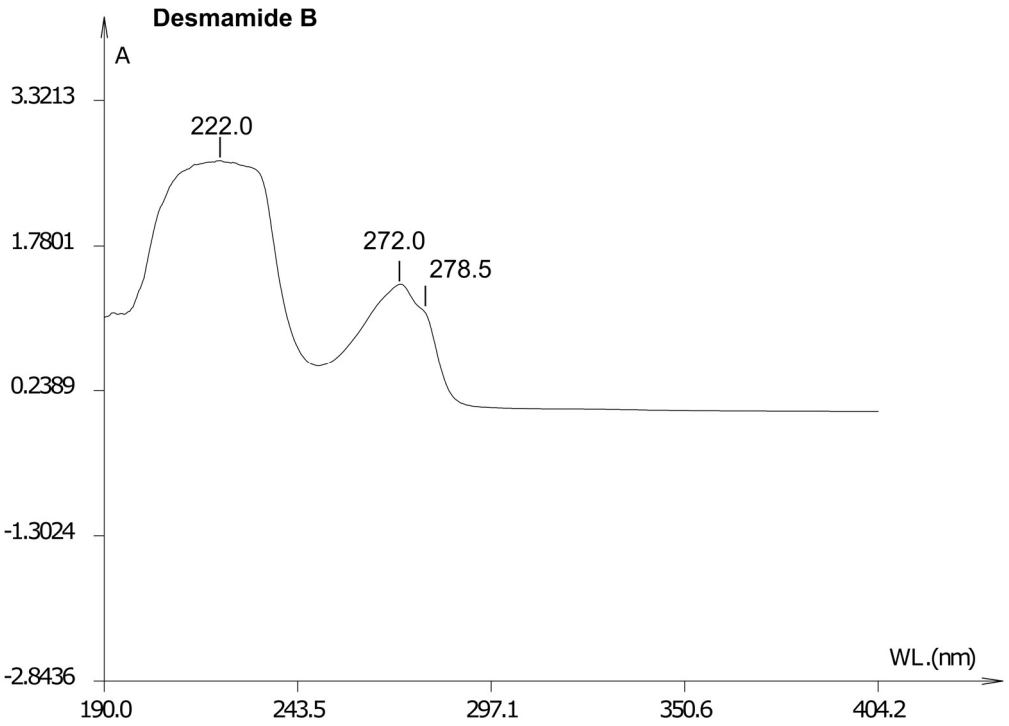


Figure S41 - UV-vis spectrum of compound **2**.

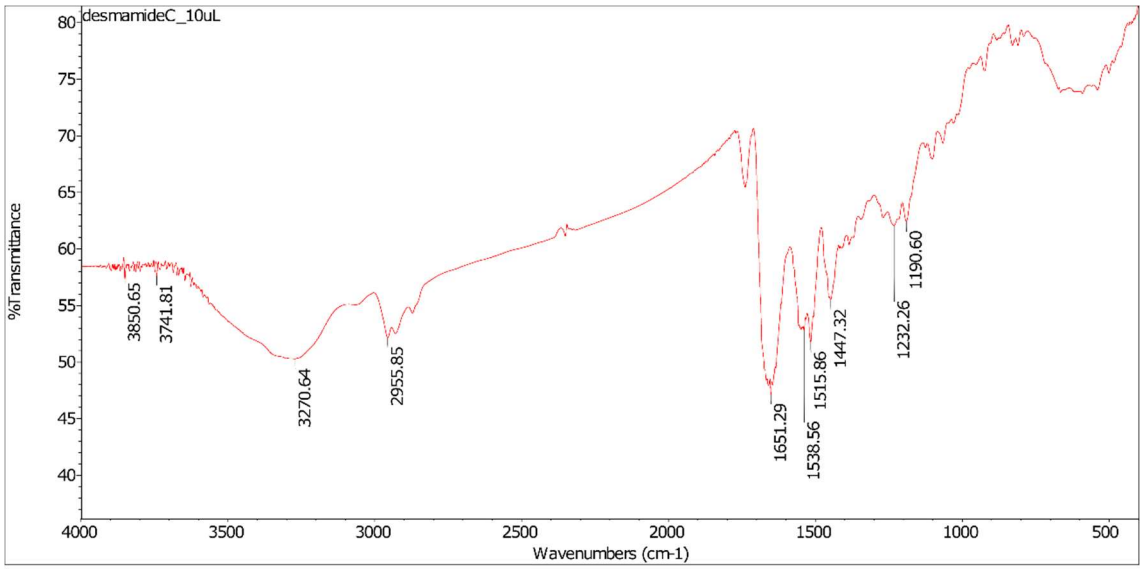


Figure S42 - FT-IR spectrum of compound **3**.

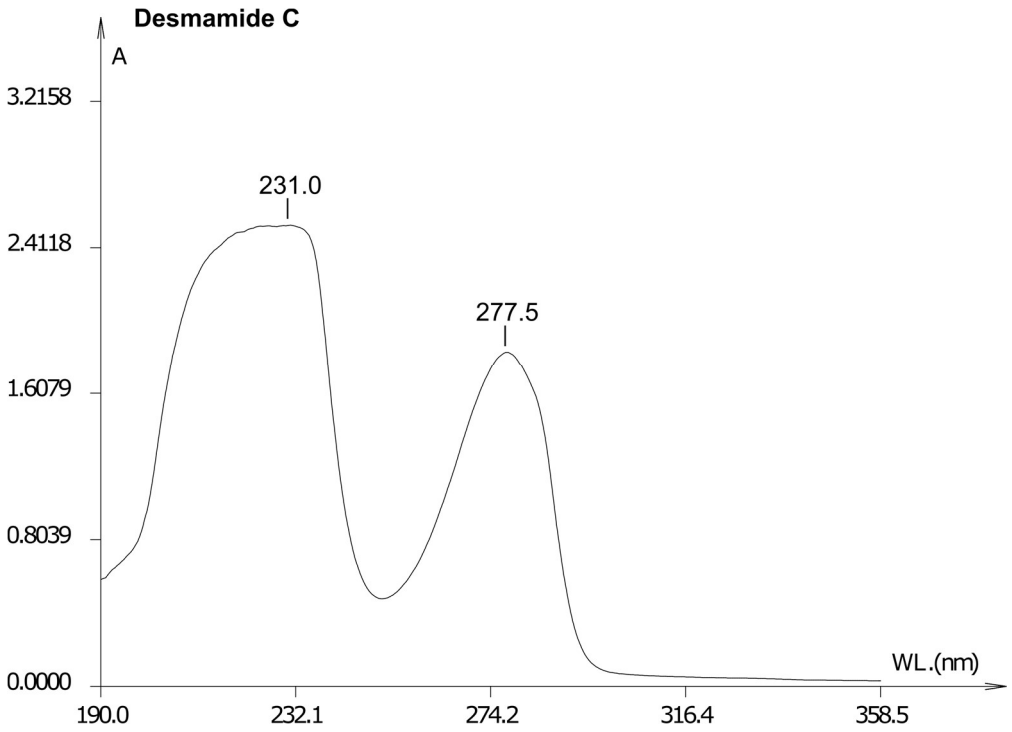


Figure S18 - UV-vis spectrum of compound **3**.

3.6. References

1. Gaysina, L. A.; Saraf, A.; Singh, P., Chapter 1 - Cyanobacteria in Diverse Habitats. In *Cyanobacteria*, Mishra, A. K.; Tiwari, D. N.; Rai, A. N., Eds. Academic Press: 2019; pp 1-28.

2. Haselkorn, R., Heterocysts. *Annual Review of Plant Physiology* **1978,** *29* (1), 319-344.

3. Chang, A. C. G.; Chen, T.; Li, N.; Duan, J., Perspectives on Endosymbiosis in Coralloid Roots: Association of Cycads and Cyanobacteria. *Frontiers in Microbiology* **2019,** *10*.

4. Lindblad, P., Cyanobacteria in Symbiosis with Cycads. In *Prokaryotic Symbionts in Plants*, Pawlowski, K., Ed. Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp 225-233.

5. NCBI genome. (accessed 29 Jan 2022).

6. Popin, R. V.; Alvarenga, D. O.; Castelo-Branco, R.; Fewer, D. P.; Sivonen, K., Mining of Cyanobacterial Genomes Indicates Natural Product Biosynthetic Gene Clusters Located in Conjugative Plasmids. *Frontiers in Microbiology* **2021,** *12*.

7. Shih, P. M.; Wu, D.; Latifi, A.; Axen, S. D.; Fewer, D. P.; Talla, E.; Calteau, A.; Cai, F.; Tandeau de Marsac, N.; Rippka, R.; Herdman, M.; Sivonen, K.; Coursin, T.; Laurent, T.; Goodwin, L.; Nolan, M.; Davenport, K. W.; Han, C. S.; Rubin, E. M.; Eisen, J. A.; Woyke, T.; Gugger, M.; Kerfeld, C. A., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences* **2013,** *110* (3), 1053-1058.

8. Moore, B. S.; Chen, J. L.; Patterson, G. M. L.; Moore, R. E.; Brinen, L. S.; Kato, Y.; Clardy, J., [7.7]Paracyclophanes from blue-green algae. *Journal of the American Chemical Society* **1990,** *112* (10), 4061-4063.

9. Hohlman, R. M.; Sherman, D. H., Recent advances in hapalindole-type cyanobacterial alkaloids: biosynthesis, synthesis, and biological activity. *Natural Product Reports* **2021,** *38* (9), 1567-1588.

10. Schwartz, R. E.; Hirsch, C. F.; Sesin, D. F.; Flor, J. E.; Chartrain, M.; Fromtling, R. E.; Harris, G. H.; Salvatore, M. J.; Liesch, J. M.; Yudin, K., Pharmaceuticals from cultured algae. *Journal of Industrial Microbiology* **1990,** *5* (2-3), 113-123.

11. Proteau, P. J.; Gerwick, W. H.; Garcia-Pichel, F.; Castenholz, R., The structure of scytonemin, an ultraviolet sunscreen pigment from the sheaths of cyanobacteria. *Experientia* **1993,** *49* (9), 825-9.

12. Kaasalainen, U.; Fewer, D. P.; Jokela, J.; Wahlsten, M.; Sivonen, K.; Rikkinen, J., Cyanobacteria produce a high variety of hepatotoxic peptides in lichen symbiosis. *Proceedings of the National Academy of Sciences* **2012,** *109* (15), 5886-5891.

13. Liaimer, A.; Jensen, J. B.; Dittmann, E., A Genetic and Chemical Perspective on Symbiotic Recruitment of Cyanobacteria of the Genus Nostoc into the Host Plant Blasia pusilla L. *Frontiers in Microbiology* **2016,** *7*.

14. Liaimer, A.; Jenke-Kodama, H.; Ishida, K.; Hinrichs, K.; Stangeland, J.; Hertweck, C.; Dittmann, E., A polyketide interferes with cellular differentiation in the symbiotic cyanobacterium Nostoc punctiforme. *Environ Microbiol Rep* **2011,** *3* (5), 550-8.

15. Liaimer, A.; Helfrich, E. J. N.; Hinrichs, K.; Guljamow, A.; Ishida, K.; Hertweck, C.; Dittmann, E., Nostopeptolide plays a governing role during cellular differentiation of the symbiotic cyanobacterium Nostoc punctiforme. *Proceedings of the National Academy of Sciences* **2015,** *112* (6), 1862-1867.

16. Fewer, D. P.; Jokela, J.; Heinilä, L.; Aesoy, R.; Sivonen, K.; Galica, T.; Hrouzek, P.; Herfindal, L., Chemical diversity and cellular effects of antifungal cyclic lipopeptides from cyanobacteria. *Physiol Plant* **2021,** *173* (2), 639-650.

17. Kahne, D.; Leimkuhler, C.; Lu, W.; Walsh, C., Glycopeptide and Lipoglycopeptide Antibiotics. *Chemical Reviews* **2005,** *105* (2), 425-448.

18. Tyurin, A. P.; Alferova, V. A.; Paramonov, A. S.; Shuvalov, M. V.; Kudryakova, G. K.; Rogozhin, E. A.; Zherebker, A. Y.; Brylev, V. A.; Chistov, A. A.; Baranova, A. A.; Biryukov, M. V.; Ivanov, I. A.; Prokhorenko, I. A.; Grammatikova, N. E.; Kravchenko, T. V.; Isakova, E. B.; Mirchink, E. P.; Gladkikh, E. G.; Svirshchevskaya, E. V.; Mardanov, A. V.; Beletsky, A. V.; Kocharovskaya, M. V.; Kulyaeva, V. V.; Shashkov, A. S.; Tsvetkov, D. E.; Nifantiev, N. E.; Apt, A. S.; Majorov, K. B.; Efimova, S. S.; Ravin, N. V.; Nikolaev, E. N.; Ostroumova, O. S.; Katrukha, G. S.; Lapchinskaya, O. A.; Dontsova, O. A.; Terekhov,

S. S.; Osterman, I. A.; Shenkarev, Z. O.; Korshun, V. A., Gausemycins A,B: Cyclic Lipoglycopeptides from Streptomyces sp.*. *Angew Chem Int Ed Engl* **2021,** *60* (34), 18694-18703.

19. Vestola, J.; Shishido, T. K.; Jokela, J.; Fewer, D. P.; Aitio, O.; Permi, P.; Wahlsten, M.; Wang, H.; Rouhiainen, L.; Sivonen, K., Hassallidins, antifungal glycolipopeptides, are widespread among cyanobacteria and are the end-product of a nonribosomal pathway. *Proceedings of the National Academy of Sciences* **2014,** *111* (18), E1909-E1917.

20. Bui, T.-H.; Wray, V.; Nimtz, M.; Fossen, T.; Preisitsch, M.; Schröder, G.; Wende, K.; Heiden, S. E.; Mundt, S., Balticidins A–D, Antifungal Hassallidin-Like Lipopeptides from the Baltic Sea Cyanobacterium Anabaena cylindrica Bio33. *Journal of Natural Products* **2014,** *77* (6), 1287-1296.

21. Somma, S.; Gastaldo, L.; Corti, A., Teicoplanin, a new antibiotic from Actinoplanes teichomyceticus nov. sp. *Antimicrob Agents Chemother* **1984,** *26* (6), 917-23.

22. Singh, M. P.; Petersen, P. J.; Weiss, W. J.; Janso, J. E.; Luckman, S. W.; Lenoy, E. B.; Bradford, P. A.; Testa, R. T.; Greenstein, M., Mannopeptimycins, new cyclic glycopeptide antibiotics produced by Streptomyces hygroscopicus LL-AC98: antibacterial and mechanistic activities. *Antimicrob Agents Chemother* **2003,** *47* (1), 62-9.

23. Ciabatti, R.; Kettenring, J. K.; Winters, G.; Tuan, G.; Zerilli, L.; Cavalleri, B., Ramoplanin (A-16686), a new glycolipodepsipeptide antibiotic. III. Structure elucidation. *J Antibiot (Tokyo)* **1989,** *42* (2), 254-67.

24. Lu, S.-E.; Novak, J.; Austin, F. W.; Gu, G.; Ellis, D.; Kirk, M.; Wilson-Stanford, S.; Tonelli, M.; Smith, L., Occidiofungin, a Unique Antifungal Glycopeptide Produced by a Strain of Burkholderia contaminans. *Biochemistry* **2009,** *48* (35), 8312-8321.

25. Butler, M. S.; Hansford, K. A.; Blaskovich, M. A. T.; Halai, R.; Cooper, M. A., Glycopeptide antibiotics: Back to the future. *The Journal of Antibiotics* **2014,** *67* (9), 631-644.

26. 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. *Nature Biotechnology* **2016,** *34* (8), 828-837.

27. Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L.-F.; Ninomiya, A.; Takada, K.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of peptidic natural products through database search of mass spectra. *Nat Chem Biol* **2017,** *13* (1), 30-37.

28. Gurevich, A.; Mikheenko, A.; Shlemov, A.; Korobeynikov, A.; Mohimani, H.; Pevzner, P. A., Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat Microbiol* **2018,** *3* (3), 319-327.

29. Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L. F.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of microbial metabolites through database search of mass spectra. *Nat Commun* **2018,** *9* (1), 4035.

30. Dictionary of Natural Products 32.0. https://dnp.chemnetbase.com (accessed 20 Dec 2021).

31. van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castaño-Espriu, L.; Chang, C.; Clark, T. N.; Cleary Little, J. L.; Delgadillo, D. A.; Dorrestein, P. C.; Duncan, K. R.; Egan, J. M.; Galey, M. M.; Haeckl, F. P. J.; Hua, A.; Hughes, A. H.; Iskakova, D.; Khadilkar, A.; Lee, J.-H.; Lee, S.; LeGrow, N.; Liu, D. Y.; Macho, J. M.; McCaughey, C. S.; Medema, M. H.; Neupane, R. P.; O'Donnell, T. J.; Paula, J. S.; Sanchez, L. M.; Shaikh, A. F.; Soldatou, S.; Terlouw,

B. R.; Tran, T. A.; Valentine, M.; van der Hooft, J. J. J.; Vo, D. A.; Wang, M.; Wilson, D.; Zink, K. E.; Linington, R. G., The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Central Science* **2019,** *5* (11), 1824-1833.

32. Claridge, T. D.; Pérez-Victoria, I., Enhanced 13C resolution in semi-selective HMbC: a band-selective, constant-time HMBC for complex organic structure elucidation by NMR. *Org Biomol Chem* **2003,** *1* (21), 3632-4.

33. Usvalampi, A. Microbial production of xylitol, L-xylulose and L-xylose Doctoral, Aalto University, 2013.

34. Zhang, Y.; Liu, S.; Liu, H.; Liu, X.; Che, Y., Cycloaspeptides F and G, Cyclic Pentapeptides from a Cordyceps-colonizing isolate of Isaria farinosa. *J Nat Prod* **2009,** *72* (7), 1364-7.

35. Ortiz-López, F. J.; Carretero-Molina, D.; Sánchez-Hidalgo, M.; Martín, J.; González, I.; Román-Hurtado, F.; de la Cruz, M.; García-Fernández, S.; Reyes, F.; Deisinger, J. P.; Müller, A.; Schneider, T.; Genilloud, O., Cacaoidin, First Member of the New Lanthidin RiPP Family. *Angew Chem Int Ed Engl* **2020,** *59* (31), 12654-12658.

36. Dittmann, E.; Gugger, M.; Sivonen, K.; Fewer, D. P., Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol* **2015,** *23* (10), 642-652.

37. Blin, K.; Shaw, S.; Kloosterman, A. M.; Charlop-Powers, Z.; van Wezel, G. P.; Medema, Marnix H.; Weber, T., antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **2021,** *49* (W1), W29-W35.

38. Keatinge-Clay, A. T., A tylosin ketoreductase reveals how chirality is determined in polyketides. *Chem Biol* **2007,** *14* (8), 898-908.

39. Fischbach, M. A.; Walsh, C. T., Assembly-line enzymology for polyketide and nonribosomal Peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev* **2006,** *106* (8), 3468-96.

40. Magarvey, N. A.; Haltli, B.; He, M.; Greenstein, M.; Hucul, J. A., Biosynthetic Pathway for Mannopeptimycins, Lipoglycopeptide Antibiotics Active against Drug-Resistant Gram-Positive Pathogens. *Antimicrobial Agents and Chemotherapy* **2006,** *50* (6), 2167-2177.

41. Favrot, L.; Blanchard, J. S.; Vergnolle, O., Bacterial GCN5-Related N-Acetyltransferases: From Resistance to Regulation. *Biochemistry* **2016,** *55* (7), 989-1002.

42. Daigle, D. M.; Hughes, D. W.; Wright, G. D., Prodigious substrate specificity of AAC(6')-APH(2"), an aminoglycoside antibiotic resistance determinant in enterococci and staphylococci. *Chem Biol* **1999,** *6* (2), 99-110.

43. Hegde, S. S.; Javid-Majd, F.; Blanchard, J. S., Overexpression and mechanistic analysis of chromosomally encoded aminoglycoside 2'-N-acetyltransferase (AAC(2')-Ic) from Mycobacterium tuberculosis. *J Biol Chem* **2001,** *276* (49), 45876-81.

44. Ramos, V.; Morais, J.; Castelo-Branco, R.; Pinheiro, Â.; Martins, J.; Regueiras, A.; Pereira, A. L.; Lopes, V. R.; Frazão, B.; Gomes, D.; Moreira, C.; Costa, M. S.; Brûle, S.; Faustino, S.; Martins, R.; Saker, M.; Osswald, J.; Leão, P. N.; Vasconcelos, V. M., Cyanobacterial diversity held in microbial biological resource centers as a biotechnological asset: the case study of the newly established LEGE culture collection. *Journal of Applied Phycology* **2018,** *30* (3), 1437-1451.

45. Kotai, J., Instructions for preparation of modified nutrient solution Z8 for algae. *Oslo: Norwegian Institute for Water Research* **1972,** *5*.

46. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **2003,** *13* (11), 2498-504.

47. Honda, S.; Akao, E.; Suzuki, S.; Okuda, M.; Kakehi, K.; Nakamura, J., High-performance liquid chromatography of reducing carbohydrates as strongly ultraviolet-absorbing and electrochemically sensitive 1-phenyl-3-methyl-5-pyrazolone derivatives. *Anal Biochem* **1989,** *180* (2), 351-7.

48. Wood, D. E.; Salzberg, S. L., Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* **2014,** *15* (3), R46.

49. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A. A.; Dvorkin, M.; Kulikov, A. S.; Lesin, V. M.; Nikolenko, S. I.; Pham, S.; Prjibelski, A. D.; Pyshkin, A. V.; Sirotkin, A. V.; Vyahhi, N.; Tesler, G.; Alekseyev, M. A.; Pevzner, P. A., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **2012,** *19* (5), 455-77.

50. Wu, Y. W.; Simmons, B. A.; Singer, S. W., MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **2016,** *32* (4), 605-7.

Chapter IV

Secondary metabolite biosynthesis in the endophytic cyanobacterium *Nostoc* sp. LEGE 12450

4. Secondary metabolite biosynthesis in the endophytic cyanobacterium *Nostoc* sp. LEGE 12450

4.1. Introduction

Natural product peptides are one of the secondary metabolite classes that encompasses more chemodiversity and are widely distributed among organisms.[1,2] Their role in microbial ecology has been explored with depth, and so has their pharmaceutical potential, being a source of pharmaceutical agents, *e.g.* vancomycin.[2] With the emergence of genome mining applied to the natural products discovery, the abundance of such peptides became clear due to the widespread presence of non-ribosomal and ribosomal biosynthetic gene clusters (BGCs) in microbial genomes.[3] The enzymes present in these biosynthetic pathways are responsible for peptide assembly and can be classified as non-ribosomal peptide synthetases (NRPSs) or ribosomal synthases, in which the peptide suffers a plethora of post-translational modifications before peptic release (RiPPs).[1] While ribosomal pathways depend on the 20 proteinogenic amino acid residues, NRPSs can use naturally-occurring non-proteinogenic amino acids into their growing oligopeptide chain.[1] This confers NRPs a huge chemical diversity, which is further expanded by tailoring enzymes, which act after the peptide scaffold is formed.[4] Additionally, interfacing of NRPS with polyketide biosynthesis is another source of structural diversification, such is the case of hybrid PKS/NRPS (polyketide synthases) lipopeptide assembly lines. In some cases, these hybrid PKS/NRPS BGCs start with a FAAL (fatty acyl AMP-ligase) enzyme that incorporates a fatty acid that can be further elongated by the PKS before passing to the NRPS for peptide incorporation.[5,6]

Cyanobacteria are a group of photosynthetic Gram-negative bacteria, rich in natural products, and widespread throughout the globe.[7] A phylogenomic analysis of this phylum revealed a high occurrence of hybrid NRPS/PKS biosynthetic pathways, especially in late-branching organisms as Nostocales.[8,9] The majority of these BGCs were not linked to a known metabolite, highlighting the potential for new (lipo)peptide discovery from this group of microorganisms.[10] The high numbers of BGCs within the order Nostocales are accommodated in their large genomes and likely provides them with an advantage for symbiotic interactions.[9,11] Besides their free-living form, Nostocalean cyanobacteria can establish facultative symbiotic relationships with a wide range of organisms, with the exception of the *Nostoc*-Azollae obligate symbiont.[11] The main reason for cyanobacteria

to enter in symbiosis, is their ability to perform nitrogen fixation in specialized cells, termed heterocysts.[12] The multicellularity of Nostocales strains is well recognized not only by the formation of heterocysts, but also, spore forming akinetes and motile filaments hormogonia, essential for symbiosis.[12] The complex lifecycle of Nostocalean cyanobacteria have been the focus of many studies, using the facultative symbiont model organism *Nostoc punctiforme* PCC 73102, in an attempt to understand the genomic and metabolomic orchestration responsible for its lifecycle regulation.[13-16] Nostopeptolides, belonging to the lipopeptide class, were identified as key players on this regulation, having a crucial role the symbiotic crosstalk with host plants, functioning as hormogonia repressors or activators, depending on the ecological environment.[14] Nostopeptolides are composed of a fatty acyl moiety connected to an heptapeptide core, with an exocyclic amino acid, and these metabolites can be also classified as depsipeptides due to the peptide macrocyclization as a macrolactone.[17] Recent discoveries regarding cyanobacteria metabolic regulation were achieved using this model organism, in which the production of specific polyketide metabolites improved the expression of peptidic secondary metabolite BGCs, namely nostopeptolide, anabaenopeptin, microviridin and lanthipeptide BGCs present in the genome of *N. punctiforme* PCC 73102.[18] Terrestrial symbiotic cyanobacteria, can be characterized as a source of new and diverse natural product peptides, as demonstrated by the diversity found in *N. punctiforme* PCC 73102.[18,19] Adding to the mentioned peptides, several others were isolated from symbiotic cyanobacteria[19], such as cryptophycins[20], nostocyclopeptides[21], aeruginosins[22] and the glycosidic peptides suomilide and banyasides[23,24]. All these non-ribosomal peptides represent unique chemical scaffolds: the cryptophycin depsipeptides can show chlorination on the tyrosine residue and an epoxide on the phenyl-octanoic acid[25]; nostocyclopeptides have the rare non-proteinogenic amino acid methyl-proline and the macrocycle is closed by forming an imine bond[26]; the linear peptides aeruginosins can be halogenated, sulphated and/or hydroxylated, besides the presence of the unique scaffolds hydroxy-phenyl lactic acid (Hpla), amino acid 2-carboxy-6-hydroxyoctahydroindole (Choi) and arginine derivatives (*e.g.* argininol)[27]; and suomilide and banayside peptides can be sulfated, glycosylated and have unique scaffolds as the azabiocyclononane (Abn) and 1-amidino-3-(2-aminoethyl)-3-pyrroline (Aeap)[23,24]. Some of these peptides were first reported in non-symbiotic strains, *e.g.* suomilide isolated from *Nodularia spumigena*[23] and aeruginosins isolated from *Microcystis aeruginosa* NIES-298[22]; however, a study focusing on *Nostoc* strains isolated from the host plant *Blasia pusilla* L. and surrounding soil samples, demonstrated the widespread distribution of these peptides in the *Nostoc* lineage, including symbiotic cyanobacterial strans.[19] A

common feature of cyanobacterial peptides, reported for the aforementioned metabolites, is the production of several variants of the same peptide, a phenomenon that can be explained by the promiscuity of the NRPS adenylation domains that can recognize multiple, structurally similar amino acids residues.[1] In terms of bioactivity, all the mentioned natural products show bioactive potential: nostopeptolides have showed antitoxin bioactivities[28]; anabaenopeptins, aeruginosins and suomilide/banyaside variants have potent inhibitory activity of serine proteases[29,30]; nostocyclopeptides were recently described to have inhibitory activity of the 20S proteosome[31] and cryptophycins have potent anticancer activity[20].

The chemical exploration of the nostocalean cyanobacterium *Nostoc* sp. LEGE 12450, a plant facultative symbiont isolated from a cycad (*Encephalartos horridus*) coralloid root, was performed based on metabolomic analysis of a pre-fractionated organic extract. For three metabolites, dereplication efforts retrieved no hits in natural products databases, therefore their isolation was undertaken, leading to the purification of nostopeptolide L5 (**1**), suomilide 1077 (**2**) and suomilide 1007 (**3**), analogues of the already known nostopeptolides L1-L4 and suomilide, respectively. The biosynthesis of **1** and **2**-**3** was investigated in the genome of *Nostoc* sp. LEGE 12450, by comparing genes found in this strain to the described biosynthetic gene clusters for these two families of compounds. Regarding bioactivity, only compound **1** showed anti-fouling activity, with an EC_{50} of 19.4 μM against the settlement of *Mytilus galoprovencialis* larvae.

4.2. Results and discussion

The initial screening to select the most promising cyanobacterial LEGEcc strains, reported in chapter II, retrieved a hit regarding another cycad symbiont, the cyanobacterium *Nostoc* sp. LEGE 12540. Molecular networking analysis using Global Natural Products Social Molecular Networking (GNPS)[32] showed the presence of four clusters specific to this cyanobacterium. The methanolic fraction of *Nostoc* sp. LEGE 12450 contained four mass features highlighted on those clusters, corresponding to one major metabolite of $[M+H]^+$ *m/z* 1100.530, one minor metabolite of $[M+H]^+$ *m/z* 824.439 and two additional main metabolites that shared a similar fragmentation pattern $[M+H]^+$ *m/z* 1007.423 and $[M+H]^+$ *m/z* 1077.466 (Figure S1, Table S1). Interestingly, these last two metabolites shared a neutral loss of -79.95 amu, possibly related to a sulfate or phosphate group. Dereplication using Dereplicator[33], Dereplicator VarQuest[34] and Dereplicator+[35], as manual searches in the Dictionary of Natural Products (DNP)[36] and Natural Product Atlas (NPAtlas)[37] resulted in no hits, confirming the novelty of the

compounds. However, a search of the ions present on the cluster of $[M+H]^+$ *m/z* 824.439 (Figure S1) using DNP and NPAtlas, led to the annotation of two anabaenopeptins, specifically Lyngbyaureidamide A ($[M+H]^+$ *m/z* 856.456) and B ($[M+H]^+$ *m/z* 842.441).[38] Since the fragmentation pattern of all three mass features was similar (Figure S2), a thorough search of the anabaenopeptin compound family resulted in the identification of the minor metabolite of *Nostoc* sp. LEGE 12450 as anabaenopeptin MM823, isolated from a *Microcystis* sp. bloom.[39] The isolation process proceeded with the three masses from the GNPS molecular networking analysis, namely $[M+H]^+$ *m/z* 1100.530, $[M+H]^+$ *m/z* 1077.466 and $[M+H]^+$ *m/z* 1007.423.

4.2.1. MS-guided isolation of Nostopeptolide L5 (**1**) & Suomilides 1077 and 1007 (**2**,**3**)

A xenic unicyanobacterial culture of *Nostoc* sp. LEGE 12450 was performed in large-scale (160 L), yielding 60 g of dry biomass that was extracted with CH_2Cl_2/MeOH (2:1, v/v) by repeated percolation. The resulting organic extract (10.4 g) was analyzed by Liquid Chromatography coupled to High Resolution Electrospray Ionization Mass Spectrometry (LC-HRESIMS), which confirmed the presence of the mass features of interest, therefore enabling a MS-guided isolation approach. After extraction, a normal-phase Vacuum Liquid Chromatography (VLC) was performed to obtain 13 fractions of increasing polarity. The subsequent LC-HRESIMS analysis showed that the compounds were abundant in fractions eluting from 3:2 to 1:9 EtOAc:MeOH (v/v). The combined fractions (3.7 g) were further purified using reversed-phase VLC, semipreparative and analytical reverse-phase HPLC (Figure S3). LC-HRESIMS was used throughout the isolation process to check the presence of the mass features. The process led to the isolation of nostopeptolide L5 (**1**, 19.0 mg, *m/z* 1100.530 for $[M+H]^+$), suomilide 1077 (**2**, 4.8 mg, *m/z* 1077.466 for $[M+H]^+$) and suomilide 1007 (**3**, 3.3 mg, *m/z* 1007.423 for $[M+H]^+$). Since these compounds were present in adjacent chromatographic fractions during the purification process, a rough quantification using LC-HRESIMS (Figure S4, Table S2) indicated that the combined mass of these three metabolites reaches ~57 mg in the initial extract (approx. 1% of dry cell material).

4.2.2. Structure elucidation of **1**

Structure elucidation was for performed first on the most abundant metabolite, **1**. High Resolution Electrospray Ionization Mass Spectrometry (HRESIMS) analysis of **1** showed

a *m/z* 1100.5286 for the $[M+H]^+$ ion (Figure S5), that was in accordance with the molecular formula $C_{55}H_{73}N_9O_{15}$, showing 24 degrees of unsaturation. Inspection of the ^1H and ^{13}C APT NMR data (400 MHz, DMSO-d_6) of **1**, shown in table 1 (Figure S6, S7), indicated the presence of a peptidic structure since multiple resonances consistent with exchangeable NH protons (δ_H 8.5-7.50), amide (δ_C 172-168) and alpha carbons (δ_C 60-50) were found. The presence of an aliphatic terminal methyl group triplet (CH_3-7, δ_C 13.8, δ_H 0.88), two methylene signals (δ_H 1.38; 2.06) and carbonyl group (δ_C 167.9) indicated the presence of an acyl chain, suggesting the lipopeptide nature of **1**. Two-dimensional NMR analysis (HSQC, HMBC, COSY and NOESY) allow the assignment of eight amino acids and one short chain fatty acid moiety (Figure 1, Figures S8-S12). On the acyl chain it was observed the presence of a quaternary alkene (δ_C 130.3) with an HMBC correlation to a methyl group (δ_C 12.5, δ_H 1.69), which in turn showed an HMBC correlation to the carbonyl group (δ_C 167.9). Both HMBC and COSY correlations of the methylene signals with the methyl group and methine proton H-3 (δ_H 6.18) allowed to build the 2-methylhex-2-enoic acid (MHEA) moiety. A search on NPAtlas using MHEA as query, indicated that the compounds nostopeptolides L1-L4 contained this scaffold. Nostopeptolides L1-L4 are depsipeptides isolated from *Nostoc* sp. UK2aImI and structurally are composed of MHEA and an eight amino acid core (Figure S13), in which five are non-proteinogenic amino acid residues (*O*meSer, PhePr, Dhb, Hyp and 4mePro).[28] The assignment of the 1D and 2D NMR data of **1**, indicated the presence of *O*meSer, PhePr (phenylalanyl-propionic acid), Dhb (dehydrobutyrine) and Hyp (4-hydroxy-proline), together with threonine (Thr), proline (Pro), tyrosine (Tyr) and Gln (glutamine), as shown in Figure 1A. Briefly, the Thr could be connected to MHEA by an HMBC correlation to C-1 (δ_C 167.9), and to *O*meSer by a NOE correlation of H-8 (δ_H 4.68) to the amino group (δ_H 8.70); PhePr shows correlation of its alpha carbon (H-16 - δ_H 4.74) to the carbonyl ketone (δ_C 201.9) and is connected to Pro by an NOE correlation of H-25 to H-29 (CH_2 – δ_H 3.88/3.46); Dhb shows an HMBC correlation to carbonyl C-32 (Pro, δ_C 171.4) and a NOE correlation of its NH proton (δ_H 9.28) to NH group of Tyr (δ_H 6.30); Gln is connected to Tyr through HMBC correlation to C-45 (δ_C 169.9), and to Hyp by a NOE correlation of H-46 (δ_H 4.24) to H-51 (CH_2, δ_H 3.54); lastly, HMBC correlations from H-54 (δ_H 4.11) and H-9 (δ_H 4.79) to C-55 (δ_C 170.9) form the macrocycle ring closure, through an ester bond, to give the full structure of **1** (Figure 1A).

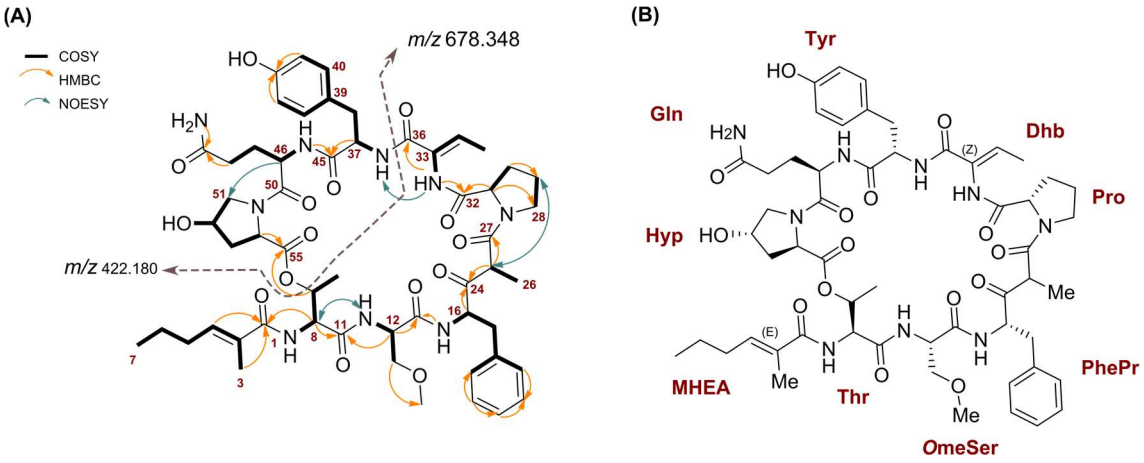


Figure 1 - Structure elucidation of **1**. (A) Key HMBC, COSY and NOESY correlations and HRMS/MS fragmentations supporting the structure proposal for **1**; (B) Compound **1** structure, evidencing the resolved stereochemistry.

The increased resolution of the amide δ_C resonance by using band-selective constant-time HMBC experiments, together with the comparison to Nostopeptolide L1 published NMR data[28] (Figure S12, Table S3), firmly established the connectivity between the amino acid residues in **1**. HRESIMS/MS analysis also confirmed the structure of **1** (Figure S14), by showing fragmentations consistent with the residue sequences [Hyp-Gln-Tyr-Dhb-Pro-PhePr-*O*meSer] and [MHEA-Thr-*O*meSer-PhePr-(Pro-Dhb)-Tyr-Gln]. The absolute configurations of the amino acid residues in **1**, were determined using Marfey's analysis (Figure S15), in which Gln was found to be in D-configuration, while others were in L-configuration, as represented in Figure 1B. Hydroxy-proline was found to be in *trans*-L-configuration (Figure 1B). For some amino acid residues, it was visible a minor conformer on the hydrolysate of **1**, specifically for Gln, Phe, Thr, and Hyp, that can correspond to a conformer with different stereochemistry (Figure S15). The double bond of Dhb was found to be in a (*Z*) configuration, while the double bond in the MHEA is in a (*E*) configuration (Figure 1B).

Table 1 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for nostopeptolide L5 (**1**) in DMSO-d_6.

unit	no.	δC, type	δH[a] (*J* in Hz)	HMBC[b]	COSY	NOESY
MHEA	1	167.9, C				
	2	130.3, C				
	3	12.5, CH_3	1.69, d (1.2)	1, 2, 4, 5, 6	4, 5	
	4	135.8, CH	6.18, dt (7.3, 1.2)	1, 3, 5, 6	3, 5	6
	5	29.8, CH_2	2.06, m	1, 2, 4, 6, 7	4, 6, 7	
	6	21.6, CH_2	1.38, m	4, 5, 7	5, 7	

	7	13.8, CH_3	0.88, t (7.3)	5, 6	4, 5, 6	
Thr	8	56.6, CH	4.68, m	1, 9, 10, 11	9, 10	NH-*O*meSer
	9	70.3, CH	4.79, m	8, 10, 55	8, 10	
	10	16.1, CH_3	0.73, d (6.3)	8, 9	9	8, NH-*O*meSer
	11	168.5, C				
	OH					
	NH		7.88, d (8.3)	11	8	
OmeSer	12	51.6, CH	4.61, dt (8.6, 4.2)	11, 13, 15	13a, 13b, NH-*O*meSer	NH-PhePr
	13	73.0, CH_2	3.50, dd (9.5)	12, 14, 15	12, 13b	NH-*O*meSer, NH-PhePr
			3.80, dd (9.5)	14, 15	12, 13a	NH-PhePr
	14	58.4, CH_3	3.28, s	13		
	15	170.2, C				
	NH		8.70, br		12	8, 10, 13a, 13b
PhePr	16	55.9*, CH	4.74, m	15, 17, 18, 24	17a, 17b, NH-Phe	25
	17	33.1, CH_2	2.82, dd (10.2, 14.8)	16, 18, 19/23, 24	16, 17b, 19/23	NH-Phe
			3.09, dd (3.8, 14.8)	16, 18, 19/23, 24	16, 17a, 19/23	NH-Phe
	18	138.0, C				
	19/23	128.2, CH	7.21, m	17, 18, 20/22, 21	17a,b	
	20/22	128.7, CH	7.23, m	18, 19/23, 21		
	21	126.2, CH	7.17, m	18, 19/23, 20/22		
	24	201.9, C				
	25	45.5*, CH	4.03, d (6.8)	24, 26, 27	26	16, 29ab, NH-PhePr
	26	11.1, CH_3	1.17, d (6.9)	24, 25, 27	25	29b
	27	170.8, C				
	NH		8.83, d (8.9)	15, 16	16	12, 13ab, 17ab, 25
Pro	28	24.8, CH_2	2.04, m		29ab, 30a	30b
			1.98, m		28a, 29ab, 30a	
	29	47.3, CH_2	3.88, m	28	28ab, 29b	25, 26, NH-Dhb
			3.46, m		28ab, 29a	25, 26, NH-Dhb
	30	29.7, CH_2	2.24, m	32	28a, 30b, 31	NH-Dhb
			1.83, m	28, 29, 32	30a, 31	NH-Dhb
	31	60.8, CH	4.09, m	28, 29, 30, 32	30ab	
	32	171.4, C				
Dhb	33	162.7, C				
	34	124.9*, CH	4.83, m		35	NH-Dhb
	35	13, CH_3	1.43, d (7.2)	32, 33, 34, 36	34, NH-Dhb	
	36	130.3, C				
	NH		9.28, s	32, 33, 36	34, 35	28a, 29ab, 31, NH-Tyr
Tyr	37	52.4, CH	4.69, m	33, 38, 39, 45	38, NH-Tyr	NH-Gln
	38	38.2, CH_2	2.71, m	45	37	40/14, NH-Tyr, NH-Gln
	39	126.8, C				
	40/44	130.2, CH	6.92, d (8.5)	38, 40/44, 41/43, 42	38, 41/43	
	41/43	114.9, CH	6.61, d (8.5)	39, 40/44, 41/43, 42	40/44	38
	42	155.7, C				
	45	169.9, C				
	OH		9.16, s			OH-Hyp, 41/43, 40/44, NH-Gln
	NH		6.30, s		37	NH-Dhb, 38

Gln	46	50.0, CH	4.24, m	45	47ab, NH-Gln	48, 51
	47	25.8, CH_2	1.50, m	48, 49	46, 47b, 48	NH-Gln
			1.76, m		46, 47a	NH-Gln, 51
	48	30.4, CH_2	1.97, m	46, 47, 49	47ab	40/44, NH-Gln
	49	173.3, C				
	50	170.7				
	NHa		6.80, s	49	Gln-NHb	
	NHb		6.98, s	49	Gln-NHa	
	NH		8.35, d (8.0)	45, 46, 47	46	37, 38, 47ab, 48
Hyp	51	54.2, CH_2	3.54, m	52, 53	52	47b, NH-Gln
	52	68.9, CH	4.28, m		51, 53b, OH-Hyp	
	53	36.4, CH_2	1.90, m	52, 54, 55	53b, 52, 54	
			1.73, m	52, 54, 55	52, 54	53a
	54	58.8, CH	4.11, s	51, 52, 53, 55	53ab	
	55	170.9, C				
	OH		5.13, d (2.6)		52	OH-Tyr

[a]from HSQC, [b]from proton to indicated carbon.

The nostopeptolide compound family is composed of nostopeptolides A1-A3[17], nostopeptolide 1052[16] and nostopeptolides L1-L4[28]. The A1-A3 compounds were discovered in *Nostoc* sp. GSV224 and are formed by a butyric acid-derived fatty acyl moiety connected to a core of eight amino acids [Ile-Ser-4mePro-LeuAc-Leu-Gly-Asn-Tyr-Pro] (Figure S13).[17] The ring formation occurs between the Pro C-terminus and the hydroxy group of Ser, having Ile as an exocyclic amino acid connected to the fatty acyl chain. Nostopeptolide 1052 was discovered in *Nostoc punctiforme* PCC 73102, together with nostopeptolide A1, showing only differences in the acyl chain, having an acetic acid-derived unit, instead of butyric acid.[16] Nostopeptolides L1-L4, discovered from *Nostoc* sp. UK2aImI are composed of an MHEA acyl chain connected to an eight amino acid core [Thr-OmeSer-PhePr-4mePro-Dhb-Ile-Gln-Hyp].[28] The macrocycle is formed by an ester bond between the C-terminus Hyp and the hydroxy group of Thr. Since **1** shows only differences in two amino acid residues, when compared to nostopeptolides L1-L4, this compound was named nostopeptolide L5.

An interesting feature of this compound family is the number of non-proteinogenic amino acids, mainly the L1-L5 versions, when compared to A1-A3/1052. The characteristic 4mePro, also found in spumigins[40], and nostocyclopeptides[21], was not found in the structure of **1**. On the other hand, *O*meSer, Dhb, and Hyp were shared between **1** and nostopeptolides L1-L4 but were not present in the nostopeptolides A1-A3. Hydroxy-proline is a non-proteinogenic amino acid found in some families of cyanobacterial natural products, mainly from the marine environment, such as

microcolins[41], laxaphycins[42], majusculamide D[43] and scytocyclamide A[44]. Still, cyanobacteria from other habitats, also synthesize natural products containing Hyp, such as the nostoweipeptins[28] and the heinamides[45]. The *O*meSer is a rare non-proteinogenic amino acid, found only on coibamide A[46] besides nostopeptolides L1-L5. Di-methylated versions of Ser (*N*,*O*-dimeSer) are found in three cyanobacterial compounds, all from marine origin (lyngbyastatin 2, norlyngbyastatin 2 and malevamide C)[47,48]. Previous authors[28] hypothesized that *O*-methylation of this amino acid residue could prevent the macrocyclic ring closure on its hydroxyl group, as occurs on nostopeptolides A1-A3[17]. Depsipeptides are usually cyclized by an ester bond formation with *α*-hydroxy groups of serine or threonine, presenting exocyclic amino acid residues connected to the fatty acyl chain, as is the case of nostopeptolides A1-A3/1052.[17,28] However, in nostopeptolides L1-L5 both amino acids are present, so *O*-methylation of the Ser residue, ensured the macrocyclization on the Thr residue, not having an exocyclic amino acid connected to MHEA. As Liu and co-workers[28] hypothesized, this could be a new mechanism of directing peptide macrocyclization to a specific amino acid.

4.2.3. Structure elucidation of **2** and **3**

The initial inspection of compounds **2** and **3**, based on LC-HRESIMS/MS data acquired for GNPS molecular networking analysis, indicated a presence of a possible phosphate or sulfate group, due to a neutral loss of 79.95 amu in Full MS spectra (Figure S16). Calculating the loss of phosphate (79.967 amu) or sulphate groups (79.957 amu), and analyzing the HRESIMS spectra of the pure compounds, confirmed the presence of a sulfate group as a part of the structures and showed a $[M+H]^+$ *m/z* 1077.4657 for compound **2** and $[M+H]^+$ *m/z* 1007.4233 for compound **3**. The molecular formula $C_{45}H_{72}O_{20}N_8S$ matched the protonated mass of **2**, with 14 degrees of unsaturation, while $C_{41}H_{66}O_{19}N_8S$ corresponded to the protonated mass of **3**, with 13 degrees of unsaturation. NMR efforts were applied to the structure elucidation of **2,** since it was more abundant. Analysis of the ^1H and ^{13}C APT NMR spectra (400 MHz, DMSO-d_6), present on Table 2 (Figure S17, S18), revealed the presence of an anomeric proton (H-1', δ_C 94.7, δ_H 4.97 $J_{1'\text{-}2'}$ 3.7 Hz), as well as the oxymethines corresponding to a sugar unit. Additionally, two-dimensional NMR analysis (HSQC, HBMC, COSY) showed the presence of two aliphatic terminal methyl groups (H-2'd, δ_C 13.4, δ_H 0.88 and H-4'f, δ_C 13.8, δ_H 0.86) connected to a small methylene envelope, suggesting the presence of two fatty acyl moieties (Table 2, Figures S19-S21). Detailed analysis of these resonances indicated that these fatty acyl chains were substituents of the glycosyl hydroxy groups,

due to the HMBC correlations of H2' (δ_H 4.84, dd 11, 3.7 Hz) and H4' (δ_H 5.34, dd 3.4, 1.0 Hz) to the carbon esters C-2'a (δ_C 172.5) and C-4'a (δ_C 172.0), respectively. Resonances consistent with exchangeable NH protons were observed, as well as amide carbons, indicating the presence of amino acid residues in the structure of **2**. Still, only isoleucine (Ile) could be assigned based on the spectral data. Due to the complexity of the carbon connectivity of **2**, a search on NPAtlas was conducted using as query the sulfate group and the glycosyl moiety. The sulfate group is a common feature in several classes of cyanobacterial secondary metabolites, such as aeruginosins[22], cyanopeptolins[49] and micropeptins[39]. However, only some compounds bearing a sulfate group were glycosylated, namely aeruginosins 205B/828A[50,51], banyasides A-B[24], suomilide[23] and varlaxins 1046A-1022A[52]. Both aeruginosins were chlorinated and not in agreement with the isotopic pattern of **2** and **3**, and varlaxins have a glucose unit decorated with two 4-hydroxyphenylacetic acid units (Hpaa), not being in agreement with the NMR data of **2** (Figure S22). On the other hand, both banyasides A-B and suomilide have glucose linked to one or two fatty acid moieties, respectively. By comparing the NMR data of **2** with banaysides/suomilide, it was evident that the compounds isolated from *Nostoc* sp. LEGE 12450 were suomilide analogues and were named suomilide 1077 (**2**) and suomilide 1007 (**3**) (Figure S22).

These metabolites are composed of a 2-*O*-methylglyceric acid 3-*O*-sulfate (MgS), an azabiocyclononane moiety (Abn), a 1-amidino-3-(2-aminoethyl)-3-pyrroline (Aaep) and a glycosyl unit decorated with two different length fatty acids at positions 2 and 4, and a carbamoyl unit at position 3. The structure, represented in Figure 2A, shows the connectivity of the MgS moiety to Ile by the HMBC correlation to the carbonyl C-4 (δ_C 169.8) by both H-2 (δ_H 3.93) and H-5 (δ_H 4.61) (Figure S20). The Abn unit is connected to Ile through the HMBC correlation of H-11 (δ_H 3.93) to the C-terminus carboxyl group (C-10, δ_C 171.9) (Figure 2A, Figure S20). The main resonances of the Abn unit, confirming its unique structure are: HMBC correlation of H-11 (δ_H 3.93) to C-12 (δ_C 53.4); COSY correlations from H-12 to H-16; HMBC correlations to the quaternary carbon C-17 (δ_c 80.5); and the COSY and HMBC correlation of the NH group to H-19 (δ_H 4.22) and C-18 (δ_c 156.6) confirming the urethane linkage (Figure 2A, Figure S20-S21). The Aaep structure is connected through COSY correlations from the vinylic proton H-24 (δ_H 5.64) to the H-27 (δ_H 4.13) and to the amino group (δ_H 7.57) (Figure 2A, Figure S21). Aaep and Abn are connected through an HMBC correlation to the carbonyl C-20 (δ_c 168.8) by the Aaep amino group and by H-11 (δ_H 3.93) (Figure 2A, Figure S20). The guanidine carbon was assigned based on comparison to the NMR data of suomilide since no correlations

was observed for **2** (Table S4). The proton H-15 of the Abn moiety (δ_H 3.73) showed an HMBC correlation to the anomeric carbon C-1' (δ_H 94.7), connecting the glycosyl unit to the Abn (Figure 2A, Figure S20). Lastly, a butyric acid (C4) chain formed an ester bond to the second hydroxy group, while hexanoic acid (C6) is connected to position four, confirmed by HMBC correlations to the carbonyls from H-2' (δ_H 4.84) and H-4' (δ_H 5.34), respectively (Figure 2A, Figure S20). On the third hydroxy group, H-3' shows an HMBC correlation to the carbamoyl moiety, finalizing the structure of **2** (Figure 2A, Figure S20). HRESIMS and HRESIMS/MS analysis of **2** was in accordance with the structure assigned by the NMR spectral data, clearly showing the loss of SO_3 (Figure S23) and the main fragments of the decorated *α*-glycosyl, MgS-Ile scaffold and Aaep (Figure S24).

Table 2 - ^{1}H (400 MHz) and ^{13}C (100 MHz) NMR Spectroscopic Data for suomilide 1077 (**2**) in DMSO-d_6.

Unit	no.	δC, type	δH[a] (*J* in Hz)	HMBC[b]	COSY
Ile-MgS	1	66.2, CH_2	3.94, m	4	1b
			3.77, dd (11.5, 7.8)	2, 4	1a
	2	80.2, CH	3.93, m	1, 3, 4	1ab
	3	57.3, CH_3	3.30, m	2	
	4	169.8, C			
	5	53.1, CH	4.61, t (6.7)	4, 6, 7, 10	6, 1-NH
	6	36.4, CH_2	1.71, m		5, 8ab, 9
	7	14.5, CH_3	0.88, m	5	6
	8	25.4, CH_2	1.29, m	6, 7, 9	6, 8a, 9
			1.18, m	7, 9	6, 8a, 9
	9	11.7, CH_3	0.89, m	8	8ab
	10	171.9, C			
	1-NH		7.93, m	4, 5	5
Abn	11	56.9, CH	4.52, d (2.2)	10, 12, 17, 19, 20	19
	12	53.4, CH	4.28, br		13, 14ab
	13	65.7, CH	3.72, m	17	12, 1-OH
	14	28.8, CH_2	2.14, dd (12.5, 4.5)	16	12, 14b, 15
			1.72, m		12, 14a, 15
	15	70.1, CH	3.73, m	1'	14ab, 16ab
	16	35.4, CH_2	2.30, m	15, 17, 19	15, 16b
			1.98, dd (10.6; 5.2)	14, 15, 17	15, 16a
	17	80.5, C			
	18	156.6, C			
	19	58.1, CH	4.22, d (1.7)	13, 17, 20	11, 2-NH
	20	168.8, C			
	1-OH		6.04, d (4.0)	12, 13, 17	13
	2-NH		7.95, m	17, 18, 19	19
Aaep	21	37.2, CH_2	3.18, m	20, 22, 23	22, 3-NH

	22	27.8, CH_2	2.27, m	21, 23, 24	21, 24, 25/27
	23	135.9, C			
	24	119.0, CH	5.64, br	22, 23, 25, 27	22, 25/27
	25	54.9, CH_2	4.13, m	23, 24	22, 24
	26	154.1, C			
	27	55.4, CH_2	4.13, m	23, 24	22, 24
	3-NH		7.57, t (5.5)	20	21
	4-NH		7.14, s		
Galactose	1'	94.7, CH	4.97, d (3.7)	2', 3', 4', 5', 5''	2'
	2'	68.0, CH	4.84, dd (11, 3.7)	3', 2'a	3'
	3'	66.6, CH	5.03, dd (11, 3.4)	4', 3'a	2', 4'
	4'	68.4, CH	5.34, dd (1.0, 3.4)	2', 3', 5', 5'', 4'a	3', 5'
	5'	67.0, CH	4.14, m	1', 4', 5''	1', 4', 5''
	5''	70.0, CH_2	3.34, m	2', 3', 4', 5'a	5', 5''
			3.26, (6.1, 10)	2', 3', 4', 5'a	5', 5''
	2'a	172.5, C			
	2'b	34.4, CH_2	2.29, m	2'a, 2'c, 2'd	2'c, 2'd
	2'c	17.9, CH_2	1.54, m	2'a, 2', 2'd	2'b, 2'd
	2'd	13.4, CH_3	0.88, m	2'b, 2'c	2'c
	3'a	155.4, C			
	NH_2	6.55, s			
	4'a	172.0, C			
	4'b	33.3, CH_2	2.35, td (7.5, 3.2)	4'a, 4'c, 4'd	4'c
	4'c	24.1, CH_2	1.55, m	4'b, 4'd, 4'e	4'b, 4'd
	4'd	30.6, CH_2	1.27, m	4'b, 4'e	4'b, 4'c, 4'e, 4'f
	4'e	21.7, CH_2	1.28, m	4'c, 4'd, 4'f	4'b, 4'c, 4'd, 4'f
	4'f	13.8, CH_3	0.86, m	4'd, 4'e	4'e
	5'a	58.5, CH_3	3.21, s	5''	

[a]from HSQC, [b]from proton to indicated carbon.

The absolute configuration was only determined for the amino acid residue Ile, by Marfey's derivatization, showing a D-configuration, as reported both in suomilide (Ile) and in banyasides A-B (Leu) (Figure S25). Determination of the absolute configuration of the glycosyl residue was not possible, so only the relative configuration is reported. Based on the small $J_{1'-2'}$ value of 3.7 Hz, and from coupling constant analysis of the ^{1}H NMR (Figure S26), *α*-galactopyranose is the glycosyl unit present in **2**. This is the first report of a galactose unit regarding this family of compounds, since both suomilide[23], banyasides[24] and varlaxins[52] have *α*-glucose residues. Searching for glycosylated aeruginosins, the only carbohydrates found to decorate these structures were xylose[53], glucuronic acid and, pentose[27] and mannuronic acid[53]. A wider search of cyanobacterial natural products containing galactose in its structure, retrieve glycolipids from *Oscillatoria* and *Scytonema* cyanobacterial strains[54-56], three mycosporine-like amino

acids (MAAs) from N*ostoc* strains[57,58] and the glycoside cyanopterin, from the *Synechocystis* PCC 6803[59], being considered an uncommon glycosyl unit in cyanobacterial products.[60]

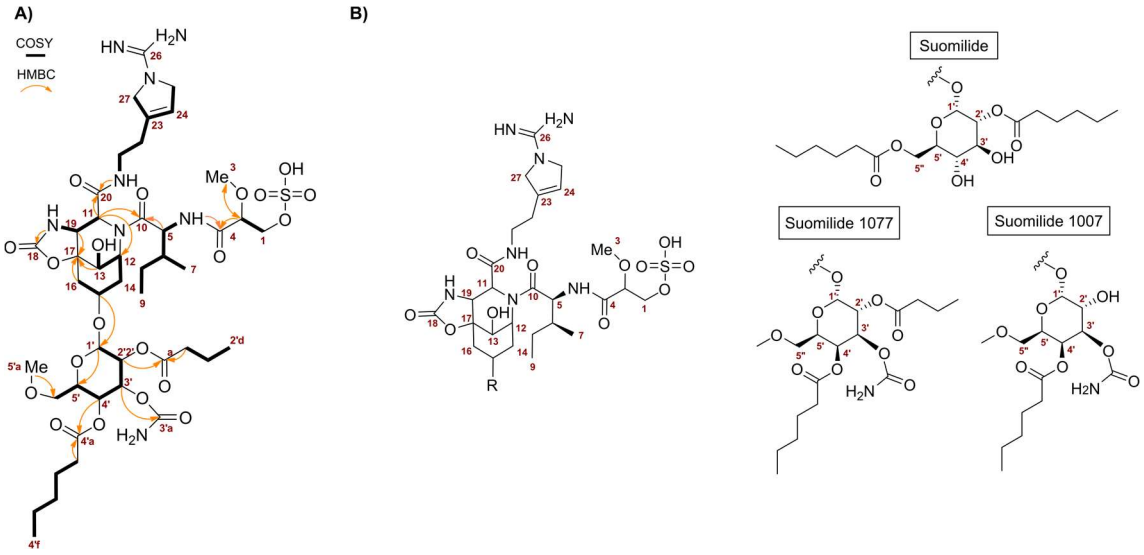


Figure 2 - Structure elucidation of **2**. Key HMBC and COSY. Comparison of the suomilide with **2** and **3**, showing the main structural differences in the sugar unit.

The ^{1}H and ^{13}C NMR data of **3** in DMSO-d_6 (Table S4, Figure S27-28) was similar to those of **2**, revealing only differences associated with the glycosyl resonances, mainly the fatty acyl chains. The lack of the methyl and methylene resonances of butyric acid on the 2D HSQC data (Figure S29), together with the presence of an OH proton, δ_H 4.94 (d, 7.6 Hz), corroborates the loss of the C4 fatty acid connected in position 2 of the glycosyl unit (Figure S30). HRESIMS/MS confirms the structure assignment of **3**, by comparison to the fragmentation spectra of **2** (Figure S31). The stereochemistry of the hydrophobic amino acid Ile was assumed to be the same as **2** and based on the coupling constants of the glycosyl unit, the relative configuration of the glycosyl unit is also *α*-galactose (Figure S32).

Suomilides are glycosylated tetrapeptides, bearing the unusual moieties MgS, Aaep and Abn, in which the first two are encountered in other metabolites from the aeruginosin family (Figure S22). Aeruginosins are also composed of a Choi unique moiety, that in suomilides and banyasides is substituted by an Abn unit. A recent study by Liu and co-workers[53], included suomilide as an integrant part of the structurally diverse aeruginosin family of natural products, and highlighted that the chemical diversity within this family of

metabolites is higher than expected. Within this family of compounds are the previously mentioned banyasides and varlaxyns, as well as pseudoaeruginosins and nostosins[52], from cyanobacterial origin. These natural products are also relevant due to the reported bioactivity, being potent inhibitors of trypsin-type serine proteases, an important characteristic for the pharmaceutical industry in developing anticancer drugs.[29]

4.2.4. Identification and analysis of the biosynthetic gene cluster of **1** and **2-3**

The biosynthetic machinery responsible for nostopepotilide production was already described both for *Nostoc* sp. GSV224 (nostopeptolides A1-A3)[61] and *Nostoc punctiforme* PCC 73102 (Nostopeptolide A1 and 1052)[16,62], corresponding to a canonical hybrid PKS/NRPS biosynthetic route. A search in the genome of *Nostoc* sp. LEGE12450 (Genbank: NZ_JADEXU010000000), annotated using antiSMASH v7.0 (beta2-86685d9d)[63], quickly revealed the presence of a hybrid PKS-NRPS biosynthetic gene cluster in accordance with the assigned structure of **1** and with the known BGCs from the other two *Nostoc* strains. The BGC is composed of two PKS and eight NRPS modules, for which the corresponding adenylation (A) domains are predicted to activate the amino acids found in **1**, as represented in Figure 3. As such, this 43.7 kb BGC was considered the *nos* gene cluster encoding the production of **1** since it contains the known *nos*ABCD cassette[61]. Nevertheless, a detailed annotation of its genes was carried out (Table S5). In contrast to what was described regarding the BGCs encoding for nostopeptolide A1-A3 and 1052 production, the BGC of *Nostoc* sp. LEGE 12450 shows two PKS modules instead of one. The first PKS module is located upstream of *nos*ABCD and is likely responsible for synthesis of the MHEA moiety, which is then condensed to Thr by the first module of the NRPS NosA. This previously unreported PKS module, NosH, features a putative *trans*-AT KS domain, followed by a dehydrogenase (DH), carbon methyl transferase (cMT) and ketoreductase (KR). The proposed biosynthesis represented on Figure 3, starts with an elongation step of a butyric fatty acyl starter unit with one unit of malonyl-CoA, after which ketoreduction by KR, dehydration by DH and methylation at the alpha position by the cMT yields MHEA. The resulting MHEA intermediate is then passed to NosA which incorporates Thr, Ser and Phe; this mega-enzyme has a methyltransferase FkbM domain responsible for the methylation of Ser to give *O*meSer. The biosynthesis continues with the type I PKS NosB, incorporating one malonyl-CoA which is also methylated at the alpha position by a cMT module. This originates the PhePr scaffold in **1**. This biosynthetic route was already described for the modified amino acid LeuAc in nostopeptolides A1-A3/1052[61]. NosC and NosD

incorporate the remaining amino acids residues into the growing chain, which is then released and cyclized by the thioesterase present on the C-terminus of NosD, to give the cyclic peptide. NosD has an epimerase domain that is in accordance with the stereochemistry of the experimentally determined Gln amino acid.

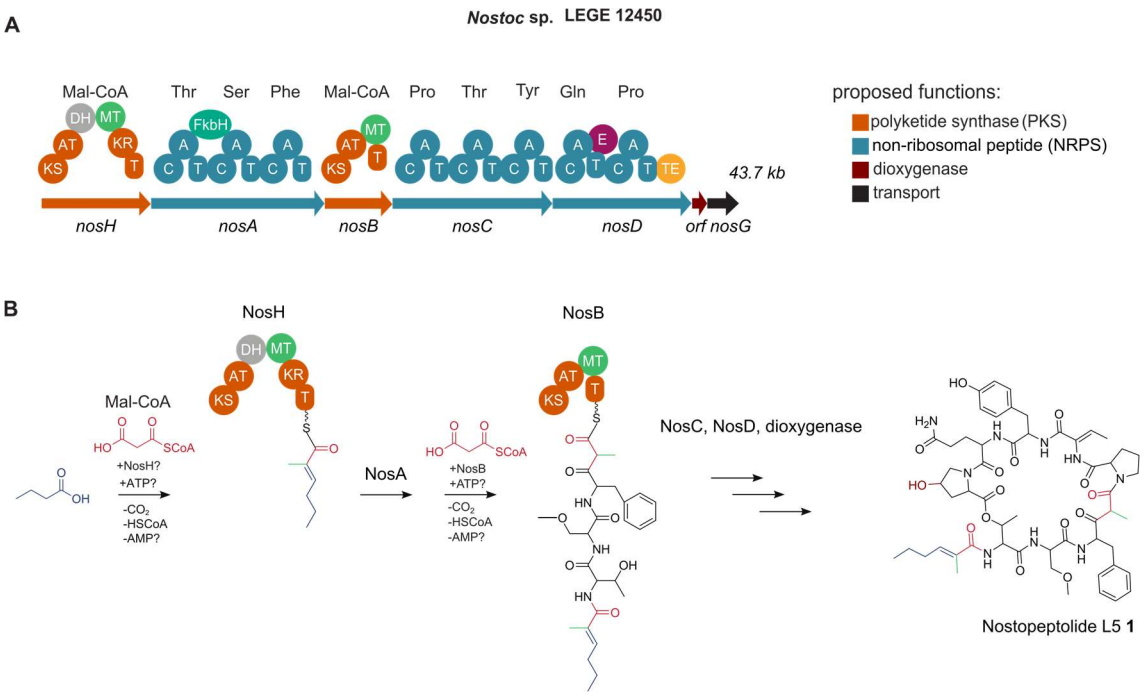


Figure 3 - Biosynthesis of the nostopeptolide L5 (**1**). **A** – Schematic representation of the *nos* BGC, including the proposed functions of the predicted open-reading frames (Orfs), domain composition of detected PKS and NRPS (or NRPS-like) modules, and bioinformatically-predicted (antiSMASH) monomers activated by acyltranferase (AT) or adenylation (A) domains. **B** – Proposed biosynthetic steps leading to **1**, highlighting the formation of the MHEA residue from butyric acid, by PKS machinery; KS – ketosynthase, DH – dehydrogenase, MT – methyltransferase, KR – ketoreductase, T – PCP; FkbH – methyltransferase; E – epimerase; TE – thioesterase.

The previously reported *nos* BGCs contain two additional genes responsible for the production of 4mePro, NosE (alcohol dehydrogenase) and NosF (MeP5C reductase)[61]. An alignment of the reported BGCs of *Nostoc* strains known to produce this class of compounds (Figure S33) with the *nos* BGC of *Nostoc* sp. LEGE 12450, showed that the latter does not contain homologues of NosE or NosF. This was expected, since **1** does not contain a 4mePro residue. Searching for additional genes within the *nos* BGC, an orf annotated herein as a dioxygenase might be involved in the hydroxylation of Pro to generate Hyp. This enzyme is characterized as a phytanoyl-CoA dioxygenase (Table S5) and a quick search on biosynthetic pathways of the other cyanobacterial natural products containing Hyp, indicated that a homologue – IxaN – from the heinamides BGC

[45], is proposed to add a hydroxyl group to proline at position 4. These proposed steps explain the biosynthesis of **1** using canonical enzymology (Figure 3). The final gene in the BGC, NosG, is annotated as an ABC transporter.

To check if the *nos* BGC associated with **1** was widely distributed among cyanobacterial genomes and could be found in genomes of cultured cyanobacteria or in metagenome data, ClusterBlast data from antiSMASH analysis was inspected. Besides the *nos* BGC for nostopeptolides A1-A3 and 1052, a BGC with a similar architecture to the BGC associated with **1** could not be retrieved. However, BlastP searches on Nos proteins revealed that *Nostoc* sp. XA010 contains a similar BGC to that of strain LEGE 12450 and thus, is likely to produce **1** or close analogues (Figure S33). This cyanobacterial strain was isolated from a terrestrial environment in a project with the aim of analyzing the genome evolution of *Nostoc* strains due to their ability to be facultative symbionts.[11] According to the authors, this cyanobacterium is part of the "real *Nostoc*" clade, comprising both free-living and symbiotic strains, but no metabolic profiling was performed in that study.

The biosynthesis of suomilide compounds was proposed based on homology to the aeruginosin biosynthetic NRPS machinery, by Ahmed and co-workers.[29] The structures of members of the aeruginosin family of compounds is composed of four main residues, an *N*-terminal 4-hydroxyphenyl lactate derivative, a hydrophobic amino acid, a Choi moiety and a C-terminal arginine derivative.[53] The putative BGC of suomilide (*suo*) from *Nodularia sphaerocarpa* UHCC 0038 is constituted by 28 proteins, of which 14 were homologous of *aer* enzymes.[29] In the genome of *Nostoc* sp. LEGE 12450, one NRPS-type BGC showed similarities to the putative *suo* BGC, comprising 32 proteins and with 39.9 kb, shown in Figure 4. A detailed annotation of these proteins and their comparison to those of the previously reported *suo* BGC was carried out (Table S6). As described by Ahmed and co-workers[29], the biosynthesis of **2** and **3** likely starts with SouB, responsible for the generation of the MgS moiety, by loading glycerate that is subsequently methylated and sulfated. In *Nostoc* sp. LEGE 12450, *souB* is followed by two orfs (Orf 3 – hypothetical protein and Orf4 - adenylyl-sulfate kinase), while in strain UHCC 0038 three orfs follow *suoB* (encoding Orfs2-4, nuclear transport factor 2, hypothetical protein and adenylyl-sulfate kinase, respectively) (Table S6). The enzymes SouC[64], SouK and SouD-F are involved in the biosynthesis of the Abn moiety, together with the $SuoH_{1-4}$ oxygenases.[29] All these enzymes seem to work in an off-line manner and the Abn residue is then loaded to the growing peptide chain by SuoG. A new transporter enzyme between SuoK and SuoD appears in the *suo* BGC of *Nostoc* sp.

LEGE 12450 (Orf 28), with unknown function. SuoG is a dimodule NRPS, in which the first activates and incorporates Abn, and the second is an off-loading module that incorporates the arginine derivative to the C-terminus of the tetrapeptide. The mechanism responsible for the conversion of Arg to Aaep is still unknonwn.[29] Decoration of the Abn moiety with a glucose unit is achieved by SouI in suomilide, however the annotated *suo* BGC in *Nostoc* sp. LEGE 12450 lacks this enzyme, despite compounds **2** and **3** being glycosylated. Orfs 21 and 22, annotated as carbohydrate binding proteins, might be involved in the incorporation of galactose in the structure of **2** and **3**. The addition of the fatty acyl chains and the carbamoyl residue to galactose are performed by Orf16 (MBOAT enzyme) and Orf17 (carbamoyl transferase), respectively. Comparing the remaining proteins of *Nostoc* sp. LEGE 12450 to those of the already reported *suo* BGC (Figure S34), Orfs 29-31 are new and possibly related to the structural differences on the glycosyl unit between suomilide and **2**, **3**. A class I SAM-dependent methyltransfese (Orf 29) might be involved in the methylation of the hydroxyl group on position 6 of the sugar residue. The extra GDSL hydrolase (Orf 30) and a SAM radical enzyme (Orf 31) might be involved in the exchange of hexanoic acid from position 6 in suomilide to position 5 in **2** and **3**. The *suo* BGC of *Nodularia sphaerocarpa* UHCC 0038 ends with two transporters (Orf 25 and SuoN2), but these were not found in the proposed *suo* BGC from *Nostoc* sp. LEGE 12450.

Once again, ClusterBlast was used to check if the *suo* BGC of **2**,**3** was widely distributed among cyanobacterial genomes and could be found in genomes of cultured cyanobacteria or in metagenome data. Two homologous BGCs were found in *Nostoc* sp. 'Peltigera membranacea cyanobiont' N6[65] and *Scytonema* sp. NIES-4073. The first cyanobacterium is a lichen cyanobiont and showed also the orfs 28-31 as found in the *suo* BGC of *Nostoc* sp. LEGE 12450, but no metabolomic analysis of the cyanobacterium was performed. Taking in consideration the similarity of the aligned BGCs (Figure S35) is possible for this strain to produce the same compounds or analogues of **2**, **3**. The second cyanobacterium is a terrestrial strain, showing the same BGC organization with the newly reported orfs, but the N-terminal amino acid seems to be valine and no sulfotransferase was present in the cluster. This starting amino acid is also found in dysinosins, isolated from a sponge, and belonging to the aeruginosin family of compounds.[53] Besides *Nostoc* sp. LEGE 12450, suomilides seem to be present in other cyanobionts, as those of the plant *Blasia plusia* L.[19] and of mosses[65].

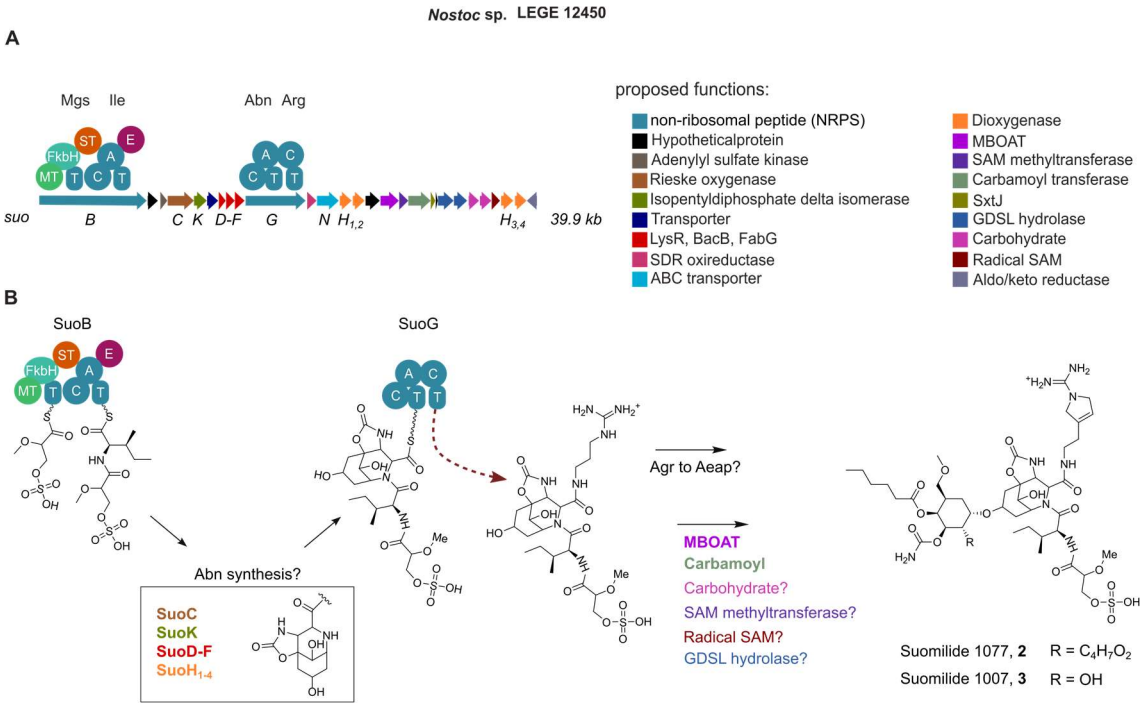


Figure 4 - Biosynthesis of the suomilides 1077 (**2**) and 1007 (**3**). A – Schematic representation of the *suo* BGC, including the proposed functions of the predicted open-reading frames (orf), domain composition of detected NRPS (or NRPS-like) modules, and the activated monomers activated by adenylation (A) domain, based on the putative *suo* BGC from *Nodularia sphaerocarpa* UHCC 0038. B – Proposed biosynthetic steps leading to **2** and **3**, following the putative *suo* BGC from *Nodularia sphaerocarpa* UHCC 0038. MT – methyltransferase; ST – sulfotransferase. Enzymes in bold are confirmed to be involved in the proposed biosynthetic steps.

4.2.5. Biological activity

Compounds **1**-**2** were tested against a panel of four bacteria, two Gram-negative and two Gram-positive (*Escherichia coli, Salmonela typhimurium*, *Staphylococcus aureus* and *Bacilus subtilis*) and one fungal strain (*Candida albicans*). Neither compound demonstrated antibacterial activity in the selected strains. Anticancer activity was tested against the liver (HepG2) and colon (HCT 116) carcinoma cells lines in a 2D model, and the for the colon carcinoma HCT 116, also in a 3D model. Both compounds, did not show bioactivity against the tested cell lines (Table S7).

Biofouling causes one of the bigger maintenance costs in the maritime industries, since it corresponds to the adhesion of organisms to the submerged surfaces. To fight such inescapable problem, the use of antifouling coatings is the common strategy applied. To check the if compounds **1**-**2** could show antifouling activity, an anti-settlement bioassay towards the macrofouling species was conducted, using *Mytilus galloprovincialis*. The compounds were tested initially at a concentration of 50 μM. Nostopeptolide L5 (**1**), showed potent activity against the settlement of *M.*

galloprovincialis larvae, leading to 0% of settlement, while **2** showed 65% of attachment of the larvae (Figure S34). Dose-response curves showed that compound **1** was moderately potent with an EC_{50} of 19.4 μM in this assay (Figure S36), being a good candidate to be developed as an antifouling agent.

4.3. Material and Methods

4.3.1. General experimental procedures

Optical rotations were measured using a JASCO P-2000 polarimeter (JASCO) with SpectraManager 2.14.02 software. Infrared spectra were collected on a Nicolet iS5 FTIR spectrometer (ThermoScientific) with OMNIC 9.8.372 software. The UV spectra were acquired on a UV-1600PC spectrometer (VWR) controlled by MWAVE 1.0.20 software. 1D and 2D NMR data were obtained in the Materials Center of the University of Porto (CEMUP) on a Bruker Avance III, 400 MHz controlled by TopSpin 3.2. Compounds were analyzed in deuterated DMSO (DMSO-d_6, Sigma). Chemical shifts (^{1}H and ^{13}C) are expressed in δ (ppm), referenced to the residual non-deuterated solvent used (^{1}H, 2.500 ppm; ^{13}C, 39.52 ppm). NMR data were analyzed in MNova 12.0.4. LC-HRESIMS and LC-HRESIMS/MS analyses were performed on an UltiMate 3000 UHPLC (Thermo Fisher Scientific) system composed of an LPG-3400SD pump, WPS-3000SL autosampler and VWD-3100 UV/VIS detector coupled to a Q Exactive Focus Hybrid Quadrupole-Orbitrap Mass Spectrometer controlled by Q Exactive Focus Tune 2.9 and Xcalibur 4.1 (Thermo Fisher Scientific). For LC-HRESIMS data, full Scan mode was used with capillary voltage set to -3.8kV, capillary temperature to 300°C and sheath gas flow rate to 35 units. HPLC separations were performed in a JASCO PU-4180 HPLC pump with MD-4010 Photo Diode Array Detector, and separations were monitored at 210, 254, 280, 366 nm wavelengths. All solvents used were ACS grade, except for HPLC solvents (HPLC gradient grade) and LC-MS solvents (MS-grade). Amino acids were purchased from Fluorochem (L-Phe, D-allo-Thr; L-allo-Thr; D-Ser; L-Ser; L-Pro; D-Pro; D-Tyr; L-Tyr; D-Thr; L-Glu; D-Glu; D-allo-Ile, L-allo-Ile), Glenthan Life Sciences (D-Phe; L-Thr; D-Ile), TCI (*cis*-4-D-Pro; *trans*-4-L-Pro, *cis*-4-L-Pro), Sigma Aldrich (*cis*-4-D-Pro) and Bachem (L-allo-Ile).

4.3.2. Cyanobacterial strain cultivation

The cyanobacterial strain *Nostoc* sp. LEGE 12450, kindly provided by LEGEcc[66], was cultured in Z8 media[67], at 25 °C, with constant aeration, under a 16h/8h light (10−30 *μ*mol photons s^{-1} m^{-2})/dark cycle. Large-scale culturing of this strain was carried out in 20 L, to a total volume of 160 L, in Nalgene containers with aeration and cells were harvested during late exponential/early stationary phase (22 months of growth) through centrifugation (7178 ×*g*), immediately frozen, and freeze-dried until further usage.

4.3.3. Molecular Networking analysis

The cyanobacterial strain *Nostoc* sp. LEGE 12450 was analyzed using GNPS Molecular Networking analysis[32], as reported in chapter II. Briefly, 4 L of culture were harvested at the exponential phase of growth, frozen and freeze-dried. A crude extraction of the dry cell material using a mixture of CH_2Cl_2/MeOH (2:1, v/v) was performed at room temperature and the resulting extract was fractionated in a Strata SI-1 Silica cartridge (5 g, 55 *μ*m, 70A, Phenomenex) using a gradient of increasing polarity from *n*-hexane (*n*-hex) to EtOAc to MeOH, to yield three fractions (A-C). For LC-HRESIMS analysis, fractions were prepared at 1.0 mg mL^{-1} in LC-MS grade MeOH, filtered through a 0.2 *μ*m syringe regenerated cellulose filter, and 5 *μ*L were injected in an ACE Ultracore 2.5 SuperC18 column (75 × 2.1 mm, Avantor). Samples were eluted at 0.35 mL min^{-1} over a linear gradient from 99.5% solution A (95% H_2O, 5% MeOH, 0.1% v/v HCOOH) and 0.5% solution B (95% isopropanol, 5% MeOH, 0.1% v/v HCOOH) to reach 10% solution B over 0.5 min, followed by an increase to 60% solution B in 8 min and then to 90% in 1 min; these conditions were kept for 6 min before returning to the initial conditions. The column oven was set to 40 °C, UV monitoring was carried out at 254.0 nm. The capillary voltage of the Heated Electrospray Ionization (HESI) was set to 3.8 kV and its temperature to 300 ºC. The sheath gas and auxiliary gas flow rate were at 35 and 10 (arbitrary units as provided by the software settings). Full MS scans were acquired at a resolution of 70,000 fwhm (range of 150–2000 *m/z*), and data dependent MS/MS ($ddMS^2$, Discovery mode) at a resolution of 17,500 fwhm (isolation window used was 3.0 amu, and normalized collision energy was 35). For the molecular network analysis, the protocol available in GNPS documentation was followed and the resulting network file was visualized using Cytoscape v3.8.0[68].

4.3.4. Extraction and isolation of **1**-**3**

Freeze-dried biomass from large-scale culturing (60 g, d.w.) was extracted by repeated percolation using a mixture of CH_2Cl_2/MeOH (2:1, v/v) at room temperature and at 40 ºC. A total extract mass of 10.4 g was obtained. Further fractionation using vacuum liquid chromatography (VLC) was performed, using silica gel 60 (0.015-0.040 mm, Merck). The extract was dry loaded onto the column using 10 g of silica. A mixture of 1:1 (v/v) EtOAc/*n*-hex was used to start the separation, and stepwise gradient to 100% EtOAc and then to 100% MeOH was used, yielding 13 fractions. Fractions containing the mass features of interest (3.7 g), eluting with 3:2 (v/v) to 1:9 (v/v) EtOAc:MeOH, were further fractionated using reversed phase VLC (Silica gel C18-RP, 17% C, ca.0.8 mmol g^{-1}, 40-63 µ, Thermo Sicentific), with a stepwise gradient from 5% MeOH (aq) to 100% MeOH, yielding 10 fractions. The mass features of interest were present in higher intensity in fraction seven (239 mg, eluted in 90% MeOH (aq)), which was further separated by reversed phase semi-preparative HPLC using an Aeris Peptide XB-C18 (100 Å, 250 × 10 mm, 5 *µ*m, Phenomenex). A gradient elution program was employed, starting with 10% MeCN (aq), increasing to 60% MeOH (aq) over 40 minutes, and then to 90% MeOH (aq) in 5 minutes for column cleaning, before returning to the initial conditions. For nostopeptolide L5 (**1**) isolation, the HPLC fraction E1 (30.3 mg, t_R 28.5-30 min) was collected and subjected to an additional purification step using as stationary phase an C18 column (100 Å, 250 × 10 mm, 5 µm, ACE) with a gradient starting in 25% MeCN (aq) and increasing to 50% MeCN (aq) in 20 min, and further to 60% MeCN (aq) in 5 min for column cleaning, before returning to the initial conditions. The purified fraction E1-A corresponded to the spectroscopically pure nostopeptolide L5 (**1**, t_R = 18.0 min, 19 mg). For suomilides isolation, the HPLC fraction C (6.1 mg, t_R 20-24 min), from the first HPLC separation, was initial used to purify compound **3**. However, since only 1 mg of suomilide 1007 (**3**) was obtained, exploration of the previous fractions was conducted. The reversed-phase VLC fraction 4 (96.7 mg), eluting with 40% MeOH (aq) was selected for further purification steps due to its low sample complexity and high mass. This fraction was subjected to a semi-preparative HPLC using a C18 column (100 Å, 250 × 10 mm, 5 µm, ACE) and for mobile phase a gradient elution was employed starting with 10% MeCN (aq), increasing to 55% MeCN (aq) in 40 min, and then to 70% MeCN (aq) for column cleaning, before returning to the initial conditions. Fraction C (3.9 mg, t_R 21.9-27 min) was used for purification of **3**, while fractions D-E (8.4 mg, t_R 27-29.5 min) were combined in an effort to isolate **2**. Starting with compound **3**, fraction C was further purified using an analytical HPLC step with a Synergy Hydro-RP (80Å, 250 x 4.6 mm, 10 µm, Phenomenex) and isocratic elution at 25% MeCN (aq) resulted in three fractions, on

which C2 corresponded to the spectroscopically pure suomilide 1007 (**3**) (2.3 mg, t_R 15-17 min). Joining to the previous purified mass from fraction C, a total of 3.3 mg of compound **3** was obtained. Finally, for compound **2**, the joined HPLC fractions DE were subjected to a semi-preparative HPLC using a C18 column (100 Å, 250 × 10 mm, 5 µm, ACE), using an isocratic eluent of 34% MeCN (aq) and fraction DE-4 (4.8 mg, t_R 12.5-14 min) corresponded to the spectroscopically pure suomilide 1077 (**2**). The physical chemistry analysis optical rotation, infrared (IF) and UV absorption were performed for the two main compounds (**1**, **2**) and are reported as follows.

Nostopeptolide L5 (**1**). Glassy material; $[\alpha]_D^{25}$ -116 (*c* 0.01, MeOH); IR ν_{max} 3271.1, 3061.9, 2933.6, 2873.2, 1732.1, 1660.1, 1651.2, 1644.5, 1633.6, 1615.0, 1537.8, 1515.1, 1505.0, 1454.7 cm^{-1} (Figure S37); UV (MeOH) λ_{max} (log ε) 272.5 (3.0), 253.5 (3.1), 240.5 (3.1) (Figure S39); ^1H and ^{13}C NMR spectroscopic data (DMSO-d_6) Table 1; HRESIMS *m/z* 1100.5286 $[M+H]^+$ (calcd. for $C_{55}H_{74}N_9O_{15}$ 1100.5304); HRESIMS/MS (Figure S14).

Suomilide 1077 (**2**). Glassy material; $[\alpha]_D^{25}$ +73 (*c* 0.01, MeOH); IR ν_{max} 3350.5, 2959.9, 2929.5, 2873.5, 2282.2, 1738.3, 1651.1, 1537.7, 1462.5, 1416.9 cm^{-1} (Figure S38); UV (MeOH) λ_{max} (log ε) 322.0 (1.9), 275.5 (2.3), 224.5 (3.1) (Figure S40); ^1H and ^{13}C NMR spectroscopic data (DMSO-d_6) Table 2; HRESIMS *m/z* 1077.4657 $[M+H]^+$ (calcd. for $C_{45}H_{73}N_8O_{20}S$ 1077.4662); HRESIMS/MS (Figure S24).

4.3.5. HRESIMS/MS analysis

HRESIMS/MS analysis of pure **1**-**3** was performed by direct injection of each compound (1.0 mg mL^{-1}, flow 0.005 mL min^{-1}) into the spectrometer, with a 35 000 FWHM resolution, using an isolation window of 1 *m/z*, loop count of 3, and AGC target of 5 × 10^4, and a collision energy of 45 (arbitrary units). For compound **3**, the fragmentation analysis was performed by LC-HRESIMS/MS following the same chromatographic and spectrometric conditions described for the molecular networking chapter (4.3.3.), but with the collision energy of 40 (arbitrary units), $ddMS_2$ in confirmation mode and using inclusion mass without the sulfate group (*m/z* 927.466).

4.3.6. Quantification of **1**-**3** in *Nostoc* sp. LEGE 12450

To quantify the production of nostopeptolide L5 and suomilides 1077/1007 in the cyanobacterial *Nostoc* sp. LEGE 12450, a calibration curve for each compound was

obtained from LC-HRESIMS analysis. Briefly, standards of each compound were prepared at 0.5, 0.1, 0.05, 0.01, 0.005 and 0.001 mg mL^{-1}. Five-microliters (5 μL) of each standard solution were injected in an LC-HRESIMS system fitted with an ACE Ultracore 2.5 SuperC18 column (75 × 2.1 mm, Avantor) following the same chromatographic and spectrometric conditions described in the molecular networking analysis method. The area under the curve from the extracted ion chromatograms (EIC) for each compound were plotted, and a linear regression was fitted. This was then used to extrapolate the concentration of each compound in the initial crude extract used for isolation and in the biomass from which it was obtained (Table S2, Figure S4).

4.3.7. Marfey's derivatization analysis

The Marfey's derivatization analysis was applied to compounds **1** and **2**. To do the acid hydrolysis, 0.5 mg of compound was placed into 1.0 mL of 6 N HCl for 24h at 110°C under continuous stirring. After the reaction, HCl was removed under a N_2 stream. Amino acids standards (Phe, Ser, Thr, allo-Thr, Gln, Tyr, hydroxy-Pro, Ile and allo-Ile), L- and D- isomer standards of each amino acid were prepared at 50 mM in 50 μL of LC-MS grade H_2O, with the exception of Tyr (2.9 mM and 2.5 mM for D- and L- isomers, respectively). For the Gln residue, standards where prepared using L-Glu and D-Glu. A Marfey's reagent (1-fluoro-2,4-dinitrophenyl-5-L-alanine amide) solution was prepared at a 1% (w/v) concentration in 100 μL acetone. For the hydrolysate derivatization, 100 μL of 1 M $NaHCO_3$ and 100 μL of the Marfey's reagent acetone solution were added to a glass vial, which was placed on an oil bath at 40 ºC for 1 h with a magnetic stirrer. The reaction was quenched with 100 μL of 1 M HCl, diluted with 200 μL MeCN (LC-MS grade) and filtered (0.2 µm). The same protocol was applied to the amino acid standards: to 50 μL of each standard were added 20 μL of 1 M NaHCO3 and 100 μL of the Marfey's reagent solution. The reaction (40 ºC, 1h) was quenched with 20 μL 1 M HCl and diluted with 810 µL MeCN. After derivatization all samples were analyzed by LC-HRESIMS, using an ACE UltraCore 2.5 SuperC18 column (75 × 2.1 mm, Avantor). Eluents were solutions A (95% H_2O, 5% MeOH, 0.1% v/v HCOOH) and B (100% MeCN, 0.1% v/v HCOOH). The separation was carried out as follows: 100% A for 1 min, then a linear gradient for 2 min to 20% B and then a steeper gradient to 50% B in 15 min and a cleaning step reaching 100% B before returning to the initial conditions. Injection volume was 2 μL for all amino acids with the exception of Tyr that was 5 μL. Retention time (t_R, min) for standards of **1**: L-Ser 7.03, D-Ser 7.15, L-Thr 7.22, D-Thr 8.10, L-*allo*-Thr 7.31, D-allo-Thr 7.67, L-Glu 7.67, D-Glu 8.06, L-Phe 11.47, D-Phe 12.91, L-Pro 8.45, D-Pro 8.89,

L-Tyr 7.20, D-Tyr 7.72, *cis*-L-Hyp 7.01, *trans*-L-Hyp 6.55, *cis*-D-Hyp 7.05, *trans*-D-Hyp 6.27. Retention time (t_R, min) for hydrolysate of **1**: L-Ser 7.04, L-Thr 7.24, D-Glu 8.08, L-Phe 11.52, L-Pro 8.41, L-Tyr 7.16 and *trans*-L-Hyp 6.53 (Figure S15). Retention time (t_R, min) for standards of **2**, L-Ile 11.28, D-Ile 13.05, L-*allo*-Ile 11.28, D-*allo*-Ile 13.09, and the retention time for the hydrolysate was D-*allo*-Ile 13.11 (Figure S25).

4.3.8. Genome sequencing and bioinformatics analysis

The total genomic DNA from cyanobacterium *Nostoc* sp. LEGE 12450 was isolated from a fresh culture pellet (50 mL) using the commercial NZY Plant/Fungi gDNA Isolation kit (Nzytech), following the manufacturer's instructions. The genome was sequenced by MicrobesNG, using Illumina technology and 2 x 250 bp paired-end reads. Because the culture was not axenic, the genomic data was treated as a metagenome: the retrieved contigs were analyzed using the binning tool MaxBin 2.0.[69] This procedure yielded a draft genome (Genbank: JADEXU000000000) with an estimated size of 8.8 Mb, distributed among 415 contigs. AntiSMASH[63] v7.0.0 (beta2-86685d9d) was used to annotate BGCs in the genome data using relaxed detection settings and all extra features selected. Manual annotation of the genome of *Nostoc* sp. LEGE 12450 was performed using BlastP. Clinker of CAGECAT was used to align the BGCs (default conditions)[70].

4.3.9. Bioactivity profiling

Antibacterial assays were performed in collaboration with Mariana Girão, using the disc diffusion method and using *Staphylococcus aureus*, *Bacilus subtilis*, *Escherichia coli*, *Salmonela typhimurium* and *Candida albicans*. Mueller-Hinton agar medium (Oxoid) was used for bacterial inoculation and Sabouraud agar medium (Oxoid) was used for fungi inoculation. Petri dishes with the media and inocula were incubated for 24h at 37 °C and the turbidity adjusted to 0.5 McFarland standard (OD600nm = 0.095-0.110). The plates were seeded with this inoculum, the 6 mm blank discs (Oxoid) distributed in the inoculated plates and impregnated with 15 μL of the 1 mg mL^{-1} stock solution of each fraction. Blank disks impregnated with 15 μL DMSO served as negative controls. After 30 min at room temperature, the plates were incubated at 37 ºC for 24h and the antimicrobial activity was estimated by the development of an inhibition halo around the disk.

The anticancer assays were performed in collaboration with Dr. Ralph Urbatzka and Leonor Ferreira, in CIIMAR. Human liver hepatocellular carcinoma (HepG2) and human colon carcinoma cell line HCT 116, were obtained from Sigma-Aldrich and cultured at 37 ºC in a humidified atmosphere of 5% CO_2, in Dulbecco's modified Eagle medium (DMEM) (Gibco, Thermo Fisher Scientific) and McCoy's 5A medium (Sigma-Aldrich) respectively, both complemented with 10% (v/v) fetal bovine serum, 1% penicillin/streptomycin (Pen-Strep 100 IU mL^{-1} and 10 mg mL^{-1}, respectively) and 0.1% amphotericin B (250 µg mL^{-1}). Cell viability on 2D cell model was assessed by the MTT (3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide) assay. HCT 116 and HepG2 cell lines were seeded in 96-well plates at a density of 3.3×10^4 cells mL^{-1} and incubated overnight before adding **1** and **2** in triplicates at a concentration of 10 *µ*M. DMSO served as solvent (0.1%) and positive control (20%). After 48h, 20 *µ*L of MTT (1 mg mL^{-1}) was added and incubated for 4h. Media was removed and formazan crystals were dissolved in 100 *µ*L of DMSO and absorption was read at 570 nm on a microplate reader (Synergy HT, Biotek, Germany). To produce 3D spheroids, the human colon carcinoma cell line HCT 116 was seeded at a concentration of 5.0 x 10^4 cells mL^{-1} on ultra-low attachment, round-bottom 96-well plates (Costar, Corning) and allowed to settle for 30 minutes at room temperature before incubation for 5 days at 37 ºC and 5% CO_2. Compounds were added in triplicates to renewed media in concentrations ranging from 50 to 0.2 *µ*M. DMSO was used as solvent control (0.5%) and staurosporine (1 *µ*M) as positive control. After 96h cell viability was assessed using the acid phosphatase assay. Cell medium was removed and spheroids washed with PBS before incubation in 100 *µ*L of p-nitrophenyl phosphate (2 mg mL^{-1}) in sodium acetate buffer (0.1 M) for 2h in the dark. NaOH at 1N was added to stop the reaction before reading the absorbance at 405 nm on a microplate reader (Synergy HT, Biotek).

The larvae anti-settlement bioassays with *M. galloprovincialis* was performed in collaboration with Dr. Joana Almeida and Sandra Pereira (CIIMAR).[71] Briefly, samples of *M. galloprovincialis* juveniles were collected from mussel beds during low tides in the intertidal zone of Memória beach, Matosinhos, Portugal (4113059" N; 843028" W). Mussel plantigrade larvae (0.5–2 mm) were sorted in a binocular magnifier (Olympus SZX2-ILLT, Tokyo, Japan) and kept in filtered seawater until bioassays preparation. Mussel plantigrades exhibiting foot exploring behavior were selected for exposure to the compounds at 50 *µ*M in 24-well microplates for 15 h, in the dark at 18 °C. Five competent larvae were included per well and four well replicates per condition. Stock solutions (50 mM) in DMSO were used to prepare test media in filtered seawater (0.1%). Negative and positive controls were used, namely DMSO (0.1%) and CuSO4 (5 M) as an efficient

antifouling agent. After exposure, the anti-settlement bioactivity was determined by the observation of efficient attached plantigrade larvae considering the existence/absence of produced byssal threads for each test condition. For the EC_{50} assay, the same protocol was followed differing on the concentration of the tested compound (100 *μ*M, 50 *μ*M, 25 *μ*M, 12.5 *μ*M, 6.2 *μ*M and 3.1 *μ*M).

4.4. Conclusions

Chemical and genomic exploration of the cycad coralloid root symbiont cyanobacterium *Nostoc* sp. LEGE 12450, led to the isolation and structure elucidation of the lipopeptide nostopeptolide L5 (**1**) and the tetrapeptides suomilide 1077 (**2**) and 1007 (**3**). Putative biosynthetic gene clusters had already been proposed for both families of compounds; exploring the genome of *Nostoc* sp. LEGE 12450 led to the identification of BGCs matching the previous proposals, which strengthens the respective assignments. For compound **1**, the *nos* BGC showed a minor difference when compared to the BGCs from cyanobacteria known to produce nostopeptolides A1-A3 (*Nostoc* sp. GSV 224) and 1052 (*N. punctiforme* PCC 73102). An additional PKS upstream of the *nosABCD* cassette, might be responsible for the difference in acyl chain between these compounds. 2-Methylhex-2-enoic acid (MHEA) is featured in Nostopeptolides L1-L4 and in (**1**), but this was the first report of the proposed biosynthesis of the MHEA residue. Another interesting feature regarding these compounds (and nostopeptolides L1-L5), as reported and suggested by Liu and co-workers, is the methylation of the Ser residue that led to the macrocycle being generated using the hydroxyl group of Thr, possibly being a new method of directing the macrocyclization to a specific residue. However, additional studies need to be performed to confirm this hypothesis.

Compounds **2** and **3** are suomilide analogues, a class of compounds that belong to the aeruginosin family. This family of cyanobacterial natural products is composed of four amino acid residues, in which three can be highly modified; in the case of suomilides glycosylation and sulfation occur. Comparing the structures of the previously reported suomilide with **2** and **3**, the only differences are in the glycosyl unit connected to the Abn residue. The compounds **2**/**3** have *α*-galactose instead of *α*-glucose, and fatty acyl decorations of butyric acid (C4), hexanoic acid (C6), carbamoyl (C_2H_2NO) and methyl groups. These differences are reflected on the enzymes present in the *suo* BGC of *Nostoc* sp. LEGE 12450 when comparing to the putative *suo* BGC of *Nodularia sphaerocarpa* UHCC 0038. Many of the enzymes present in both clusters are yet

uncharacterized, which opens up the opportunity of discovering new enzyme functions and eventually new biosynthetic pathways in which these are involved. In terms of bioactivity, only compound **1** showed antifouling activity with moderate potency.

Symbiotic cyanobacteria proved to be a good niche to discover new chemical scaffolds and biosynthetic enzymes. Although the role of nostopeptolides in cyanobacterial symbiosis is already reported, the chemical ecology behind these ancient relationships is still underexplored. The detection of suomilides in more symbiotic cyanobacterial strains, might indicate the potential role of these chemically diverse compounds in the symbiotic cross-talk between the cyanobacterium and its host.

4.5. Supplementary material

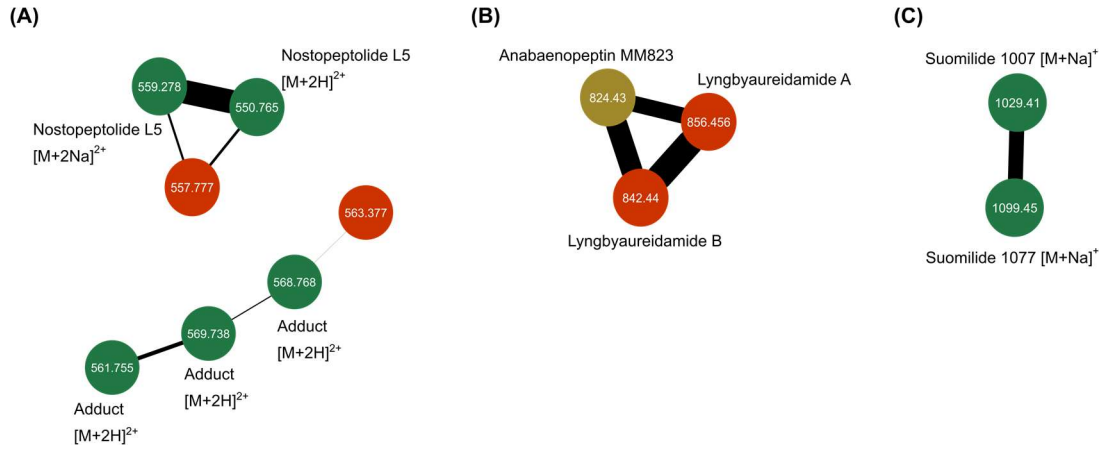


Figure S1 - GNPS molecular networking nodes present on the methanolic fraction of *Nostoc* sp. LEGE 12450. **(A)** Adducts of nostopeptolide L5; (**B**) Mass features correspondent to the known anabaenopeptins; Sodium adduct of suomilides 1077 and 1007; The mass features highlighted in green were isolated in the present work.

Table S1 - Mass features from the molecular networking analysis obtained using LC-HRESIMS/MS data for fractions C of the organic extracts of *Nostoc* sp. LEGE 12450. The more intense masses in the crude extract are shown in bold.

Mass value	**Intensity**		**Adducts (calc.)**					
m/z	**Fraction C**	**Crude**	**$[M+NH_4]^+$**	**$[M+Na]^+$**	**$[M+2H]^{2+}$**	**$[M+H+NH4]^{2+}$**	**$[M+H+Na]^{2+}$**	**$[M+2Na]^{2+}$**
1100.530	8.17E+07	9.30E+07	1117.556	1122.512	550.769	559.282	572.751	561.760
1114.546	1.85E+06	1.05E+06	1131.572*	1136.528	557.776	566.290	568.767	579.758
563.376	2.33E+06	-	580.403*	585.358*	282.192*	290.705*	293.183*	304.174*
824.430	2.39E+07	2.61E+06	841.456*	846.412	412.718	421.232*	423.709	434.700
856.456	1.28E+07	1.45E+06	873.482*	878.438	428.732	437.245*	439.722*	450.713
842.441	7.71E+06	8.47E+05	859.467	864.423	421.724*	430.237*	432.715*	443.706*
1007.423	2.31E+07	1.88E+07	1024.450	1029.405	504.215*	512.729*	515.206*	526.197*
1077.466	5.17E+07	7.52E+07	1094.492	1099.448	539.236*	547.750*	550.227*	561.218

*Masses not found in the MS spectra.

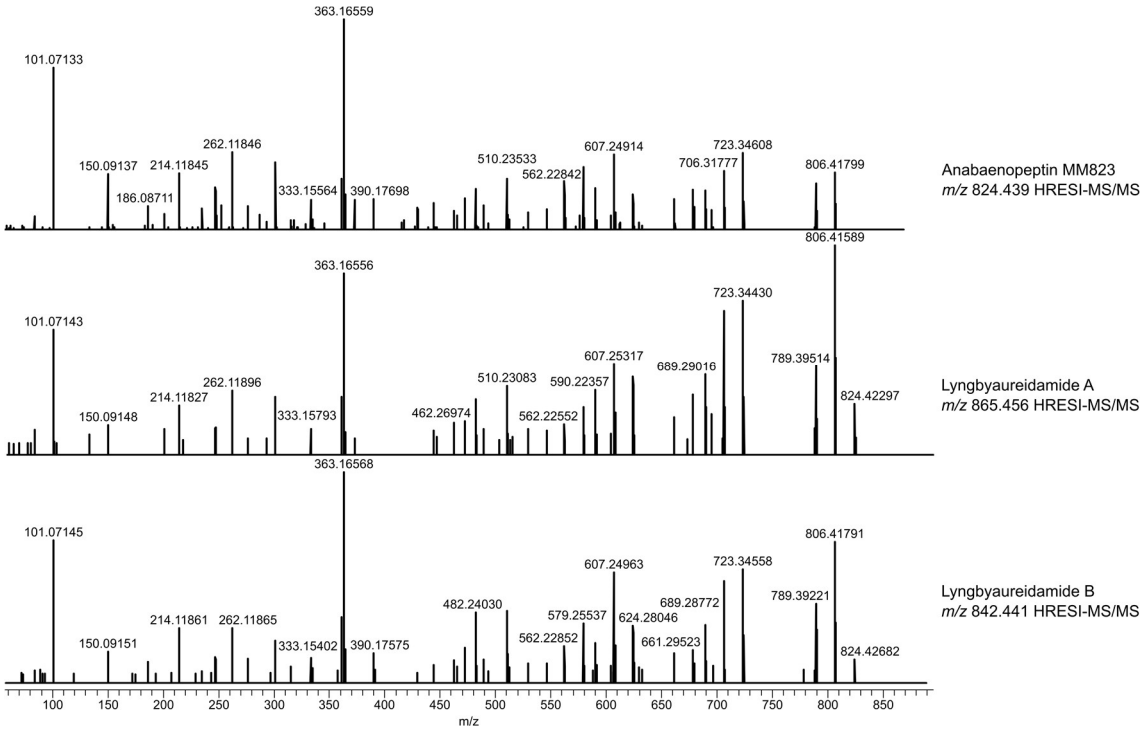


Figure S2 - HRESIMS/MS of the anabaenopeptins found in the crude extract of *Nostoc* sp. LEGE 12450, showing the identical fragmentation pattern of three metabolites.

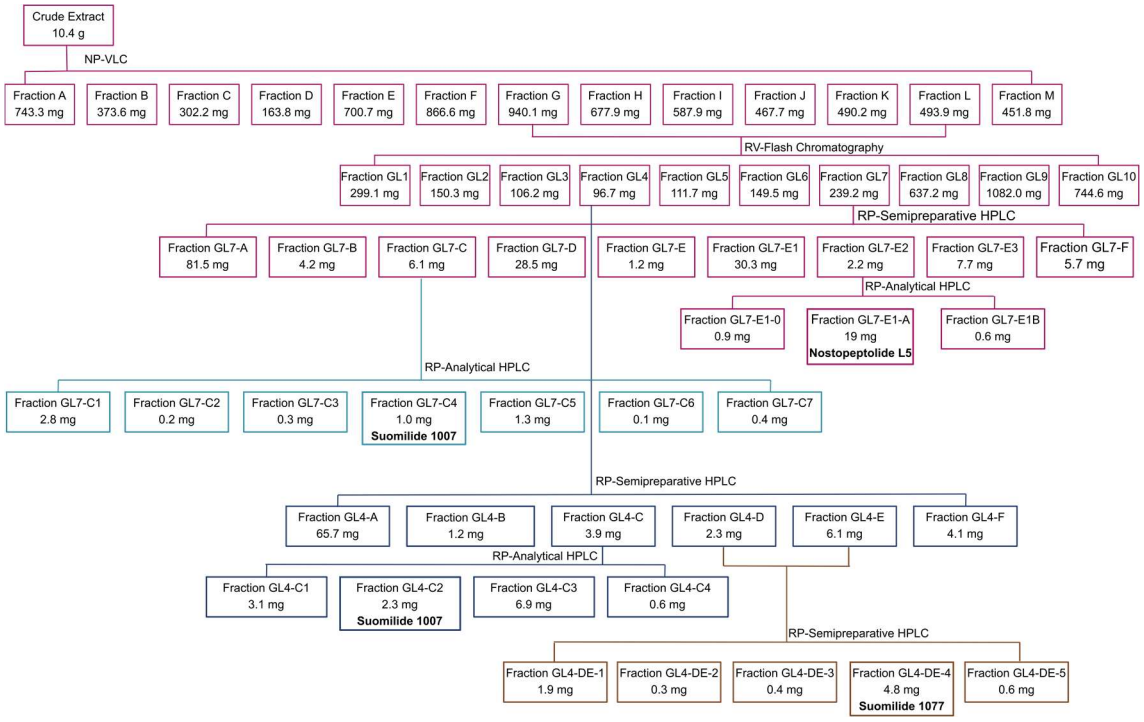


Figure S3 - Fractionation scheme leading to the purification of **1**-**3**.

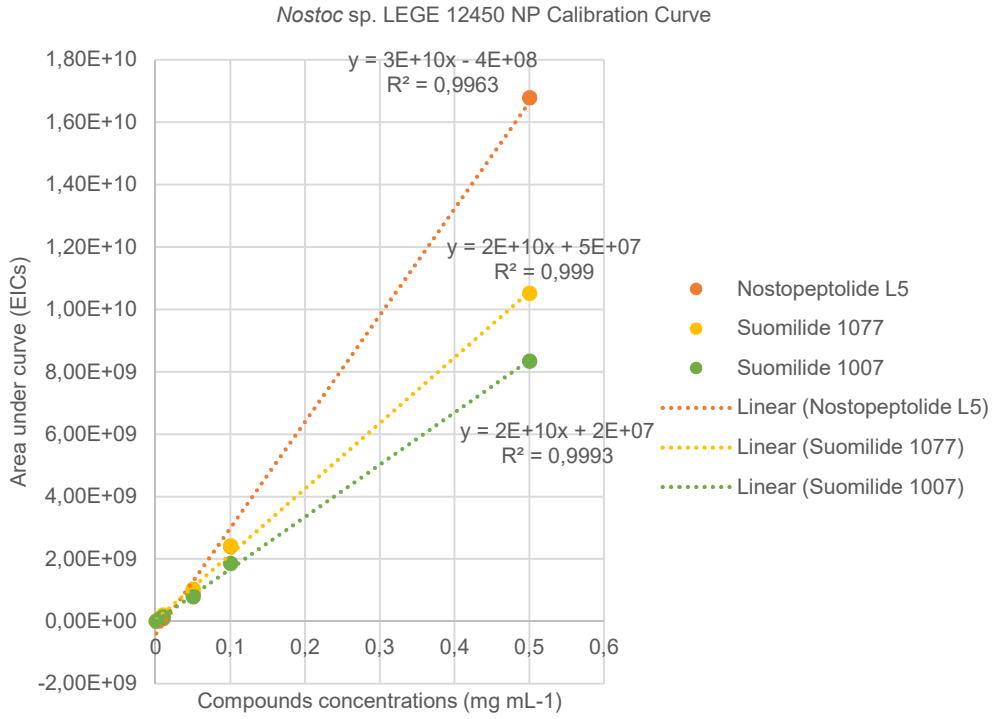


Figure S4 - LC-HRESIMS (positive mode) calibration curves (using standards of **1**-**3**) used for estimating the amount of metabolites produced by *Nostoc* sp. LEGE 12450 in the organic extract used for isolation.

Table S2 – LC-HRESIMS data for Nostopeptolide L5 (**1**) and suomilides 1077/1007 (**2**,**3**) standards used for quantification of the compounds in the cyanobacterial extract/biomass. Also shown are the data in the organic extract.

Concentration	**1 (EIC)**	**2 (EIC)**	**3 (EIC)**
mg mL^{-1}	**Area under curve**	**Area under curve**	**Area under curve**
0.5	1.68E+10	1.05E+10	8.35E+09
0.1	2.40E+09	2.42E+09	1.86E+09
0.05	8.89E+08	1.04E+09	7.90E+08
0.01	9.93E+07	2.06E+08	1.47E+08
0.005	5.49E+07	9.93E+07	7.61E+07
0.001	8.74E+06	1.90E+07	7.67E+07
Crude Extract (1 mg mL^{-1})	3.62E+08	4.33E+08	1.22E+08
g of compound per gram of extract	0.025	0.024	0.007
g of compound per gram of biomass	0.441	0.419	0.123

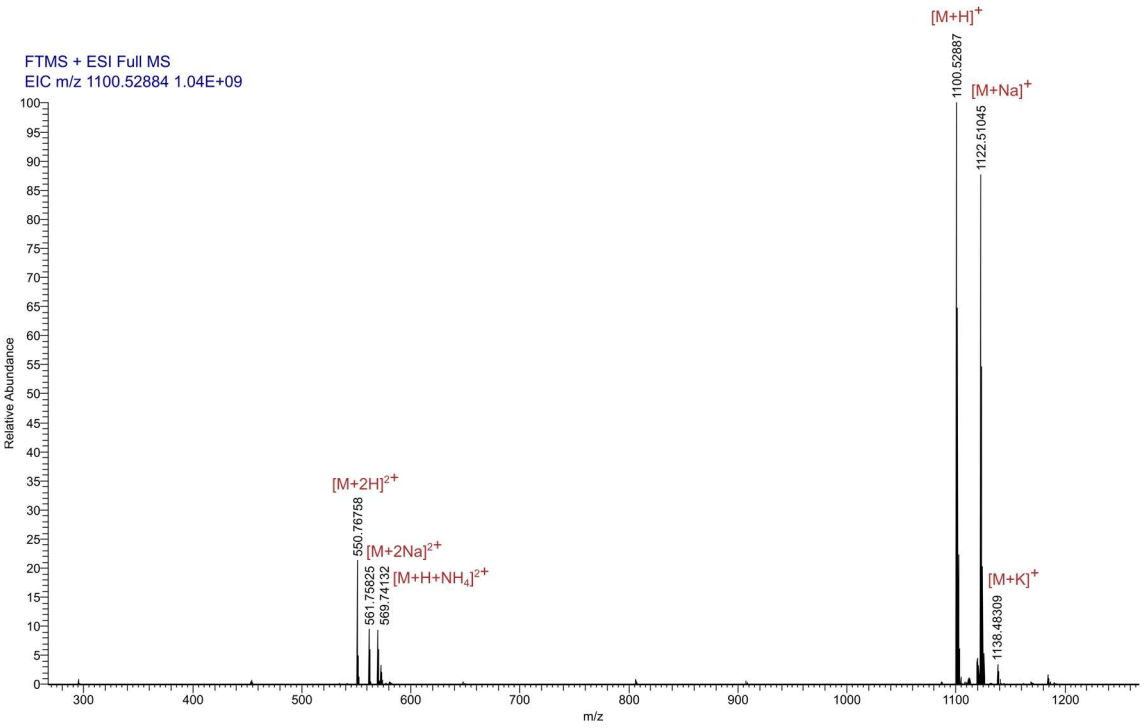


Figure S5 - HRESIMS spectrum of **1**.

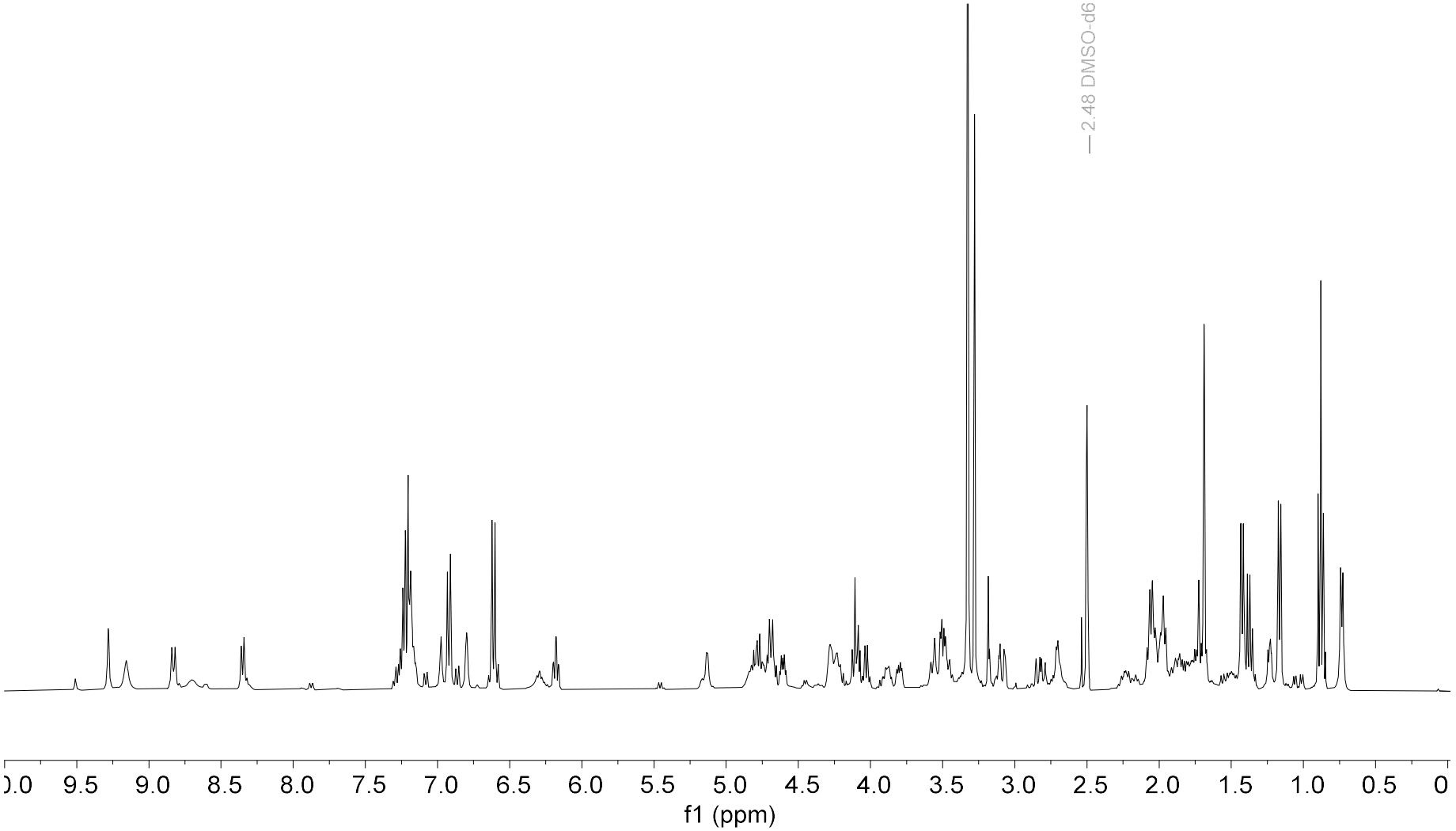


Figure S6 - ^{1}H NMR (DMSO-d_6, 400 MHz) spectrum of compound **1**.

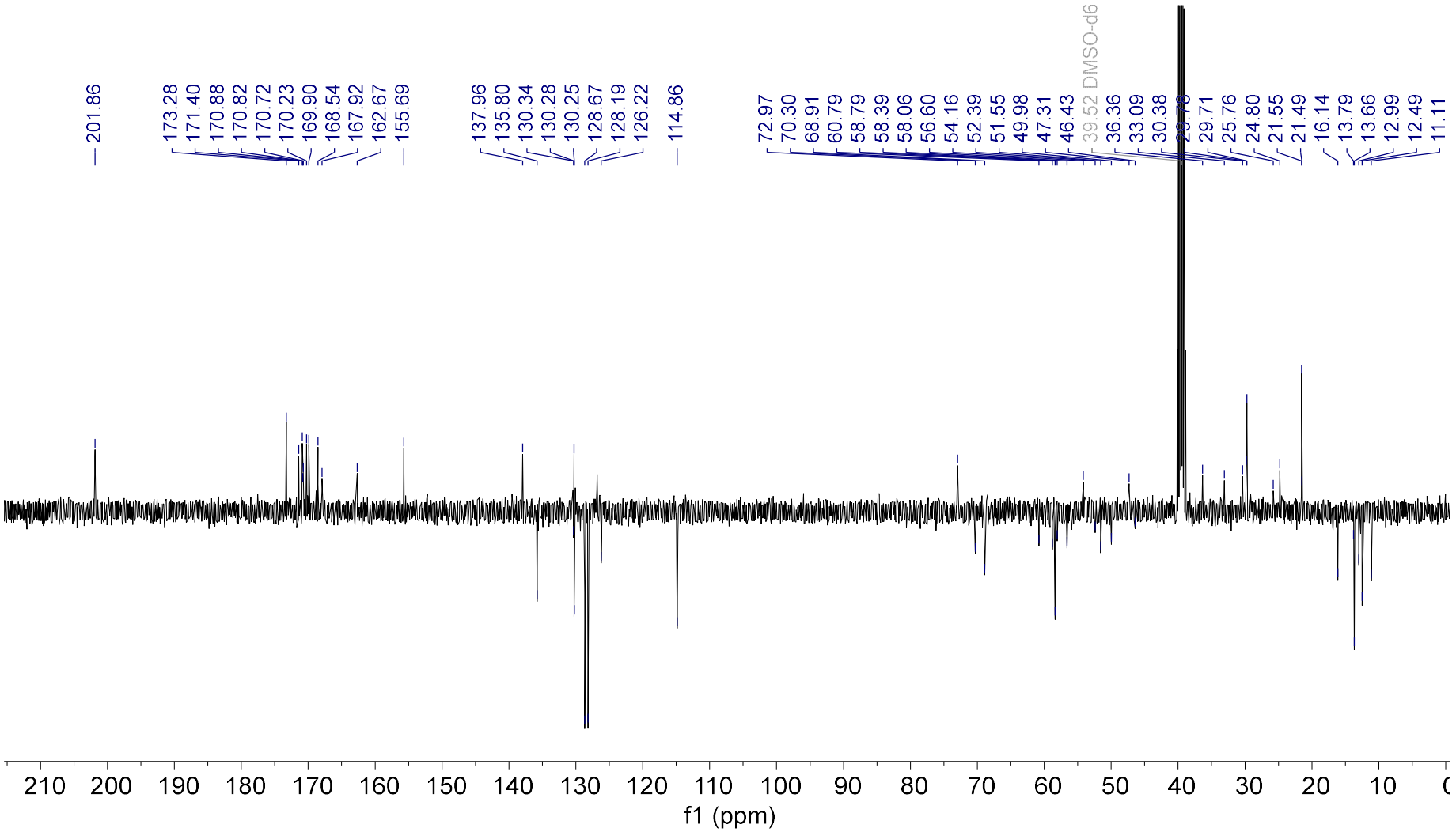


Figure S7 - ^{13}C NMR (APT, DMSO-d_6, 100 MHz) spectrum of compound **1**.

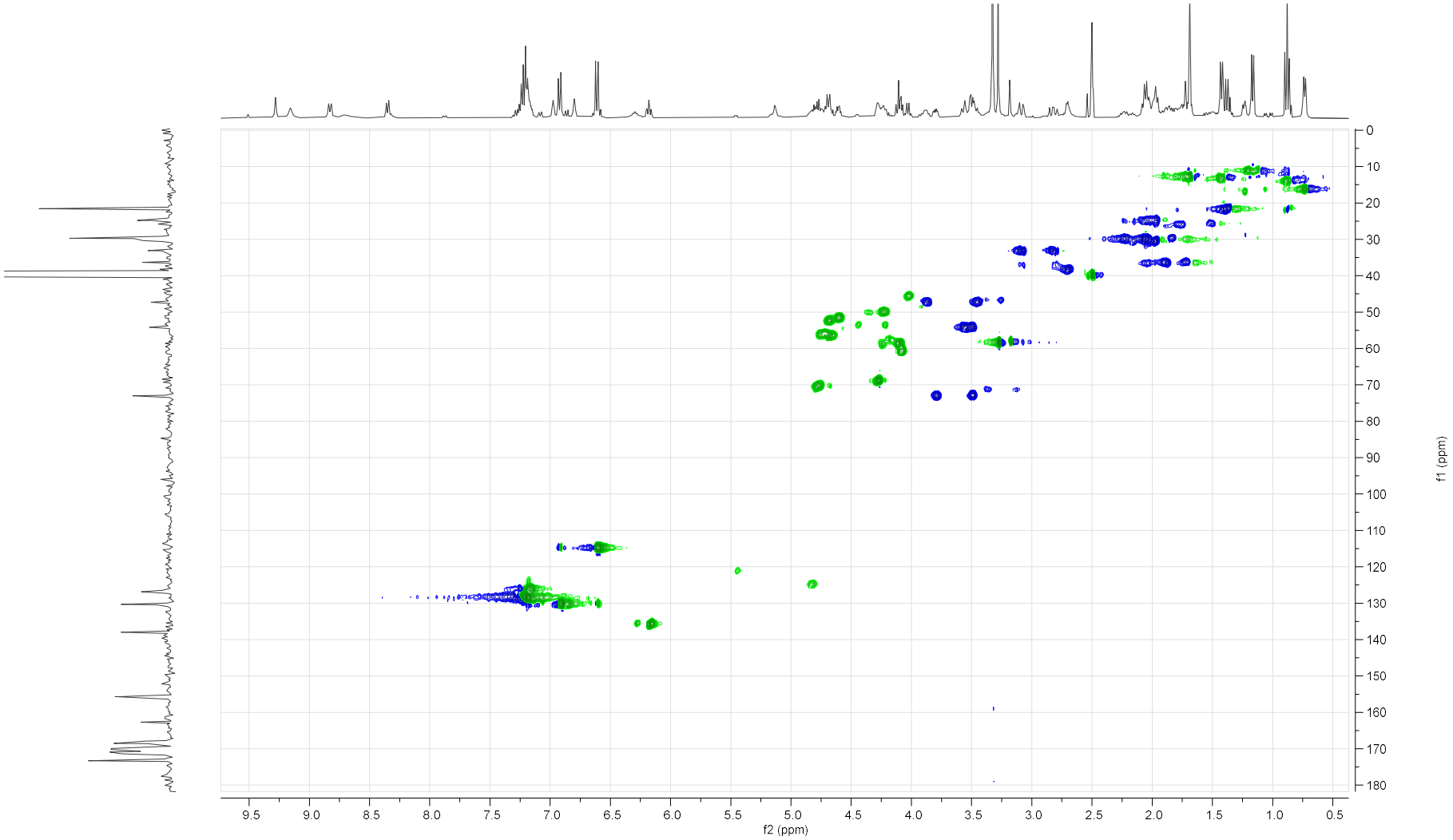


Figure S8 - HSQC (DMSO-d_6, 400 MHz) spectrum of compound **1**.

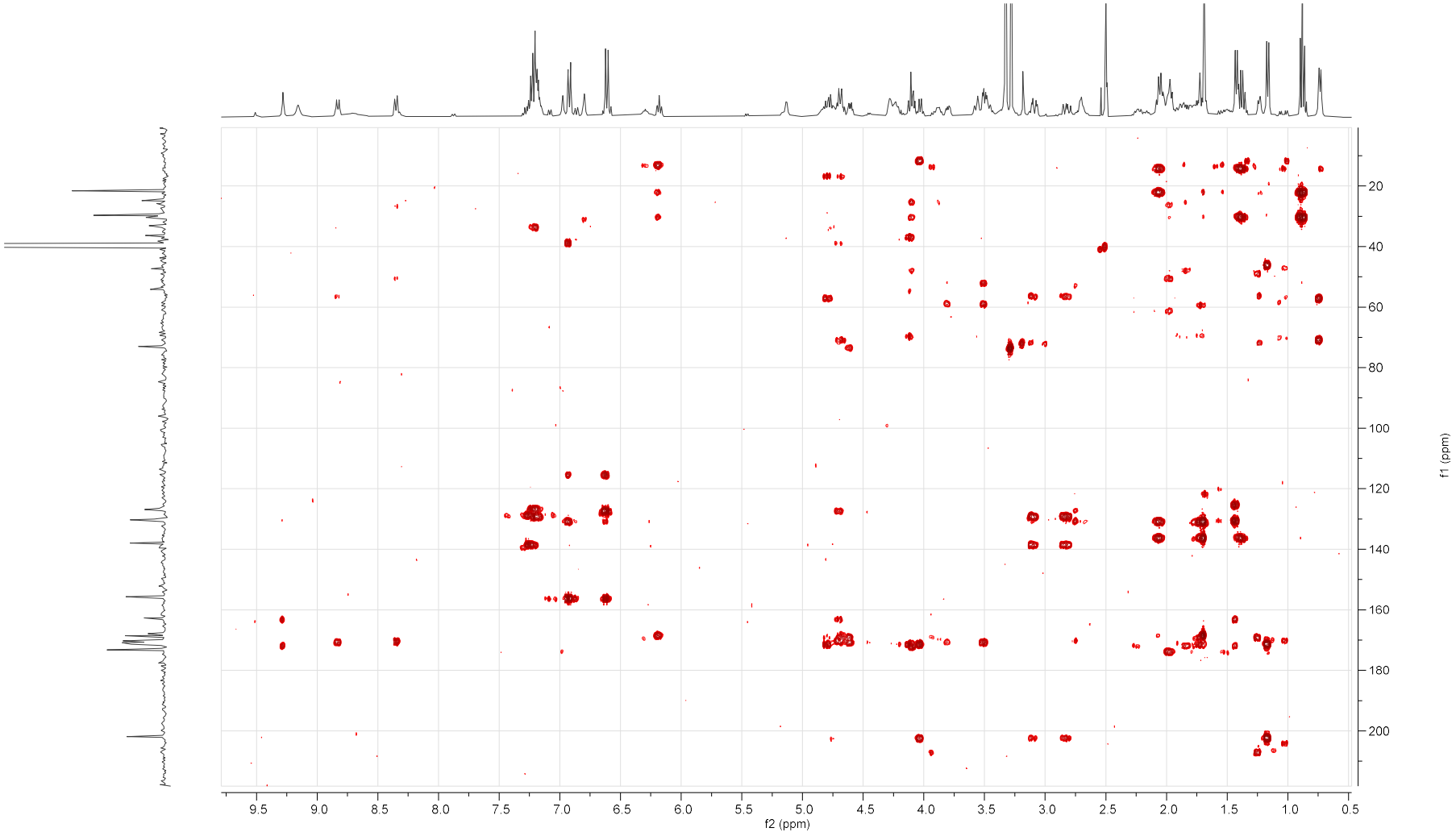


Figure S9 - HMBC (DMSO-d_6, 400 MHz) spectrum of compound **1**.

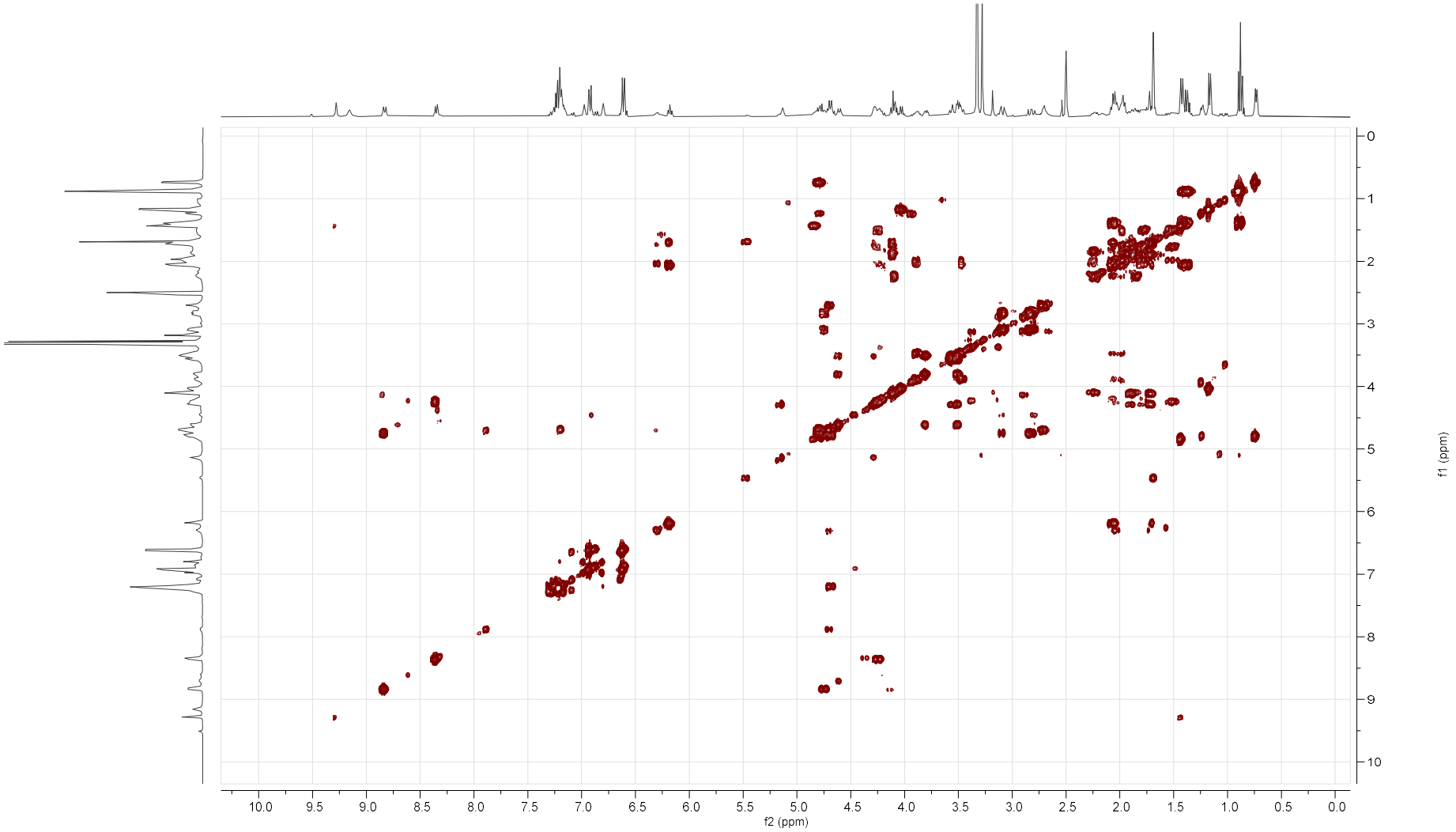


Figure S10 - COSY (DMSO-d_6, 400 MHz) spectrum of compound **1**.

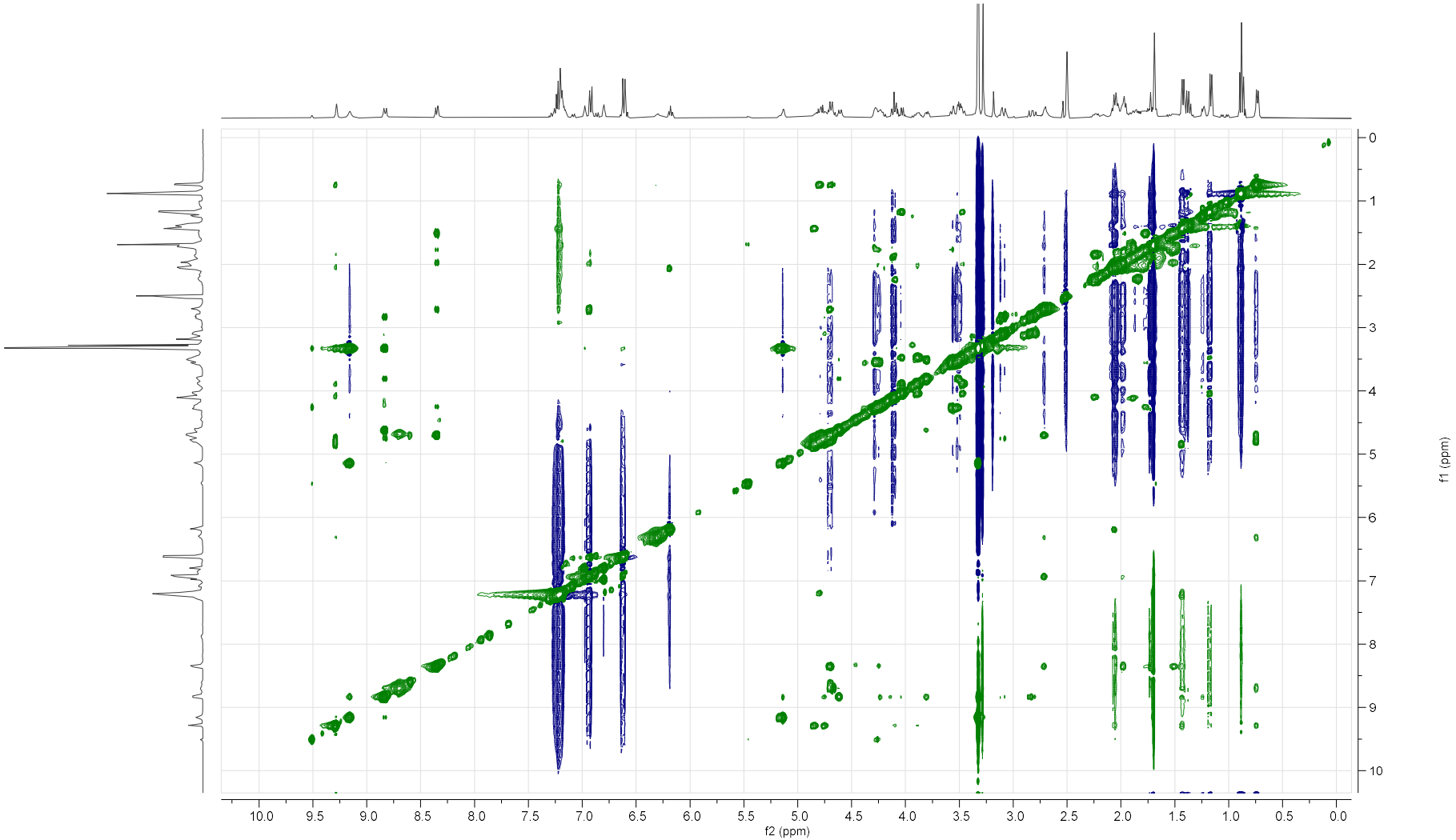


Figure S11 - NOESY (DMSO-d_6, 400 MHz) spectrum of compound **1**.

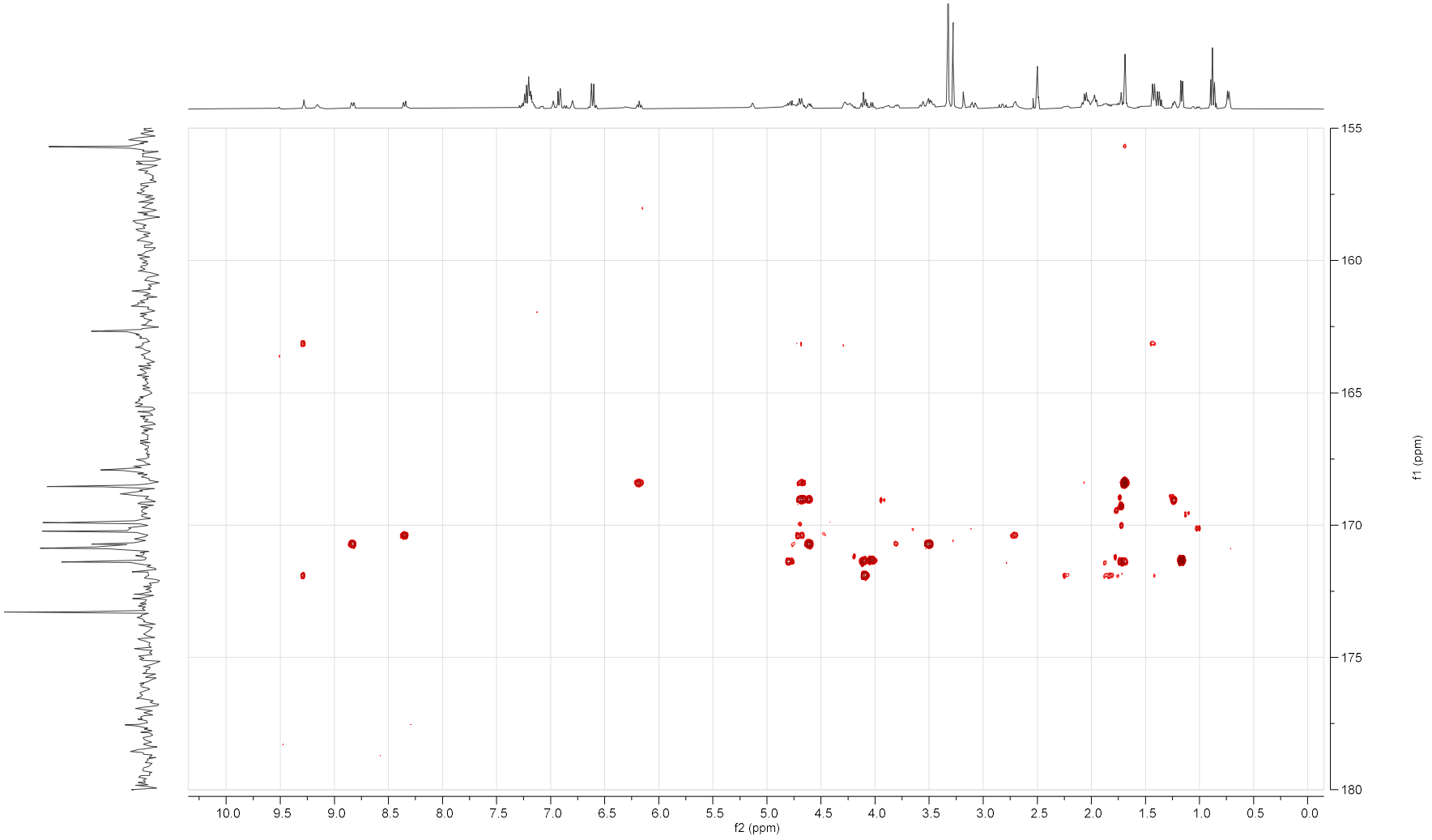


Figure S12 - Band-selective HMBC (DMSO-d_6, 600 MHz) spectrum of compound **1**.

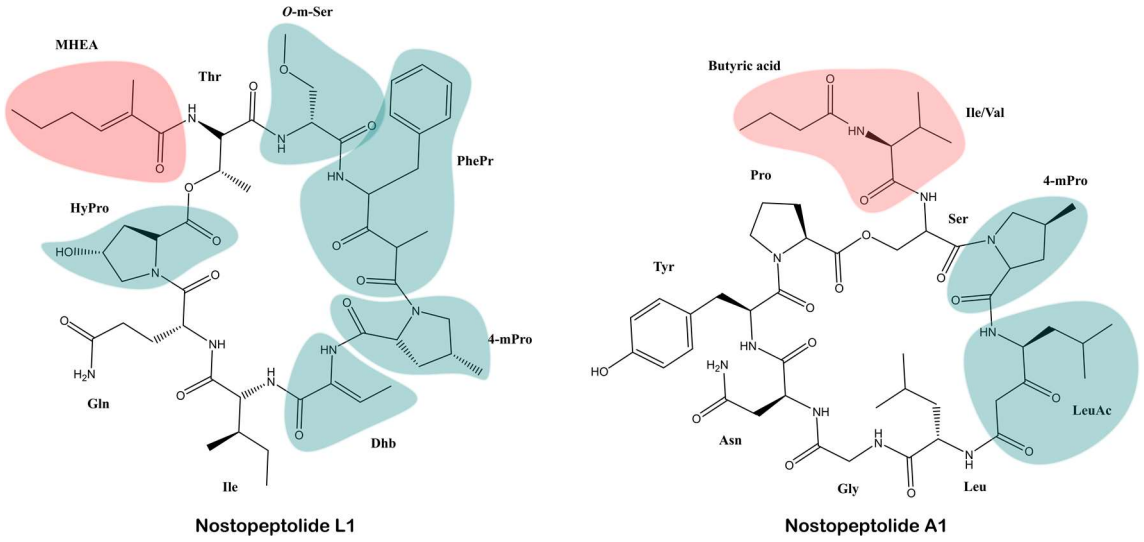


Figure S13 - Structures of nostopeptolide L1 isolated from *Nostoc* sp. UK2aImI, having an MHEA moiety connected to eight amino acids [Thr-*O*meSer-PhePr-4mePro-Dhb-Ile-Gln-Hyp]; and nostopeptolide A1, isolated from *Nostoc* sp. GSV224 and *Nostoc punctiforme* PCC 73102, showing the butyric acid unit and the eight amino acids (Thr-Ser-4mePro-LeuAc-Leu-Gly-Asn-Tyr-Pro].

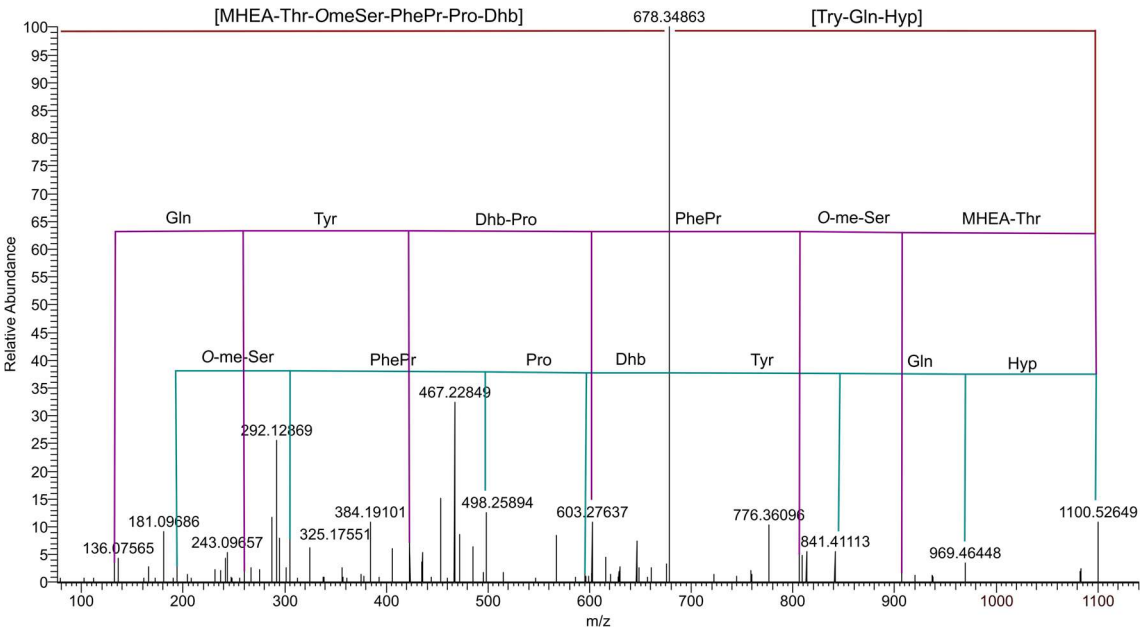


Figure S14 - HRESIMS/MS spectrum of **1**, highlighting key fragmentations supporting the NMR-based residue sequence proposal.

Table S3 - ^{1}H and ^{13}C NMR Data comparison of nostopeptolide L5 (**1**) and nostopeptolide L1 in DMSO-*d*6.

Unit	no	Type	Carbon		δH[a] (*J* in Hz)	
			L5	L1	L5	L1
MHEA	1	C	167.9	168.0		
	2	C	130.3	130.2		
	3	CH_3	12.5	12.3	1.69, d (1.2)	1.70, s
	4	CH	135.8	135.6	6.18, dt (7.3, 1.2)	6.18, dt (7.3, 1.5)
	5	CH_2	29.8	29.4	2.06, m	2.06, q (7.3)
	6	CH_2	21.6	21.3	1.38, m	1.39, q (7.3)
	7	CH_3	13.8	13.5	0.88, t (7.3)	0.89, t (7.3)
Thr	8	CH	56.6	56.5	4.68, m	4.63, dd (9.3, 9.8)
	9	CH	70.3	70.2	4.79, m	4.82, (6.6, 9.8)
	10	CH_3	16.1	16.0	0.73, d (6.3)	0.69, d (6.6)
	11	C	168.5	168.6		
		OH				
		NH			7.88	7.34, d (9.3)
*O*MeSer	12	CH	51.6	51.6	4.61, dt (8.6, 4.2)	4.56, ddd (5.6)
	13	CH_2	73.0	71.9	3.50, dd (9.5)	3.58
					3.80, dd (9.5)	3.70, dd (5.9, 9.8)
	14	CH_3	58.4	58.1	3.28, s	3.29
		NH			8.70, br	8.60, d (7.8)
	15	C	170.2	170.4		
PhePr	16	CH	55.9	55.4	4.74, m	4.75
	17	CH_2	33.1	32.7	2.82, dd (10.2, 14.8)	2.83, dd (10.7, 15.1)
					3.09, dd (3.8, 14.8)	3.09, dd (3.4, 15.1)
	18	C	138.0	137.6		
	19/23	CH	128.2	128.0	7.21, m	7.22
	20/22	CH	128.7	128.4	7.23, m	7.20
	21	CH	126.2	126.0	7.17, m	7.17
	24	C	201.9	201.8		
	25	CH	45.5	45.2	4.03, d (6.8)	3.95, q (6.8)
	26	CH_3	11.1	11.1	1.17, d (6.9)	1.17, d (6.8)
	27	C	170.8	170.5		
		NH			8.83, d (8.9)	8.98, d (9.3)
Pro/4mePro	28	CH_2	24.8	54.0	2.04, m	3.63
					1.98, m	3.26, (9.8)
	29	CH_2 (CH) (CH_3)	47.3	32.6 16.8	3.88, m 3.46, m	2.3 1.18, d (6.4)
	30	CH_2	29.7	37.9	2.24, m	2.41
					1.83, m	1.44
	31	CH	60.8	61.1	4.09, m	4.02, t (8.5)
	32	C	171.4	171.1		
Dhb	33	C	162.7	162.6		
	34	CH	124.9	124.9	4.83, m	4.70, q (7.3)
	35	CH_3	13.0	12.7	1.43, d (7.2 Hz)	1.36, d (7.3)
	36	C	130.3	130.0		
		NH			9.28, s	9.37
Tyr/Ile	37	CH	52.4	54.6	4.69, m	4.39, dd (9.4, 7.6)
	38	CH_2 (CH)	38.2	38.0	2.71, m	1.45
	39	C (CH_3)	126.8	15.2		0.81, d (6.8)
	40/44	CH (CH_2)	130.2	23.7	6.92, d (8.5)	0.96/1.50
	41/43	CH (CH_3)	114.9	11.0	6.61, d (8.5)	0.80, d (7.3)
	42	C	155.7			

	45	C	169.9	170.3		
		OH			9.16, s	
		NH			6.30, s	6.07, d (9.4)
Gln	46	CH	50.0	49.6	4.24, m	4.33
	47	CH_2	25.8	25.3	1.50, m	1.59
					1.76, m	1.83
	48	CH_2	30.4	30.2	1.97, m	2.16, t (7.3)
	49	C	173.3	173.0		
		Nha			6.80, s	6.71, 7.20, s
		NHb			6.98, s	
		NH			8.35, d (8.0)	8.45, d (7.8)
	50		170.7	170.7		
Hyp	51	CH_2	54.2	54.0	3.54, m	3.54
						3.60
	52	CH	68.9	68.8	4.28, m	4.29
	53	CH_2	36.4	36.2	1.90, m	1.91
					1.73, m	1.75
	54	CH	58.8	58.6	4.11, s	4.13, t (8.0)
	55	C	170.9	170.7		
		OH			5.13, d (2.6)	5.12

[a]from HSQC.

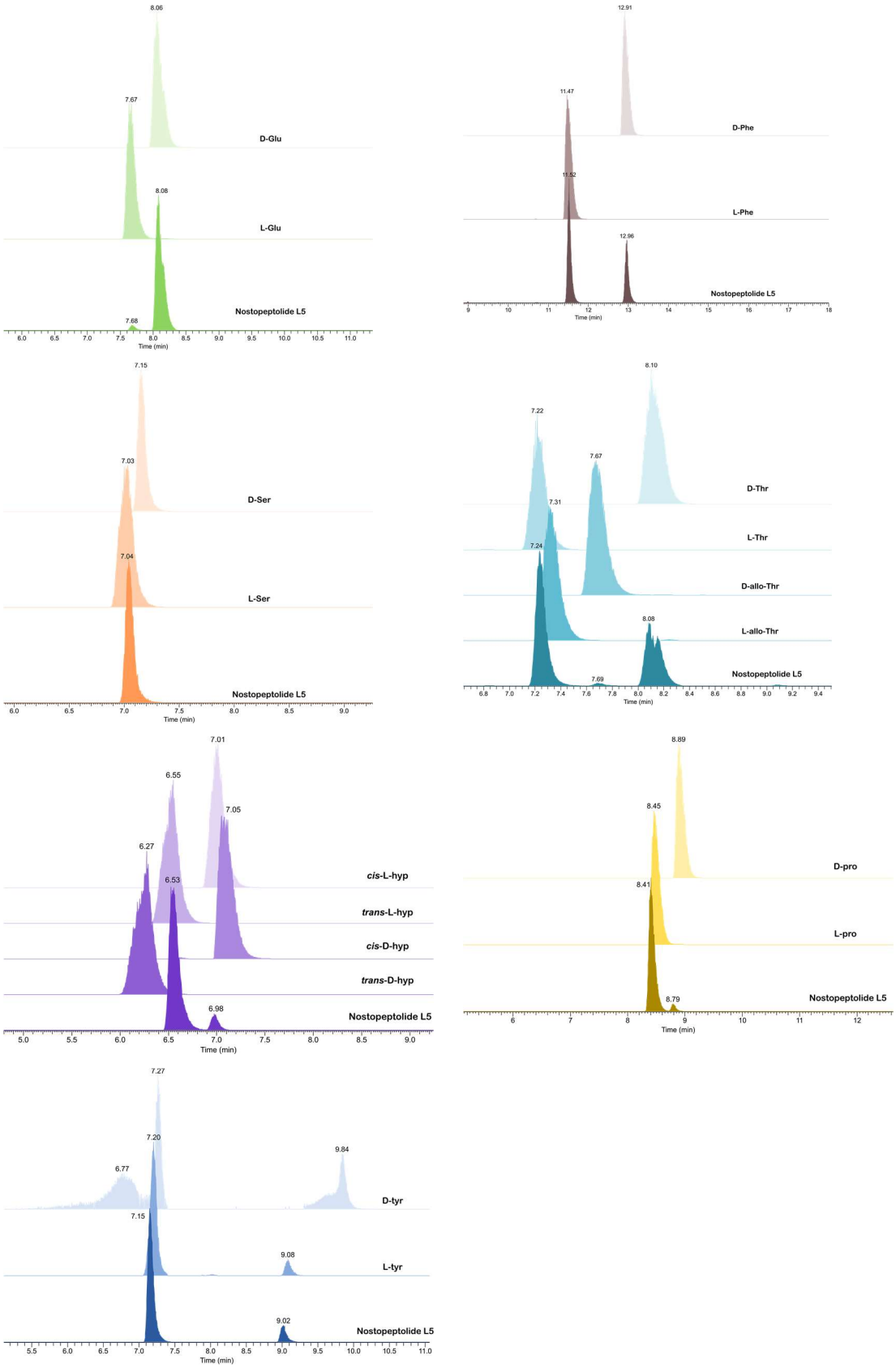


Figure S15 - LC-HRESIMS analysis of the Marfey's derivatization products. EICs of $[M+H]^+$ ions for Glu (*m/z* 400.11), Phe (*m/z* 418.14), Ser (*m/z* 358.10), Hyp (*m/z* 384.12), Pro (*m/z* 368.12), Tyr (*m/z* 434.13) Thr and allo-Thr (*m/z* 372.11) are shown for the standards and hydrolysate of nostopeptolide L5.

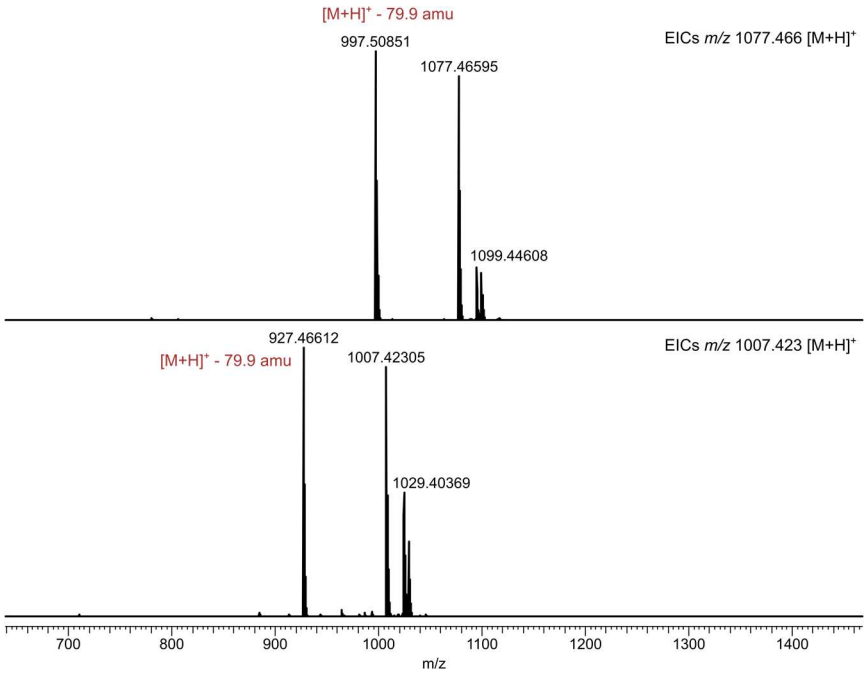


Figure S16 - Comparison of the EIC from the compounds **2** and **3**, showing the loss of 79.95 amu (atomic mass units) for both compounds, prior isolation.

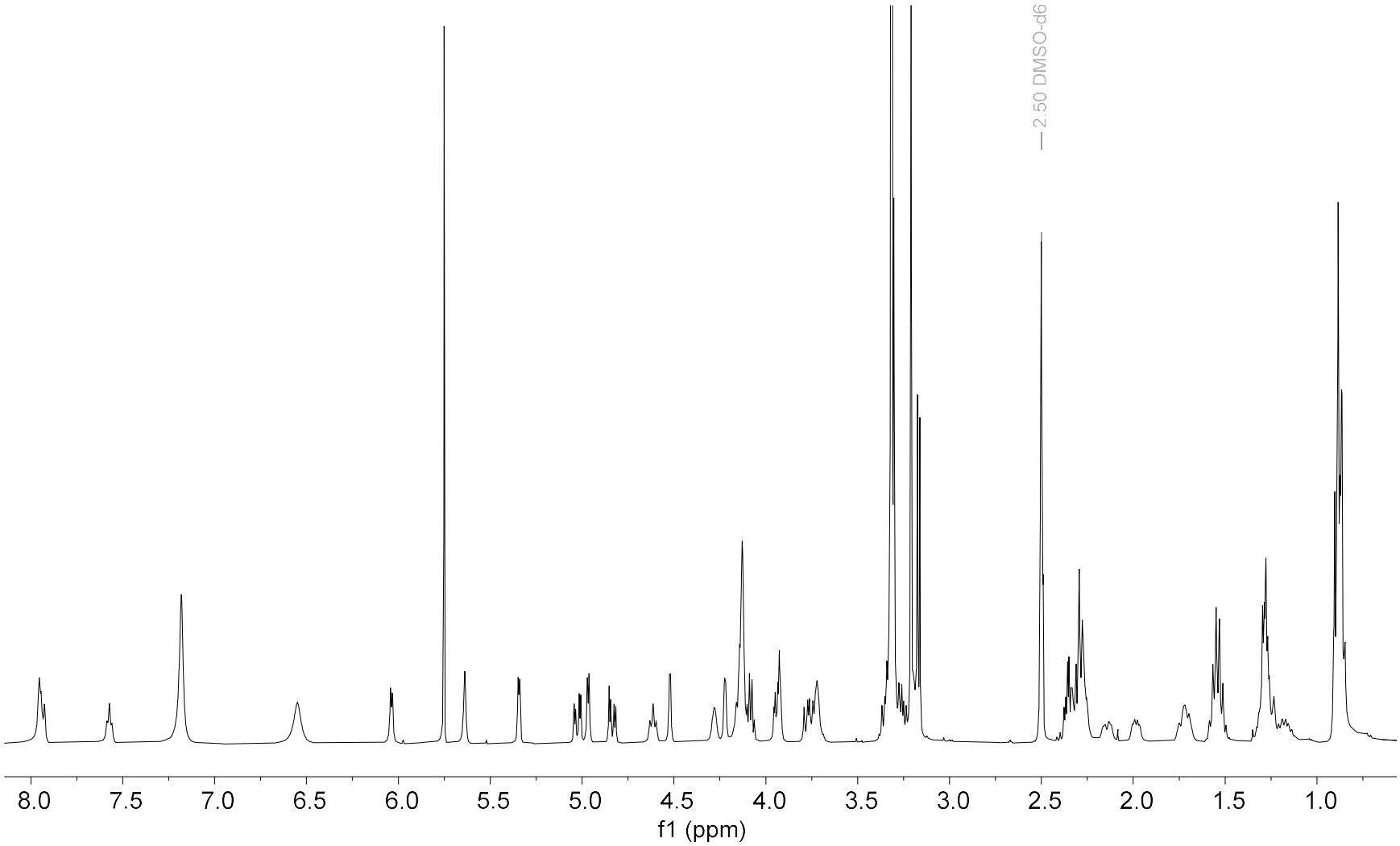


Figure S17 - ^{1}H NMR (DMSO-d_6, 100 MHz) spectrum of compound **2**.

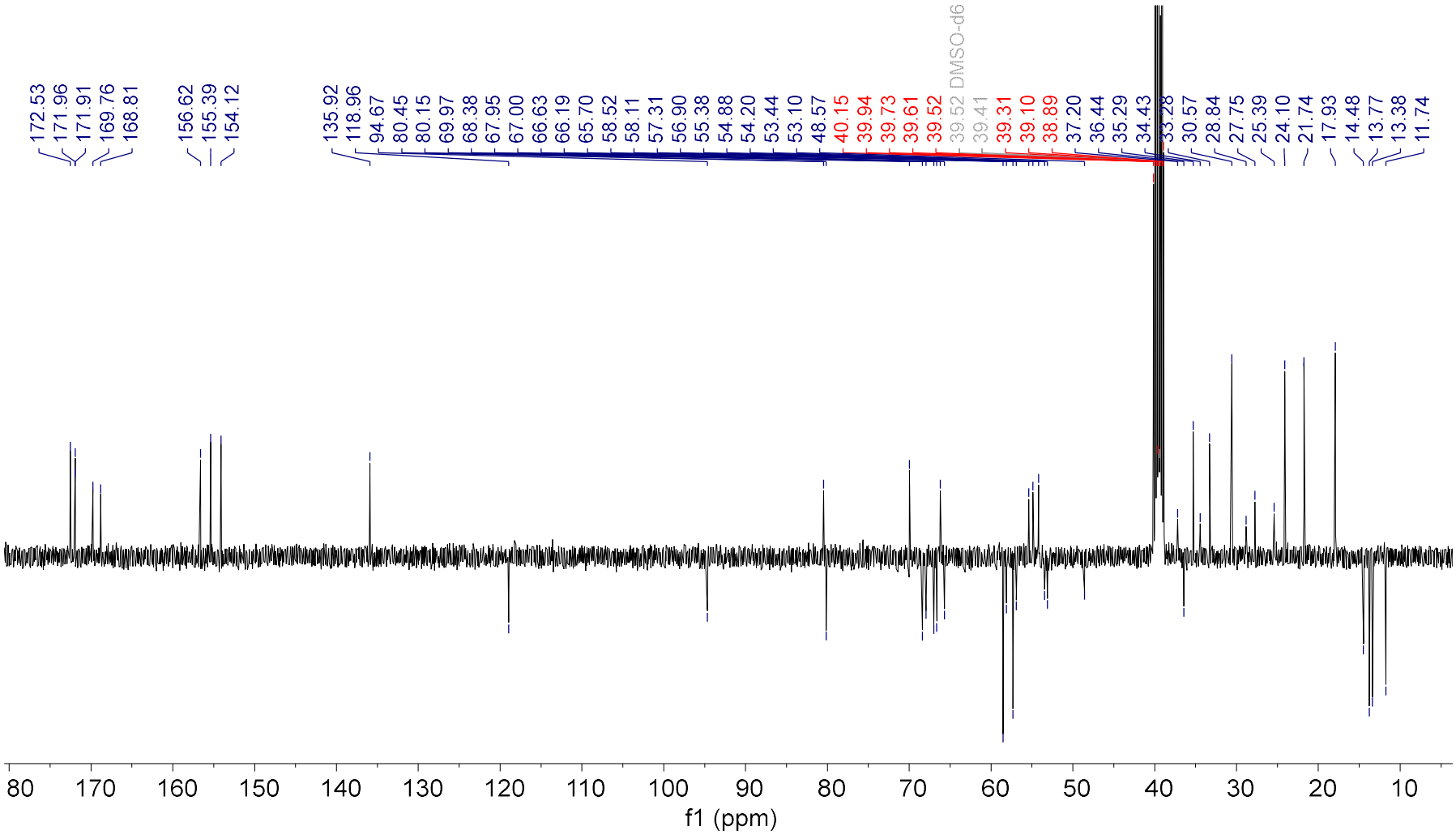


Figure S18 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **2**.

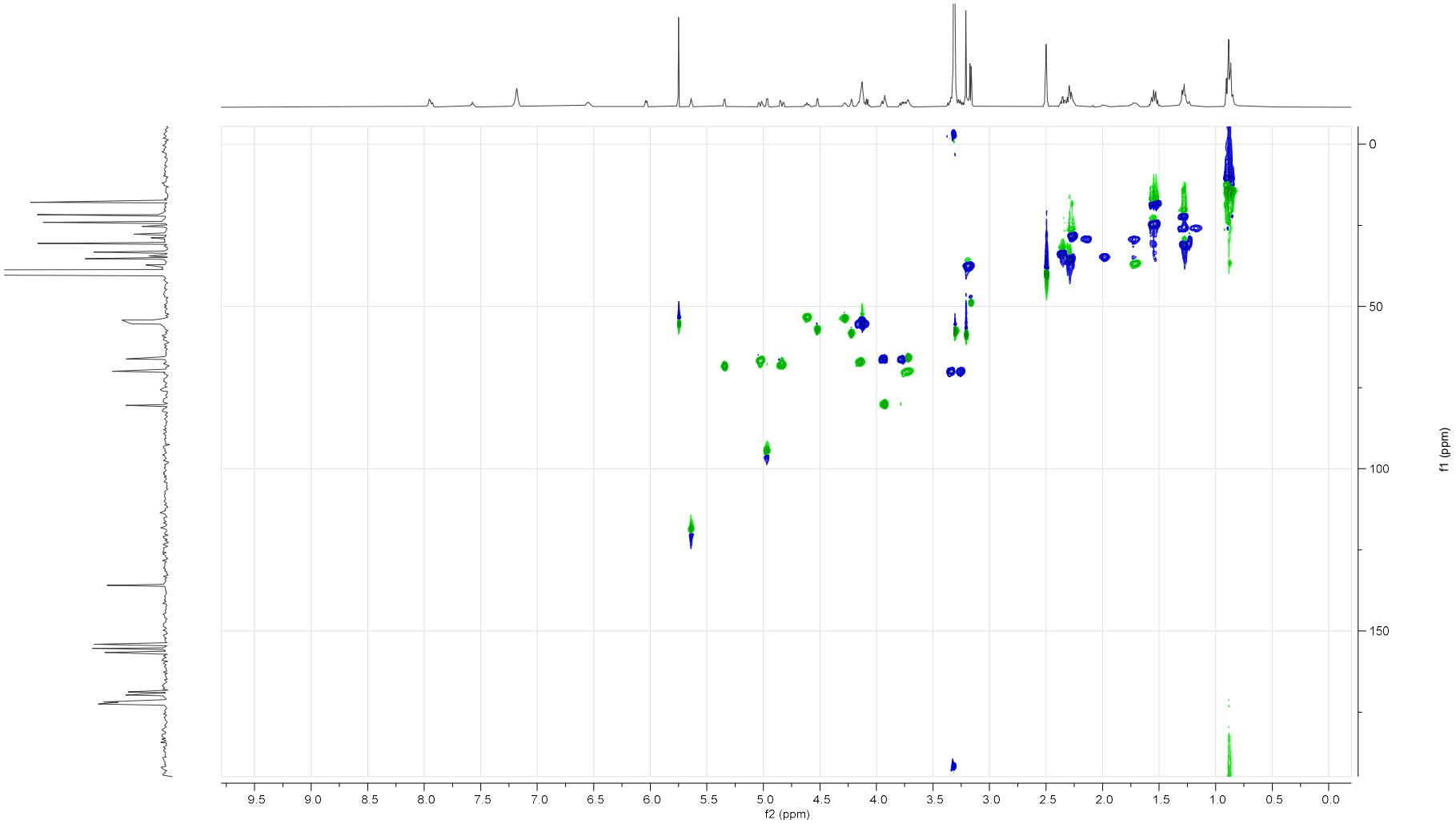


Figure S19 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **2**.

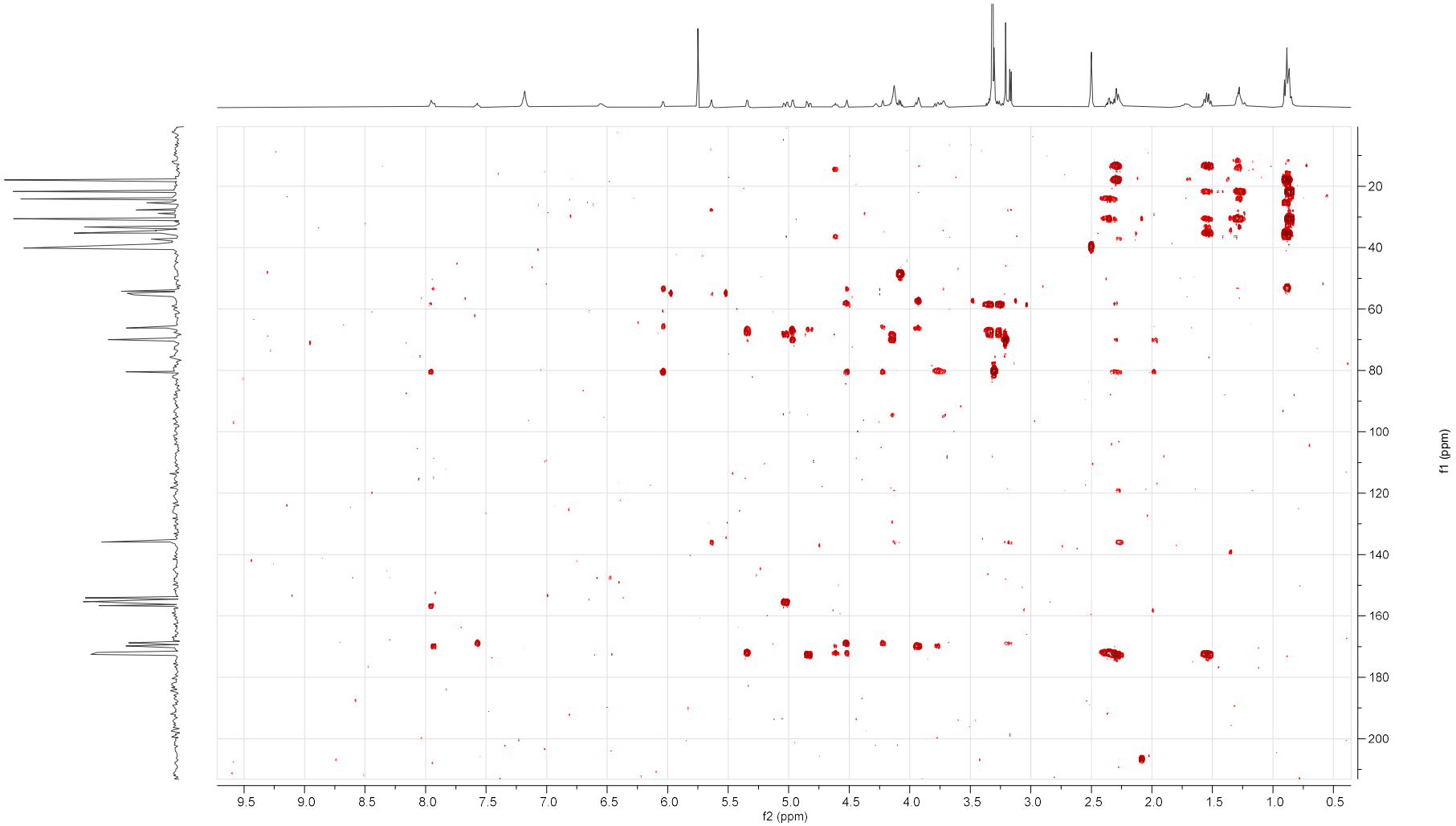


Figure S20 - HMBC (DMSO-d_6, 600 MHz) spectrum of compound **2**.

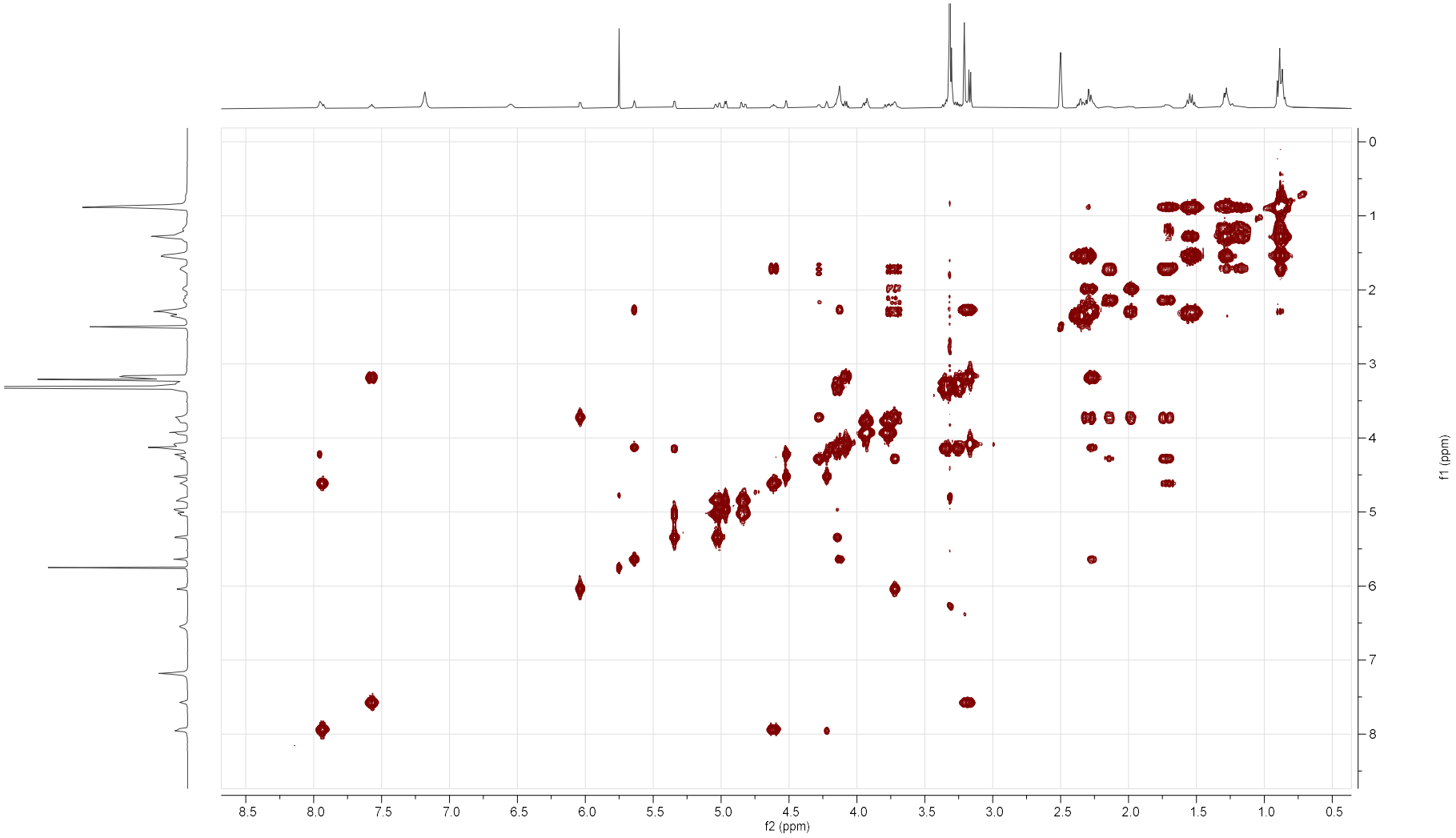


Figure S21 - COSY (DMSO-d_6, 600 MHz) spectrum of compound **2**.

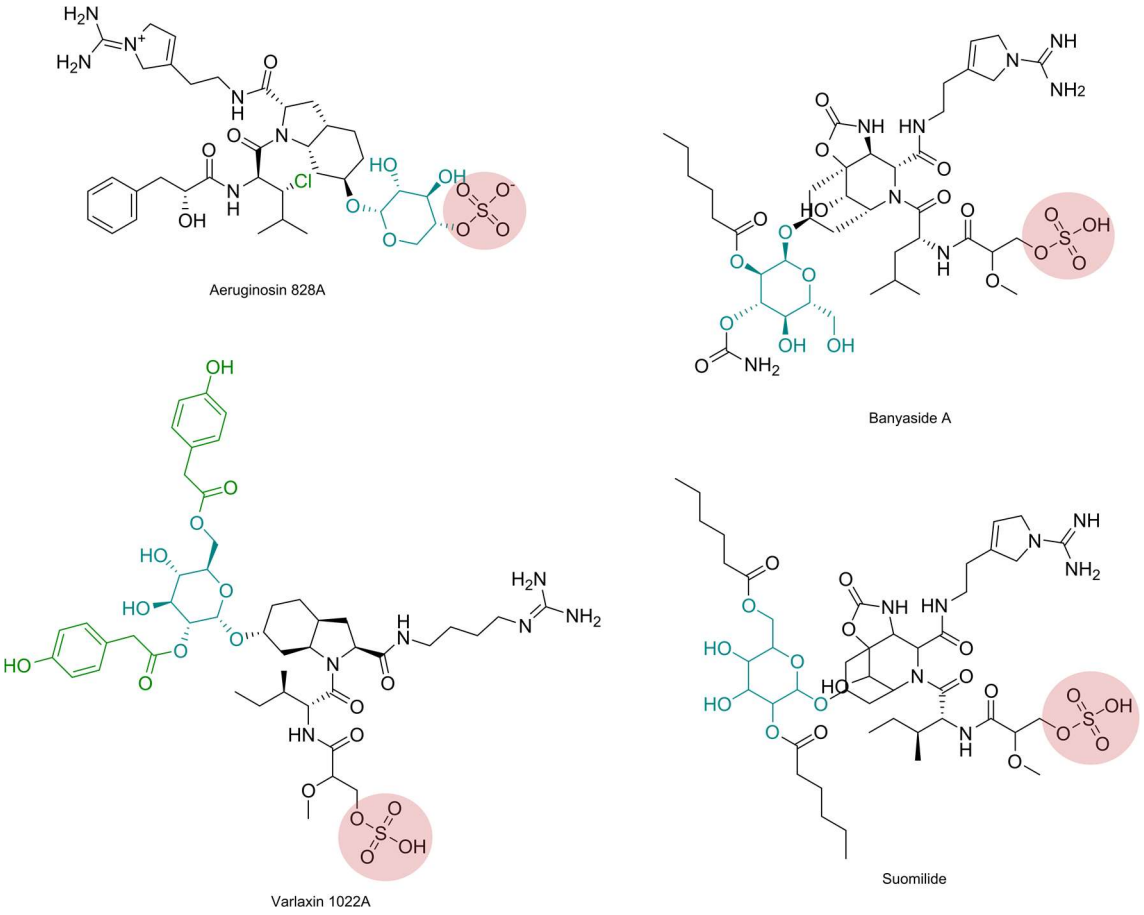


Figure S22 - NPAtlas search of cyanobacterial natural products, using the sulfate and glycosyl group as query. Four compound classes were retrieved and an example of each is represented. The sulfate group is highlighted by a red shade and the glycosyl unit drawn in blue. For aeruginosin 828A, the chlorine atom is highlighted in green as well as the Hpaa units of varlaxin 1022A.

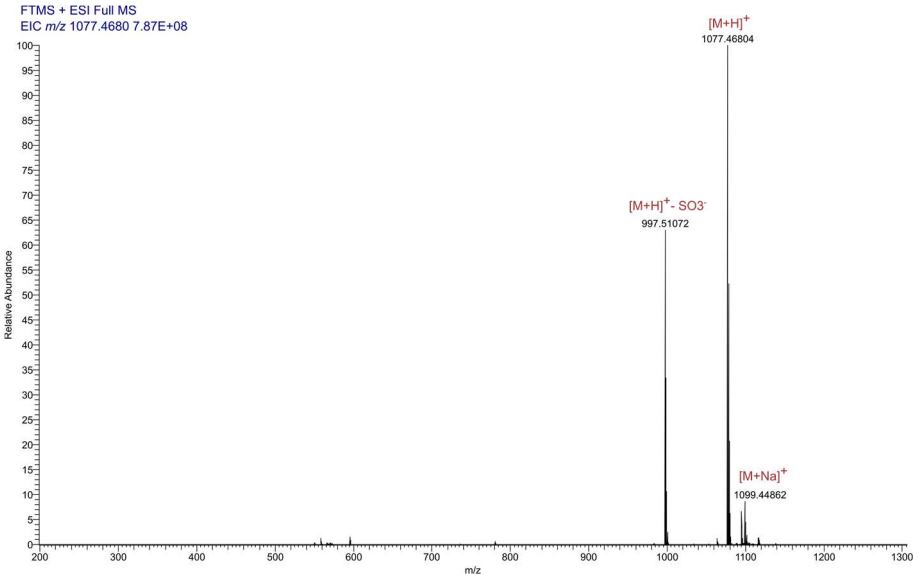


Figure S23 - HRESIMS spectrum of **2**.

Table S4 - Comparison of the NMR data of Suomilide and **2** and **3**, in DMSO-d_6. In bold are the signals assigned based on comparison with suomilide NMR data. Signals in red are only present in the suomilide structure.

		δC			**δH**		
Unit	**no.**	**Suomilide**	**1077**	**1007**	**Suomilide**	**1077**	**1007**
Ile-MgS	1	66.2	66.2	66.3	3.75	3.94/3.77	3.94/3.78
	2	80.0	80.2	80.1	3.94	3.93	3.96
	3	57.3	57.3	57.4	3.29	3.30	3.30
	4	169.9	169.8	170.0			
	5	53.3	53.1	53.2	4.57	4.61	4.60
	6	36.3	36.4	36.3	1.70	1.71	1.75
	7	14.5	14.5	14.6	0.87	0.88	0.89
	8	25.4	25.4	25.5	1.17	1.18	1.19
					1.29	1.29	1.30
	9	11.7	11.7	11.8	0.89	0.89	0.90
	10	172.0	171.9	171.8			
	1-NH				8.03	7.93	8.02
Abn	11	56.6	56.9	56.8	4.59	4.52	4.59
	12	53.5	53.4	53.5	4.23	4.28	4.30
	13	65.7	65.7	65.9	3.69	3.72	3.71
	14	28.7	28.8	28.8	1.68	1.72	1.82
					2.14	2.14	2.20
	15	69.8	70.1	69.7	3.62	3.73	3.72
	16	34.5	35.4	34.5	1.94	1.98	1.96
	17	80.4	80.5	80.6			
	18	156.6	156.6	156.7			
	19	58.1	58.1	58.2	4.27	4.22	4.26
	20	168.7	168.8	168.7			
	2-NH				7.95	7.95	8.14
	1-OH				6.02	6.04	6.03
Aaep	21	37.2	37.2	37.3	3.16	3.18	3.21
	22	27.7	27.8	27.8	2.23	2.27	2.29
	23	135.9	135.9	136.0			
	24	118.8	119.0	118.9	5.61	5.64	5.64
	25	54.2	54.9	54.1	4.11	4.13	4.11
	26	154.1	**154.1**	**154.3**			
	27	55.4	55.4	55.3	4.09	4.13	4.11
					4.11		
	3-NH				7.53	7.57	7.62
	4-NH				7.13	**7.14**	**7.41**
Glycosyl	1'	94.7	94.7	98.3	4.78	4.97	4.74
	2'	72.9	68.0	65.7	4.35	4.84	3.68
	3'	69.9	66.6	69.7	3.56	5.03	4.73
	4'	70.2	68.4	68.6	3.14	5.34	5.28

	5'	70.0	67.0	66.9	3.66	4.14	4.03
	5''	63.3	70.0	70.2	4.29	3.34	3.29
					4.02	3.26	3.22
	OHa				5.23		
	OHb				5.31		
	2'a	172.8	172.5				
	2'b	33.2	34.4		2.3	2.28	
	2'c	24.1	17.9		1.50	1.54	
	2'd	30.6	13.4		1.25	0.88	
	2'e	21.8			1.25		
	2'f	13.7			0.84		
	OHc						4.94
	3'a		155.4	156.1			
	NH_2					6.55	6.43
	4'a		172.0	172.0			
	4'b		33.3	33.4		2.35	2.31
	4'c		24.1	24.2		1.55	1.54
	4'd		30.6	30.6		1.27	1.27
	4'e		21.7	21.8		1.28	1.27
	4'f		13.8	13.8		0.86	0.86
	5'a	172.9	58.5	58.5		3.21	3.20
	5'b	33.4			2.3		
	5'c	24.1			1.5		
	5'd	30.6			1.25		
	5'e	21.8			1.25		
	5'f	13.8			0.84		

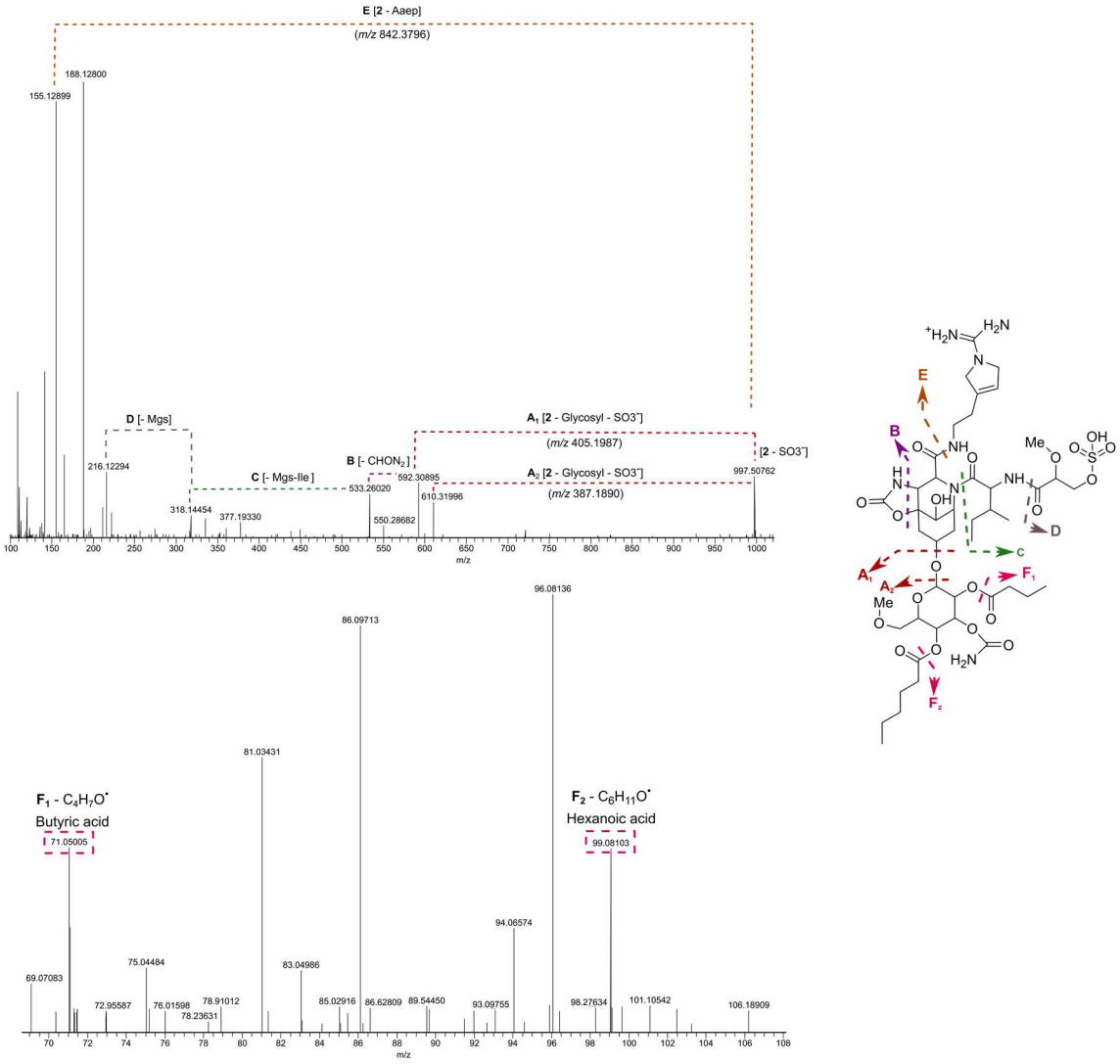


Figure S24 - HRESIMS/MS analysis of **2**. The main fragments are represented in the structure on the right side and are: **A** – loss of the sulphate group and sugar unit (*m/z* 592.309); **B** – loss of the oxazolidin-4-1 ring (*m/z* 59.00); **C** – loss of MgS-Ile moiety (*m/z* 215.12); **D** – loss of the Mgs, fragmenting from Ile (*m/z* 74.04); E – loss of the Aaep moiety (*m/z* 155.13) and $\mathbf{F_1}$ and $\mathbf{F_2}$ corresponding to the fragmentation of the C4 and C6 from the glycosyl unit, respectively.

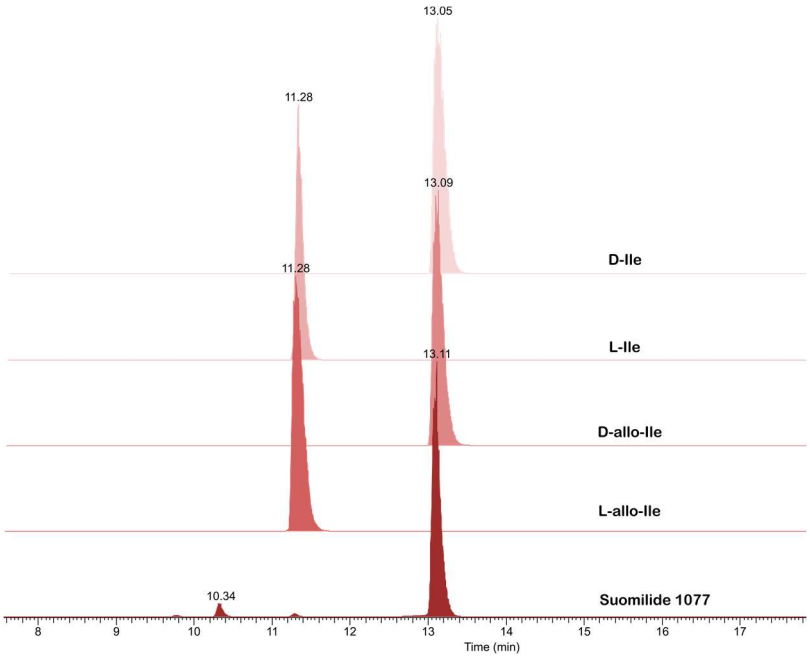


Figure S25 - LC-HRESIMS analysis of the Marfey's derivatization products. EIC of $[M+H]^+$ ion for isoleucine (*m/z* 284.15) is shown for the standards and hydrolysate of **2**.

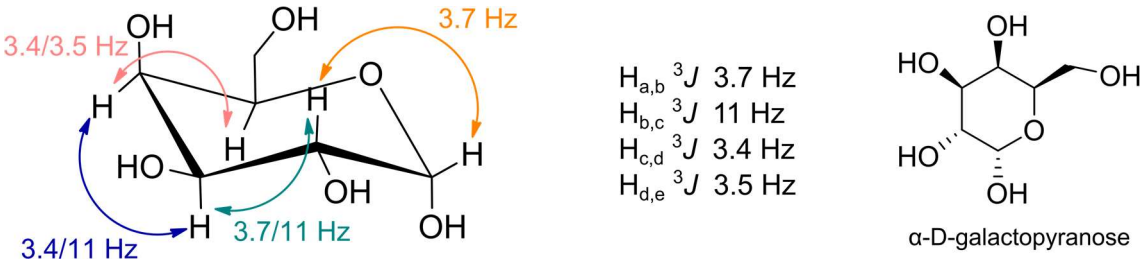


Figure S26 - Coupling constant analysis (^{1}H and COSY data, DMSO-d_6, 600 MHz) underscoring the proposed sugar configuration of compound **2**.

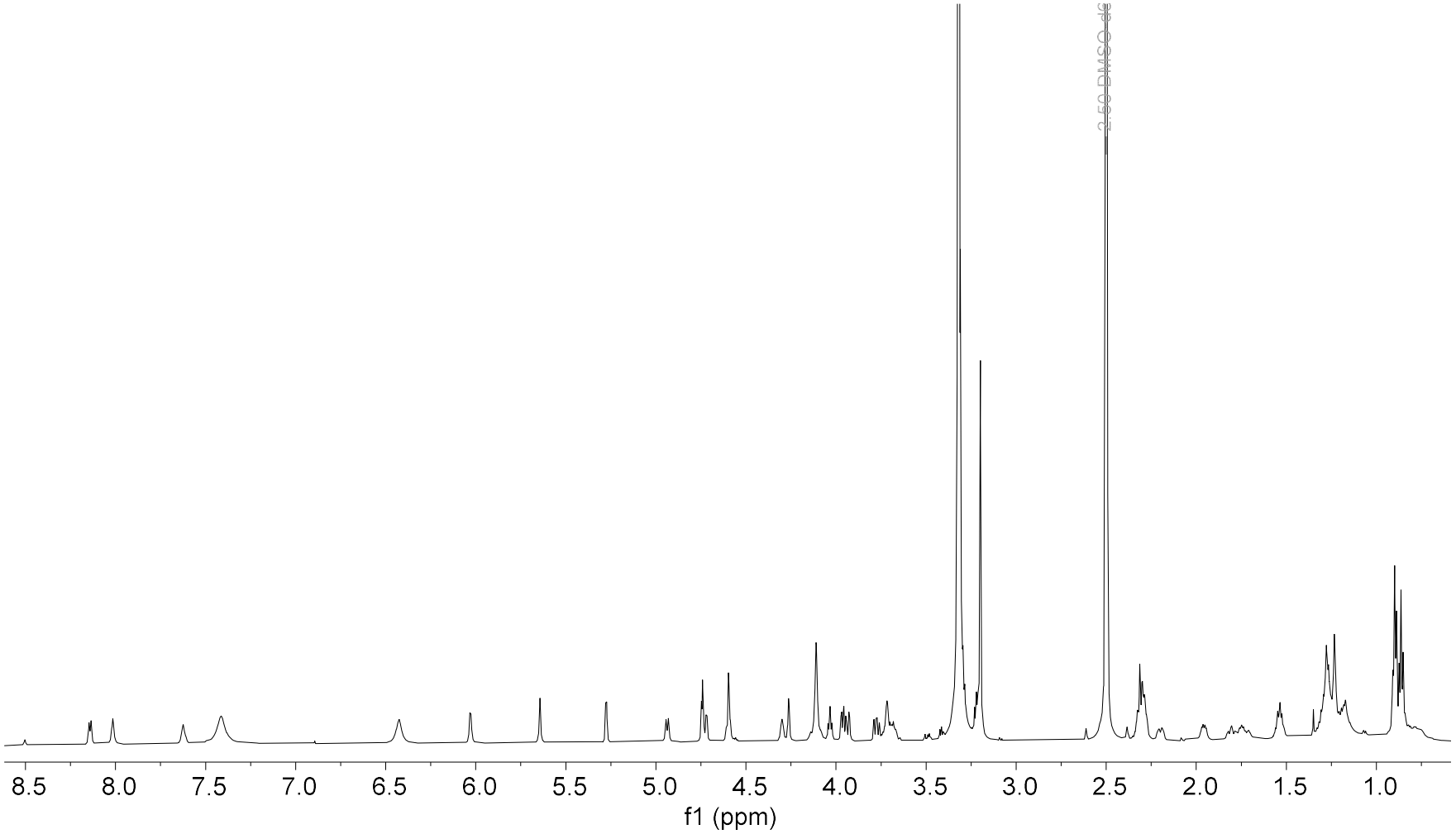


Figure S27 - ^{1}H NMR (DMSO-d_6, 150 MHz) spectrum of compound **3**.

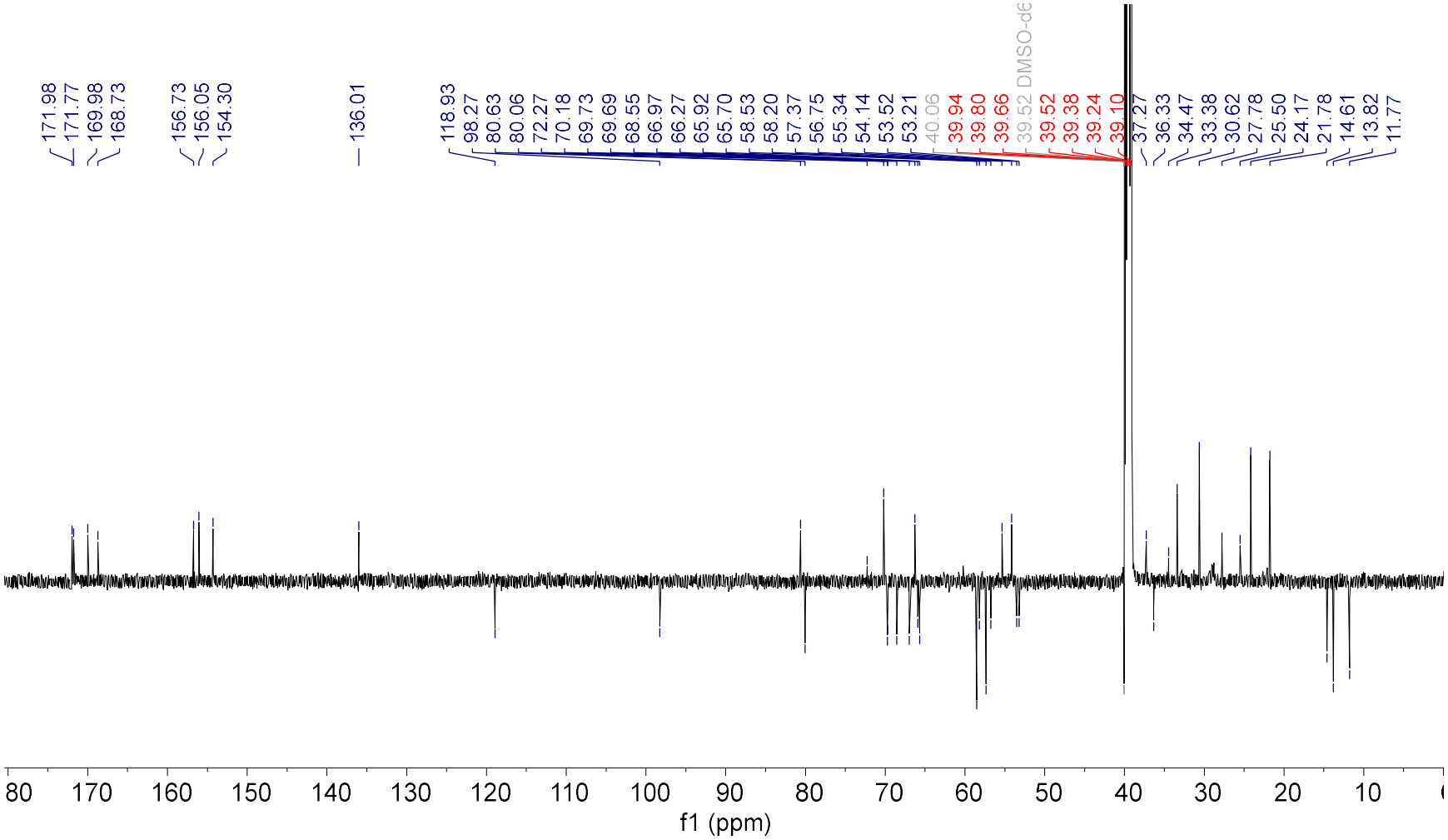


Figure S28 - ^{13}C NMR (APT, DMSO-d_6, 150 MHz) spectrum of compound **3**.

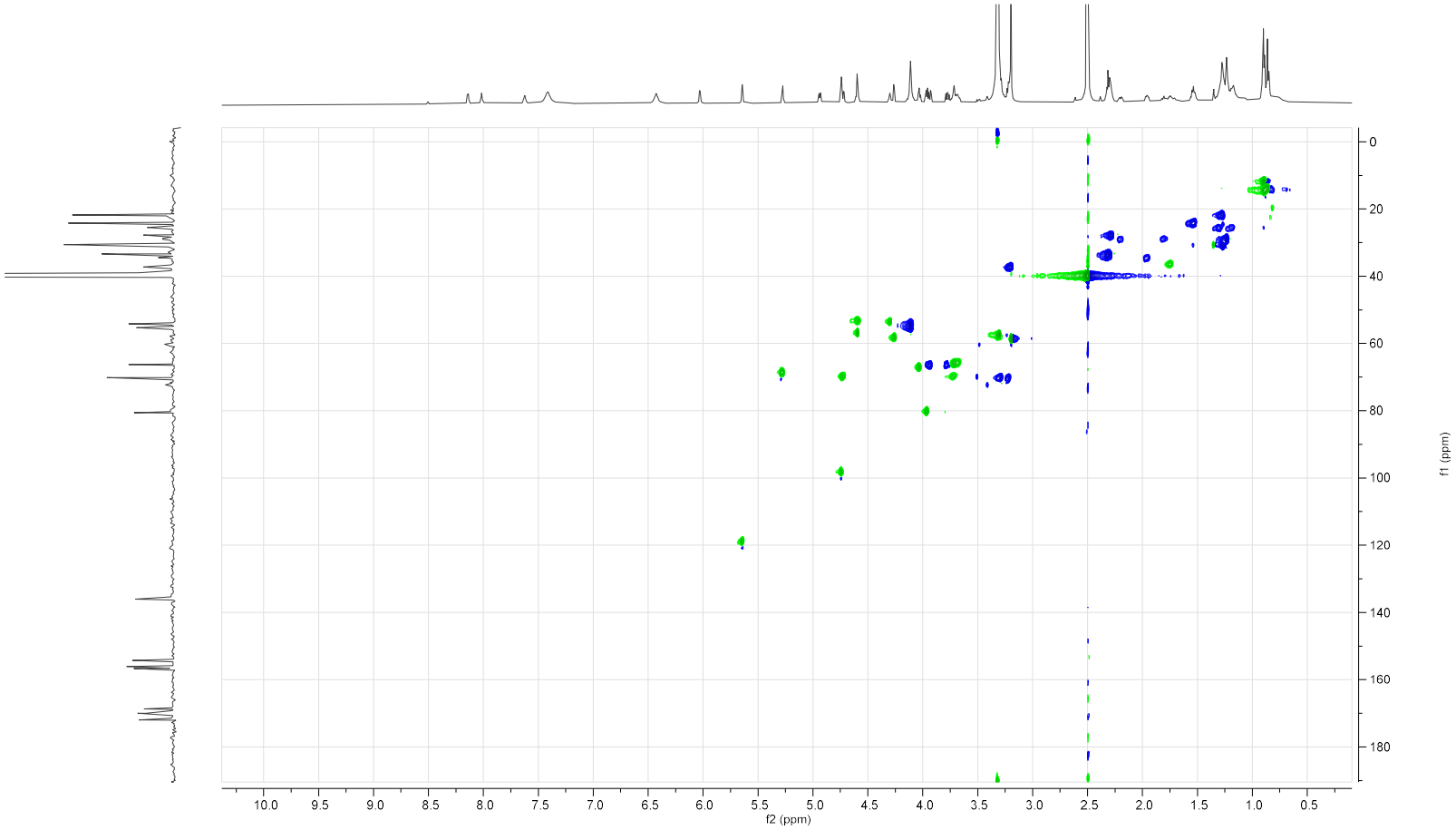


Figure S29 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **3**.

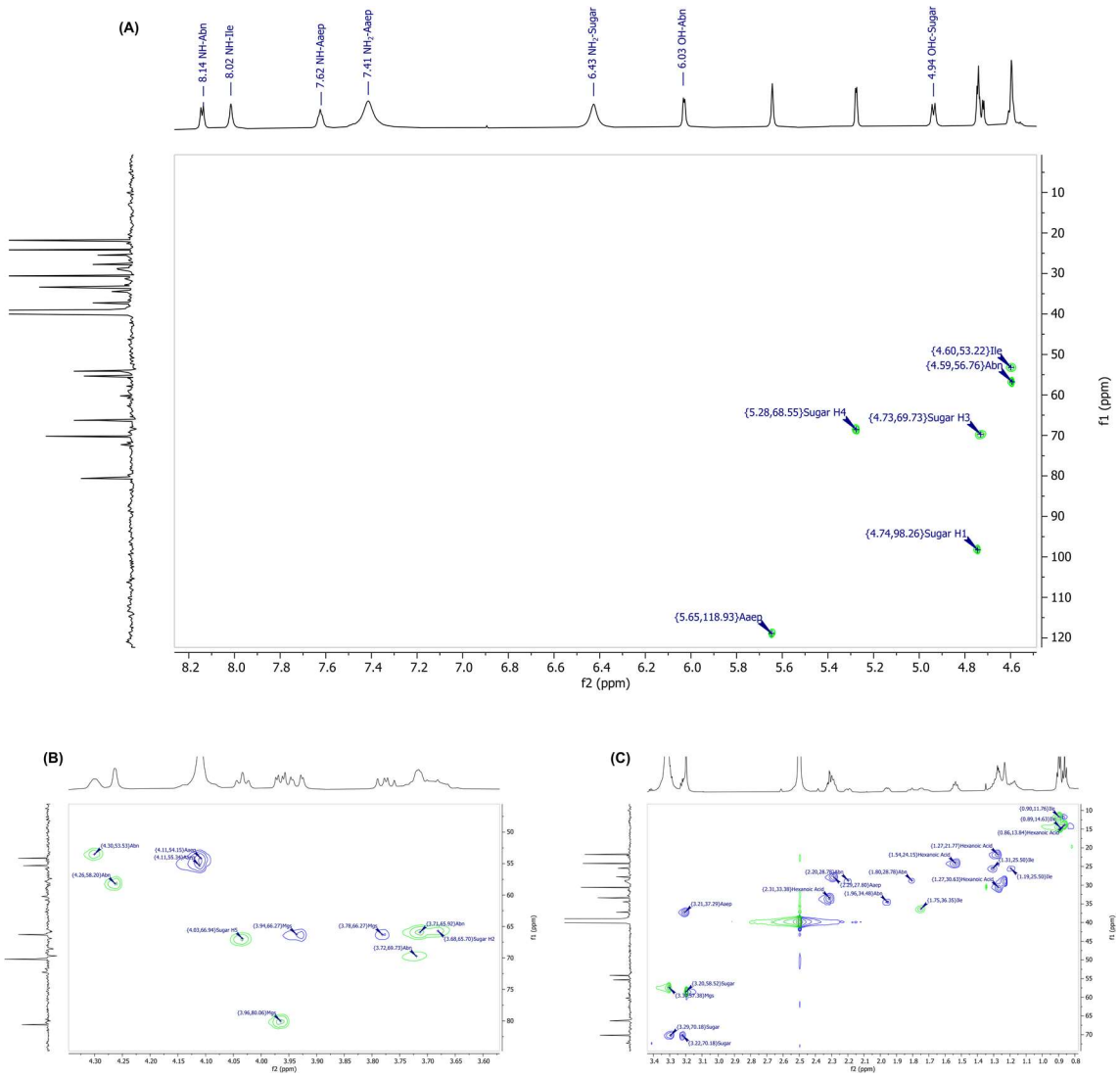


Figure S30 - HSQC (DMSO-d_6, 600 MHz) spectrum of compound **3**, showing the proton annotation. **(A)** Exchangeable protons, highlighting the OH proton, δ^H 4.94 (d, 7.6 Hz) of position 2 of the sugar unit; **(B)** and **(C)** the remaining resonances of the structure.

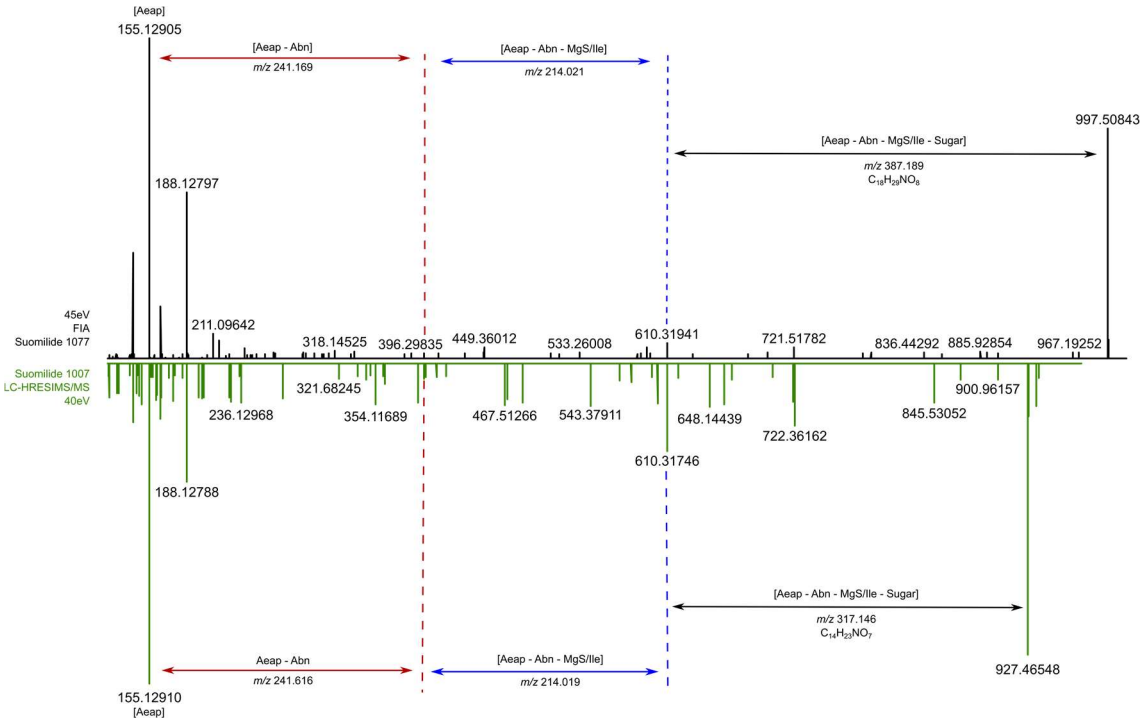


Figure S31 - Comparison of the MS_2 from the suomilide 1077 (**2**) with soumilide 1007 (**3**), confirming the structural difference in the glycosyl unit connected to the Abn moiety.

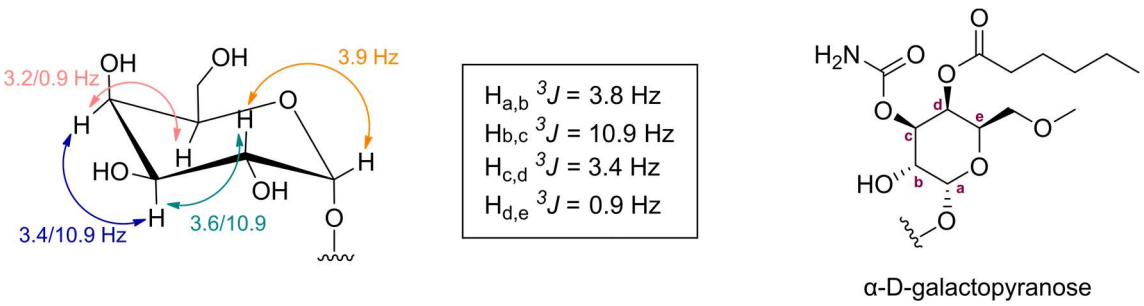


Figure S32 - Coupling constant analysis (^{1}H and COSY data, DMSO-d_6, 600 MHz) underscoring the proposed sugar configuration of compound **3**.

Table S5 - Annotation of the *nos* biosynthetic gene cluster.

Protein	Lenght (aa)	Accesion number	Locus ID	Predicted funtion	closest homolog (refseq)	organism	identity/ similarity (%)	accession no
NosH	2243	WP_193945304.1	IQ275_03335	PKS	type I polyketide synthase [Nostoc sp. XA010]	Nostoc sp. XA010	96/98	WP_229555179.1
NosA	3612	WP_193945303.1	IQ275_03330	NRPS	non-ribosomal peptide synthetase [Nostoc sp. XA010]	Nostoc sp. XA010	95/96	WP_229555178.1
NosB	1428	WP_193945302.1	IQ275_03325	PKS	type I polyketide synthase [Nostoc sp. XA010]	Nostoc sp. XA010	96/98	WP_229555177.1
NosC	3292	WP_193945301.1	IQ275_03320	NRPS	non-ribosomal peptide synthetase [Nostoc sp. XA010]	Nostoc sp. XA010	94/95	WP_229555176.1
NosD	2883	WP_193945300.1	IQ275_03315	NRPS	non-ribosomal peptide synthetase [Nostoc sp. XA010]	Nostoc sp. XA010	95/96	WP_229555175.1
Orf	277	WP_193945299.1	IQ275_03310	phytanoyl-CoA dioxygenase family protein	phytanoyl-CoA dioxygenase family protein [Nostoc sp. XA010]	Nostoc sp. XA010	97/99	WP_229555174.1
NosG	675	WP_193945298.1	IQ275_03305	ABC transporter ATP-binding protein	ABC transporter ATP-binding protein/permease [Nostoc sp. JL34]	Nostoc sp. JL34	96/97	MBN3883111.1

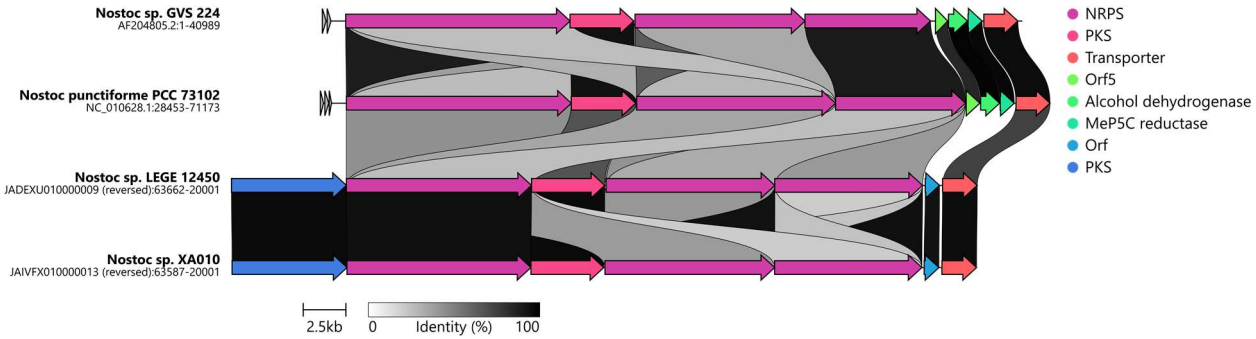


Figure S33 - Alignment of the biosynthetic gene clusters (BGCs) from cyanobacterial strains known to produce nostopeptolides: *Nostoc* sp. GVS 224, *Nostoc punctiforme* PCC 73102 and *Nostoc* sp. LEGE 12450. *Nostoc* sp. XA010 genome was annotated by antiSMASH and aligned as well, since it was the closest homologue in the performed BGC annotation (Table S5).

Table S6 - Annotation of the *suo* biosynthetic gene cluster. Proteins at bold are new, comparing to the *suo* BGC of *Nodularia sphaerocarpa* UHCC 0038 (Orf28-31).

Protein	Lenght (aa)	Accesion number	Locus ID	Predicted funtion	closest homolog (refseq)	organism	identity/ similarity (%)	accession no
SuoB	2943	WP_228038150.1	IQ275_05600	NRPS (FkbH - methoxymalonyl ACP biosynthesis)	amino acid adenylation domain-containing protein [Nostoc sp. B(2019)]	Nostoc sp. B(2019)	88/93	WP_238404877.1
	2943		IQ275_05605	NRPS (Condensation of L-amino acids)	non-ribosomal peptide synthetase [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	90/94	WP_146110921.1
	2943		IQ275_05610	NRPS (Amino acid adenylation domain)	non-ribosomal peptide synthetase [Nostoc sp. KVJ20]	Nostoc sp. KVJ20	91/06	QQH17589.1
	2943		IQ275_05615	NRPS (Epimerization domain)	non-ribosomal peptide synthetase [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	94/96	WP_228038150.1
Orf3	274	WP_193945667.1	IQ275_05620	Hypothetical protein	hypothetical protein [Nostoc sp. KVJ20]	Nostoc sp. KVJ20	92/97	WP_069074352.1
Orf4	190	WP_193945668.1	IQ275_05625	Adenylyl-sulfate kinase	adenylyl-sulfate kinase [Brasilonema octagenarum]	Brasilonema octagenarum	93/97	WP_169266147.1
SuoC	720	WP_193945669.1	IQ275_05630	Rieske non-heme iron oxygenase	Rieske 2Fe-2S domain-containing protein [Brasilonema octagenarum]	Brasilonema octagenarum	93/95	WP_169266146.1
SuoK	351	WP_193945670.1	IQ275_05635	Type 2 isopentenyl-diphosphate Delta-isomerase	type 2 isopentenyl-diphosphate Delta-isomerase [Scytonema sp. NIES-4073]	Scytonema sp. NIES-4073	96/98	WP_096564146.1
Orf28	302	WP_193945671.1	IQ275_05640	DMT Transporter	DMT family transporter [Brasilonema sennae]	Brasilonema sennae	95/97	WP_171978127.1
SouD	202	WP_193945672.1	IQ275_05645	LysR family transcriptional regulator	LysR family transcriptional regulator [unclassified Nostoc]	unclassified Nostoc	99/99	WP_143856977.1
SouE	235	WP_193945673.1	IQ275_05650	Cupin domain-containing protein (BacB or AerE)	cupin domain-containing protein [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	87/93	WP_104900369.1
SouF	264	WP_193945674.1	IQ275_05655	NAD(P)-dependent dehydrogenase (fabG)	SDR family oxidoreductase [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	95/96	WP_104900370.1

SuoG	1642	WP_193945675.1	IQ275_05660	NRPS (DCL and LCL)	non-ribosomal peptide synthetase [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	89/92	WP_104900371.1
Orf11	258	WP_228038157.1	IQ275_05665	SDR family oxidoreductase	SDR family NAD(P)-dependent oxidoreductase [Nostoc sp. 'Peltigera membranacea cyanobiont' 213]	Nostoc sp. 'Peltigera membranacea cyanobiont' 213	96/97	WP_094328879.1
SuoN1	596	WP_193945677.1	IQ275_05670	ABC Transporter	ABC transporter substrate-binding protein [Nostoc sp. 'Peltigera membranacea cyanobiont' 213]	Nostoc sp. 'Peltigera membranacea cyanobiont' 213	94/96	WP_094328878.1
SuoH1	324	WP_193945678.1	IQ275_05675	Isopenicillin N oxygenase	2-oxoglutarate and iron-dependent oxygenase domain-containing protein [Nostoc sp. KVJ20]	Nostoc sp. KVJ20	90/93	WP_069070596.1
SuoH2	328	WP_193945679.1	IQ275_05680	Isopenicillin N oxygenase	2-oxoglutarate and iron-dependent oxygenase domain-containing protein [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	96/97	WP_104900375.1
Orf15	385	WP_193945680.1	IQ275_05685	Hypothetical protein	hypothetical protein [unclassified Nostoc]	unclassified Nostoc	94/95	WP_094328875.1
Orf16	502	WP_193945681.1	IQ275_05690	MBOAT (membrane-bound O-acyltransferase)	MBOAT family O-acyltransferase [unclassified Nostoc]	unclassified Nostoc	90/94	WP_094328874.1
Orf29	240	WP_193945682.1	IQ275_05695	Class I SAM-dependent methyltransferase	class I SAM-dependent methyltransferase [Nostoc sp. 'Peltigera membranacea cyanobiont' 213]	Nostoc sp. 'Peltigera membranacea cyanobiont' 213	96/97	WP_094328873.1
Orf17	612	WP_193945683.1	IQ275_05700	Carbamoyltransferase	carbamoyltransferase [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	99/99	WP_104900377.1
Orf18	137	WP_228038151.1	IQ275_05705	SxtJ (Saxitoxin)	SxtJ family membrane protein [Nostoc sp. 'Peltigera membranacea cyanobiont' 213]	Nostoc sp. 'Peltigera membranacea cyanobiont' 213	96/97	WP_094328871.1
Orf19	54	WP_190233927.1	IQ275_05710	Hypothetical protein	DUF5989 family protein [Nostocales]	Nostocales	98/98	WP_096564133.1
Orf30	432	WP_193945685.1	IQ275_05715	SGNH/GDSL hydrolase	SGNH/GDSL hydrolase family protein [Nostoc sp. 'Peltigera membranacea cyanobiont' 213]	Nostoc sp. 'Peltigera membranacea cyanobiont' 213	97/98	WP_143856976.1

Orf20	326	WP_193945687.1	IQ275_05720	SGNH/GDSL hydrolase	SGNH/GDSL hydrolase family protein [Nostoc sp. 'Peltigera membranacea cyanobiont' 213]	Nostoc sp. 'Peltigera membranacea cyanobiont' 213	95/97	WP_094328869.1
Orf21	319	WP_193945689.1	IQ275_05725	Carbohydrate binding protein	carbohydrate-binding protein [Nostoc sp. 'Peltigera membranacea cyanobiont' N6]	Nostoc sp. 'Peltigera membranacea cyanobiont' N6	94/97	WP_146110923.1
Orf22	313	WP_193945691.1	IQ275_05730	Carbohydrate binding protein	carbohydrate-binding protein [Brasilonema sennae]	Brasilonema sennae	96/96	WP_171975716.1
Orf31	227	WP_193945693.1	IQ275_05735	Radical SAM	radical SAM protein [Nostoc sp. KVJ20]	Nostoc sp. KVJ20	96/98	WP_084227114.1
SuoH3	335	WP_193945695.1	IQ275_05740	Isopenicillin N oxygenase	2-oxoglutarate and iron-dependent oxygenase domain-containing protein [Nostoc sp. KVJ20]	Nostoc sp. KVJ20	94/96	WP_069070585.1
SuoH4	317	WP_193945697.1	IQ275_05745	Isopenicillin N oxygenase	2OG-Fe(II) oxygenase family protein [Nostoc sp. 'Peltigera membranacea cyanobiont' 213]	Nostoc sp. 'Peltigera membranacea cyanobiont' 213	94/96	WP_094328864.1
Orf25	277	WP_193945699.1	IQ275_05750	Aldo/keto reductase	aldo/keto reductase [unclassified Nostoc]	unclassified Nostoc	95/96	WP_094328863.1

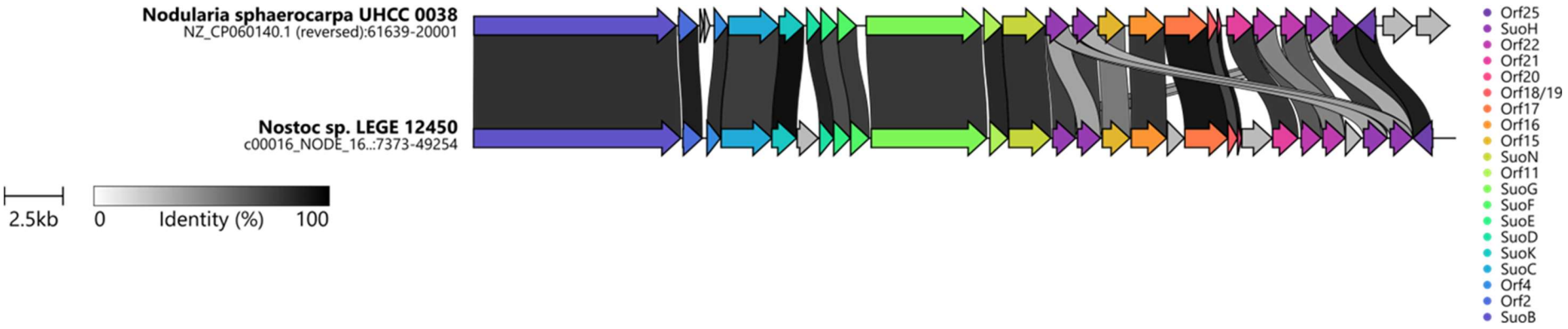


Figure S19 - Alignment of the biosynthetic gene clusters (BGCs) from *Nostoc* sp. LEGE 12450 and with *Nodularia sphaerocarpa* UHCC 0038 from which the *suo* BGC was annotated first.

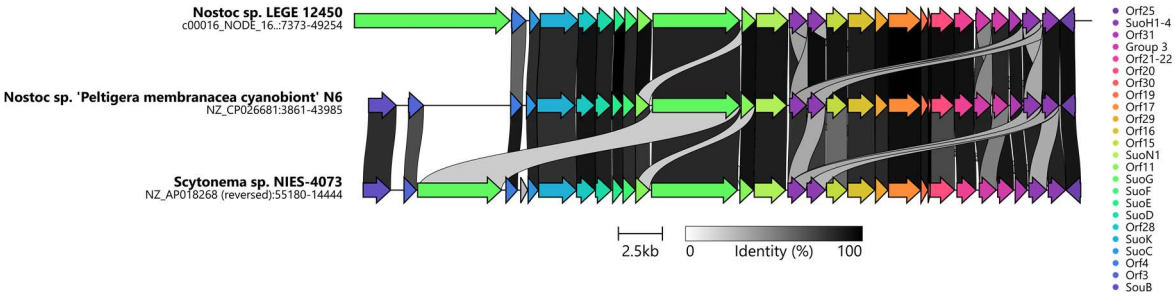


Figure S35 - Alignment of the BGCs with similarity retrieved by using ClusterBlast, the antiSMASH tool.

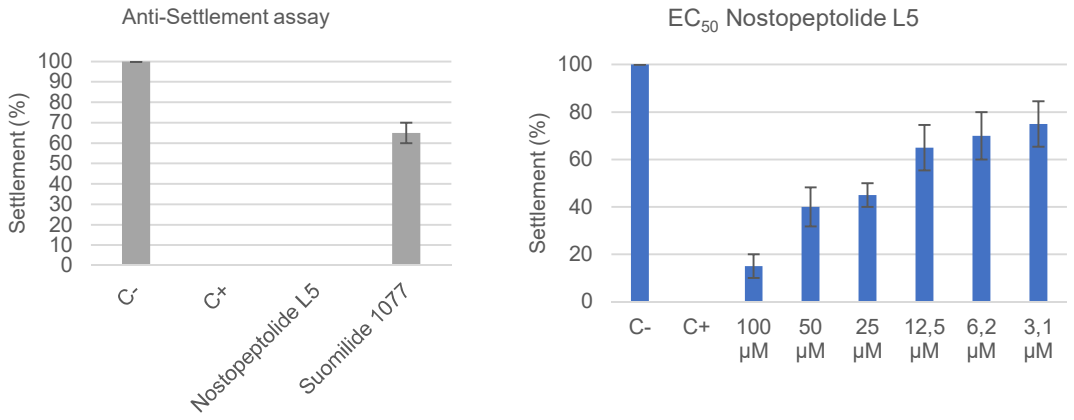


Figure S36 - Antifouling activity of **1**, with the anti-settlement assay using *M. galloprovincialis* larvae and assay performed for the EC_{50} calculations.

Table S7 - Anticancer assays using the hepatocellular (HepG2) and colon (HCT 116) carcinoma cell lines in monolayer cultures (2D) and with HCT 116 cell line in 3D cultures. The compounds 1 and 2 were tested at a concentration of 10 μM.

	2D		3D
	HepG2	HCT 116	HCT 116
Nostopeptolide L5	128.5 ± 12.9	98.9 ± 5.6	97.2 ± 2.8
Suomilide 1077	109.7 ± 20.6	93.3 ± 4.5	100.9 ± 2.9

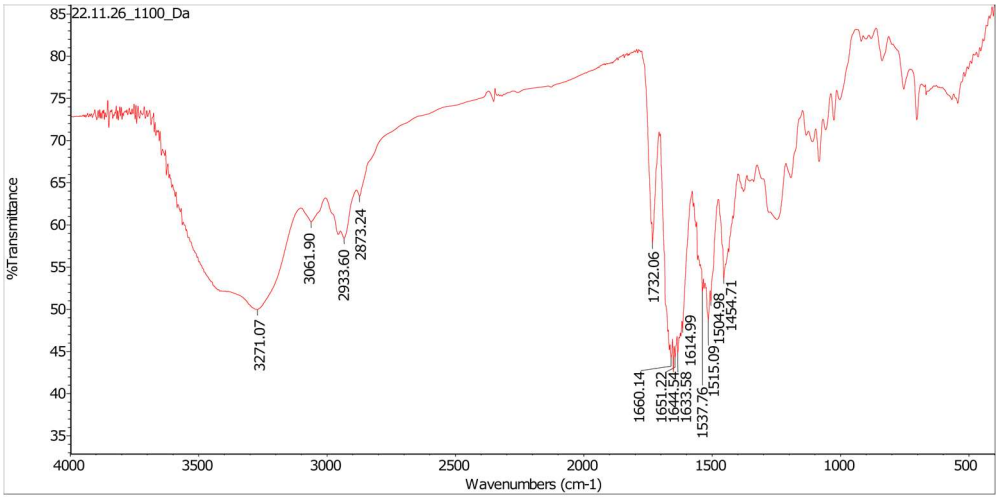


Figure S37 - FT-IR spectrum of compound **1**.

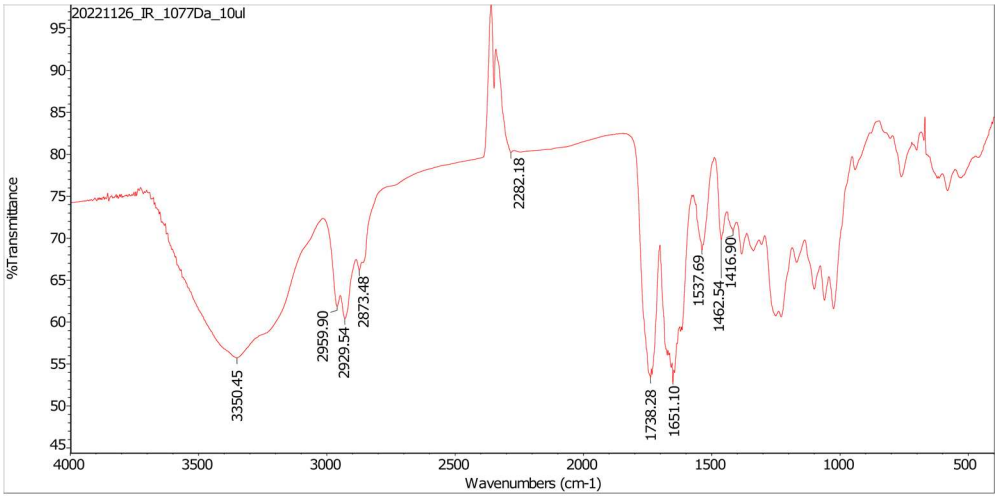


Figure S38 - FI-IR spectrum of compound **2**.

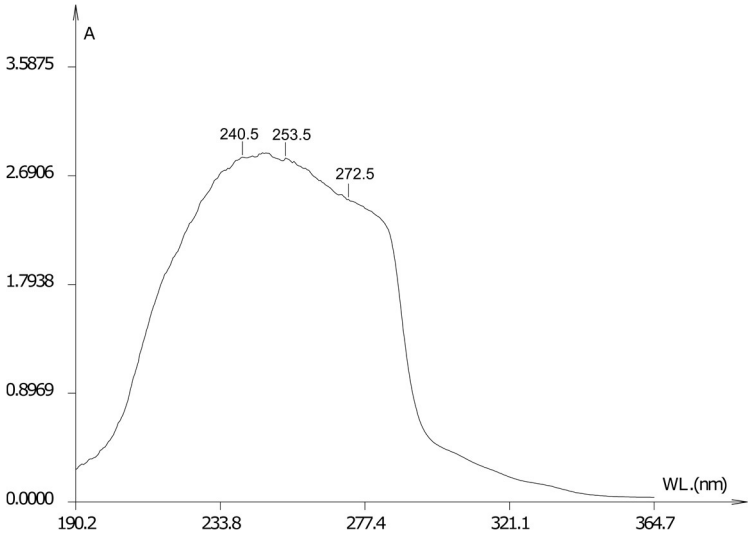


Figure S39 - UV-vis spectrum of compound **1**.

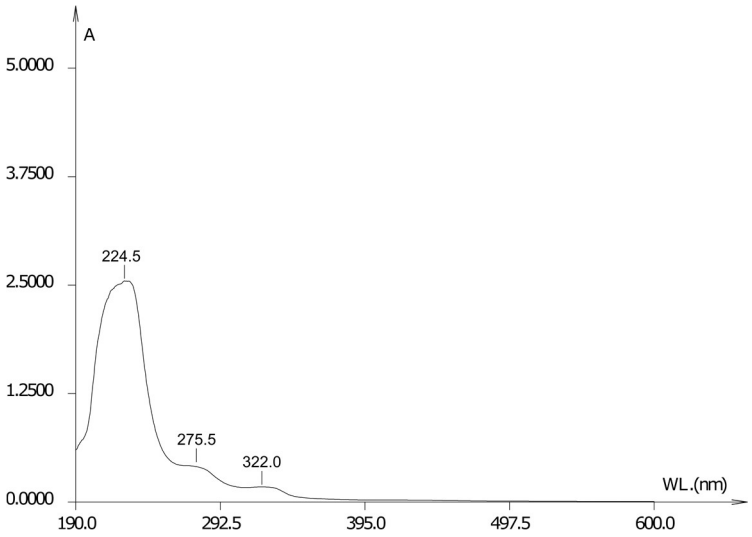


Figure S40 - UV-vis spectrum of compound **2**.

4.6. References

1. Wenski, S. L.; Thiengmag, S.; Helfrich, E. J. N., Complex peptide natural products: Biosynthetic principles, challenges and opportunities for pathway engineering. *Synthetic and Systems Biotechnology* **2022,** *7* (1), 631-647.

2. Baunach, M.; Chowdhury, S.; Stallforth, P.; Dittmann, E., The Landscape of Recombination Events That Create Nonribosomal Peptide Diversity. *Molecular Biology and Evolution* **2021,** *38* (5), 2116-2130.

3. Ziemert, N.; Alanjary, M.; Weber, T., The evolution of genome mining in microbes - a review. *Nat Prod Rep* **2016,** *33* (8), 988-1005.

4. Dittmann, E.; Gugger, M.; Sivonen, K.; Fewer, D. P., Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. *Trends Microbiol* **2015,** *23* (10), 642-652.

5. Fewer, D. P.; Jokela, J.; Heinilä, L.; Aesoy, R.; Sivonen, K.; Galica, T.; Hrouzek, P.; Herfindal, L., Chemical diversity and cellular effects of antifungal cyclic lipopeptides from cyanobacteria. *Physiol Plant* **2021,** *173* (2), 639-650.

6. Mareš, J.; Hájek, J.; Urajová, P.; Kust, A.; Jokela, J.; Saurav, K.; Galica, T.; Čapková, K.; Mattila, A.; Haapaniemi, E.; Permi, P.; Mysterud, I.; Skulberg, O. M.; Karlsen, J.; Fewer, D. P.; Sivonen, K.; Tønnesen, H. H.; Hrouzek, P., Alternative Biosynthetic Starter Units Enhance the Structural Diversity of Cyanobacterial Lipopeptides. *Applied and Environmental Microbiology* **2019,** *85* (4), e02675-18.

7. Shestakov, S. V.; Karbysheva, E. A., The origin and evolution of cyanobacteria. *Biology Bulletin Reviews* **2017,** *7* (4), 259-272.

8. Calteau, A.; Fewer, D. P.; Latifi, A.; Coursin, T.; Laurent, T.; Jokela, J.; Kerfeld, C. A.; Sivonen, K.; Piel, J.; Gugger, M., Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genomics* **2014,** *15* (1), 977.

9. Popin, R. V.; Alvarenga, D. O.; Castelo-Branco, R.; Fewer, D. P.; Sivonen, K., Mining of Cyanobacterial Genomes Indicates Natural Product Biosynthetic Gene Clusters Located in Conjugative Plasmids. *Frontiers in Microbiology* **2021,** *12*.

10. Shih, P. M.; Wu, D.; Latifi, A.; Axen, S. D.; Fewer, D. P.; Talla, E.; Calteau, A.; Cai, F.; Tandeau de Marsac, N.; Rippka, R.; Herdman, M.; Sivonen, K.; Coursin, T.; Laurent, T.; Goodwin, L.; Nolan, M.; Davenport, K. W.; Han, C. S.; Rubin, E. M.; Eisen, J. A.; Woyke, T.; Gugger, M.; Kerfeld, C. A., Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proceedings of the National Academy of Sciences* **2013,** *110* (3), 1053-1058.

11. Huo, D.; Li, H.; Cai, F.; Guo, X.; Qiao, Z.; Wang, W.; Yu, G.; Li, R., Genome Evolution of Filamentous Cyanobacterium Nostoc Species: From Facultative Symbiosis to Free Living. *Microorganisms* **2021,** *9* (10), 2015.

12. Gaysina, L. A.; Saraf, A.; Singh, P., Chapter 1 - Cyanobacteria in Diverse Habitats. In *Cyanobacteria*, Mishra, A. K.; Tiwari, D. N.; Rai, A. N., Eds. Academic Press: 2019; pp 1-28.

13. Liaimer, A.; Jenke-Kodama, H.; Ishida, K.; Hinrichs, K.; Stangeland, J.; Hertweck, C.; Dittmann, E., A polyketide interferes with cellular differentiation in the symbiotic cyanobacterium Nostoc punctiforme. *Environ Microbiol Rep* **2011,** *3* (5), 550-8.

14. Liaimer, A.; Helfrich, E. J. N.; Hinrichs, K.; Guljamow, A.; Ishida, K.; Hertweck, C.; Dittmann, E., Nostopeptolide plays a governing role during cellular differentiation of the symbiotic cyanobacterium Nostoc punctiforme. *Proceedings of the National Academy of Sciences* **2015,** *112* (6), 1862-1867.

15. Guljamow, A.; Kreische, M.; Ishida, K.; Liaimer, A.; Altermark, B.; Bähr, L.; Hertweck, C.; Ehwald, R.; Dittmann, E., High-Density Cultivation of Terrestrial Nostoc Strains Leads to Reprogramming of Secondary Metabolome. *Appl Environ Microbiol* **2017,** *83* (23).

16. Dehm, D.; Krumbholz, J.; Baunach, M.; Wiebach, V.; Hinrichs, K.; Guljamow, A.; Tabuchi, T.; Jenke-Kodama, H.; Süssmuth, R. D.; Dittmann, E., Unlocking the Spatial Control of Secondary Metabolism Uncovers Hidden Natural Product Diversity in Nostoc punctiforme. *ACS Chem Biol* **2019,** *14* (6), 1271-1279.

17. Golakoti, T.; Yoshida, W. Y.; Chaganty, S.; Moore, R. E., Isolation and Structures of Nostopeptolides A1, A2 and A3 from the Cyanobacterium Nostoc sp. GSV224. *Tetrahedron* **2000,** *56* (46), 9093-9102.

18. Krumbholz, J.; Ishida, K.; Baunach, M.; Teikari, J. E.; Rose, M. M.; Sasso, S.; Hertweck, C.; Dittmann, E., Deciphering Chemical Mediators Regulating Specialized Metabolism in a Symbiotic Cyanobacterium. *Angew Chem Int Ed Engl* **2022,** *61* (26), e202204545.

19. Liaimer, A.; Jensen, J. B.; Dittmann, E., A Genetic and Chemical Perspective on Symbiotic Recruitment of Cyanobacteria of the Genus Nostoc into the Host Plant Blasia pusilla L. *Frontiers in Microbiology* **2016,** *7*.

20. Kerksiek, K.; Mejillano, M. R.; Schwartz, R. E.; Georg, G. I.; Himes, R. H., Interaction of cryptophycin 1 with tubulin and microtubules. *FEBS Letters* **1995,** *377* (1), 59-61.

21. Golakoti, T.; Yoshida, W. Y.; Chaganty, S.; Moore, R. E., Isolation and structure determination of nostocyclopeptides A1 and A2 from the terrestrial cyanobacterium Nostoc sp. ATCC53789. *J Nat Prod* **2001,** *64* (1), 54-9.

22. Ishida, K.; Okita, Y.; Matsuda, H.; Okino, T.; Murakami, M., Aeruginosins, protease inhibitors from the cyanobacterium Microcystis aeruginosa. *Tetrahedron* **1999,** *55* (36), 10971-10988.

23. Fujii, K.; Sivonen, K.; Adachi, K.; Noguchi, K.; Shimizu, Y.; Sano, H.; Hirayama, K.; Suzuki, M.; Harada, K.-i., Comparative study of toxic and non-toxic cyanobacterial products: A novel glycoside, suomilide, from non-toxic Nodularia spumigena HKVV. *Tetrahedron Letters* **1997,** *38* (31), 5529-5532.

24. Pluotno, A.; Carmeli, S., Banyasin A and banyasides A and B, three novel modified peptides from a water bloom of the cyanobacterium Nostoc sp. *Tetrahedron* **2005,** *61* (3), 575-583.

25. Magarvey, N. A.; Beck, Z. Q.; Golakoti, T.; Ding, Y.; Huber, U.; Hemscheidt, T. K.; Abelson, D.; Moore, R. E.; Sherman, D. H., Biosynthetic Characterization and Chemoenzymatic Assembly of the Cryptophycins. Potent Anticancer Agents from Nostoc Cyanobionts. *ACS Chemical Biology* **2006,** *1* (12), 766-779.

26. Kopp, F.; Mahlert, C.; Grünewald, J.; Marahiel, M. A., Peptide Macrocyclization: The Reductase of the Nostocyclopeptide Synthetase Triggers the Self-Assembly of a Macrocyclic Imine. *Journal of the American Chemical Society* **2006,** *128* (51), 16478-16479.

27. Fewer, D. P.; Jokela, J.; Paukku, E.; Österholm, J.; Wahlsten, M.; Permi, P.; Aitio, O.; Rouhiainen, L.; Gomez-Saez, G. V.; Sivonen, K., New structural variants of aeruginosin produced by the toxic bloom forming cyanobacterium Nodularia spumigena. *PLoS One* **2013,** *8* (9), e73618.

28. Liu, L.; Jokela, J.; Herfindal, L.; Wahlsten, M.; Sinkkonen, J.; Permi, P.; Fewer, D. P.; Døskeland, S. O.; Sivonen, K., 4-Methylproline Guided Natural Product Discovery: Co-Occurrence of 4-Hydroxy- and 4-Methylprolines in Nostoweipeptins and Nostopeptolides. *ACS Chemical Biology* **2014,** *9* (11), 2646-2655.

29. Ahmed, M. N.; Wahlsten, M.; Jokela, J.; Nees, M.; Stenman, U.-H.; Alvarenga, D. O.; Strandin, T.; Sivonen, K.; Poso, A.; Permi, P.; Metsä-Ketelä, M.; Koistinen, H.; Fewer, D. P., Potent Inhibitor of Human Trypsins from the Aeruginosin Family of Natural Products. *ACS Chemical Biology* **2021,** *16* (11), 2537-2546.

30. Entfellner, E.; Frei, M.; Christiansen, G.; Deng, L.; Blom, J.; Kurmayer, R., Evolution of Anabaenopeptin Peptide Structural Variability in the Cyanobacterium Planktothrix. *Frontiers in Microbiology* **2017,** *8*.

31. Fidor, A.; Cekała, K.; Wieczerzak, E.; Cegłowska, M.; Kasprzykowski, F.; Edwards, C.; Mazur-Marzec, H., Nostocyclopeptides as New Inhibitors of 20S Proteasome. *Biomolecules* **2021,** *11* (10).

32. 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. *Nature Biotechnology* **2016,** *34* (8), 828-837.

33. Mohimani, H.; Gurevich, A.; Mikheenko, A.; Garg, N.; Nothias, L.-F.; Ninomiya, A.; Takada, K.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of peptidic natural products through database search of mass spectra. *Nat Chem Biol* **2017,** *13* (1), 30-37.

34. Gurevich, A.; Mikheenko, A.; Shlemov, A.; Korobeynikov, A.; Mohimani, H.; Pevzner, P. A., Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. *Nat Microbiol* **2018,** *3* (3), 319-327.

35. Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L. F.; Dorrestein, P. C.; Pevzner, P. A., Dereplication of microbial metabolites through database search of mass spectra. *Nat Commun* **2018,** *9* (1), 4035.

36. Dictionary of Natural Products 32.0. https://dnp.chemnetbase.com (accessed 20 Dec 2021).

37. van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castaño-Espriu, L.; Chang, C.; Clark, T. N.; Cleary Little, J. L.; Delgadillo, D. A.; Dorrestein, P. C.; Duncan, K. R.; Egan, J. M.; Galey, M. M.; Haeckl, F. P. J.; Hua, A.; Hughes, A. H.; Iskakova, D.; Khadilkar, A.; Lee, J.-H.; Lee, S.; LeGrow, N.; Liu, D. Y.; Macho, J. M.; McCaughey, C. S.; Medema, M. H.; Neupane, R. P.; O'Donnell, T. J.; Paula, J. S.; Sanchez, L. M.; Shaikh, A. F.; Soldatou, S.; Terlouw, B. R.; Tran, T. A.; Valentine, M.; van der Hooft, J. J. J.; Vo, D. A.; Wang, M.; Wilson, D.; Zink, K. E.; Linington, R. G., The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Central Science* **2019,** *5* (11), 1824-1833.

38. Zi, J.; Lantvit, D. D.; Swanson, S. M.; Orjala, J., Lyngbyaureidamides A and B, two anabaenopeptins from the cultured freshwater cyanobacterium Lyngbya sp. (SAG 36.91). *Phytochemistry* **2012,** *74*, 173-177.

39. Zafrir-Ilan, E.; Carmeli, S., Eight novel serine proteases inhibitors from a water bloom of the cyanobacterium Microcystis sp. *Tetrahedron* **2010,** *66* (47), 9194-9202.

40. Fujii, K.; Sivonen, K.; Adachi, K.; Noguchi, K.; Sano, H.; Hirayama, K.; Suzuki, M.; Harada, K.-i., Comparative study of toxic and non-toxic cyanobacterial products: Novel peptides from toxic Nodularia spumigena AV1. *Tetrahedron Letters* **1997,** *38* (31), 5525-5528.

41. Zhang, L.-H.; Longley, R. E.; Koehn, F. E., Antiproliferative and immunosuppressive properties of microcolin A, a marine-derived lipopeptide. *Life Sciences* **1997,** *60* (10), 751-762.

42. Bornancin, L.; Alonso, E.; Alvariño, R.; Inguimbert, N.; Bonnard, I.; Botana, L. M.; Banaigs, B., Structure and biological evaluation of new cyclic and acyclic laxaphycin-A type peptides. *Bioorganic & Medicinal Chemistry* **2019,** *27* (10), 1966-1980.

43. Moore, R. E.; Entzeroth, M., Majusculamide D and deoxymajusculamide D, two cytotoxins from Lyngbya majuscula. *Phytochemistry* **1988,** *27* (10), 3101-3103.

44. Heinilä, L. M. P.; Fewer, D. P.; Jokela, J. K.; Wahlsten, M.; Jortikka, A.; Sivonen, K., Shared PKS Module in Biosynthesis of Synergistic Laxaphycins. *Frontiers in Microbiology* **2020,** *11*.

45. Heinilä, L. M. P.; Fewer, D. P.; Jokela, J. K.; Wahlsten, M.; Ouyang, X.; Permi, P.; Jortikka, A.; Sivonen, K., The structure and biosynthesis of heinamides A1–A3 and B1–B5, antifungal members of the laxaphycin lipopeptide family. *Organic & Biomolecular Chemistry* **2021,** *19* (25), 5577-5588.

46. Medina, R. A.; Goeger, D. E.; Hills, P.; Mooberry, S. L.; Huang, N.; Romero, L. I.; Ortega-Barría, E.; Gerwick, W. H.; McPhail, K. L., Coibamide A, a Potent Antiproliferative Cyclic Depsipeptide from the Panamanian Marine Cyanobacterium Leptolyngbya sp. *Journal of the American Chemical Society* **2008,** *130* (20), 6324-6325.

47. Luesch, H.; Yoshida, W. Y.; Moore, R. E.; Paul, V. J., Lyngbyastatin 2 and Norlyngbyastatin 2, Analogues of Dolastatin G and Nordolastatin G from the Marine Cyanobacterium Lyngbya majuscula. *Journal of Natural Products* **1999,** *62* (12), 1702-1706.

48. Horgen, F. D.; Yoshida, W. Y.; Scheuer, P. J., Malevamides A−C, New Depsipeptides from the Marine Cyanobacterium Symploca laete-viridis. *Journal of Natural Products* **2000,** *63* (4), 461-467.

49. Jakobi, C.; Rinehart, K. L.; Neuber, R.; Mez, K.; Weckesser, J., Cyanopeptolin SS, a disulphated depsipeptide from a water bloom: structural elucidation and biological activities. *Phycologia* **1996,** *35* (sup6), 111-116.

50. Shin, H. J.; Matsuda, H.; Murakami, M.; Yamaguchi, K., Aeruginosins 205A and -B, Serine Protease Inhibitory Glycopeptides from the Cyanobacterium Oscillatoria agardhii (NIES-205). *The Journal of Organic Chemistry* **1997,** *62* (6), 1810-1813.

51. Kohler, E.; Grundler, V.; Häussinger, D.; Kurmayer, R.; Gademann, K.; Pernthaler, J.; Blom, J. F., The toxicity and enzyme activity of a chlorine and sulfate containing aeruginosin isolated from a non-microcystin-producing Planktothrix strain. *Harmful Algae* **2014,** *39*, 154-160.

52. Heinilä, L. M. P.; Jokela, J.; Ahmed, M. N.; Wahlsten, M.; Kumar, S.; Hrouzek, P.; Permi, P.; Koistinen, H.; Fewer, D. P.; Sivonen, K., Discovery of varlaxins, new aeruginosin-type inhibitors of human trypsins. *Organic & Biomolecular Chemistry* **2022,** *20* (13), 2681-2692.

53. Liu, J.; Zhang, M.; Huang, Z.; Fang, J.; Wang, Z.; Zhou, C.; Qiu, X., Diversity, Biosynthesis and Bioactivity of Aeruginosins, a Family of Cyanobacteria-Derived Nonribosomal Linear Tetrapeptides. *Marine Drugs* **2023,** *21* (4), 217.

54. Reshef, V.; Mizrachi, E.; Maretzki, T.; Silberstein, C.; Loya, S.; Hizi, A.; Carmeli, S., New acylated sulfoglycolipids and digalactolipids and related known glycolipids from cyanobacteria with a potential to inhibit the reverse transcriptase of HIV-1. *J Nat Prod* **1997,** *60* (12), 1251-60.

55. Rho, M.-C.; Matsunaga, K.; Yasuda, K.; Ohizumi, Y., A Novel Monogalactosylacylglycerol with Inhibitory Effect on Platelet Aggregation from the Cyanophyceae Oscillatoria rosea. *Journal of Natural Products* **1996,** *59* (3), 308-309.

56. Son, B. W.; Kim, J. C.; Choi, H. D., A new diacylgalactolipid containing 4Z-16:1 from the marine cyanobacterium Oscillatoria sp. *Lipids* **2001,** *36* (4), 427-429.

57. Nazifi, E.; Wada, N.; Asano, T.; Nishiuchi, T.; Iwamuro, Y.; Chinaka, S.; Matsugo, S.; Sakamoto, T., Characterization of the chemical diversity of glycosylated mycosporine-like amino acids in the terrestrial cyanobacterium Nostoc commune. *Journal of Photochemistry and Photobiology B: Biology* **2015,** *142*, 154-168.

58. Ishihara, K.; Watanabe, R.; Uchida, H.; Suzuki, T.; Yamashita, M.; Takenaka, H.; Nazifi, E.; Matsugo, S.; Yamaba, M.; Sakamoto, T., Novel glycosylated mycosporine-like amino acid, 13-O-(β-galactosyl)-porphyra-334, from the edible cyanobacterium Nostoc sphaericum-protective activity on human keratinocytes from UV light. *J Photochem Photobiol B* **2017,** *172*, 102-108.

59. Lee, H. W.; Oh, C. H.; Geyer, A.; Pfleiderer, W.; Park, Y. S., Characterization of a novel unconjugated pteridine glycoside, cyanopterin, in Synechocystis sp. PCC 6803. *Biochim Biophys Acta* **1999,** *1410* (1), 61-70.

60. Li, K.; Cai, J.; Su, Z.; Yang, B.; Liu, Y.; Zhou, X.; Huang, J.; Tao, H., Glycosylated Natural Products From Marine Microbes. *Frontiers in Chemistry* **2020,** *7*.

61. Hoffmann, D.; Hevel, J. M.; Moore, R. E.; Moore, B. S., Sequence analysis and biochemical characterization of the nostopeptolide A biosynthetic gene cluster from Nostoc sp. GSV224. *Gene* **2003,** *311*, 171-80.

62. Hunsucker, S. W.; Klage, K.; Slaughter, S. M.; Potts, M.; Helm, R. F., A preliminary investigation of the Nostoc punctiforme proteome. *Biochem Biophys Res Commun* **2004,** *317* (4), 1121-7.

63. Blin, K.; Shaw, S.; Kloosterman, A. M.; Charlop-Powers, Z.; van Wezel, G. P.; Medema, Marnix H.; Weber, T., antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research* **2021,** *49* (W1), W29-W35.

64. Ishida, K.; Christiansen, G.; Yoshida, W. Y.; Kurmayer, R.; Welker, M.; Valls, N.; Bonjoch, J.; Hertweck, C.; Börner, T.; Hemscheidt, T.; Dittmann, E., Biosynthesis and structure of aeruginoside 126A and 126B, cyanobacterial peptide glycosides bearing a 2-carboxy-6-hydroxyoctahydroindole moiety. *Chem Biol* **2007,** *14* (5), 565-576.

65. Gagunashvili, A. N.; Andrésson, Ó. S., Distinctive characters of Nostoc genomes in cyanolichens. *BMC Genomics* **2018,** *19* (1), 434.

66. Ramos, V.; Morais, J.; Castelo-Branco, R.; Pinheiro, Â.; Martins, J.; Regueiras, A.; Pereira, A. L.; Lopes, V. R.; Frazão, B.; Gomes, D.; Moreira, C.; Costa, M. S.; Brûle, S.; Faustino, S.; Martins, R.; Saker, M.; Osswald, J.; Leão, P. N.; Vasconcelos, V. M., Cyanobacterial diversity held in microbial biological

resource centers as a biotechnological asset: the case study of the newly established LEGE culture collection. *Journal of Applied Phycology* **2018,** *30* (3), 1437-1451.

67. Wu, Y. W.; Simmons, B. A.; Singer, S. W., MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **2016,** *32* (4), 605-7.

68. CAGECAT. https://cagecat.bioinformatics.nl (accessed April).

69. Almeida, J. R.; Palmeira, A.; Campos, A.; Cunha, I.; Freitas, M.; Felpeto, A. B.; Turkina, M. V.; Vasconcelos, V.; Pinto, M.; Correia-da-Silva, M.; Sousa, E., Structure-Antifouling Activity Relationship and Molecular Targets of Bio-Inspired(thio)xanthones. *Biomolecules* **2020,** *10* (8), 1126.

Chapter V

General Conclusion

5. General Conclusion

In this project, cyanobacteria proved once more their potential for the discovery of new natural products as well as its responsible biosynthetic pathways. The metabolomic screening conducted here, showed the potential of the late-branching cyanobacterial orders, mainly Nostocales. Also, exploring well stablished culture collections, as it is the case of LEGEcc, opens the opportunity to work with a wide morphological and phylogenetic diversity of cyanobacteria, for example from symbiotic ecosystems. The two symbiotic cyanobacteria, isolated from *Cycas* plants, explored in this project contributed to the enrichment of new natural products of this phylum, as well as the biosynthetic machinery responsible for its assembly.

Desmamides A-C are a new class of lipoglycopeptides from the cyanobiont *Desmonostoc muscorum* LEGE 12446. These encompass unique chemical features, as the fatty acyl moiety dhmda connected to the core of amino acids, resulting of the modification of a hexanoic acid starter unit. The putative *dsm* biosynthetic gene cluster was also identified, being in agreement with the chemical structure of desmamides. However, it was not possible to identify the enzyme responsible for the glycosylation of these compounds, opening the door for the discovery of a new biosynthetic mechanism and/or enzyme capable of adding the monosaccharide unit to the structure. Lipoglycopeptides are known for their antibacterial and antifungal activities, a characteristic also demonstrated by desmamide B that was active against the black rot disease-causing plant pathogen *Xanthomonas campestris*. This antibiotic activity, together with the high production of these metabolites by the cyanobacterium and its ecological settings might indicate their ecological role in the host plant.

Nostopeptolide L5 and suomilides 1077 and 1007 are new peptides from the cyanobiont *Nostoc* sp. LEGE 12450, analogues of two family of compounds well recognized in cyanobacteria. Nostopeptolide L5 is a new depsipeptide that shows unique chemical scaffolds, as the non-proteinogenic amino acids Hyd and *O*meSer, and also the modified fatty acyl unit MHEA, only found in nostopeptolides L1-L5. The analysis of the genomic content of this cyanobacterium allowed the identification of the *nos* biosynthetic gene cluster, by comparison to BGCs annotated for other nostopeptolide producing cyanobacteria. However, the biosynthesis of the MHEA scaffold was proposed here for the first time, and it would involve the modification of a butyric acyl unit. Suomilides are tetrapeptides belonging to the aeruginosins family of cyanobacterial secondary metabolites. These can be highly modified as is the case of suomilide 1077 and 1007 that are sulfated and glycosylated, showing also a variety of non-proteinogenic

aminoc acids (Aeap and Abn). The main differences of these two new metabolites with the known suomilide are in glycosyl unit (galactose) and respective decorations. This fact was translated in the biosynthetic machinery leading to the production of these compounds, by comparing the putative *suo* biosynthetic gene cluster from *Nodularia sphaerocarpa* UHCC 0038, with the one annotated in the genome of *Nostoc* sp. LEGE 12450. Main differences were related with the glycosylation enzymes, showing also new enzymes that might be involved in the biosynthesis of suomilides 1077, 1007. In fact, most of the enzymes proposed to be involved in the production of these compounds are not well studied, opening once again the door for discovery of new enzymatic activities. Although the auroginosin compound family is well known by their potent serine proteases inhibitory activity, no bioactivity was detected for suomilides 1077, 1007 with the tested bioassays. On the other hand, nostopeptolide L5 demonstrated promising activity as an antifouling agent.

Exploring niche-specific habitats, as symbiotic partnerships between cyanobacteria and cycads, can provide a deeper knowledge of the chemical cues and biosynthetic pathways responsible for the production of new natural products. Understanding the ecologic role of such metabolites will also improve their applicability into new pharmacological and/or biotechnological routes. Since evolution is always causing new adaptations and diversifications, it is only expectable that cyanobacteria will continue to provide a huge chemical and enzymatic diversity, in the natural products research field.