

**Comparative Genomics and Metabolomics
in the Genus *Nocardia***

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Daniel Männle
aus Offenburg

Tübingen
2020

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	28.07.2020
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	JProf. Dr. Leonard Kaysser
2. Berichterstatter:	Prof. Dr. Nadine Ziemert

Eigenständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit „Comparative Genomics to Metabolomics: Nocobactin Production in *Nocardia*“ selbständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, wörtlich sowie inhaltlich übernommenen Stellen und Zitate eindeutig gekennzeichnet habe. Ich erkläre, dass die Richtlinien zur Sicherung guter wissenschaftlicher Praxis der Universität Tübingen (Beschluss des Senats vom 25.05.2000) beachtet wurden. Ich versichere an Eides statt, dass diese Angaben wahr sind und dass ich nichts verschwiegen habe. Mir ist bekannt, dass die falsche Abgabe einer Versicherung an Eides statt mit Freiheitsstrafe bis zu drei Jahren oder mit Geldstrafe bestraft wird.

Bad Säckingen, den

Daniel Männle

Contents

Abbreviations.....	6
List of Publications	9
Zusammenfassung.....	10
Summary	11
Personal Contribution	12
1. Introduction	16
1.1. Background Information	16
1.1.1. Natural product discovery	16
1.1.2. The phylum Actinobacteria.....	16
1.1.3. Pathogenicity of <i>Nocardia</i>	17
1.1.4. Natural products from <i>Nocardia</i>	17
1.2. Siderophores	18
1.2.1. General features and application of siderophores	18
1.2.2. Nocobactin NA and nocobactin derivatives from <i>Nocardia</i>	18
1.2.3. Nocobactin BGC and nocobactin biosynthesis.....	21
1.3. Bioinformatics for natural product discovery	23
1.3.1. Preliminary genomic studies on the genus <i>Nocardia</i>	23
1.3.2. Brief history of genome mining	23
1.3.3. Bioinformatic tools for natural product discovery.....	24
1.3.4. Challenges of bioinformatic predictions for natural product discovery	25
2. Goals.....	26
3. Results & Discussion.....	27
3.1. Comparative Genomics and Metabolomics in the Genus <i>Nocardia</i>	27
3.1.1. Phylogeny and biosynthetic potential of the genus <i>Nocardia</i>	27

3.1.2. Metabolic diversity and gene cluster families in <i>Nocardia</i>	29
3.1.3. The gene cluster family of nocobactins	34
3.1.4. Nocobactin-type siderophores produced by selected <i>Nocardia</i> strains	36
3.1.5. Identification of a mycobactin-type siderophore	40
3.2. New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A-C, Revealed by Genome Mining of <i>Nocardia terpenica</i> IFM 0406.....	41
3.2.1. Identification and organization of the terpenibactin BGC.....	41
3.2.2. LC/MS screening for the products of the terpenibactin BGC	42
3.2.3. NMR structure elucidation of terpenibactins A-C.....	45
3.2.4. Biological properties of terpenibactins A-C.....	47
4. Conclusion & Outlook.....	48
5. References	49
6. Acknowledgements	61
7. Curriculum Vitae	62
8. Appendix.....	63

Abbreviations

μM	micromolar
AMP	adenosine-5'-O-monophosphate
ANI	average nucleotide identity
AT	acyltransferase
autoMLST	automated multi-locus species tree
BGC	biosynthetic gene cluster
BLAST	Basic Local Alignment Search Tool
C	condensation
c	cutoff
COSY	correlation spectroscopy
Cy	heterocyclization
<i>d</i> 4-MeOH	methanol- <i>d</i> ₄
Da	Dalton
E	Epimerization
GCF	gene cluster family
GNPS	Global Natural Product Social Molecular Networking
HMBC	heteronuclear multiple-bond correlation spectroscopy
HMM	hidden markov model
HPLC	high performance liquid chromatography
HSQC	heteronuclear single-quantum correlation spectroscopy
JGI-IMG	Joint Genome Institute – Integrated Microbial Genomes & Microbiomes
kbp	kilo base pairs
KR	ketoreductase
KS	ketosynthase
L	Liter
LC	liquid chromatography
M	mass
<i>M.</i>	<i>Mycobacterium</i>
<i>m/z</i>	mass-to-charge ratio
Mbp	mega base pairs
MeOH	methanol
MHz	Megahertz
MIBiG	Minimum Information about a Biosynthetic Gene cluster
MIC	minimal inhibitory concentration
min	minute
mL	milliliter
MLSA	multilocus sequence analysis
MS	mass spectrometry
<i>N.</i>	<i>Nocardia</i>
NCBI	National Center for Biotechnology Information
NMR	nuclear magnetic resonance spectroscopy
NOESY	nuclear Overhauser effect spectroscopy
NRPS	non-ribosomal peptide synthetase

Pfam	protein family
PKS	polyketide synthase
ppm	parts per million
R	coefficient of correlation
SI	supporting information
sp.	species
spp.	plural of species
SSN	sequence similarity network
T	thiolation
TIC	total ion chromatogram
TOCSY	total correlation spectroscopy
UID	unique identifier
Δ	deletion

List of Publications

Publications:

Chen, J.; Frediansyah, A.; **Männle, D.**; Straetener, J.; Brötz-Oesterhelt, H.; Ziemert, N.; Kaysser, L.; Gross, H.;

“New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A-C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406.”

ChemBioChem **2020**, 10.1002/cbic.202000062.

Männle, D.; McKinnie, S. M. K.; Mantri, S. S.; Steinke, K.; Lu, Z.; Moore, B. S.; Ziemert, N.; Kaysser, L.;

“Comparative Genomics and Metabolomics in the Genus *Nocardia*”

mSystems **2020**.

Oral presentations:

Männle, D.; McKinnie, S. M. K.; Steinke, K.; Lu, Z.; Moore, B. S.; Ziemert, N.; Kaysser, L.;

“Comparative Genomics to Metabolomics: Nocobactin Production in *Nocardia*”

International VAAM-Workshop 2019: Biology of natural product producing microorganisms.
September 2019, Jena (Germany)

Poster presentations:

Chen, J.; **Männle, D.**; Ziemert, N.; Gross, H.;

“Genome-driven isolation of nocobactin NA derivatives from *Nocardia terpenica* IFM 0406.”

International VAAM-Workshop 2017: Biology of natural product producing microorganisms.
September 2017, Tübingen (Germany)

Männle, D.; Ziemert, N.; Kaysser, L.;

“Genome-guided Drug Discovery”

International VAAM-Workshop 2016: Biology of natural product producing microorganisms.
September 2016, Freiburg (Germany)

Zusammenfassung

Ursprüngliche Methoden der Entdeckung neuer Naturstoffe aus Bakterien, waren zunächst davon abhängig, diese im Labor zu kultivieren, wodurch das Auffinden neuer chemischer Strukturen sehr arbeits- und zeitintensiv war. Oftmals kam es dabei zur Wiederentdeckung schon bekannter Naturstoffe. Die heutzutage stetig steigende Verfügbarkeit an Genomsequenzen, kommt für Wissenschaftler einem wahren Schatzfund gleich. Mittels bioinformatischen Genanalyseprogrammen wie antiSMASH, können Biosynthesewege aus genetischen Sequenzen identifiziert werden.¹ Leider ist es mit solchen automatisierten Programmen schwierig herauszufinden, wie sich geringfügige genetische Unterschiede in Biosynthesegenclustern, auf die daraus resultierenden chemischen Strukturen auswirken. Dies erschwert die Priorisierung von Stämmen und kann oftmals zu falschen Vorhersagen chemischer Strukturen führen.

Im Rahmen dieser Doktorarbeit wurde das metabolische Potential von *Nocardia*, einem wenig untersuchten, aber vielversprechendem actinobakteriellen Genus, mittels genomgestützter Vorgehensweise, untersucht.² Durch BiG-SCAPE (Biosynthetic Genes Similarity Clustering and Prospecting Engine) konnten Sequenzähnlichkeitsnetzwerke generiert werden, die eine deutliche Unterteilung in Genclusterfamilien zeigten und eine Fülle an biosynthetischen Genclustern zahlreicher Klassen offenbarten (z.B. Polyketid-, nicht-ribosomale Peptid- und Terpenoidstoffwechselwege).³ Exemplarisch wurde untersucht, wie sich geringfügige genetische Unterschiede in hoch konservierten nocobactinähnlichen Biosynthesestoffwechselwegen, auf die chemisch strukturelle Diversität auswirken.⁴

Die Analyse ausgewählter Nocardia-Stämme mittels LC-MS (Flüssigchromatographie mit Massenspektrometrie-Kopplung) und GNPS (Global Natural Product Social Molecular Networking) konnte zeigen, dass sich geringfügige Unterschiede in nocobactinähnlichen Genclustern, bestimmten nocobactinähnlichen Siderophoren zuordnen lassen.^{5, 6} Die neuartigen Siderophore Terpenibactin A-C wurden mittels LC-MS, NMR-Spektroskopie (Kernspinresonanzspektroskopie) und Bioassays charakterisiert.⁷ Des Weiteren wurden neuartige lösliche Nocobactine mittels LC-MS identifiziert. Interessanterweise produzierten manche Nocardia-Stämme mycobactinähnliche Siderophore. Durch NMR-Spektroskopie konnte die zugehörige chemische Struktur aufgeklärt werden, welche eine hohe Ähnlichkeit zu bekannten Virulenzfaktoren aus Mycobakterien aufwies.^{4, 8}

Das folgende genomgestützte Verfahren verdeutlicht das hohe Potential des Genus *Nocardia* und weist auf die konstitutive Rolle der Nocobactine hin. Diese Doktorarbeit setzt den Grundstein für zukünftige *Nocardia* Genanalysen, zukünftige Untersuchungen anderer seltener Actinobakterien, verhindert Wiederentdeckung von Naturstoffen und macht Priorisierung von Stämmen möglich.

Summary

Traditional methods for natural product discovery relied exclusively on cultivation of bacteria in the laboratory. Discovery of novel chemistry was laborious and time consuming and frequently resulted in the rediscovery of known natural products. The increasing availability of genomic sequences represents a huge treasure trove for researchers. Biosynthetic pathways can be readily identified from sequence data using genome mining tools such as antiSMASH.¹ However, using automated bioinformatic genome analysis tools, it is often unclear to what degree genetic variability in homologous biosynthetic pathways relates to chemical structural diversity. This makes prioritization of microbial strains and compound identification extremely difficult and often leads to incorrect prediction of natural product structural diversity.

In the scope of this thesis, the metabolic potential of *Nocardia*, an under-investigated but highly prolific actinobacterial genus, was assessed using a genomics-guided approach.² Sequence similarity networks generated by BiG-SCAPE (Biosynthetic Genes Similarity Clustering and Prospecting Engine) showed the presence of distinct gene cluster families including a plethora of biosynthetic gene clusters of various classes including: polyketide; non-ribosomal peptide; and terpenoid pathways.³ Highly conserved biosynthetic pathways encoding for nocobactin-type siderophores were used to exemplify how specific differences in highly related gene clusters correlate to structural diversity in the produced compounds.⁴

Metabolic profiling of selected *Nocardia* strains using LC-MS (liquid chromatography-mass spectrometry) metabolomics data and GNPS (Global Natural Product Social Molecular Networking) revealed related nocobactin-type biosynthetic gene clusters that can indeed be assigned to distinct structural types of nocobactin-type siderophores.^{5, 6} The new nocobactin-type siderophores terpenibactins A-C were characterized using LC-MS, NMR (nuclear magnetic resonance) spectroscopy and bioassays.⁷ Furthermore, novel soluble forms of nocobactin-type siderophores were identified using LC-MS. Interestingly, some *Nocardia* strains produced mycobactin-type siderophores, where NMR spectroscopy revealed a chemical structure highly similar to known virulence factors from mycobacteria.^{4, 8}

The subsequent comparative genomics and metabolomics approach highlights the potential of the highly promising genus *Nocardia* and points out the constitutive role of nocobactins. It can set the foundation for future *Nocardia* genome mining approaches and thorough assessment of other rare actinobacteria, circumventing rediscovery of natural products and facilitating strain prioritization.

Personal Contribution

“Comparative Genomics and Metabolomics in the Genus *Nocardia*”

Männle, D.; McKinnie, S. M. K.; Mantri, S. S.; Steinke, K.; Lu, Z.; Moore, B. S.; Ziemert, N.; Kaysser, L.;

Author contribution:

- **Daniel Männle**
 - Schreiben des Manuskripts
 - Überprüfung und Überarbeitung des Manuskripts
 - Erstellung von Abbildungen und Tabellen
 - Erstellung des SI-Teils
 - Download und Inventarisierung der *Nocardia* Genome
 - Durchführung und Auswertung von *Nocardia* antiSMASH
 - Durchführung und Auswertung von *Nocardia* autoMLST
 - Durchführung und Auswertung von BiG-SCAPE
 - Kultivierung und Erhaltung selektierter *Nocardia*
 - Extraktion der *Nocardia* Kulturen
 - Durchführung und Auswertung von HPLC, LC-MS und MS/MS Experimenten
 - Durchführung und Auswertung von GNPS
 - Probenvorbereitung für NMR
 - Durchführung der NMR Spektroskopie und Analyse
 - Isolation von *Nocardia* gDNA
 - 16S rDNA Sequenzierung zur Ausschließung von Kontaminationen
 - Durchführung und Auswertung von isotoopenmarkierten Studien
- Shaun McKinnie
 - Hilfestellung bei Isolierung der Nocobactine
 - Verantwortliche Durchführung von NMR Spektroskopie und Analyse
 - Schreiben des Manuskripts (NMR-Teil)
- Katharina Steinke
 - Bachelorarbeit zum Thema „Genome mining within the actinomycete genus *Nocardia*“
- Zeyin Lu
 - Erste Kultivierung ausgewählter *Nocardia*
 - Erste HPLC und LC-MS Messungen
 - Masterarbeit zum Thema „Discovery of new siderophores by *Nocardia* spp. genome mining“
- Shrikant Mantri
 - Statistische Auswertung in Bezug auf „pathogenic *Nocardia* BGCs“
- Bradley Moore
 - Betreuung des Projekts

- Nadine Ziemert
 - Schreiben des Manuskripts
 - Hilfestellung bei bioinformatische Fragestellungen
 - Betreuung des Projekts
 - Entscheidend an der anfänglichen Hypothesenbildung beteiligt
 - Entscheidend an allen Diskussionen von Daten und Analysen beteiligt
- Leonard Kaysser
 - Schreiben des Manuskripts
 - Erstellen von Abbildungen
 - Betreuung des Projekts
 - Entscheidend an der anfänglichen Hypothesenbildung beteiligt
 - Entscheidend an allen Diskussionen von Daten und Analysen beteiligt

“New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A-C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406”

Chen, J.; Frediansyah, A.; **Männle, D.**; Straetener, J.; Brötz-Oesterheld, H.; Ziemert, N.; Kaysser, L.; Gross, H.;

Author contribution:

- Julia Chen
 - Isolation der terpenibactine A-C
 - Durchführung und Auswertung von LC-MS Experimenten der terpenibactinen A-C
 - Strukturaufklärung der Terpenibactine mittels NMR
 - Erstellen von Abbildungen
 - Überprüfung und Bearbeitung des Manuskripts
- Andri Frediansyah
 - Durchführung und Auswertung des „Muscarinic M3 receptor calcium flux assay“
 - Überprüfung und Bearbeitung des Manuskripts
- **Daniel Männle**
 - **Hilfestellung bei bioinformatischen Rückfragen**
 - **Identifikation und Lokalisation eines *nbtF*-ähnlichen Gens auf dem Genom von *N. terpenica* IFM 0406**
 - **Überprüfung und Bearbeitung des Manuskripts**
- Jan Straetener
 - Durchführung und Auswertung von Cytotoxizitäts und mikrobiellen Inhibitionsassays
- Heike Brötz-Oesterheld
 - Überprüfung und Bearbeitung des Manuskripts
- Nadine Ziemert
 - Überprüfung und Bearbeitung des Manuskripts
- Leonard Kaysser
 - Überprüfung und Bearbeitung des Manuskripts
- Harald Gross
 - Schreiben des Manuskripts
 - Erstellen von Abbildungen
 - Betreuung des Projekts

1. Introduction

1.1. Background Information

1.1.1. Natural product discovery

Natural products are small organic compounds produced by living organisms. The ecological roles can be as diverse as their chemical structures and include defense, communication and nutrient acquisition.⁹ Natural products can be derived from primary, but are mainly derived from secondary metabolism. Secondary metabolites are considered non-essential for growth or reproduction of bacteria and were thought to give the producer a survival advantage over others.¹⁰

Natural products are highly valuable and many of today's most important compounds for medical applications or the pharmaceutical industry.¹¹⁻¹³ These unparalleled successful sources for drug discovery find use as antibiotics, immunosuppressants or anticancer compounds.^{11, 14} New antibiotics are urgently needed, as bacteria gain resistance to routinely used antibiotics quickly.¹⁵ The unique chemical stereostructures of some natural products make them extremely hard or impossible to synthesize.¹⁶ Therefore, scientists often rely on microorganisms that naturally produce natural products and often use them as precursors for commercial drug products.¹⁷

1.1.2. The phylum Actinobacteria

One of the most promising sources for natural products is the bacterial phylum of Actinobacteria.¹⁸ It includes the deadliest bacterial pathogen *Mycobacterium tuberculosis*, which is the third leading cause of death worldwide after deaths by the HI-Virus and the human pathogen *Corynebacterium diphtheriae*.¹⁹ However, the phylum also contains *Corynebacterium glutamicum*, the industrial producer of glutamic acid and contains the genus *Streptomyces* that has been known as a rich resource for discovery of secondary metabolites and is the source of over half of the bioactive metabolites from bacteria used in medicine.^{13, 20, 21} Underexplored bacterial taxa were shown to harbor prolific producers of novel bioactive chemistry.²² *Nocardia* is one underexplored genus from the order Actinomycetales with potential for production of bioactive compounds.²³ They are closely related to *Corynebacterium*, *Rhodococcus* and *Mycobacterium*. *Nocardia* are aerobic, Gram-positive and partially acid-fast actinomycetes, ubiquitously distributed in both aquatic and terrestrial habitats.²⁴⁻²⁶ These saprophytes are able to form branching and filamentous mycelia that can fragment into rod-shaped bacteroid non-motile elements.²⁷

1.1.3. Pathogenicity of *Nocardia*

Originally, *Nocardia* were foremost known for their facultative and opportunistic pathogenicity. They were first described by Edward Nocard in 1888 as causative agents of severe infections of the skin, lung or central nervous system in humans and animals causing nocardiosis and have been extensively studied by Beaman and others.²⁸⁻³¹ Nocardiosis mainly affects immunocompromised patients and leads to high mortality rates. Nevertheless, nocardiosis was also reported from immunocompetent patients.² Despite many *Nocardia* species being facultative pathogens, some species are clinically more frequently associated with nocardiosis patients than others.^{30, 32-34} These strains are further referred to as group 1 pathogens and include *N. asteroides*, *N. abscessus*, *N. brasiliensis*, *N. cyriacigeorgica*, *N. farcinica*, *N. nova*, and *N. otitidiscaviarum*. Further frequently mentioned pathogens include *N. africana*, *N. brevicatena*, *N. kruczakiae*, *N. paucivorans*, *N. transvalensis*, and *N. veterana*. Group 1 pathogenic *Nocardia* are analyzed in this study.

1.1.4. Natural products from *Nocardia*

Few studies have focused on secondary metabolites and respective biosynthetic genes in *Nocardia* strains.² Reports of natural products produced by members of *Nocardia* are reported scarcely and include the β -lactone antibiotics nocardiolactone, nocardicin A and B and the immunosuppressive terpenoid antibiotic brasiliardin A amongst few other natural products (Figure 1).³⁵⁻³⁷

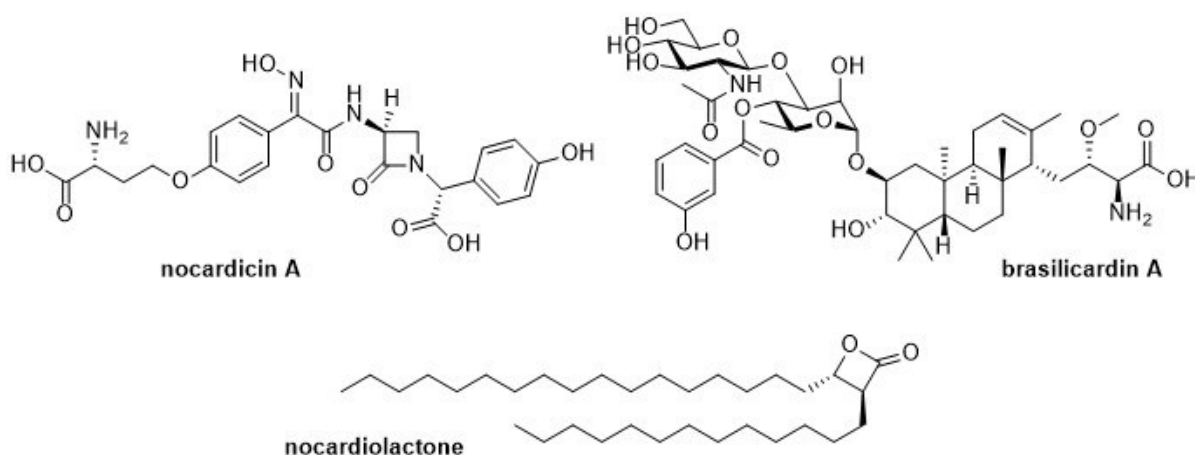


Figure 1: selected known natural products produced by members of the genus *Nocardia*. Adapted from Männle et al. 2020.⁶

1.2. Siderophores

1.2.1. General features and application of siderophores

Natural products include important chelating molecules, so-called siderophores that are one of the few examples of bacterial secondary metabolites where the ecological function is known.^{38, 39} Siderophores not only play an important role in microbial iron acquisition, they are also known to be virulence factors associated with host colonization in a host-pathogen competition for scarce iron (e.g. in *Mycobacterium*).^{40, 41} However, interest in the investigation of bacteria extends beyond a siderophore's capacity to bind ferric iron for bacterial survival. In fact, siderophores can find application in the field of medicine, as part of antibiotic drug conjugates and iron overload therapies.⁴² They are also used in bioremediation of polluted habitats.⁴³

1.2.2. Nocobactin NA and nocobactin derivatives from *Nocardia*

The lipophilic, membrane-associated siderophore nocobactin NA (Figure 2) has been first described in 1974 and was found as a UV-active siderophore from cultures of *Nocardia asteroides* ATCC 3318 (later reclassified as *Nocardia farcinica* IFM 10152) by Ratledge and Snow.⁴⁴ Structurally, nocobactin NA-a and NA-b share a 2-hydroxyphenyl-5-methyl-oxazoline or -oxazole moiety, a *N*-acetylated or -formylated *N*-hydroxylysine that is linked via an ester moiety to a long-chain α -methyl, β -hydroxy fatty acid, and a C-terminal *N*-hydroxy α -amino ϵ -caprolactam ring (Figure 2). Furthermore, cytotoxicity of nocobactin NA has been observed to be higher against cancer cell lines than normal cells, which makes the nocobactins also interesting as anticancer compounds.⁴⁵ Nocobactins were reported from other *Nocardia* spp..⁴⁶ Differently substituted derivatives of cytotoxic nocobactin NA-a and NA-b have been found and include amamistatin A and B, formobactin, brasilibactin A, and nocardimicins A-I (Figure 2).^{7, 44, 45, 47-51} Nocobactin derivatives show variation in the substitution pattern of the 2-OH-phenyl residue, of the oxazole ring system, of the *N*-hydroxylysine residues and in the nature of the fatty acid. It has been reported that nocobactin NA and nocardimicins A-F feature an acetyl moiety, whereas nocardimicins G-I, amamistatin A and B, brasilibactin and formobactin feature a formyl moiety. Nocobactin-type siderophores exhibit two types of heterocyclic ring variants. Nocobactin NA, formobactin, the amamistatins, and nocardimicins A-F show oxazole moieties, whereas brasilibactin A and nocardimicins G-I contain oxazoline moieties. Nocobactin-type siderophores have already shown outstanding cytotoxic properties (nocobactin NA-a and NA-b, brasilibactin A) and have been described as growth inhibitors of tumor cell lines (amamistatins A and B).^{45, 49, 50} Furthermore, the nocardimicins have been found to effectively inhibit muscarinic M3 receptor binding, a target in the treatment of certain respiratory and gastrointestinal disorders.^{47, 48}

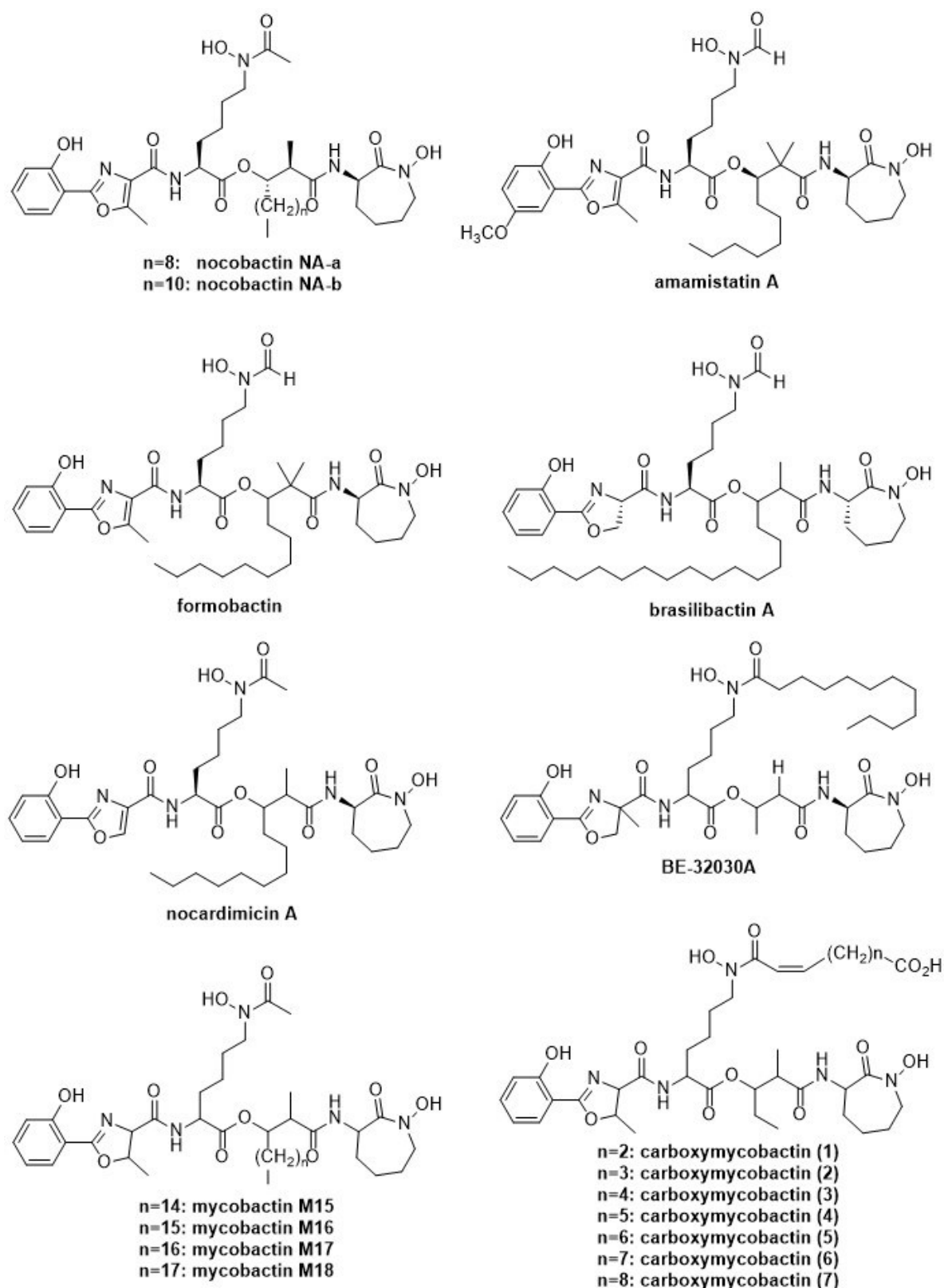


Figure 2: selected nocobactin-type and mycobactin-type siderophores by *Nocardia* spp.. Nocobactin NA-a and Na-b are produced by *Nocardia farcinica* IFM 10152 (formerly known as *Nocardia asteroides* ATCC 3318).^{4, 44} Amamistatin A and B are produced by *Nocardia asteroides* SCRC-A2359.⁴⁹ Formobactin is produced by *Nocardia* sp. ND20.⁵¹ Brasilibactin A is produced by *Nocardia brasiliensis* IFM 0995.^{50, 52} Nocardimicins A-F are produced by *Nocardia* sp. TP-A0674 and nocardimicins G-I are produced by *Nocardia nova* JCM 6044.^{47, 48} BE-32030A-E (mycobactin-type siderophores) are produced by *Nocardia* sp. A32030.⁵³ Mycobactin M15, M16, M17 and M18 are produced by *Mycobacterium marinum*.⁸ Carboxymycobactin is produced by *Mycobacterium avium* CR1/69.⁵⁴

Introduction

Despite *Mycobacterium* mainly producing closely related mycobactins, nocobactins also have been reported previously from *Mycobacterium marinum* (termed as mycobactin M).⁸ Vice-versa, BE32030 A-E are the only known mycobactin-type siderophores reported from *Nocardia* sp. A32030 where no associated BGC (biosynthetic gene cluster) was reported (Figure 2).⁵³ Furthermore, soluble forms of mycobactins, the so-called carboxymycobactins were reported that feature a carboxylic acid moiety attached to the fatty acyl chain.⁵⁴⁻⁵⁶ Several other structurally similar siderophores have been reported from *Nocardia* spp. and other closely related genera including asterobactin, heterobactin A, JBIR-16, nocardichelin A, transvalencin Z, and ornibactin-C4 (Figure 3).⁵⁷⁻⁶²

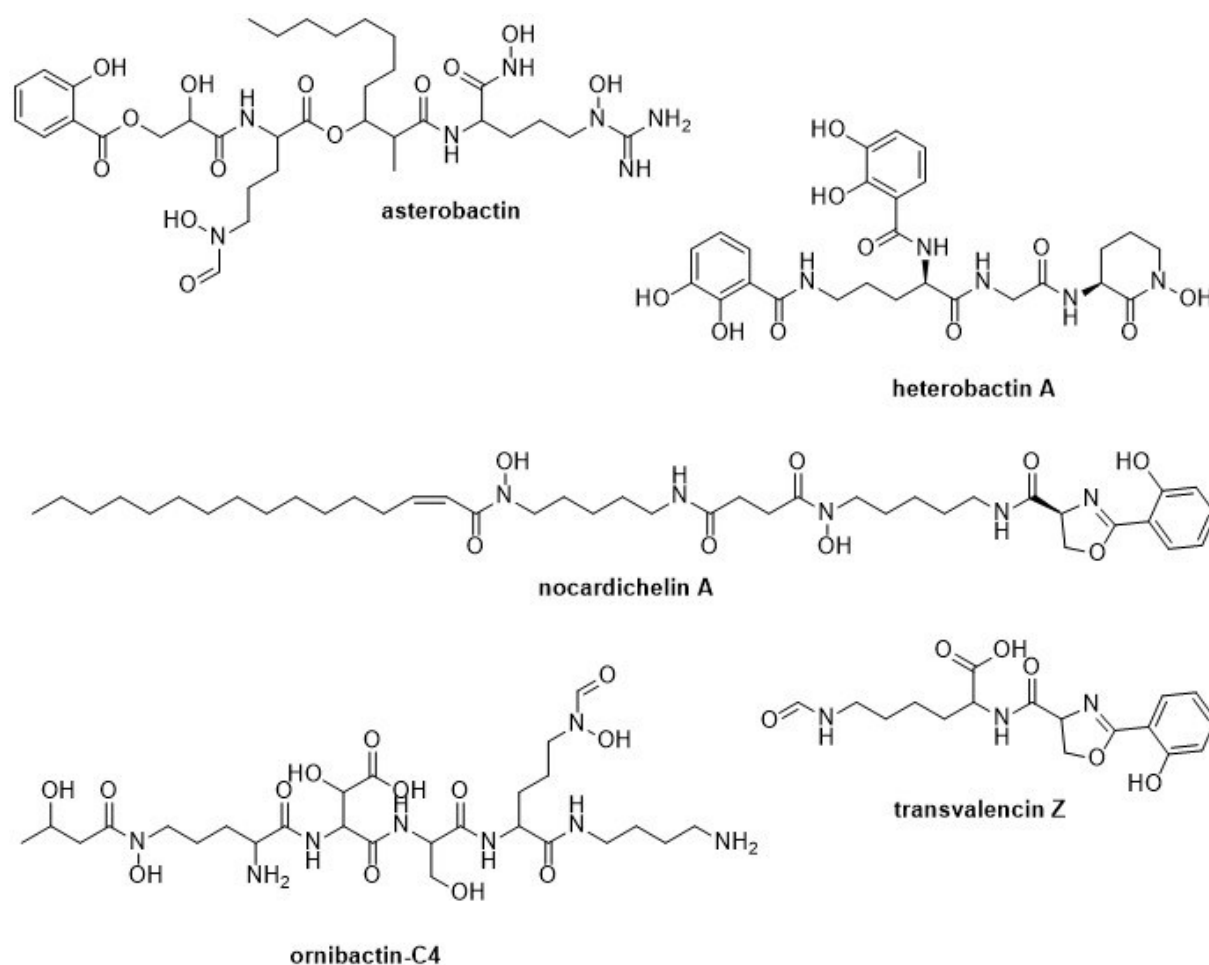


Figure 3: siderophores produced by *Nocardia* spp. and closely related genera. Asterobactin is produced by *Nocardia asteroides* IFM 0959⁵⁷. Heterobactin A is produced by *Rhodococcus erythropolis* PR4.⁵⁸ Nocardichelins A and B are produced by *Nocardia acta* 3026.⁶⁰ Ornibactin-C4, -C6, and -C8 are produced by *Pseudomonas cepacia*.⁶¹ Transvalencin Z is produced by *Nocardia transvalencis* IFM 10065.⁶²

1.2.3. Nocobactin BGC and nocobactin biosynthesis

The biosynthetic pathway for nocobactin NA was first reported from *Nocardia farcinica* IFM 10152 by Hoshino et al.⁴ It was found to consist of ten genes distributed between two genetic loci, whereof the first locus consists of eight genes (subcluster I) and the second locus involves two genes (subcluster II) located ~180 kb upstream of subcluster I. This split BGC involves a non-linear assembly line consisting of three non-ribosomal peptides (NRPSs) and one type I polyketide synthase (PKS) module (Figure 4). Subcluster I consists of *nbtA*, *nbtB*, *nbtC*, *nbtD*, *nbtE*, *nbtF*, *nbtG* and *nbtH* biosynthetic genes, whereas subcluster II consists of *nbtS* and *nbtT* genes. The nocobactin NA BGC is deposited in the MIBiG (Minimum Information about a Biosynthetic Gene cluster) database, a repository for BGCs linked to its respective natural product.⁶³ It is to mention that subcluster II is not deposited in MIBiG.

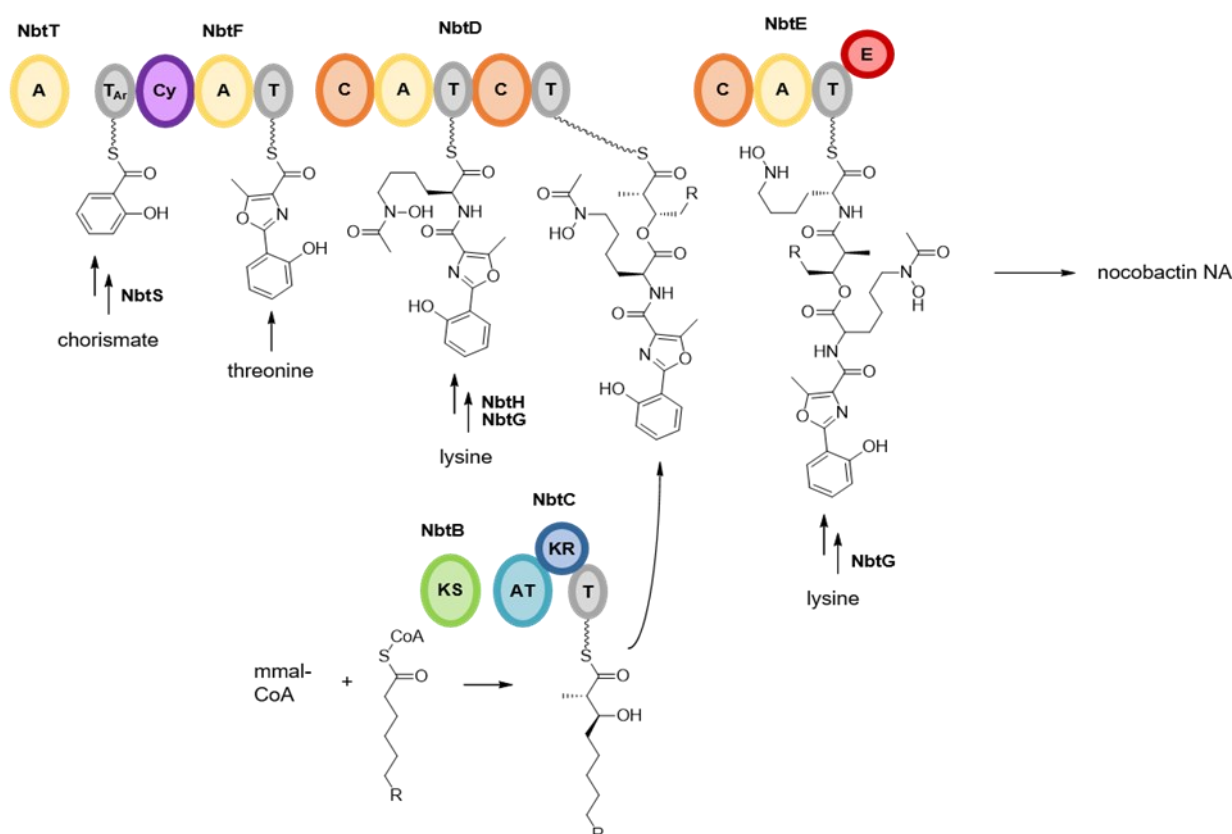


Figure 4: Model for nocobactin NA biosynthesis. NRPS/PKS domains: adenylation (A), thiolation (T), heterocyclization (Cy), condensation (C), epimerization (E), ketosynthase (KS), acyltransferase (AT), ketoreductase (KR). NbtA (type II thioesterase) not shown, which presumably rids aberrantly loaded substrates from the PKS or NRPS.^{4, 64} Adapted from Männle et al..⁶

Nocobactin NA biosynthesis commences by loading salicylate, synthesized from chorismate by the putative salicylate synthase NbtS and activated by NbtT, onto the ACP domain of NbtF.⁴ NbtF then catalyzes the condensation and heterocyclization of salicylic acid and threonine. Subsequently, NbtD condensates lysine with the growing chain. NbtH and NbtG likely direct N-acylation and N-hydroxylation, respectively. To this point, it is unclear whether this occurs before or after incorporation of the lysine moiety. NbtB and NbtC are proposed to activate a fatty acid, which is

extended by Claisen-type condensation with methylmalonyl-CoA, subsequently reduced at the β -position.⁴ The incorporation of the long chain fatty acyl moiety by ester linkage is probably mediated by NbtD before NbtE attaches another lysine. Epimerization of lysine to the D-isomer is followed by *N*-hydroxylation and formation of the ϵ -caprolactam.⁶⁵ Based on sequence analysis, NbtA is presumed to be a thioesterase type II, which presumably rids aberrantly loaded substrates from the PKS or NRPS.^{4, 64} Interestingly, Hoshino and co-workers state that a $\Delta nbtE$ mutant was found to be impaired in cytotoxicity against J774A.1 cells, suggesting that nocobactin NA production is required for virulence of *Nocardia farcinica* IFM 10152 similar to mycobactin-conferred virulence in *Mycobacterium tuberculosis*.^{4, 41, 66} Notably, both structurally and biosynthetically the nocobactins closely resemble the mycobactins from *Mycobacterium*. Consequently, as nocobactin NA has a considerable high similarity to this major virulence factor in *M. tuberculosis*, nocobactins in general may have a similar function as virulence factors in *Nocardia*, contributing to the progression of nocardiosis.^{2, 4}

The biosynthesis of mycobactins has been studied intensively.⁶⁷⁻⁶⁹ Mycobactins and nocobactins differ in the position and chain length of their long chain fatty acyl substituents. In mycobactins, fatty acyl chains are attached to the *N*-hydroxylysine moiety whereas nocobactin-like siderophores feature *N*-formylated/*N*-acetylated *N*-hydroxylysine residues. Both membrane-bound mycobactins and extracellular carboxymycobactins are known to be important for the lifestyle of pathogenic *Mycobacterium* strains and have been shown to play a fundamental role in survival and growth of *M. tuberculosis* in human hosts.^{55, 66, 70}

1.3. Bioinformatics for natural product discovery

1.3.1. Preliminary genomic studies on the genus *Nocardia*

In 2014, Komaki and coworkers already analyzed seven *Nocardia* genomes and found that *Nocardia* strains carry as many PKS-I and NRPS gene clusters as those of *Streptomyces* strains.² A hidden potential of putative new natural products from a bacterial actinomycete genome can exceed >10% of some genomes.^{71, 72} Furthermore, they state that the number of PKS-I and NRPS BGCs in *Nocardia* strains varies substantially depending on species, with *N. brasiliensis* strains carrying the largest numbers of BGCs.² Two prevalent pathogens namely *Nocardia farcinica* IFM 10152 and *Nocardia brasiliensis* HUJEG-1 were analyzed by Komaki and coworkers with a genome size of 6.0 Mb and 9.4 Mb, respectively. These genome sizes are similar to those of representative *Streptomyces* strains (5.0 – 11.9 Mb).² The authors further suggest that there is no clear relation between genome size and pathogenicity in *Nocardia* strains, presumably, due to their nature as facultative pathogens. In comparison, pathogenic *Mycobacterium* and *Rhodococcus* strains were found to possess smaller genomes, less BGCs, and less diverse BGCs, due to their adaptation to hosts.^{73, 74} Living inside a host, reduces the need to adapt to sudden environmental changes or to fight off microbial competitors, reducing the need of different biosynthetic pathways that contributes to genome minimization.⁷⁵ Komaki and coworkers highlight *Nocardia* as highly promising resource in the search of novel secondary metabolites.² Accompanied by the ever increasing amount of publicly available *Nocardia* genome sequencing data, more than 100 *Nocardia* genomes are now freely available from the National Center for Biotechnology Information (NCBI) or the Joint Genome Institute – Integrated Microbial Genomes & Microbiomes (JGI-IMG) database, representing a treasure trove for natural product discovery.⁷⁶⁻⁸¹ A recent study on the closely related genus *Rhodococcus* also emphasizes the huge genomic potential of Actinobacteria for secondary metabolite production.⁸²

1.3.2. Brief history of genome mining

In the pregenomic era, traditional methods for natural product discovery relied on cultivation of bacteria in the laboratory and discovery of novel chemistry was laborious and time consuming. Rediscovery of already known natural products was highly likely. Furthermore, due to strict regulation mechanisms under standard laboratory conditions, the full secondary metabolite potential of a bacterial genome is not reflected.

Today, genome mining greatly facilitates the prioritization of finding novel chemistry.⁸³ With (next-generation) genome sequencing on the rise, genomes can be mined for their potential to produce secondary metabolites.⁸⁴ Genome mining has set the foundation for the identification of an enormous diversity of natural products even within closely related species.⁸⁵ It enables researchers

to uncover the hidden potential of a genome and facilitates the discovery of novel secondary metabolites through computational approaches.^{83, 86}

Genome mining has revealed that the potential of bacterial genomes goes far beyond what bacteria produce under laboratory conditions.⁸⁷⁻⁸⁹ Genes encoding biosynthetic pathways for natural products are often organized in BGCs.^{20, 90, 91} Surprisingly, even well studied strains were found to harbor many more BGCs than compounds therefrom discovered.^{84, 91, 92} However, many BGCs remain transcriptionally inactive or 'silent' under standard laboratory conditions.^{93, 94} Doroghazi and coworkers recently observed that BGCs from one gene cluster family (GCF) were silent in 77% of their strains examined.⁹⁵ A GCF is defined as a group of BGCs that exhibits similar BGC organization with high sequence similarity.⁹⁶

1.3.3. Bioinformatic tools for natural product discovery

It is essential to prioritize and dereplicate putative BGCs bioinformatically, as later rediscovery of already known natural products can be frustrating, expensive, and time consuming. Recently, suitable genome mining tools have been developed to circumvent rediscovery and to simplify the search for new natural products by rapid straightforward detection of secondary metabolite BGCs.⁹⁷⁻¹⁰⁰ Currently, the gold standard in genome mining is antiSMASH that allows a rapid genome-wide identification, annotation, and analysis of secondary metabolite BGCs in bacterial and fungal genomes.^{1, 98, 101} Several other tools have been developed, such as BAGEL 3, PRISM 3, RODEO, and SMURF facilitating genomics based discovery of secondary metabolites.¹⁰²⁻¹⁰⁵

The bioinformatics tool BiG-SCAPE is able to read putative BGC GenBank files predicted by antiSMASH.³ BiG-SCAPE predicts Pfam composition similarity by hmmscan from the HMMer suite and a pairwise distance calculation.¹⁰⁶ Pfam is a large collection database of protein families, represented by multiple sequence alignments using profile hidden markov models (HMMs).^{107, 108} A pairwise BGC distance matrix is calculated based on a weighted combination of Jaccard Index, Adjacency Index, and Domain Sequence Similarity Score. In principle, BiG-SCAPE creates a distance matrix from protein domain content, copy numbers, and sequence similarity comparable to a BLAST alignment that can be readily imported into sequence similarity networking tools, such as Cytoscape or Gephi.^{71, 109-111} Sequence similarity networks (SSNs) have been widely utilized to compare BGCs from large sequencing datasets.^{71, 81, 85, 95}

The MIBiG database can be a valuable repository linking BGCs to their natural products. However, MIBiG only links known BGCs to its known compounds. BGCs that have not yet been linked to a natural product have been termed 'orphan'.¹¹² For the purpose of easier linkage of BGCs to compounds, a new community-based database called "iOMEGA Paired Data platform" was created

(unpublished, personal contribution). The platform links BGCs to tandem mass spectrometry data (MS/MS) for structure predictions. Its intention is to link molecules to their producers, find large scale genome-metabolome associations, use genomic data to assist in structural elucidation of molecules and provide a centralized database for paired datasets. Findings of this thesis are already integrated into iOMEGA to contribute and accelerate the field of natural products discovery.

1.3.4. Challenges of bioinformatic predictions for natural product discovery

Although genome mining has led to various natural product discoveries to date, its promise of revolutionizing natural product discovery has yet to be fulfilled on a large scale, as the process of mining genetically encoded natural products is not keeping pace with the rate of genome sequencing.⁹⁶ One of the major bottlenecks in genome mining is that it takes a significant amount of time and money to connect a molecule to its biosynthetic signature.⁹⁶

Bioinformatic prediction tools such as antiSMASH can be reliably used to predict functions of genes, but predictions of new compound classes or chemical structures remain challenging.¹¹³ Tools are forced with challenges related to the complex and repetitive nature of the biosynthetic genes involved and more knowledge about enzymatic functionalities must be generated.⁹⁵ In addition, quality of BGC predictions are highly dependent on quality of the input data, where difficulties can arise when genome sequences are highly fragmented or only partial sequences are available.^{1, 95} Vice-versa, antiSMASH uses a conservative approach and rather overpredicts BGC borders, resulting in hybrid BGCs, consisting of more than one BGC resulting in falsified chemical structure predictions. BLAST and Pfam searches for domains in the so-called core genes of polyketide synthase (PKS) and nonribosomal peptide synthetase (NRPS) have served as powerful means in identifying BGCs.¹¹⁴ AntiSMASH predictions, first identify core genes using their conserved motifs and then extends predictions to flanking genes encoding tailoring enzymes including hydroxylases, oxidases, methylases, transcription factors, and transporter genes. Ziemert and coworkers stated that differences in BGCs within a GCF occurred largely among genes predicted to encode tailoring enzymes.⁸⁵ Despite resulting in minor chemical structure differences, it can have a major impact on biological activity. As tailoring enzymes are often encoded at distal positions within a BGC and BGC border prediction remains challenging, it is often unclear, whether a tailoring enzyme still belongs to a BGC, hampering precise chemical structure prediction. Thus, it remains extremely difficult to bioinformatically predict the correlation of genetic variability within a GCF in respect to its chemical structural diversity. Fortunately, advances in analytical mass spectrometry are revolutionizing the ways in which natural products are detected and first attempts were made, linking chemical molecular families to respective GCFs of biosynthetic pathways for refined natural product discovery.^{86, 96}

2. Goals

This thesis pursues the goal to uncover the biosynthetic potential of the bacterial genus *Nocardia* by application of suitable bioinformatic tools for strain prioritization in natural product discovery. Until now, it remains unclear to what extent genetic variability in closely related biosynthetic gene clusters relates to chemical structural diversity. Overall, my thesis aims to reveal what adjustments need to be made in bioinformatic tools for genetic variability reflection to accurately reflect the chemical structural diversity.

In detail, a comparative genomics to metabolomics approach in the genus *Nocardia* is conducted by evaluating state of the art computational approaches for their application in natural product research. The genus *Nocardia* is examined bioinformatically for its capability to produce novel chemistry. In particular, genetic variability in the biosynthetic gene cluster family of nocobactins will be highlighted by definition of thresholds in BiG-SCAPE sequence similarity networks to accurately reflect the diversity of metal-chelating molecules produced. In cooperation, nocobactin-type terpenibactins produced by *Nocardia terpenica* IFM 0406 will be characterized by MS/MS, NMR, and bioassays. Further molecular diversity of nocobactin-like siderophores produced by selected *Nocardia* strains is uncovered by MS/MS and visualized in GNPS (Global Natural Product Social Molecular Networking). The resulting chemical structural diversity of this siderophore family is utilized to evaluate user-defined BiG-SCAPE thresholds for the prediction of gene cluster families and biosynthetic gene cluster subfamilies. Thresholds that most exactly represent genetic variability of nocobactin-type BGC subfamilies are identified that best reflect chemical structural diversity of nocobactin-type siderophores.

3. Results & Discussion

3.1. Comparative Genomics and Metabolomics in the Genus *Nocardia*

3.1.1. Phylogeny and biosynthetic potential of the genus *Nocardia*

103 non-redundant *Nocardia* genomes were downloaded from the NCBI or the JGI-IMG database.⁷⁶⁷⁷ In order to assess the evolutionary relationships between *Nocardia* genomes, a multilocus sequence analysis (MLSA) was performed using autoMLST with implemented JSpecies ANI analysis.¹¹⁵⁻¹¹⁷ A maximum likelihood phylogenetic tree was generated with autoMLST that identified six major phylogenetic lineages referred to as clades A, B, C, D, E, and F (Figure 5 left). One strain, namely *Nocardia harenae* NBRC 108248 formed a distinct single membered phylogenetic lineage referred to as clade X.

Discrepancies were observed, as two strains, namely *Nocardia* sp. NRRL S-836 and *Nocardia* sp. 348MFTsu5.1 claded with the outgroups such as *Rhodococcus erythropolis* NBRC 15567. Indeed, MASH ANI analysis revealed that these strains belonged to a different genus.¹¹⁸ Estimated average nucleotide identity (ANI) values showed a 91.12% ANI closest to *Lentzea guizhouensis* and 78.91% ANI closest to *Williamsia muralis* NBRC 105860 for *Nocardia* sp. NRRL S-836 and *Nocardia* sp. 348MFTsu5.1, respectively.¹¹⁹ CheckM analysis supported this finding, as both NRRL S-836 and 348MFTsu5.1 fell into marker lineages others from UID1815 (*Nocardia*), namely UID1746 and UID1814, respectively.¹²⁰ As the assignment of NRRL S-836 and 348MFTsu5.1 to the genus *Nocardia* was wrong, they were disregarded from further analysis.^{118, 121} The completeness of all genomes was assessed using checkM and revealed a completeness of at least 98.24% with an average number of contigs of 147 (1 to 615 contigs per genome).

The remaining 101 *Nocardia* strains were isolated from different sources. 50 strains were clinical isolates, 31 were derived from environmental sources (freshwater, seawater, or soil), seven were isolated from animals, whereas six were found as plant pathogens. All *Nocardia* genomes were analyzed for their biosynthetic capability to produce natural products using antiSMASH that revealed a huge diversity of putative BGCs of various classes (Figure 5, right).¹²² *Nocardia* showed a high putative BGC potential, comparable to the potential of *Amycolatopsis* and *Streptomyces*, the latter being the most promising producers of natural products to date.^{74, 123, 124} AntiSMASH revealed an average of 34 BGCs per *Nocardia* genome. This can be compared to the average number of BGCs present in the genus *Streptomyces*, underpinning findings from Komaki and coworkers.² In comparison, the average numbers of BGCs detected in genomes of *Micromonospora* strains were 20 BGCs.¹²⁵

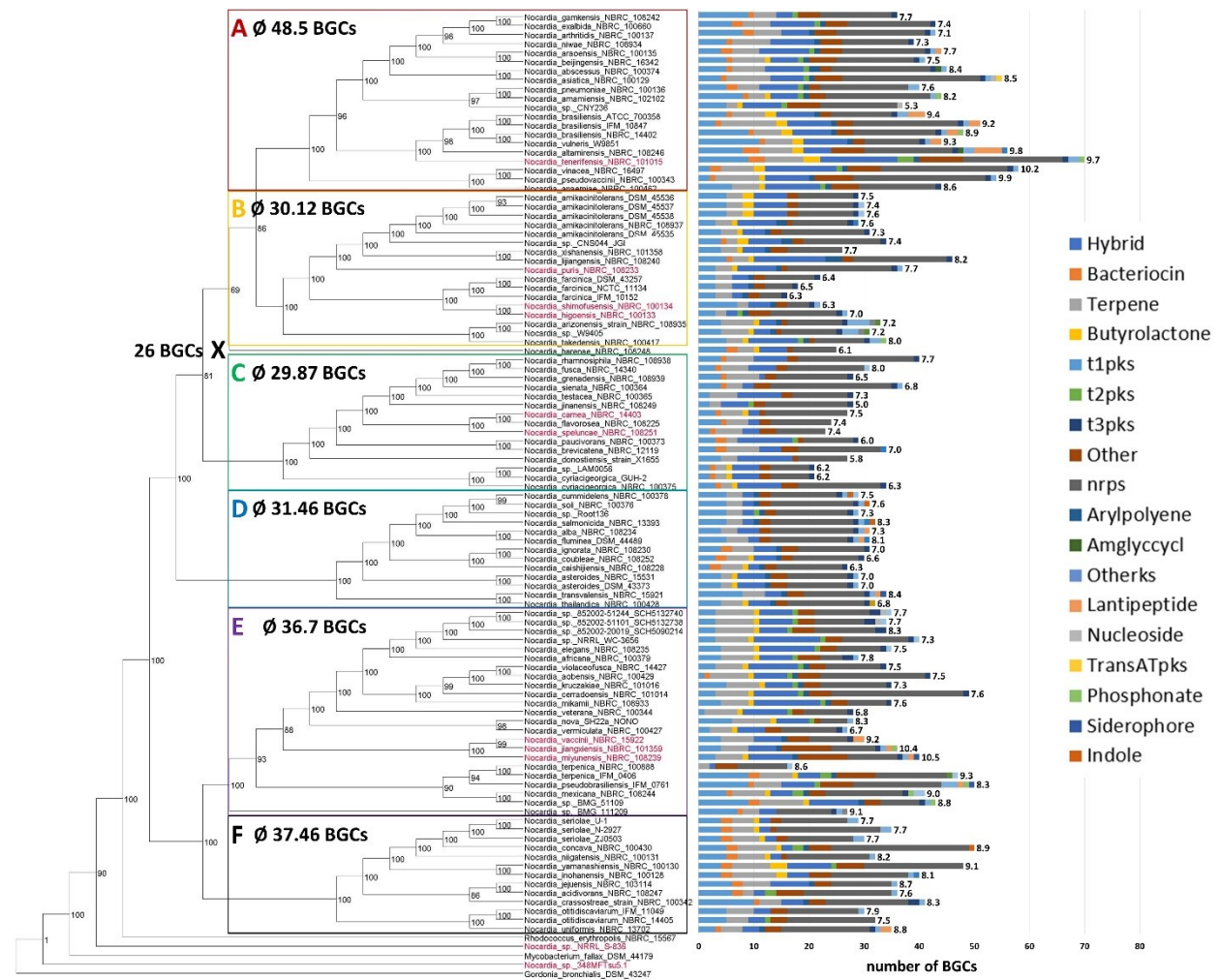


Figure 5: *Nocardia* autoMLST phylogeny and antiSMASH BGC distribution. Maximum likelihood tree of 103 members of the genus *Nocardia* based on 63 concatenated housekeeping genes identified with autoMLST.¹¹⁷ Ultrafast bootstrap values were calculated using 1000 replicates. Numbers next to the bars (predicted BGCs by antiSMASH) represent genome sizes in Mbp. Average numbers of BGCs per clade are shown. X labels a distinct branch formed by a single *Nocardia* strain *Nocardia harenae* NBRC 108248. *Nocardia* genomes marked in red did not contain a nocobactin-like BGC. Adapted from Männle et al..⁶

Numbers of putative BGCs correlated with clades, where clade A had the highest number of BGCs per genome with an average of 48.5 BGCs per genome and where clade C had the least number of BGCs per genome with an average of 29.87 BGCs per genome. Most prevalent BGCs in *Nocardia* were from the class of NRPS.⁷⁴ It can be concluded that *Nocardia* phylogeny is a more important indicator of BGC distribution than geographic origin or source of isolation.^{85, 123}

A positive correlation was found between numbers of BGCs per strain and respective genome size ($R^2=0.4057$; Figure 6). Doroghazi et al. found comparable correlations for selected Actinomycetes such as *Mycobacterium*, *Rhodococci*, and *Streptomyces*.⁸² Larger genomes can accommodate more gene clusters devoted to secondary metabolism.⁸² A similar correlation was also found by Schorn et al., where several rare actinobacterial genera were examined for secondary metabolite production.⁸¹

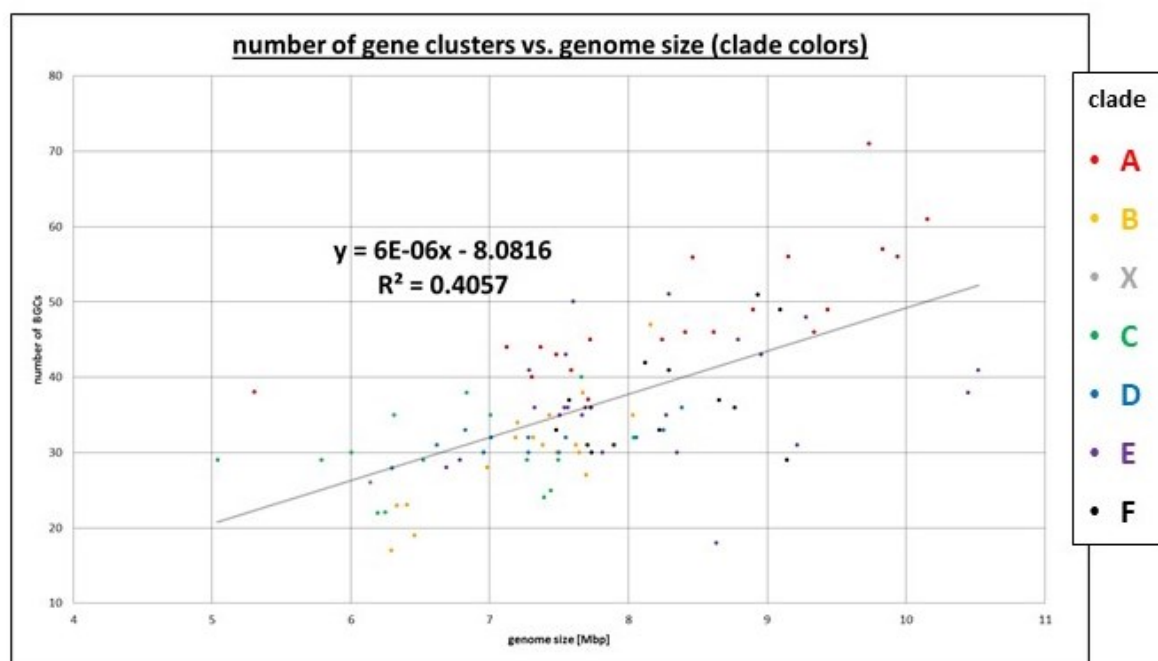


Figure 6: Number of putative *Nocardia* BGCs per strain and respective genome size [Mbp]. Genomes were downloaded from NCBI or JGI-IMG in *.gbk format. Putative BGCs were predicted with antiSMASH 4.0. Genome size was plotted against the number of BGCs. Each dot corresponds to one *Nocardia* strain. Colors indicate phylogenetic clade A-F and X. Trendline indicates a positive correlation between number of BGCs present in a strain and its genome size (coefficient of determination ($R^2=0.4057$)). Adapted from Männle et al. 2020.⁶

When group 1 pathogens (7.6 Mbp; 32.4 BGCs) were compared to the remaining genomes of the genus *Nocardia* (7.8 Mbp; 36.5 BGCs), on average group 1 pathogenic *Nocardia* genomes were smaller and contained less BGCs. *N. farcinica* IFM 10152, a group 1 pathogen, was found with the smallest genome of all *Nocardia* species analyzed and the least number of BGCs. Surprisingly, this finding contradicted results from Komaki and coworkers, who found no clear relation between genome size and *Nocardia* pathogenicity.²

3.1.2. Metabolic diversity and gene cluster families in *Nocardia*

The BiG-SCAPE algorithm was assessed to uncover the diversity of BGCs in the genus *Nocardia* by sequence similarity networks (SSNs). Processed antiSMASH BGCs 101 *Nocardia* strains were incorporated as inputs for BiG-SCAPE.⁷¹ SSNs of various thresholds (25-95%) were generated and visualized in Cytoscape 3.6.1.¹⁰⁹ A *Nocardia* putative BGC SSN with a threshold of 50% is shown in Figure 7A and Figure 8. This 50% SSN consisted of 2836 BGCs as unique nodes that represented single putative BGCs from antiSMASH and 22455 edges. In total, the 50% SSN formed 258 independent gene cluster families (GCFs) of two or more nodes of varying classes of natural product BGCs. A similar threshold of 40% ($c = 0.6$) was used by Schorn et al. for the definition of GCFs in the analysis of rare marine actinomycete genomes.⁸¹

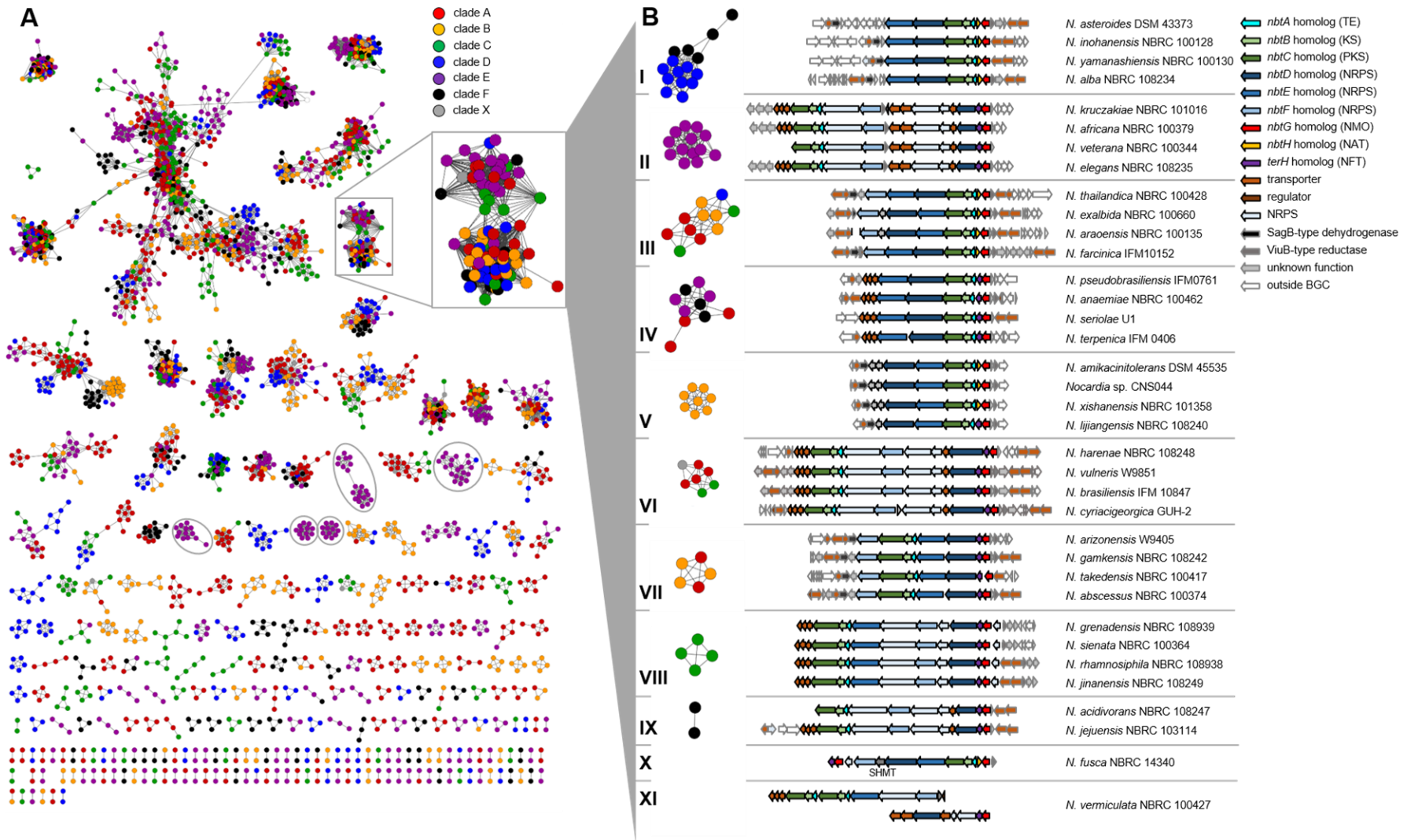


Figure 7: BiG-SCAPE sequence similarity network (SSN) and nocobactin-like subfamilies visualized in Cytoscape. A: Sequence similarity network (SSN; $c = 0.5$) of *Nocardia* spp.. Each node represents one BGC identified by antiSMASH 4.0. Colors represent phylogenetic clades A-F and X. Clade-specific GCFs are highlighted. B: SSN ($c = 0.3$) of nocobactin-like BGCs. Representative pathways of different subfamilies (I-IX). Two singletons (X, XI) are shown. TE=thioesterase, KS=ketosynthase, PKS=polyketide synthase, NRPS=non-ribosomal peptide synthetase, NMO=*N*-monooxygenase, NAT=*N*-acetyltransferase, NFT=*N*-formyltransferase, SHMT=serine hydroxymethyltransferase. Adapted from Männle et al. 2020.⁶

Interestingly, Figure 7A clearly showed a correlation between phylogenetic clades A-F and BGC distribution within the SSN. *Nocardia* strains from clade E encoded a number of BGCs that were uniquely found in this clade and were not found elsewhere. This included BGCs for terpenoid, NRPS, and PKS pathway biosynthesis. A gene cluster with high similarity to hopene biosynthesis was found in all *Nocardia* of clade E and otherwise only in *Nocardia transvalensis* NBRC 15921 (clade D).¹²⁶ In contrast, *Nocardia* spp. from clade F did not form a single characteristic GCF that featured more than seven strains.

Figure 8 showed several large superclusters that consisted of BGCs with different biosynthetic features. For example, putative NRPS BGCs (light blue) clustered with putative NRPS-terpene hybrid BGCs (white) or putative terpene BGCs (blue). These superclusters could not be satisfyingly resolved, even when higher thresholds were applied. Ceniceros et al. reported such problems before and attributed them to deficient separation of individual BGCs by antiSMASH input data.⁷⁴ Hence, a closer manual inspection was necessary and revealed a number of distinct subclusters, e.g. GCF 3-6, GCF 9, and GCF 10 (Figure 8). In addition, the analysis was repeated with the new BiG-SCAPE “glocal” mode option and found no major differences in generated SSNs. In the default “global” mode, all domains from both BGCs were taken into account, whereas in the “glocal” mode, only a subset of the domains was chosen.³

GCFs harbored BGCs that were highly abundant in investigated *Nocardia* genomes (Figure 8). This included several GCFs with mono-modular type I PKS-like ketosynthase pathways, e.g. GCF 1, GCF 11, and GCF 14, potentially involved in the modification of fatty acid precursors.

GCF 1 was highly conserved throughout the genus and likely directed the synthesis of mycolic acids. Mycolic acids are characteristic and essential cell wall components of *Mycobacterium* and related genera, e.g. *Corynebacterium*, *Rhodococcus*, *Gordonia*, and *Nocardia*.¹²⁷

The lower subcluster of GCF 11 contained BGCs homologous to nocardiolactone (Figure 8). Nocardiolactone (Figure 1) is a fatty acid derivative with similarity to lipstatin that has been isolated from different *Nocardia* strains.³⁵ Recently, Robinson et al. reported the identification of the nocardiolactone BGC featuring a β -lactone synthetase.¹²⁸ β -lactone-containing natural products are potent inhibitors of diverse hydrolytic enzymes, hence their biosynthesis has drawn a lot of attention.¹²⁹⁻¹³¹

BGCs from GCF 2 were found in all *Nocardia* genomes and encoded enzymes for the synthesis of ectoine, which stabilizes biopolymers under extreme conditions or osmotic stress.^{74, 132} In fact, the capability to produce ectoine was found in other actinobacterial genera such as *Streptomyces*,

Mycobacterium and in the sister genus *Rhodococcus* before.^{133, 134} GCF 3 included pathways highly conserved throughout *Nocardia* which are likely responsible for the formation of carotenoids of unknown chemical structure. It encoded polyprenyl synthase (CrtE), phytoene synthase (CrtB), and phytoene desaturase (CrtI) homologs for the production of lycopene. An obvious candidate gene coding for a lycopene cyclase was not found within BGCs from GCF 3.

However, a putative lycopene cyclase gene present in all *Nocardia* species was the main feature of the conserved terpene GCF 10. Together, both GCFs could direct the concerted biosynthesis of mono- or bicyclic carotenes. Such monocyclic carotenoid glycosides and their esters have been isolated from *Nocardia kirovani* and different *Rhodococcus* species before.^{135, 136} Another common class of pigments are aryl polyenes, generated by an unusual type II PKS system (GCF 8). These molecules are mostly known from Gram-negative γ -Proteobacteria and Bacteroides and are widespread in *Nocardia* (82 of 101 genomes).⁷¹

The gene clusters organized in GCF 4 were found in all *Nocardia* species and encoded a single NRPS with four modules, co-located with a putative *N*-acyltransferase and two transporters. The produced peptide likely consists of an *N*-terminal serine, three additional amino acids, and presumably feature aliphatic or aromatic side chains.

More conserved NRPS pathways, present in 75 out of 101 *Nocardia* genomes, were identified in GCF 5 that consisted of a trimodular NRPS gene with epimerization (E) domains, encoded in the first and in the terminal module.

Similarly conserved were large NRPS assembly lines represented by GCF 6 that always encoded a putative nitropropane dioxygenase, a metallopeptidase, an esterase, an epimerase, and an enoyl-CoA dehydratase.

Linocin M18 has been described as a secreted bacteriolysin from *Brevibacterium linens* M18 and has been shown to form multimeric microcompartments.^{137, 138} Respective orthologous BGCs were encoded in 103 *Nocardia* genomes and were summarized in GCF 9. Bacteriolysins were associated with pathogen-host cell attachment and immunogenicity in *M. tuberculosis* and *Burkholderia cepacia* complex.^{139, 140}

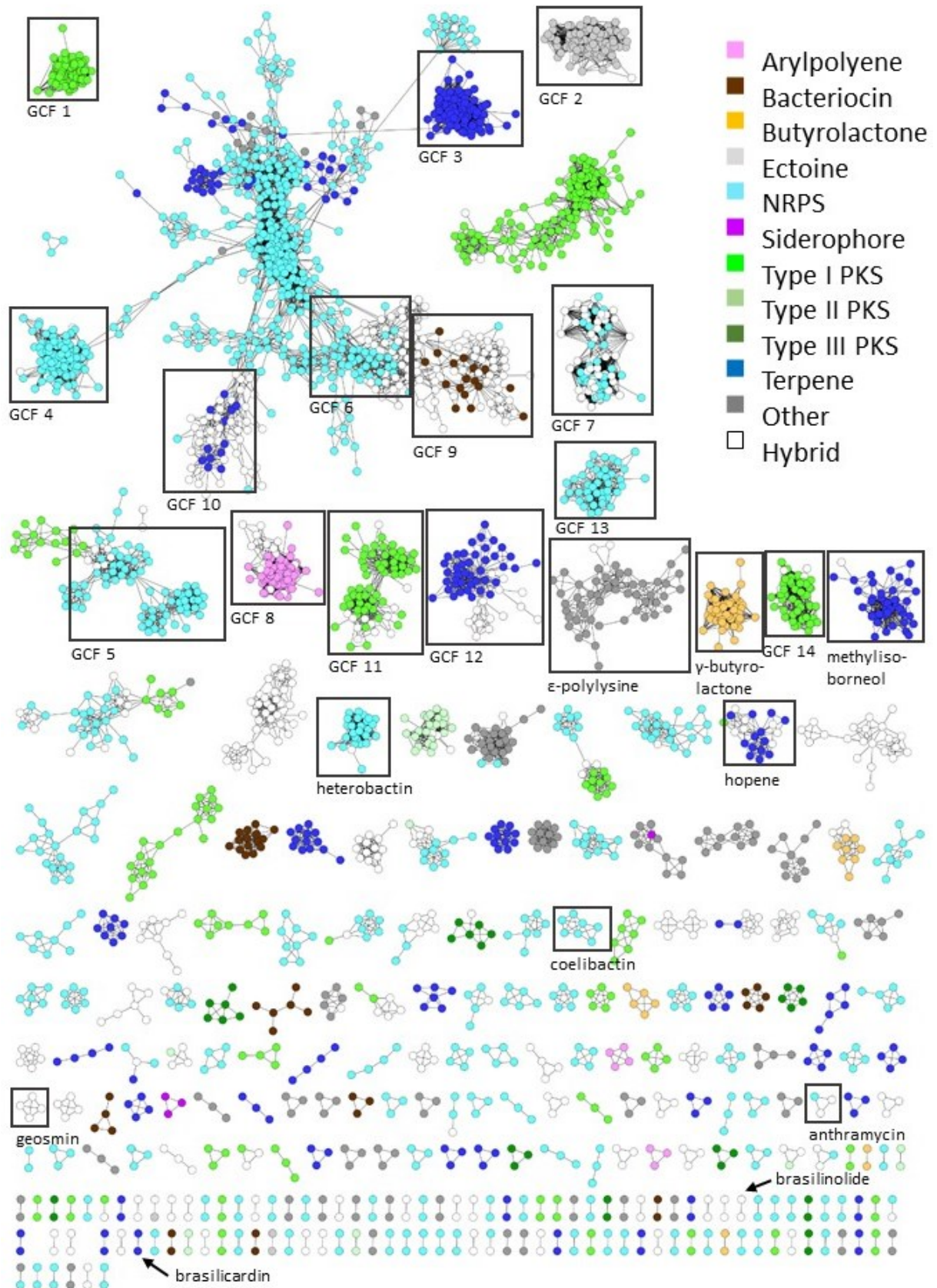


Figure 8: BiG-SCAPE cluster type sequence similarity network of all BGCs from *Nocardia* strains found with antiSMASH 4.0. Colors show cluster types from *Nocardia* BGCs. Colorless nodes are hybrid BGCs found in *Nocardia*. Adapted from Männle et al. 2020.⁶

3.1.3. The gene cluster family of nocobactins

Most interestingly, another highly abundant BGC family (GCF 7) was found in 92 of 101 *Nocardia* genomes (Figure 7A, Figure 8). It encoded hybrid PKS/NRPS pathways annotated by antiSMASH as putative nocobactin-like BGCs. Over 90% of all *Nocardia* genomes analyzed encoded such BGCs. This widespread occurrence indicated a constitutive role of nocobactins in the biology of the genus. Therefore, the genetic variability of the nocobactin-like siderophore GCF 7 was assessed. GCF 7 was found to harbor subclusters I of putative nocobactin-type BGCs only, whereas GCF 13 harbored subcluster II in cases where a putative *nbtF* gene was co-located. In *Nocardia* genomes, *nbtF* was either found on subcluster I or subcluster II, never on both subclusters. AntiSMASH only predicted subcluster II, when *nbtF* was co-located with cluster II and falsely termed it as a putative mycobactin-like BGC. Otherwise, subcluster II was completely absent from the analysis. A manual concatenation of subcluster I and subcluster II and reanalysis via BiG-SCAPE did not change the nocobactin-like GCFs at a threshold of 70%. The splitting of the nocobactin-type BGCs into subclusters I and II was reflected in the 50% BiG-SCAPE SSN. In Figure 8, GCF 7 represented nocobactin-type subcluster I, whereas GCF 13 represented nocobactin-type subcluster II. Interestingly, GCF 13 was annotated by antiSMASH as a putative mycobactin-like BGC with 30% similarity. GCF 13 contained a small pathway of three conserved genes with homology to *nbtF*, *nbtS*, and *nbtT*. These genomes lacked a *nbtF*-like gene in their respective nocobactin-like subclusters I and *nbtF* in subcluster II was not only a copy thereof. However, subcluster II was only recognized bioinformatically, when *nbtF* was co-located with *nbtS* and *nbtT*. This BGC organization was found to be shared by almost 45% of *Nocardia* strains analyzed.

The nocobactin family GCF 7 in Figure 8A formed two distinct subareas at a threshold of 50% that pointed out genetic differences. Indeed, each subarea was characterized by the exclusive presence of either a gene for an *N*-acetyltransferase (NbtH homolog) or an *N*-formyltransferase (TerH homolog), suggesting that the subdivision of the nocobactin family GCF 7 reflected the production of structurally different derivatives. These derivatives suggested either an acetyl or a formyl group attached to the *N*-hydroxylysine residue.

To further refine the putative nocobactin-like GCF 7 from the BiG-SCAPE 50% threshold SSN, a threshold of 70% was applied (Figure 8B). This resulted in the segmentation of GCF 7 into nine independent families of more than two BGCs (referred to as nocobactin-like BGC subfamilies I-IX) and two singletons (X and XI). Interestingly, nocobactin-like BGC subfamilies correlated with the phylogenetic clades observed by autoMLST analysis. Subfamilies II, V, VIII, and IX consisted solely of pathways from clade E, B, C, and F, respectively. Subfamilies I and VII were dominated by clades D, F, and A, B, respectively. Presence or absence of *nbtH/terH* homologous genes also determined these

subfamilies at a threshold of 70% by means of subfamilies I and III that encoded a *nbtH* homologous gene, whereas subfamilies II, IV, and VI-IX encoded a gene homologous to *terH*. Only subfamily V comprised pathways that harbored both *nbtH* and *terH* homologous genes. In addition, splitting of subfamilies was based on the presence/absence of the *nbtF* homolog in subcluster I. Subfamilies II, III, VI-VIII, and IX encoded a putative *nbtF* gene in subcluster I, whereas in strains of subfamilies I, IV, and V, a putative *nbtF* gene was found to be present in subcluster II.

Further characteristic features of GCF 7 were additional NRPS modules in subfamilies II, VI, VIII, and IX. Here, the nocobactin-like pathways encoded a discrete C-T didomain protein, probably a functional substitute to the missing terminal C-T domains of the NbtD homolog and another NRPS enzyme with a C-A-T-E-C domain structure. Unfortunately, it was not possible to predict the specificity of the corresponding A-domain bioinformatically.

Known nocobactin-type molecules showed diversity in stereochemical configurations. Nocobactin NA features a central *N*-acylated *N*-hydroxy-L-lysine and a terminal *N*-hydroxy-D-lysine-1,6-lactam. In contrast, amamistatin A incorporates D- and L-lysine and brasilibactin A D- and D-lysine derivatives at these positions, respectively.^{4, 49, 52} It was likely that the presence of additional epimerization (E) domain-containing NRPS-enzymes found in subfamilies II (including *N. terpenica* IFM 0406), VI, VIII, and IX might produce different stereoisomers. Exemplary, the NRPS TerE from *N. terpenica* IFM 0406 epimerizes and condensates a *N*-hydroxylysine, which is subsequently lactamized, most likely in a non-enzymatic way.⁶⁵

Moreover, in subfamilies VI and VIII the PKS enzymes with homology to NbtC were extended by a C-methyltransferase domain to an AT-KR-T-cMT module with predicted AT-domain specificity to mmal-CoA. Notably, formobactin A and the amamistatins contain a rare α -dimethylated fatty acid which could derive from such a biosynthetic set-up.

The co-localization of a gene for a putative SagB-type dehydrogenase was characteristic for the clusters subsumed in subfamilies I, III, V, and VII. Furthermore, the subfamilies were distinguished based on transporter location likely associated with the BGCs. In subfamilies II, IV, VI, VIII, and IX, transporters were located within nocobactin-type BGCs, whereas in subfamilies I, III, V, and VII, transporters were located immediately adjacent to the clusters.

In total, nine nocobactin-type BGCs appeared as singletons in the BiG-SCAPE SSN with a threshold of 70% and did not cluster with other BGCs from previously defined subfamilies. Their nocobactin-type BGC architecture looked different from other nocobactin-type BGCs, e.g. by means of gene duplication. *N. fusca* NBRC 14340 encoded one of these unique pathways and appeared as the only

strain with a nocobactin-type BGC in subfamily X. This BGC contained homologs to all genes of the nocobactin NA gene cluster with an additional copy of *nbtF*, a *terH*-like gene that coded for a putative *N*-acetyltransferase, a *nbtH*-like gene that coded for a putative *N*-formyltransferase, a gene for a discrete A domain and a putative serine hydroxymethyltransferase gene with a 59% sequence identity on protein level to FmoH from *Streptomyces* sp. Sp080513GE-23. Interestingly, as FmoH serves as a biosynthetic precursor for an unusual 4-methyloxazoline moiety, such a moiety was also found in the mycobactin-type siderophore BE-32030 produced by *Nocardia* sp. A32030.^{53, 141} Furthermore, *Nocardia vermiculata* NBRC 100427 (subfamily XI) showed two copies of an *nbtABC*-like operon and encoded *nbtD*-, *nbtG*-, and *terH*-like genes together with an additional NRPS module and transporters at another independent location on the genome different from subcluster I or II.

3.1.4. Nocobactin-type siderophores produced by selected *Nocardia* strains

In order to verify the BiG-SCAPE nocobactin-type subfamilies, representative *Nocardia* strains were ordered that encoded pathways with characteristic biosynthetic functionalities. For this purpose, selected *Nocardia* strains were cultivated under iron limiting conditions and intracellular membrane-associated culture extracts were subjected to positive ion mode LC-MS analysis. Extracts were treated with gallium sulfate ($\text{Ga}_2(\text{SO}_4)_3$) under reducing conditions to stabilize metal-chelating nocobactin-type siderophores and to simplify MS identification via their distinct gallium isotopic pattern. Metal-free siderophores and Fe^{3+} -bound nocobactin-type siderophores were largely replaced by Ga^{3+} .

MS/MS spectra of twelve *Nocardia* strains and pure samples of terpenibactins A-C were used to construct molecular networks with GNPS visualized in Cytoscape.^{5, 109} This resulted in a molecular network of ionizable metabolites with 3567 nodes and 4177 edges (Figure 9). At a cosine value of 0.6, the GNPS network formed 108 clusters of two or more nodes and showed 2470 singletons. The terpenibactins clustered with one molecular family from seven *Nocardia* strains that likely represented molecular families of compounds with similar chemical structures (Figure 9a). Each of these seven *Nocardia* strains produced a variety of different nocobactin-type siderophores with varying fatty acyl chain length and saturation that clustered with the terpenibactins. *N*-hydroxylysine *N*-acetylation or *N*-formylation, as well as oxazole or oxazoline moieties were consistently present. Exemplary, one postulated nocobactin-type siderophore is shown per nocobactin-type BGC subfamily (Figure 10).

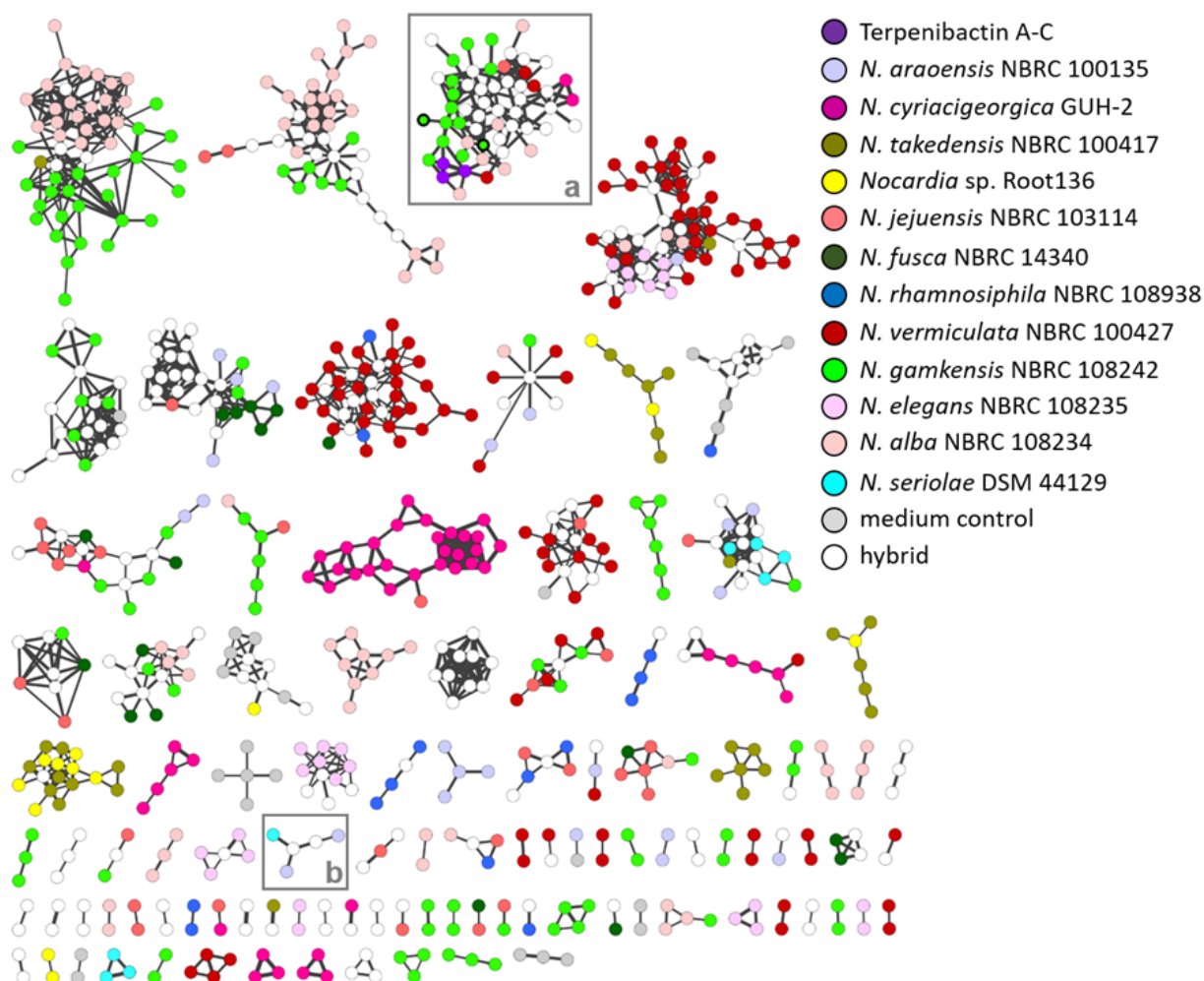


Figure 9: Global Natural Product Social Molecular Networking (GNPS) MS/MS network of culture extracts from selected *Nocardia* strains. Node color indicates the source strain of each ion. Bold node borders indicate putative carboxynocobactins. Molecular families a and b containing putative siderophores are highlighted by gray framed boxes. Medium was prepared after Ratledge et. al.¹⁴² Adapted from Männle et al. 2020.⁵

These nocobactin-type siderophores from seven different *Nocardia* strains represented molecules with minor structural differences to published nocobactin NA. All MS/MS predicted siderophores from *N. alba* NBRC 108234 (subfamily I), *N. elegans* NBRC 108235 (subfamily II), *N. cyriacigeorgica* GUH-2 (subfamily VI), *N. gamkensis* NBRC 108242 (subfamily VII), *N. rhamnosiphila* NBRC 108938 (subfamily VIII), *N. jejuensis* NBRC 103114 (subfamily IX), *N. fusca* NBRC 14340 (subfamily X), and *N. vermiculata* NBRC 100427 (subfamily XI), as well as the known compounds from literature produced by *N. farcinica* IFM 10152 (subfamily III) and the previously reported siderophores from *N. terpenica* IFM 0406 (subfamily IV), all included a phenoxazoline/-oxazole, an *N*-acylated *N*-hydroxylysine, a hydroxy fatty acid, and a ϵ -caprolactam moiety.^{4, 7} This core structure was independent of the BGC composition by means of presence of additional NRPS genes (subfamilies II, VI, VIII, and IX), duplication of PKS genes (subfamily XI) or absence of the *nbtF* homologous gene in subcluster I (subfamilies I, IV, V, and X).

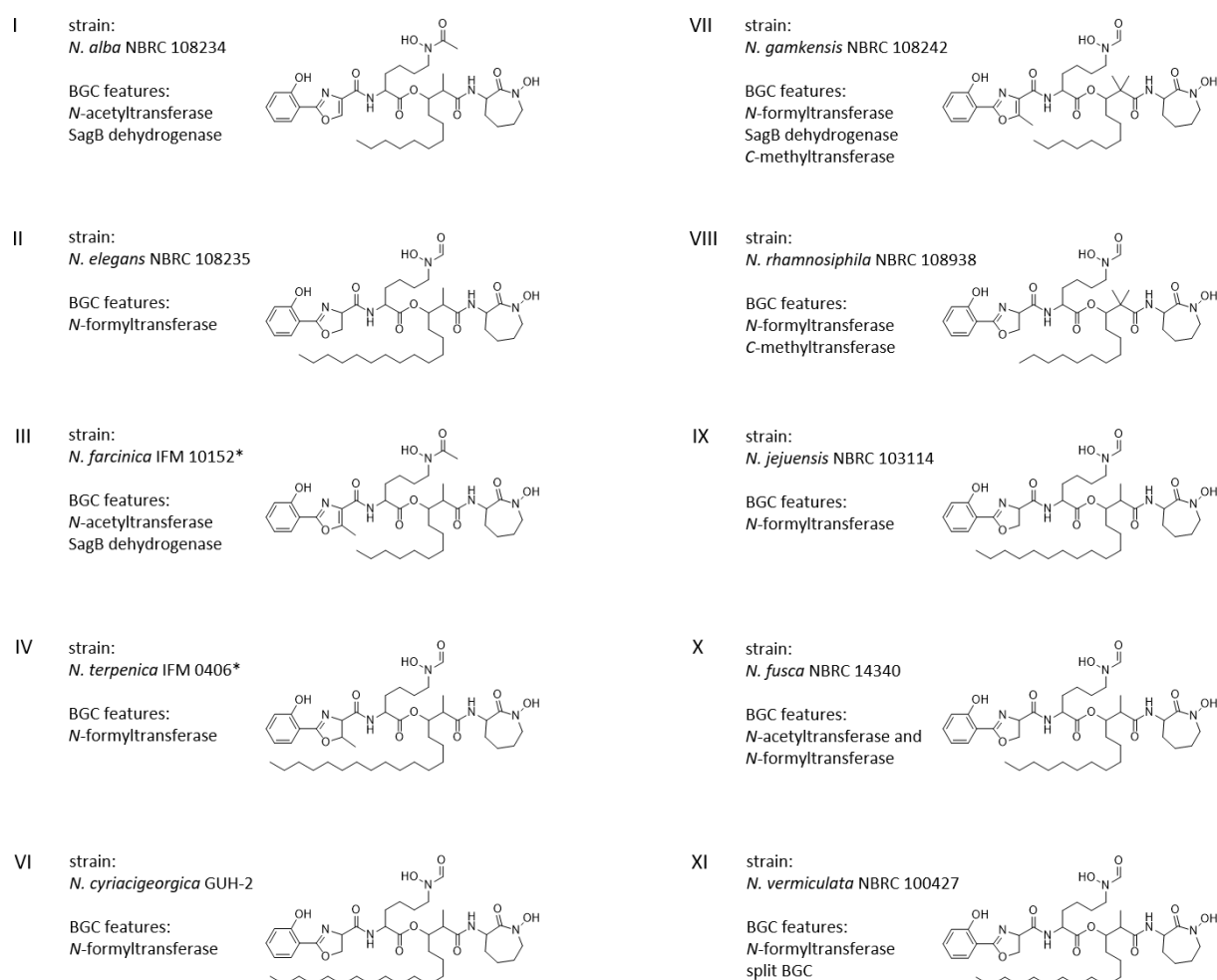


Figure 10: nocobactin-type siderophores produced by selected *Nocardia* strains from subfamilies I – XI. Selected representative compounds from each subfamily are shown. Structures shown in I, II, and VI-XI subfamilies were predicted from high resolution MS/MS experiments. Adapted from Männle et al. 2020.⁶ *chemical structure shown have been adopted from Hoshino et al. and Chen et al..^{4,7}

The presence of NbtH and TerH homologs, as well as PKS *C*-methyltransferase domains within nocobactin-type BGCs was reflected in MS/MS fragmentation patterns of the identified nocobactin-type molecules. In accordance with bioinformatic predictions, pathways from subfamilies I, III, and V produced *N*-acetylated and pathways from subfamilies II, IV, VI, VII, and VIII-IX produced *N*-formylated nocobactin derivatives. Furthermore, collocation of a *sagB*-type dehydrogenase gene with nocobactin-type BGCs in subfamilies I, III, and VII led to the production of nocobactin derivatives with a phenoxazole moiety. Vice-versa, the absence of a *sagB*-type dehydrogenase gene in subfamilies II, IV, VI, and VIII-XI correlated with the production of phenoxazoline moiety containing nocobactin derivatives. This indicated that this *SagB*-type dehydrogenase was involved in the oxidation of the oxazoline ring in nocobactin-type siderophore biosynthesis.

Interestingly, MS/MS structure predictions suggested that *N. alba* NBRC 108234 (subfamily I) produced nocobactin-type siderophores with a non-methylated oxazole moiety and an *N*-acetyl group that matched nocardimicin A, B, and D (Figure 10). Additionally, *N. elegans* NBRC 108235

(subfamily II) and *N. jejuensis* NBRC 103114 (subfamily IX) produced *N*-formylated, non-methylated oxazoline-containing molecules that matched chemical structures of nocardimicin G and nocardimicin H, respectively. Extracts of *N. cyriacigeorgica* GUH-2 (subfamily VI) were analyzed and an ion was found that matched the MS/MS predictions for brasilibactin A.⁵⁰ Additionally, the nocobactin-type BGC from *N. gamkensis* NBRC 108242 (subfamily VII) revealed a compound that matched formobactin.⁵¹ Notably, *N. rhamnosiphila* NBRC 108938 (subfamily VIII) produced a molecule likely containing a non-methylated oxazoline, an *N*-formyl group, and a *gem*-dimethylated hydroxy fatty acid moiety that was not described before.

In order to determine the correct position of the methylene unit in fragment ions of the nocobactin-type siderophore from *N. alba* NBRC 108234, ¹³C₁-L-serine was fed, and culture extracts were analyzed by LC-MS. A characteristic shift indicated the incorporation of ¹³C₁-L-serine into the oxazoline ring that indicated the presence of an *N*-acetyl group. For the biosynthesis of mycobactin, it has been proposed that the oxazoline ring derives from demethylation of a methyloxazoline moiety.¹⁴³ For nocobactin-type siderophores in *Nocardia* spp., data suggested that methylation of the oxazoline ring rather depended on the specificity of the respective A-domain towards serine or threonine in the NbtF homolog.

In particular, two metabolites identified from *N. gamkensis* NBRC 108242 caught my attention. Exact masses indicated modification of the central hydroxy fatty acyl unit by an additional carboxylic acid group. In analogy to carboxymycobactins, these two molecules may represent similarly oxidized “carboxynocobactins”.^{55, 144} It was assumed that missing genes for carboxynocobactin-type siderophore production (e.g. carboxylase genes) were located on another genomic region.

3.1.5. Identification of a mycobactin-type siderophore

Interestingly, MS/MS spectra from *N. araoensis* NBRC 100135 (subfamily III) and *N. takedensis* NBRC 100417 (subfamily VII) correlated with mycobactin-type siderophores with the long fatty acyl moiety attached to the *N*-hydroxylysine side chain. In order to verify mycobactin-type siderophore production, a compound from culture extracts of *N. araoensis* NBRC 100135 was isolated. NMR structural assignments of the gallium-bound siderophore identified key spin systems that revealed that the fatty acid indeed was attached to the linear *N*-hydroxylysine side chain (Figure 11). As NBRC 100135 and NBRC 100417 were genuine representatives of their respective subfamilies and no mycobactin-type BGCs were found on respective genomes, it was difficult to explain why they produced mycobactins instead of nocobactins. It was speculated for mycobactin biosynthesis in *Nocardia* spp. that in trans-encoded enzymes were recruited from other metabolic pathways that encoded a long-chain fatty acid *N*-acyltransferase.

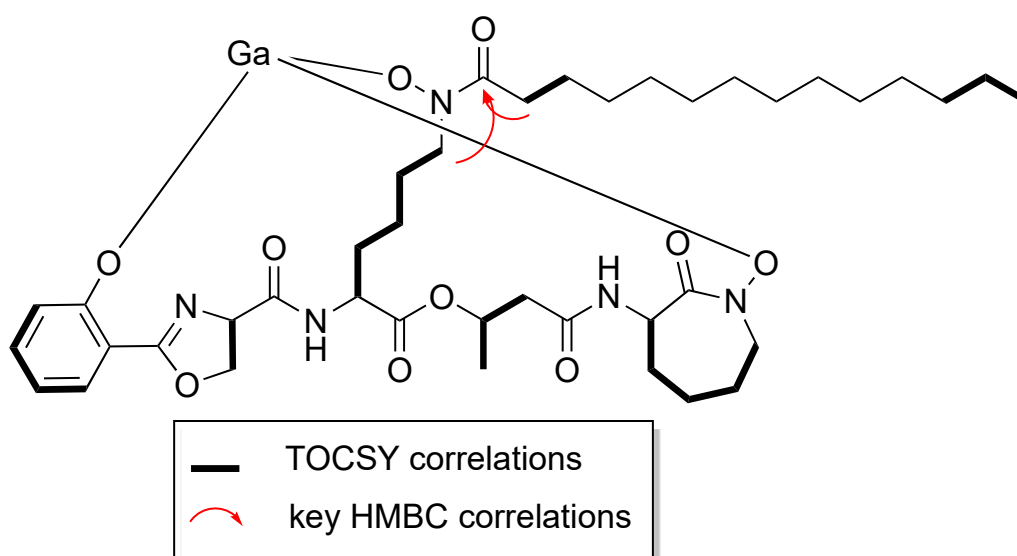


Figure 11: TOCSY correlations and key HMBC correlations of a mycobactin-type siderophore with m/z 840.3669 $[M+Ga-2H]^+$. Adapted from Männle et al. 2020.⁶

3.2. New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A-C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406

3.2.1. Identification and organization of the terpenibactin BGC

Genome mining of *Nocardia terpenica* IFM 0406 using antiSMASH revealed the presence of 38 BGCs including a split BGC (separated by 282 kbp) with predicted high similarity to the nocobactin NA-a and NA-b BGC from *Nocardia farcinica* IFM 10152 (termed as subcluster I and subcluster II) (Figure 12). From now on it will be termed terpenibactin BGC.⁹⁸ The requisite genetic machinery for nocobactin biosynthesis was present. However, within the terpenibactin BGC, a *nbtH*-like gene was missing. In *Nocardia farcinica* IFM 10152 *nbtH* encoded a *N*-hydroxylysine *N*-acetyltransferase. Instead, the terpenibactin BGC encoded a *N*-hydroxylysine *N*-formyltransferase gene in place of *nbtH*, termed *terH* (Figure 12).⁷

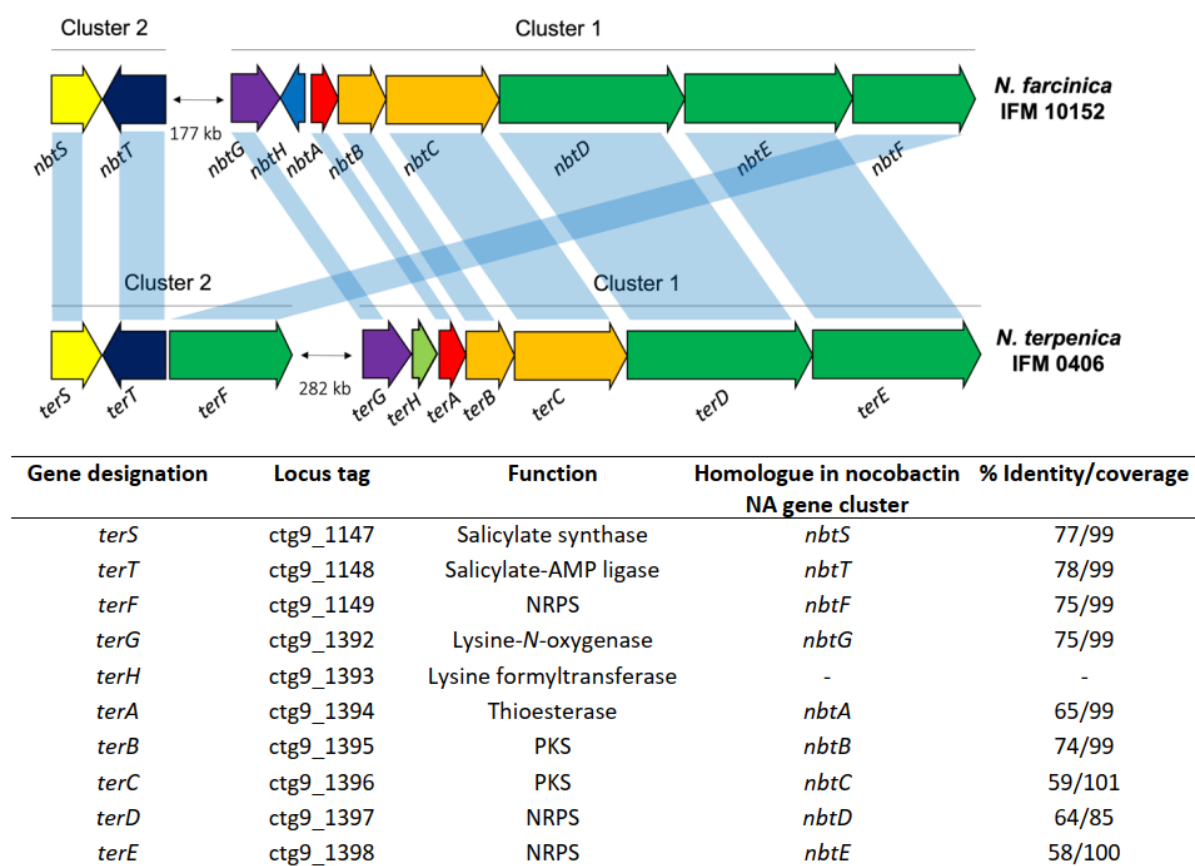


Figure 12: Comparison of the nocobactin NA biosynthetic gene cluster (top, *Nocardia farcinica* IFM 10152) with a novel BGC (bottom, *Nocardia terpenica* IFM 0406). Blue shading links homologous genes. After Chen et al.⁷

Additionally, the terpenibactin BGC encoded a homologous gene to *nbtF* (termed *terF*) in subcluster II together with *nbtS*- and *nbtT*-like genes (termed *terS* and *terT*), whereas respective *terF* gene was absent in subcluster I. In *Nocardia farcinica* IFM 10152, *nbtF* was located at the first locus. The

relocation of the *nbtF*-like gene in *Nocardia terpenica* IFM 0406 did not preclude the production of new nocobactin congeners, since in nocobactin biosynthesis, NbtF acts directly after NbtS and NbtT, which are also located at locus two.⁴ The resultant biosynthesis of terpenibactins A-C was analogous to the delineated biosynthesis of nocobactin NA with exceptions addressed to presence of a *terH* gene and absence of a putative SagB-type dehydrogenase in the terpenibactin BGC.⁴ SagB is involved in the biosynthesis of streptolysin S toxin by catalyzing the oxidation of oxazoline and thiazoline heterocycles to form oxazole and thiazole ring structures.¹⁴⁵

Since terpenibactin biosynthesis involved the utilization of fatty acids derived from primary metabolism, it was challenging to predict the lengths of acyl chains by bioinformatics. Considering the range of known nocobactin-type compounds, chain lengths were assumed between C₇-C₁₇ units long. From bioinformatics predictions, one or more brasilibactin-type compounds were postulated and masses between 701 and 841 Da were suggested.

Stereochemical predictions of the absolute configuration of the amino acid-related moieties were conducted on the basis of *in silico* predictions from the terpenibactin BGC. The E domain in TerE indicated the conversion of the activated lysine into the D-configuration, whereas absence of C/E and E domains in TerF and TerD suggested the incorporation and processing of L-configured threonine and lysine moieties. This suggested an R configuration at position 1 and S configurations at positions 12, 14 and 15 (Figure 17).

3.2.2. LC/MS screening for the products of the terpenibactin BGC

Nocardia terpenica IFM 0406 cell pellet methanol extracts were generated as reported before for nocobactin NA and similar compounds.^{44, 47-50} Extract MS analysis led to the detection of three peaks in the sought-after mass range, with UV spectra maxima at 219, 255, and 339 nm as expected from nocobactin NA UV profiles indicative of a substituted benzenoid chromophore with an extended conjugated system. All three peaks absorbed at 460 nm when analyzed via HPLC, as would be expected of Fe³⁺-siderophore complexes similar to Fe³⁺-nocobactin NA (Figure 13).⁴⁴ In total, each compound was detected as single charged pseudo-molecular ion [M+H]⁺, ferrated form [M-3H+Fe³⁺+H]⁺, and sodium adduct [M-3H+Fe³⁺+Na]⁺. As MS data did not match any known compound, new nocobactin-type siderophores were termed terpenibactins A-C named after the producing strain *Nocardia terpenica* IFM 0406 (Figure 14).

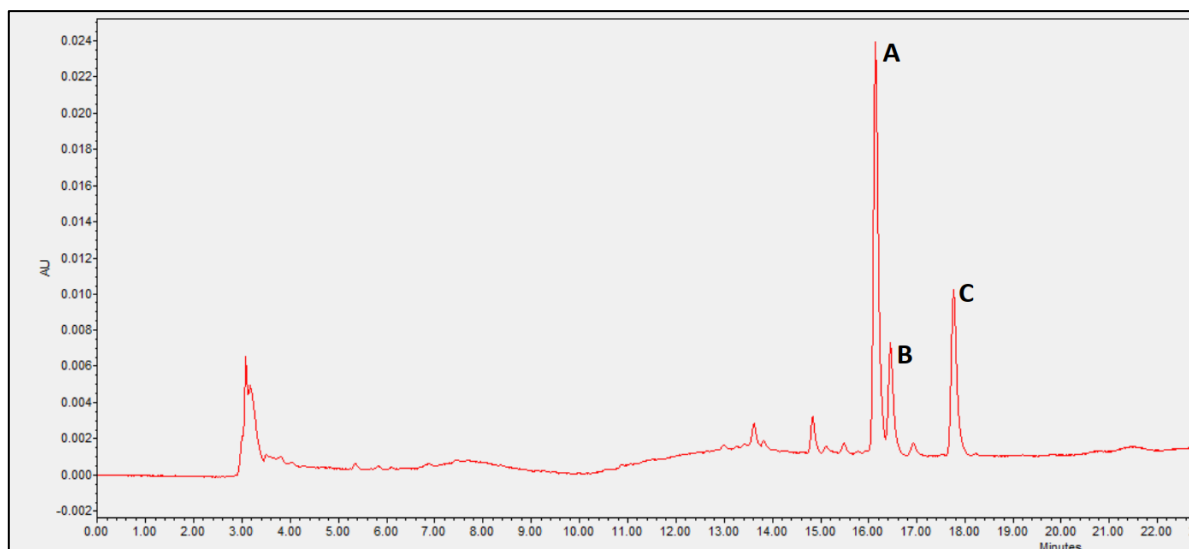


Figure 13: Terpenibactins A-C HPLC profile. Terpenibactins appeared between 16-18 min at 460 nm in HPLC. After Chen.¹⁴⁶

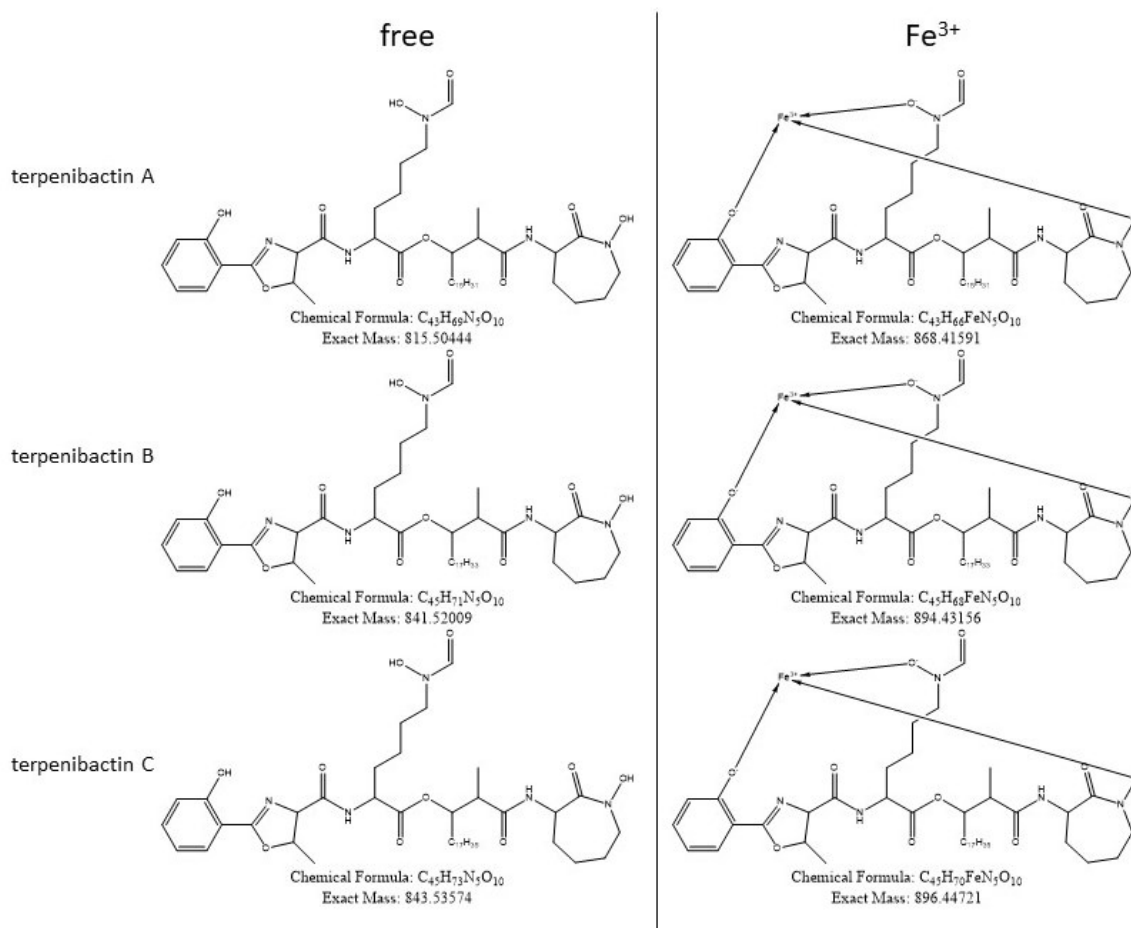


Figure 14: free terpenibactins A-C (left) and ferrated terpenibactins A-C (right) produced by *Nocardia terpenica* IFM 0406.

All attempts for the decomplexation of ferric nocobactin after Ratledge and Snow failed.⁴⁴ As a consequence, extracts were treated with $\text{Ga}_2(\text{SO}_4)_3$ to generate stable complexes, even facilitating high-quality NMR measurements.^{61, 147, 148} This method simplified MS analysis of culture extracts, due to the characteristic gallium isotope pattern of gallium-complexed siderophores. Fragmentation patterns of $\text{Ga}_2(\text{SO}_4)_3$ treated terpenibactins A-C were consistent along the ester linkage between the

modified lysine residue and the β -hydroxy fatty acid moiety. After removal of the ϵ -caprolactam, gallium-bound substructures remained as a bidentate ligand and produced characteristic fragment ions for terpenibactin A (Figure 15). Further characteristic fragment ions observed represent the ϵ -caprolactam moiety. In analogy, similar fragmentation patterns were observed for terpenibactin B and C with fragment ions that stemmed from a slightly different fatty acyl moiety (Figure 16).

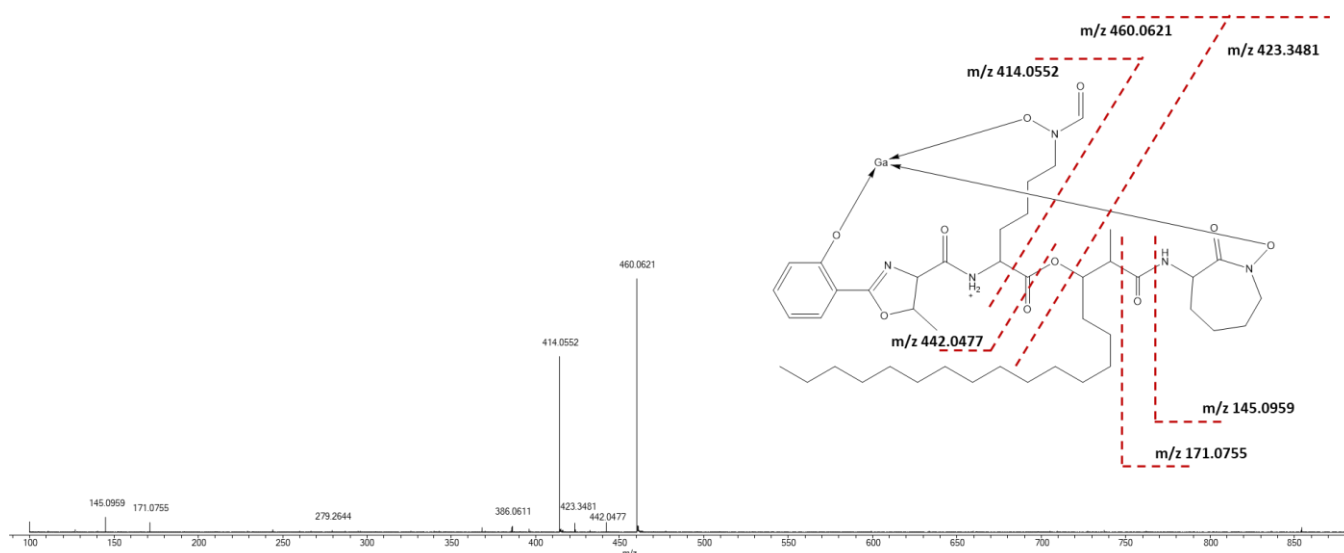
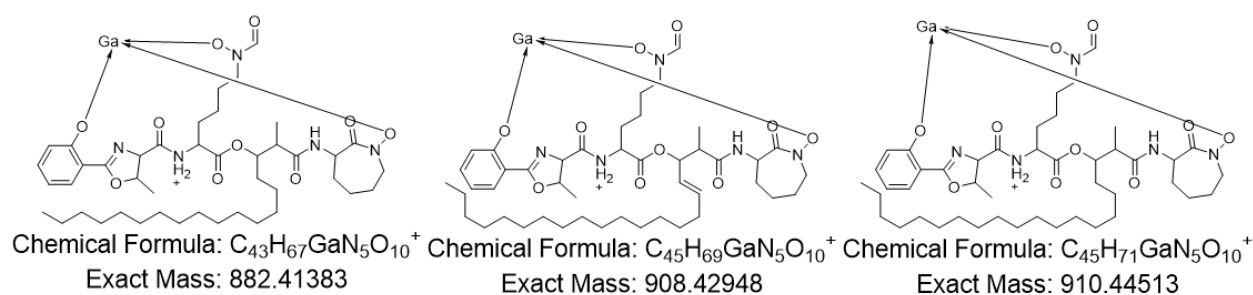


Figure 15: MS/MS fragmentation pattern of gallium-bound terpenibactin A. Precursor mass (m/z 882.4135 [$M-2H+Ga$] $^+$) is marked with a blue diamond.

N-formylation of the *N*-hydroxylysine residue in terpenibactin biosynthesis was postulated bioinformatically according to presence of *terH* within the terpenibactin BGC instead of presence of a gene homologous to *nbth*. Brasilibactin A (isolated from *Nocardia terpenica* IFM 0995) also featured a *N*-formylated *N*-hydroxylysine residue.⁵⁰ Additionally, absence of a putative SagB-type dehydrogenase encoded within the terpenibactin BGC was characteristic for oxazoline ring formation. According to findings from Männle et al., a methylated oxazoline ring was the only plausible structure prediction in accordance with the terpenibactin BGC.⁶ Furthermore, MS/MS fragmentation spectra showed fragments common to terpenibactins A-C, namely those of the ϵ -caprolactam ring with the neighboring amide group alone and with the full amide bond (Figure 15). Predictions of their exact masses closely matched observed masses for terpenibactin A, B, and C found with high resolution MS (Table 1). In total, MS/MS analysis confirmed the nocobactin-like structure of terpenibactins A-C (Figure 16).

Table 1: Terpenibactin A, B, and C produced by *Nocardia terpenica* IFM 0406.

name	exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	Δ ppm error
Terpenibactin A	815.50444	882.41383	882.4135	0.37397
Terpenibactin B	841.52009	908.42948	908.4282	1.40903
Terpenibactin C	843.53574	910.44513	910.4442	1.02148

**Figure 16: Chemical structures and chemical formulae of gallium-bound terpenibactins A-C.** Gallium-bound terpenibactin A (m/z 882.41383 [M-2H+Ga]⁺), gallium-bound terpenibactin B (m/z 908.42948 [M-2H+Ga]⁺), and gallium-bound terpenibactin C (m/z 910.44513 [M-2H+Ga]⁺).

3.2.3. NMR structure elucidation of terpenibactins A-C

¹H and ¹³C NMR spectra suggested compounds similar to nocobactin NA with aromatic protons and a singlet attributed to a formyl group. The *N*-formyl-*N*-hydroxylysine substructure was determined by COSY and confirmed by NOESY correlations (Figure 17 **Error! Reference source not found.**). Furthermore, terpenibactin A ¹³C NMR spectra showed signals that gave rise to a methylated oxazoline. ¹H-¹H COSY and ¹H-¹³C HSQC NMR analysis revealed that signals observed in ¹³C NMR belonged to a spin system including a methyl group (Figure 17). To confirm unsaturation in the extended alkyl chains of terpenibactin B, further COSY/HSQC-TOCSY correlations were elucidated (Figure 17). The precise location of this double bond was proposed through HSQC-TOCSY NMR and suggested that the double bond occurred between carbons at positions 31 and 32. The terpenibactin A-C ε-caprolactam ring common with nocobactin NA was confirmed by COSY and HSQC-TOCSY NMR correlations (Figure 17). COSY NMR confirmed a 2-methyl-3-hydroxyoctadecanoic moiety in terpenibactin A by COSY correlations (Figure 17). Further COSY correlations between methane carbon 8 and 10 also confirmed the presence of the 2-methyl-3-hydroxyoctadecanoic moiety. All three methane carbons at positions 8, 9, and 10 in turn showed HMBC correlations with the carbonyl carbon at position 7. COSY correlations were found to link the carbon at position 10 with carbons 29 to 31, and carbon 42 with a terminal methyl group at position 43. A precise listing of all terpenibactins A-C chemical shifts is shown in Table 2.

Results & Discussion

Table 2: ^1H and ^{13}C NMR spectroscopic table for terpenibactins A-C. Measured at 400 MHz (^1H) and 100 MHz (^{13}C) in d_4 -MeOH). After Chen.¹⁴⁶

No.	Terpenibactin A		Terpenibactin B		Terpenibactin C	
	δ_{C}	δ_{H}	δ_{C}	δ_{H}	δ_{C}	δ_{H}
1	51.0	4.78 (m)	51.0	4.79 (m)	51.0	4.79 (m)
2	31.5	1.95 (m)	31.7	1.94 (m)	31.5	1.95 (m)
3	28.7	1.82 (m)	28.7	1.82 (m)	28.8	1.82 (m)
4	25.8	1.8 (m)	25.4	1.80 (m)	25.4	1.81 (m)
5	54.2	3.99 (t), 3.86 (m)	54.2	3.99 (t), 3.85 (m)	54.2	4.00 (t), 3.86 (m)
6	166.2		166.2		166.1	
7	173.6		173.6		173.6	
8	45.1	2.63 (qd)	45.1	2.63 (qd)	45.1	2.63 (qd)
9	14.0	1.15 (d)	14.0	1.15 (d)	14.3	1.15 (d)
10	77.8	5.11 (dt)	77.8	5.11 (dt)	77.8	5.11 (dt)
11	174.6		174.6		174.6	
12	54.7	4.38 (m)	54.7	4.39 (m)	54.7	4.39 (m)
13	170.6		170.6		170.6	
14	73.0	4.31 (d)	73.0	4.31 (d)	73.0	4.31 (d)
15	80.8	5.27 (m)	80.8	5.27 (m)	80.8	5.27 (m)
16	21.1	1.49 (d)	21.1	1.48 (d)	21.1	1.50 (d)
17	171.7		171.7		171.7	
18	110.3		110.3		110.3	
19	167.3		167.2		167.2	
20	130.6	7.67 (dd)	130.6	7.67 (dd)	130.7	7.67 (dd)
21	117.1	6.68 (t)	117.1	6.68 (t)	117.0	6.68 (t)
22	136.8	7.36 (ddd)	136.8	7.36 (ddd)	136.8	7.36 (ddd)
23	122.6	6.81 (d)	122.6	6.81 (d)	122.6	6.82 (d)
24	30.7	2.04 (m)	30.6	2.04 (m)	30.7	2.04 (m)
25	23.6	1.31 (m)	23.5	1.31 (m)	23.5	1.32 (m)
26	29.7	1.97 (m), 1.60 (m)	19.5	1.96 (m), 1.60 (m)	19.5	1.97 (m), 1.61 (m)
27	49.3	3.95 (m), 3.70 (m)	49.3	3.95 (m), 3.70 (m)	49.3	3.95 (m), 3.70 (m)
28	155.5	8.13 (s)	155.1	8.13 (s)	155.1	8.13 (s)
29	30.4	1.49 (m)	30.4	1.49 (m)	30.4	1.49 (m)
30	28.0	2.04 (m)	28.0	2.04 (m)	28.0	2.04 (m)
31	30.7 - 29.9	1.30 (m)	130.7	5.35 (t)	30.7 - 29.9	1.30 (m)
32	30.7 - 29.9	1.30 (m)	130.7	5.35 (t)	30.7 - 29.9	1.30 (m)
33	30.7 - 29.9	1.30 (m)	28.0	2.04 (m)	30.7 - 29.9	1.30 (m)
34	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
35	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
36	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
37	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
38	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
39	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
40	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
41	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
42	23.6	1.31 (m)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
43	14.3	0.91 (t)	30.7 - 29.9	1.30 (m)	30.7 - 29.9	1.30 (m)
44			23.6	1.31 (m)	23.6	1.31 (m)
45			14.3	0.91 (t)	14.3	0.91 (t)

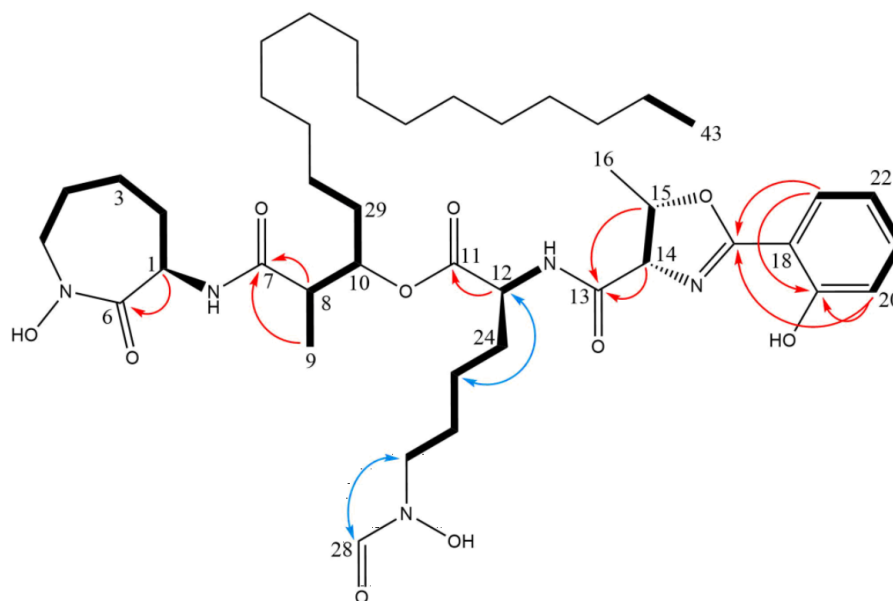


Figure 17: Proposed structure of terpenibactin A. Bold bonds show COSY/HSQC-TOCSY correlations; red arrows show HMBC correlations; blue arrows show NOESY correlations. Terpenibactin C has the same structure with a fatty acid chain extended by 2 x CH₂ units. Terpenibactin B shows COSY/HSQC-TOCSY correlations from positions 29 to 33, including a double bond between position 31 and 32. *In silico* predictions of the incorporations of D-lysine, L-lysine, and L-threonine into the terpenibactin backbone resulted in a relative stereochemistry of 1R, 12S, 14S, and 15S. After Chen.¹⁴⁶

3.2.4. Biological properties of terpenibactins A-C

Gallium-chelated terpenibactins A-C were evaluated in antimicrobial, cytotoxic, and muscarinic M3 receptor calcium flux assays (Table 3). In antimicrobial assays, terpenibactins A-C were inactive at an MIC level of 64 µg/mL. In cytotoxicity assays, terpenibactins A-C showed moderate bioactivity. However, Sakagami et al. demonstrated that iron-chelated nocobactins exhibit cytotoxic activity two orders lower than its corresponding iron-free form.⁴⁵ It is possible that gallium-chelated terpenibactins were affected similarly. All three compounds showed potent muscarinic M3 receptor inhibitory activity. These findings were in good agreement with IC₅₀ values reported from nocardimicins A-H (0.37 to 5.89 µM).^{47, 48}

Table 3: Biological activities of terpenibactins A-C. Modified after Chen et al..⁷

Compound	Muscarinic M3 receptor inhibitory activity	Cytotoxicity
	IC ₅₀	IC ₅₀ (HeLa)
Terpenibactin A (1)	1.15 ± 0.13 µM	16.7 µg/mL
Terpenibactin B (2)	1.77 ± 0.26 µM	45.9 µg/mL
Terpenibactin C (3)	1.59 ± 0.13 µM	33.3 µg/mL
Atropine	8.10 ± 0.80 nM	N/A

4. Conclusion & Outlook

The computational investigation of the prolific genus *Nocardia* revealed valuable insights into its highly diverse biosynthetic capacity to produce natural products. The longstanding question of the relationship of BGC genetic variability within GCFs and chemical structural diversity was addressed by relation of observed nocobactin-type BGC variation to respective predicted nocobactin-type siderophore diversity by chemical analytics data.

In conclusion, antiSMASH with subsequent BiG-SCAPE SSN analysis were shown to be suitable bioinformatic tools to successfully predict *Nocardia* GCFs (50% threshold) and BGC subfamilies (70% threshold for nocobactins) on a large scale. Comparative genomics data was successfully linked to respective chemical analytics data for refined chemical structure prediction of nocobactins. Observed *Nocardia* nocobactin-type siderophore chemical structural variations predicted by MS/MS matched distinct nocobactin-type BGC subfamilies at a 70% threshold. Thus, *Nocardia* nocobactin-type BGC genetic variability between strains indeed was reflected as chemically diverse nocobactin-type siderophores. It was shown that substitution of single biosynthetic genes, e.g. *N*-formyltransferase instead of a *N*-acetyltransferase gene, lead to distinct chemical structure. I was able to describe new nocobactin variants, new carboxynocobactins, and mycobactin variants from several *Nocardia* species. No BGCs were reported before that were associated with the siderophores brasilibactin A, nocardimicins A-I, and formobactin. Chemical structures of three new nocobactin-type siderophores and one mycobactin-type siderophore, were elucidated using NMR. In particular, the terpenibactins were found to possess inhibitory activity to the muscarinic M3 receptor, while exhibiting only a low cytotoxicity, providing suggestions for beneficial medical applications.

In summary, the use of SSNs allowed the distinction and prediction of natural product structural variations in the genus *Nocardia*. The putative BGC potential per genome of *Nocardia* indeed rivals better investigated genera such as *Streptomyces* and *Amycolatopsis*.^{74, 123, 124} Some GCFs were widespread throughout *Nocardia*, raising questions about their role in nature. One such GCF encoded nocobactin-type siderophores that might play a similar role in virulence, as mycobactins do in *Mycobacterium*. First evidence for nocobactins as virulence factors in *Nocardia* was already provided by Hoshino et al., but further confirmation is needed.⁴ As a vast majority of BGCs from *Nocardia* were orphan and were not associated with a discovered natural product so far, results from this thesis will facilitate future genome mining approaches and lower the chance of natural product rediscovery by prioritizing *Nocardia* BGCs and strains for a future genomics-driven targeted exploitation of the genus *Nocardia* for natural product drug discovery campaigns, biomedical applications, or the chemical industry.

5. References

1. Blin, K.; Shaw, S.; Steinke, K.; Villebro, R.; Ziemert, N.; Lee, S. Y.; Medema, M. H.; Weber, T., antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Research* **2019**, *47*, W81-W87.
2. Komaki, H.; Ichikawa, N.; Hosoyama, A.; Takahashi-Nakaguchi, A.; Matsuzawa, T.; Suzuki, K.; Fujita, N.; Gonoj, T., Genome based analysis of type-I polyketide synthase and nonribosomal peptide synthetase gene clusters in seven strains of five representative *Nocardia* species. *BMC Genomics* **2014**, *15*, 323.
3. Navarro-Munoz, J. C.; Selem-Mojica, N.; Mullowney, M. W.; Kautsar, S. A.; Tryon, J. H.; Parkinson, E. I.; De Los Santos, E. L. C.; Yeong, M.; Cruz-Morales, P.; Abubucker, S.; Roeters, A.; Lokhorst, W.; Fernandez-Guerra, A.; Cappellini, L. T. D.; Goering, A. W.; Thomson, R. J.; Metcalf, W. W.; Kelleher, N. L.; Barona-Gomez, F.; Medema, M. H., A computational framework to explore large-scale biosynthetic diversity. *Nature Chemical Biology* **2020**, *16*, 60-68.
4. Hoshino, Y.; Chiba, K.; Ishino, K.; Fukai, T.; Igarashi, Y.; Yazawa, K.; Mikami, Y.; Ishikawa, J., Identification of Nocobactin NA Biosynthetic Gene Clusters in *Nocardia farcinica*. *Journal of Bacteriology* **2011**, *193*, 441-448.
5. Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kaponov, 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.; P, C. A. B.; 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. O.; Pogliano, K.; Lington, 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*, 828-837.
6. Männle, D. M. S. M. K. M., S. S.; Steinke, K.; Lu, Z.; Moore, B. S.; Ziemert, N.; Kaysser, L., Comparative genomics to metabolomics in the genus *Nocardia*. *mSystems* **2020**.
7. Chen, J. L. F., A.; Männle, D.; Straetener, J.; Brötz-Oesterhelt, H.; Ziemert, N.; Kaysser, L.; Gross, H., New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A-C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406. *Chembiochem : a European journal of chemical biology* **2020**, *21*, 1-10.
8. Snow, G. A.; White, A. J., Chemical and biological properties of mycobactins isolated from various mycobacteria. *The Biochemical journal* **1969**, *115*, 1031-1050.

References

9. Davies, J. K., Specialized microbial metabolites: functions and origins. *The Journal of antibiotics* **2013**, *66*, 361.
10. Breitling, R.; Cenicerros, A.; Jankevics, A.; Takano, E., Metabolomics for secondary metabolite research. *Metabolites* **2013**, *3*, 1076-1083.
11. Newman, D. J.; Cragg, G. M., Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of natural products* **2016**, *79*, 629-661.
12. Katz, L.; Baltz, R. H., Natural product discovery: past, present, and future. *Journal of industrial microbiology & biotechnology* **2016**, *43*, 155-176.
13. Bérdy, J., Bioactive microbial metabolites. *The Journal of antibiotics* **2005**, *58*, 1-26.
14. Koehn, F. E.; Carter, G. T., The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery* **2005**, *4*, 206-20.
15. Spellberg, B., The future of antibiotics. *Critical Care* **2014**, *18*, 228.
16. Beutler, J. A., Natural Products as a Foundation for Drug Discovery. *Curr Protoc Pharmacol* **2009**, *46*, 9.11.1-9.11.21.
17. Bharel, S.; Gulati, A.; Abdin, M. Z.; Srivastava, P. S.; Vishwakarma, R. A.; Jain, S. K., Enzymatic Synthesis of Artemisinin from Natural and Synthetic Precursors. *Journal of natural products* **1998**, *61*, 633-636.
18. Bérdy, J., Thoughts and facts about antibiotics: where we are now and where we are heading. *The Journal of antibiotics* **2012**, *65*, 385-95.
19. Goodfellow, M.; Whitman, W. B.; Parte, A. C., Bergey's manual of systematic bacteriology. *Springer* **2012**, *5*.
20. Ikeda, H.; Ishikawa, J.; Hanamoto, A.; Shinose, M.; Kikuchi, H.; Shiba, T.; Sakaki, Y.; Hattori, M.; Ōmura, S., Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nature biotechnology* **2003**, *21*, 526.
21. Chater, K. F., *Streptomyces* inside-out: a new perspective on the bacteria that provide us with antibiotics. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **2006**, *361*, 761-768.
22. Challinor, V. L.; Bode, H. B., Bioactive natural products from novel microbial sources. *Annals of the New York Academy of Sciences* **2015**, *1354*, 82-97.
23. Dhakal, D.; Chaudhary, A. K.; Pokhrel, A.; Jha, A.; Darsandhari, S.; Shrestha, B.; Sohng, J. K., Underpinning the secondary metabolites from *Nocardia* spp. *Journal of Biomolecule Reconstruction* **2013**, *10*, 9-17.
24. Kurup, P. V.; Sandhu, R. S., Isolation of *Nocardia caviae* from Soil and Its Pathogenicity for Laboratory Animals. *Journal of Bacteriology* **1965**, *90*, 822-823.
25. Yamamura, H.; Hayakawa, M.; Nakagawa, Y.; Tamura, T.; Kohno, T.; Komatsu, F.; Imura, Y., *Nocardia takedensis* sp. nov., isolated from moat sediment and scumming activated sludge. *International Journal of Systematic and Evolutionary Microbiology* **2005**, *55*, 433-436.

26. El-Gendy, M. M.; Hawas, U. W.; Jaspars, M., Novel bioactive metabolites from a marine derived bacterium *Nocardia* sp. ALAA 2000. *The Journal of antibiotics* **2008**, *61*, 379-386.
27. Muricy, E. C. M.; Lemes, R. A.; Bombarda, S.; Ferrazoli, L.; Chimara, E., Differentiation between *Nocardia* spp. and *Mycobacterium* spp.: Critical aspects for bacteriological diagnosis. *Rev Inst Med Trop Sao Paulo* **2014**, *56*, 397-401.
28. Nocard, E., Note sur la maladie des boeufs de la Guadeloupe connue sous le nom de Farcin. *Annales de l'Institut Pasteur* **1888**, *2*, 293-302.
29. Fatahi-Bafghi, M., Nocardiosis from 1888 to 2017. *Microbial Pathogenesis* **2018**, *114*, 369-384.
30. Wilson, J. W., Nocardiosis: Updates and Clinical Overview. *Mayo Clinic Proceedings* **2012**, *87*, 403-407.
31. Beaman, B. L.; Beaman, L., *Nocardia* species: host-parasite relationships. *Clinical Microbiology Reviews* **1994**, *7*, 213-264.
32. Paige, E. K.; Spelman, D., Nocardiosis: 7-year experience at an Australian tertiary hospital. *Internal Medicine Journal* **2019**, *49*, 373-379.
33. Minero, M. V.; Marin, M.; Cercenado, E.; Rabadan, P. M.; Bouza, E.; Munoz, P., Nocardiosis at the turn of the century. *Medicine (Baltimore)* **2009**, *88*, 250-261.
34. Ott, S. R.; Meier, N.; Kolditz, M.; Bauer, T. T.; Rohde, G.; Presterl, E.; Schürmann, D.; Lepper, P. M.; Ringshausen, F. C.; Flick, H.; Leib, S. L.; Pletz, M. W., Pulmonary nocardiosis in Western Europe—Clinical evaluation of 43 patients and population-based estimates of hospitalization rates. *International Journal of Infectious Diseases* **2019**, *81*, 140-148.
35. Mikami, Y.; Yazawa, Y.; Tanaka, Y.; Ritzau, M.; Gräfeb, U., Isolation and Structure of Nocardiolactone, A New Dialkyl-Substituted β -Lactone From Pathogenic *Nocardia* Strains. *Natural Product Letters* **1999**, *13*, 277-284.
36. Hashimoto, M.; Komori, T.; Kamiya, T., Nocardicin A and B, monocyclic .beta.-lactam antibiotics from a *Nocardia* species. *Journal of the American Chemical Society* **1976**, *98*, 3023-3025.
37. Komaki, H.; Nemoto, A.; Tanaka, Y.; Takagi, H.; Yazawa, K.; Mikami, Y.; Shigemori, H.; Kobayashi, J.; Ando, A.; Nagata, Y., Brasilicardin A, a new terpenoid antibiotic from pathogenic *Nocardia brasiliensis*: fermentation, isolation and biological activity. *The Journal of antibiotics* **1999**, *52*, 13-9.
38. Saha, R.; Saha, N.; Donofrio, R. S.; Bestervelt, L. L., Microbial siderophores: a mini review. *Journal of basic microbiology* **2013**, *53*, 303-317.
39. Seyedsayamdost, M. R.; Cleto, S.; Carr, G.; Vlamakis, H.; João Vieira, M.; Kolter, R.; Clardy, J., Mixing and Matching Siderophore Clusters: Structure and Biosynthesis of Serratiochelins from *Serratia* sp. V4. *Journal of the American Chemical Society* **2012**, *134*, 13550-13553.
40. Price-Whelan, A.; Dietrich, L. E. P.; Newman, D. K., Rethinking 'secondary' metabolism: physiological roles for phenazine antibiotics. *Nature Chemical Biology* **2006**, *2*, 71.
41. Golden, C. A.; Kochan, I.; Spriggs, D. R., Role of mycobactin in the growth and virulence of tubercle bacilli. *Infection and immunity* **1974**, *9*, 34-40.

References

42. Kontoghiorghes, G. J.; Kolnagou, A.; Skiada, A.; Petrikkos, G., The Role of Iron and Chelators on Infections in Iron Overload and Non Iron Loaded Conditions: Prospects for the Design of New Antimicrobial Therapies. *International Journal for Hemoglobin Research* **2010**, *34*, 227-239.
43. Saha, M.; Sarkar, S.; Sarkar, B.; Sharma, B.; Bhattacharjee, S.; Tribedi, P., Microbial siderophores and their potential applications: A review. *Environmental science and pollution research international* **2015**, *23*, 3984–3999.
44. Ratledge, C.; Snow, G. A., Isolation and structure of nocobactin NA, a lipid-soluble iron-binding compound from *Nocardia asteroides*. *The Biochemical journal* **1974**, *139*, 407-413.
45. Sakagami, H.; Ishihara, M.; Hoshino, Y.; Ishikawa, J.; Mikami, Y.; Fukai, T., Cytotoxicity of nocobactins NA-a, NA-b and their ferric complexes assessed by semiempirical molecular orbital method. *In vivo (Athens, Greece)* **2005**, *19*, 277-282.
46. Ratledge, C.; Patel, P. V., Isolation, properties and taxonomic relevance of lipid-soluble, iron-binding compounds (the nocobactins) from *Nocardia*. *J. Gen. Microbiol.* **1976**, *93*, 141-152.
47. Ikeda, Y.; Nonaka, H.; Furumai, T.; Onaka, H.; Igarashi, Y., Nocardimicins A, B, C, D, E, and F, siderophores with muscarinic M3 receptor inhibiting activity from *Nocardia* sp. TP-A0674. *Journal of natural products* **2005**, *68*, 1061-1065.
48. Ikeda, Y.; Furumai, T.; Igarashi, Y., Nocardimicins G, H and I, siderophores with muscarinic M3 receptor binding inhibitory activity from *Nocardia nova* JCM 6044. *The Journal of antibiotics* **2005**, *58*, 566-572.
49. Kokubo, S.; Suenaga, K.; Shinohara, C.; Tsuji, T.; Uemura, D., Structures of Amamistatins A and B, Novel Growth Inhibitors of Human Tumor Cell Lines from *Nocardia asteroides*. *Tetrahedron* **2000**, *56*, 6435-6440.
50. Tsuda, M.; Yamakawa, M.; Oka, S.; Tanaka, Y.; Hoshino, Y.; Mikami, Y.; Sato, A.; Fujiwara, H.; Ohizumi, Y.; Kobayashi, J., Brasilibactin A, a cytotoxic compound from actinomycete *Nocardia brasiliensis*. *Journal of natural products* **2005**, *68*, 462-464.
51. Murakami, Y.; Kato, S.; Nakajima, M.; Matsuoka, M.; Kawai, H.; Shinya, K.; Seto, H., Formobactin, a novel free radical scavenging and neuronal cell protecting substance from *Nocardia* sp. *The Journal of antibiotics* **1996**, *49*, 839-845.
52. Mitchell, J. M.; Shaw, J. T., Synthesis and Stereochemical Assignment of Brasilibactin A. *Organic Letters* **2007**, *9*, 1679-1681.
53. Tsukamoto, M.; Murooka, K.; Nakajima, S.; Abe, S.; Suzuki, H.; Hirano, K.; Kondo, H.; Kojiri, K.; Suda, H., BE-32030 A, B, C, D and E, new antitumor substances produced by *Nocardia* sp. A32030. *The Journal of antibiotics* **1997**, *50*, 815-821.
54. Lane, S. J.; Marshall, P. S.; Upton, R. J.; Ratledge, C.; Ewing, M., Novel extracellular mycobactins, the carboxymycobactins from *Mycobacterium avium*. *Tetrahedron Letters* **1995**, *36*, 4129-4132.
55. Ratledge, C.; Ewing, M., The occurrence of carboxymycobactin, the siderophore of pathogenic mycobacteria, as a second extracellular siderophore in *Mycobacterium smegmatis*. *Microbiology (Reading, England)* **1996**, *142*, 2207-2212.

56. Ratledge, C.; Dover, L., Iron Metabolism in Pathogenic Bacteria. *Annual review of microbiology* **2000**, *54*, 881-941.
57. Nemoto, A.; Hoshino, Y.; Yazawa, K.; Ando, A.; Mikami, Y.; Komaki, H.; Tanaka, Y.; Grafe, U., Asterobactin, a new siderophore group antibiotic from *Nocardia asteroides*. *The Journal of antibiotics* **2002**, *55*, 593-597.
58. Bosello, M.; Zeyadi, M.; Kraas, F. I.; Linne, U.; Xie, X.; Marahiel, M. A., Structural Characterization of the Heterobactin Siderophores from *Rhodococcus erythropolis* PR4 and Elucidation of Their Biosynthetic Machinery. *Journal of natural products* **2013**, *76*, 2282-2290.
59. Mukai, A.; Komaki, H.; Takagi, M.; Shin-ya, K., Novel siderophore, JBIR-16, isolated from *Nocardia tenerifensis* NBRC 101015. *The Journal of antibiotics* **2009**, *62*, 601-603.
60. Schneider, K.; Rose, I.; Vikineswary, S.; Jones, A. L.; Goodfellow, M.; Nicholson, G.; Beil, W.; Sussmuth, R. D.; Fiedler, H. P., Nocardichelins A and B, siderophores from *Nocardia strain acta* 3026. *Journal of natural products* **2007**, *70*, 932-935.
61. Stephan, H.; Freund, S.; Beck, W.; Jung, G.; Meyer, J.; Winkelmann, G., Ornibactins - a new family of siderophores from *Pseudomonas*. *Biometals : an international journal on the role of metal ions in biology, biochemistry, and medicine* **1993**, *6*, 93-100.
62. Mukai, A.; Fukai, T.; Matsumoto, Y.; Ishikawa, J.; Hoshino, Y.; Yazawa, K.; Harada, K.; Mikami, Y., Transvalencin Z, a new antimicrobial compound with salicylic acid residue from *Nocardia transvalensis* IFM 10065. *The Journal of antibiotics* **2006**, *59*, 366-369.
63. Kautsar, S. A.; Blin, K.; Shaw, S.; Navarro-Muñoz, J. C.; Terlouw, B. R.; van der Hooft, J. J. J.; van Santen, J. A.; Tracanna, V.; Suarez D.; G., H.; V., Pascal A.; Selem-Mojica, N.; Alanjary, M.; Robinson, S. L.; Lund, G.; Epstein, S. C.; Sisto, A. C.; Charkoudian, L. K.; Collemare, J.; Lington, R. G.; Weber, T.; Medema, M. H., MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Research* **2019**, *48*, D454-D458.
64. Heathcote, M. L.; Staunton, J.; Leadlay, P. F., Role of type II thioesterases: evidence for removal of short acyl chains produced by aberrant decarboxylation of chain extender units. *Cell Chemical Biology* **2001**, *8*, 207-220.
65. Liu, X.; Jin, Y.; Cui, Z.; Nonaka, K.; Baba, S.; Funabashi, M.; Yang, Z.; Van Lanen, S. G., The Role of a Nonribosomal Peptide Synthetase in L-Lysine Lactamization During Capuramycin Biosynthesis. *Chembiochem : a European journal of chemical biology* **2016**, *17*, 804-810.
66. Reddy, P. V.; Puri, R. V.; Chauhan, P.; Kar, R.; Rohilla, A.; Khera, A.; Tyagi, A. K., Disruption of mycobactin biosynthesis leads to attenuation of *Mycobacterium tuberculosis* for growth and virulence. *The Journal of Infectious Diseases* **2013**, *208*, 1255-1265.
67. Quadri, L. E.; Sello, J.; Keating, T. A.; Weinreb, P. H.; Walsh, C. T., Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Cell Chemical Biology* **1998**, *5*, 631-45.
68. Chavadi, S. S.; Stirrett, K. L.; Edupuganti, U. R.; Vergnolle, O.; Sadhanandan, G.; Marchiano, E.; Martin, C.; Qiu, W. G.; Soll, C. E.; Quadri, L. E. N., Mutational and Phylogenetic Analyses of the *Mycobacterium tuberculosis* mbt Gene Cluster. *Journal of Bacteriology* **2011**, *193*, 5905-5913.

References

69. Krithika, R.; Marathe, U.; Saxena, P.; Ansari, M. Z.; Mohanty, D.; Gokhale, R. S., A genetic locus required for iron acquisition in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 2069-2074.
70. De Voss, J. J.; Rutter, K.; Schroeder, B. G.; Su, H.; Zhu, Y.; Barry, C. E., 3rd, The salicylate-derived mycobactin siderophores of Mycobacterium tuberculosis are essential for growth in macrophages. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97*, 1252-1257.
71. Cimermancic, P.; Medema, M. H.; Claesen, J.; Kurita, K.; Wieland Brown, L. C.; Mavrommatis, K.; Pati, A.; Godfrey, P. A.; Koehrsen, M.; Clardy, J.; Birren, B. W.; Takano, E.; Sali, A.; Lington, R. G.; Fischbach, M. A., Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **2014**, *158*, 412-421.
72. Nett, M.; Ikeda, H.; Moore, B. S., Genomic basis for natural product biosynthetic diversity in the actinomycetes. *Natural product reports* **2009**, *26*, 1362-1384.
73. Batut, B.; Knibbe, C.; Marais, G.; Daubin, V., Reductive genome evolution at both ends of the bacterial population size spectrum. *Nature Reviews Microbiology* **2014**, *12*, 841.
74. Cenicerros, A.; Dijkhuizen, L.; Petrusma, M.; Medema, M. H., Genome-based exploration of the specialized metabolic capacities of the genus Rhodococcus. *BMC Genomics* **2017**, *18*, 593.
75. Moran, N. A., Microbial minimalism: genome reduction in bacterial pathogens. *Cell* **2002**, *108*, 583-586.
76. Markowitz, V. M.; Chen, I. M. A.; Palaniappan, K.; Chu, K.; Szeto, E.; Grechkin, Y.; Ratner, A.; Jacob, B.; Huang, J.; Williams, P. G.; Huntemann, M.; Anderson, I.; Mavromatis, K.; Ivanova, N. N.; Kyrpides, N. C., IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* **2012**, *40*, D115-D122.
77. Agarwala R.; Barrett T.; Beck J.; Benson D. A.; Bollin C.; Bolton E.; Bourexis D.; Brister J. R.; Bryant S. H.; Canese K.; Cavanaugh M.; Charowhas C.; Clark K.; Dondoshansky I.; Feolo M.; Fitzpatrick L.; Funk K.; Geer L. Y.; Gorelenkov V.; Graeff A.; Hlavina W.; Holmes B.; Johnson M.; Kattman B.; Khotomlianski V.; Kimchi A.; Kimelman M.; Kimura M.; Kitts P.; Klimke W.; Kotliarov A.; Krasnov S.; Kuznetsov A.; Landrum M. J.; Landsman D.; Lathrop S.; Lee J. M.; Leubsdorf C.; Lu Z.; Madden T. L.; Marchler-Bauer A.; Malheiro A.; Meric P.; Karsch-Mizrachi I.; Mnev A.; Murphy T.; Orris R.; Ostell J.; O'Sullivan C.; Palanigobu V.; Panchenko A. R.; Phan L.; Pierov B.; Pruitt K. D.; Rodarmer K.; Sayers E. W.; Schneider V.; Schoch C. L.; Schuler G. D.; Sherry S. T.; Siyan K.; Soboleva A.; Soussov V.; Starchenko G.; Tatusova T. A.; Thibaud-Nissen F.; Todorov K.; Trawick B. W.; Vakarov D.; Ward M.; Yaschenko E.; Zasytkin A.; Zbicz K., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **2018**, *46*, D8-D13.
78. Vera-Cabrera, L.; Ortiz-Lopez, R.; Elizondo-Gonzalez, R.; Ocampo-Candiani, J., Complete Genome Sequence Analysis of Nocardia brasiliensis HUJEG-1 Reveals a Saprobic Lifestyle and the Genes Needed for Human Pathogenesis. *PLoS One* **2013**, *8*, e65425.
79. Imajoh, M.; Fukumoto, Y.; Yamane, J.; Sukeda, M.; Shimizu, M.; Ohnishi, K.; Oshima, S., Draft Genome Sequence of Nocardia seriolae Strain N-2927 (NBRC 110360), Isolated as the Causal Agent of Nocardiosis of Yellowtail (Seriola quinqueradiata) in Kochi Prefecture, Japan. *Genome Announcements* **2015**, *3*.
80. Vautrin, F.; Bergeron, E.; Dubost, A.; Abrouk, D.; Martin, C.; Cournoyer, B.; Louzier, V.; Winiarski, T.; Rodriguez-Nava, V.; Pujic, P., Genome Sequences of Three Nocardia cyriacigeorgica

Strains and One *Nocardia asteroides* Strain. *Microbiology Resource Announcements* **2019**, *8*, e00600-e00619.

81. Schorn, M. A.; Alanjary, M. M.; Aguinaldo, K.; Korobeynikov, A.; Podell, S.; Patin, N.; Lincecum, T.; Jensen, P. R.; Ziemert, N.; Moore, B. S., Sequencing rare marine actinomycete genomes reveals high density of unique natural product biosynthetic gene clusters. *Microbiology (Reading, England)* **2016**, *162*, 2075-2086.

82. Doroghazi, J. R.; Metcalf, W. W., Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* **2013**, *14*, 1-13.

83. Ziemert, N.; Alanjary, M.; Weber, T., The evolution of genome mining in microbes - a review. *Natural product reports* **2016**, *33*, 988-1005.

84. Hopwood, D. A.; Bibb, M. J.; Chater, K. F.; Kieser, T.; Bruton, C. J.; Kieser, H. M.; Lydiate, D. J.; Smith, C. P.; Ward, J. M.; Schrempf, H., *Genetic Manipulation of Streptomyces: a laboratory manual*. International Union of Biochemistry and Molecular Biology Inc.1985; Vol. 14.

85. Ziemert, N.; Lechner, A.; Wietz, M.; Millán-Aguiñaga, N.; Chavarria, K. L.; Jensen, P. R., Diversity and evolution of secondary metabolism in the marine actinomycete genus *Salinispora*. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, E1130-E1139.

86. Medema, M. H.; Fischbach, M. A., Computational approaches to natural product discovery. *Nature Chemical Biology* **2015**, *11*, 639-648.

87. Bachmann, B. O.; Van Lanen, S. G.; Baltz, R. H., Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *Journal of industrial microbiology & biotechnology* **2014**, *41*, 175-184.

88. Challis, G. L., Mining microbial genomes for new natural products and biosynthetic pathways. *Microbiology (Reading, England)* **2008**, *154*, 1555-1569.

89. Baltz, R. H., Renaissance in antibacterial discovery from actinomycetes. *Current Opinion in Pharmacology* **2008**, *8*, 557-563.

90. Ikeda, H.; Nonomiya, T.; Usami, M.; Ohta, T.; Omura, S., Organization of the biosynthetic gene cluster for the polyketide anthelmintic macrolide avermectin in *Streptomyces avermitilis*. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 9509-9514.

91. Bentley, S. D.; Chater, K. F.; Cerdeno-Tarraga, A. M.; Challis, G. L.; Thomson, N. R.; James, K. D.; Harris, D. E.; Quail, M. A.; Kieser, H.; Harper, D.; Bateman, A.; Brown, S.; Chandra, G.; Chen, C. W.; Collins, M.; Cronin, A.; Fraser, A.; Goble, A.; Hidalgo, J.; Hornsby, T.; Howarth, S.; Huang, C. H.; Kieser, T.; Larke, L.; Murphy, L.; Oliver, K.; O'Neil, S.; Rabinowitsch, E.; Rajandream, M. A.; Rutherford, K.; Rutter, S.; Seeger, K.; Saunders, D.; Sharp, S.; Squares, R.; Squares, S.; Taylor, K.; Warren, T.; Wietzorrek, A.; Woodward, J.; Barrell, B. G.; Parkhill, J.; Hopwood, D. A., Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **2002**, *417*, 141-147.

92. Udvary, D. W.; Zeigler, L.; Asolkar, R. N.; Singan, V.; Lapidus, A.; Fenical, W.; Jensen, P. R.; Moore, B. S., Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proceedings of the National Academy of Sciences of the United States of America* **2007**, *104*, 10376-10381.

References

93. Seyedsayamdost, M. R., High-throughput platform for the discovery of elicitors of silent bacterial gene clusters. *Proceedings of the National Academy of Sciences of the United States of America* **2014**, *111*, 7266-7271.
94. Abdelmohsen, U. R.; Grkovic, T.; Balasubramanian, S.; Kamel, M. S.; Quinn, R. J.; Hentschel, U., Elicitation of secondary metabolism in actinomycetes. *Biotechnology advances* **2015**, *33*, 798-811.
95. Doroghazi, J. R.; Albright, J. C.; Goering, A. W.; Ju, K. S.; Haines, R. R.; Tchalukov, K. A.; Labeda, D. P.; Kelleher, N. L.; Metcalf, W. W., A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nature Chemical Biology* **2014**, *10*, 963-968.
96. Nguyen, D. D.; Wu, C. H.; Moree, W. J.; Lamsa, A.; Medema, M. H.; Zhao, X.; Gavilan, R. G.; Aparicio, M.; Atencio, L.; Jackson, C.; Ballesteros, J.; Sanchez, J.; Watrous, J. D.; Phelan, V. V.; van de Wiel, C.; Kersten, R. D.; Mehnaz, S.; De Mot, R.; Shank, E. A.; Charusanti, P.; Nagarajan, H.; Duggan, B. M.; Moore, B. S.; Bandeira, N.; Palsson, B.; Pogliano, K.; Gutiérrez, M.; Dorrestein, P. C., MS/MS networking guided analysis of molecule and gene cluster families. *Proceedings of the National Academy of Sciences of the United States of America* **2013**, *110*, E2611-E2620.
97. Skinnider, M. A.; Dejong, C. A.; Rees, P. N.; Johnston, C. W.; Li, H.; Webster, A. L.; Wyatt, M. A.; Magarvey, N. A., Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). *Nucleic Acids Research* **2015**, *43*, 9645-9662.
98. Weber, T.; Blin, K.; Duddela, S.; Krug, D.; Kim, H. U.; Bruccoleri, R.; Lee, S. Y.; Fischbach, M. A.; Muller, R.; Wohlleben, W.; Breitling, R.; Takano, E.; Medema, M. H., antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research* **2015**, *43*, W237-W243.
99. Ziemert, N.; Podell, S.; Penn, K.; Badger, J. H.; Allen, E.; Jensen, P. R., The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* **2012**, *7*, e34064.
100. Blin, K.; Kim, H. U.; Medema, M. H.; Weber, T., Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics* **2019**, *20*, 1103-1113.
101. Blin, K.; Wolf, T.; Chevrette, M. G.; Lu, X.; Schwalen, C. J.; Kautsar, S. A.; Suarez Duran, H. G.; de Los Santos, E. L. C.; Kim, H. U.; Nave, M.; Dickschat, J. S.; Mitchell, D. A.; Shelest, E.; Breitling, R.; Takano, E.; Lee, S. Y.; Weber, T.; Medema, M. H., antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research* **2017**, *45*, W36-W41.
102. van Heel, A. J.; de Jong, A.; Montalban-Lopez, M.; Kok, J.; Kuipers, O. P., BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Research* **2013**, *41*, W448-W453.
103. Skinnider, M. A.; Merwin, N. J.; Johnston, C. W.; Magarvey, N. A., PRISM 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Research* **2017**, *45*, W49-W54.
104. Tietz, J. I.; Schwalen, C. J.; Patel, P. S.; Maxson, T.; Blair, P. M.; Tai, H. C.; Zakai, U. I.; Mitchell, D. A., A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nature Chemical Biology* **2017**, *13*, 470-478.

105. Khaldi, N.; Seifuddin, F. T.; Turner, G.; Haft, D.; Nierman, W. C.; Wolfe, K. H.; Fedorova, N. D., SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology* **2010**, *47*, 736-741.
106. Eddy, S. R., A new generation of homology search tools based on probabilistic inference. *Genome Informatics* **2009**, *23*, 205-211.
107. El-Gebali, S.; Mistry, J.; Bateman, A.; Eddy, S. R.; Luciani, A.; Potter, S. C.; Qureshi, M.; Richardson, L. J.; Salazar, G. A.; Smart, A.; Sonnhammer, E. L. L.; Hirsh, L.; Paladin, L.; Piovesan, D.; Tosatto, S. C. E.; Finn, R. D., The Pfam protein families database in 2019. *Nucleic Acids Research* **2018**, *47*, D427-D432.
108. Sonnhammer, E. L. L.; Eddy, S. R.; Birney, E.; Bateman, A.; Durbin, R., Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Research* **1998**, *26*, 320-322.
109. 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 Research* **2003**, *13*, 2498-2504.
110. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *Journal of Molecular Biology* **1990**, *215*, 403-410.
111. Bastian, M.; Heymann, S.; Jacomy, M., Gephi: An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media* **2009**, *3*, 361-362.
112. Gross, H., Strategies to unravel the function of orphan biosynthesis pathways: recent examples and future prospects. *Applied microbiology and biotechnology* **2007**, *75*, 267-277.
113. Fisch, K. M.; Schäberle, T. F., Toolbox for Antibiotics Discovery from Microorganisms. *Arch Pharm (Weinheim)* **2016**, *349*, 683-691.
114. Takeda, I.; Umemura, M.; Koike, H.; Asai, K.; Machida, M., Motif-independent prediction of a secondary metabolism gene cluster using comparative genomics: application to sequenced genomes of *Aspergillus* and ten other filamentous fungal species. *DNA Research* **2014**, *21*, 447-457.
115. Richter, M.; Rossello-Mora, R., Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America* **2009**, *106*, 19126-19131.
116. Maiden, M. C.; Bygraves, J. A.; Feil, E.; Morelli, G.; Russell, J. E.; Urwin, R.; Zhang, Q.; Zhou, J.; Zurth, K.; Caugant, D. A.; Feavers, I. M.; Achtman, M.; Spratt, B. G., Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, *95*, 3140-3145.
117. Alanjary, M.; Steinke, K.; Ziemert, N., AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. *Nucleic Acids Research* **2019**, *47*, W276-W282.
118. Zhang, W.; Du, P.; Zheng, H.; Yu, W.; Wan, L.; Chen, C., Whole-genome sequence comparison as a method for improving bacterial species definition. *The Journal of general and applied microbiology* **2014**, *60*, 75-78.

References

119. Ondov, B. D.; Treangen, T. J.; Melsted, P.; Mallonee, A. B.; Bergman, N. H.; Koren, S.; Phillippy, A. M., Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **2016**, *17*, 132.
120. Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W., CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **2015**, *25*, 1043-1055.
121. Luo, Q.; Hiessl, S.; Steinbuchel, A., Functional diversity of *Nocardia* in metabolism. *Environmental Microbiology* **2014**, *16*, 29-48.
122. Medema, M. H.; Blin, K.; Cimermancic, P.; de Jager, V.; Zakrzewski, P.; Fischbach, M. A.; Weber, T.; Takano, E.; Breitling, R., antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* **2011**, *39*, W339-W346.
123. Adamek, M.; Alanjary, M.; Sales-Ortells, H.; Goodfellow, M.; Bull, A. T.; Winkler, A.; Wibberg, D.; Kalinowski, J.; Ziemert, N., Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. *BMC Genomics* **2018**, *19*, 426.
124. Ward, A. C.; Allenby, N. E. E., Genome mining for the search and discovery of bioactive compounds: the *Streptomyces* paradigm. *FEMS Microbiology Letters* **2018**, *365*, 1-20.
125. Carro, L.; Nouioui, I.; Sangal, V.; Meier-Kolthoff, J. P.; Trujillo, M. E.; Montero-Calasanz, M. C.; Sahin, N.; Smith, D. L.; Kim, K. E.; Peluso, P.; Deshpande, S.; Woyke, T.; Shapiro, N.; Kyrpides, N. C.; Klenk, H.; Göker, M.; Goodfellow, M., Genome-based classification of micromonosporae with a focus on their biotechnological and ecological potential. *Scientific reports* **2018**, *8*, 525.
126. Poralla, K.; Muth, G.; Hartner, T., Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). *FEMS Microbiology Letters* **2000**, *189*, 93-95.
127. Marrakchi, H.; Laneelle, M. A.; Daffe, M., Mycolic acids: structures, biosynthesis, and beyond. *Cell Chemical Biology* **2014**, *21*, 67-85.
128. Robinson, S. L.; Terlouw, B. R.; Smith, M. D.; Pidot, S. J.; Stinear, T. P.; Medema, M. H.; Wackett, L. P., Global analysis of adenylate-forming enzymes reveals β -lactone biosynthesis pathway in pathogenic *Nocardia*. *bioRxiv* **2019**.
129. Wolf, F.; Bauer, J. S.; Bendel, T. M.; Kulik, A.; Kalinowski, J.; Gross, H.; Kaysser, L., Biosynthesis of the beta-Lactone Proteasome Inhibitors Belactosin and Cystargolide. *Angewandte Chemie - International Edition* **2017**, *56*, 6665-6668.
130. Kaysser, L., Built to bind: biosynthetic strategies for the formation of small-molecule protease inhibitors. *Natural product reports* **2019**, *36*, 1654-1686.
131. McGlinchey, R. P.; Nett, M.; Eustaquio, A. S.; Asolkar, R. N.; Fenical, W.; Moore, B. S., Engineered biosynthesis of antiprotealide and other unnatural salinosporamide proteasome inhibitors. *Journal of the American Chemical Society* **2008**, *130*, 7822-7823.
132. Peters, P.; Galinski, E. A.; Trüper, H. G., The biosynthesis of ectoine. *FEMS Microbiology Letters* **1990**, *71*, 157-162.

133. Bursy, J.; Kuhlmann, A. U.; Pittelkow, M.; Hartmann, H.; Jebbar, M.; Pierik, A. J.; Bremer, E., Synthesis and uptake of the compatible solutes ectoine and 5-hydroxyectoine by *Streptomyces coelicolor* A3(2) in response to salt and heat stresses. *Appl. Environ. Microbiol.* **2008**, *74*, 7286-7296.
134. Ofer, N.; Wishkautzan, M.; Meijler, M.; Wang, Y.; Speer, A.; Niederweis, M.; Gur, E., Ectoine biosynthesis in *Mycobacterium smegmatis*. *Appl. Environ. Microbiol.* **2012**, *78*, 7483-7486.
135. Guinand, M.; Vacheron, M. J.; Michel, G., Structure des parois cellulaires des nocardia. I- isolement et composition des parois de *Nocardia kirovani*. *FEBS Letters* **1970**, *6*, 37-39.
136. Osawa, A.; Kasahara, A.; Mastuoka, S.; Gassel, S.; Sandmann, G.; Shindo, K., Isolation of a novel carotenoid, OH-chlorobactene glucoside hexadecanoate, and related rare carotenoids from *Rhodococcus* sp. CIP and their antioxidative activities. *Bioscience, Biotechnology and Biochemistry* **2011**, *75*, 2142-2147.
137. Valdes-Stauber, N.; Scherer, S., Isolation and characterization of Linocin M18, a bacteriocin produced by *Brevibacterium linens*. *Appl. Environ. Microbiol.* **1994**, *60*, 3809-3914.
138. Sutter, M.; Boehringer, D.; Gutmann, S.; Gunther, S.; Prangishvili, D.; Loessner, M. J.; Stetter, K. O.; Weber-Ban, E.; Ban, N., Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology* **2008**, *15*, 939-947.
139. Rosenkrands, I.; Rasmussen, P. B.; Carnio, M.; Jacobsen, S.; Theisen, M.; Andersen, P., Identification and characterization of a 29-kilodalton protein from *Mycobacterium tuberculosis* culture filtrate recognized by mouse memory effector cells. *Infection and immunity* **1998**, *66*, 2728-2735.
140. McClean, S.; Healy, M. E.; Collins, C.; Carberry, S.; O'Shaughnessy, L.; Dennehy, R.; Adams, A.; Kennelly, H.; Corbett, J. M.; Carty, F.; Cahill, L. A.; Callaghan, M.; English, K.; Mahon, B. P.; Doyle, S.; Shinoy, M., Linocin and OmpW Are Involved in Attachment of the Cystic Fibrosis-Associated Pathogen *Burkholderia cepacia* Complex to Lung Epithelial Cells and Protect Mice against Infection. *Infection and Immunology* **2016**, *84*, 1424-1437.
141. Muliandi, A.; Katsuyama, Y.; Sone, K.; Izumikawa, M.; Moriya, T.; Hashimoto, J.; Kozono, I.; Takagi, M.; Shin-ya, K.; Ohnishi, Y., Biosynthesis of the 4-methyloxazoline-containing nonribosomal peptides, JBIR-34 and -35, in *Streptomyces* sp. Sp080513GE-23. *Cell Chemical Biology* **2014**, *21*, 923-934.
142. Ratledge, C.; Hall, M. J., Influence of Metal Ions on the Formation of Mycobactin and Salicylic Acid in *Mycobacterium smegmatis* Grown in Static Culture. *Journal of Bacteriology* **1971**, *108*, 314-319.
143. McMahon, M. D.; Rush, J. S.; Thomas, M. G., Analyses of MbtB, MbtE, and MbtF suggest revisions to the mycobactin biosynthesis pathway in *Mycobacterium tuberculosis*. *Journal of Bacteriology* **2012**, *194*, 2809-2818.
144. McQueen, C. F.; Groves, J. T., A reevaluation of iron binding by Mycobactin J. *Journal of Biological Inorganic Chemistry* **2018**, *23*, 995-1007.
145. Lee, S. W.; Mitchell, D. A.; Markley, A. L.; Hensler, M. E.; Gonzalez, D.; Wohlrab, A.; Dorrestein, P. C.; Nizet, V.; Dixon, J. E., Discovery of a widely distributed toxin biosynthetic gene cluster. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 5879-5884.

References

146. Chen, J., Genom-getriebene Untersuchungen und Isolierung von NRPS-basierten Siderophoren. *PhD thesis* **2018**.
147. Lautru, S.; Deeth, R. J.; Bailey, L. M.; Challis, G. L., Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nature Chemical Biology* **2005**, *1*, 265-269.
148. Emery, T., Exchange of iron by gallium in siderophores. *Biochemistry* **1986**, *25*, 4629-4633.

6. Acknowledgements

I would like to thank JProf. Dr. Leonard Kaysser and Prof. Dr. Nadine Ziemert, who gave me the possibility to write this dissertation and to work interdisciplinary in two working groups with such a highly motivating topic. Furthermore, I would like to thank them for being such great supervisors, who were always willing to help patiently, when problems with any concern arose.

I would like to thank Prof. Dr. Bradley Moore for giving me the opportunity to broaden my horizon during my stay as a visiting research scholar at Scripps Institution of Oceanography (University of San Diego, UCSD) in San Diego, California, USA.

Extraordinary thanks go to Prof. Dr. Shaun McKinnie for his great support in NMR, motivation, and reachability throughout the total time of research visit at Scripps Institution of Oceanography.

I would like to thank Prof. Dr. Harald Gross and PD Dr. Bertolt Gust, who were always open for questions related to laboratory work or general academia related questions. Furthermore, I would like to thank them for their kind agreement to examine my oral PhD presentation.

A big thanks also goes to Dr. Martina Adamek and Dr. Mohammad Alanjary for help when questions in bioinformatics arose. Your hints were a big help in progress of this thesis.

Thank you so much Corinna Fischer, Emmanuel Wemakor, Kerstin Seeger, and Lucy Westrich for your great support in laboratory and administrative organization.

Furthermore, I want to thank everybody from the Microbiology/ Biotechnology and Pharmaceutical Biology working groups who made my work possible by helping me with any concerns imaginable. I would like to highlight the hard-working times when a short distraction with a coffee break was saving.

Special thanks go to the Fulbright Association for funding my research at UCSD and for providing such a great community of talented people.

Finally, I would like to thank my family and friends, in particular Sophia Singer for distraction from work, especially when free time was rare during writing of my thesis.

7. Curriculum Vitae

Personal Data

Name: Daniel Christoph Männle
Date of Birth: 04.11.1987
Place of Birth: Offenburg
Nationality: German
e-mail: daniel-maennle@web.de

Postdoc

Analytics, Science & Technology Microbiology
7/2020 – today
Novartis Stein AG Switzerland

PhD

Pharmaceutical Biology (Pharmacy Department) / Microbiology/Biotechnology (Biology Department)
11/2015 – 06/2020
Thesis: "Comparative Genomics to Metabolomics: Nocobactin Production in Nocardia"
Eberhard Karls University of Tübingen/ Scripps Institution of Oceanography (La Jolla, California, USA)
Supervisor: JProf. Dr. Leonard Kaysser/ Prof. Dr. Nadine Ziemert

Studies

Master of Science Microbiology (Biology Department)
10/2013 – 10/2015
Thesis: "Localization of plasmid DNA in *Streptomyces lividans* by Fluorescence *in situ* Hybridization (FISH)"
Eberhard Karls University of Tübingen
Supervisor: Dr. Guenther Muth

Bachelor of Science Geoecology (Geology & Biology Departments)
10/2010 – 9/2013
Thesis: "Microbial PAH degradation potential in biochar amended soil"
Eberhard Karls University of Tübingen
Supervisor: Prof. Dr. Andreas Kappler

School

Schiller Gymnasium Offenburg
08/1998 - 06/2007

8. Appendix

“Comparative Genomics and Metabolomics in the Genus *Nocardia*”

Männle, D.; McKinnie, S. M. K.; Mantri, S. S.; Steinke, K.; Lu, Z.; Moore, B. S.; Ziemert, N.; Kaysser, L.; mSystems **2020**, 5, e00125-e00125-20, <https://doi.org/10.1128/mSystems.00125-20>

“New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A-C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406”

Chen, J.; Frediansyah, A.; Männle, D.; Straetener, J.; Brötz-Oesterhelt, H.; Ziemert, N.; Kaysser, L.; Gross, H.;

Chembiochem **2020**, 21, 1-10, <https://doi.org/10.1002/cbic.202000062>



Comparative Genomics and Metabolomics in the Genus *Nocardia*

 Daniel Männle,^{a,b,c}  Shaun M. K. McKinnie,^d  Shrikant S. Mantri,^{b,c} Katharina Steinke,^{b,c} Zeyin Lu,^{a,b}
 Bradley S. Moore,^{e,f}  Nadine Ziemert,^{b,c}  Leonard Kaysser^{a,b}

^aPharmaceutical Biology, Eberhard Karls University Tübingen, Tübingen, Germany

^bGerman Centre for Infection Research (DZIF), Tübingen, Germany

^cInterfaculty Institute for Microbiology and Infection Medicine Tübingen (IMIT), Microbiology and Biotechnology, University of Tübingen, Tübingen, Germany

^dDepartment of Chemistry and Biochemistry, University of California, Santa Cruz, California, USA

^eScripps Institution of Oceanography, University of California, San Diego, California, USA

^fSkaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, California, USA

ABSTRACT Using automated genome analysis tools, it is often unclear to what degree genetic variability in homologous biosynthetic pathways relates to structural variation. This hampers strain prioritization and compound identification and can lead to overinterpretation of chemical diversity. Here, we assessed the metabolic potential of *Nocardia*, an underinvestigated actinobacterial genus that is known to comprise opportunistic human pathogens. Our analysis revealed a plethora of putative biosynthetic gene clusters of various classes, including polyketide, nonribosomal peptide, and terpenoid pathways. Furthermore, we used the highly conserved biosynthetic pathway for nocobactin-like siderophores to investigate how gene cluster differences correlate to structural differences in the produced compounds. Sequence similarity networks generated by BiG-SCAPE (Biosynthetic Gene Similarity Clustering and Prospecting Engine) showed the presence of several distinct gene cluster families. Metabolic profiling of selected *Nocardia* strains using liquid chromatography-mass spectrometry (LC-MS) metabolomics data, nuclear magnetic resonance (NMR) spectroscopy, and GNPS (Global Natural Product Social molecular networking) revealed that nocobactin-like biosynthetic gene cluster (BGC) families above a BiG-SCAPE threshold of 70% can be assigned to distinct structural types of nocobactin-like siderophores.

IMPORTANCE Our work emphasizes that *Nocardia* represent a prolific source for natural products rivaling better-characterized genera such as *Streptomyces* or *Amycolatopsis*. Furthermore, we showed that large-scale analysis of biosynthetic gene clusters using similarity networks with high stringency allows the distinction and prediction of natural product structural variations. This will facilitate future genomics-driven drug discovery campaigns.

KEYWORDS *Nocardia*, biosynthetic gene cluster, genome mining, genomics, metabolomics, natural products, siderophores

Bacterial natural products are an important source for small-molecule pharmaceuticals. Historically, they have played a major role in what is arguably the greatest success story of modern medicine: the containment of microbial infectious diseases. Nowadays, natural products still represent frequently used lead structures for the development of antibiotic, anticancer, and immunosuppressive agents (1). However, over the last decades, the pharmaceutical industry has shown dwindling interest in classic natural product screening programs. This is primarily due to the increasing rediscovery rate of already known molecules especially in well-investigated bacterial

Citation Männle D, McKinnie SMK, Mantri SS, Steinke K, Lu Z, Moore BS, Ziemert N, Kaysser L. 2020. Comparative genomics and metabolomics in the genus *Nocardia*. mSystems 5:e00125-20. <https://doi.org/10.1128/mSystems.00125-20>.

Editor David F. Savage, University of California, Berkeley

Copyright © 2020 Männle et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Nadine Ziemert, nadine.ziemert@uni-tuebingen.de, or Leonard Kaysser, leonard.kaysser@uni-tuebingen.de.

Received 10 February 2020

Accepted 13 May 2020

Published 2 June 2020

genera such as *Streptomyces*. In recent years, several developments have revived interest in microorganisms as a source of new drug leads. On the one hand, underinvestigated bacterial taxa were shown to comprise prolific producers of novel bioactive chemistry (2). On the other hand, genome sequencing revealed that even in extensively studied bacteria, the molecules isolated so far represent only the tip of the iceberg (3). The biosynthetic capacity of filamentous actinobacteria, such as *Streptomyces*, and myxobacteria, cyanobacteria, and many others, often exceeds 30 or more predicted secondary metabolites per strain. Notably, most of these orphan pathways cannot be associated with known compounds, making them a formidable source for new chemistry and drug leads. However, with more than 400,000 putative gene clusters in the database of the Joint Genome Institute (JGI), the development of bioinformatic tools to assess and prioritize these pathways has become crucial (4). For this purpose, a number of different web-based genome analysis platforms have been established allowing gene-to-molecule predictions of varying accuracy (5). Firm knowledge of the limitations of such approaches is of critical importance to evaluate the novelty of encoded compounds with sufficient confidence.

Here, we explore the potential of the genus *Nocardia* for the biosynthesis of natural products. Moreover, we use the highly conserved nocobactin pathway to probe thresholds and limits of popular gene cluster analysis tools. *Nocardia* are rod-shaped partially acid-fast actinobacteria of the order *Corynebacteriales*, closely related to *Rhodococcus*, *Corynebacterium*, and *Mycobacterium*. They are able to form branching, filamentous mycelia that can fragment into bacteroid nonmotile elements. *Nocardia* are ubiquitous saprophytes often found in organic-rich soil, but they are most prominently known as opportunistic pathogens. Nocardiosis is an infectious disease affecting the immunocompromised with a high mortality rate, but it has also been reported for immunocompetent patients (6). Commonly, *Nocardia* cause localized cutaneous or pulmonary infections depending on the nature of exposure to the pathogen. From here, the bacteria may disseminate, leading to systemic forms of nocardiosis and severe brain infections. The pathogenicity of *Nocardia* has been extensively studied by Beaman and others (7). Still, major aspects are poorly understood, including the role of virulence factors in strain-specific development and characteristics of the disease. Over the last decades, *Nocardia* have also served as the occasional source for natural products (Fig. 1) (8–11). However, in comparison to other genera of the phylum, the chemistry of *Nocardia* is vastly underexplored. Notably, the preliminary analysis of the first seven available *Nocardia* genomes by Komaki et al. in 2014 showed a diverse repertoire of nonribosomal peptide synthetase (NRPS) and type I polyketide synthase (PKS) pathways (12). The authors found that the content of respective gene clusters varies considerably between different clinically relevant *Nocardia* pathogens, ranging from 12 such pathways in *Nocardia farcinica* IFM 10152 to 30 in some investigated *Nocardia brasiliensis* strains (IFM 10847 and HUJEG-1). NRPS and type I PKS are modular biosynthetic machineries in which every module is responsible for the incorporation of a building block by assembly line logic. They are important driving forces for the generation of chemical diversity and structural complexity in bacterial natural products. Other relevant classes of secondary metabolites include type II and type III polyketides, terpenoids, ribosomally synthesized and posttranslationally modified peptides (RiPPs), alkaloids, and nucleosides. In the current report, we present a comprehensive analysis of the biosynthetic capacity of the genus *Nocardia* based on 103 published genomes. To this end, we used an integrated workflow that combines the most popular gene cluster analysis tool antiSMASH and the recently developed BiG-SCAPE algorithm to generate sequence similarity networks (13, 14). With this setup, we observed tremendous potential for the production of bioactive small molecules rivaling *Streptomyces* in number and diversity. Several biosynthetic pathways were found to be conserved throughout the genus including a gene cluster for the formation of nocobactin-type siderophores. Using this pathway as a proxy, we probed different stringencies to construct the sequence similarity networks and correlated the generated gene cluster families (GCFs) with a mass spectrometry (MS) metabolic network.

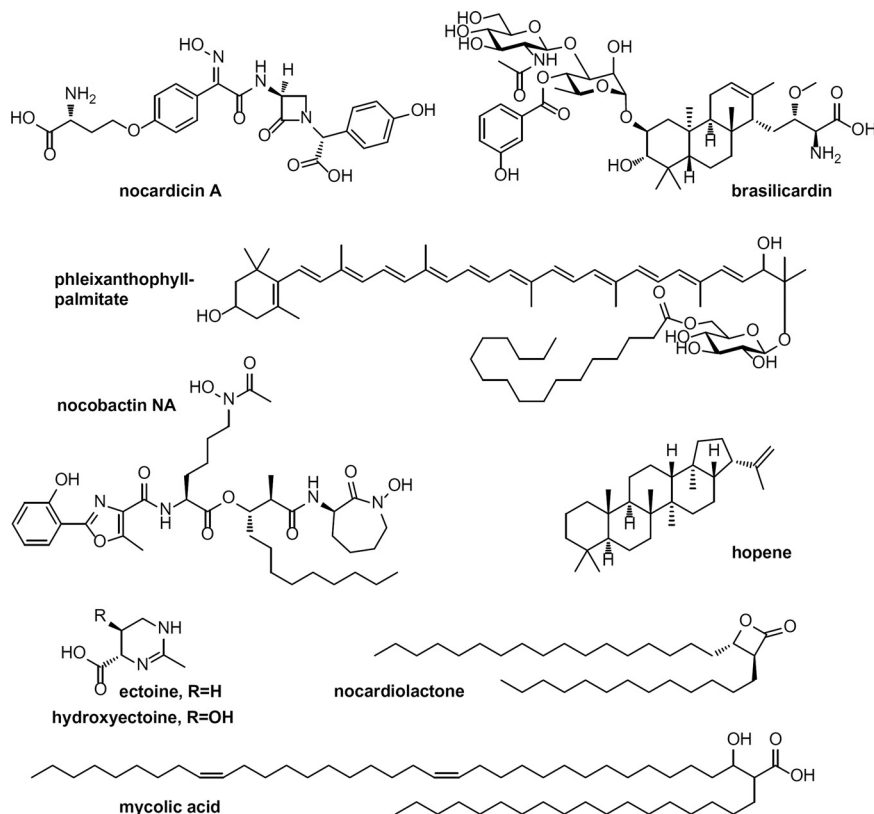


FIG 1 Selected known and predicted compounds from *Nocardia* spp.

Strikingly, by applying a BiG-SCAPE threshold of 70%, we were able to produce GCFs that could be assigned to the synthesis of structurally distinct groups of nocobactin derivatives.

RESULTS AND DISCUSSION

Phylogeny and biosynthetic potential of the genus *Nocardia*. A search in public databases such as NCBI and JGI found 103 nonredundant genome sequences from organisms named *Nocardia*, as of 2017. The genomes had an average of 147 contigs (1 to 615) with an estimated completeness of at least 98.24%, as analyzed by checkM (see Table S20 at <https://doi.org/10.5281/zenodo.3784407>). In order to obtain an overview about the diversity of the genus and evolutionary relationships within the genus, we built phylogenetic trees using the web tool autoMLST (15). To this end, 63 conserved housekeeping genes with neutral dN/dS (ratio of nonsynonymous to synonymous evolutionary changes) values were identified by autoMLST and used for the concatenated alignment as a basis for tree building. The resulting tree (Fig. 2) shows six major clades (A to F) within the genus and one distinct branch (X) formed by the single strain *Nocardia harenae* NBRC 108248. Interestingly, two strains formed a clade outside the outgroups, indicating that they do not belong to the genus *Nocardia*. MASH analysis (16) to estimate the average nucleotide identity (ANI) showed that both strains, named *Nocardia* sp. strain NRRL 5-836 and *Nocardia* sp. strain 348MFTsu5.1, are more similar to *Lentzea guizhouensis* and *Williamsia muralis* NBRC 105860 with an ANI of 91.12% and 78.91%, respectively, thus providing further evidence that these strains are incorrectly assigned and do not belong to the genus *Nocardia* (17). Furthermore, these results were supported by checkM analysis, as both strains fall into different lineage markers. We therefore did not include these two strains in further analysis. From the 101 remaining analyzed *Nocardia* strains, 50 are clinical isolates, seven have been isolated as animal and seven as plant pathogens, and 31 are derived from environmental sources.

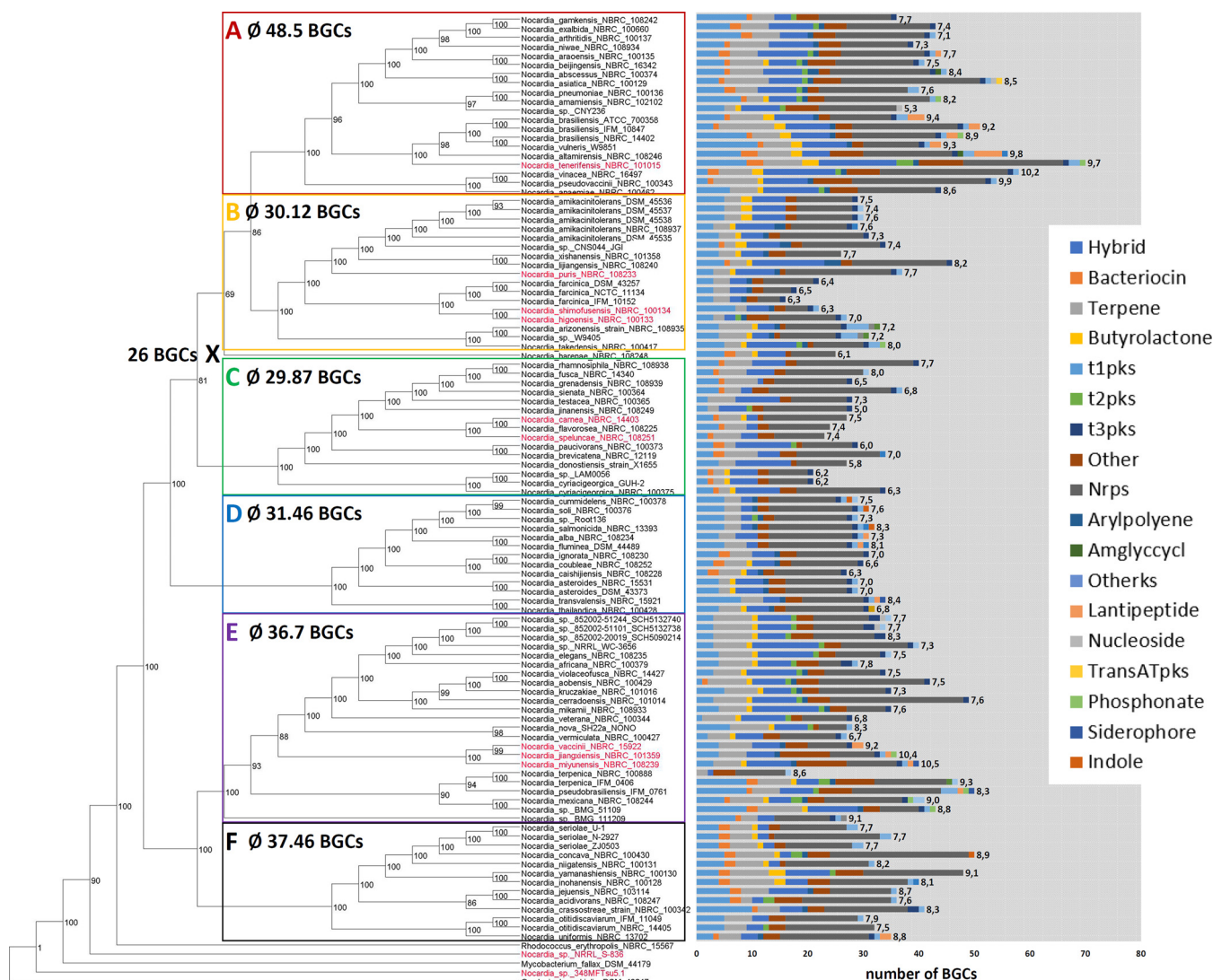


FIG 2 *Nocardia* phylogeny and BGC distribution. Maximum likelihood tree of 103 members of the genus *Nocardia* based on 63 concatenated housekeeping genes identified with autoMLST (see Fig. S2 and Table S3 at <https://doi.org/10.5281/zenodo.3784407>). Ultrafast bootstrap values were calculated using 1,000 replicates. Numbers next to the bars (predicted BGCs by antiSMASH) represent genome sizes (in megabase pairs). Average numbers of BGCs per clade are shown. X labels a distinct branch formed by a single *Nocardia* strain, *Nocardia harenae* NBRC 108248.

Nocardia of all origins are evenly spread across the six phylogenetic clades, suggesting that pathogenicity of *Nocardia* strains is not reflected by evolutionary relationship (see Fig. S30 at <https://doi.org/10.5281/zenodo.3784407>). To assess the biosynthetic potential of the strains, we performed computational genome mining analysis using the program antiSMASH (13). All analyzed genomes contain a variety of biosynthetic gene clusters (BGCs) belonging to the major biosynthetic classes (Fig. 2). On average, a *Nocardia* genome features 36 BGCs and a predicted overall chemical/metabolic diversity of 35% NRPs, 18% PKs, 13% terpenoids, 12% hybrids, 2.2% bacteriocins, 2.2% arylpolyenes, 2% butyrolactones, 1.1% RiPPs, and 14.5% others. Gratifyingly, the number of identified pathways did not correlate with the number of contigs per genome ($R^2 = 0.28$; see Fig. S32 at <https://doi.org/10.5281/zenodo.3784407>). Although this warrants a certain degree of robustness to our data, actual BGC numbers might vary slightly, as antiSMASH results were not reinspected for broken or merged clusters. Interestingly, we could observe a correlation between the phylogenetic clade and BGC abundance. Clades A, E, and F contain on average more BGCs than clades B and C. However, even in strains with relatively low BGC numbers, at least 16 BGCs were

detected. Similar to their sister genus *Rhodococcus*, the most prevalent BGCs in *Nocardia* are NRPSs (18). A positive correlation was found between the number of biosynthetic gene clusters in each strain predicted by antiSMASH and the respective genome size (see Fig. S1 at <https://doi.org/10.5281/zenodo.3784407>). Overall, the established biosynthetic capabilities of *Nocardia* strains are comparable to those of *Streptomyces* and *Amycolatopsis*, based on recently reported diversity plots (18–20).

Genomic features of clinically relevant *Nocardia* strains. *Mycobacterium* and *Rhodococcus* pathogenic strains generally have smaller genomes as well as fewer and less diverse BGCs than the rest of the respective genera (21). The opportunistic nature of nocardiosis, the high diversity of causative species, and the insufficient knowledge on *Nocardia* virulence factors makes the distinction of pathogenic and nonpathogenic strains difficult. Accordingly, we were not able to observe a correlation between the genome size or BGC content and the different sources of isolation (see Fig. S30 at <https://doi.org/10.5281/zenodo.3784407>). It is fair to assume that in principle, many *Nocardia* strains can act as facultative pathogens. Nevertheless, some *Nocardia* species are reported more often from nocardiosis patients than others, i.e., *N. asteroides*, *N. cyriacigeorgica*, *N. brasiliensis*, *N. abscessus*, *N. farcinica*, *N. nova*, and *N. otitidiscaviarum*, further referred to as group 1 pathogens (6, 22–24). Indeed, the genomes of these pathogens are on average slightly smaller than the rest of the genus (7.4 Mbp versus 7.8 Mbp) and contain fewer BGCs (32.5 versus 36.5). A remarkable exception are the different strains of *N. brasiliensis* with genomes of 9.4 (ATCC 700358), 9.2 (IFM 10847), and 8.9 (NBRC 101014) Mbp in size and an outstanding chemical potential with 49, 56, and 49 predicted BGCs, respectively. It is notable that *N. farcinica* IFM 10152, a clinically prevalent group 1 pathogen, with 6.3 Mbp contains the smallest genome and the lowest number of predicted BGCs (16 BGCs) of all *Nocardia*. Intriguingly, a majority of these clusters are conserved in other *Nocardia* spp., sometimes throughout the genus (see Table S22 at <https://doi.org/10.5281/zenodo.3784407>). It could be speculated that the BGCs found in *N. farcinica* represent some version of a minimal secondary metabolome required to support both a lifestyle as a soil-dwelling saprophyte and as a successful opportunistic pathogen. A preliminary survey of putative *Nocardia* virulence factors such as catalases and superoxide dismutases showed no significant trend regarding number and identity of respective genes in group 1 strains versus the rest of the genus (see Table S21 at <https://doi.org/10.5281/zenodo.3784407>). This could suggest that strain-specific pathogenicity is mainly driven by other characteristics.

Metabolic diversity and gene cluster families in *Nocardia*. To assess the diversity of BGCs in the genus in more detail, we performed a network analysis using the BiG-SCAPE algorithm. BiG-SCAPE employs Pfam composition similarity to calculate BGC distances that takes a weighted combination of the Jaccard index, adjacency index, and domain sequence similarity score into account (14, 25). The generated sequence similarity network with a default similarity score cutoff of $c = 0.5$ (equals threshold of 50%) consisted of 2,836 BGCs as unique nodes and 22,455 edges (Fig. 3 and 4). Altogether, 258 independent gene cluster families (GCFs) were formed with two or more nodes and 750 singletons. Notably, several large superclusters are visible, which are composed of subclusters with apparently different biosynthetic features (Fig. 4). These superclusters could not be satisfyingly resolved using higher similarity cutoffs or different alignment modes. Such problems have been reported from other comparative genome analyses and attributed to fragmented genome assemblies and deficient separation of individual BGCs by antiSMASH. (18). Indeed, manual inspection of the *Nocardia* BGC similarity network revealed a number of distinct subclusters we considered additional GCFs (Fig. 4). Next, we wanted to assess whether the clinically most relevant *Nocardia* species exhibit a characteristic secondary metabolism and conducted a number of statistical analyses. Boxplots along with t test indicate that the BGC class frequency distribution between group 1 and other *Nocardia* genomes is insignificant (see Fig. S33 at <https://doi.org/10.5281/zenodo.3784407>). Moreover, principal-component analysis (PCA) plots show that the GCF presence/absence infor-

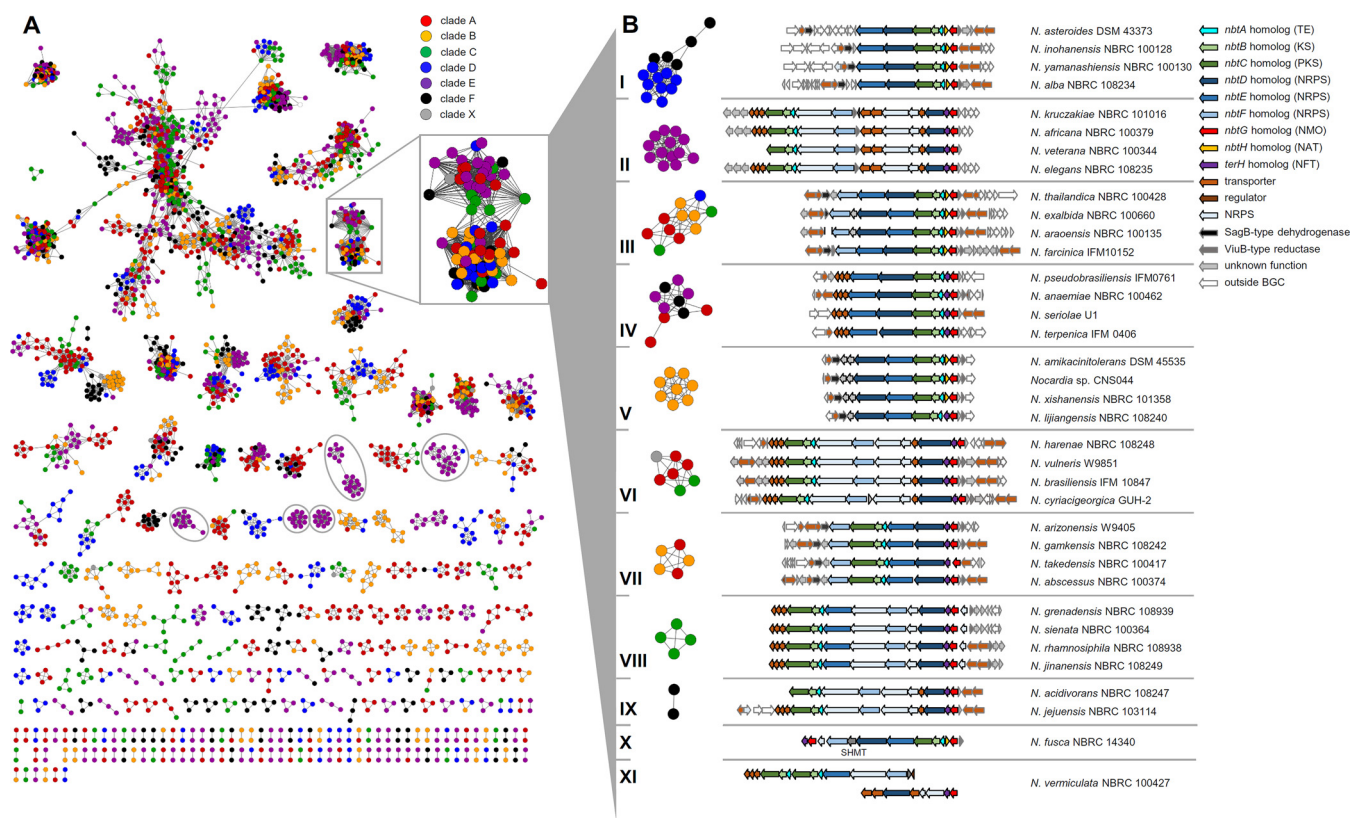


FIG 3 BiG-SCAPE sequence similarity network (SSN) and nocobactin-like subfamilies visualized in Cytoscape. (A) SSN ($c = 0.5$) of *Nocardia* spp. Each node represents one BGC identified by antiSMASH 4.0. Colors represent phylogenetic clades A to F and X. Clade-specific GCFs are highlighted. (B) SSN ($c = 0.7$) of nocobactin-like BGCs. Representative pathways of different subfamilies (I to IX). Two singletons (X and XI) are shown. Abbreviations: TE, thioesterase; KS, ketosynthase; PKS, polyketide synthase; NRPS, nonribosomal peptide synthetase; NMO, *N*-monoxygenase; NAT, *N*-acetyltransferase; NFT, *N*-formyltransferase; SHMT, serine hydroxymethyltransferase.

mation is insufficient for clustering the *Nocardia* genomes into separate groups based on clinical annotations (group 1; see Fig. S34 at <https://doi.org/10.5281/zenodo.3784407>). Overall, it appears that none of the BGCs are significantly more abundant or even specific to prevalent *Nocardia* pathogens in comparison to the rest of the genus. However, correlation can be observed between phylogenetic clades and BGC distribution within the sequence similarity network (Fig. 3). In particular, strains from clade E feature a number of clusters that are not found in bacteria from other clades, including pathways for terpenoid, NRPS, and PKS biosynthesis. For example, a gene cluster with high similarity to hopene biosynthesis is found in all *Nocardia* of clade E and otherwise only in *Nocardia transvalensis* NBRC 15921. Hopanoids (Fig. 1) are pentacyclic triterpenoid components of membranes from diverse prokaryotes (26). They have stabilizing functions similar to sterols in eukaryotes and might be of use to members of clade E in specific environments. In contrast, BGCs of *Nocardia* spp. attributed to clade F do not form a single characteristic GCF with representatives from more than seven strains.

A number of GCFs comprise clusters, which are highly abundant in the investigated *Nocardia* genomes (Fig. 4). This includes several GCFs that represent pathways with mono-modular type I PKS-like ketosynthases, i.e., GCF 1, 11, and 14, potentially involved in the modification of fatty acid precursors. For example, GCF 1 represents a highly conserved cluster found throughout the genus, which likely directs the synthesis of mycolic acids. Mycolic acids (Fig. 1) are characteristic and essential cell wall components of mycobacteria and related genera, e.g., *Corynebacterium*, *Rhodococcus*, *Gordonia*, and *Nocardia* (27). Depending on the producing organism, these molecules differ in chain length and degree of unsaturation. *Nocardia* mycolic acids usually consist of a C_{32-40} β -hydroxy chain and a C_{12-18} α -alkyl side chain. A typical cluster of GCF 1 encodes a

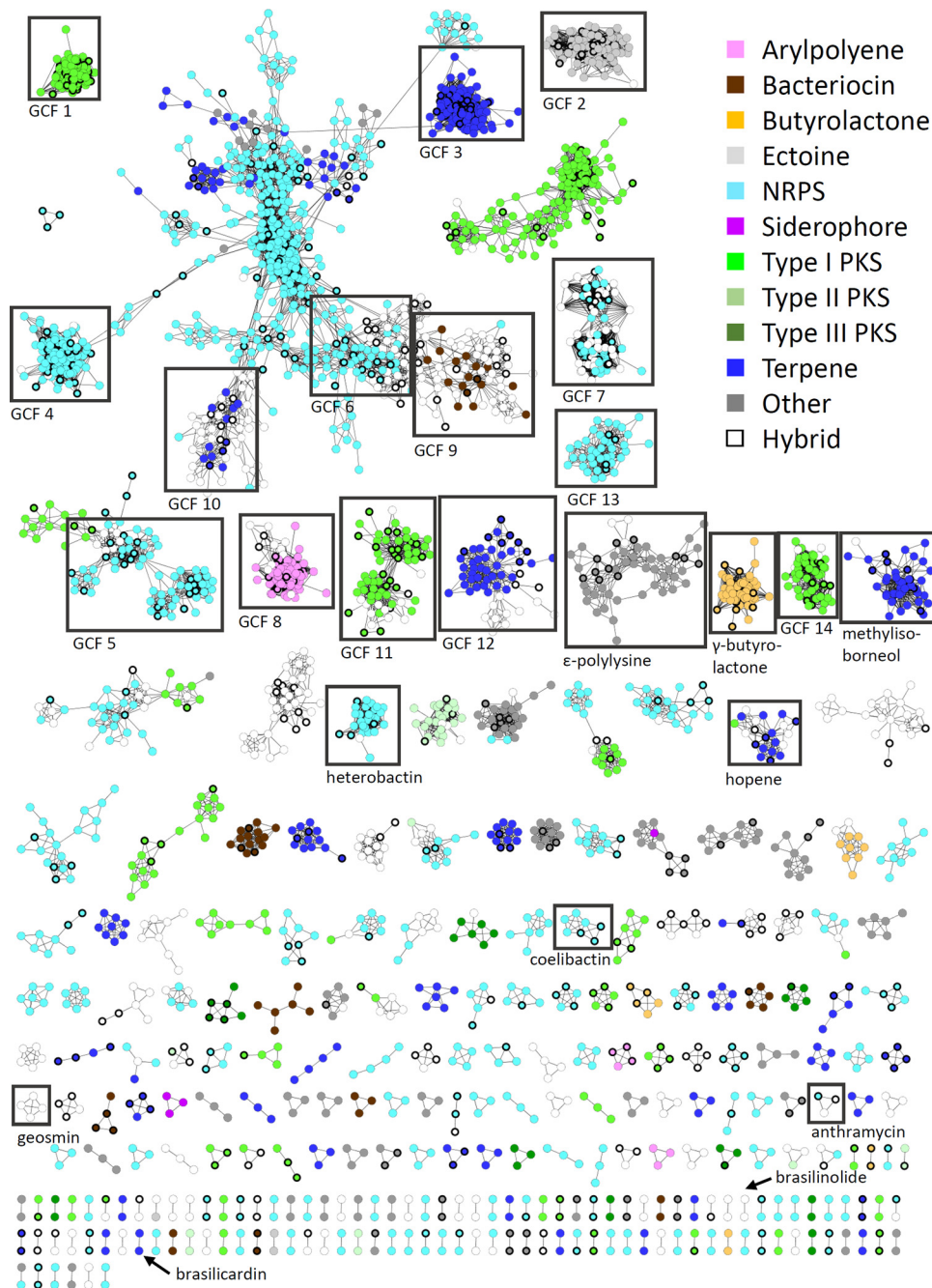


FIG 4 BiG-SCAPE cluster type sequence similarity network of all BGCs from *Nocardia* strains found with antiSMASH 4.0. Colors show cluster types from *Nocardia* BGCs. Colorless nodes are hybrid BGCs found in *Nocardia*. Bold node borders refer to group 1 pathogens.

homolog of Pks13, the essential condensing enzyme to form mycolic acid from two fatty acyl coenzyme A (acyl-CoA) precursors (28). It contains additional genes that are assigned to the biosynthesis and transfer of arabinose and galactose. Together, these enzymes presumably participate in the assembly of the *Nocardia* mycolyl-arabino-galactan-peptidoglycan complex. Nocardiolactone (Fig. 1) is a fatty acid derivative with similarity to lipstatin that has been isolated from different *Nocardia* strains (29). Recently, Robinson and coworkers reported on the identification of the nocardiolactone gene cluster featuring a β -lactone synthetase (30). β -Lactone-containing natural products are potent inhibitors of diverse hydrolytic enzymes, and their biosynthesis has

drawn a lot of attention (31–33). The lower subcluster of GCF 11 contains 36 homologous nocardiolactone BGCs of which six have been initially assigned by antiSMASH as putative lipstatin-type pathways (Fig. 4).

The BGCs that form GCF 2 are found in all *Nocardia* genomes and encode enzymes for the synthesis of ectoine (Fig. 1), a compatible solute which stabilizes biopolymers (DNA, proteins, etc.) under extreme conditions (34). The capability to produce ectoine or 5-hydroxyectoine is not uncommon to actinobacterial genera such as *Streptomyces* and *Mycobacterium* (35, 36). The strict conservation of this pathway in *Nocardia* seems to reflect a general requirement of all members of the genus to be able to function under osmotic stress. It was recently shown that the ectoine BGC is similarly conserved in the sister genus *Rhodococcus* (18).

Nocardia are known to produce colorful pigments ranging from yellow to red, but a systematic analysis of the carotenoid content in this genus has not been conducted (7). GCF 3 summarizes pathways that are likely responsible for the formation of a carotenoid of unknown structure. The BGC is highly conserved in all *Nocardia* genomes. It encodes polyprenyl synthase (CrtE), phytoene synthase (CrtB), and phytoene desaturase (CrtI) homologs for the production of lycopene. An obvious candidate for a cyclase is not found in the cluster. However, a putative lycopene cyclase present in all *Nocardia* species is the main feature of the conserved terpene GCF 10. Together, both BGCs could direct the concerted biosynthesis of mono- or bicyclic carotenes. Further modifications may involve the attachment of a glycosyl group similar to sioxanthin (37). Such monocyclic carotenoid glycosides and their esters (Fig. 1) have been isolated from *Nocardia kirovani* and different *Rhodococcus* species (38, 39). Arylpolyenes are another common class of pigments and generated by an unusual type II PKS system. These molecules are mostly known from Gram-negative gammaproteobacteria and bacteroides (25). A pathway with signature motifs for the synthesis of aryl polyenes is widespread in *Nocardia* (82 of 101 genomes) and clustered in GCF 8.

The gene clusters organized in GCF 4 are also found in all *Nocardia* species. Here, a single NRPS with four modules is colocated with a putative *N*-acyltransferase and two transporters. The produced peptide likely consists of an N-terminal serine and three additional amino acids. The three residues cannot be predicted with certainty from the respective incorporating A domains but presumably feature aliphatic or aromatic side chains (see Table S4 at <https://doi.org/10.5281/zenodo.3784407>). Two other conserved NRPS pathways could be identified. One of them (GCF 5) consists of a trimodular NRPS gene with epimerization (E) domains encoded in the first module and in the terminal module. A peptide sequence could not be predicted on the basis of the A domains. This BGC is present in 75 out of 101 *Nocardia* genomes. Similarly conserved is a large NRPS assembly line of different size represented by GCF 6. The enzyme complex may contain 7 to 13 modules, but the gene cluster environment is invariant and always encodes a putative nitropropane dioxygenase, a metallopeptidase, an esterase, an epimerase, and an enoyl-CoA dehydratase.

Linocin M18 has been described as a secreted bacteriolysin (formerly type III bacteriocin) from *Brevibacterium linens* M18 (40). Similar proteins were found in various other bacteria, including *Mycobacterium tuberculosis* and *Burkholderia cepacia* complex. They could be associated with pathogen-host cell attachment and immunogenicity (41, 42). Moreover, linocins have been shown to form multimeric microcompartments (43). A respective orthologous gene is encoded in all *Nocardia* genomes and summarized in GCF 9.

The gene cluster family of nocobactins. Another highly abundant BGC family (GCF 7) is found in 92 of 101 *Nocardia* genomes and encodes a conserved hybrid NRPS/PKS pathway annotated by antiSMASH as putative nocobactin NA BGC (>50% homology; Fig. 3A). Nocobactin NA (Fig. 1) was first reported in 1974 as a UV-active siderophore from cultures of *N. asteroides* by Ratledge and Snow (44). Nocobactins were subsequently found in other *Nocardia* spp., but not in *Rhodococcus* (45). A number of derivatives have been isolated since, including the amamistatins, nocardimicins,

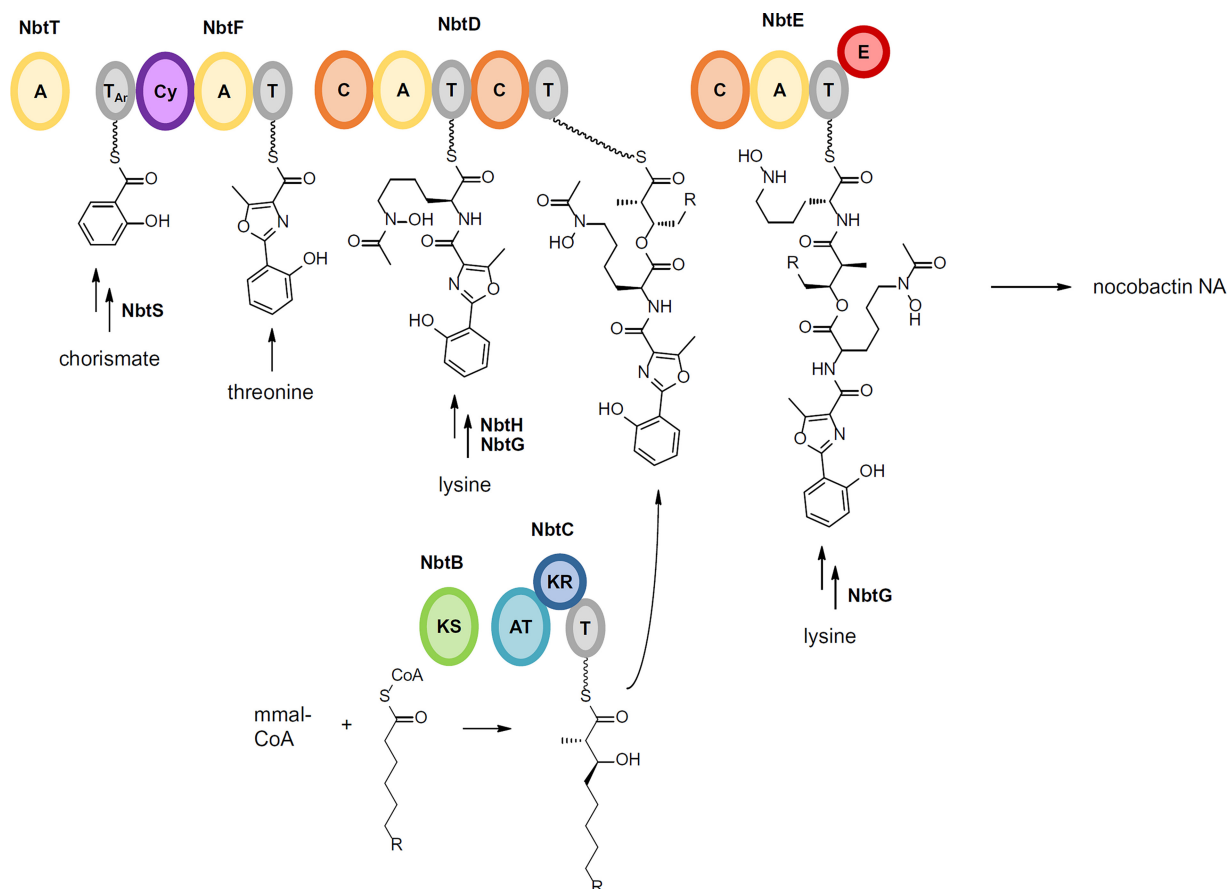


FIG 5 Model for nocobactin NA biosynthesis. NRPS/PKS domains: adenylation (A), thiolation (T), heterocyclization (Cy), condensation (C), epimerization (E), ketosynthase (KS), acyltransferase (AT), ketoreductase (KR), aryl carrier protein (T_A), methylmalonyl CoA (mm-CoA).

brasilibactin, formobactin, and terpenibactins (46–50), which all share a hydroxyphenyloxazoline/-oxazole moiety, a *N*-acetylated or *N*-formylated *N*-hydroxylysine that is linked via an ester moiety to a long-chain α -methyl, β -hydroxy fatty acid, and a C-terminal *N*-hydroxy α -amino ϵ -caprolactam ring. The biosynthetic pathway for nocobactin NA reported from *N. farcinica* IFM 10152 involves a nonlinear assembly line consisting of three NRPS modules and one PKS module (Fig. 5) (51). NbtF (T-cyclization domain [Cy]-A-T) is proposed to catalyze the condensation and heterocyclization of salicylic acid and threonine with the help of the stand-alone A-domain NbtT. NbtD (C-A-T-C-T) activates and adds lysine, or a derivative, to the growing chain. NbtH and NbtG, either before or after incorporation of the lysine moiety, likely direct *N*-acylation and *N*-hydroxylation, respectively. NbtB (KS) and NbtC (AT-ketoreductase [KR] domain-T) are proposed to be responsible for the activation of a fatty acid which is extended by Claisen-type condensation with methylmalonyl-CoA and subsequently reduced at the β -position. The incorporation of the fatty acyl moiety by ester linkage is probably mediated by NbtD before NbtE (C-A-T-E) attaches another lysine. Epimerization of this amino acid to the D isomer is followed by *N*-hydroxylation and formation of the ϵ -caprolactam. Notably, the nocobactin pathway is split in two gene clusters, which are located 180 kb apart on the *N. farcinica* IFM 10152 chromosome. Cluster I contains the majority of the biosynthetic genes, whereas cluster II codes for NbtT and a putative salicylate synthase, NbtS. This splitting is reflected in the BiG-SCAPE similarity network by GCF 13 that is annotated by antiSMASH as mycobactin-like with 30% similarity. In fact, this GCF contains a small pathway of three conserved genes with high homology to *nbtT*, *nbtS*, and *nbtF*. It appears that nocobactin cluster II (which is not deposited in the MIBiG database) is recognized bioinformatically only if the NRPS module NbtF is

encoded. Indeed, all strains containing such a cluster II variant lack an *nbtF* homolog in their respective nocobactin cluster I. This pathway organization is shared by almost 45% of *Nocardia* spp. The widespread occurrence of the nocobactin NA-like pathway in *Nocardia* genomes indicates a constitutive role for this siderophore in the biology of the genus. Structurally and biosynthetically, the nocobactins are closely related to the mycobactins which have similar prominence in mycobacteria (52). Both compound classes differ in the chain length of their acyl substituents. Mycobactins feature a long-chain fatty acid attached to the *N*-hydroxylysine moiety instead of the formyl/acetyl group in nocobactin-like molecules and a central 3-hydroxybutyrate group. They are the quintessential siderophores for the genus *Mycobacterium* and are produced by all species except *M. paratuberculosis*. These molecules exist typically as membrane-bound iron scavengers but can also be produced in a more polar form as extracellular carboxymycobactins (53). Mycobactins of both variants are important for the lifestyle of pathogenic mycobacteria and have been shown to play a fundamental role in the survival and growth of *M. tuberculosis* in human macrophages (54, 55). It is tempting to speculate that nocobactins may have a similar function as virulence factors in the progression of nocardiosis. The first evidence for this theory was obtained by Hoshino et al. studying $\Delta nbtE$ deficient mutants in infection assays with the macrophage-like cell line J774A.1 (51).

To assess the metabolic diversity of nocobactins in the genus *Nocardia*, we analyzed the composition of the BGC family generated by BiG-SCAPE in more detail. Notably, the nocobactin GCF obtained with a threshold of 50% shows two prominent subclusters (Fig. 3A). Both subclusters are characterized on a genetic level by the exclusive presence of a gene for an *N*-acetyltransferase (NbtH homolog) or an *N*-formyltransferase. This finding suggests that the subdivision of the nocobactin GCF reflects the production of structurally different derivatives with either an acetyl or formyl group attached to the *N*-hydroxylysine residue. From the reported nocobactin-like siderophores, nocobactin NA and nocardimicins A to F feature an acetyl moiety, whereas nocardimicins G to I, formobactin, brasilibactin, amamistatins, and terpenibactins feature a formyl moiety. A candidate enzyme for this reaction, TerH, has been identified in the terpenibactin pathway (50). Further refinement of the BiG-SCAPE similarity network with a threshold of 70% resulted in the segmentation of the nocobactin cluster I GCF to nine independent families of more than two BGCs (further referred to as nocobactin-like BGC subfamilies) and nine singletons (Fig. 3B). Interestingly, the composition of these new subfamilies correlates to some degree with the phylogenetic clades observed by autoMLST analysis. Subfamilies II, V, VIII, and IX consist solely of pathways from clades E, B, C, and F, respectively. Subfamilies I and VII are dominated by clades D and F and clades A and B, respectively. Again, the absence or presence of NbtH/TerH homologs is a major discriminating factor that determines these subfamilies. Subfamilies I and III carry a gene that encodes an NbtH, and subfamilies II, IV, VI, and VII to IX carry a gene that encodes a TerH homolog. Only subfamily V comprises pathways that contain either an *nbtH* or *terH* gene. Additionally, splitting of subfamilies is based on the clustering of the *nbtF* homolog. Subfamilies II, III, VI, VII, VIII, and IX exhibit a putative *nbtF* gene in cluster I, whereas in subfamilies I, IV, and V, *nbtF* is collocated with cluster II. Further characteristic family features are additional and/or alternative NRPS modules in subfamilies II, VI, VIII, and IX. Here, the nocobactin-like pathways have a gene that encodes a discrete C-T didomain protein, probably a functional substitute to the missing terminal C-T domains of the NbtD homolog, and another NRPS enzyme with a C-A-T-E-C domain structure. The specificity of the corresponding A domain could not be predicted bioinformatically. It is worth mentioning that the previously reported nocobactin-type molecules exhibit diverse stereochemical configurations. Nocobactin NA features a central acylated *N*-hydroxy L-lysine and a terminal *N*-hydroxy D-lysine 1,6-lactam. In contrast, amamistatin A incorporates D- and L-lysine and brasilibactin A incorporates D- and D-lysine derivatives at these positions, respectively (51, 56, 57). Possibly, the production of different stereoisomers might rely on the presence of additional E-domain-containing NRPS enzymes as found in subfamilies II, VI, VIII, and IX.

Moreover, in subfamilies VI and VIII, the PKS enzymes with NbtC homology are extended by a C-methyltransferase domain to an AT-KR-T-cMT module with predicted AT domain specificity to methylmalonyl-CoA (mmal-CoA). Notably, formobactin A and the amamistatins contain a rare α -dimethylated fatty acid which could derive from such a biosynthetic setup. It is also worth mentioning that the colocalization of a gene for a putative SagB-type dehydrogenase is characteristic for the clusters subsumed in subfamilies I, III, V, and VII. SagB is involved in the biosynthesis of streptolysin S toxin by catalyzing the oxidation of oxazoline and thiazoline heterocycles to form oxazole and thiazole ring structures (58). Interestingly, both heterocyclic variants exist in nocobactin-type siderophores, with oxazole moieties in nocobactin NA, formobactin, the amamistatins and nocardimicins A to F, and oxazolines in brasilibactin A, nocardimicins G to I, and the terpenibactins. The established subfamilies can also be distinguished based on the association of the pathways with transporter genes. In subfamilies II, IV, VI, VIII, and IX, such functionalities are encoded in the same operon as the biosynthetic genes. In subfamilies I, III, V, and VII, transporters are encoded immediately adjacent to the clusters. Nine gene clusters were not clustered with the other subfamilies. One of these pathways (X), found in *Nocardia fusca* NBRC 14340, contains homologs to all genes of the nocobactin NA gene cluster. Additionally, it features a second copy of *nbtF*, a *terH* homolog, a gene for a discrete A domain and a putative serine hydroxymethyltransferase gene. Another singleton cluster (XI) is encoded in *Nocardia vermiculata* NBRC 100427 exhibiting two copies of an *nbtABC*-like operon but lacking *nbtD*, *nbtG*, and *nbtH/terH*-like genes. The missing homologs are located in an independent cluster together with genes for a second NRPS C-A-T-E-C module and several transporters.

Nocobactin-like molecules produced by different *Nocardia* strains. Clearly, the different families obtained from the BiG-SCAPE similarity network are composed of pathways with characteristic biosynthetic functionalities. We next wanted to inquire how these distinct cluster organizations correspond to structural diversity in the produced nocobactin-type siderophores. Therefore, we examined the intracellular and membrane-associated culture fractions from a selection of *Nocardia* strains representative of the identified nocobactin-like BGC subfamilies (see Table S19 at <https://doi.org/10.5281/zenodo.3784407>). The culture extracts were subjected to positive ion mode liquid chromatography-mass spectrometry (LC-MS) analysis. To stabilize molecules with metal-chelating properties and facilitate their identification, we supplemented the samples with $\text{Ga}_2(\text{SO}_4)_3$ under reducing conditions. This treatment resulted in the almost complete equipment of metal-free siderophores with Ga(III) and largely replaced Fe(III) in already loaded compounds. The generated gallium complexes could be detected by mass spectrometry via their distinct isotopic pattern. The tandem mass spectrometry (MS/MS) spectra obtained from data-dependent acquisition experiments were used to construct a molecular network with the help of the GNPS (Global Natural Product Social Molecular Networking) platform (59). For an internal reference, we included pure samples of the recently discovered terpenibactins in our analysis (see Table S5 at <https://doi.org/10.5281/zenodo.3784407>) (50). The resulting network of ionizable metabolites from 12 *Nocardia* strains together with the terpenibactins represented 3,567 nodes and 4,177 edges (Fig. 6). At a cosine value of 0.6, it forms 108 clusters of two or more nodes and shows 2,470 singletons. The observed clusters likely represent molecular families of compounds with similar chemical structures. The terpenibactins clustered with one molecular family containing 21 putative metabolites of seven analyzed *Nocardia* strains. A closer inspection showed that this association is based on the consistent fragmentation of the terpenibactins along the ester linkage between the modified lysine residue and the β -hydroxy fatty acid moiety. This produces characteristic fragment ions of m/z 460.0621 (major), m/z 414.0552 (major), and m/z 442.0477 (minor) for terpenibactin A (see Fig. S4 at <https://doi.org/10.5281/zenodo.3784407>). A similar pattern can be observed for terpenibactin B and terpenibactin C with deviating fragment ions containing the variable fatty acyl moiety (see Fig. S3 at

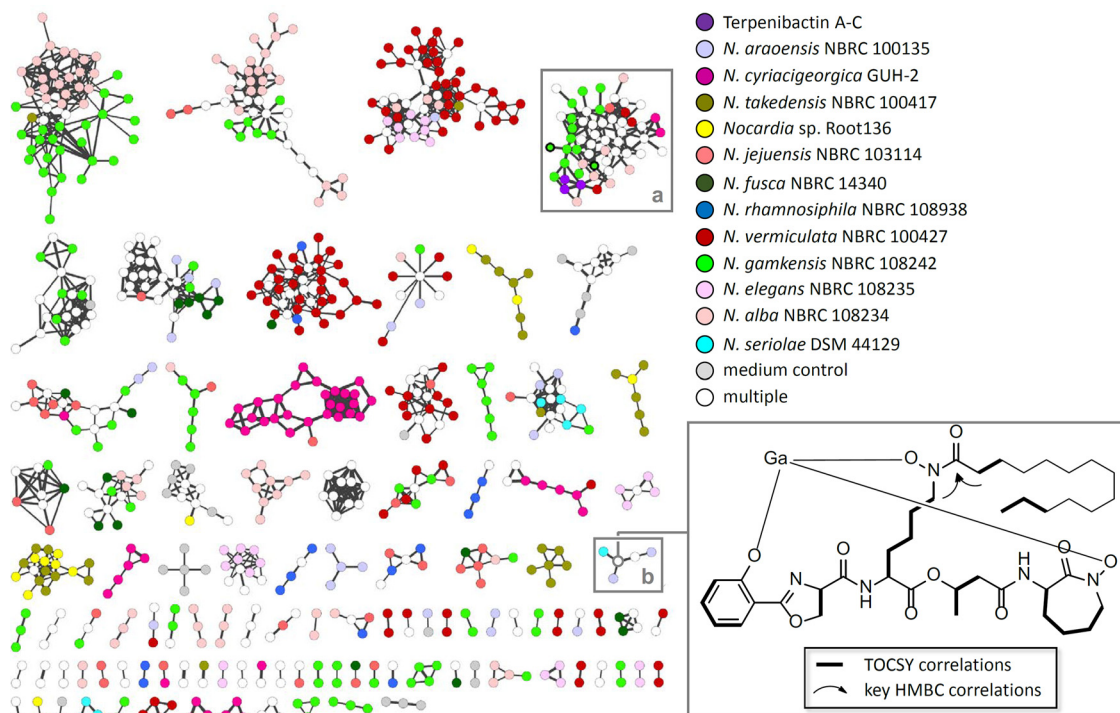


FIG 6 Molecular MS/MS network of culture extracts from selected *Nocardia* strains. The color of the node indicates the source of the ions. Bold node borders indicate putative carboxynocobactins. Molecular families containing putative siderophores a (nocobactin-type) and b (mycobactin-type) are highlighted by gray framed boxes. The structure of the identified mycobactin-type compound from *N. araoensis* NBRC 100135 with an exact mass of 773.4569 Da and m/z 840.3669 $[M-2H+Ga]^+$ is shown in complex with gallium.

<https://doi.org/10.5281/zenodo.3784407>). Additional prominent fragments of m/z 145.1 and m/z 171.08 are found in most MS/MS spectra and likely represent the caprolactam moiety. Using these results as fingerprints, we interpreted the other spectra clustered with the terpenibactins. In consideration of the biosynthetic information deduced from the gene clusters and the already known chemical diversity, we postulated possible structures for each of the produced nocobactin derivatives (Fig. 7; see Fig. S5 to S21 and Tables S6 to S15 at <https://doi.org/10.5281/zenodo.3784407>). The first thing to note is that representatives of all investigated subfamilies seem to synthesize molecules with only minor structural differences to nocobactin NA. The predicted compounds all comprise a phenoxazoline/-oxazole, an acylated *N*-hydroxylysine, a hydroxy fatty acid, and a caprolactam moiety. The integrity of this core structure was independent of the composition of the respective BGCs, e.g., the presence of additional NRPS genes (subfamilies II, VI, VIII, and IX), duplication of the PKS genes (subfamily XI) or absence of the *nbtF* homolog in cluster I (subfamilies I, IV, V, and X). However, the MS/MS characteristics of the different identified nocobactin-type molecules are in accordance with the presence of *nbtH* (*N*-acetyltransferase) or *terH* (*N*-formyltransferase) homologs and PKS C-methyltransferase domains in their respective gene clusters. This supports our bioinformatic predictions that generally pathways from subfamilies I, III, and V produce acetylated nocobactin derivatives and pathways from subfamilies II, IV, and VI to IX produce formylated nocobactin derivatives. Strikingly, colocation of a *sagB*-like dehydrogenase gene with the nocobactin biosynthetic pathway in subfamilies I, III, and VII led to the almost exclusive production of molecules with a predicted phenoxazoline moiety. The absence of such a gene in subfamilies II, IV, VI, and VIII to XI, on the other hand, correlated with the accumulation of phenoxazoline-containing compounds. This strongly indicates that the *SagB*-type dehydrogenase is involved in the oxidation of the oxazoline ring in the biosynthesis of this class of siderophores. Our genomics-to-metabolomics analysis of *Nocardia alba* NBRC 108234 (subfamily I) suggests that this

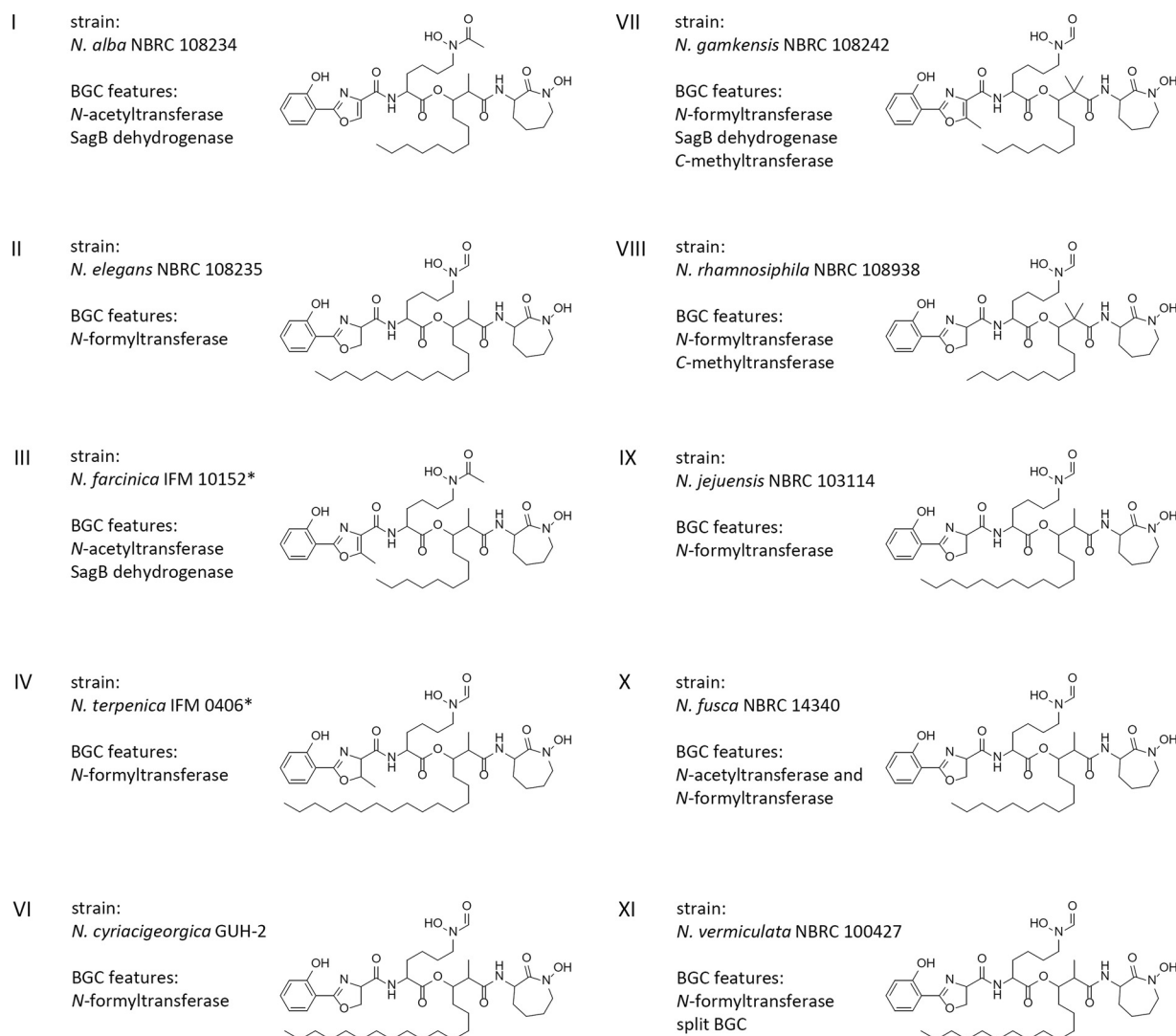


FIG 7 Production of nocobactin-type siderophores by selected *Nocardia* strains from subfamilies I to XI. Representative compounds are depicted. Structures shown in I, II, and VI to XI were predicted from HRMS and MS/MS experiments. An asterisk indicates that the structures shown in III and IV have been adopted from literature (50, 51).

strain synthesizes nocobactin-type siderophores with a nonmethylated oxazole moiety and a *N*-acetyl group (Fig. 7). This closely resembles the structures of the already known nocardimicins A to F. Indeed, exact mass and MS/MS spectra of ions with m/z 796.3043 [M-2H+Ga]⁺, m/z 824.3356 [M-2H+Ga]⁺, and m/z 852.3669 [M-2H+Ga]⁺ perfectly match those of nocardimicins A, B, and D, respectively (see Fig. S18 to S20 at <https://doi.org/10.5281/zenodo.3784407>). In addition, we would postulate that *Nocardia elegans* NBRC 108235 (subfamily II) and *Nocardia jejuensis* NBRC 103114 (subfamily IX) produce *N*-formylated nonmethylated oxazoline-containing molecules with MS characteristics identical to those of nocardimicin G and nocardimicin H (see Fig. S15 and S16 at <https://doi.org/10.5281/zenodo.3784407>). Similar molecules are produced by *N. cyriacigeorgica* GUH-2, the representative of subfamily VI. This subfamily also comprises the three BGCs found in *N. brasiliensis* strains ATCC 700358, IFM 10847, and NBRC 14402, respectively. It is noteworthy that a nocardimicin analog, brasilibactin A, has previously been isolated from this species (48). Closer inspection of the *N. cyriacigeorgica* GUH-2 extract revealed an ion with m/z 868.3966 [M-2H+Ga]⁺ matching the predicted MS properties of brasilibactin A (see Fig. S21 at <https://doi.org/10.5281/zenodo.3784407>). Moreover, gene cluster and MS analysis of *Nocardia gamkensis* NBRC

108242 (subfamily VII) conform to the synthesis of nocobactin derivatives comprising a methyloxazole, an *N*-formyl group, and a central *gem*-dimethylated hydroxy fatty acid moiety. Such a core skeleton is known from formobactin and the amamistatins. Based on high-resolution MS (HRMS) and MS/MS spectra, the detected compound with m/z 810.3199 $[M-2H+Ga]^+$ does indeed correlate with formobactin (see Fig. S12 at <https://doi.org/10.5281/zenodo.3784407>). Overall, we have identified candidate gene clusters for the biosynthesis of formobactin, the nocardimicins, and brasilibactin and provided the first evidence for their function. The pathways for these molecules have not been reported so far. Moreover, the production of these and the other known nocobactin-like siderophores, nocobactin NA and terpenibactin, corresponds to the assignment of their gene clusters to different subfamilies. Notably, the biosynthetic pathways clustered in subfamily VIII appear to produce variants of this structural class that have not been described before. The representative investigated, *Nocardia rhamnosiphila* NBRC 108938, accumulates a molecule likely containing a nonmethylated oxazoline, an *N*-formyl group, and a *gem*-dimethylated hydroxy fatty acid moiety. We are currently pursuing isolation and characterization of this new compound in detail. Another interesting finding are two metabolites from *N. gamkensis* NBRC 108242 which exhibit the characteristic MS fragmentation pattern of nocobactins but have a mass indicating the presence of a carboxyl group. The detection of MS fragments with m/z 458.0477, m/z 412.0432, m/z 171.0688, and m/z 145.0981 for parent ion m/z 840.2982 $[M-2H+Ga]^+$ and fragments m/z 458.0450, m/z 412.0419, m/z 171.0480, and m/z 145.0894 for parent ion m/z 812.2634 $[M-2H+Ga]^+$ implies intact phenyl methyloxazole, *N*-formylated *N*-hydroxylysine, and ϵ -caprolactam moieties. This suggests that it is the central hydroxy fatty acyl unit, which is modified by an additional carboxylic acid group. Such structural variations are known from carboxymycobactins, the extracellular version of mycobactins, which contain α,ω -dicarboxylic acids (53). Analogously, the two molecules detected may represent “carboxynocobactins,” the first examples of similarly oxidized nocobactin-like siderophores. We are currently further exploring this hypothesis.

Next, we wanted to validate our MS- and genomic data-based structure prediction strategy. We were particularly interested to examine molecular variations that are invisible to our MS/MS analysis design and are thus mainly deduced from gene cluster content. One such example is the location of an additional methylene unit (+14 Da) in fragment ions representing the western part of the molecule, e.g., in m/z 838.3512 $[M-2H+Ga]^+$ from *N. alba* NBRC 108234 (see Fig. S23 at <https://doi.org/10.5281/zenodo.3784407>). In theory, this could derive either from the presence of a methyloxazoline or an *N*-acetyl group in the respective compounds. Therefore, we fed $L-[^{13}C_1]$ serine to cultures of *N. alba* NBRC 108234 and analyzed the culture extracts by LC-MS. Detection of parent ion m/z 839.5489 $[M-2H+Ga+1]^+$ and fragments m/z 413.2523 and m/z 441.2829 indicate the incorporation of *L*-serine into the oxazoline ring and, hence, the presence of an *N*-acetyl group (see Fig. S23 and S24 at <https://doi.org/10.5281/zenodo.3784407>). For mycobactin biosynthesis, it has been proposed that the oxazoline ring derives from demethylation of a methyloxazoline moiety (60). However, our data suggest that for nocobactin-type siderophores in *Nocardia* spp., the selected methylation of the oxazoline ring instead depends on the specificity of the respective A domain in the NbtF homolog toward threonine or serine.

Identification of a mycobactin-type molecule produced by *N. araoensis* NBRC 100135. We were not able to identify nocobactin-type molecules from *Nocardia araoensis* NBRC 100135 and *N. takedensis* NBRC 100417, although both have the biosynthetic capacity in the form of pathways belonging to subfamilies III and VII, respectively (see Table S18 at <https://doi.org/10.5281/zenodo.3784407>). This finding made us curious, and we reexamined the culture extracts of these strains to identify putative alternative siderophores. More precisely, we searched the respective MS spectra for molecular ions with a mass shift and isotope pattern that indicated the scavenging of Ga^{3+} . Indeed, we found the ions of putative siderophores in the same mass range as that for the nocobactins, e.g., m/z 824.3353 $[M-2H+Ga]^+$, m/z 840.3669

$[M-2H+Ga]^+$, and m/z 854.3817 $[M-2H+Ga]^+$. Intriguingly, MS/MS experiments produced signal spectra that would instead correlate with mycobactin-type molecules, wherein the long fatty acid moiety is *N*-acylated on the *N*-hydroxylysine instead of being incorporated by NbtB and NbtC in the polyketide extension comprising the ester linkage. To explore this possibility, we isolated a compound with the mass of 773.4569 Da from culture extracts of *N. araoensis* NBRC 100135. Following overnight incubation of this unknown putative mycobactin-like siderophore with $Ga_2(SO_4)_3$ dissolved in methanol, we removed the residual solvent and resuspended it in $CDCl_3$ for a suite of one-dimensional (1D) and 2D nuclear magnetic resonance (NMR) spectroscopic analyses. Structural assignment of the gallium-bound siderophore (Fig. 6; see Table S17 and Fig. S25 to S29 at <https://doi.org/10.5281/zenodo.3784407>) allowed us to identify key spin systems by total correlation spectroscopy (TOCSY): an *ortho*-disubstituted phenol, a serine-derived oxazoline, a linear saturated fatty acid, two *N*-hydroxylysines, and most notably, a β -hydroxybutyrate moiety. The presence of the relatively simple β -hydroxybutyrate spin system instead of the much more complex β -hydroxy fatty acid observed in the nocobactins further validated the prediction that this molecule is indeed a mycobactin-like siderophore. It was determined that the fatty acid was attached to the linear *N*-hydroxylysine side chain via key 1H - ^{13}C heteronuclear multiple bond correlation (HMBC) cross peaks from the same amide carbonyl carbon to both the amino acid ϵ - CH_2 and the fatty acid α - CH_2 , further supporting the MS/MS fragmentation data regarding the location of this moiety. A direct NMR comparison of this novel metabolite to the most structurally similar molecules (mycobactin S or BE-32030 B from another *Nocardia* species [61]) was not performed due to the presence of a chelating Ga^{3+} ion within our analyses, while previous isolates contained the metal-free forms. One beneficial outcome of the addition of gallium is that it minimized rotameric forms of this peptide and easily identified groups adjacent to metal-chelating moieties (the ϵ - CH_2 groups of both *N*-hydroxylysines and the α - CH_2 of the fatty acid) due to the strong diastereotopic shifts from chiral coordination to the metal ion. The nocobactin-type BGCs from *N. araoensis* NBRC 100135 and *N. takedensis* NBRC 100417 are genuine representatives of subfamilies III and VII, respectively. They are almost identical to the pathways of *N. farcinica* IFM 10152 and *N. gamkensis* NBRC 108242, respectively, the other investigated members of those clades. On the basis of these gene clusters, it is difficult to explain why they would support the production of mycobactins instead of nocobactins. We would speculate that for the biosynthesis of mycobactins in *Nocardia* spp., *trans*-encoded enzymes are recruited from other metabolic pathways, e.g., a long-chain fatty acid *N*-acyltransferase.

Conclusion. Our comparative genomic survey of the actinobacterial genus *Nocardia* revealed a highly diverse biosynthetic capacity for natural products which rivals better investigated genera such as *Streptomyces* and *Amycolatopsis*. Most of the GCFs identified by BGC similarity networks could not be correlated with known compounds, making *Nocardia* a promising source for the discovery of new drug leads. Other GCFs are widespread in or even constitutive for this genus, raising interesting questions about their role in bacterial physiology, ecology, and pathogenesis. By using the highly conserved biosynthetic pathway for nocobactin-type siderophores as a proxy, we sought to address the longstanding question of the relationship of BGC homology and chemical diversity. We could show that at least for this case, the BiG-SCAPE gene cluster subfamilies generated by a similarity cutoff of 0.7 encoded structurally distinct nocobactin derivatives. These congeners differed in oxidation state and the selected presence of various methyl groups. By using this approach, we were able to assign for the first time BGC candidates to the nocardimicins, formobactin, and brasilibactin. Moreover, we identified the *Nocardia* strains of subfamily VIII that likely produce so far undescribed nocobactin variants. Our study shows the successful linkage of comparative genomics data to chemical analytics data for refined chemical structure prediction. It exemplifies how multi-omics data can be implemented into drug discovery and aid dereplication efforts. With public repositories such as MassIVE and MetaboLights for

MS/MS data and additional community tools, which provide links to genomic data (e.g., <http://pairedomicsdata.bioinformatics.nl>), this approach will be even more powerful in the future. Our results will thus facilitate future genome mining approaches and open the door for the targeted exploitation of the genus *Nocardia* for natural product discovery.

MATERIALS AND METHODS

Strains and growth conditions. Chemicals, microorganisms, and molecular biological agents were purchased from standard commercial sources. *Nocardia araoensis* NBRC 100135, *Nocardia cyriacigeorgica* GUH-2, *Nocardia takedensis* NBRC 100417, *Nocardia* sp. strain Root136, *Nocardia terpenica* IFM 0406, *Nocardia jejuensis* NBRC 103114, *Nocardia fusca* NBRC 14340, *Nocardia rhamnosiphila* NBRC 108938, *Nocardia vermiculata* NBRC 100427, *Nocardia gamkensis* NBRC 108242, *Nocardia elegans* NBRC 108235, *Nocardia seriolae* DSM 44129, and *Nocardia alba* NBRC 108234 were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ). Strains were grown and maintained either on tryptic soy broth (TSB) (pancreatic digest of casein [17 g/liter], papaic digest of soybean [3 g/liter], sodium chloride [5 g/liter], dipotassium phosphate [2.5 g/liter], dextrose [2.5 g/liter]) purchased from Becton Dickinson or brain heart infusion (BHI), purchased from Sigma-Aldrich.

***Nocardia* genomic data.** All *Nocardia* genome sequences used in this study were downloaded from the National Center for Biotechnology Information (NCBI) database or the Department of Energy (DOE) Joint Genome Institute – Integrated Microbial Genomes & Microbiomes (JGI IMG) database (62, 63). Genomes that consisted of more than 615 contigs were discarded from analysis. The completeness of the genomes were assessed using checkM (see Table S20 at <https://doi.org/10.5281/zenodo.3784407>) (64).

Phylogenetic analysis (autoMLST). A multilocus sequence analysis (MLSA) (65) was performed with autoMLST using the web-based application at <http://automlst.ziemertlab.com/analyze#> with the “denovo mode” and default settings (15). The resulting phylogenetic tree was visualized and modified in Geneious.

***In silico* assessment of biosynthetic potential of *Nocardia* strains.** *In silico* analysis was performed with BLAST, antiSMASH 3.0 and 4.0 versions, and Geneious 9.1.8. BLASTP searches were performed using the NCBI Protein BLAST program against the nonredundant protein sequence database (66).

BIG-SCAPE sequence similarity network. All files were processed using antiSMASH 4.0 including the ClusterFinder border prediction algorithm (13, 25). The ClusterFinder border prediction algorithm was used to automatically trim BGCs where gene cluster borders were possible to predict. To the resulting 3,614 putative BGCs, 1,817 more were added from the MIBiG database as reference data (release 1.4, August 2018; see Fig. S31 at <https://doi.org/10.5281/zenodo.3784407>) (67). The final BGC set was analyzed using BIG-SCAPE version 201804 implemented in Python 2.7.13 (14). The “hybrids” mode, which allows BGCs with mixed annotations to be analyzed together, was enabled, and several cutoffs from 0.05 to 0.95 were tested. The program was run in both global and glocal alignment modes (see supplemental material at <https://doi.org/10.5281/zenodo.3784407>). Obtained sequence similarity matrices were then visualized using Cytoscape 3.6.1 (68).

Production conditions. Production of nocobactin- and mycobactin-like siderophores was achieved by using *Nocardia* mycelium from petri dishes (80 mm by 2 mm) containing BHI agar for preculture inoculation. *Nocardia* strains were cultured at 30°C and 200 rpm for at least 1 week. Five milliliters of preculture was then used to inoculate a main culture and shaken at 30°C at 200 rpm for at least 20 days. As production medium, a minimal medium (glycerol [20 ml/liter], L-asparagine [5 g/liter], KH₂PO₄ [5 g/liter or 50 g/liter] [pH 7.5]) was used as described previously (69). Chelex 100 sodium form, purchased from Sigma-Aldrich, was added, and the medium was filtered through Whatman no. 541 filter papers to remove Chelex. After autoclaving, media were supplemented with Zn²⁺ (0.5 μg/ml), Mn²⁺ (0.1 μg/ml), Mg²⁺ (0.04 mg/ml), and Fe²⁺ (0.05 μg/ml) or Fe²⁺ (2 μg/ml) for iron-deficient (main culture) or iron-sufficient (preculture) growth, respectively. All *Nocardia* strains were cultured in 300-ml precleaned Erlenmeyer flasks with 50 ml medium each at 30°C at 200 rpm. All glassware was precleaned by 2% (wt/vol) alcoholic KOH/8 N HNO₃ treatments according to Ratledge and Winder (70).

Extraction of mycobactin and nocobactin-like siderophores. After at least 2 weeks of inoculation, mycelia were separated from the culture supernatant by centrifugation. Twenty-five milliliters of LC-MS grade methanol was added to the cell pellet. After sonication and centrifugation, the organic phase was isolated and evaporated under reduced pressure. The resulting crude extracts were dissolved in methanol (MeOH). To obtain siderophores bound to gallium, a solution of gallium sulfate (20 mg in 30 ml MeOH) was prepared. Ten milliliters was added to crude extracts and evaporated under reduced pressure.

Isolation and purification of a mycobactin-like siderophore by HPLC. For the isolation of a mycobactin-like siderophore from *Nocardia araoensis* NBRC 100135, an Agilent Technologies ProStar 410 Auto Sampler equipped with a Phenomenex Luna 5-μm C18(2) 100-Å high-performance liquid chromatography (HPLC) column (100 by 21.2 mm) was used with an Agilent Technologies 440-LC fraction collector and a flow rate of 10 ml/min with water plus 0.02% formic acid as solution A and methanol with 0.02% formic acid as solution B (see Table S1 at <https://doi.org/10.5281/zenodo.3784407>). Fractions generated were subjected to an Agilent Technologies 1100 HPLC system equipped with a Phenomenex Luna 5-μm C8(2) column (250 by 10 mm; 100 Å) at a flow rate of 3.5 ml/min (see Table S2 at <https://doi.org/10.5281/zenodo.3784407>). The mycobactin-like siderophore with the mass of 840.367 [M-2H+Ga]⁺ from *N. araoensis* NBRC 100135 eluted after 15.6 min. The peak was collected and dried to yield an orange-to-red oil.

Isolation of gDNA. To isolate genomic DNA (gDNA), *Nocardia* precultures from either BHI or TSB agar plates were used to inoculate 50 ml of liquid TSB or BHI medium and cultivated for at least 2 weeks at 30°C and 200 rpm. Cells were harvested, and genomic DNA was isolated from *Nocardia* with standard protocols for *Streptomyces* by phenol-chloroform extraction. Protocols were followed as described in reference 71 with additional treatment of 40 mg proteinase K and subsequent incubation at 60°C for at least 30 min, but no longer than 1 h, after the addition of sodium dodecyl sulfate (SDS).

16S rRNA gene analysis. To rule out contamination of *Nocardia araoensis* NBRC 100135 production cultures, gDNA was isolated from respective precultures. The 16S rRNA gene was amplified by PCR with the 27F and 1492R universal eubacterial 16S rRNA gene primer set and high-fidelity Q5 polymerase (see Table S16 at <https://doi.org/10.5281/zenodo.3784407>). PCR products were analyzed by gel electrophoresis. The Zero Blunt PCR cloning kit was used to clone PCR fragments, and 20 positive clones were sequenced. A BLASTn search against the NCBI nucleotide database showed “*Nocardia araoensis* NBRC 100135” as the first hit for all 20 clones. In addition, production cultures were examined visually by microscopy and inoculation of selective media. No contamination could be observed.

Mycobactin-like siderophore structure elucidation. NMR solvents were obtained from Sigma-Aldrich and used without further purification.

For NMR analysis, 1.5 mg gallium sulfate was dissolved in 3 ml deuterated methanol and then incubated with the mycobactin-like siderophore overnight. The solvent was removed *in vacuo*, and the Ga-bound species was resuspended in CDCl₃ for NMR analyses. Chemical shift signals (δ) are reported in parts per million and are referenced to the internal chloroform signal ($\delta = 7.26$ ppm for ¹H; $\delta = 77.2$ ppm for ¹³C).

LC-HRMS and MS/MS analysis. High-resolution mass spectra (HRMS) were collected on an Agilent 6530 Accurate-Mass Quadrupole Time-of-flight (QTOF) LC-MS instrument (Agilent Technologies) with an electrospray ionization (ESI) interface. Ten to 15 μ l of each crude extract was injected onto a reversed-phase C₁₈ column [Phenomenex Luna 5- μ m C18(2) column [4.6 \times 150 mm; 100 Å]] on a 1260 Infinity LC-System (Agilent Technologies) using a flow rate of 0.75 ml/min to separate the analytes. MS analysis was performed by ESI in positive ionization mode (50 eV). A gradient from 80% solvent A (water with 0.1% formic acid) to 97% solvent B (acetonitrile [ACN] with 0.06% formic acid) was applied.

Global Natural Products Social Molecular Networking (GNPS) molecular MS/MS network. A molecular network was created to simplify the discovery of new nocobactin and mycobactin derivatives using the online workflow at Global Natural Products Social Molecular Networking (GNPS) (59). High-resolution LC-MS data were converted to .mzxml format using MSConvert version 3.0.18353-2de7e414e. The data were clustered with MS-Cluster with a parent ion mass tolerance of 0.04 Da and a fragment ion mass tolerance of 0.04 Da to create consensus spectra. Networks were created where edges were filtered to have a cosine score above 0.6 and more than three matched peaks. A peak intensity of 50 was set as minimum peak intensity. The resulting molecular network was visualized in Cytoscape 3.6.1 (68). The MS/MS data were deposited in MassIVE (MSV000084771) and linked with the genomic data at the iOMEGA Pairing Omics Data Platform (<http://pairedomicsdata.bioinformatics.nl/projects>).

Stable isotope labeling. For feeding experiments with L-[1-¹³C]serine, labeled substrates were added after 2 days of cultivation in a 5 mM final concentration, respectively. High-resolution mass spectra for feeding experiments as well as nocobactin-like siderophore metabolomics data from *Nocardia cyriacigeorgica* GUH-2 were recorded on an HR-ESI-TOF-MS maXis 4G mass spectrometer (Bruker Daltonik GmbH, Bremen, Germany).

Statistical analysis. BiG-SCAPE annotations were used to compute the BGC class frequency for all the *Nocardia* genomes. The statistical significance for differences in the BGC class frequency between group 1 pathogens and other *Nocardia* genomes was calculated using *t* test in Rstudio version 1.1.383. Boxplots were generated in Rstudio using the package ggpubr version 0.2.5 and ggplot2. BiG-SCAPE presence/absence matrix containing information on gene cluster families (GCFs) across all the *Nocardia* genomes was used to perform principal-component analysis (PCA) using ClustVis web tool.

ACKNOWLEDGMENTS

We thank the DFG Research Training Group (RTG1708) and the NIH (R01-GM085770) for support and the German-American Fulbright Commission for a scholarship to D.M.

We thank Harald Gross (University of Tübingen) for helpful discussions and the kind provision of terpenibactin standards.

REFERENCES

1. Newman DJ, Cragg GM. 2016. Natural products as sources of new drugs from 1981 to 2014. *J Nat Prod* 79:629–661. <https://doi.org/10.1021/acs.jnatprod.5b01055>.
2. Challinor VL, Bode HB. 2015. Bioactive natural products from novel microbial sources. *Ann N Y Acad Sci* 1354:82–97. <https://doi.org/10.1111/nyas.12954>.
3. Doroghazi JR, Metcalf WW. 2013. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genomics* 14:611. <https://doi.org/10.1186/1471-2164-14-611>.
4. Ziemert N, Alanjary M, Weber T. 2016. The evolution of genome mining in microbes – a review. *Nat Prod Rep* 33:988–1005. <https://doi.org/10.1039/c6np00025h>.
5. Blin K, Kim HU, Medema MH, Weber T. 2019. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform* 20:1103–1113. <https://doi.org/10.1093/bib/bbx146>.
6. Wilson JW. 2012. Nocardiosis: updates and clinical overview. *Mayo Clin Proc* 87:403–407. <https://doi.org/10.1016/j.mayocp.2011.11.016>.

7. Beaman BL, Beaman L. 1994. *Nocardia* species: host-parasite relationships. Clin Microbiol Rev 7:213–264. <https://doi.org/10.1128/cmr.7.2.213>.
8. Shigemori H, Komaki H, Yazawa K, Mikami Y, Nemoto A, Tanaka Y, Sasaki T, In Y, Ishida T, Kobayashi J. 1998. Brasilicardin A. A novel tricyclic metabolite with potent immunosuppressive activity from actinomycete *Nocardia brasiliensis*. J Org Chem 63:6900–6904. <https://doi.org/10.1021/jo9807114>.
9. Hashimoto M, Komori TA, Kamiya T. 1976. Letter: Nocardicin A and B, novel monocyclic beta-lactam antibiotics from a *Nocardia* species. J Am Chem Soc 98:3023–3025. <https://doi.org/10.1021/ja00426a063>.
10. Mikami Y. 2007. Biological work on medically important *Nocardia* species. Actinomycetologica 21:46–51. <https://doi.org/10.3209/saj.SAJ210107>.
11. Dhakal D, Rayamajhi V, Mishra R, Sohng JK. 2019. Bioactive molecules from *Nocardia*: diversity, bioactivities and biosynthesis. J Ind Microbiol Biotechnol 46:385–407. <https://doi.org/10.1007/s10295-018-02120-y>.
12. Komaki H, Ichikawa N, Hosoyama A, Takahashi-Nakaguchi A, Matsuzawa T, Suzuki K-I, Fujita N, Gono T. 2014. Genome based analysis of type-I polyketide synthase and nonribosomal peptide synthetase gene clusters in seven strains of five representative *Nocardia* species. BMC Genomics 15:323. <https://doi.org/10.1186/1471-2164-15-323>.
13. Blin K, Wolf T, Chevrete MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de Los Santos ELC, Kim HU, Nave M, Dickschat JS, Mitchell DA, Shelest E, Breitling R, Takano E, Lee SY, Weber T, Medema MH. 2017. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res 45:W36–W41. <https://doi.org/10.1093/nar/gkx319>.
14. Navarro-Muñoz JC, Selem-Mojica N, Mallowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos EL, Yeong M, Cruz-Morales P, Abubucker S, Roeters A, Lokhorst W, Fernandez-Guerra A, Cappellini LTD, Goering AW, Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16:60–68. <https://doi.org/10.1038/s41589-019-0400-9>.
15. Alanjary M, Steinke K, Ziemert N. 2019. AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. Nucleic Acids Res 47:W276–W282. <https://doi.org/10.1093/nar/gkz282>.
16. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132. <https://doi.org/10.1186/s13059-016-0997-x>.
17. Zhang W, Du P, Zheng H, Yu W, Wan L, Chen C. 2014. Whole-genome sequence comparison as a method for improving bacterial species definition. J Gen Appl Microbiol 60:75–78. <https://doi.org/10.2323/jgam.60.75>.
18. Cenicerós A, Dijkhuizen L, Petrusma M, Medema MH. 2017. Genome-based exploration of the specialized metabolic capacities of the genus *Rhodococcus*. BMC Genomics 18:593. <https://doi.org/10.1186/s12864-017-3966-1>.
19. Adamek M, Alanjary M, Sales-Ortells H, Goodfellow M, Bull AT, Winkler A, Wibberg D, Kalinowski J, Ziemert N. 2018. Comparative genomics reveals phylogenetic distribution patterns of secondary metabolites in *Amycolatopsis* species. BMC Genomics 19:426. <https://doi.org/10.1186/s12864-018-4809-4>.
20. Ward A, Allenby NE. 2018. Genome mining for the search and discovery of bioactive compounds: the *Streptomyces* paradigm. FEMS Microbiol Lett 365:240. <https://doi.org/10.1093/femsle/fny240>.
21. Batut B, Knibbe C, Marais G, Daubin V. 2014. Reductive genome evolution at both ends of the bacterial population size spectrum. Nat Rev Microbiol 12:841–850. <https://doi.org/10.1038/nrmicro3331>.
22. Paige EK, Spelman D. 2019. Nocardiosis: 7-year experience at an Australian tertiary hospital. Intern Med J 49:373–379. <https://doi.org/10.1111/imj.14068>.
23. Minero MV, Marín M, Cercenado E, Rabadán PM, Bouza E, Muñoz P. 2009. Nocardiosis at the turn of the century. Medicine (Baltimore) 88:250–261. <https://doi.org/10.1097/MD.0b013e3181afa1c8>.
24. Ott SR, Meier N, Kolditz M, Bauer TT, Rohde G, Presterl E, Schürmann D, Lepper PM, Ringshausen FC, Flick H, Leib SL, Pletz MW, OPINION Study Group. 2019. Pulmonary nocardiosis in Western Europe—clinical evaluation of 43 patients and population-based estimates of hospitalization rates. Int J Infect Dis 81:140–148. <https://doi.org/10.1016/j.ijid.2018.12.010>.
25. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, Birren BW, Takano E, Sali A, Lington RG, Fischbach MA. 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158:412–421. <https://doi.org/10.1016/j.cell.2014.06.034>.
26. Poralla K, Muth G, Hartner T. 2000. Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). FEMS Microbiol Lett 189:93–95. <https://doi.org/10.1111/j.1574-6968.2000.tb09212.x>.
27. Marrakchi H, Laneelle MA, Daffe M. 2014. Mycolic acids: structures, biosynthesis, and beyond. Chem Biol 21:67–85. <https://doi.org/10.1016/j.chembiol.2013.11.011>.
28. Portevin D, De Sousa-D'Auria C, Houssin C, Grimaldi C, Chami M, Daffé M, Guilhot C. 2004. A polyketide synthase catalyzes the last condensation step of mycolic acid biosynthesis in mycobacteria and related organisms. Proc Natl Acad Sci U S A 101:314–319. <https://doi.org/10.1073/pnas.0305439101>.
29. Mikami Y, Yazawa Y, Tanaka Y, Ritzau M, Gräfe U. 1999. Isolation and structure of nocardiolactone, a new dialkyl-substituted β -lactone from pathogenic *Nocardia* strains. Nat Prod Lett 13:277–284. <https://doi.org/10.1080/10575639908048798>.
30. Robinson SL, Terlouw BR, Smith MD, Pidot SJ, Stinear T, Medema MH, Wackett LP. 2019. Global analysis of adenylate-forming enzymes reveals β -lactone biosynthesis pathway in pathogenic *Nocardia*. bioRxiv <https://doi.org/10.1101/856955>.
31. Wolf F, Bauer JS, Bendel TM, Kulik A, Kalinowski J, Gross H, Kaysser L. 2017. Biosynthesis of the β -lactone proteasome inhibitors belactosin and cystargolide. Angew Chem Int Ed Engl 56:6665–6668. <https://doi.org/10.1002/anie.201612076>.
32. Kaysser L. 2019. Built to bind: the biosynthetic strategies for the formation of small-molecule protease inhibitors. Nat Prod Rep 36:1654–1686. <https://doi.org/10.1039/c8np00095f>.
33. McGlinchey RP, Nett M, Eustaquio AS, Asolkar RN, Fenical W, Moore BS. 2008. Engineered biosynthesis of antiprotealide and other unnatural salinosporamide proteasome inhibitors. J Am Chem Soc 130:7822–7823. <https://doi.org/10.1021/ja8029398>.
34. Peters P, Galinski E, Trüper H. 1990. The biosynthesis of ectoine. FEMS Microbiol Lett 71:157–162. <https://doi.org/10.1111/j.1574-6968.1990.tb03815.x>.
35. Bursy J, Kuhlmann AU, Pittelkow M, Hartmann H, Jebbar M, Pierik AJ, Bremer E. 2008. Synthesis and uptake of the compatible solutes ectoine and 5-hydroxyectoine by *Streptomyces coelicolor* A3(2) in response to salt and heat stresses. Appl Environ Microbiol 74:7286–7296. <https://doi.org/10.1128/AEM.00768-08>.
36. Ofer N, Wishkautzan M, Meijler M, Wang Y, Speer A, Niederweis M, Gur E. 2012. Ectoine biosynthesis in *Mycobacterium smegmatis*. Appl Environ Microbiol 78:7483–7486. <https://doi.org/10.1128/AEM.01318-12>.
37. Richter TK, Hughes CC, Moore BS. 2015. Sioxanthin, a novel glycosylated carotenoid, reveals an unusual subclustered biosynthetic pathway. Environ Microbiol 17:2158–2171. <https://doi.org/10.1111/1462-2920.12669>.
38. Guinand M, Vacheron M, Michel G. 1970. Structure des parois cellulaires des nocardia. I-isolement et composition des parois de *Nocardia kirovani*. FEBS Lett 6:37–39. [https://doi.org/10.1016/0014-5793\(70\)80036-9](https://doi.org/10.1016/0014-5793(70)80036-9).
39. Osawa A, Kasahara A, Mastuoka S, Gassel S, Sandmann G, Shindo K. 2011. Isolation of a novel carotenoid, OH-chlorobactene glucoside hexadecanoate, and related rare carotenoids from *Rhodococcus* sp. CIP and their antioxidative activities. Biosci Biotechnol Biochem 75:2142–2147. <https://doi.org/10.1271/bbb.110441>.
40. Valdes-Stauber N, Scherer S. 1994. Isolation and characterization of Linocin M18, a bacteriocin produced by *Brevibacterium linens*. Appl Environ Microbiol 60:3809–3814. <https://doi.org/10.1128/AEM.60.10.3809-3814.1994>.
41. Rosenkrands I, Rasmussen PB, Carnio M, Jacobsen S, Theisen M, Andersen P. 1998. Identification and characterization of a 29-kilodalton protein from *Mycobacterium tuberculosis* culture filtrate recognized by mouse memory effector cells. Infect Immun 66:2728–2735. <https://doi.org/10.1128/IAI.66.6.2728-2735.1998>.
42. McClean S, Healy ME, Collins C, Carberry S, O'Shaughnessy L, Dennehy R, Adams Á, Kennelly H, Corbett JM, Carty F, Cahill LA, Callaghan M, English K, Mahon BP, Doyle S, Shinoy M. 2016. Linocin and OmpW are involved in attachment of the cystic fibrosis-associated pathogen *Burkholderia cepacia* complex to lung epithelial cells and protect mice against infection. Infect Immun 84:1424–1437. <https://doi.org/10.1128/IAI.01248-15>.
43. Sutter M, Boehringer D, Gutmann S, Günther S, Prangishvili D, Loessner MJ, Stetter KO, Weber-Ban E, Ban N. 2008. Structural basis of enzyme encapsulation into a bacterial nanocompartment. Nat Struct Mol Biol 15:939–947. <https://doi.org/10.1038/nsmb.1473>.

44. Ratledge C, Snow GA. 1974. Isolation and structure of nocobactin NA, a lipid-soluble iron-binding compound from *Nocardia asteroides*. *Biochem J* 139:407–413. <https://doi.org/10.1042/bj1390407>.
45. Ratledge C, Patel P. 1976. The isolation, properties and taxonomic relevance of lipid-soluble, iron-binding compounds (the nocobactins) from *Nocardia*. *Microbiology* 93:141–152. <https://doi.org/10.1099/00221287-93-1-141>.
46. Suenaga K, Kokubo S, Shinohara C, Tsuji T, Uemura D. 1999. Structures of amamistatins A and B, novel growth inhibitors of human tumor cell lines from an actinomycete. *Tetrahedron Lett* 40:1945–1948. [https://doi.org/10.1016/S0040-4039\(99\)00050-7](https://doi.org/10.1016/S0040-4039(99)00050-7).
47. Ikeda Y, Nonaka H, Furumai T, Onaka H, Igarashi Y. 2005. Nocardimicins A, B, C, D, E, and F, siderophores with muscarinic M3 receptor inhibiting activity from *Nocardia* sp. TP-A0674. *J Nat Prod* 68:1061–1065. <https://doi.org/10.1021/np050091j>.
48. Tsuda M, Yamakawa M, Oka S, Tanaka Y, Hoshino Y, Mikami Y, Sato A, Fujiwara H, Ohizumi Y, Kobayashi J. 2005. Brasilibactin A, a cytotoxic compound from actinomycete *Nocardia brasiliensis*. *J Nat Prod* 68:462–464. <https://doi.org/10.1021/np0496385>.
49. Murakami Y, Kato S, Nakajima M, Matsuoka M, Kawai H, Shin-Ya K, Seto H. 1996. Formobactin, a novel free radical scavenging and neuronal cell protecting substance from *Nocardia* sp. *J Antibiot (Tokyo)* 49:839–845. <https://doi.org/10.7164/antibiotics.49.839>.
50. Paterson J, Frediansyah A, Männle D, Straetener J, Brötz-Oesterhelt H, Ziemert N, Kaysser L, Gross H. 20 March 2020. New nocobactin derivatives with antimuscarinic activity, terpenibactins A–C, revealed by genome mining of *Nocardia terpenica* IFM 0406. *ChemBioChem* <https://doi.org/10.1002/cbic.202000062>.
51. Hoshino Y, Chiba K, Ishino K, Fukai T, Igarashi Y, Yazawa K, Mikami Y, Ishikawa J. 2011. Identification of nocobactin NA biosynthetic gene clusters in *Nocardia farcinica*. *J Bacteriol* 193:441–448. <https://doi.org/10.1128/JB.00897-10>.
52. Quadri LE, Sello J, Keating TA, Weinreb PH, Walsh CT. 1998. Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem Biol* 5:631–645. [https://doi.org/10.1016/s1074-5521\(98\)90291-5](https://doi.org/10.1016/s1074-5521(98)90291-5).
53. Ratledge C, Ewing N. 1996. The occurrence of carboxymycobactin, the siderophore of pathogenic mycobacteria, as a second extracellular siderophore in *Mycobacterium smegmatis*. *Microbiology* 142:2207–2212. <https://doi.org/10.1099/13500872-142-8-2207>.
54. De Voss JJ, Rutter K, Schroeder BG, Su H, Zhu Y, Barry CE. 2000. The salicylate-derived mycobactin siderophores of *Mycobacterium tuberculosis* are essential for growth in macrophages. *Proc Natl Acad Sci U S A* 97:1252–1257. <https://doi.org/10.1073/pnas.97.3.1252>.
55. Reddy PV, Puri RV, Chauhan P, Kar R, Rohilla A, Khera A, Tyagi AK. 2013. Disruption of mycobactin biosynthesis leads to attenuation of *Mycobacterium tuberculosis* for growth and virulence. *J Infect Dis* 208:1255–1265. <https://doi.org/10.1093/infdis/jit250>.
56. Yokokawa F, Izumi K, Omata J, Shioiri T. 2000. Total synthesis of amamistatin A, an antiproliferative linear peptide from an actinomycete. *Tetrahedron* 56:3027–3034. [https://doi.org/10.1016/S0040-4020\(00\)00170-8](https://doi.org/10.1016/S0040-4020(00)00170-8).
57. Mitchell JM, Shaw JT. 2007. Synthesis and stereochemical assignment of brasilibactin A. *Org Lett* 9:1679–1681. <https://doi.org/10.1021/ol070355o>.
58. Lee SW, Mitchell DA, Markley AL, Hensler ME, Gonzalez D, Wohlrab A, Dorrestein PC, Nizet V, Dixon JE. 2008. Discovery of a widely distributed toxin biosynthetic gene cluster. *Proc Natl Acad Sci U S A* 105:5879–5884. <https://doi.org/10.1073/pnas.0801338105>.
59. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapon CA, Luzzatto-Knaan T, Porto C, Bouslimani A, Melnik AV, Meehan MJ, Liu W-T, Crüsemann M, Boudreau PD, Esquenazi E, Sandoval-Calderón M, Kersten RD, Pace LA, Quinn RA, Duncan KR, Hsu C-C, Floros DJ, Gavilan RG, Kleigrewe K, Northen T, Dutton RJ, Parrot D, Carlson EE, Aigle B, Michelsen CF, Jelsbak L, Sohlenkamp C, Pevzner P, Edlund A, McLean J, Piel J, Murphy BT, Gerwick L, Liaw C-C, Yang Y-L, Humpf H-U, Maansson M, Keyzers RA, Sims AC, Johnson AR, Sidebottom AM, Sedio BE, et al. 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* 34:828–837. <https://doi.org/10.1038/nbt.3597>.
60. McMahon MD, Rush JS, Thomas MG. 2012. Analyses of MbtB, MbtE, and MbtF suggest revisions to the mycobactin biosynthesis pathway in *Mycobacterium tuberculosis*. *J Bacteriol* 194:2809–2818. <https://doi.org/10.1128/JB.00088-12>.
61. Tsukamoto M, Murooka K, Nakajima S, Abe S, Suzuki H, Hirano K, Kondo H, Kojiri K, Suda H. 1997. BE-32030 A, B, C, D and E, new antitumor substances produced by *Nocardia* sp. A32030. *J Antibiot (Tokyo)* 50:815–821. <https://doi.org/10.7164/antibiotics.50.815>.
62. NCBI Resource Coordinators. 2018. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 46:D8–D13. <https://doi.org/10.1093/nar/gkx1095>.
63. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpidis NC. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. *Nucleic Acids Res* 40:D115–D122. <https://doi.org/10.1093/nar/gkr1044>.
64. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. <https://doi.org/10.1101/gr.186072.114>.
65. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A* 95:3140–3145. <https://doi.org/10.1073/pnas.95.6.3140>.
66. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:405–410.
67. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Dürstherus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich E, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kötter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nützmann H-W, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, Yan X, et al. 2015. Minimum information about a biosynthetic gene cluster. *Nat Chem Biol* 11:625–631. <https://doi.org/10.1038/nchembio.1890>.
68. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504. <https://doi.org/10.1101/gr.1239303>.
69. Ratledge C, Hall MJ. 1971. Influence of metal ions on the formation of mycobactin and salicylic acid in *Mycobacterium smegmatis* grown in static culture. *J Bacteriol* 108:314–319. <https://doi.org/10.1128/JB.108.1.314-319.1971>.
70. Ratledge C, Winder FG. 1962. The accumulation of salicylic acid by mycobacteria during growth on an iron-deficient medium. *Biochem J* 84:501–506. <https://doi.org/10.1042/bj0840501>.
71. Hopwood DA, Bibb MJ, Chater KF, Kieser T, Bruton CJ, Kieser HM, Lydiate DJ, Smith CP, Ward JM, Schrepf H. 1985. Genetic manipulation of *Streptomyces*: a laboratory manual, vol 14. The John Innes Foundation, Norwich, United Kingdom.

Comparative Genomics to Metabolomics in the genus *Nocardia*

Supplemental Material

Comparative genomics to metabolomics in the genus *Nocardia*

Daniel Männle^{1,2,3}, Shaun McKinnie⁴, Shrikant S. Mantri^{2,3}, Katharina Steinke^{2,3}, Zeyin Lu^{1,2}, Bradley S. Moore^{5,6}, Nadine Ziemert^{2,3*}, Leonard Kaysser^{1,2*}

¹ Pharmaceutical Biology, Eberhard Karls University Tübingen, 72076 Germany

² German Centre for Infection Research (DZIF), partner site Tübingen

³ Interfaculty Institute for Microbiology and Infection Medicine Tübingen (IMIT), Microbiology and Biotechnology, University of Tübingen, Germany.

⁴ Department of Chemistry and Biochemistry, University of California, Santa Cruz, CA 95064

⁵ Scripps Institution of Oceanography, University of California, San Diego, CA 92093

⁶ Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, CA 92093

Table of Contents

Tables	6
Table S 1: gradient for preparative HPLC.....	6
Table S 2: gradient for analytical HPLC.....	6
Table S 3: genes considered in autoMLST phylogenetic tree.	7
Table S 4: A-domain analysis of NRPS gene encoded in GCF 4.	8
Table S 5: MS properties of terpenibactin A, B and C produced by <i>Nocardia terpenica</i> IFM 0406...	9
Table S 6: MS properties of putative siderophores produced by <i>Nocardia araoensis</i> NBRC 100135.	9
Table S 7: MS properties of putative siderophores produced by <i>Nocardia takedensis</i> NBRC 100417.	10
Table S 8: MS properties of putative siderophores produced by <i>Nocardia jejuensis</i> NBRC 103114.	10
Table S 9: MS properties of putative siderophores produced by <i>Nocardia fusca</i> NBRC 14340.....	11
Table S 10: MS properties of putative siderophore produced by <i>Nocardia rhamnosiphila</i> NBRC 108938.	11
Table S 11: MS properties of putative siderophores produced by <i>Nocardia vermiculata</i> NBRC 100427.	12
Table S 12: MS properties of putative siderophores produced by <i>N. gamkensis</i> NBRC 108242.....	12
Table S 13: MS properties of putative siderophores produced by <i>Nocardia elegans</i> NBRC 108235.	13
Table S 14: MS properties of putative siderophores produced by <i>Nocardia alba</i> NBRC 108234. ...	13
Table S 15: MS properties of putative siderophores produced by <i>Nocardia cyriacigeorgica</i> GUH-2.	14
Table S 16: PCR Primers.....	15
Table S 17: NMR spectroscopic data of a mycobactin-like siderophore from <i>N. araoensis</i> NBRC 100135.	16
Table S 18: Gene cluster table of nocobactin-like BGC from <i>N. araoensis</i> NBRC 100135.....	18
Table S 19: <i>Nocardia</i> strains used for metabolic analysis.	19

Table S 20: checkM analysis on <i>Nocardia</i> genomes used in this study.	20
Table S 21: Analysis of putative virulence factors in <i>Nocardia</i> spp..	21
Table S 22: Presence/absence map of <i>N. farcinica</i> IFM 10152 BGCs in group 1 pathogens.	22
Figures.....	23
Figure S 1: Genome size plotted against the number of predicted BGCs in <i>Nocardia</i> strains.	23
Figure S 2: <i>Nocardia</i> maximum likelihood autoMLST tree.....	24
Figure S 3: Terpenibactins A-C produced by <i>N. terpenica</i> IFM 0406.	25
Figure S 4: MS/MS analysis of terpenibactin A.	26
Figure S 5: MS/MS analysis of a putative mycobactin-like siderophore with 824.3353 [M-2H+Ga] ⁺ produced by <i>Nocardia araoensis</i> NBRC 100135.....	27
Figure S 6: MS/MS analysis of a putative mycobactin-like siderophore with 840.3669 [M-2H+Ga] ⁺ produced by <i>Nocardia araoensis</i> NBRC 100135.....	28
Figure S 7: MS/MS analysis of 854.3817 [M-2H+Ga] ⁺ produced by <i>Nocardia takedensis</i> NBRC 100417.	29
Figure S 8: MS/MS analysis of putative nocardimicin G with 840.3676 [M-2H+Ga] ⁺ produced by <i>Nocardia jejuensis</i> NBRC 103114.....	30
Figure S 9: MS/MS analysis of 812.3442 [M-2H+Ga] ⁺ produced by <i>Nocardia fusca</i> NBRC 14340. ...	31
Figure S 10: MS/MS analysis of 812.3352 [M-2H+Ga] ⁺ produced by <i>Nocardia rhamnosiphila</i> NBRC 108938.	32
Figure S 11: MS/MS analysis of 840.3685 [M-2H+Ga] ⁺ produced by <i>Nocardia vermiculata</i> NBRC 100427.	33
Figure S 12: MS/MS analysis of putative formobactin with 810.3192 [M-2H+Ga] ⁺ produced by <i>Nocardia gamkensis</i> NBRC 108242.....	34
Figure S 13: MS/MS analysis of putative carboxynocobactin with 812.2634 [M-2H+Ga] ⁺ produced by <i>Nocardia gamkensis</i> NBRC 108242.	35
Figure S 14: MS/MS analysis of putative carboxynocobactin with 840.2954 [M-2H+Ga] ⁺ produced by <i>Nocardia gamkensis</i> NBRC 108242.	36

Figure S 15: MS/MS analysis of putative nocardimicin G with 840.3683 [M-2H+Ga] ⁺ produced by <i>Nocardia elegans</i> NBRC 108235.....	37
Figure S 16: MS/MS analysis of putative nocardimicin H with 868.3985 [M-2H+Ga] ⁺ produced by <i>Nocardia elegans</i> NBRC 108235.....	38
Figure S 17 MS/MS analysis of 810.3227 [M-2H+Ga] ⁺ produced by <i>Nocardia alba</i> NBRC 108234..	39
Figure S 18: MS/MS analysis of putative nocardimicin A with 796.3046 [M-2H+Ga] ⁺ produced by <i>Nocardia alba</i> NBRC 108234.....	40
Figure S 19: MS/MS analysis of putative nocardimicin B with 824.3355 [M-2H+Ga] ⁺ produced by <i>Nocardia alba</i> NBRC 108234.....	41
Figure S 20: MS/MS analysis of putative nocardimicin D with 852.3668 [M-2H+Ga] ⁺ produced by <i>Nocardia alba</i> NBRC 108234.....	42
Figure S 21: MS/MS analysis of putative brasilibactin A with 868.3989 [M-2H+Ga] ⁺ produced by <i>Nocardia cyriacigeorgica</i> GUH-2.	43
Figure S 22: BiG-SCAPE cluster-type sequence similarity network of all <i>nocobactin-type</i> BGCs from <i>Nocardia</i> strains found with antiSMASH 4.0.....	44
Figure S 23: Isotopic peak analysis of unlabeled and ¹³ C-labeled nocobactin-like siderophore.	45
Figure S 24: MS/MS analysis of a ¹³ C-labeled/ unlabeled nocobactin-like siderophore from <i>Nocardia alba</i> NBRC 108234.....	46
Figure S 25: TOCSY correlations and key HMBC correlations of a mycobactin-type siderophore with m/z 840.3669 [M+Ga-2H] ⁺	47
Figure S 26: ¹ H NMR spectrum of a gallium-bound mycobactin-like siderophore with m/z 840.3669 [M+Ga-2H] ⁺ from <i>Nocardia araoensis</i> NBRC 100135 in CDCl ₃ (600 MHz).	47
Figure S 27: ¹ H- ¹³ C multiplicity edited HSQC NMR spectrum of gallium-bound mycobactin-like siderophore from <i>Nocardia araoensis</i> NBRC 100135 in CDCl ₃ (600 MHz) with m/z 840.3669 [M+Ga-2H] ⁺	48
Figure S 28: ¹ H- ¹ H TOCSY NMR spectrum of gallium-bound mycobactin-like siderophore from <i>Nocardia araoensis</i> NBRC 100135 in CDCl ₃ (600 MHz) with m/z 840.3669 [M+Ga-2H] ⁺	49
Figure S 29: ¹ H- ¹³ C HMBC NMR spectrum of gallium-bound mycobactin-like siderophore from <i>Nocardia araoensis</i> NBRC 100135 in CDCl ₃ (600 MHz) with m/z 840.3669 [M+Ga-2H] ⁺	50
Figure S 30: <i>Nocardia</i> autoMLST tree. Colors indicate <i>Nocardia</i> spp. source of isolation. Information regarding isolation source was taken from respective published genomes.....	51

Figure S 31: BiG-SCAPE cluster type sequence similarity network of all BGCs from Nocardia strains found with antiSMASH 4.0 including the MIBiG database.....	52
Figure S 32: Number of contigs plotted against the number of predicted BGCs in Nocardia strains.....	53
Figure S 33: Boxplot of BGC class frequency distribution across Nocardia genome groups.	54
Figure S 34: PCA-analysis on GCF presence/absence.....	55
References.....	56

Tables

Table S 1: gradient for preparative HPLC.

time [min]	solvent A [%]	solvent B [%]
0	75	25
3	75	25
6	50	50
11	1	99
20	1	99
25	75	25
30	75	25

Table S 2: gradient for analytical HPLC.

time [min]	solvent A [%]	solvent B [%]
0	75	25
3	75	25
6	50	50
11	1	99
20	1	99
22	75	25
25	75	25

Table S 3: genes considered in autoMLST phylogenetic tree.

protein family	gene	associated function	associated enzyme
TIGR01798	cit_synth_I	Energy metabolism	citrate (Si)-synthase
TIGR00043	TIGR00043	Protein synthesis	rRNA maturation RNase YbeY
TIGR00970	leuA_yeast	Amino acid biosynthesis	2-isopropylmalate synthase
TIGR01855	IMP_synth_hisH	Amino acid biosynthesis	imidazole glycerol phosphate synthase, glutamine amidotransferase subunit
TIGR00115	tig	Protein fate	trigger factor
TIGR00138	rsmG_gidB	Protein synthesis	16S rRNA (guanine(527)-N(7))-methyltransferase RsmG
TIGR01978	sufC	Biosynthesis of cofactors, prosthetic groups, and carriers	FeS assembly ATPase SufC
TIGR02673	FtsE	Cellular processes	cell division ATP-binding protein FtsE
TIGR00338	serB	Amino acid biosynthesis	phosphoserine phosphatase SerB
TIGR00154	ispE	Biosynthesis of cofactors, prosthetic groups, and carriers	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase
TIGR00031	UDP-GALP_mutase	Cell envelope	UDP-galactopyranose mutase
TIGR00922	nusG	Transcription	transcription termination/antitermination factor NusG
TIGR02091	glgC	Energy metabolism	glucose-1-phosphate adenyltransferase
TIGR03723	T6A_TsaD_YgjD	Protein synthesis	tRNA threonylcarbamoyl adenosine modification protein TsaD
TIGR00158	L9	Protein synthesis	ribosomal protein bl9
TIGR01083	nth	DNA metabolism	endonuclease III
TIGR00090	rsfS_ojap_ybeB	Protein synthesis	ribosome silencing factor
TIGR00091	TIGR00091	Protein synthesis	tRNA (guanine-N(7))-methyltransferase
TIGR02692	tRNA_CCA_actino	Protein synthesis	CCA tRNA nucleotidyltransferase
TIGR01072	murA	Cell envelope	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
TIGR00190	thiC	Biosynthesis of cofactors, prosthetic groups, and carriers	phosphomethylpyrimidine synthase
TIGR00096	TIGR00096	Protein synthesis	16S rRNA (cytidine(1402)-2'-O)-methyltransferase
TIGR00218	manA	Energy metabolism	mannose-6-phosphate isomerase, class I
TIGR01737	FGAM_synth_I	Purines, pyrimidines, nucleosides, and nucleotides	phosphoribosylformylglycinamide synthase I
TIGR01994	SUF_scaf_2	Biosynthesis of cofactors, prosthetic groups, and carriers	SUF system FeS assembly protein, NifU family
TIGR03800	PLP_synth_Pdx2	Biosynthesis of cofactors, prosthetic groups, and carriers	pyridoxal 5'-phosphate synthase, glutaminase subunit Pdx2
TIGR00963	secA	Protein fate	preprotein translocase, SecA subunit
TIGR00964	secE_bact	Protein fate	preprotein translocase, SecE subunit
Ribosomal_L10	PF00466.16	Unclassified	Ribosomal protein L10
TIGR01280	xseB	DNA metabolism	exodeoxyribonuclease VII, small subunit
TIGR00168	infC	Protein synthesis	translation initiation factor IF-3
TIGR00731	bl25_bact_ctc	Protein synthesis	ribosomal protein bl25, Ctc-form
TIGR02727	MTHFS_bact	Central intermediary metabolism	5-formyltetrahydrofolate cyclo-ligase
TIGR01966	RNasePH	Transcription	ribonuclease PH
TIGR00382	clpX	Protein fate	ATP-dependent Clp protease, ATP-binding subunit ClpX
TIGR01357	aroB	Amino acid biosynthesis	3-dehydroquinate synthase
TIGR02729	Obg_CgtA	Protein synthesis	Obg family GTPase CgtA
TIGR01683	thiS	Biosynthesis of cofactors, prosthetic groups, and carriers	thiamine biosynthesis protein ThiS
TIGR00020	prfB	Protein synthesis	peptide chain release factor 2
TIGR00751	menA	Biosynthesis of cofactors, prosthetic groups, and carriers	1,4-dihydroxy-2-naphthoate octaprenyltransferase
TIGR03534	RF_mod_PrmC	Protein fate	protein-(glutamine-N5) methyltransferase, release factor-specific
TIGR01393	lepA	Unknown function	elongation factor 4
TIGR00564	trpE_most	Amino acid biosynthesis	anthranilate synthase component I
TIGR00759	aceE	Unclassified	pyruvate dehydrogenase (acetyl-transferring), homodimeric type
TIGR00302	TIGR00302	Purines, pyrimidines, nucleosides, and nucleotides	phosphoribosylformylglycinamide synthase, purS protein
TIGR00330	glpX	Energy metabolism	fructose-1,6-bisphosphatase, class II
TIGR00237	xseA	DNA metabolism	exodeoxyribonuclease VII, large subunit
TIGR00613	reco	DNA metabolism	DNA repair protein RecO
TIGR01498	folK	Biosynthesis of cofactors, prosthetic groups, and carriers	2-amino-4-hydroxy-6-hydroxymethylidihydropteridine diphosphokinase
TIGR00713	hemL	Biosynthesis of cofactors, prosthetic groups, and carriers	glutamate-1-semialdehyde-2,1-aminomutase
TIGR00525	folB	Biosynthesis of cofactors, prosthetic groups, and carriers	dihydroneopterin aldolase
TIGR00615	recR	DNA metabolism	recombination protein RecR
TIGR00362	DnaA	DNA metabolism	chromosomal replication initiator protein DnaA
TIGR01980	sufB	Biosynthesis of cofactors, prosthetic groups, and carriers	FeS assembly protein SufB
TIGR01051	topA_bact	DNA metabolism	DNA topoisomerase I
TIGR00082	rbfA	Transcription	ribosome-binding factor A
TIGR01039	atpD	Energy metabolism	ATP synthase F1, beta subunit
TIGR00739	yajC	Protein fate	preprotein translocase, YajC subunit
TIGR03284	thym_sym	Purines, pyrimidines, nucleosides, and nucleotides	thymidylate synthase
TIGR01137	cysta_beta	Amino acid biosynthesis	cystathionine beta-synthase
TIGR00103	DNA_YbaB_EbFC	Unknown function	DNA-binding protein, YbaB/EbFC family
TIGR01032	rplT_bact	Protein synthesis	ribosomal protein bl20
TIGR01296	asd_B	Amino acid biosynthesis	aspartate-semialdehyde dehydrogenase

Table S 4: A-domain analysis of NRPS gene encoded in GCF 4.

GCF 4	A1	AA	A2	AA	A3	AA	A4	AA
<i>N. farcinica</i>	DVWHFSXX	S	DALFIGMV	-	DALFMGAI	F/W	DALFLGVI	Hpg
<i>N. brasiliensis</i>	DVWHFSLV	S	DALFIGMV	-	DALFMGGI	F/W	DALALGVI	Hpg
<i>N. abscessus</i>	DVWHFSLV	S	DALFIGMV	-	DALFMGVV	F/W	?	?
<i>N. otitidiscaviarum</i>	DVWHFSLV	S	DALFIGMV	-	DAFFMGGV	L/I/V	DALFLGVI	Hpg
<i>N. kruczakiae</i>	DVWHFSLV	S	DALFIGMV	-	DALFMGGI	F/W	DALFLGVI	Hpg
<i>N. asteroides</i>	DVWHFSLV	S	DALFIGMV	-	DVLFMGAV	-	DGLFLGVI	L
<i>N. brevicatena</i>	DVWHFSLV	S	DALFMGMV	-	DALFMGVI	F/W	DALFLGAI	F/W
<i>N. nova</i>	DVWHFSLV	S	DALFIGMV	-	DALFMGGI	F/W	DALFLGVI	Hpg
<i>N. cyriacigeorgica</i>	DVWHFSLV	S	DALFMGMV	-	DALFMGVV	F/W	DALALGVI	Hpg
<i>N. transvalensis</i>	DVWHFSLV	S	DALFIGMV	-	DAFFVGGV	V	DALFLGVI	Hpg
<i>N. veterana</i>	DVWHFSLV	S	DALFIGMV	-	DALFMGGI	F/W	DALFLGVI	Hpg

Table S 5: MS properties of terpenibactin A, B and C produced by *Nocardia terpenica* IFM 0406.

name	exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
Terpenibactin A	815.50444	882.41383	882.4135	0.37397
Terpenibactin B	841.52009	908.42948	908.4282	1.40903
Terpenibactin C	843.53574	910.44513	910.4442	1.02148

Table S 6: MS properties of putative siderophores produced by *Nocardia araoensis* NBRC 100135.

exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
745.4262	812.33558	812.3379	2.85597
757.4262	824.33558	824.3353	0.33967
773.4575	840.36688	840.3669	0.02379
771.4418	838.35123	838.3508	0.51291
785.4575	852.36688	852.3668	0.09386

Table S 7: MS properties of putative siderophores produced by *Nocardia takedensis* NBRC 100417.

exact mass [M] calculated	exact mass [M- 2H+Ga] ⁺ calculated	exact mass [M- 2H+Ga] ⁺ observed	ppm error
771.44184	838.35123	838.3511	0.15507
773.45749	840.36688	840.3657	1.40415
787.47314	854.38253	854.3817	0.97146
801.48879	868.39818	868.3977	0.55274

Table S 8: MS properties of putative siderophores produced by *Nocardia jejuensis* NBRC 103114.

name	exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
	743.41054	810.32047	810.3201	0.4566
	745.42619	812.33612	812.336	0.14772
	771.44184	838.35177	838.3499	2.2306
Nocardimicin G	773.45749	840.36742	840.3676	0.21419

Table S 9: MS properties of putative siderophores produced by *Nocardia fusca* NBRC 14340.

exact mass [M] calculated	exact mass [M- 2H+Ga] ⁺ calculated	exact mass [M- 2H+Ga] ⁺ observed	ppm error
731.41054	798.31993	798.3198	0.16284
745.42619	812.3356	812.3386	3.69304
757.42619	824.33558	824.3362	0.75212
771.44184	838.35123	838.3539	3.18481
801.48879	868.3982	868.3987	0.57577
815.50444	882.4138	882.4138	0.79327

Table S 10: MS properties of putative siderophore produced by *Nocardia rhamnosiphila* NBRC 108938.

exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
745.42619	812.33558	812.3352	0.46778

Table S 11: MS properties of putative siderophores produced by *Nocardia vermiculata* NBRC 100427.

exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
745.42619	812.33558	812.335	0.71399
771.44184	838.35123	838.3535	2.7077
773.45749	840.36688	840.3685	1.92772
801.48879	868.39818	868.3961	2.39521

Table S 12: MS properties of putative siderophores produced by *N. gamkensis* NBRC 108242.

name	exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
	715.37924	782.28863	782.2887	0.08948
	729.39489	796.30428	796.3024	2.36090
	741.39489	808.30428	808.304	0.34640
Formobactin	743.41054	810.31993	810.3192	0.77747
	745.35342	812.26280	812.2634	0.73868
	757.42619	824.33558	824.3355	0.09705
	773.38472	840.29410	840.2954	1.54710

Table S 13: MS properties of putative siderophores produced by *Nocardia elegans* NBRC 108235.

name	exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
	731.41054	798.31993	798.3231	3.97082
	759.44184	826.35123	826.3531	2.26296
Nocardimicin G	773.45749	840.36688	840.3683	1.68974
Nocardimicin H	801.48879	868.39818	868.3985	0.36850

Table S 14: MS properties of putative siderophores produced by *Nocardia alba* NBRC 108234.

name	exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
Nocardimicin A	729.39489	796.30428	796.3046	0.27628
	743.41054	810.31993	810.3227	2.75200
	755.41054	822.31993	822.3221	1.98220
Nocardimicin B	757.42619	824.33558	824.3355	0.75212
	771.44184	838.35123	838.3512	0.03578
Nocardimicin D	785.45749	852.3669	852.3668	0.09386
	813.48879	880.39818	880.3982	0.59064
	827.50444	894.41383	894.4131	1.41993

Table S 15: MS properties of putative siderophores produced by *Nocardia cyriacigeorgica* GUH-2.

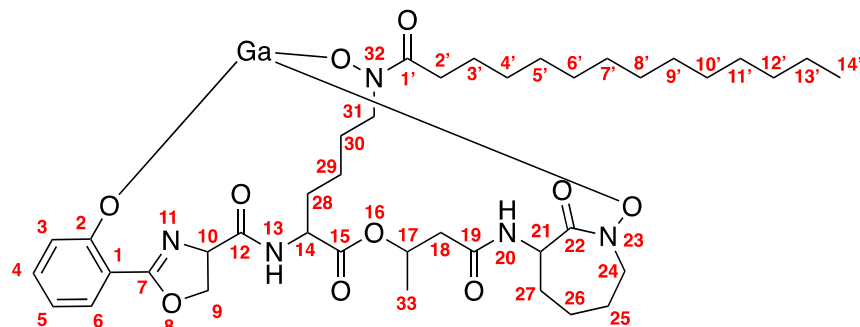
name	exact mass [M] calculated	exact mass [M-2H+Ga] ⁺ calculated	exact mass [M-2H+Ga] ⁺ observed	ppm error
	717.39489	784.30428	784.3044	0.15300
	743.41054	810.31993	810.3187	1.51791
	745.42619	812.33558	812.335	0.71399
	759.44184	826.35123	826.3516	0.44775
	773.45749	840.36688	840.3669	0.02380
	799.47314	866.38253	866.3818	0.84258
Brasilibactin A	801.48879	868.39818	868.3989	0.82911

Table S 16: PCR Primers.

Primer name	Primer sequence 5' → 3'
27F	AGAGTTTGATCMTGGCTCAG
1492R	GGTTACCTTGTTACGACTT

Table S 17: NMR spectroscopic data of a mycobactin-like siderophore from *N. araoensis* NBRC 100135.

Mycobactin-like siderophore incubated with gallium sulphate in CD₃OD, concentrated *in vacuo*, then NMR data recorded in CDCl₃. Numbering scheme adapted from Hoshino et al. 2011.¹



<u>Assignment</u>	<u>¹H-NMR</u>	<u>¹³C-NMR</u>	<u>Number</u>
phenol			
H-3	6.59	115.4	3
H-4	7.32	135.9	4
H-5	6.92	122.8	5
H-6	7.60	129.6	6
oxazoline			
H-9	4.84, 4.61	70.7	9
H-10	4.62	66.8	10
Lys1 (linear)			
NH	7.95	-	13
α	4.69	50.3	14
β	2.13	30.3	28
γ	1.32	29.8	29
δ	2.00, 1.87	27.3	30
ε	4.04, 3.84	53.4	31
3-hydroxybutanoate			
α	2.82, 2.45	41.8	18
β	5.50	67.6	17
γ	1.38	18.4	33
Lys2 (cyclic)			
NH	9.75	-	20
α	4.32	53.7	21
β	2.12, 2.09	26.6	27
γ	1.87, 1.55	27.1	26
δ	1.82, 1.57	18.3	25
ε	3.96, 3.74	48.4	24
Fatty acid			
C=O	-	165.1	1'
H-2'	2.49, 2.33	29.5	2'

H-3'	1.78	25.7	3'
H-4' to H-11'	1.29 (multiple)	29.8 (multiple)	4'-11'
H-12'	1.27	32.1	12'
H-13'	1.30	22.8	13'
H-14'	0.89	14.1	14'

Table S 18: Gene cluster table of nocobactin-like BGC from *N. araoensis* NBRC 100135.

gene	AA	protein homolog	accession number	ID	predicted function
<i>nbtS</i>	436	salicylate synthase [<i>Nocardia farcinica</i> IFM 10152]	BAD55464.1	87.7	salicylate synthase
<i>nbtT</i>	538	putative salicylate-AMP ligase [<i>Nocardia farcinica</i> IFM 10152]	BAD55465.1	84.56	salicylate-AMP ligase
<i>nbtG</i>	428	putative lysine-N-oxygenase [<i>Nocardia farcinica</i> IFM 10152]	BAD55606.1	79.43	lysine-N-oxygenase
<i>nbtH</i>	220	putative N6-hydroxylysine acetyltransferase [<i>Nocardia farcinica</i> IFM 10152]	BAD55607.1	58.41	N6-hydroxylysine acetyltransferase
<i>nbtA</i>	257	thioesterase [<i>Nocardia farcinica</i> IFM 10152]	BAD55608.1	62.55	thioesterase
<i>nbtB</i>	442	putative polyketide synthase [<i>Nocardia farcinica</i> IFM 10152]	BAD55609.1	72.54	polyketide synthase
<i>nbtC</i>	1011	putative polyketide synthase [<i>Nocardia farcinica</i> IFM 10152]	BAD55610.1	55.11	polyketide synthase
<i>nbtD</i>	1508	non-ribosomal peptide synthetase [<i>Nocardia farcinica</i> IFM 10152]	BAD55612.1	58.6	non-ribosomal peptide synthetase
<i>nbtE</i>	1671	putative non-ribosomal peptide synthetase [<i>Nocardia farcinica</i> IFM 10152]	BAD55611.1	57.22	non-ribosomal peptide synthetase
<i>nbtX</i>	360	putative membrane protein [<i>Nocardia farcinica</i> IFM 10152]	BAD60682.1	35.82	Unknown/ hypothetical protein
<i>nbtF</i>	958	putative non-ribosomal peptide synthetase [<i>Nocardia farcinica</i> IFM 10152]	BAD55613.1	81.31	non-ribosomal peptide synthetase

*Abbreviations

AA protein length in amino acid

ID amino acid identity [%]

Sim amino acid similarity

N. *Nocardia*

Table S 19: *Nocardia* strains used for metabolic analysis.

genome	GenBank accession number	Source of Isolation	Biosafety level
<i>Nocardia araoensis</i> NBRC 100135	BAFR00000000	Human	1
<i>Nocardia cyriaciageorgica</i> GUH-2	NC_016887	human	2
<i>Nocardia takedensis</i> NBRC 100417	NZ_BAGG00000000	Human	1
<i>Nocardia</i> sp. Root136	NZ_LMFE00000000	Soil	1
<i>Nocardia jejuensis</i> NBRC 103114	NZ_BDBU00000000	Human	1
<i>Nocardia fusca</i> NBRC 14340	NZ_BDCU00000000	Freshwater	1
<i>Nocardia rhamnosiphila</i> NBRC 108938	NZ_BDCL00000000	Animal	1
<i>Nocardia vermiculata</i> NBRC 100427	NZ_BDCA00000000	Soil	1
<i>Nocardia gamkensis</i> NBRC 108242	NZ_BDBM00000000	Soil	1
<i>Nocardia elegans</i> NBRC 108235	NZ_BDBF00000000	Human	1
<i>Nocardia alba</i> NBRC 108234	NZ_BDAX00000000	plant	1
<i>Nocardia seriolae</i> DSM 44129	n.a.	animal	1

Table S 20: checkM analysis on *Nocardia* genomes used in this study.

bin id	marker lineage	# markers	# marker sets	completeness	contamination	strain heterogeneity	number of contigs
<i>Nocardia abscessus</i> NBRC_100374	o Actinomycetales (UID1815)	574	266	99.37	1.69	0	274
<i>Nocardia acidivorans</i> NBRC_108247	o Actinomycetales (UID1815)	567	264	99.56	2.43	7.14	161
<i>Nocardia africana</i> NBRC_100379	o Actinomycetales (UID1815)	571	265	99.37	1.32	15.38	93
<i>Nocardia alba</i> NBRC_108234	o Actinomycetales (UID1815)	574	266	98.96	2.07	0	62
<i>Nocardia altamirensis</i> NBRC_108246	o Actinomycetales (UID1815)	574	266	99.75	4.19	7.14	167
<i>Nocardia amamiensis</i> NBRC_102102	o Actinomycetales (UID1815)	574	266	99.09	6.31	2.7	487
<i>Nocardia amikacinintolerans</i> DSM_45535	o Actinomycetales (UID1815)	574	266	99.75	0.89	0	33
<i>Nocardia amikacinintolerans</i> DSM_45536	o Actinomycetales (UID1815)	574	266	99.75	1.12	0	28
<i>Nocardia amikacinintolerans</i> DSM_45537	o Actinomycetales (UID1815)	574	266	99.37	0.93	0	95
<i>Nocardia amikacinintolerans</i> DSM_45538	o Actinomycetales (UID1815)	574	266	99.75	0.93	0	30
<i>Nocardia amikacinintolerans</i> NBRC_108937	o Actinomycetales (UID1815)	574	266	99.56	1.03	0	108
<i>Nocardia anaemiae</i> NBRC_100462	o Actinomycetales (UID1815)	574	266	99.75	2.51	18.18	134
<i>Nocardia aobensis</i> NBRC_100429	o Actinomycetales (UID1815)	571	265	98.81	1.04	15.38	291
<i>Nocardia araoensis</i> NBRC_100135	o Actinomycetales (UID1815)	574	266	99.44	2.16	11.76	352
<i>Nocardia arthritidis</i> NBRC_100137	o Actinomycetales (UID1815)	574	266	99.62	1.1	0	113
<i>Nocardia asiatica</i> NBRC_100129	o Actinomycetales (UID1815)	574	266	99.75	4.35	10.34	475
<i>Nocardia asteroides</i> DSM_43373	o Actinomycetales (UID1815)	572	265	99.75	1.74	9.09	25
<i>Nocardia asteroides</i> NBRC_15531	o Actinomycetales (UID1815)	572	265	99.37	2.11	16.67	39
<i>Nocardia beijingensis</i> NBRC_16342	o Actinomycetales (UID1815)	574	266	99.62	1.95	0	113
<i>Nocardia brasiliensis</i> ATCC_700358	o Actinomycetales (UID1815)	574	266	99.37	4.84	10	1
<i>Nocardia brasiliensis</i> IFM_10947	o Actinomycetales (UID1815)	574	266	99.37	4.84	10	223
<i>Nocardia brasiliensis</i> NBRC_14402	o Actinomycetales (UID1815)	574	266	99.75	3.66	8.82	128
<i>Nocardia brevicatena</i> NBRC_12119	o Actinomycetales (UID1815)	574	266	99.25	2.71	18.18	243
<i>Nocardia calshijiensis</i> NBRC_108228	o Actinomycetales (UID1815)	572	265	98.43	0.86	0	60
<i>Nocardia carnea</i> NBRC_14403	o Actinomycetales (UID1815)	574	266	99.32	1.48	0	126
<i>Nocardia cerraensis</i> NBRC_101014	o Actinomycetales (UID1815)	571	265	98.85	1.17	16.67	388
<i>Nocardia concava</i> NBRC_100430	o Actinomycetales (UID1815)	567	264	99.76	4.41	0	206
<i>Nocardia couleae</i> NBRC_108252	o Actinomycetales (UID1815)	572	265	98.81	2.13	0	40
<i>Nocardia crassostreae</i> strain NBRC_100342	o Actinomycetales (UID1815)	572	265	99.51	2.6	9.52	57
<i>Nocardia cummideiensis</i> NBRC_100378	o Actinomycetales (UID1815)	574	266	99.75	1.65	0	60
<i>Nocardia cyriacigeorgica</i> GH-2	o Actinomycetales (UID1815)	574	266	99.37	0.39	0	1
<i>Nocardia cyriacigeorgica</i> NBRC_100375	o Actinomycetales (UID1815)	574	266	98.25	1	11.11	328
<i>Nocardia donostiensis</i> strain X1655	o Actinomycetales (UID1815)	574	266	99.25	0.56	0	49
<i>Nocardia elegans</i> NBRC_108235	o Actinomycetales (UID1815)	571	265	99.2	1.63	0	117
<i>Nocardia exalbida</i> NBRC_100660	o Actinomycetales (UID1815)	574	266	99.69	1.11	8.33	165
<i>Nocardia farcinica</i> DSM_43257	o Actinomycetales (UID1815)	574	266	99.87	0.9	25	43
<i>Nocardia farcinica</i> IFM_10152	o Actinomycetales (UID1815)	574	266	99.87	0.3	28.57	3
<i>Nocardia farcinica</i> NCTC_11134	o Actinomycetales (UID1815)	574	266	99.75	1.24	20	5
<i>Nocardia flavorosea</i> NBRC_108225	o Actinomycetales (UID1815)	574	266	99.32	0.5	0	70
<i>Nocardia fluminea</i> DSM_44489	o Actinomycetales (UID1815)	574	266	99.37	2.49	12.5	4
<i>Nocardia fusca</i> NBRC_14340	o Actinomycetales (UID1815)	574	266	99.29	3.06	0	105
<i>Nocardia gamkensis</i> NBRC_108242	o Actinomycetales (UID1815)	574	266	99.62	1.39	9.09	120
<i>Nocardia grenadensis</i> NBRC_108939	o Actinomycetales (UID1815)	574	266	99.32	2.25	52	71
<i>Nocardia harenae</i> NBRC_108248	o Actinomycetales (UID1815)	572	265	99.15	1.07	0	25
<i>Nocardia higoensis</i> NBRC_100133	o Actinomycetales (UID1815)	574	266	99.69	1.03	0	185
<i>Nocardia ignorata</i> NBRC_108230	o Actinomycetales (UID1815)	572	265	98.43	2.22	0	69
<i>Nocardia inohanensis</i> NBRC_100128	o Actinomycetales (UID1815)	567	264	99.46	2.1	0	26
<i>Nocardia jejuensis</i> NBRC_103114	o Actinomycetales (UID1815)	567	264	99.75	1.55	0	89
<i>Nocardia jiangxiensis</i> NBRC_101359	o Actinomycetales (UID1815)	567	264	99.94	2.95	5.26	163
<i>Nocardia jinanensis</i> NBRC_108249	o Actinomycetales (UID1815)	574	266	63.75	0.63	0	107
<i>Nocardia kruzakiae</i> NBRC_101016	o Actinomycetales (UID1815)	571	265	98.99	1.7	11.11	103
<i>Nocardia lijiangensis</i> NBRC_108240	o Actinomycetales (UID1815)	574	266	99.23	1.14	0	230
<i>Nocardia mexicana</i> NBRC_108244	o Actinomycetales (UID1815)	571	265	98.24	2.63	10	180
<i>Nocardia mikamii</i> NBRC_108933	o Actinomycetales (UID1815)	571	265	98.99	1.91	0	42
<i>Nocardia miyuenensis</i> NBRC_108239	o Actinomycetales (UID1815)	567	264	99.67	3.14	4.35	208
<i>Nocardia nilgatensis</i> NBRC_100131	o Actinomycetales (UID1815)	567	264	99.65	2.65	4.17	114
<i>Nocardia niwae</i> NBRC_108934	o Actinomycetales (UID1815)	574	266	99.81	1.58	7.69	105
<i>Nocardia nova</i> SH22a NONO	o Actinomycetales (UID1815)	571	265	99.37	0.86	0	1
<i>Nocardia otitidiscaeviarum</i> IFM_11049	o Actinomycetales (UID1815)	567	264	99.29	1.63	5.26	65
<i>Nocardia otitidiscaeviarum</i> NBRC_14405	o Actinomycetales (UID1815)	567	264	99.29	1.61	8.33	239
<i>Nocardia paucivorans</i> NBRC_100373	o Actinomycetales (UID1815)	574	266	98.87	0.78	0	128
<i>Nocardia pneumoniae</i> NBRC_100136	o Actinomycetales (UID1815)	574	266	99.44	1.65	8.33	150
<i>Nocardia pseudobrasiliensis</i> IFM_0761	o Actinomycetales (UID1815)	567	264	96.62	3.51	15.38	204
<i>Nocardia pseudovaccinii</i> NBRC_100343	o Actinomycetales (UID1815)	574	266	99.75	4.77	16.67	492
<i>Nocardia puris</i> NBRC_108233	o Actinomycetales (UID1815)	572	265	99.32	2.99	0	175
<i>Nocardia rhamnosiphila</i> NBRC_108938	o Actinomycetales (UID1815)	574	266	99.32	1.44	15.79	93
<i>Nocardia salmonicida</i> NBRC_13393	o Actinomycetales (UID1815)	574	266	99.81	3.36	11.76	99
<i>Nocardia seriolae</i> N-2927	o Actinomycetales (UID1815)	567	264	99.81	2.81	0	593
<i>Nocardia seriolae</i> U-1	o Actinomycetales (UID1815)	567	264	99.61	2.81	0	313
<i>Nocardia seriolae</i> ZJ0503	o Actinomycetales (UID1815)	567	264	99.81	2.62	0	319
<i>Nocardia shimofuensis</i> NBRC_100134	o Actinomycetales (UID1815)	574	266	99.69	0.75	0	75
<i>Nocardia sienata</i> NBRC_100364	o Actinomycetales (UID1815)	574	266	99.29	1.39	4.76	135
<i>Nocardia soli</i> NBRC_100376	o Actinomycetales (UID1815)	574	266	99.56	1.72	0	111
<i>Nocardia</i> sp. 348MFTsu.1	o Actinomycetales (UID1814)	572	276	99.31	0.38	0	35
<i>Nocardia</i> sp. 852002-20019 SCH5090214	o Actinomycetales (UID1815)	571	265	98.99	2.13	20	467
<i>Nocardia</i> sp. 852002-51101 SCH5132738	o Actinomycetales (UID1815)	571	265	99.37	1.07	18.18	217
<i>Nocardia</i> sp. 852002-51244 SCH5132740	o Actinomycetales (UID1815)	571	265	99.37	1.44	16.67	297
<i>Nocardia</i> sp. BMC_111209	o Actinomycetales (UID1815)	567	264	99.75	1.29	0	5
<i>Nocardia</i> sp. BMC_51109	o Actinomycetales (UID1815)	571	265	99.37	2.28	12.5	4
<i>Nocardia</i> sp. CNS044	o Actinomycetales (UID1815)	574	266	99.75	1.51	14.29	13
<i>Nocardia</i> sp. CNY236	o Actinomycetales (UID1815)	572	265	98.62	1.63	0	75
<i>Nocardia</i> sp. LAM0056	o Actinomycetales (UID1815)	574	266	99.37	0.76	0	5
<i>Nocardia</i> sp. NRRL_3-836	o Actinomycetales (UID1746)	368	206	99.51	1.53	0	188
<i>Nocardia</i> sp. NRRL_WC-3656	o Actinomycetales (UID1815)	571	265	98.99	1.17	29.41	152
<i>Nocardia</i> sp. Root136	o Actinomycetales (UID1815)	574	266	99.81	1.9	0	14
<i>Nocardia</i> sp. W9405	o Actinomycetales (UID1815)	572	265	99.31	3.08	0	110
<i>Nocardia spelunca</i> NBRC_108251	o Actinomycetales (UID1815)	574	266	99.32	0.54	0	61
<i>Nocardia takedensis</i> NBRC_100417	o Actinomycetales (UID1815)	572	265	99.51	3.41	4.35	265
<i>Nocardia tenerifensis</i> NBRC_101015	o Actinomycetales (UID1815)	572	265	98.88	4.46	15.62	615
<i>Nocardia terpenica</i> IFM_0406	o Actinomycetales (UID1815)	567	264	99.37	5.84	10	21
<i>Nocardia terpenica</i> NBRC_100888	o Actinomycetales (UID1815)	567	264	90.51	11.47	6.94	1
<i>Nocardia testacea</i> NBRC_100365	o Actinomycetales (UID1815)	574	266	98.94	1.79	21.74	271
<i>Nocardia thailandica</i> NBRC_100428	o Actinomycetales (UID1815)	572	265	99.18	3.3	0	312
<i>Nocardia transversalis</i> NBRC_15921	o Actinomycetales (UID1815)	572	265	99.18	3.3	0	128
<i>Nocardia uniformis</i> NBRC_13702	o Actinomycetales (UID1815)	567	264	99.67	2.05	0	144
<i>Nocardia vaccinii</i> NBRC_15922	o Actinomycetales (UID1815)	567	264	99.48	2.33	0	131
<i>Nocardia vermiculata</i> NBRC_100427	o Actinomycetales (UID1815)	571	265	98.99	0.28	0	80
<i>Nocardia veterana</i> NBRC_100344	o Actinomycetales (UID1815)	571	265	99.37	1.01	50	210
<i>Nocardia vinacea</i> NBRC_16497	o Actinomycetales (UID1815)	574	266	99.09	4.37	14.29	429
<i>Nocardia violaceofusca</i> NBRC_14427	o Actinomycetales (UID1815)	571	265	98.96	0.5	10	77
<i>Nocardia vulneris</i> W9851	o Actinomycetales (UID1815)	574	266	99.75	3.84	8.82	137
<i>Nocardia xishanensis</i> NBRC_101358	o Actinomycetales (UID1815)	574	266	99.56	2.22	0	124
<i>Nocardia yamanashiensis</i> NBRC_100130	o Actinomycetales (UID1815)	567	264	99.62	3.06	5	80

Table S 21: Analysis of putative virulence factors in *Nocardia* spp..

Genes for catalases and superoxide dismutases were manually searched by hmmsearch (HMMER 3.1b2). Catalases included searches for Catalase, Mn_catalase Catalase-rel, Catalase_C, HTHP and Peroxidase Hidden Markov Models. Superoxide Dismutases (SoDs) included searches for Sod_Cu, Sod_Fe_C, Sod_Fe_N and Sod_Ni Hidden Markov Models. Virulence factor counts were in accordance with Vera-Cabrera *et al.* 2013.²

strain	source of isolation	catalases	SoDs	number of BGCs	autoMLST clade	genome size [Mbp]
<i>Mycobacterium tuberculosis</i> H37Rv	human	1	2	15	n.a.	4.4
<i>Nocardia abscessus</i> NBRC 100374	human	5	2	46	A	8.4
<i>Nocardia acidovorans</i> NBRC 108247	soil	3	2	37	F	7.6
<i>Nocardia africana</i> NBRC 100379	human	6	2	30	E	7.8
<i>Nocardia alba</i> NBRC 108234	soil	3	2	32	D	7.3
<i>Nocardia altamirensis</i> NBRC 108246	soil	n.a.	n.a.	57	A	9.8
<i>Nocardia amamiensis</i> NBRC 102102	soil	3	2	45	A	8.2
<i>Nocardia amikacinotolerans</i> DSM 45535	human	3	2	32	B	7.3
<i>Nocardia amikacinotolerans</i> DSM 45536	human	4	2	30	B	7.5
<i>Nocardia amikacinotolerans</i> DSM 45537	human	4	2	31	B	7.4
<i>Nocardia amikacinotolerans</i> DSM 45538	human	4	2	31	B	7.6
<i>Nocardia amikacinotolerans</i> NBRC 108937	human	4	2	30	B	7.6
<i>Nocardia anaeimae</i> NBRC 100462	human	5	2	46	A	8.6
<i>Nocardia abensis</i> NBRC 100429	human	4	2	43	E	7.5
<i>Nocardia arvensis</i> NBRC 100135	human	4	2	45	A	7.7
<i>Nocardia arthritis</i> NBRC 100137	human	5	2	44	A	7.1
<i>Nocardia asiatica</i> NBRC 100129	human	5	2	56	A	8.5
<i>Nocardia asteroides</i> DSM 43373	human	3	2	30	D	7.0
<i>Nocardia asteroides</i> NBRC 15531	soil	3	2	30	D	7.0
<i>Nocardia beijingensis</i> NBRC 16342	soil	4	2	43	A	7.5
<i>Nocardia brasiliensis</i> ATCC 700358	human	5	2	49	A	9.4
<i>Nocardia brasiliensis</i> IFM 10847	human	5	2	56	A	9.2
<i>Nocardia brasiliensis</i> NBRC 14402	human	6	2	49	A	8.9
<i>Nocardia brevicatenata</i> NBRC 12119	human	3	2	35	C	7.0
<i>Nocardia caishijensis</i> NBRC 108228	soil	2	2	28	D	6.3
<i>Nocardia carneae</i> NBRC 14463	human	3	2	29	C	7.5
<i>Nocardia cerasioides</i> NBRC 101014	soil	5	2	50	E	7.6
<i>Nocardia concava</i> NBRC 100430	human	4	2	51	F	8.9
<i>Nocardia coulebae</i> NBRC 108252	soil	3	2	31	D	6.6
<i>Nocardia crassostreae</i> NBRC 100342	animal	3	2	41	F	8.3
<i>Nocardia cummideiensis</i> NBRC 100378	plant	3	2	30	D	7.5
<i>Nocardia cyriacigeorgica</i> GUH-2	human	3	2	22	C	6.2
<i>Nocardia cyriacigeorgica</i> NBRC 100375	human	4	3	35	C	6.3
<i>Nocardia donostiensis</i> X1655	human	4	2	29	C	5.8
<i>Nocardia elegans</i> NBRC 108235	human	3	2	36	E	7.5
<i>Nocardia exalida</i> NBRC 100660	human	5	2	44	A	7.4
<i>Nocardia farcinica</i> DSM 43257	human	5	2	23	B	6.4
<i>Nocardia farcinica</i> IFM 10152	human	4	2	17	B	6.3
<i>Nocardia farcinica</i> NCTC 11134	animal	5	2	19	B	6.5
<i>Nocardia flavovaseae</i> NBRC 108225	soil	3	3	25	C	7.4
<i>Nocardia fluminea</i> DSM 44489	freshwater	2	2	32	D	8.1
<i>Nocardia fusca</i> NBRC 14340	soil	4	2	32	C	8.0
<i>Nocardia gamkensis</i> NBRC 108242	soil	4	2	37	A	7.7
<i>Nocardia grenadensis</i> NBRC 108939	soil	3	2	29	C	6.5
<i>Nocardia harense</i> NBRC 108248	soil	3	2	26	X	6.1
<i>Nocardia higoensis</i> NBRC 100133	human	2	2	28	B	7.0
<i>Nocardia ignarata</i> NBRC 108230	human	3	2	32	D	7.0
<i>Nocardia imohaniensis</i> NBRC 100128	human	6	2	42	F	8.1
<i>Nocardia lianyuensis</i> NBRC 101359	soil	5	3	38	E	10.4
<i>Nocardia linumensis</i> NBRC 108249	soil	4	3	29	C	5.0
<i>Nocardia kruszchiae</i> NBRC 101016	human	4	2	36	E	7.3
<i>Nocardia lijiangensis</i> NBRC 108240	soil	4	2	47	B	8.2
<i>Nocardia mexicana</i> NBRC 108244	human	4	2	43	E	9.0
<i>Nocardia mikamii</i> NBRC 108933	human	4	2	36	E	7.6
<i>Nocardia miyuenensis</i> NBRC 108239	soil	9	2	41	E	10.5
<i>Nocardia niigatensis</i> NBRC 100131	human	4	2	33	F	8.2
<i>Nocardia niwaie</i> NBRC 108934	human	4	2	40	A	7.3
<i>Nocardia nova</i> SH22a	plant	3	2	30	E	8.3
<i>Nocardia otitidiscaviarum</i> IFM 11049	human	5	3	31	F	7.9
<i>Nocardia otitidiscaviarum</i> NBRC 14445	animal	6	3	33	F	7.5
<i>Nocardia pouvairensis</i> NBRC 100373	human	5	2	30	C	6.0
<i>Nocardia pneumoniae</i> NBRC 100136	human	4	2	41	A	7.6
<i>Nocardia pseudobrasiliensis</i> IFM 0761	human	4	3	51	E	8.3
<i>Nocardia pseudovaccinii</i> NBRC 100343	soil	4	2	56	A	9.9
<i>Nocardia puris</i> NBRC 108233	human	3	2	38	B	7.7
<i>Nocardia rhamnosiphila</i> NBRC 108938	soil	3	2	40	C	7.7
<i>Nocardia salmonicida</i> NBRC 13393	animal	3	2	33	D	8.3
<i>Nocardia seriolae</i> N-2927	animal	2	2	36	F	7.7
<i>Nocardia seriolae</i> U-1	animal	2	2	30	F	7.7
<i>Nocardia seriolae</i> ZJ0503	animal	2	2	31	F	7.7
<i>Nocardia shimizuensis</i> NBRC 100134	soil	2	2	23	B	6.3
<i>Nocardia sinensis</i> NBRC 100364	human	4	2	38	C	6.8
<i>Nocardia soil</i> NBRC 100376	freshwater	3	2	32	D	7.6
<i>Nocardia</i> sp. 852002-20019_SCH5090214	?	4	2	35	E	8.3
<i>Nocardia</i> sp. 852002-51101_SCH5132738	?	4	2	35	E	7.7
<i>Nocardia</i> sp. 852002-51244_SCH5132740	?	4	2	36	E	7.7
<i>Nocardia</i> sp. BMG 111209	plant	3	2	29	F	9.1
<i>Nocardia</i> sp. BMG 51109	plant	4	2	45	E	8.8
<i>Nocardia</i> sp. CNS044	marine	5	2	35	B	7.4
<i>Nocardia</i> sp. CNY236	marine	2	2	38	A	5.3
<i>Nocardia</i> sp. NRRL WC-3656	soil	3	2	41	E	7.3
<i>Nocardia</i> sp. Roat136	plant	3	2	30	D	7.3
<i>Nocardia</i> sp. W9805	human	2	2	32	B	7.2
<i>Nocardia spelunceae</i> NBRC 108251	soil	5	2	24	C	7.4
<i>Nocardia takendensis</i> NBRC 100417	soil	3	2	35	B	8.0
<i>Nocardia tenerifensis</i> NBRC 101015	soil	5	3	71	A	9.7
<i>Nocardia terpenica</i> IFM 0406	human	4	2	48	E	9.3
<i>Nocardia terpenica</i> NBRC 100888	human	4	2	18	E	8.6
<i>Nocardia testacea</i> NBRC 100365	human	4	2	29	C	7.3
<i>Nocardia thailandica</i> NBRC 100428	human	3	2	33	D	6.8
<i>Nocardia transvalensis</i> NBRC 15921	human	4	2	36	D	8.4
<i>Nocardia uniformis</i> NBRC 13702	soil	4	2	36	F	8.8
<i>Nocardia vaccinii</i> NBRC 15922	plant	3	2	31	E	9.2
<i>Nocardia vermiculata</i> NBRC 103427	human	4	2	28	E	6.7
<i>Nocardia veterana</i> NBRC 100344	human	4	2	29	E	6.8
<i>Nocardia vinacea</i> NBRC 16497	soil	6	2	61	A	10.2
<i>Nocardia violaceofusca</i> NBRC 14427	soil	4	2	35	E	7.5
<i>Nocardia vulneris</i> W9851	human	6	2	46	A	9.3
<i>Nocardia xishanensis</i> NBRC 101358	soil	5	2	27	B	7.7
<i>Nocardia yamanashiensis</i> NBRC 100130	human	3	2	49	F	9.1
<i>Nocardia jejuensis</i> NBRC 103114	soil	4	2	37	F	8.7
<i>Nocardia</i> sp. LAM0056	soil	3	2	22	C	6.2

Table S 22: Presence/absence map of *N. farcinica* IFM 10152 BGCs in group 1 pathogens.

Putative antiSMASH (4.0.0) BGCs from *Nocardia farcinica* IFM 10152 were used as reference to manually analyze the BGC content of the other group 1 *Nocardia* strains. Presence/ absence in other *Nocardia* was marked in color, when a homologous BGC was found. Color indicates BGC cluster types: Light yellow (type I pks), dark blue (other), purple (terpene), dark green (nrps), dark yellow (arylpolyene), light green (ectoine), light blue (type I pks-nrps).

group 1 pathogen	antiSMASH 4.0.0 predicted BGCs																			
	1	2	3a	3b	3c	4	5	6	7	8	9	10	11	12	13	14	15	16a	16b	16c
<i>N. farcinica</i> IFM 10152	Light yellow	Dark blue	Purple	Dark green	Dark green	Dark green	Purple	Dark yellow	Dark blue	Light green	Dark green	Light blue	Dark green	Purple	Light yellow	Purple	Light yellow	Light blue	Dark green	Dark green
<i>N. cyriaciageorgica</i> GUH-2	Light yellow	Dark blue	Purple	Dark green	Dark green	Dark green	Purple	Dark yellow		Light green			Dark green		Light yellow			Light blue	Dark green	Dark green
<i>N. asteroides</i> NBRC 15531	Light yellow	Dark blue	Purple	Dark green	Dark green	Dark green	Purple	Dark yellow		Light green			Dark green	Purple	Light yellow			Light blue	Dark green	Dark green
<i>N. nova</i> SH22a	Light yellow		Purple	Dark green	Dark green		Purple			Light green			Dark green		Light yellow			Light blue	Dark green	Dark green
<i>N. otitidiscaviarum</i> NBRC 14405	Light yellow		Purple	Dark green	Dark green	Dark green	Purple	Dark yellow		Light green			Dark green		Light yellow			Light blue	Dark green	Dark green
<i>N. abscessus</i> NBRC 100374	Light yellow	Dark blue	Purple	Dark green	Dark green	Dark green	Purple	Dark yellow		Light green			Dark green	Purple	Light yellow			Light blue	Dark green	
<i>N. brasiliensis</i> IFM 10847	Light yellow	Dark blue	Purple	Dark green	Dark green	Dark green	Purple	Dark yellow		Light green			Dark green	Purple	Light yellow			Light blue	Dark green	

Figures

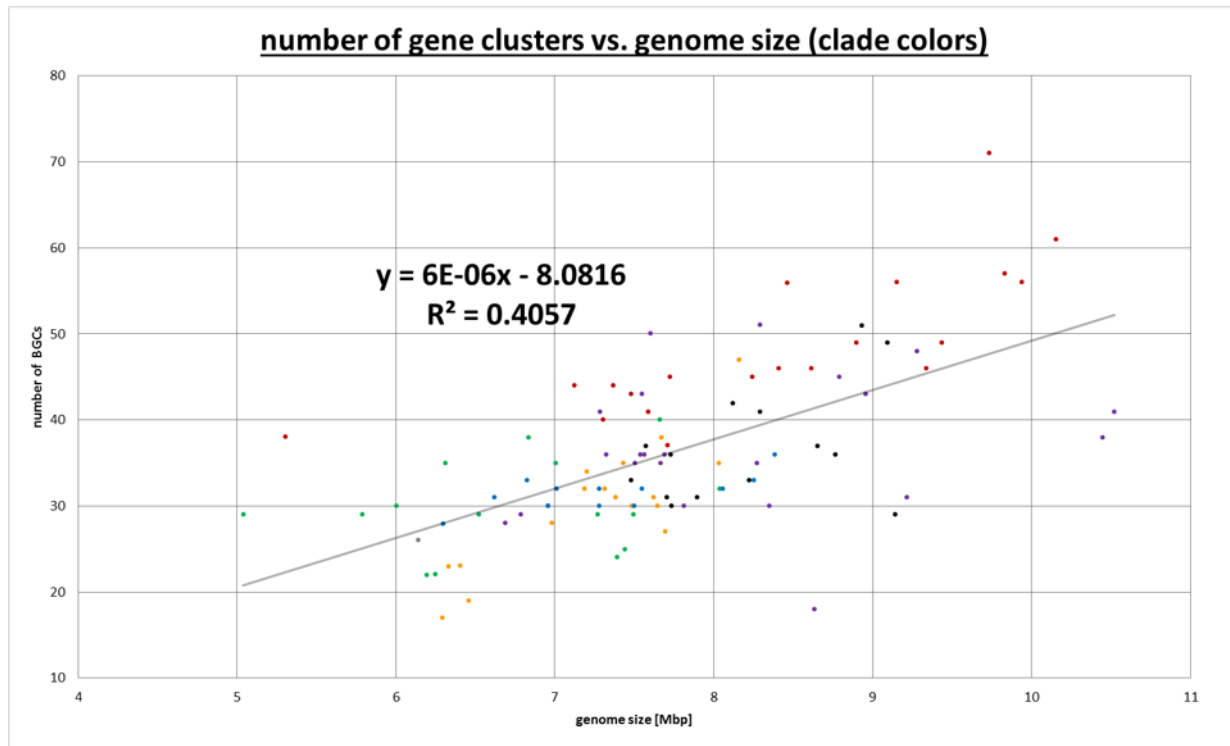


Figure S 1: Genome size plotted against the number of predicted BGCs in *Nocardia* strains.

Each dot corresponds to one *Nocardia* strain. Colors indicate phylogenetic clade A-F and X. Trendline indicates a positive correlation between number of BGCs present in a strain and its genome size (coefficient of determination ($R^2=0.4057$)).

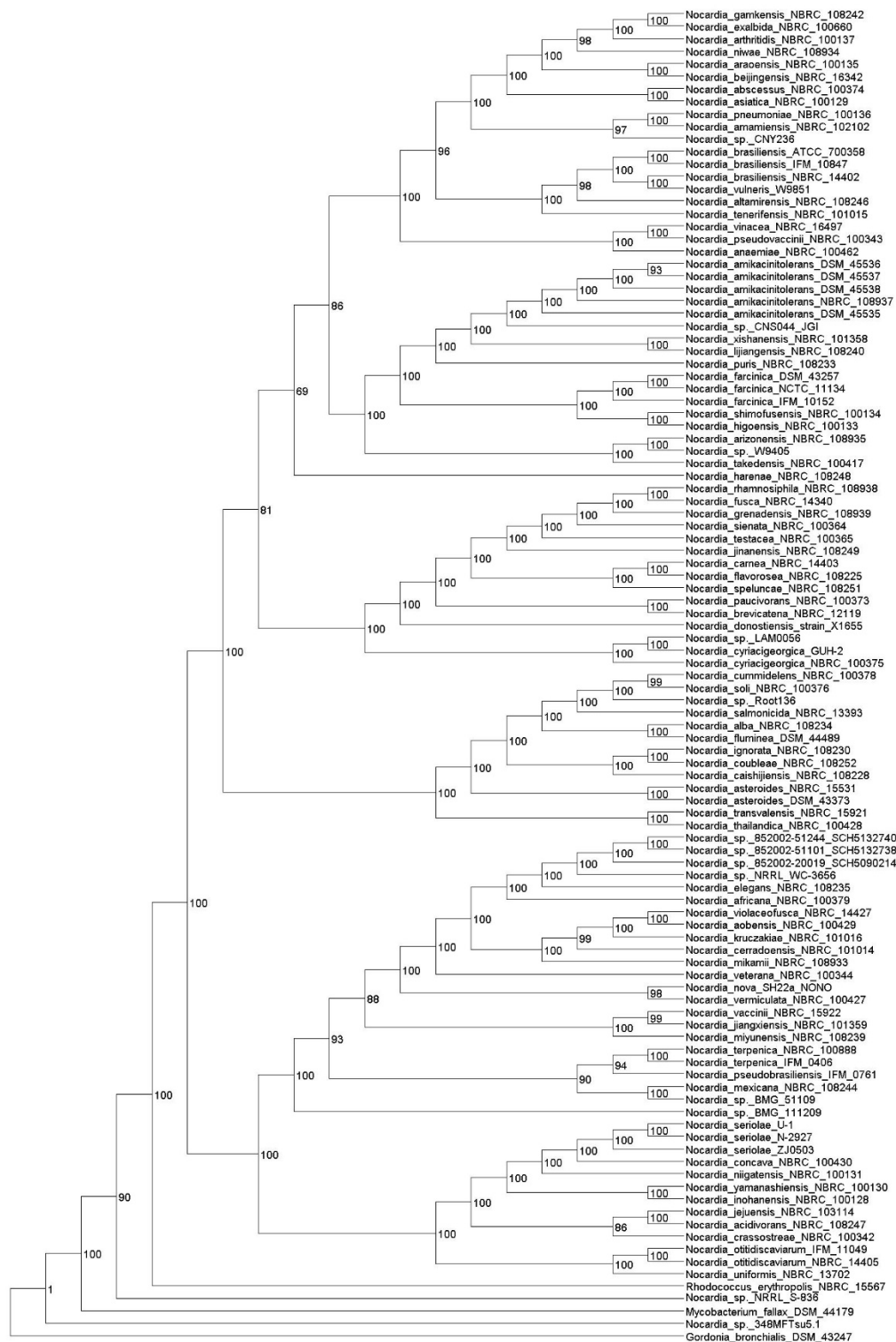


Figure S 2: *Nocardia* maximum likelihood autoMLST tree.

Analysis of 104 genomes obtained from JGI specified as *Nocardia* based on 63 concatenated housekeeping genes identified with autoMLST. Ultrafast bootstrap values were calculated using 1000 replicates. *Mycobacterium fallax* DSM 44179 and *Gordonia bronchialis* DSM 43247 were used as outgroups.

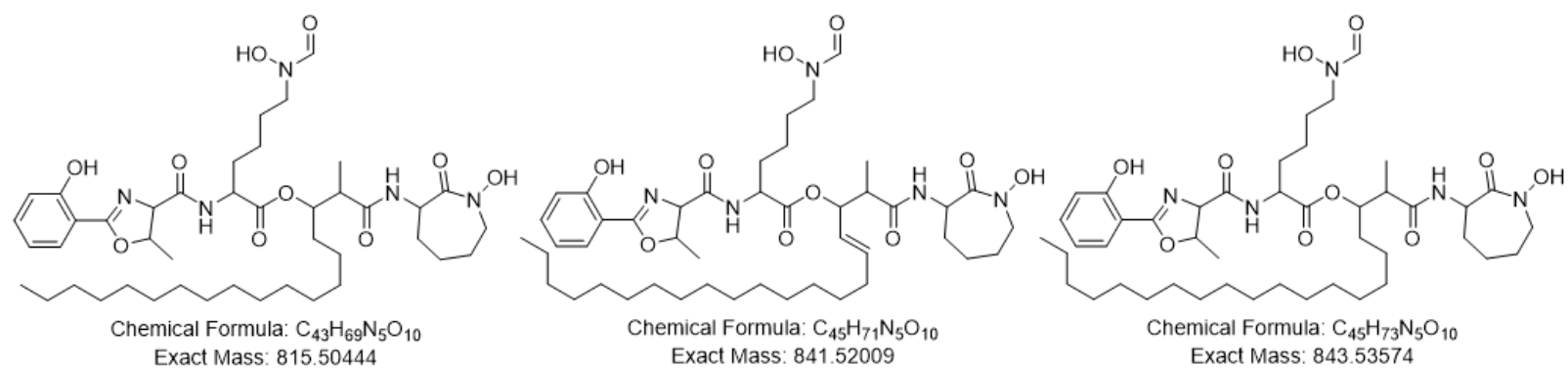


Figure S 3: Terpenibactins A-C produced by *N. terpenica* IFM 0406.

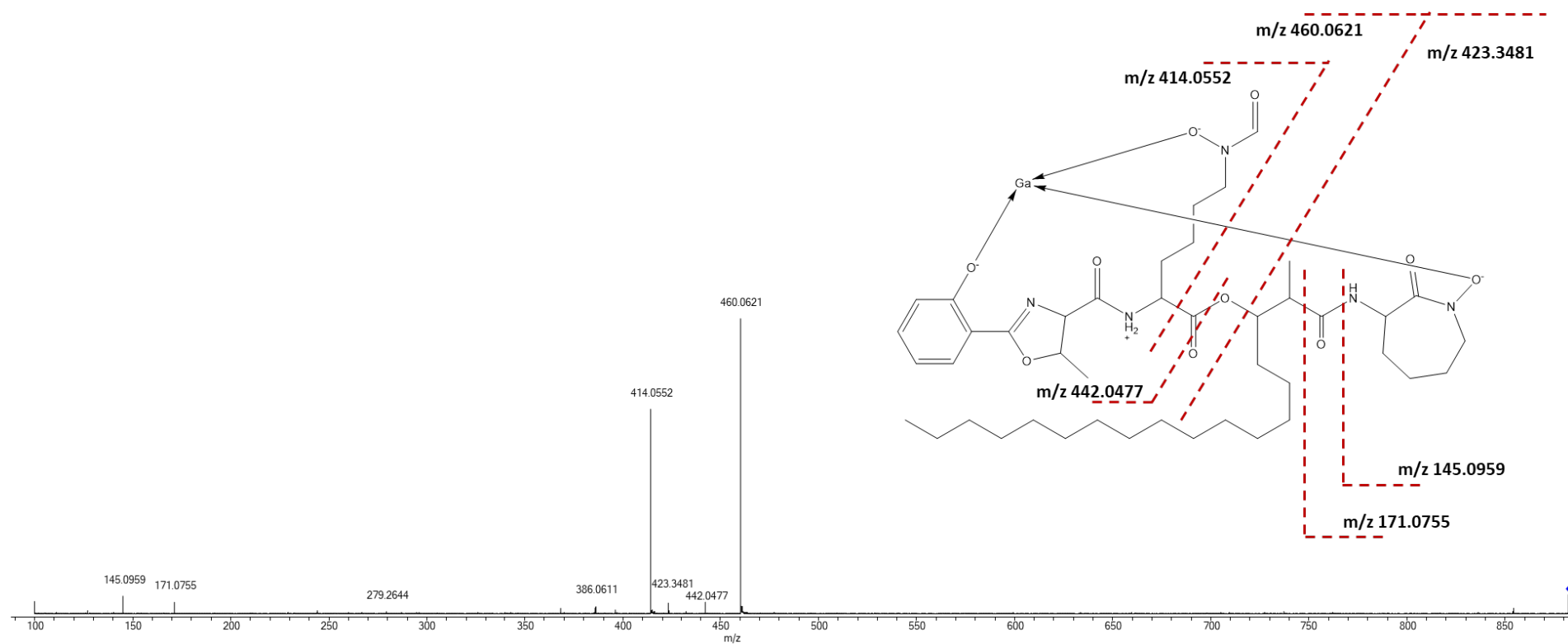


Figure S 4: MS/MS analysis of terpenibactin A.

Precursor mass (882.4135 [M-2H+Ga]⁺) is marked with a blue diamond.

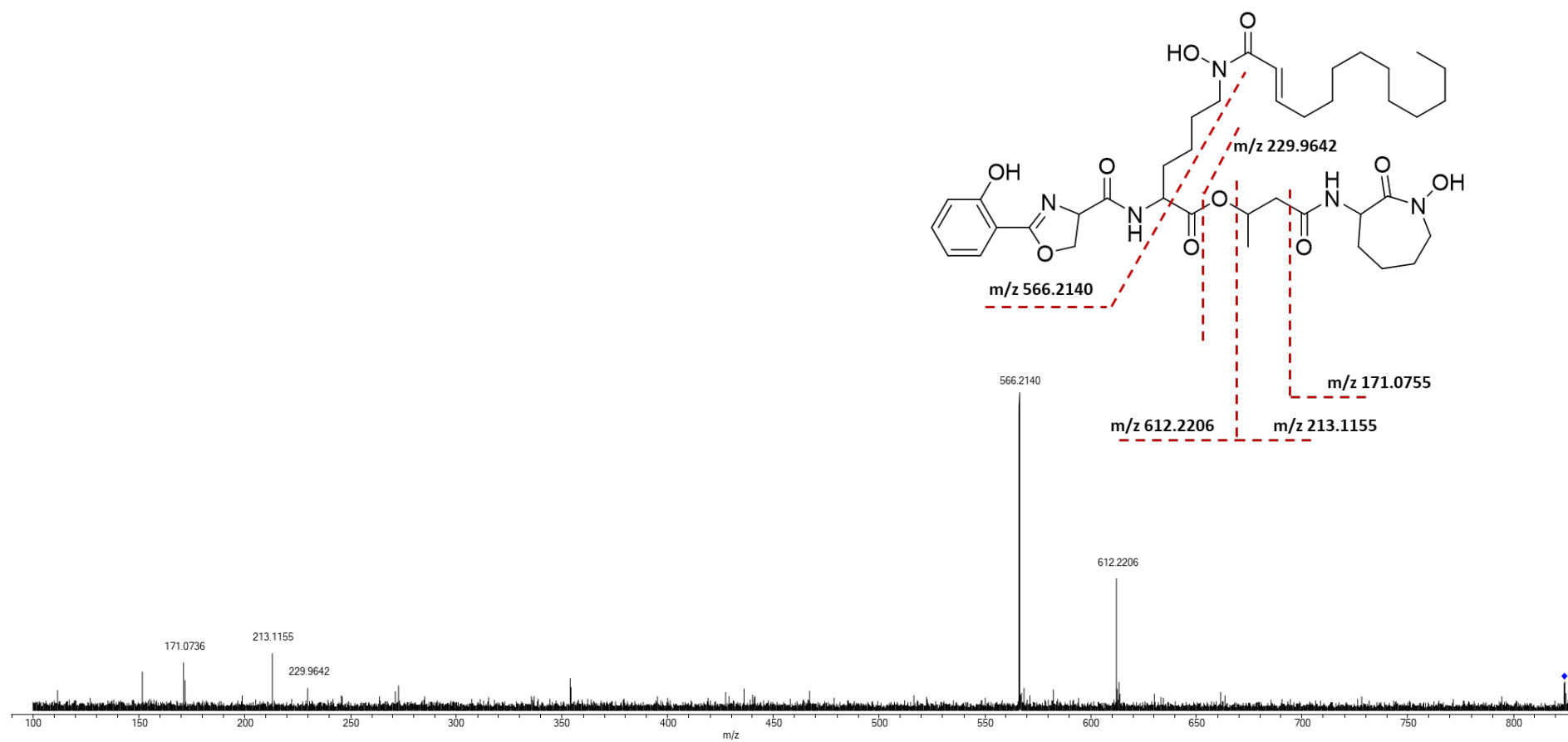


Figure S 5: MS/MS analysis of a putative mycobactin-like siderophore with 824.3353 [M-2H+Ga]⁺ produced by *Nocardia araoensis* NBRC 100135.

Precursor mass is marked with a blue diamond.

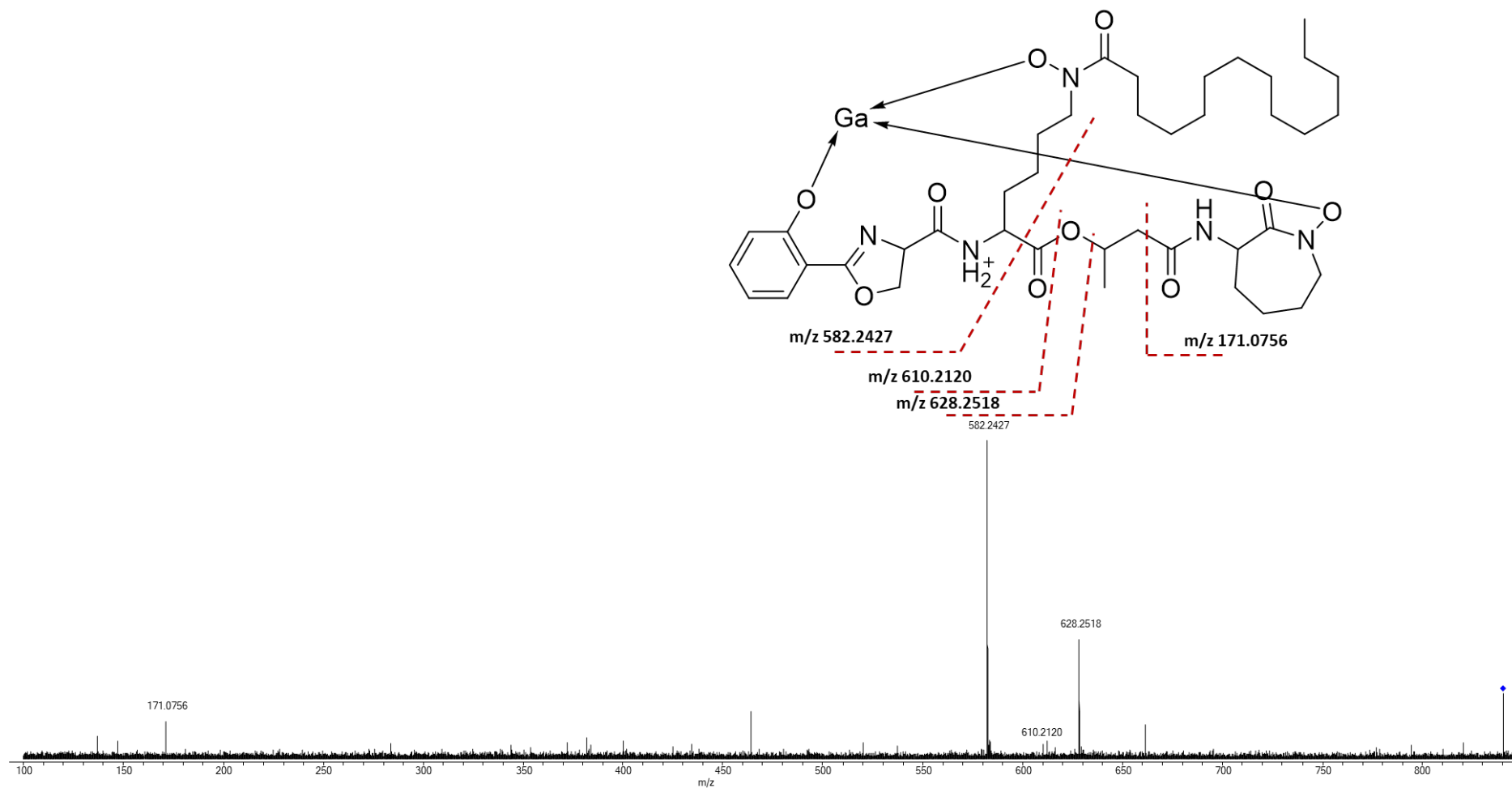


Figure S 6: MS/MS analysis of a putative mycobactin-like siderophore with 840.3669 [M-2H+Ga]⁺ produced by *Nocardia araoensis* NBRC 100135.

Precursor mass is marked with a blue diamond.

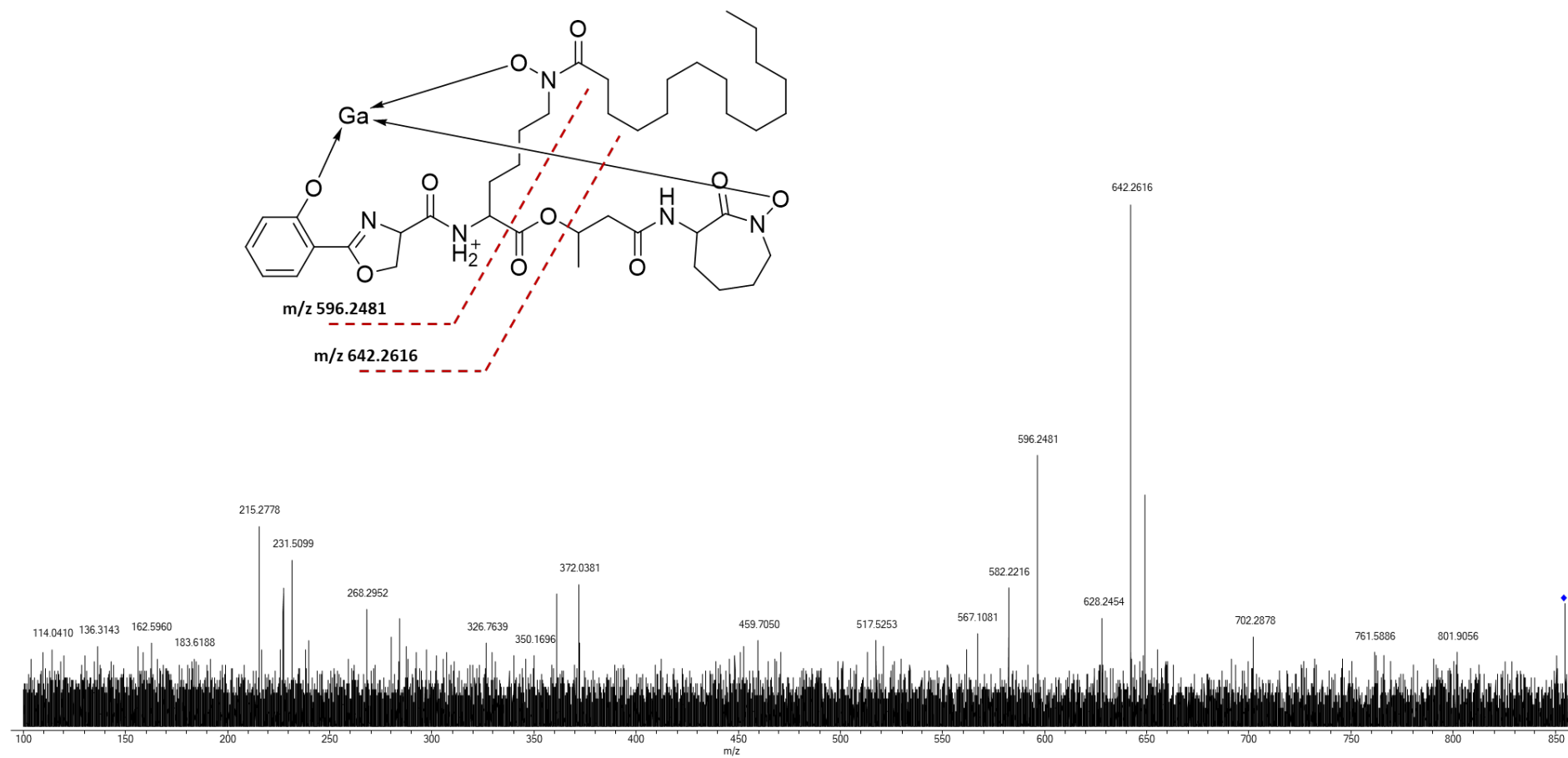


Figure S 7: MS/MS analysis of 854.3817 $[M-2H+Ga]^+$ produced by *Nocardia takedensis* NBRC 100417.

Precursor mass is marked with a blue diamond.

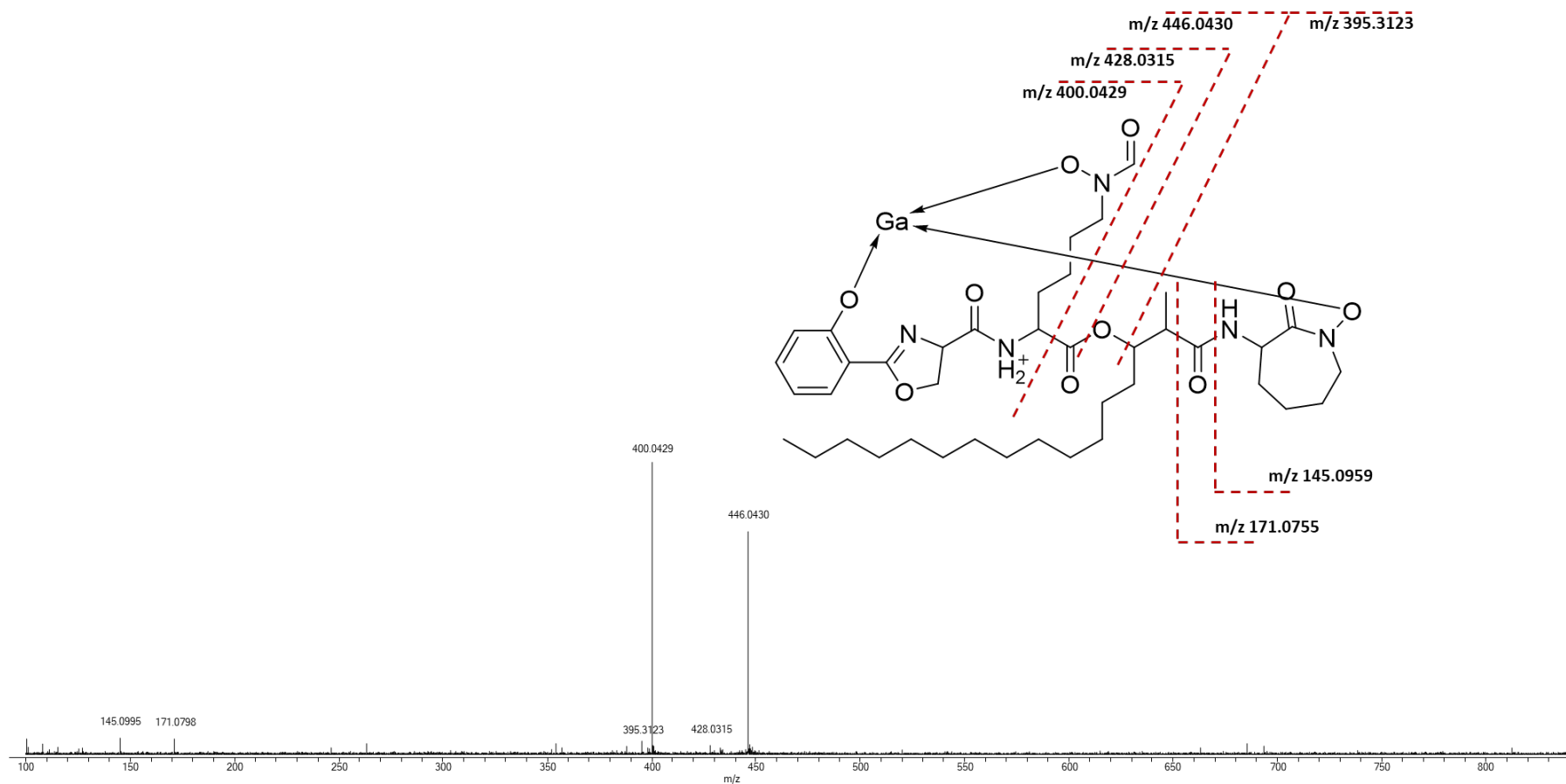


Figure S 8: MS/MS analysis of putative nocardimicin G with 840.3676 $[M-2H+Ga]^+$ produced by *Nocardia jejuensis* NBRC 103114.

Precursor mass is marked with a blue diamond.

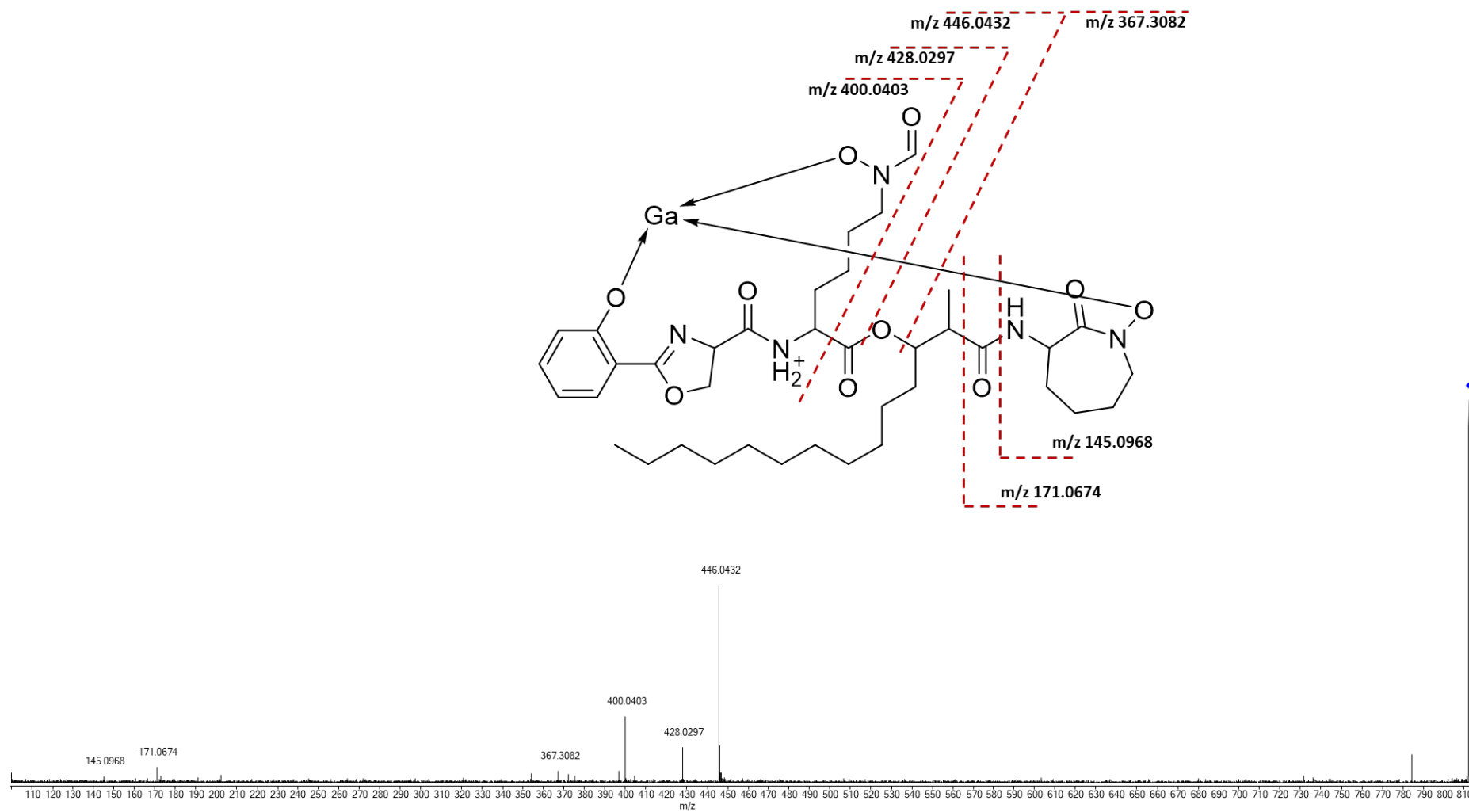


Figure S 9: MS/MS analysis of 812.3442 $[M-2H+Ga]^+$ produced by *Nocardia fusca* NBRC 14340.
Precursor mass is marked with a blue diamond.

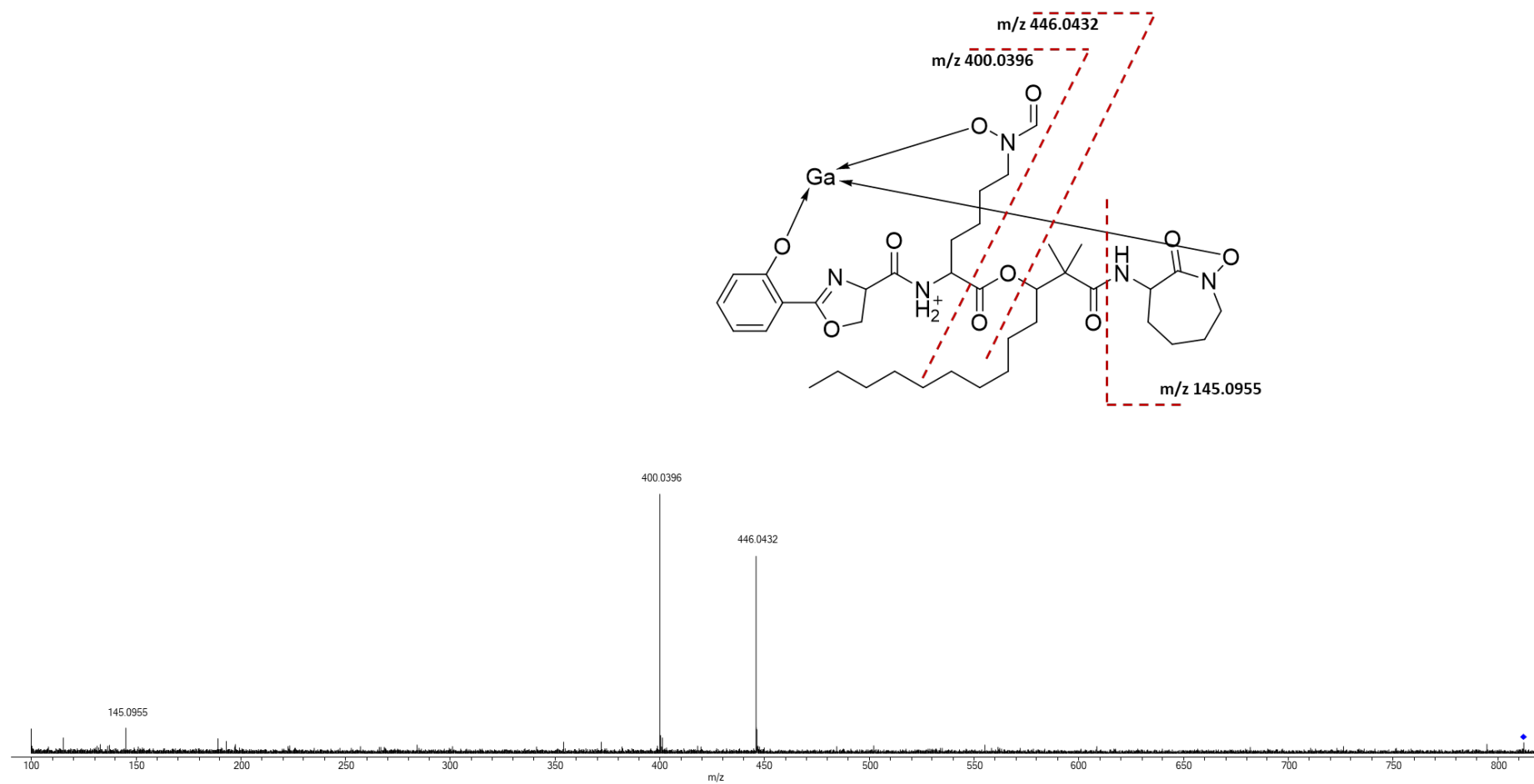


Figure S 10: MS/MS analysis of 812.3352 $[M-2H+Ga]^+$ produced by *Nocardia ramosiphila* NBRC 108938. Precursor mass is marked with a blue diamond.

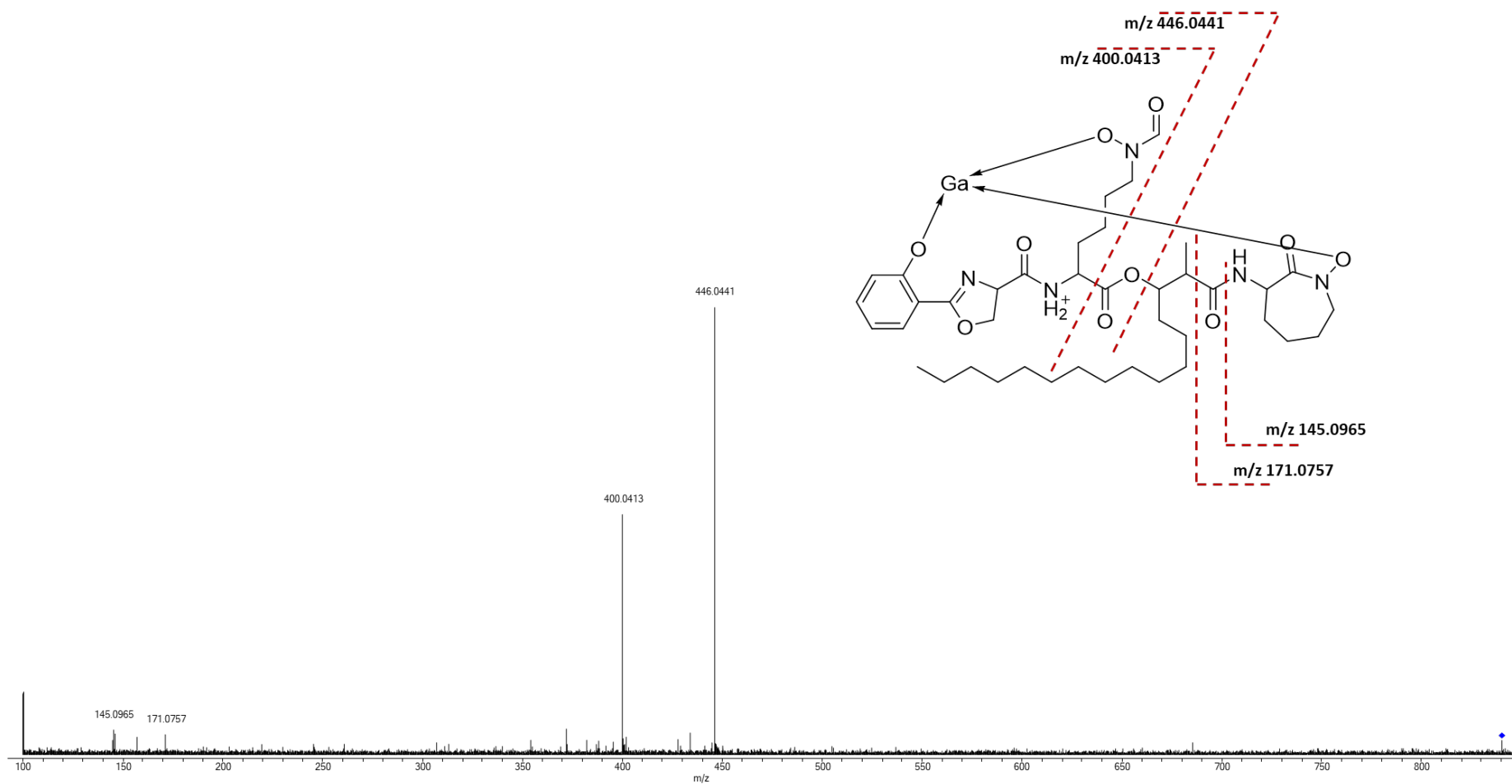


Figure S 11: MS/MS analysis of 840.3685 [M-2H+Ga]⁺ produced by *Nocardia vermiculata* NBRC 100427.
Precursor mass is marked with a blue diamond.

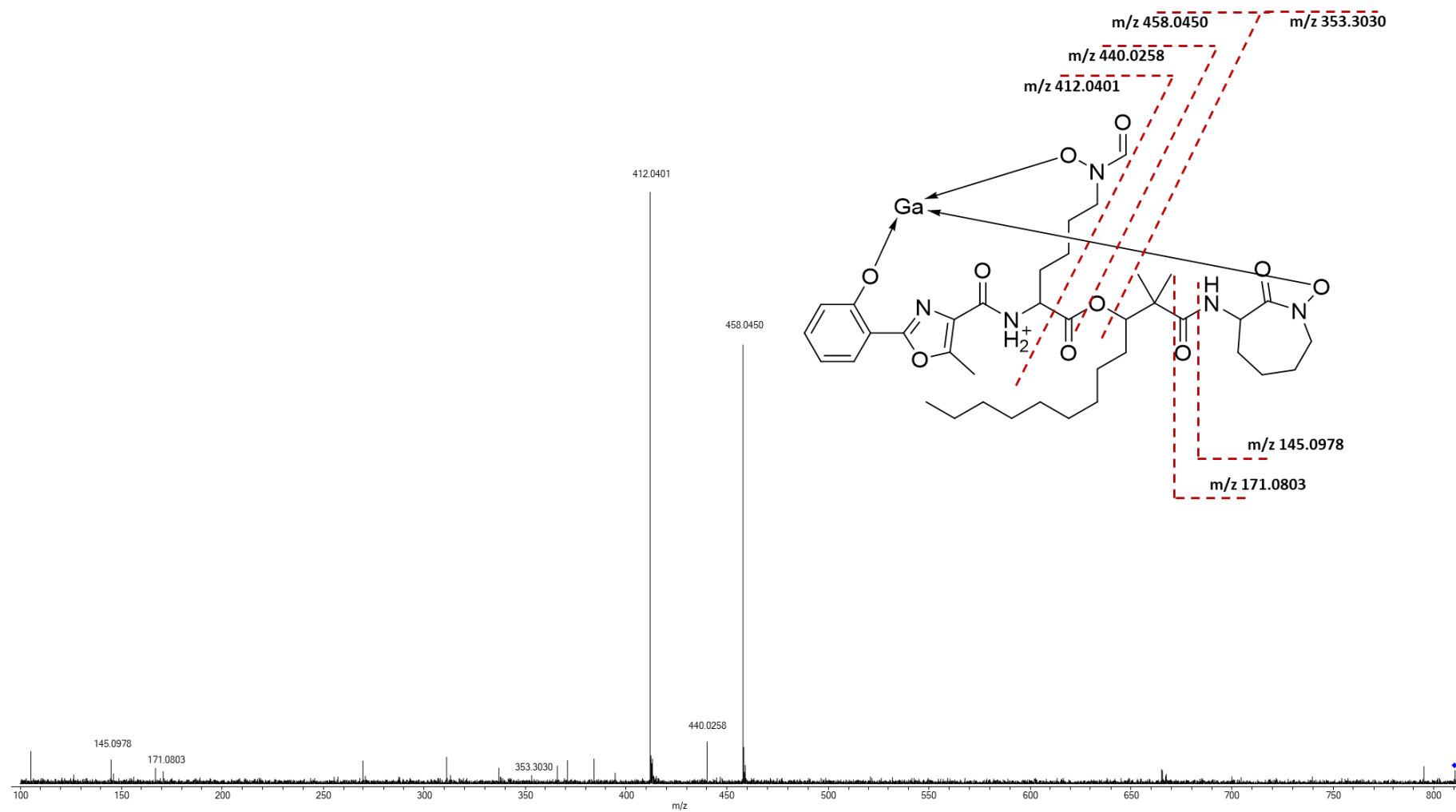


Figure S 12: MS/MS analysis of putative formobactin with 810.3192 [M-2H+Ga]⁺ produced by *Nocardia gamkensis* NBRC 108242.

Precursor mass is marked with a blue diamond.

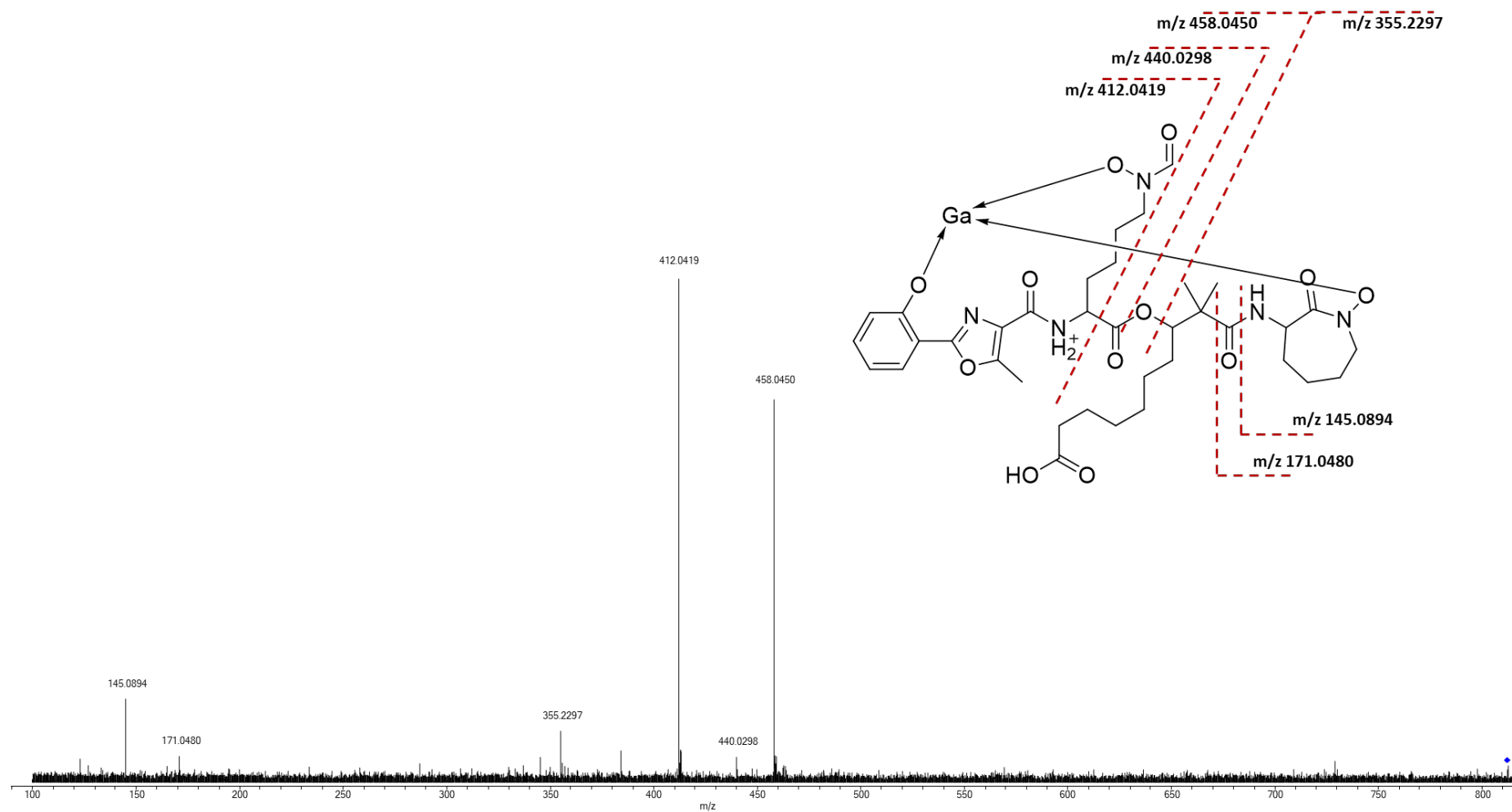


Figure S 13: MS/MS analysis of putative carboxynocobactin with 812.2634 [M-2H+Ga]⁺ produced by *Nocardia gamkensis* NBRC 108242. Precursor mass is marked with a blue diamond.

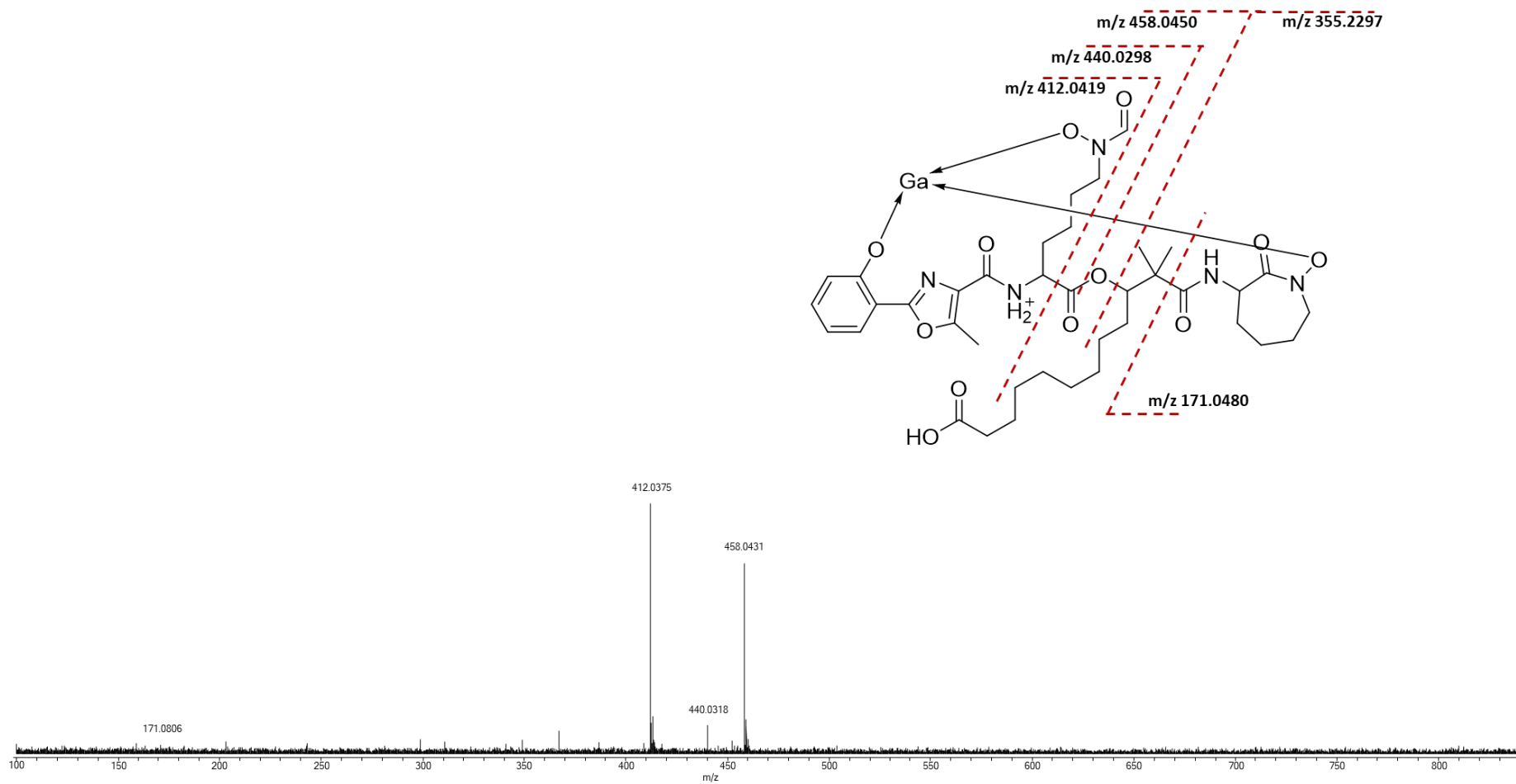


Figure S 14: MS/MS analysis of putative carboxynocobactin with 840.2954 [M-2H+Ga]⁺ produced by *Nocardia gamkensis* NBRC 108242. Precursor mass is marked with a blue diamond.

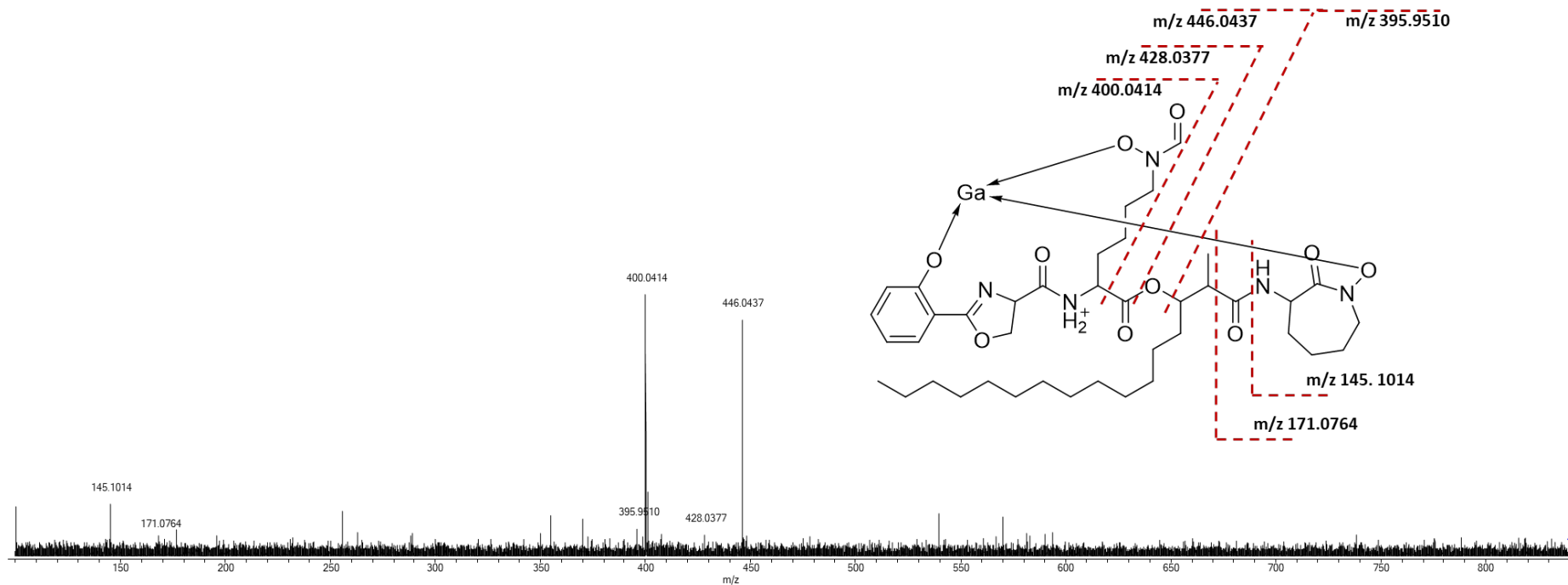


Figure S 15: MS/MS analysis of putative nocardimicin G with 840.3683 $[M-2H+Ga]^+$ produced by *Nocardia elegans* NBRC 108235.

Precursor mass is marked with a blue diamond.

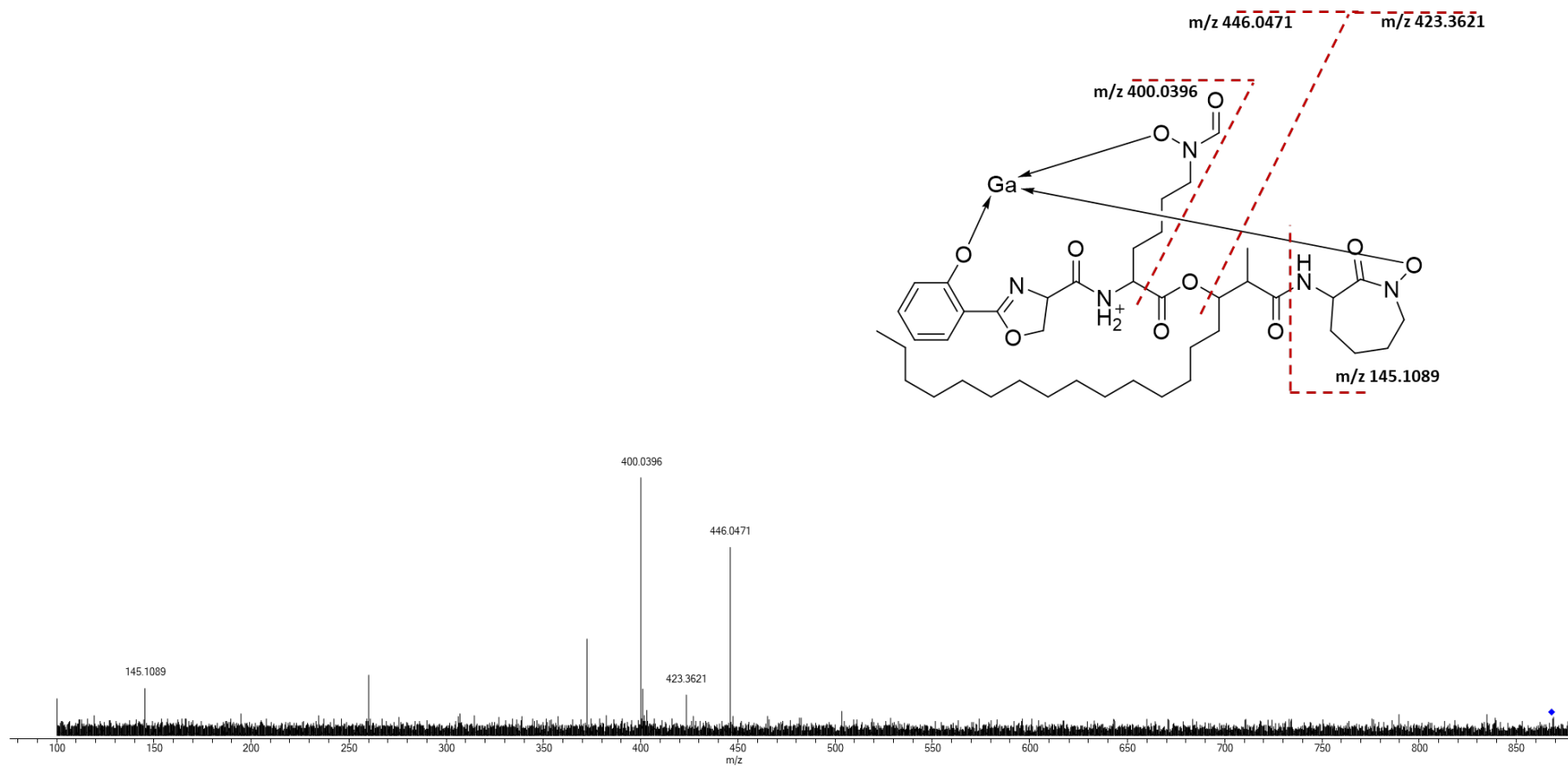


Figure S 16: MS/MS analysis of putative nocardimicin H with 868.3985 $[M-2H+Ga]^+$ produced by *Nocardia elegans* NBRC 108235. Precursor mass is marked with a blue diamond.

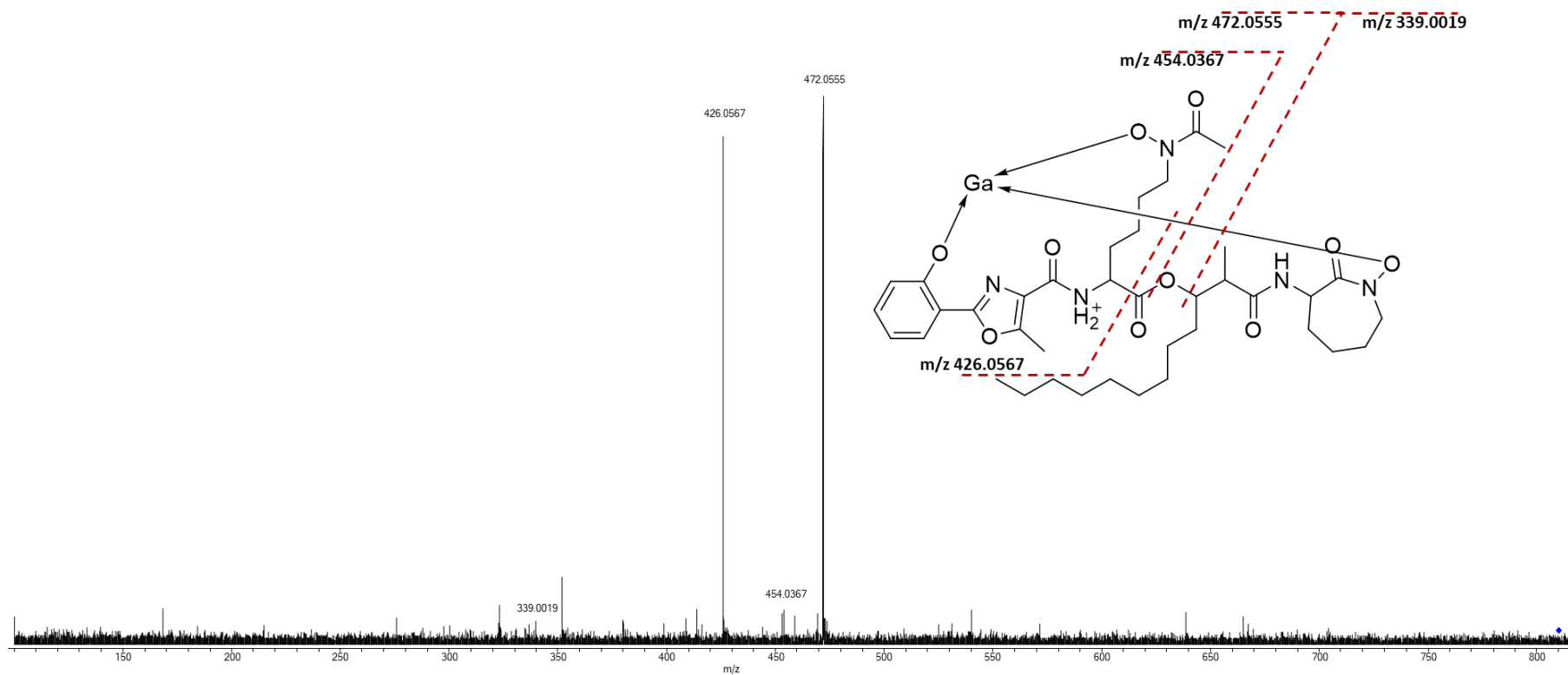


Figure S 17 MS/MS analysis of 810.3227 [M-2H+Ga]⁺ produced by *Nocardia alba* NBRC 108234.
Precursor mass is marked with a blue diamond.

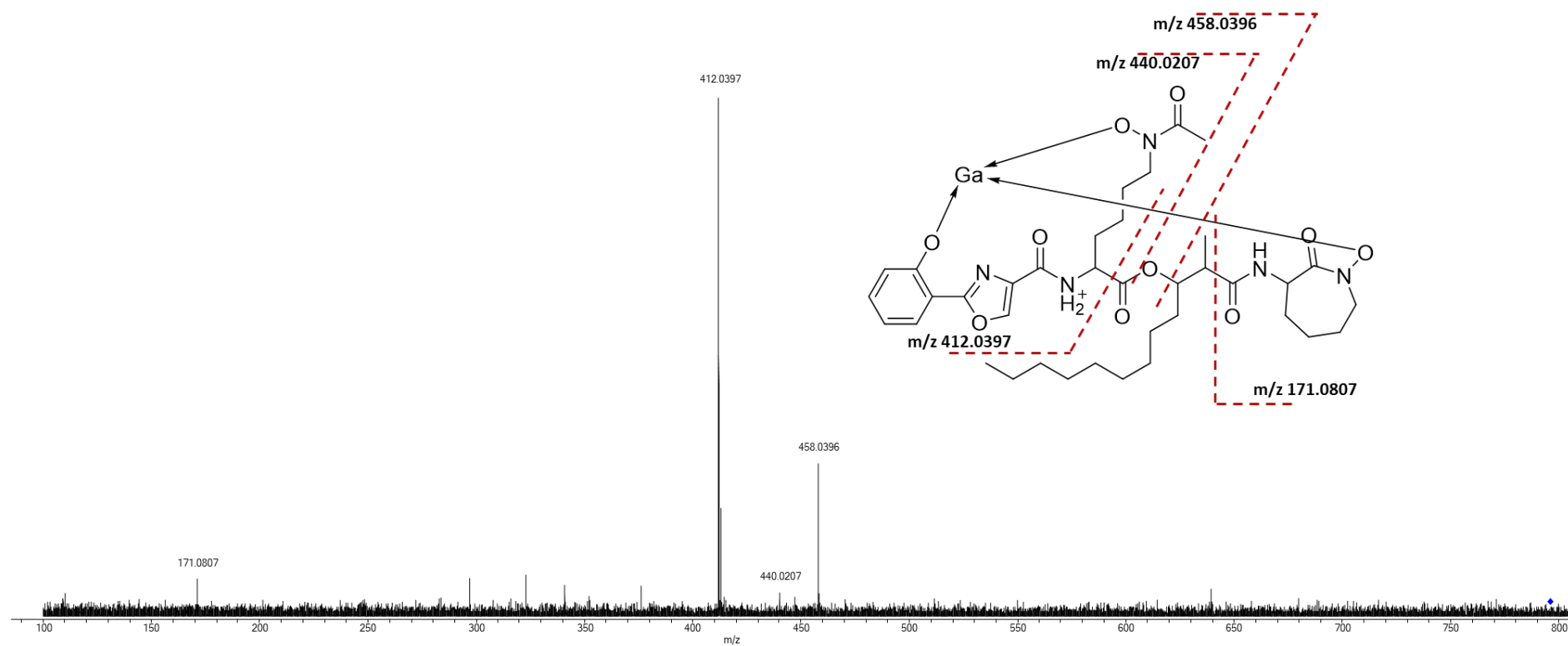


Figure S 18: MS/MS analysis of putative nocardimicin A with 796.3046 [M-2H+Ga]⁺ produced by *Nocardia alba* NBRC 108234.

Precursor mass is marked with a blue diamond.

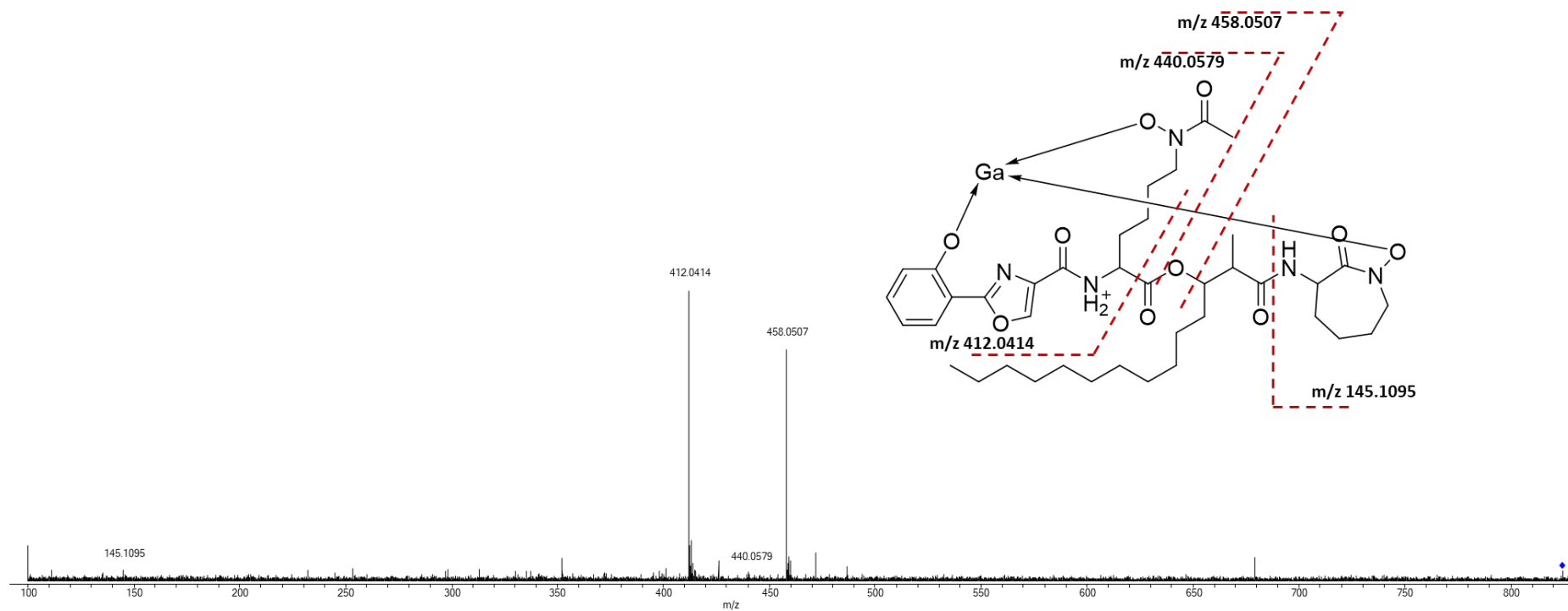


Figure S 19: MS/MS analysis of putative nocardimicin B with 824.3355 [M-2H+Ga]⁺ produced by *Nocardia alba* NBRC 108234.

Precursor mass is marked with a blue diamond.

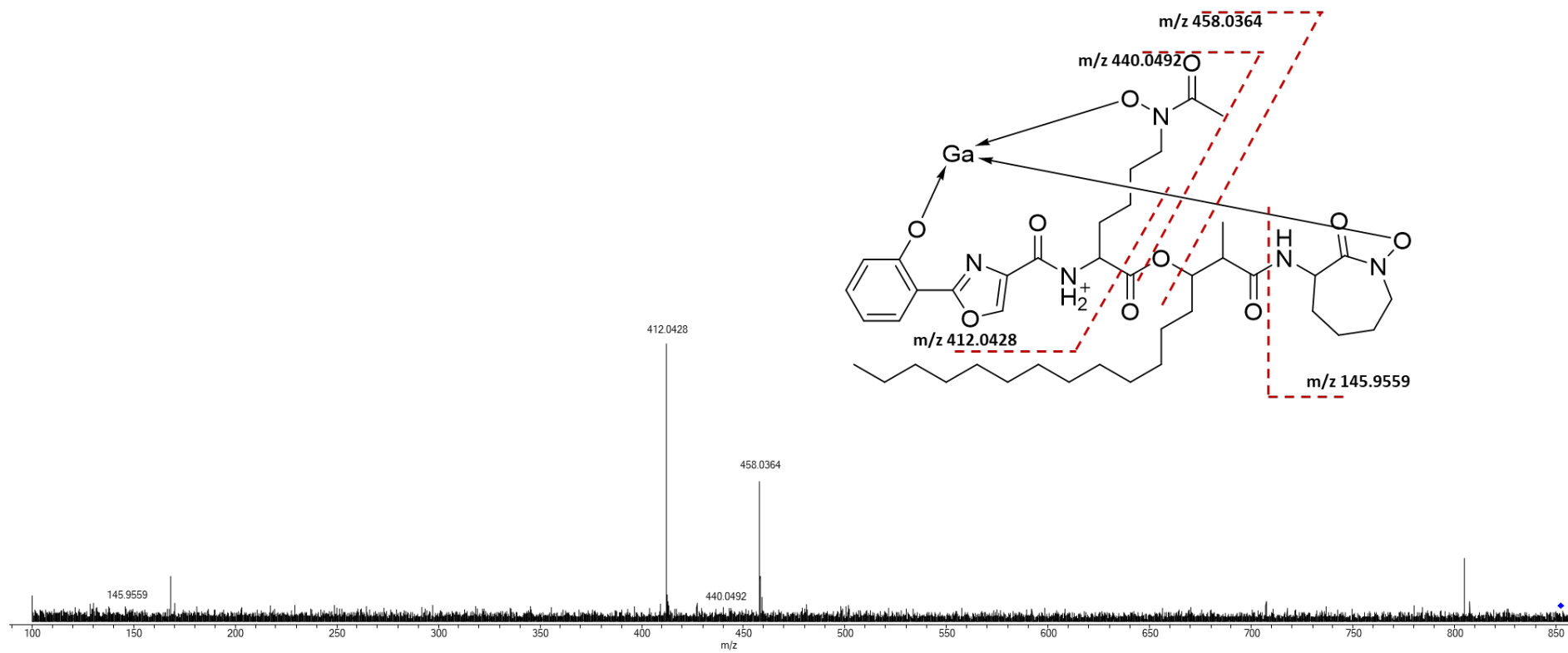


Figure S 20: MS/MS analysis of putative nocardimicin D with 852.3668 [M-2H+Ga]⁺ produced by *Nocardia alba* NBRC 108234.

Precursor mass is marked with a blue diamond.

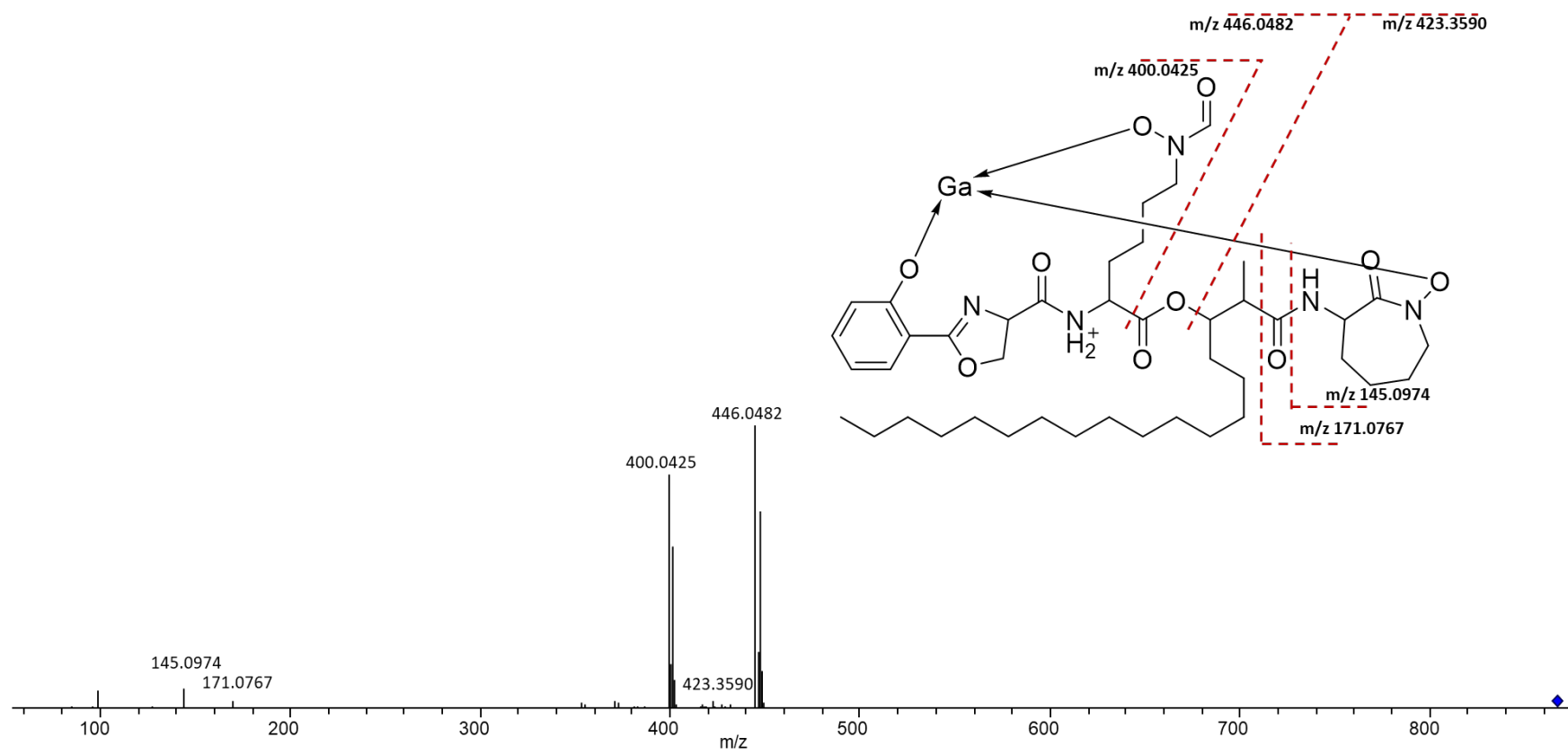


Figure S 21: MS/MS analysis of putative brasilibactin A with 868.3989 [M-2H+Ga]⁺ produced by *Nocardia cyriacigeorgica* GUH-2.

Precursor mass is marked with a blue diamond.

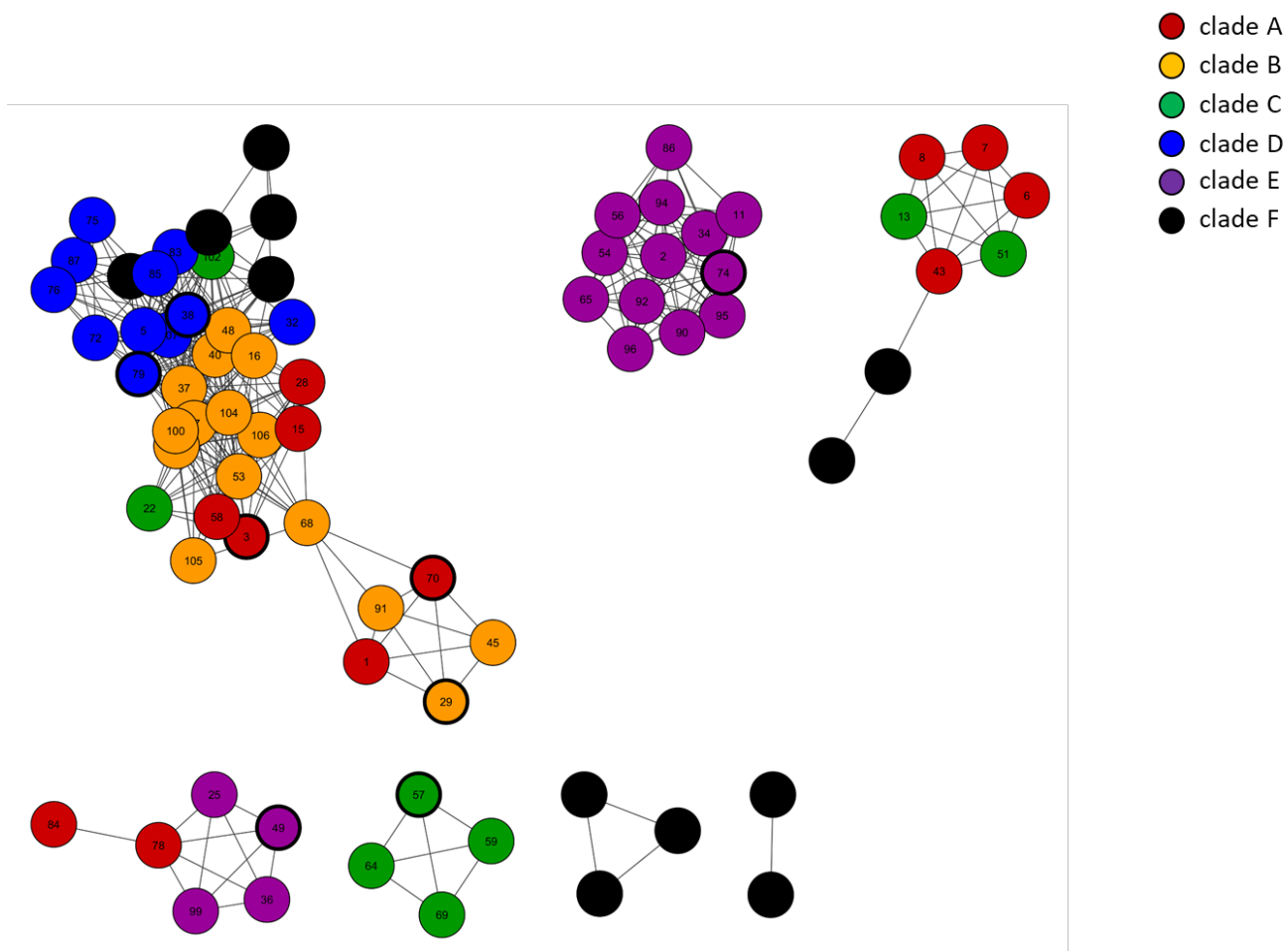
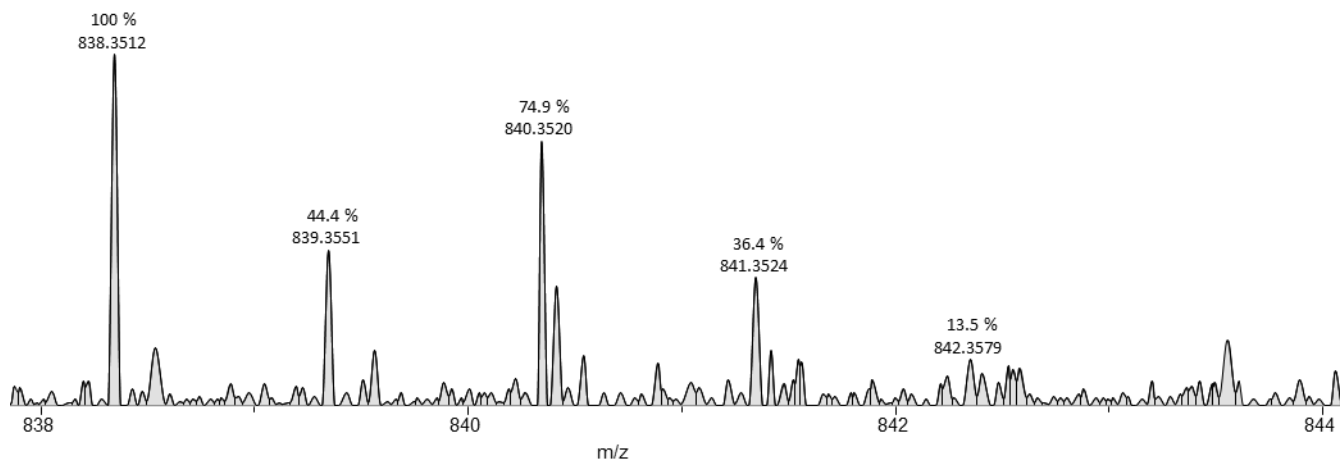


Figure S 22: BiG-SCAPE cluster-type sequence similarity network of all *nocobactin*-type BGCs from *Nocardia* strains found with antiSMASH 4.0.

A BiG-SCAPE threshold of 70 % is shown without singletons. Nocobactin-like BGCs and respective cluster parts II (*nbtS/nbtT*) were trimmed manually and concatenated. Bold circles highlighted show strains cultivated for metabolomics studies. Numbers represent strain identification numbers. Filling colors represent phylogenetic clades A-F.

(1) Isotopic peak analysis of 838.3512 [M-2H+Ga]⁺ non-supplemented cultures



(2) Isotopic peak analysis of 839.3540 [M-2H+Ga]⁺ from L-serine-1-¹³C-supplemented cultures

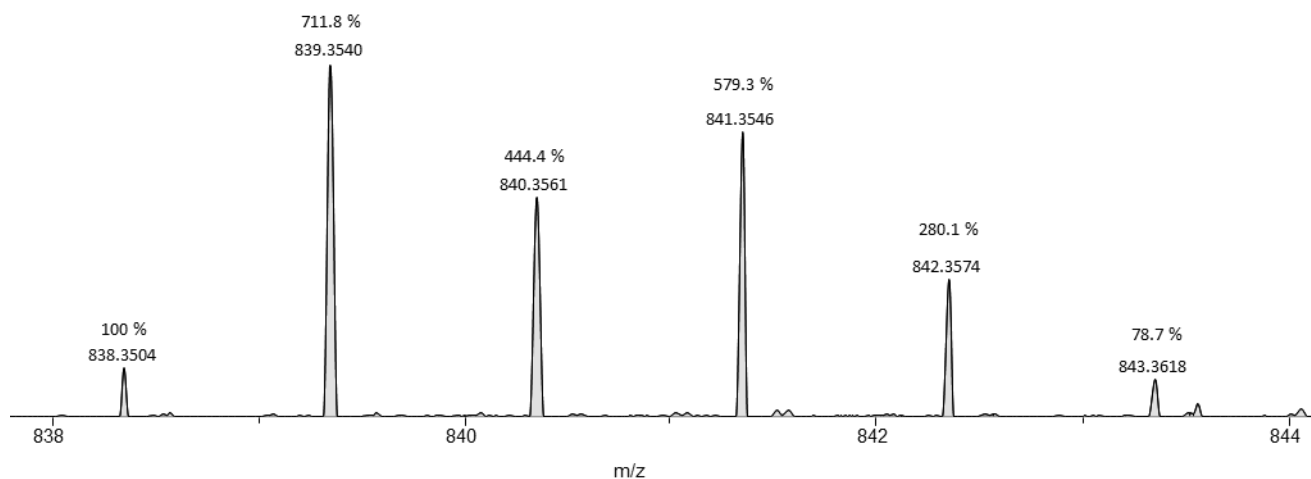
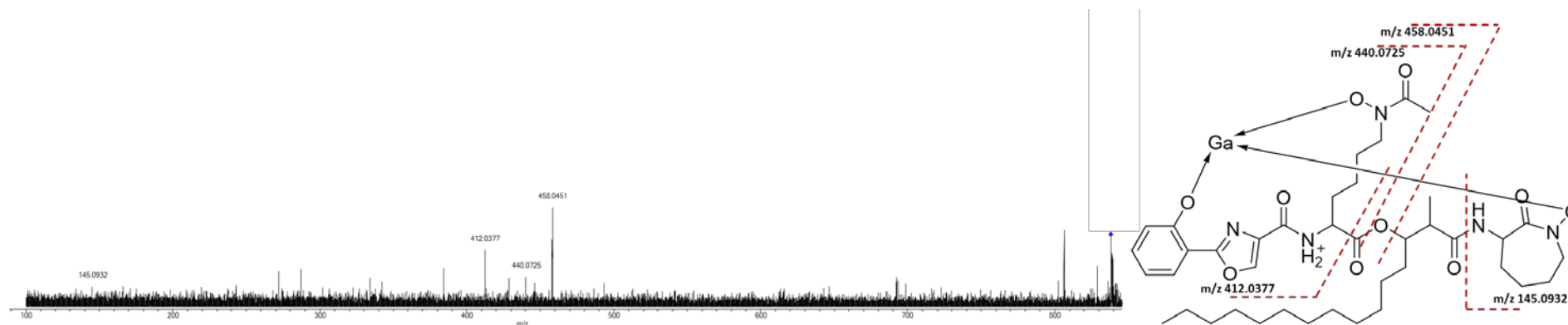


Figure S 23: Isotopic peak analysis of unlabeled and ¹³C-labeled nocobactin-like siderophore.

Isotopic peak analysis of *Nocardia alba* NBRC 108234 extracts; (1) unfed *Nocardia alba* produces unlabeled nocobactin-like siderophore 838.3512 [M-2H+Ga]⁺; feeding *Nocardia alba* with ¹³C₁-labeled L-serine results in m/z shifts of 839.3540 [M-2H+Ga+1]⁺, 840.3561 [M-2H+Ga+2]⁺, 841.3546 [M-2H+Ga+3]⁺, 842.3574 [M-2H+Ga+4]⁺, 843.3618 [M-2H+Ga+5]⁺.

(1) MS/MS analysis of an unlabeled nocobactin-like siderophore; precursor ion m/z 838.3608 $[M-2H+Ga]^+$



(2) MS/MS analysis of a L-serine-1- ^{13}C labeled nocobactin-like siderophore; precursor mass ion m/z 839.5489 $[M-2H+Ga+1]^+$

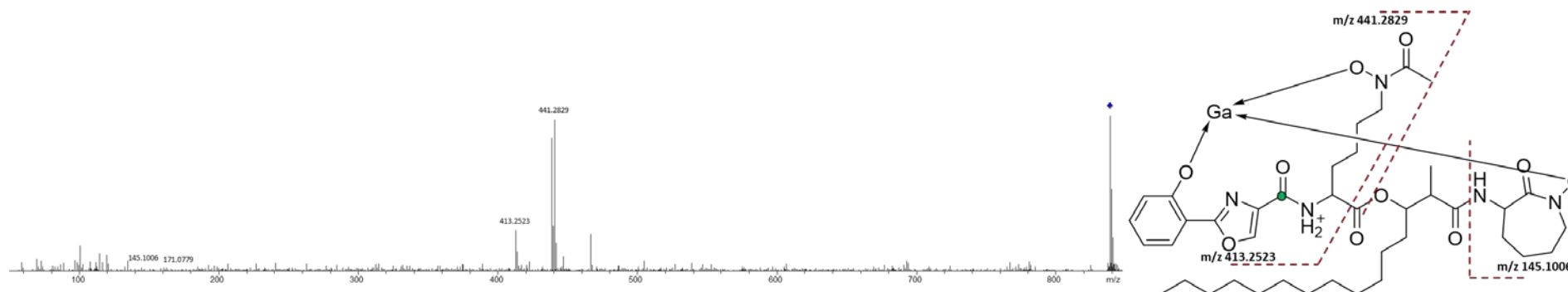


Figure S 24: MS/MS analysis of a ^{13}C -labeled/ unlabeled nocobactin-like siderophore from *Nocardia alba* NBRC 108234.

Labeled ^{13}C is marked in green. Precursor mass is marked with a blue diamond.

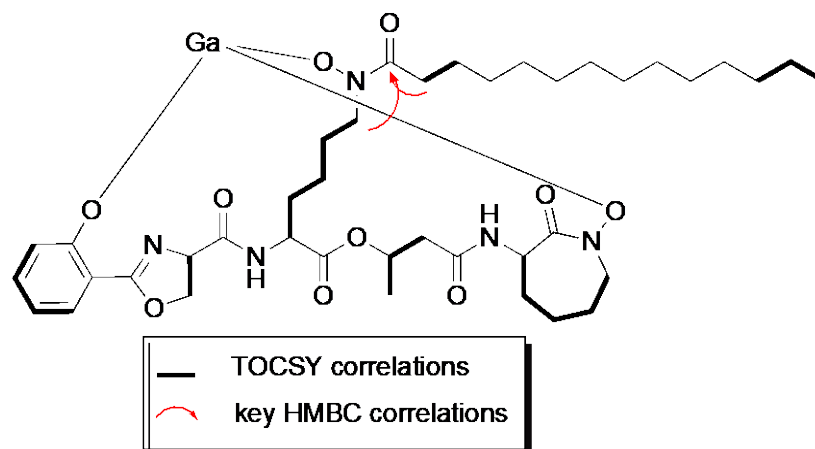


Figure S 25: TOCSY correlations and key HMBC correlations of a mycobactin-type siderophore with m/z 840.3669 $[M+Ga-2H]^+$.

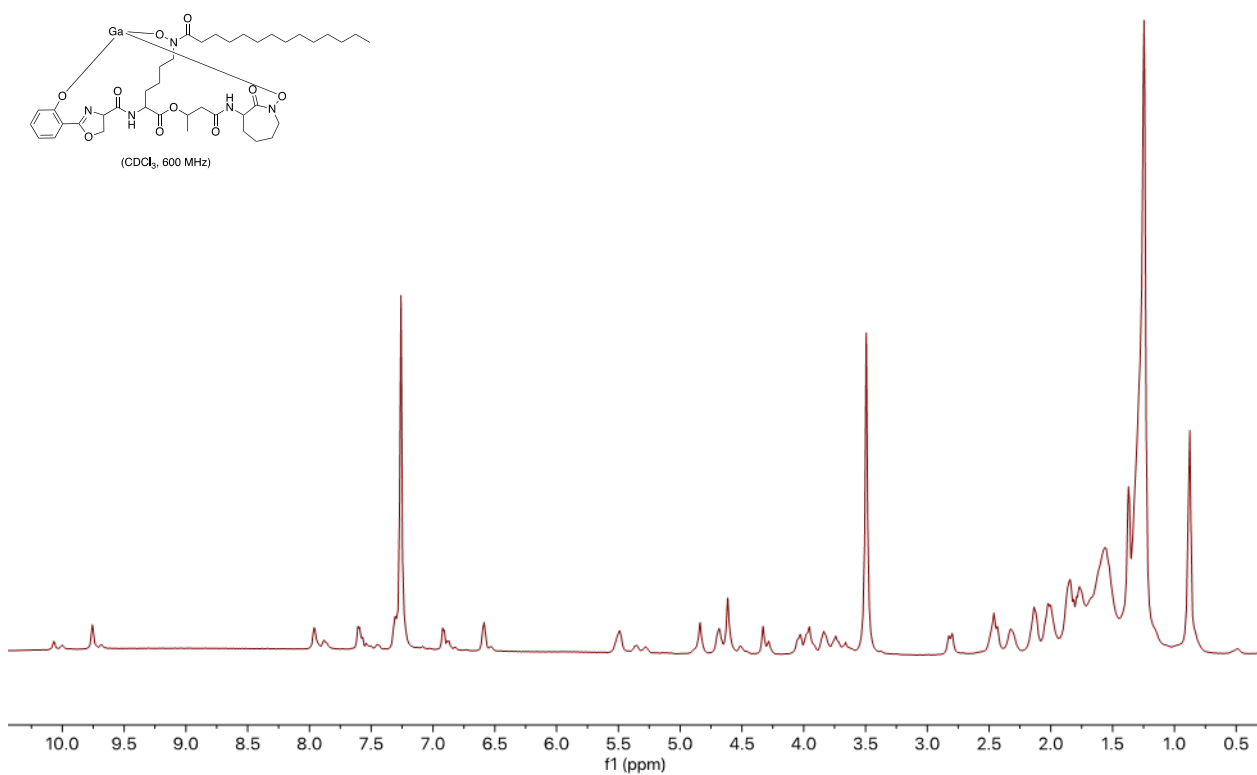


Figure S 26: 1H NMR spectrum of a gallium-bound mycobactin-like siderophore with m/z 840.3669 $[M+Ga-2H]^+$ from *Nocardia araoensis* NBRC 100135 in $CDCl_3$ (600 MHz).

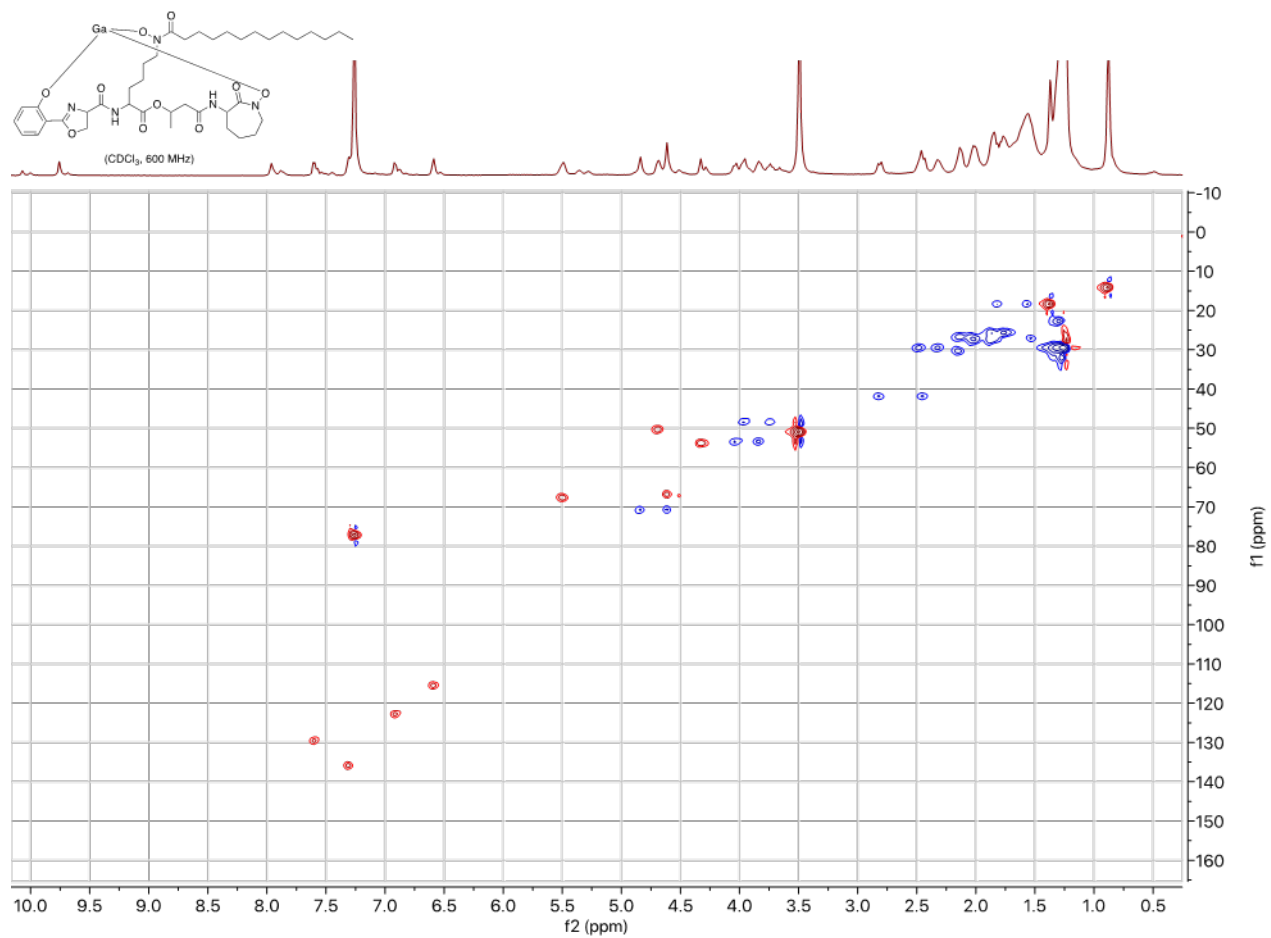


Figure S 27: ^1H - ^{13}C multiplicity edited HSQC NMR spectrum of gallium-bound mycobactin-like siderophore from *Nocardia araoensis* NBRC 100135 in CDCl_3 (600 MHz) with m/z 840.3669 $[\text{M}+\text{Ga}-2\text{H}]^+$.

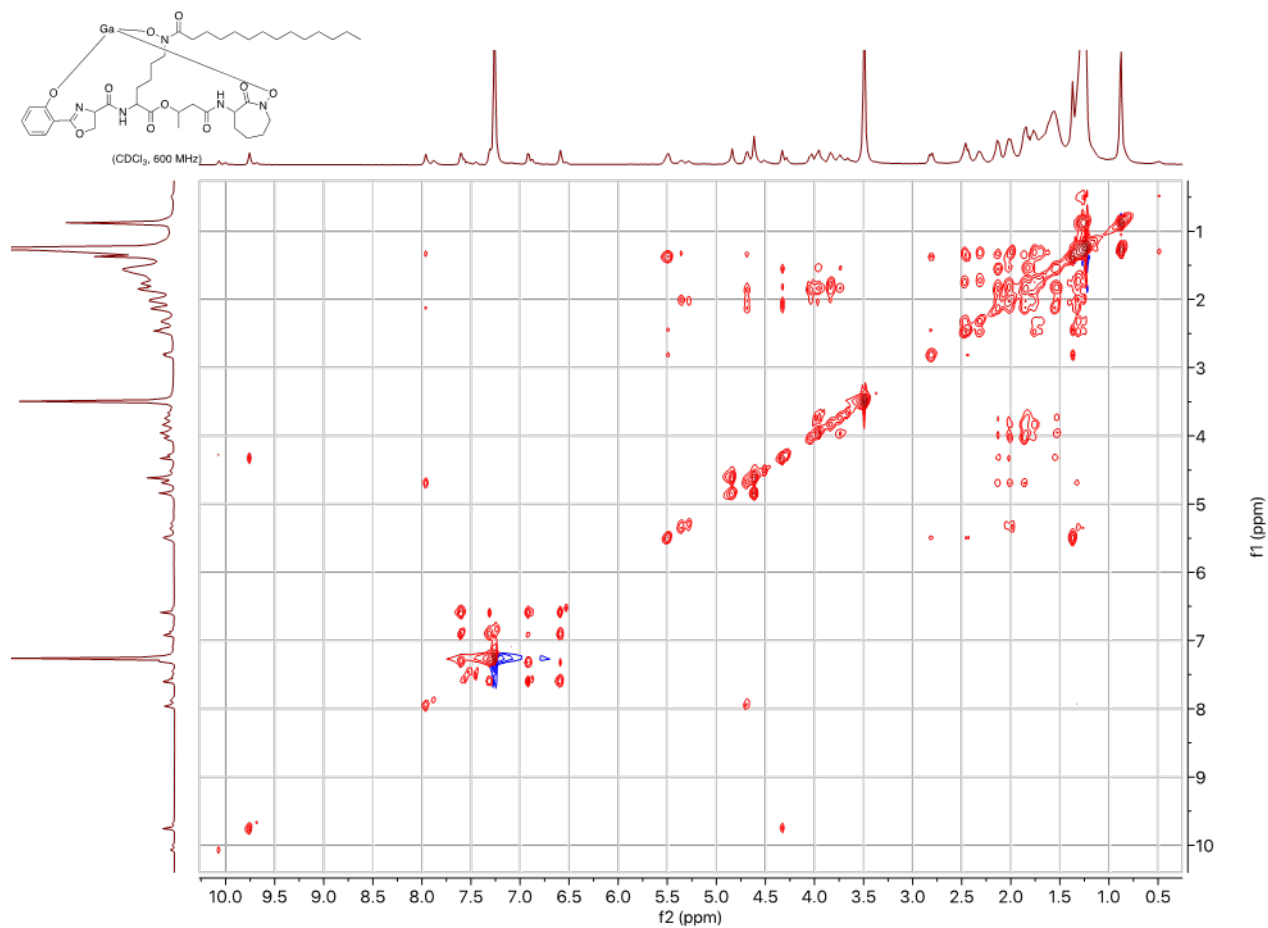


Figure S 28: ^1H - ^1H TOCSY NMR spectrum of gallium-bound mycobactin-like siderophore from *Nocardia araoensis* NBRC 100135 in CDCl_3 (600 MHz) with m/z 840.3669 $[\text{M}+\text{Ga}-2\text{H}]^+$.

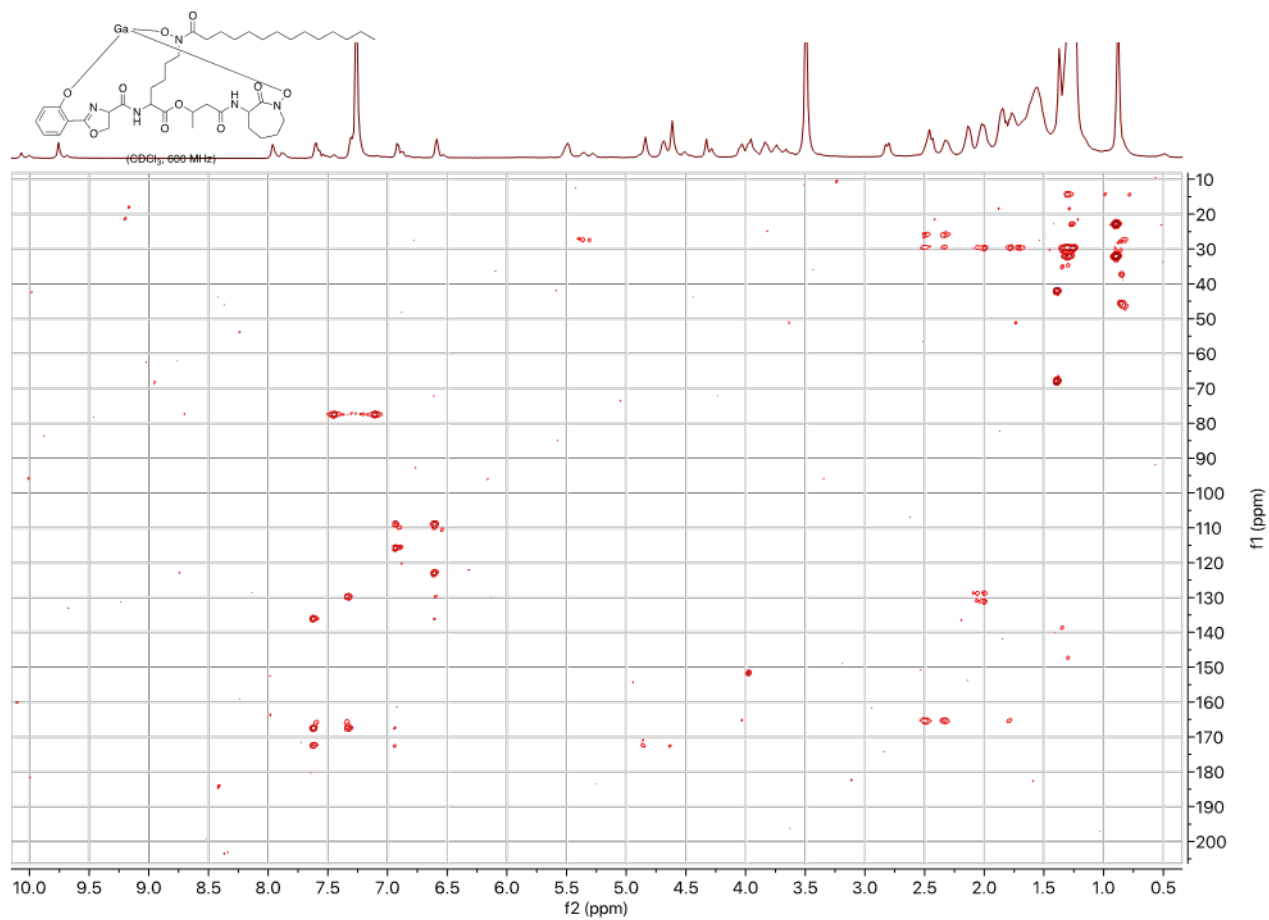


Figure S 29: ^1H - ^{13}C HMBC NMR spectrum of gallium-bound mycobactin-like siderophore from *Nocardia araoensis* NBRC 100135 in CDCl_3 (600 MHz) with m/z 840.3669 $[\text{M}+\text{Ga}-2\text{H}]^+$.

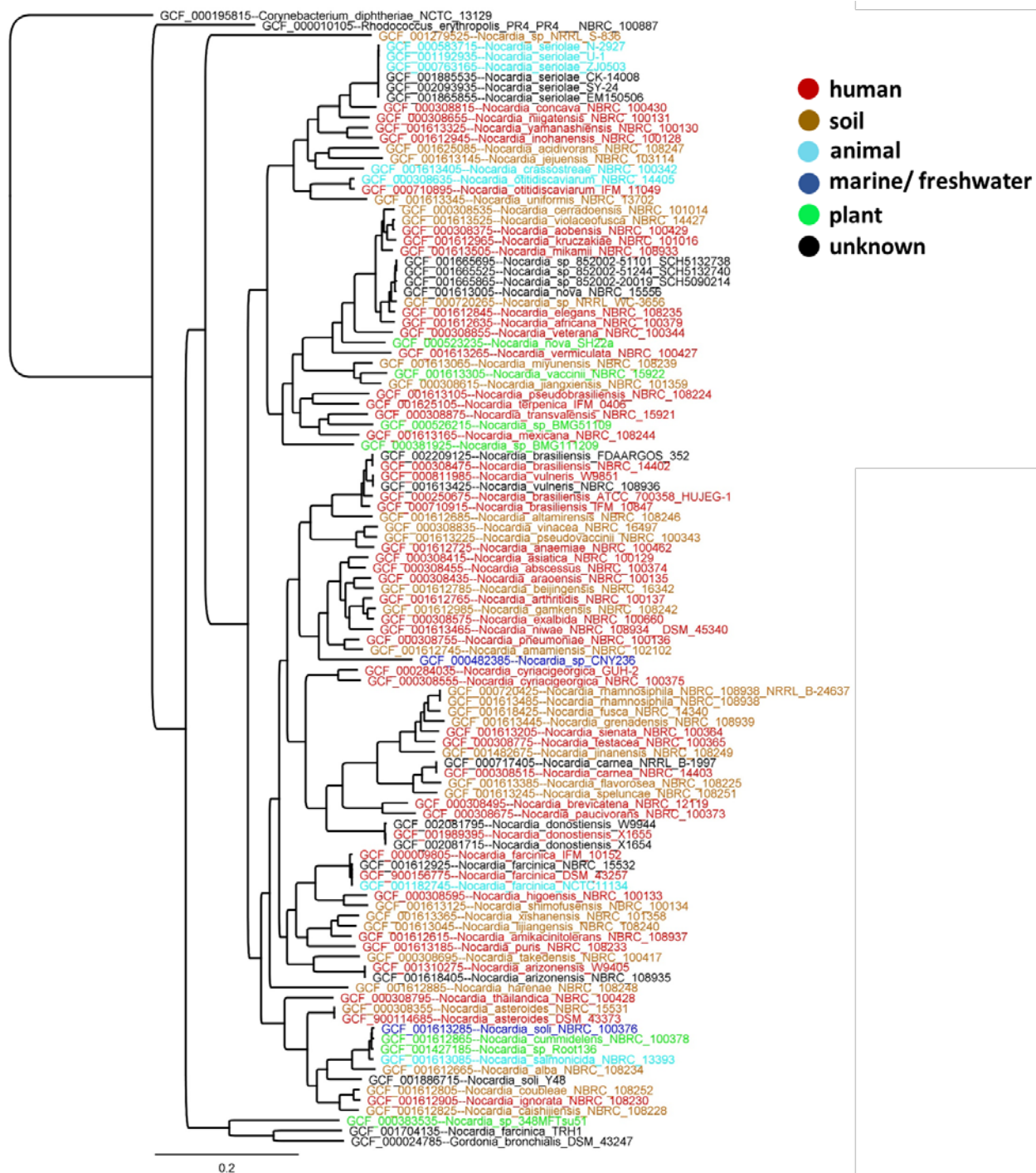


Figure S 30: *Nocardia* autoMLST tree. Colors indicate *Nocardia* spp. source of isolation. Information regarding isolation source was taken from respective published genomes.

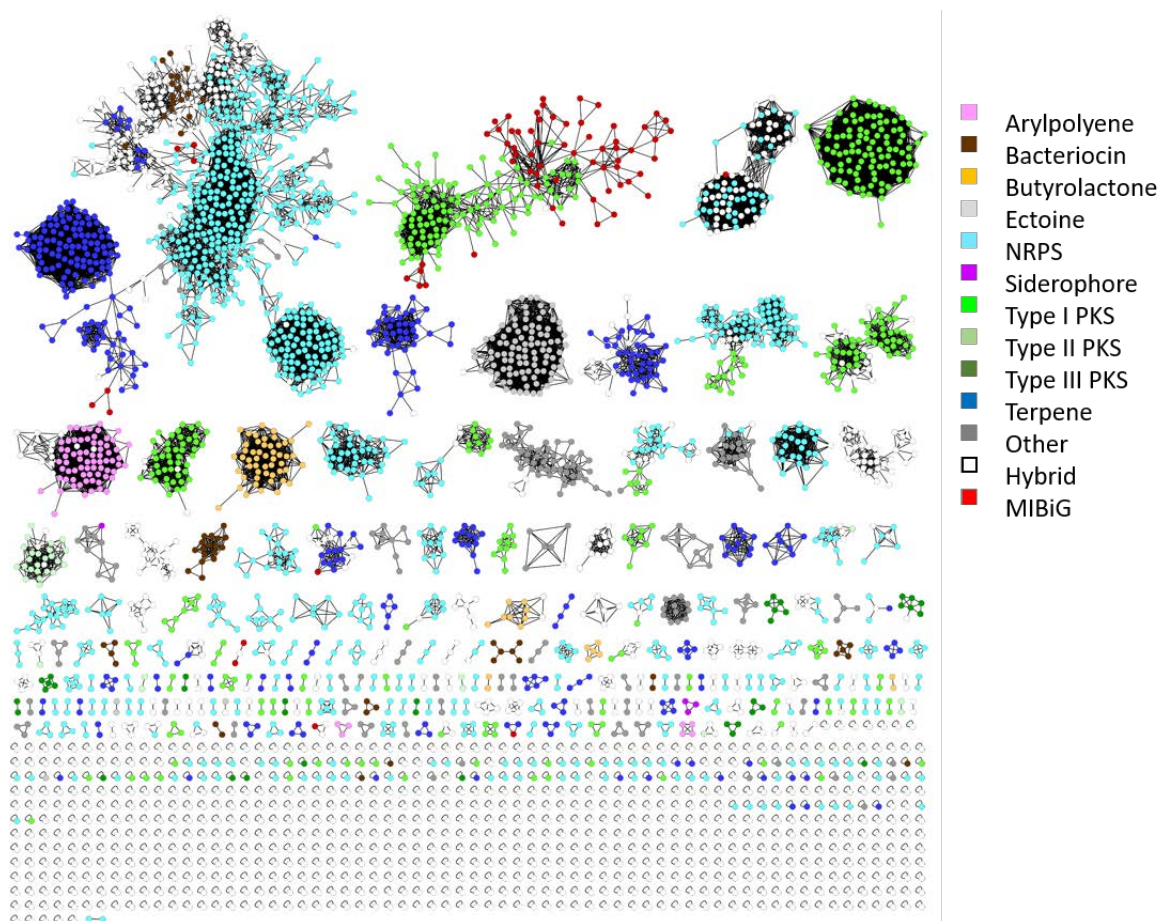


Figure S 31: BiG-SCAPE cluster type sequence similarity network of all BGCs from *Nocardia* strains found with antiSMASH 4.0 including the MIBiG database.

Colors show cluster types from *Nocardia* BGCs. Colorless nodes are hybrid BGCs found in *Nocardia*. Bold node borders refer to group 1 pathogens. MIBiG BGCs (version 1.4) were hidden when no *Nocardia* BGC showed a similarity.

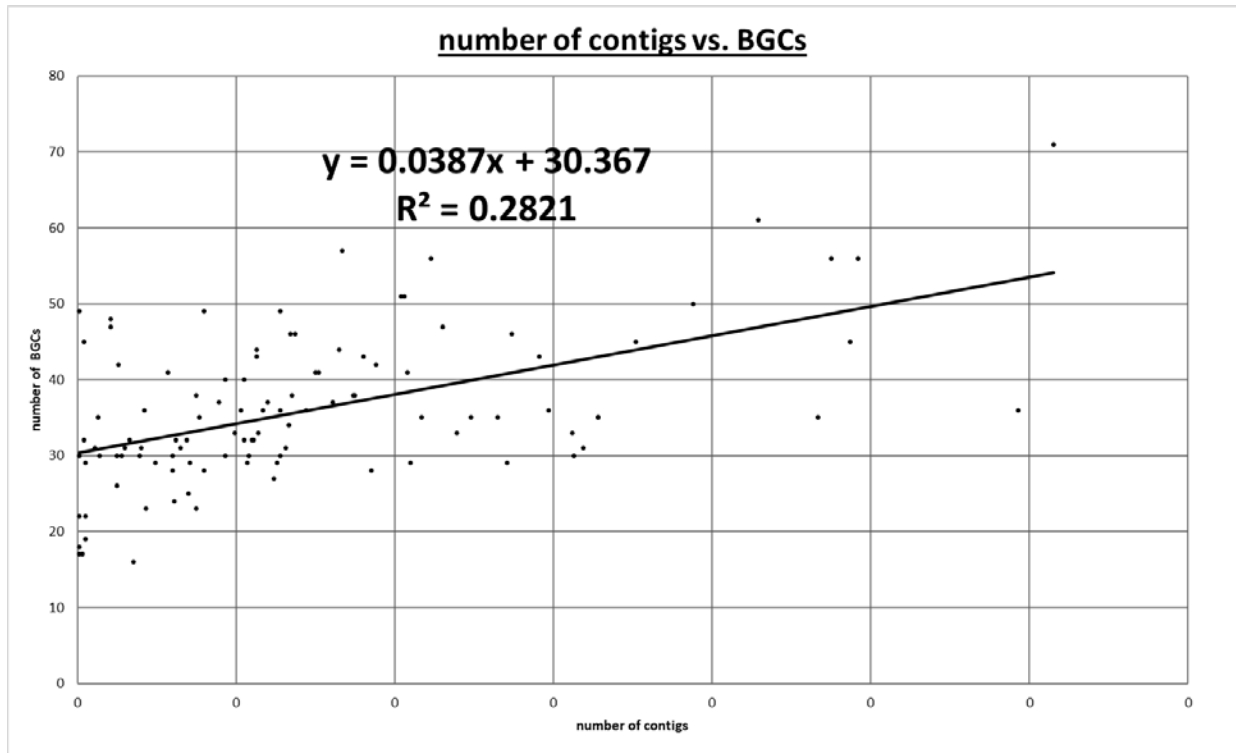


Figure S 32: Number of contigs plotted against the number of predicted BGCs in *Nocardia* strains. Each dot corresponds to one *Nocardia* strain. Trendline indicates no correlation between number of BGCs present in a strain and its genome size (coefficient of determination ($R^2=0.2821$)).

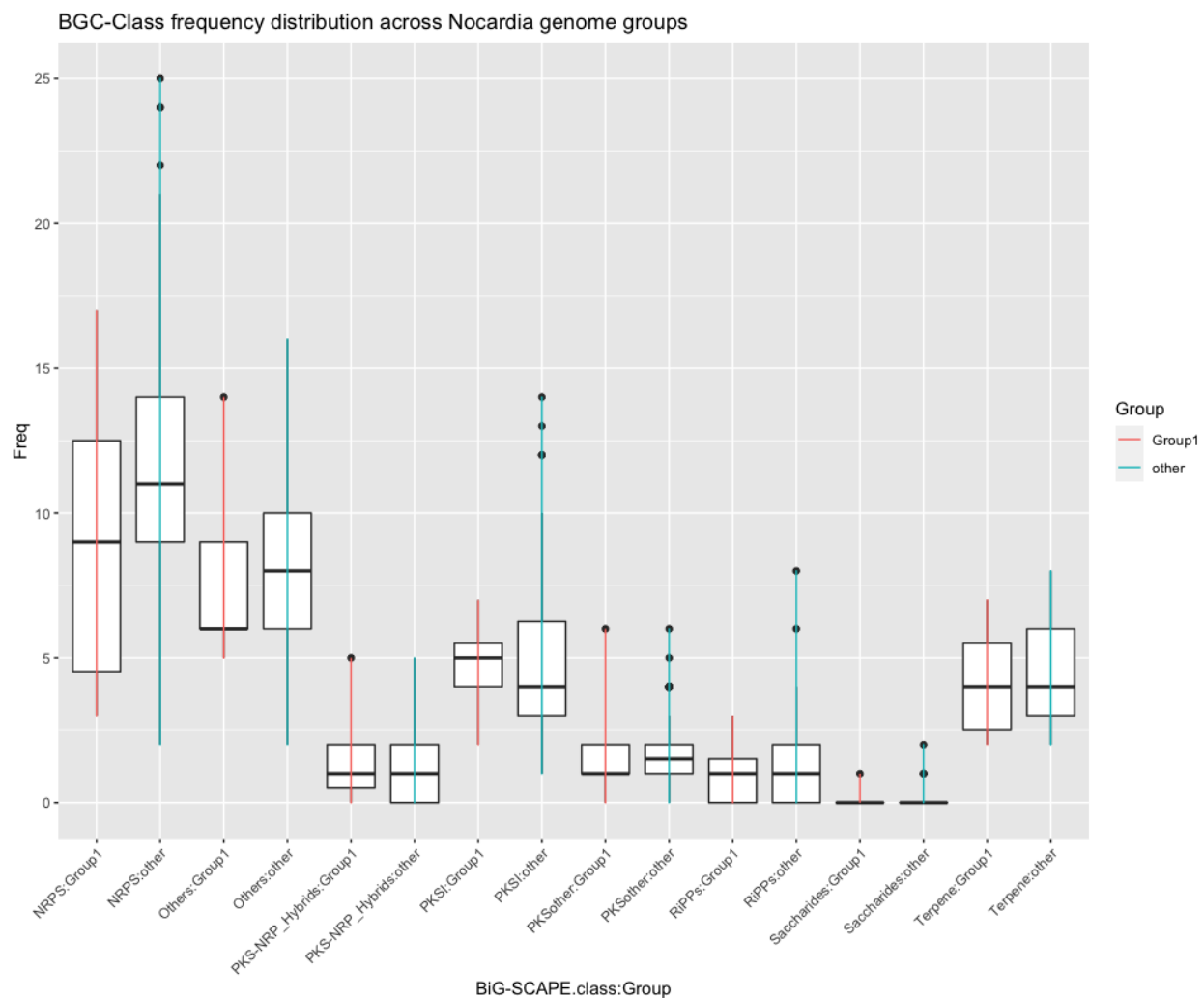


Figure S 33: Boxplot of BGC class frequency distribution across *Nocardia* genome groups.

Group1: Group 1 pathogenic *Nocardia*. Other: “other” *Nocardia* in analysis not in group 1 pathogens. X-axis represents BiG-SCAPE groups across *Nocardia* natural product classes. Y-axis represents total numbers of BGC-class frequencies. Black bars within boxplots show median BGC-class frequencies. Black dots represent outliers. Red/blue lines show whiskers of respective groups. Box edges represent lower and upper quartile. For identifying the group 1 individual genome specific clusters, a raw distance (cutoff = 0.75) was used as given in BiG-SCAPE network matrices.

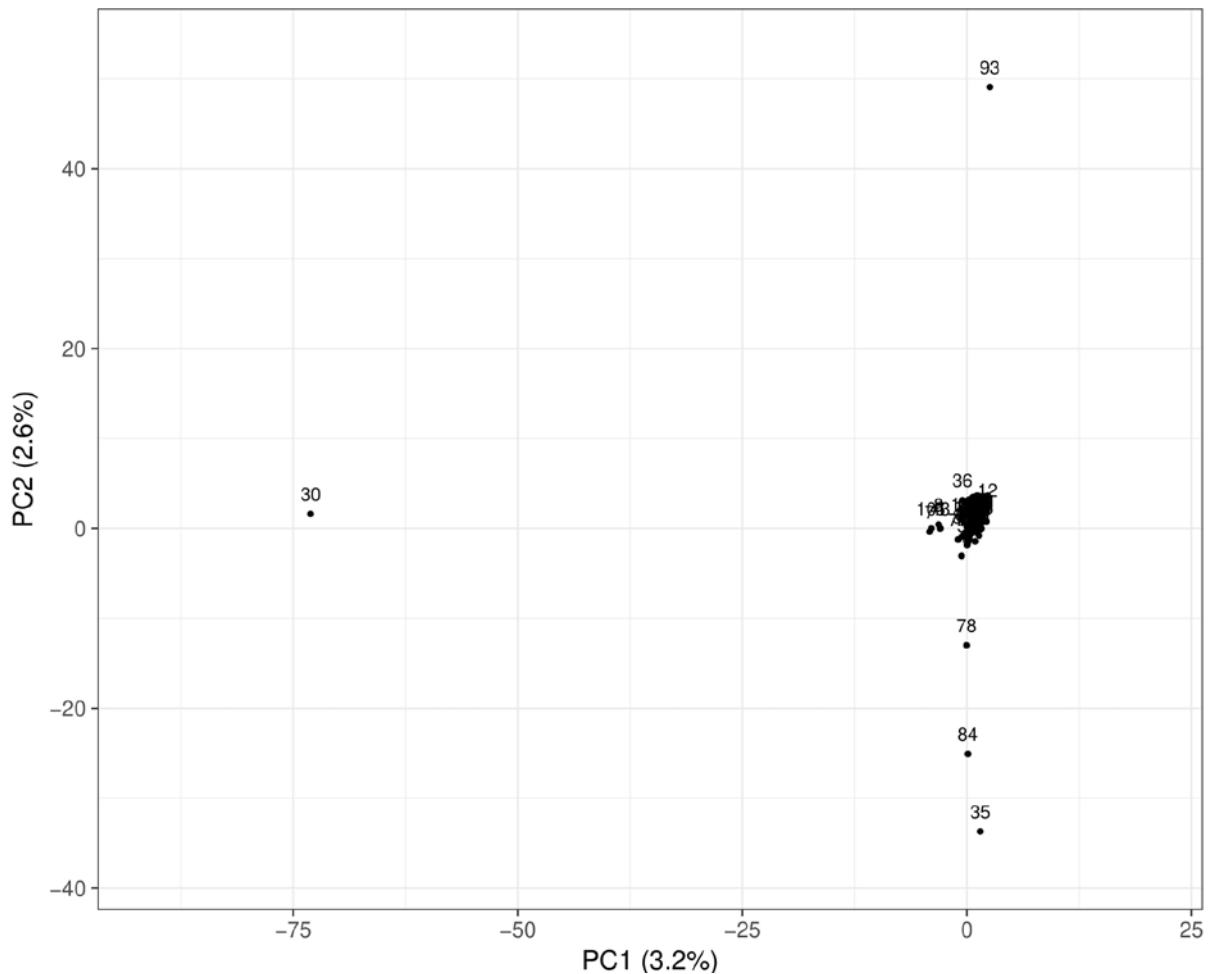


Figure S 34: PCA-analysis on GCF presence/absence.

GCF presence/absence information for clustering the *Nocardia* genomes into separate groups based on clinical annotations (*Nocardia* group 1 pathogens: 1 = *Nocardia abscessus* NBRC 100374; 5 = *Nocardia asteroides* NBRC 15531; 6 = *Nocardia brasiliensis* ATCC 700358; 13 = *Nocardia cyriacigeorgica* GUH-2; 16 = *Nocardia farcinica* IFM 10152; 21 = *Nocardia otitidiscaviarum* NBRC 14405; 47 = *Nocardia nova* SH22a (NONO); “other” *Nocardia* genomes not in group). X and Y axis show principal component 1 (PC1) and principal component 2 (PC2) that explain 3.2 % and 2.6 % of the total variance, respectively. Sample size = 107 data points. Genome analysis IDs are annotated alongside the data points. The PCA plot was performed using ClustVis web tool.

References

1. Hoshino, Y.; Chiba, K.; Ishino, K.; Fukai, T.; Igarashi, Y.; Yazawa, K.; Mikami, Y.; Ishikawa, J., Identification of Nocobactin NA Biosynthetic Gene Clusters in *Nocardia farcinica*. *J Bacteriol* **2011**, *193*, 441-8.
2. Vera-Cabrera, L.; Ortiz-Lopez, R.; Elizondo-Gonzalez, R.; Ocampo-Candiani, J., Complete Genome Sequence Analysis of *Nocardia brasiliensis* HUJEG-1 Reveals a Saprobic Lifestyle and the Genes Needed for Human Pathogenesis. *PLoS One* **2013**, *8*.



New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A–C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406

Julia Chen,^[a, b] Andri Frediansyah,^[a, b, c] Daniel Männle,^[a, b, d] Jan Straetener,^[b, e] Heike Brötz-Oesterhelt,^[b, e] Nadine Ziemert,^[b, d] Leonard Kaysser,^[a, b] and Harald Gross^{*[a, b]}

We report a genomics-guided exploration of the metabolic potential of the brasilicardin producer strain *Nocardia terpenica* IFM 0406. Bioinformatics analysis of the whole genome sequence revealed the presence of a biosynthetic gene cluster presumably responsible for the generation of formerly unknown nocobactin derivatives. Mass spectrometry-assisted isolation led

to the identification of three new siderophores, terpenibactins A (1), B (2) and C (3), which belong to the class of nocobactins. Their structures were elucidated by employing spectroscopic techniques. Compounds 1–3 demonstrated inhibitory activity towards the muscarinic M3 receptor, while exhibiting only a low cytotoxicity.

Introduction

In the past, bacteria of the genus *Nocardia* were principally associated with a pathogenic context, having been described in 1888 by Edwin Nocard^[1] as the causative agent of severe infections of the skin, lungs or the central nervous system in humans and animals.^[2,3] However, within the last two decades, it has become apparent that this genus also possesses a tremendous biotechnological and pharmaceutical potential, with species that produce numerous versatile enzymes^[4–6] and structurally novel and highly bioactive small molecules.^[7,8] This

observation is also supported by analyses of the ever-increasing amount of publicly available *Nocardia* genome sequencing data.^[9–12] Commonly, *Nocardia* spp. genomes comprise 16 to 49 biosynthetic gene clusters, which is why *Nocardia* spp. are considered 'biosynthetically talented' producers.

A prime example of such is the strain *Nocardia terpenica* IFM 0406 (formerly referred to as *Nocardia brasiliensis* IFM 0406).^[13] It is known to produce the anti-staphylococcal lipolanthine nocavionin,^[14] the antifungal brasilinolides^[15–17] and the immunosuppressive brasilicardins.^[18–21] In order to develop a safe biotechnological production platform for brasilicardins, we recently revisited this strain and reinvestigated the corresponding biosynthetic gene cluster, reported by Dairi and co-workers.^[22] Within the framework of this study, we sequenced the whole genome of IFM 0406,^[23] were able to redefine the borders of the gene cluster and achieved the heterologous expression of brasilicardin C.^[24] Further analysis of the whole genome sequence indicated a gene cluster encoding a new nocobactin-like compound.

Structurally, nocobactins feature a C-terminal *N*-OH-cyclo-lysine bound to a long chain 3-OH fatty acid, whose hydroxy group is esterified by *N*^ε-acyl-*N*^ε-hydroxy-L-lysine. The α -amino group of lysine is in turn linked to a 2-OH-phenyl-5-methyl-oxazole moiety. To date, a series of differently substituted congeners has been isolated from *Nocardia* species,^[25–30] showing variation in the substitution pattern of the 2-OH-phenyl residue, of the oxazole ring system and of the ϵ -nitrogen atoms of both lysine residues, or in the nature of the fatty acid (Table 1).

In this study, we describe the *in silico* identification of a new nocobactin biosynthetic gene cluster. Subsequently, the isolation, structure elucidation and biological evaluation of the resultant compounds, which we termed terpenibactins A–C, is presented.

[a] J. Chen, A. Frediansyah, D. Männle, Dr. L. Kaysser, Prof. Dr. H. Gross
Pharmaceutical Institute, Dept. of Pharmaceutical Biology
University of Tübingen
Auf der Morgenstelle 8, 72076 Tübingen (Germany)
E-mail: harald.gross@uni-tuebingen.de

[b] J. Chen, A. Frediansyah, D. Männle, J. Straetener, Prof. Dr. H. Brötz-
Oesterhelt, Prof. Dr. N. Ziemert, Dr. L. Kaysser, Prof. Dr. H. Gross
German Center for Infection Research (DZIF)
Partner site Tübingen, 72076 Tübingen (Germany)

[c] A. Frediansyah
Research Division for Natural Product Technology (BPTBA)
Indonesian Institute of Sciences (LIPI), Wonosari 55861 (Indonesia)

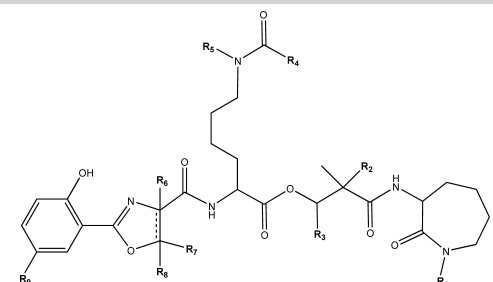
[d] D. Männle, Prof. Dr. N. Ziemert
Department of Applied Natural Products Genome Mining
Interfaculty Institute of Microbiology and Infection Medicine (IMIT)
University of Tübingen
Auf der Morgenstelle 28, 72076 Tübingen (Germany)

[e] J. Straetener, Prof. Dr. H. Brötz-Oesterhelt
Department of Microbial Bioactive Compounds
Interfaculty Institute of Microbiology and Infection Medicine (IMIT)
University of Tübingen
Auf der Morgenstelle 28, 72076 Tübingen (Germany)

Supporting information for this article is available on the WWW under
<https://doi.org/10.1002/cbic.202000062>

This article is part of a Special Collection on Microbial Biosynthesis and
Interactions. To view the complete collection, visit our homepage

© 2020 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA. This
is an open access article under the terms of the Creative Commons Attri-
bution Non-Commercial NoDerivs License, which permits use and distribu-
tion in any medium, provided the original work is properly cited, the use is
non-commercial and no modifications or adaptations are made.

Table 1. The nocobactin compound family.


Compound	R ¹	R ²	R ³	R ⁴	R ⁵	R ⁶	R ⁷	R ⁸	R ⁹	<i>m/z</i>	Ref.
nocobactin NA-a	OH	H	C ₉ H ₁₉	CH ₃	OH	–	CH ₃	–	H	743.4	[25]
nocobactin NA-b	OH	H	C ₁₁ H ₂₃	CH ₃	OH	–	CH ₃	–	H	771.4	[25]
formobactin	OH	CH ₃	C ₉ H ₁₉	H	OH	–	CH ₃	–	H	743.4	[26]
amamistatin A	OH	CH ₃	C ₇ H ₁₅	H	OH	–	CH ₃	–	OCH ₃	745.4	[27]
amamistatin B	H	CH ₃	C ₇ H ₁₅	H	H	–	CH ₃	–	OCH ₃	713.4	[27]
brasilibactin A	OH	H	C ₁₅ H ₃₁	H	OH	H	H	H	H	801.5	[28]
nocardimicin A	OH	H	C ₉ H ₁₉	CH ₃	OH	–	H	–	H	729.4	[29]
nocardimicin B	OH	H	C ₁₁ H ₂₃	CH ₃	OH	–	H	–	H	757.4	[29]
nocardimicin C	H	H	C ₁₁ H ₂₃	CH ₃	OH	–	H	–	H	741.4	[29]
nocardimicin D	OH	H	C ₁₃ H ₂₇	CH ₃	OH	–	H	–	H	785.5	[29]
nocardimicin E	H	H	C ₁₃ H ₂₇	CH ₃	OH	–	H	–	H	769.5	[29]
nocardimicin F	OH	H	C ₁₅ H ₃₁	CH ₃	OH	–	H	–	H	813.5	[29]
nocardimicin G	OH	H	C ₁₃ H ₂₇	H	OH	H	H	H	H	773.5	[30]
nocardimicin H	OH	H	C ₁₅ H ₃₁	H	OH	H	H	H	H	801.5	[30]
nocardimicin I	OH	H	C ₁₇ H ₃₅	H	OH	H	H	H	H	829.5	[30]
predicted cpd.	OH	H	C ₇ H ₁₅ –C ₁₇ H ₃₅	H	OH	H	CH ₃	H	H	703–844	this study

Results and Discussion

Identification and organization of the terpenibactin (*ter*) biosynthetic gene cluster

In silico whole genome sequence analysis of *N. terpenica* IFM 0406 using the bioinformatics web-tool antiSMASH 5.1.0^[31] revealed the presence of 38 biosynthetic gene clusters (BGCs). Besides the readily identifiable known BGCs coding for nocavionin,^[14] brasilinolide,^[17] brasilicardin,^[23,24] and ectoine,^[32] two NRPS-PKS hybrid-based BGCs, both located on contig_9 (accession no. LWGR01000021.1), were postulated to direct the biosynthesis of a nocobactin/mycobactin-like siderophore. Closer inspection of this BGC showed that it extensively parallels the nocobactin NA cluster.^[33] Both share nine homologues genes out of ten genes in total and both are divided into two sub-gene clusters (sub-gene cluster 1 and 2). In the case of the *ter* gene cluster, the inter-subcluster region spans 282 kb (Figure 1). Slight differences exist in the fact that the NRPS *nbtF* homologue *terF* is located in sub-gene cluster 2 as opposed to sub-gene cluster 1, and instead of a lysine acetyltransferase similar to NbtH, the corresponding gene *terH* codes for a lysine formyltransferase that is located at the equivalent position in cluster one (Figure 1).

Having identified the split BGC, we predicted the resultant biosynthetic route.^[33] Salicylate is synthesised from chorismate by the salicylate synthase TerS and activated by the salicylate-AMP ligase TerT, being tethered to the aryl carrier protein (ArCP) domain on the nonribosomal peptide synthetase (NRPS) TerF (Figure 2).

Here, threonine is also activated and cyclised, giving either 2-hydroxyphenyl-5-methyloxazoline or 2-hydroxyphenyl-5-methyloxazole. Recently, the candidate gene *sagB*, encoding a dehydrogenase, was recognised to be responsible for the oxidative function necessary to transform an oxazoline into an oxazole (personal communication D.M, N.Z.). Since, neither the related BGC nor the genetic environment surrounding contained a homologue of *sagB*, we predicted that terpenibactins bear a 2-hydroxyphenyl-5-methyloxazoline residue. Subsequently, the peptide backbone is extended through the condensation of *N*^ε-formyl-*N*^ε-hydroxy-L-lysine by the NRPS TerD. Since Thomas and co-workers demonstrated that the A domain of a close homologue of TerD is specific for *N*^ε-acyl-*N*^ε-hydroxy-L-lysine,^[34] the tailoring of the lysine residue by the lysine-*N*-oxygenase TerG and the lysine formyltransferase TerH occurs prior to their insertion into the nocobactin skeleton. The additional C domain within the NRPS *terD* is predicted to condense a 3-hydroxy-2-methyl fatty acid unit into the growing skeleton. The polyketide synthases TerB and TerC are expected to work in conjunction to incorporate the fatty acid into the biosynthesis. Due to the fact that fatty acids of this length cannot be synthesised by a single type I PKS module, it is hypothesised that TerC can select and carry long-chain fatty acids.^[33] Since the biosynthesis utilises fatty acids from primary metabolism, it is typically challenging to predict the lengths of the acyl chains on the basis of information gleaned from bioinformatics alone; however, considering the range of the nocobactin compound family (Table 1), we assumed the precedence of fatty acid chains between C₇–C₁₇ units long. Finally, the NRPS TerE then serves to epimerise and condense a

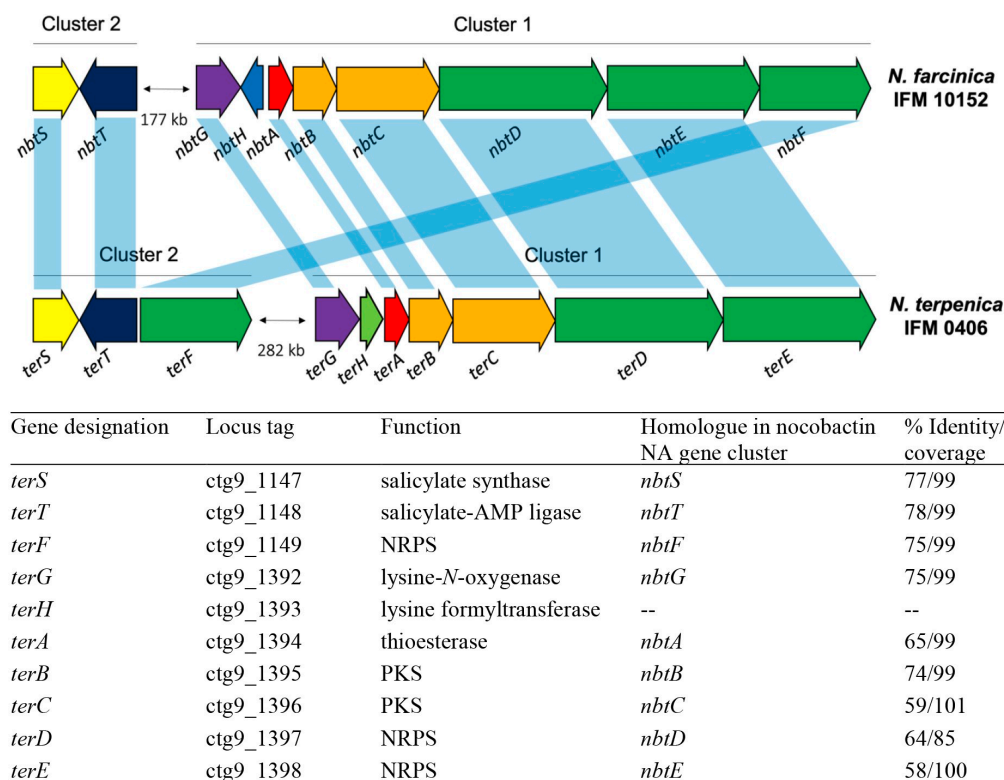


Figure 1. Comparison of the biosynthetic gene clusters encoding (top) nocobactin NA and (bottom) novel terpenibactins.

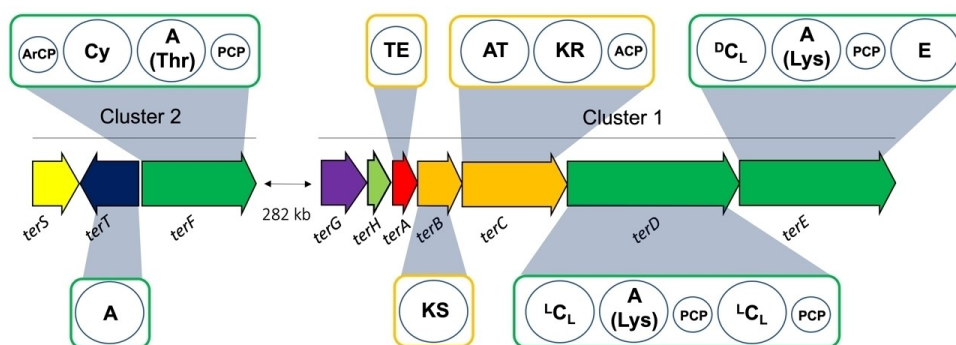


Figure 2. Detailed view of PKS (orange boxes) and NRPS (green boxes) modules in the terpenibactin gene cluster of *N. terpenica* IFM 0406, including A domain substrate specificities.

N-hydroxy-lysine residue, which is subsequently lactamised, most likely in a non-enzymatic way.^[35]

In conclusion, the newly identified gene cluster in strain IFM 0406 was predicted to produce nocobactin-like compounds with $R^1=R^5=OH$, $R^2=R^4=R^6=R^8=R^9=H$, $R^7=CH_3$ and R^3 ranging from C_7H_{15} to $C_{17}H_{35}$. With these substituents, the resultant compounds should give rise to a suggested mass between 703 and 844 Da. The predicted compounds most closely resemble brasilibactin A or nocardimicins G–I (Table 1), but bear a methyl-oxazoline instead of an oxazoline ring system. The finding that *terF* codes for a threonine-specific A domain that leads, in combination with the absence of *sagB*, to the formation of a methyl-oxazoline ring system, was pivotal.

This fact enabled us to conclude that, regardless of the length of the lipid side chain, the *ter* BGC had to give rise to production of one or more new nocobactin-like compound(s).

LC/MS screening for the products of the terpenibactin (*ter*) gene cluster

Methanol extracts of *N. terpenica* IFM 0406 cell pellets were tested for the presence of nocobactin derivatives. A similar strategy was adopted to that used in the isolation of nocobactin NA and similar compounds.^[25–30] It was intended that extracts would be monitored for compounds in the mass range of 703–

844 Da by use of low-resolution (LR)-LC/MS. The use of chrome azurol S (CAS) assays^[36] to detect the siderophore of interest proved inapplicable in this isolation effort, given that control extracts of rich, orange-coloured production medium generated false-positive CAS results. Foreseeing the possibility that any secreted siderophores might be complexed with iron at the point of analysis, and knowing from previous literature that ferric nocobactin NA has an absorbance maximum at around 470 nm, extracts were also monitored at this wavelength in HPLC analysis, in addition to the expected maxima produced by the free ligand of 210, 260 and 310 nm.^[25]

Analysis of cell pellet methanol extracts via LR-LC/MS led to the detection of three peaks with absorbance at 460 nm and masses within the range anticipated for the sought-after compounds ($M=815.8$, 841.9 and 843.8 Da). Each compound was detected as their singly charged $[M+H]^+$ pseudomolecular ion along with their ferrated form $[M-3H+Fe^{3+}+H]^+$ and the sodium adduct $[M-3H+Fe^{3+}+Na]^+$ of the latter (see Figures S1 and S2 in the Supporting Information). In summary, the MS data did not match any known compound from the nocobactin family (Table 1) and therefore suggested that IFM 0406 produced at least three new siderophores, which were given the trivial names terpenibactin A (1), B (2) and C (3).

Isolation and structure elucidation of terpenibactins A–C

To obtain sufficient material to support structural characterization, a 1 L fermentation was performed. Fractionation and purification of the MeOH extract of the cell pellet was carried out by both reversed-phase (RP) vacuum liquid chromatography (VLC) and RP-HPLC. In this way, compounds 1–3 were obtained in a ferrated form (Figure S3). However, due to paramagnetic iron's interference with the magnetic field, which leads to peak broadening and poorly resolved NMR spectra, complete iron removal was necessary in order to perform structure elucidation. Since the method originally described by Ratledge and Snow for full decomplexation of ferric nocobactin^[25] proved to be inefficient in our hands, we converted the ferric forms of 1–3 into their corresponding

gallium salts and re-purified them by employing HPLC (Figures S4 and S5).^[37,38]

The analysis of the Ga^{III} adduct of terpenibactin A (1) by HR-ESI-MS determined its exact mass to be m/z 882.4156, consistent with the molecular formula $C_{43}H_{67}GaN_5O_{10}$ (Δ 1.4 ppm), indicating a structure with 12 degrees of unsaturation. The IR spectrum (Figure S6) possessed absorptions for ester (1738 cm^{-1}) and amide (1672 cm^{-1}) carbonyl groups and an aromatic ring system (1617, 1464, 857, 757, 699 cm^{-1}). UV maxima observed at 219 and 255 and 339 nm (Figure S7) were indicative of a substituted benzenoid chromophore with an extended conjugated system. In the ^{13}C and DEPT135 NMR spectra (Figures S10 and S11), signals assignable to 43 carbons were detected, including those for seven quaternary carbons, one *N*-formyl (δ 155.3, $^1J_{C,H}=209$ Hz), four sp^2 methines, six sp^3 methines, 22 methylenes and three methyl groups (Table 2). These data also revealed the presence of nine double bonds ($3\times CC$, $1\times CN$, $5\times CO$); thus, 1 was found to be tricyclic. Analysis of the $^1H,^1H$ COSY and $^1H,^{13}C$ HSQC-TOCSY spectrum resolved the partial structures A–E (Figure 3), which were connected by using $^1H,^{13}C$ HMBC and $^1H,^1H$ NOESY correlations (Figures S13–S16).

The $^1H,^1H$ connectivities from H-3 to H-6 and their 1H and ^{13}C chemical shifts in combination with the $^1H,^1H$ coupling pattern indicated the presence of a 1,2-disubstituted benzene ring. The chemical shift of C-2 (δ 167.5) and $^1H,^{13}C$ HMBC cross correlations between H-3/C-1, H-4/C-2 and H-5/C-1 showed that fragment A consisted of a benzene ring that was substituted with a phenolic hydroxyl group at C-2. The ^{13}C NMR chemical shifts for C-7, C-9, and C-10, together with the doublet proton resonance at δ 1.49 (H-13), as well as the long-range CH couplings observed between H-10/C-7, H-10/C-12 and H-9/C-12, were all indicative of the presence of a methyl-dihydrooxazole carboxylic acid moiety, which represents fragment B. A long-range CH coupling between H-6 and C-7 allowed the connectivity between C-1 to C-7, and thus that between fragment A and B, to give 9,10-dihydro-asteroidic acid. The typical 1H and ^{13}C chemical shifts, together with the connected spin system, identified fragment C as a lysine residue. $^1H,^{13}C$ HMBC long-range correlations between its α -proton and Lys1-

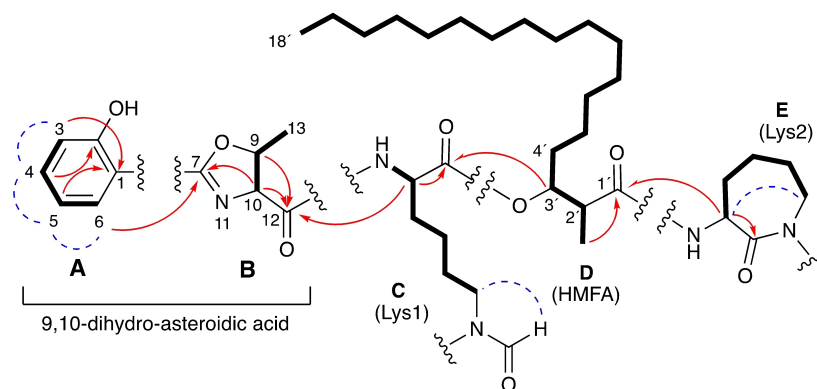


Figure 3. $^1H,^1H$ COSY and $^1H,^{13}C$ HSQC-TOCSY (bold lines), selected $^1H,^{13}C$ long-range (red arrows) and selected $^1H,^1H$ NOESY (dashed blue lines) correlations for terpenibactin A (1). HMFA: hydroxymethyl fatty acid.

Table 2. ¹ H and ¹³ C NMR spectroscopic data for terpenibactins A–C (1–3). ^[a]						
Atom	Terpenibactin A (1) ^[b]		Terpenibactin B (2) ^[b]		Terpenibactin C (3) ^[b]	
	δ_{H}	$\delta_{\text{C}}^{[c]}$	δ_{H}	$\delta_{\text{C}}^{[c]}$	δ_{H}	$\delta_{\text{C}}^{[c]}$
9,10-Dihydro-asteroidic acid						
1		110.5, qC		110.5, qC		110.5, qC
2		167.5, qC		167.5, qC		167.5, qC
3	6.81, d (8.5)	122.8, CH	6.81, d (8.5)	122.8, CH	6.81, d (8.5)	122.8, CH
4	7.36, ddd (8.7, 7.9, 1.8)	137.0, CH	7.36, ddd (8.7, 7.9, 1.8)	137.0, CH	7.36, ddd (8.7, 7.9, 1.8)	137.0, CH
5	6.68, t (7.6)	117.3, CH	6.68, t (7.6)	117.3, CH	6.68, t (7.6)	117.3, CH
6	7.67, dd (8.1, 1.8)	130.8, CH	7.67, dd (8.1, 1.8)	130.8, CH	7.67, dd (8.1, 1.8)	130.8, CH
7		171.89, qC ^[d]		171.89, qC ^[d]		171.89, qC ^[d]
9	5.27, dd (5.1, 6.3)	81.0, CH	5.27, dd (5.1, 6.3)	81.0, CH	5.27, dd (5.1, 6.3)	81.0, CH
10	4.31, d (5.1)	73.2, CH	4.31, d (5.1)	73.2, CH	4.30, d (5.2)	73.2, CH
12		170.8, qC		170.8, qC		170.8, qC
13	1.49, d (6.3)	21.3, CH ₃	1.49, d (6.3)	21.3, CH ₃	1.49, d (6.3)	21.3, CH ₃
Lys1						
C=O		171.95, qC ^[d]		171.96, qC ^[d]		171.95, qC ^[d]
α	4.39, m	54.9, CH [54.8]	4.39, m	54.9, CH	4.39, m	54.9, CH [54.8]
β	2.10, m	27.5, CH ₂	2.10, m	27.5, CH ₂	2.10, m	27.5, CH ₂
γ	1.60, m	19.7, CH ₂	1.60, m	19.7, CH ₂	1.60, m	19.7, CH ₂
	1.97, m		1.97, m		1.97, m	
δ	1.89, m	29.9, CH ₂	1.89, m	29.9, CH ₂	1.89, m	29.9, CH ₂
ϵ	3.70, m	49.5, CH ₂	3.70, m	49.5, CH ₂	3.70, m	49.5, CH ₂
	3.95, m		3.95, m		3.95, m	
NH	9.21, d (4.7)		9.20, d (4.7)		9.21, d (4.7)	
N ^F –CH=O	8.13, s	155.3, CHO		155.3, CHO		155.3, CHO
Hydroxymethyl fatty acid						
1'		173.8, qC [173.9]		173.8, qC		173.8, qC [173.9]
2'	2.62, m	45.26, CH [45.31]	2.62, m	45.3, CH	2.62, m	45.26, CH [45.31]
3'	5.11, dt (2.0, 10.0)	78.0, CH	5.11, dt (2.0, 10.0)	78.0, CH	5.11, dt (2.0, 10.0)	78.0, CH
4'	1.49, m	30.2, CH ₂	1.49, m	30.2, CH ₂	1.49, m	30.2, CH ₂
5'	1.21, m	26.9, CH ₂	2.04, m	28.2, CH ₂	1.21, m	26.9, CH ₂
6'	1.30, m	30.6, CH ₂	5.35, t (4.5)	130.8, CH	1.30, m	30.6, CH ₂
7'	1.30, m	30.6, CH ₂	5.35, t (4.5)	130.8, CH	1.30, m	30.6, CH ₂
8'	1.30, m	30.6, CH ₂	2.04, m	28.2, CH ₂	1.30, m	30.6, CH ₂
9'	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
10'	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
11'	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
12'	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
13'	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
14'	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
15'	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
16'	1.30, m	33.1, CH ₂	1.30, m	30.6, CH ₂	1.30, m	30.6, CH ₂
17'	1.32, m	23.7, CH ₂	1.30, m	30.6, CH ₂	1.30, m	33.1, CH ₂
18'	0.92, br t (6.9)	14.5, CH ₃	1.30, m	33.1, CH ₂	1.32, m	23.7, CH ₂
19'			1.32, m	23.7, CH ₂	1.30, m	30.6, CH ₂
20'			0.92, br t (6.9)	14.5, CH ₃	0.92, br t (6.9)	14.5, CH ₃
2'-Me	1.14, d (7.2)	14.2, CH ₃	1.14, d (7.3)	14.2, CH ₃	1.14, d (7.2)	14.2, CH ₃
Lys2						
C=O		166.4, qC		166.4, qC		166.4, qC
α	4.79, m	51.2, CH [51.3]	4.79, m	51.2, CH	4.79, m	51.2, CH [51.3]
β	1.48, m	31.68, CH ₂ [31.70]	1.48, m	31.7, CH ₂	1.48, m	31.68, CH ₂ [31.70]
	1.95, m		1.95, m		1.95, m	
γ	2.05, m	28.9, CH ₂	2.05, m	28.9, CH ₂	2.05, m	28.9, CH ₂
δ	1.80, m	25.6, CH ₂	1.80, m	25.6, CH ₂	1.80, m	25.6, CH ₂
ϵ	3.85, m	54.4, CH ₂	3.85, m	54.4, CH ₂	3.85, m	54.4, CH ₂
	3.99, m		3.99, m		3.99, m	
NH	n. o. ^[e]		n. o. ^[e]		n. o. ^[e]	
N ^F –C=O		155.3, qC		155.3, qC		155.3, qC

[a] Measured at 400 (¹H) and 100 MHz (¹³C) in *d*₄-MeOH. [b] Coupling constants (*J*) are in parentheses and reported in Hz; chemical shifts are given in ppm. [c] Multiplicities were deduced from DEPT135 and multiplicity edited ¹H,¹³C HSQC NMR experiments. ¹³C NMR shift values in squared brackets represent a minor conformer (ratio 3:1). [d] Assignments within a column may be interchanged. [e] n.o.: not observed.

CO, and $^1\text{H},^1\text{H}$ NOESY through-space interactions between the ϵ -methylene protons and the formyl hydrogen, defined the complete moiety as *N*-formyl-Lys1. In a similar fashion, fragment E was established as a second lysine residue (Lys2). However, the observation of a correlation between the α -proton and the corresponding the ϵ -methylene protons in the $^1\text{H},^1\text{H}$ NOESY spectrum indicated that Lys2 was cyclic, hence forming an ϵ -aminocaprolactam ring.

The last major fragment, D, was deduced as 2-methyl-3-hydroxyoctadecanoic acid on the basis of the analysis of COSY and HSQC-TOCSY spectra, together with the $^1\text{H},^{13}\text{C}$ long-range couplings from both H-2' and its 2'-methyl hydrogens to C-1'. The length of the alkyl chain was corroborated by ESI-MS/MS fragment ions. The connectivity of fragment D and E via Lys2-NH resulted from cross peaks in the HMBC spectrum between the α -proton of Lys2 and C-1'. Coupling between H-3' and Lys1-CO placed partial structure D adjacent to C through an ester bond. Furthermore, the α -proton of Lys1 exhibited HMBC couplings to C-12, which allowed the connection of partial structure C to B through an amide bond. Taking the molecular formula of 1 into consideration, the remaining two hydroxyl groups were assigned to substituents at the ϵ -nitrogen atoms of each lysine residue. This deduction was also supported by MS/MS fragment ions at m/z 145 and 423 (Figure S17). With the planar structure of 1 determined, the configuration at C-2', C-3', C-9, C-10 and the lysine residues required elucidation.

The absolute configuration of the amino acid-related moieties was inferred *in silico* from the BGC. In an NRPS context, the epimerization (E) domains and combined condensation/epimerization (C/E) domains determine the absolute configuration during the biosynthesis process. An E domain was identified in TerE, thus indicating conversion of the activated amino acid Lys2 into the D-configuration. The absence of C/E and E domains in the nonribosomal peptide synthetases TerF and TerD suggests the incorporation and processing of L-configured Thr and Lys1 moieties. In conclusion, from bioinformatics, the stereochemistry was predicted as 10*S*, 9*R*, L-(*S*)-Lys1 and D-(*R*)-Lys2. Concerning the lysine residues, these findings are in agreement with the absolute configuration determined for nocobactin NA. The relative stereochemistry at C-2'-C-3' of the 2-methyl-3-hydroxyoctadecanoic acid was assigned as *erythro* on the basis of the high similarity of ^1H and ^{13}C NMR shifts and $J_{\text{H,H}}$ values between 1 and brasilibactin A. Thus, the structure of terpenibactin A was elucidated to be 1 as shown in Figure 4. With these features, terpenibactin A (1) is most closely related to nocardimicin H, but bears a methyl group instead of a hydrogen atom at R⁷ (Figure 4 and Table 1).

Terpenibactin B (2) gave a quasi-molecular ion [$M-3\text{H} + \text{Ga}^{3+} + \text{H}$]⁺ at m/z 908.4335 in the HR-ESI-MS, appropriate for the molecular formula $\text{C}_{45}\text{H}_{69}\text{GaN}_5\text{O}_{10}$ (Δ 3.8 ppm) and corresponding to 13 degrees of unsaturation. Thus, terpenibactin B (2) differs from terpenibactin A (1) by a C_2H_2 group and an additional double bond equivalent. The ^1H and ^{13}C NMR spectra of 2 were highly similar to those of 1. However, the ^{13}C NMR in particular showed elevated signal intensity in the double bond region at δ_{C} 130.9 and additional, heightened resonances at δ_{C} 28.2. Since both resonances integrated for two carbons, it was

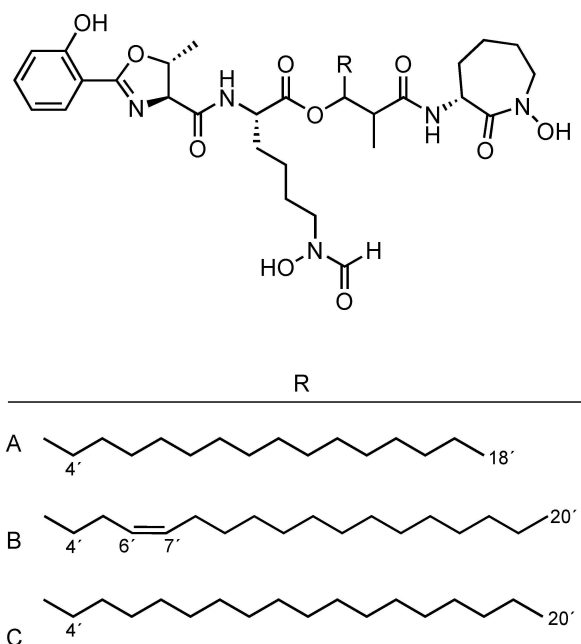


Figure 4. Chemical structures of terpenibactins A–C.

hypothesised that 2 constitutes a double bond between two magnetically equivalent carbon atoms. The HSQC-TOCSY NMR experiment proved that both resonances were part of the spin system of the hydroxymethyl fatty acid of 2. Detailed analysis of the $^1\text{H},^1\text{H}$ COSY and CH long-range correlations enabled the assignment of $\Delta^{6,7}$ double bond in 2. Its olefinic coupling constant ($^3J_{\text{H,H}} = 4.5$ Hz) showed that $\Delta^{6,7}$ has *Z* geometry. For biogenetic reasons, the same absolute configuration, as given in 1, is suggested for the remaining chiral centres of 2. Terpenibactin B (2) thus most closely resembles nocardimicin I. However, by comparison, 2 bears a methyl group at R⁷ of the oxazoline ring system and a double bond in the lipid side chain R³ (Figure 4 and Table 1).

The molecular formula of terpenibactin C (3) was established as $\text{C}_{45}\text{H}_{71}\text{GaN}_5\text{O}_{10}$ on the basis of HR-ESI-MS. The ^1H and ^{13}C NMR spectra were virtually identical with those of terpenibactin A(1). The MS/MS fragmentation pattern of 3 readily unveiled that the mass shift of 28 mass units (2 methylenes) is attributable to a difference in length of the incorporated fatty acid. Thus, 3 possesses a 2-methyl-3-hydroxyeicosanoic acid moiety instead of a 2-methyl-3-hydroxyoctadecanoic acid in 1. Consequently, terpenibactin C represents a highly similar congener of nocardimicin I, but 3 possesses a methyl-oxazoline instead of an oxazoline ring system (Figure 4 and Table 1).

Biological properties of terpenibactins A–C

As several nocobactins have shown outstanding antimicrobial,^[28] cytotoxic^[25,28] and highly selective muscarinic M3 receptor inhibitory^[29,30] properties, we evaluated the gallium

(III)-chelated terpenibactins A–C (1–3) in corresponding assays. In standardised antimicrobial assays by broth microdilutions,^[39] compounds 1–3 were all inactive up to the highest concentration tested (64 µg/mL), while in a cytotoxicity assay using HeLa cells, the compounds exhibited moderate bioactivity. By contrast, all three compounds showed potent muscarinic M3 receptor inhibitory activity, with IC₅₀ values ranging from 1.15 to 1.77 µM (Table 3). The latter results are in good agreement with the IC₅₀ values reported for the structurally related nocardimicins A–H, which ranged from 0.37 to 5.89 µM.^[29,30] Sakagami and co-workers demonstrated that iron(III)-chelated nocobactins exhibit cytotoxic activity that is two orders lower than of the corresponding iron-free forms.^[25] It is conceivable that the gallium-chelated terpenibactins A–C are affected in a similar fashion, which could explain why compounds 1–3 were found to be inactive in the antimicrobial assay and only moderately active in the cytotoxicity assay.

Conclusion

Using a bioinformatics approach, we initially deduced already from the *in silico* analysis that the resultant secondary metabolite is new and confirmed the prediction by spectroscopic means. Considering the large number of nocobactins, terpenibactins represent another example of Nature's aptitude in combinatorial chemistry, that is, in this case, through the natural diversification of molecules simply by modification of A-domain substrate specificity, C-domain promiscuity and the presence of dehydrogenases or the efficacy of hydroxylation enzymes. Furthermore, we confirmed the antimuscarinic effect of this compound class, which was already reported for nocardimicins. The addition of terpenibactins to the compound family allowed us to deduce that the antimuscarinic effect of nocobactins appears not to be influenced by the change of the oxidation status of the oxazoline ring, neither by the methylation (serine vs. threonine incorporation) nor by the type of N^ε-acylation (formylation vs. acetylation). Together with the M3-receptor subtype specificity, found with the nocardimicins, and the moderate cytotoxicity, nocobactins possibly represent an interesting pharmacophore for the development of M3-receptor selective muscarinic inhibitors.

Experimental Section

General procedures: Optical rotation values were measured on a Jasco P-2000 polarimeter, using a 3.5 mm × 10 mm cylindrical

Table 3. Biological activities of 1–3.

Compound	Muscarinic M3 receptor inhibitory activity IC ₅₀	Cytotoxicity IC ₅₀ (HeLa)
terpenibactin A (1)	1.15 ± 0.13 µM	16.7 µg/mL
terpenibactin B (2)	1.77 ± 0.26 µM	45.9 µg/mL
terpenibactin C (3)	1.59 ± 0.13 µM	33.3 µg/mL
atropine	8.10 ± 0.80 nM	N/A

quartz cell. UV spectra were recorded on a PerkinElmer Lambda 25 UV/Vis spectrometer. Infrared spectra were obtained by employing a Jasco FTIR 4200 spectrometer, interfaced with a MIRacle ATR device (ZnSe crystal). HPLC was performed with a Waters system comprising a Waters 1525 binary pump with a 7725i Rheodyne injection port, a Kromega solvent degasser and a Waters 996 photodiode array detector. For low-resolution LC/MS analysis, an 1100 Series HPLC system (Agilent Technologies) was fitted with a G1322 A degasser, a G1312 A binary pump, a G1329 A autosampler, and a G1315 A diode array detector. The Agilent HPLC components were connected with an ABSCIEX 3200 Q TRAP LC/MS/MS mass spectrometer (Sciex, Darmstadt, Germany). 1D and 2D NMR spectra were measured on a Bruker Avance II 400 HD spectrometer using a 5 mm SMART probe head. All NMR spectra were recorded in d₄-MeOH, processed with TopSpin 3.5 and MestReNova 12.0.4 and calibrated to the residual solvent signals (δ_{H/C} 3.31/49.0). High-resolution mass spectra were acquired on an HR-ESI-TOF-MS maXis 4G mass spectrometer (Bruker). All solvents were purchased as HPLC or LC/MS grade. Chemicals were purchased from Sigma-Aldrich or VWR. Steam sterilization of medium components and waste were performed at 121 °C for 15 min at 2.1 bar using a Systec VX-150 autoclave, equipped with air exhaust filtration.

Bacterial material and cultivation conditions: The producer strain *N. terpenica* IFM 0406 was purchased from the culture collection of the Medical Mycology Research Center (MMRC), Chiba University, Chiba, Japan. For cultivation, a protocol conceived by Ikeda et al.^[29a] was adapted. 100 mL of preculture medium (consisting of 10 g/L soluble starch, 5 g/L glucose, 3 g/L NZ-case, 2 g/L yeast extract, 5 g/L tryptone, 1 g/L K₂HPO₄, 0.5 g/L MgSO₄ and 3 g/L CaCO₃, adjusted to pH 7.0 before sterilization) in a 500 mL baffled Erlenmeyer flask was inoculated with a loopful of a 3-day-old agar culture of the strain, and shaken at 150 rpm at 37 °C for five days, using an INFORS HT Multitron Pro orbital incubator shaker. 3 mL of this culture was in turn used to inoculate 200 mL volumes of production medium (consisting of 5 g/L glucose, 20 g/L glycerol, 20 g/L soluble starch, 15 g/L Pharmamedia and 3 g/L yeast extract, adjusted to pH 7.0 before sterilization) in 500 mL Erlenmeyer flasks. Fermentation was carried out at 30 °C at 150 rpm for five days.

LC/MS screening: Following small scale cultivation (200 mL), cultures were centrifuged in a Thermo Scientific Heraeus Multifuge 4KR centrifuge at 5346 g for 10 min. Subsequently, the supernatant was extracted with ethyl acetate (1:1), while the cell pellet was extracted with 200 mL methanol. Upon evaporation of each extract under reduced pressure, the resultant crude extracts of each experiment were dissolved in methanol with a few drops of water, centrifuged to remove detritus prior to analysis, and profiled by LC/MS using a MeOH/H₂O gradient, increasing the MeOH portion from 50 to 100% over 20 min, followed by isocratic elution at 100% MeOH for an additional 20 min (Phenomenex Luna C18 (2)–100 Å column, 250 × 2.0 mm, 5 µm; 0.2 mL/min flow rate; with total ion current [Q3 MS positive scan mode] and photodiode array monitoring). The following MS parameters were used for this analysis: declustering potential: 70 V; entrance potential: 10 V; curtain gas: 10 psi; collision gas: 2 psi; ion spray source voltage: 4500 V; additional temperature: 450 °C, ion source gas 1 (nebuliser gas): 50 psi; ion source gas 2 (drying gas): 50 psi.

Isolation and extraction: All the broth from a 1 L fermentation of IFM 0406 was centrifuged, the cell pellet extracted with 6 × 200 mL MeOH, and evaporated to dryness. The cell pellet extract was resuspended and further fractionated via vacuum liquid chromatography (VLC). The reversed-phase (RP) C₁₈ column (dimensions: 10 × 5 cm; material: Macherey-Nagel Polygoprep 50–60 C₁₈ RP silica gel) was eluted stepwise under vacuum with solvents of decreasing polarity (300 mL per fraction), ranging from a 50:50 MeOH/H₂O mixture to pure methanol, followed by 100% dichloromethane to

give seven fractions (A–G). Fraction G, eluting with 100% dichloromethane, was further purified by RP-HPLC using a linear gradient from 50:50 to 100:0 (MeOH/H₂O) over a period of 10 min, followed by isocratic elution at 100% MeOH for an additional 25 min (Phenomenex Aeris Peptide XB–C18 column, 4.6 × 250 mm, 3.6 μm; 0.6 mL/min flow rate, UV monitoring at 340 and 460 nm), which yielded pure terpenibactins A (13.6 mg), B (10.9 mg) and C (7.4 mg).

Ga^{III}-complex formation: A suspension of 5 mg gallium sulphate in 30 mL MeOH:H₂O (1:1) was added to a near-dry fresh methanol pellet extract, and the sample rotated for one hour. This was then evaporated to dryness, resuspended in methanol and analysed using LR-LC/MS. Results showed that three new peaks had formed alongside the ferrated components, with higher intensities and masses corresponding to the gallium-complexed forms of their ferrated neighbours. Notably, repeated LR-LC/MS analysis, two days after the run that first evidenced the formation of gallium complexes, demonstrated that only trace amounts remained of the ferrated forms of the compounds of interest, suggesting a time-dependent complete interconversion of iron to gallium complexes even within a sample dried and stored at –20 °C.

Chrome azurol S (CAS)-assay: The CAS assay agar plates for the detection of siderophores were prepared on a 100 mL scale as previously described.^[36]

Antibacterial assays: The minimal inhibitory concentration (MIC) was determined in cation-adjusted Mueller-Hinton medium that contains casein, beef extract and starch by using a twofold serial dilution method according to the standards and guidelines of the Clinical and Laboratory Standards Institute (CLSI).^[39] In brief, a twofold serial dilution of the test compound was prepared in microtiter plates and seeded using a final test concentration of bacteria of 5 × 10⁵ colony-forming units per mL. After the overnight incubation at 37 °C, the MIC was determined as the lowest compound concentration preventing visible bacterial growth. The strain panel included representative species of nosocomial pathogens, which are known as “ESKAPE” bacteria. Specifically, the following strains were used: *Enterococcus faecium* BM 4147–1, *Staphylococcus aureus* ATCC 29213, *Klebsiella pneumoniae* ATCC 12657, *Acinetobacter baumannii* 09987, *Pseudomonas aeruginosa* ATCC 27853 and *Enterobacter aerogenes* ATCC 13048. *Bacillus subtilis* 168 and *Escherichia coli* ATCC 25922 were used as further reference strains. The ATCC strains were provided by the American Type Culture Collection. *A. baumannii* 09987 was obtained from the University of Bonn, Germany.

Cytotoxicity assay: The cytotoxicity test against the HeLa human cervical carcinoma cell line was performed in RPMI cell culture medium supplemented with 10% foetal bovine serum using the 7-hydroxy-3H-phenoxazin-3-one-10-oxide (resazurin) assay. A twofold serial dilution of the test compounds was prepared in duplicates in a microtiter plate and seeded with trypsinised HeLa cells to a final cell concentration of 1 × 10⁴ cells per well. After 24 h incubation at 37 °C, 5% CO₂, 95% relative humidity, resazurin was added at a final concentration of 200 μM, and cells were again incubated overnight. Cell viability was assessed by determining the reduction of resazurin to the fluorescent resorufin. Fluorescence was measured in a TECAN M200 reader at an excitation wavelength of 560 nm and an emission wavelength of 600 nm in relation to an untreated control.

Muscarinic M3 receptor calcium flux assay: Chem-1 frozen cells, stably transfected with a clone of the human M3 muscarinic acetylcholine receptor (CHRM3 cDNA, HTS116RTA), were purchased from Eurofins Pharma Bioanalytical Services US Inc. They were thawed, suspended and seeded in 96-well plates containing the media component according to the manufacturer's instructions.

The assay plates were then incubated overnight at 37 °C in a 5% CO₂ humidified atmosphere, during which the cell adhered to the bottom of the wells, followed by the removal of media components prior to use in the calcium assay.

Cytosolic-free calcium concentration ([Ca²⁺]_i) was measured using the calcium mobilization assay with fluorescent dyes. The dye-loading solution containing Fluo-8 NW Ca²⁺ dye, 10X Pluronic® F127 Plus and Hanks' buffer with 20 mM HEPES was prepared according to the Screen Quest™ Fluo-8 NW calcium assay kit (AAT Bioquest). Probenecid acid was not applied to this method. At the start of the assay, the 96-well plates containing Chem-1 cells were loaded with an equal volume of dye-loading solutions at 37 °C for 1 h. Afterwards, the cells were pre-incubated with the GPCR-antagonist test compounds terpenibactin A, B, and C for 30 min prior to the injection of the acetylcholine agonist (Sigma) at the final concentration of 250 nM. The final volume was 100 μL per well. The changes in fluorescence intensities in response to the indicated ligand were measured using a Tecan Infinite M200® injection system with a filter set for excitation and emission at 490 and 525 nm, respectively. IC₅₀ values and associated SEMs were calculated by curve-fitting the relative fluorescence unit data to an sigmoidal model using GraphPad Prism 7.00.

Acknowledgements

We kindly thank Dr. Wistuba (Mass Spectrometry Department, Institute for Organic Chemistry, University of Tübingen, Germany) for HR-MS measurements and Dr. Luqman (Microbial Genetics Department, Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Germany) for technical assistance in conducting the calcium mobilization assay. A. F. gratefully acknowledges the Program for Research and Innovation in Science and Technology (RISET-Pro)/World Bank Loan No. 8245 for a granted Ph.D. scholarship. H.B.-O., N.Z., L.K. and H.G. gratefully acknowledge support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project ID 398967434-TRR 261.

Conflict of Interest

The authors declare no conflict of interest.

- [1] E. Nocard, *Ann. Inst. Pasteur* **1888**, *2*, 293–302.
- [2] M. Fatahi-Bafghi, *Microb. Pathog.* **2018**, *114*, 369–384.
- [3] J. W. Wilson, *Mayo Clin. Proc.* **2012**, *87*, 403–407.
- [4] Q. Luo, S. Hiessl, A. Poehlein, R. Daniel, A. Steinbüchel, *Appl. Environ. Microbiol.* **2014**, *80*, 3895–3907.
- [5] K. A. El Sayed, *J. Nat. Prod.* **1998**, *61*, 149–151.
- [6] Z. H. Cheng, B. Y. Yu, G. A. Cordell, S. X. Qiu, *Org. Lett.* **2004**, *6*, 3163–3165.
- [7] D. Dhakal, A. K. Chaudhary, A. R. Pokhrel, A. K. Jha, S. Darsandhari, B. Shrestha, J. K. Sohng, *J. Biomol. Reconstruct.* **2013**, *10*, 9–17.
- [8] D. Dhakal, V. Rayamajhi, R. Mishra, J. K. Sohng, *J. Ind. Microbiol. Biotechnol.* **2019**, *46*, 385–407.
- [9] L. Vera-Cabrera, R. Ortiz-Lopez, R. Elizondo-Gonzalez, J. Ocampo-Candiani, *PLoS One* **2013**, *8*, e65425.
- [10] M. Imajoh, Y. Fukumoto, J. Yamane, M. Sakeda, M. Shimizu, K. Ohnishi, S.-I. Oshima, *Genome* **2015**, *3*, e00082–15.
- [11] F. Vautrin, E. Bergeron, A. Dubost, D. Abrouk, C. Martin, B. Cournoyer, V. Louzier, T. Winiarski, V. Rodriguez-Nava, P. Pujic, *Microbiol. Res.* **2019**, *8*, e00600–19.

- [12] M. A. Schorn, M. M. Alanjary, K. Aguinaldo, A. Korobeynikov, S. Podell, N. Patin, T. Lincecum, P. R. Jensen, N. Ziemert, B. S. Moore, *Microbiology* **2016**, *162*, 2075–2086.
- [13] Y. Hoshino, K. Watanabe, S. Iida, S. Suzuki, T. Kudo, T. Kogure, K. Yazawa, J. Ishikawa, R. M. Kroppenstedt, Y. Mikami, *Int. J. Syst. Evol. Microbiol.* **2007**, *57*, 1456–1460.
- [14] V. Wiebach, A. Mainz, M.-A. J. Siegert, N. A. Jungmann, G. Lesquame, S. Tirat, A. Dreux-Zigha, J. Aszodi, D. Le Beller, R. D. Süssmuth, *Nat. Chem. Biol.* **2018**, *14*, 652–654.
- [15] Y. Tanaka, H. Komaki, K. Yazawa, Y. Mikami, A. Nemoto, T. Tojyo, K. Kadowaki, H. Shigemori, J. Kobayashi, *J. Antibiot.* **1997**, *50*, 1036–1041.
- [16] K. Komatsu, M. Tsuda, Y. Tanaka, Y. Mikami, J. Kobayashi, *J. Org. Chem.* **2004**, *69*, 1535–1541.
- [17] H. T. Chiu, C. P. Weng, Y. C. Lin, K. H. Chen, *Org. Biomol. Chem.* **2016**, *14*, 1988–2006.
- [18] H. Shigemori, H. Komaki, K. Yazawa, Y. Mikami, A. Nemoto, Y. Tanaka, T. Sasaki, Y. In, T. Ishida, J. Kobayashi, *J. Org. Chem.* **1998**, *63*, 6900–6904.
- [19] H. Komaki, A. Nemoto, Y. Tanaka, H. Takagi, K. Yazawa, Y. Mikami, H. Shigemori, J. Kobayashi, A. Ando, Y. Nagata, *J. Antibiot.* **1999**, *52*, 13–19.
- [20] K. Komatsu, M. Tsuda, M. Shiro, Y. Tanaka, Y. Mikami, J. Kobayashi, *Bioorg. Med. Chem.* **2004**, *12*, 5545–5551.
- [21] T. Usui, Y. Nagumo, A. Watanabe, T. Kubota, K. Komatsu, J. Kobayashi, H. Osada, *Chem. Biol.* **2006**, *13*, 1153–1160.
- [22] Y. Hayashi, N. Matsuura, H. Toshima, N. Itoh, J. Ishikawa, Y. Mikami, T. Dairi, *J. Antibiot.* **2008**, *61*, 164–174.
- [23] A. Buchmann, M. Eitel, P. Koch, P. N. Schwarz, E. Stegmann, W. Wohlleben, M. Wolański, M. Krawiec, J. Zakrzewska-Czerwińska, C. Méndez, A. Botas, L. E. Núñez, F. Moris, J. Cortés, H. Gross, *Genome* **2016**, *4*, e01391–16.
- [24] P. N. Schwarz, A. Buchmann, L. Roller, A. Kulik, H. Gross, W. Wohlleben, E. Stegmann, *Biotechnol. J.* **2018**, *13*, 1700527.
- [25] a) C. Rattledge, G. A. Snow, *Biochem. J.* **1974**, *139*, 407–413; b) H. Sakagami, M. Ishihara, Y. Hoshino, J. Ishikawa, Y. Mikami, T. Fukai, *In Vivo* **2005**, *19*, 277–282.
- [26] Y. Murakami, S. Kato, M. Nakajima, M. Matsuoka, H. Kawai, K. Shin-Ya, H. Seto, *J. Antibiot.* **1996**, *49*, 839–845.
- [27] a) K. Suenaga, S. Kokubo, C. Shinohara, T. Tsuji, D. Uemura, *Tetrahedron Lett.* **1999**, *40*, 1945–1948; b) K. Kokubo, K. Suenaga, C. Shinohara, T. Tsuji, D. Uemura, *Tetrahedron* **2000**, *56*, 6435–6440.
- [28] M. Tsuda, M. Yamakawa, S. Oka, Y. Tanaka, Y. Hoshino, Y. Mikami, A. Sato, H. Fujiwara, Y. Ohizumi, J. Kobayashi, *J. Nat. Prod.* **2005**, *68*, 462–464.
- [29] a) Y. Ikeda, H. Nonaka, T. Furumai, H. Onaka, Y. Igarashi, *J. Nat. Prod.* **2005**, *68*, 1061–1065; b) J. C. Banks, C. J. Moody, *Tetrahedron Lett.* **2009**, *50*, 3371–3373.
- [30] Y. Ikeda, T. Furumai, Y. Igarashi, *J. Antibiot.* **2005**, *58*, 566–572.
- [31] K. Blin, S. Shaw, K. Steinke, R. Villebro, N. Ziemert, S. Y. Lee, M. H. Medema, T. Weber, *Nucleic Acids Res.* **2019**, *47*, W81–W87.
- [32] J. Bursy, A. U. Kuhlmann, M. Pittelkow, H. Hartmann, M. Jebbar, A. J. Pierik, E. Bremer, *Appl. Environ. Microbiol.* **2008**, *74*, 7286–7296.
- [33] Y. Hoshino, K. Chiba, K. Ishino, T. Fukai, Y. Igarashi, K. Yazawa, Y. Mikami, J. Ishikawa, *J. Bacteriol.* **2011**, *193*, 441–448.
- [34] M. D. McMahon, J. S. Rush, M. G. Thomas, *J. Bacteriol.* **2012**, *194*, 2809–2818.
- [35] X. Liu, Y. Jin, Z. Cui, K. Nonaka, S. Baba, M. Funabashi, Z. Yang, S. G. Van Lanen, *ChemBioChem* **2016**, *17*, 804–810.
- [36] B. Schwyn, J. B. Neilands, *Anal. Biochem.* **1987**, *160*, 47–56.
- [37] T. Emery, *Biochemistry* **1986**, *25*, 4629–33.
- [38] a) H. Stephan, S. Freund, W. Beck, G. Jung, J. M. Meyer, G. Winkelmann, *BioMetals* **1993**, *6*, 93–100; b) S. Lautru, R. J. Deeth, L. M. Bailey, G. L. Challis, *Nat. Chem. Biol.* **2005**, *1*, 265–269.
- [39] J. B. Patel, F. R. Cockerill, P. A. Bradford, G. M. Eliopoulos, J. A. Hindler, S. G. Jenkins, J. S. Lewis, B. Limbago, L. A. Miller, D. P. Nicolau, M. Powell, J. M. Swenson, J. D. Turnidge, M. P. Weinstein, B. L. Zimmer, *Methods for Dilution Antimicrobial Susceptibility Tests for Bacteria that Grow Aerobically. Approved Standard, Vol. 35*, 10th ed., Clinical and Laboratory Standards Institute, USA, **2015**.

Manuscript received: January 31, 2020

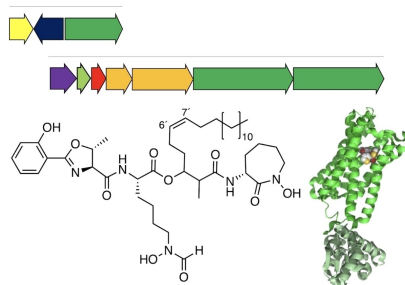
Revised manuscript received: March 7, 2020

Accepted manuscript online: March 20, 2020

Version of record online: ■■■, ■■■■

FULL PAPERS

Nature's combinatorial biosynthesis produces new bioactive nocobactin derivatives. Genome mining led to the discovery and characterization of terpenibactins A–C, three new nocobactins from the bacterium *Nocardia terpenica* IFM 0406 that exhibit unprecedented combinations of asteroic acid, lipid and modified lysine residues. They demonstrate low cytotoxicity but potent inhibitory activity towards the muscarinic M3 receptor.



J. Chen, A. Frediansyah, D. Männle, J. Straetener, Prof. Dr. H. Brötz-Oesterhelt, Prof. Dr. N. Ziemert, Dr. L. Kaysser, Prof. Dr. H. Gross*

1 – 10

New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A–C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406



ChemBioChem

Supporting Information

New Nocobactin Derivatives with Antimuscarinic Activity, Terpenibactins A–C, Revealed by Genome Mining of *Nocardia terpenica* IFM 0406

Julia Chen, Andri Frediansyah, Daniel Männle, Jan Straetener, Heike Brötz-Oesterhelt, Nadine Ziemert, Leonard Kaysser, and Harald Gross*© 2020 The Authors. Published by Wiley-VCH Verlag GmbH & Co. KGaA. This is an open access article under the terms of the Creative Commons Attribution Non-Commercial NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made. This article is part of a Special Collection on Microbial Biosynthesis and Interactions. To view the complete collection, visit our

Table of Contents

LC/MS Screening

Figure S01. LC/MS chromatogram of crude extract

Figure S02. Expected structural formula of Fe(III)-bound nocobactin-derivatives

Compound Isolation

Figure S03. HPLC-chromatogram of Fe(III)-bound terpenibactins

Figure S04. Monitoring of the conversion from Fe(III)- into Ga(III)-terpenibactins by LC/MS

Figure S05. HPLC-chromatogram of Ga(III)-bound terpenibactins

Spectroscopic data for terpenibactin A (1)

Optical rotation value of **1**

Figure S06 IR-spectrum of **1**

Figure S07 UV-spectrum of **1**

Figure S08. HR-ESI-MS spectrum of Fe(III)-terpenibactin A (**1**)

Figure S09. ¹H NMR spectrum

Figure S10. ¹³C NMR spectrum

Figure S11. DEPT135 NMR spectrum

Figure S12. ¹H-¹³C HSQC NMR spectrum

Figure S13. ¹H-¹H COSY NMR spectrum

Figure S14. ¹H-¹³C HSQC-TOCSY NMR spectrum

Figure S15. ¹H-¹³C HMBC NMR spectrum

Figure S16. ¹H-¹H NOESY NMR spectrum

Figure S17 MS fragment analysis

Spectroscopic data for terpenibactin B (2)

Optical rotation value of **2**

Figure S18. ¹H NMR spectrum

Figure S19. ¹³C NMR spectrum

Spectroscopic data for terpenibactin C (3)

Optical rotation value of **3**

Figure S20. ¹H NMR spectrum

Figure S21. ¹³C NMR spectrum

Figure S22. ¹H-¹³C HSQC NMR spectrum

Figure S23. ¹H-¹H COSY NMR spectrum

Figure S24. ¹H-¹³C HSQC-TOCSY NMR spectrum

Figure S25. ¹H-¹³C HMBC NMR spectrum

Bioassay data

Figure S26. Calcium mobilisation assay

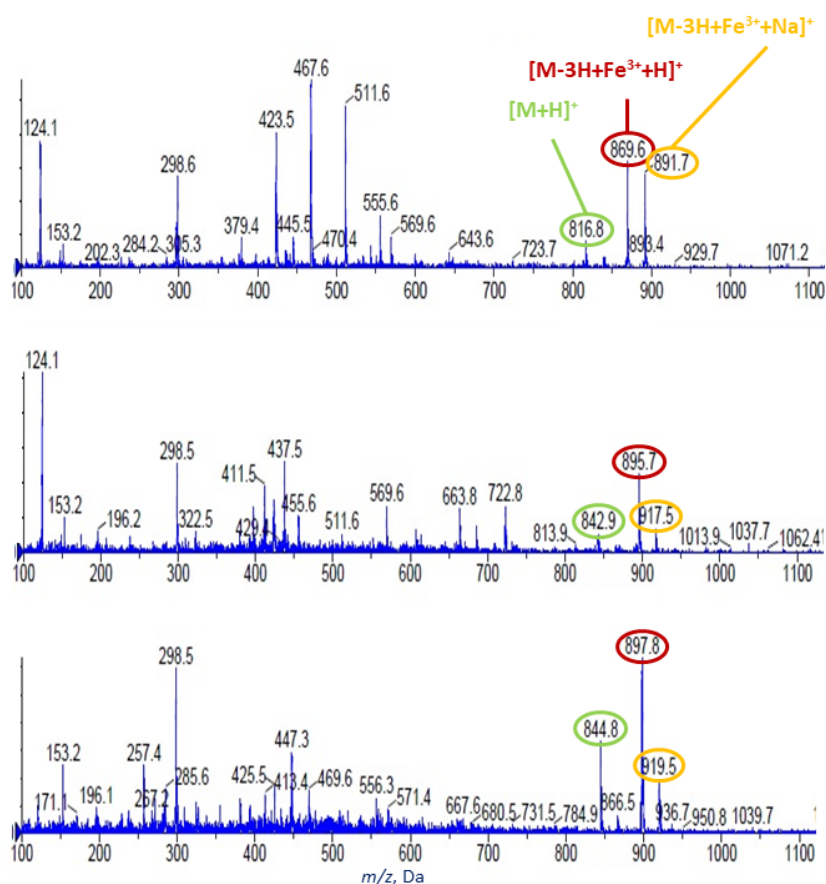


Figure S01. LR-LC/MS Q3 positive mode scan of *N. terpenica* IFM 0406 methanol extract showing (top to bottom) ions at m/z 816, 842 and 844 (green) accompanied by ions at m/z 869, 895 and 897 (red) and their sodium adducts (orange) respectively, predicted to belong to novel nocobactin-like siderophores. Subsequent HR-ESI-MS analysis of these three peaks showed their sodium adducts as base peaks, with exact masses of exact masses of m/z 891.4043, 917.4201 and 919.4358, matching closely with predicted chemical formulae of $C_{43}H_{66}FeN_5NaO_{10}$ (Δ -1.6 ppm), $C_{45}H_{68}FeN_5NaO_{10}$ (Δ -1.4 ppm) and $C_{45}H_{70}FeN_5NaO_{10}$ (Δ -1.3 ppm).

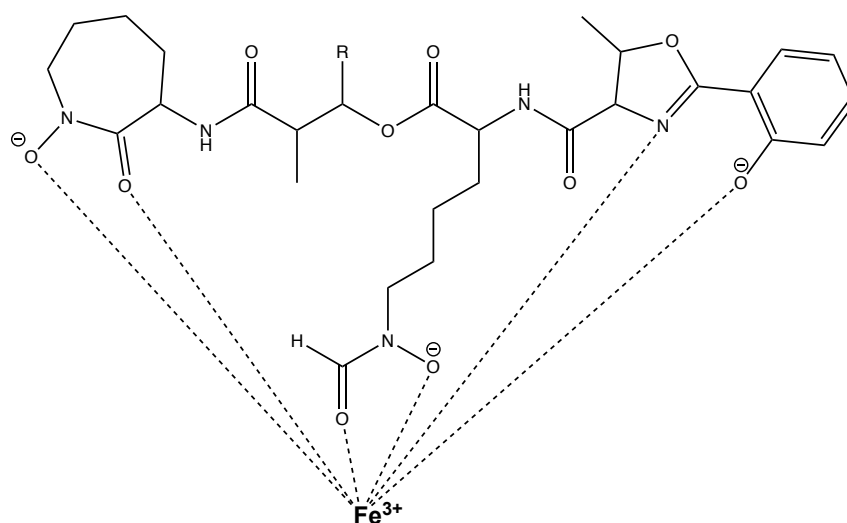


Figure S02. Expected structural formula of Fe(III)-bound nocobactin-like novel siderophores $[M-3H+Fe^{3+}+H]^+$, where R = fatty acid chain of variable length and saturation.

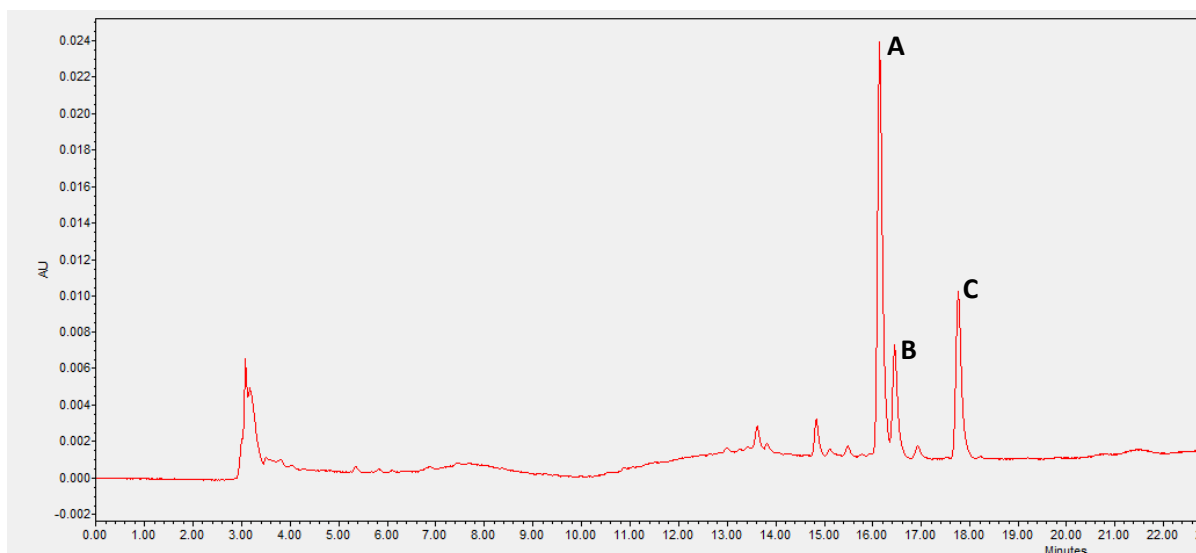


Figure S03. HPLC run, monitoring at 460 nm. Ferrated terpenibactins A-C appear at $t_R = 16$ to 18 min.

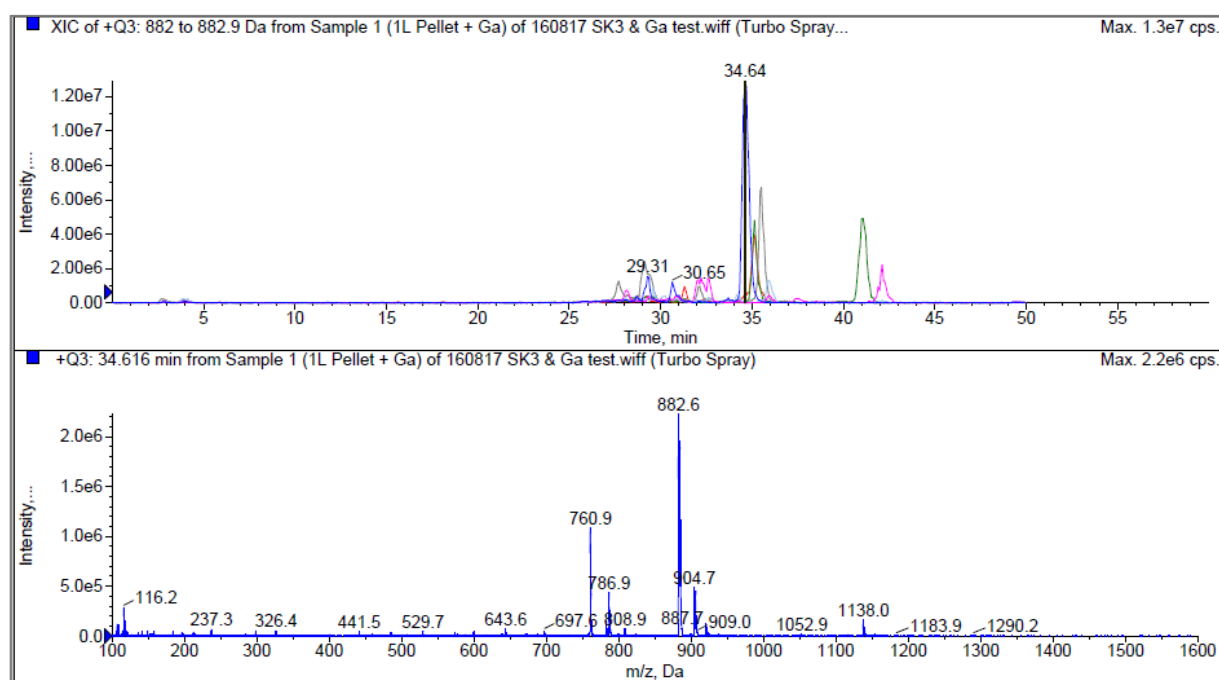


Figure S04: LR-LC/MS of gallium-bound terpenibactins in positive mode. Top panel: extracted ion chromatogram of both iron- and gallium-bound novel siderophores, with gallium-bound forms (blue, red and green traces) showing higher intensities; bottom panel: base peak at m/z 882 represents mass of Ga-bound terpenibactin A ($[M-3H+Ga^{3+}+H]^+$).

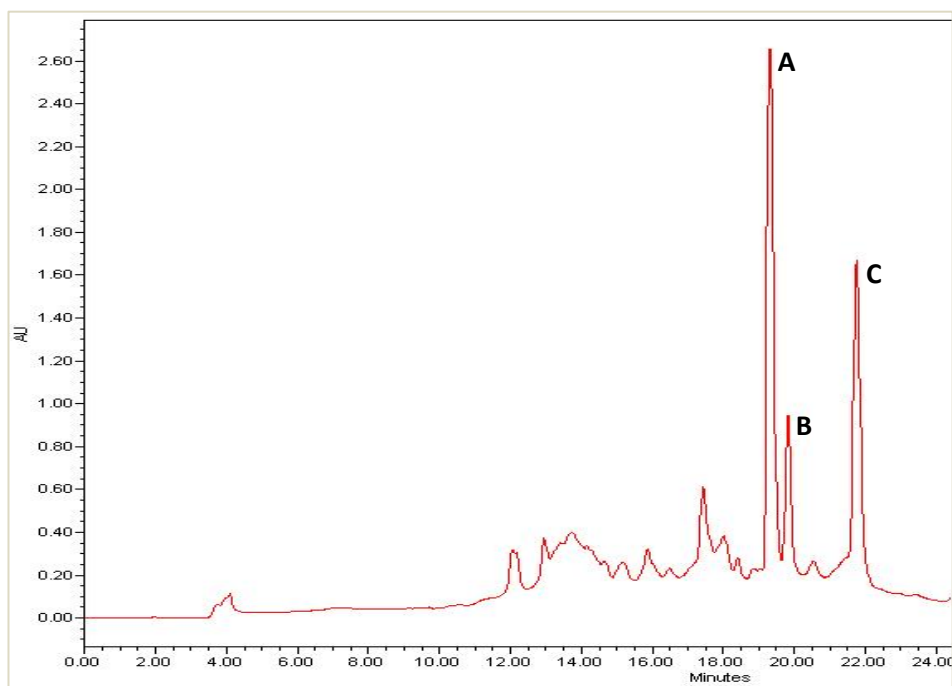


Figure S05. HPLC run, monitoring at 340 nm. Gallium-bound terpenibactins A-C appear at t_R = 19 to 23 min.

Optical rotation value of 1: $[\alpha]_{23}^D +93.9^\circ$ (c 0.33, MeOH)

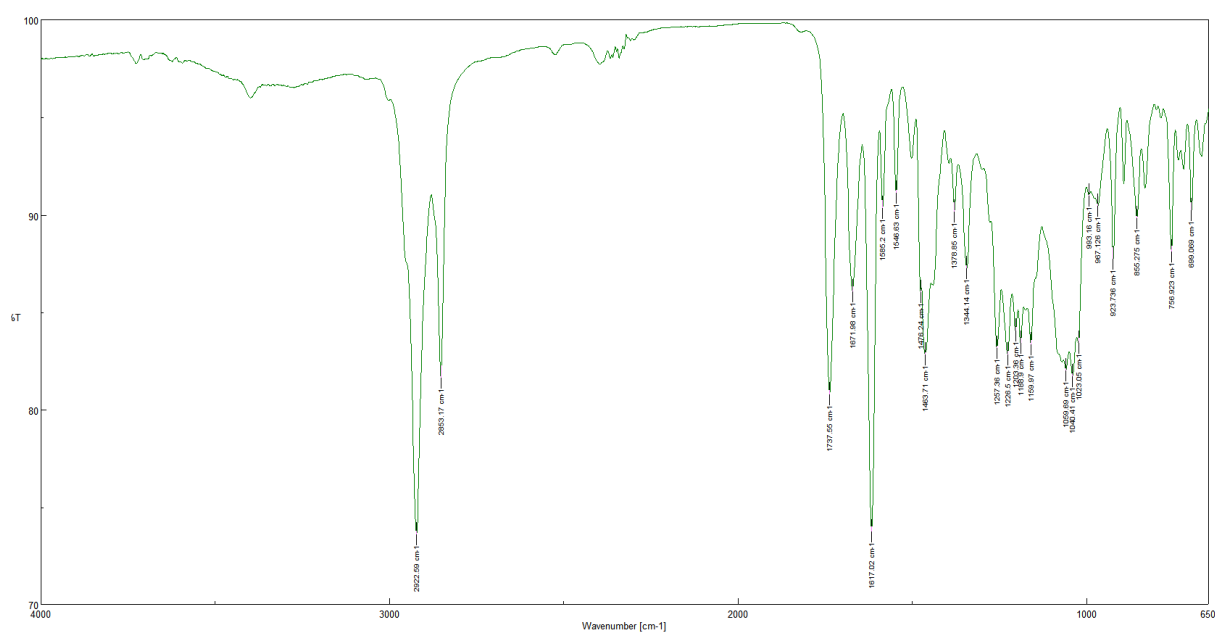


Figure S06. IR spectrum of 1.

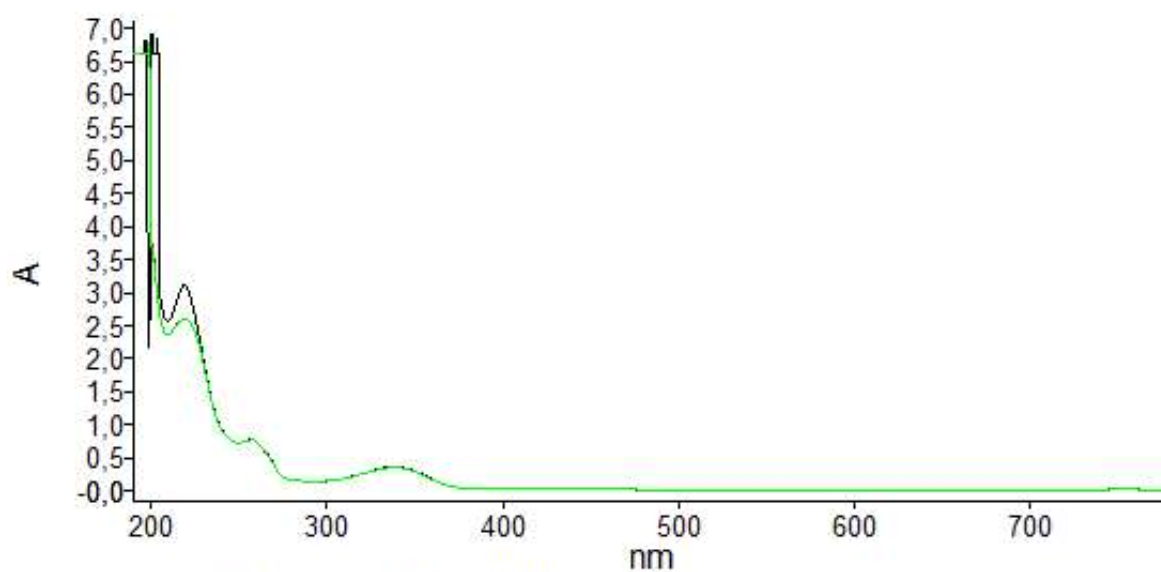


Figure S07. UV spectrum of **1** in MeOH: λ_{max} (log ϵ) 197 (3.06), 219 (2.64), 255 (2.11), 339 (1.83)

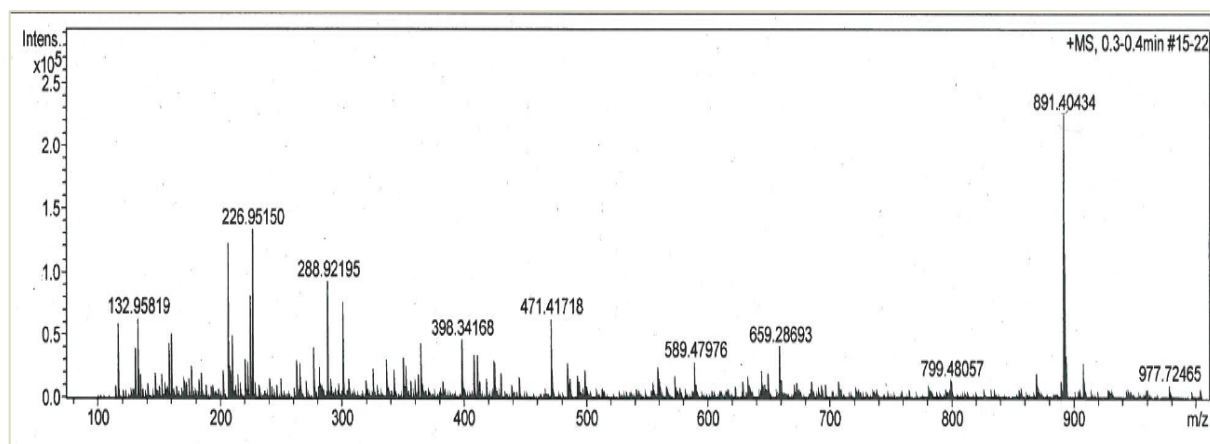


Figure S08. HR-ESI-MS analysis of Fe(III)-terpenibactin A with a base peak at m/z 891 $[\text{M}-3\text{H}+\text{Fe}^{3+}+\text{Na}]^+$.

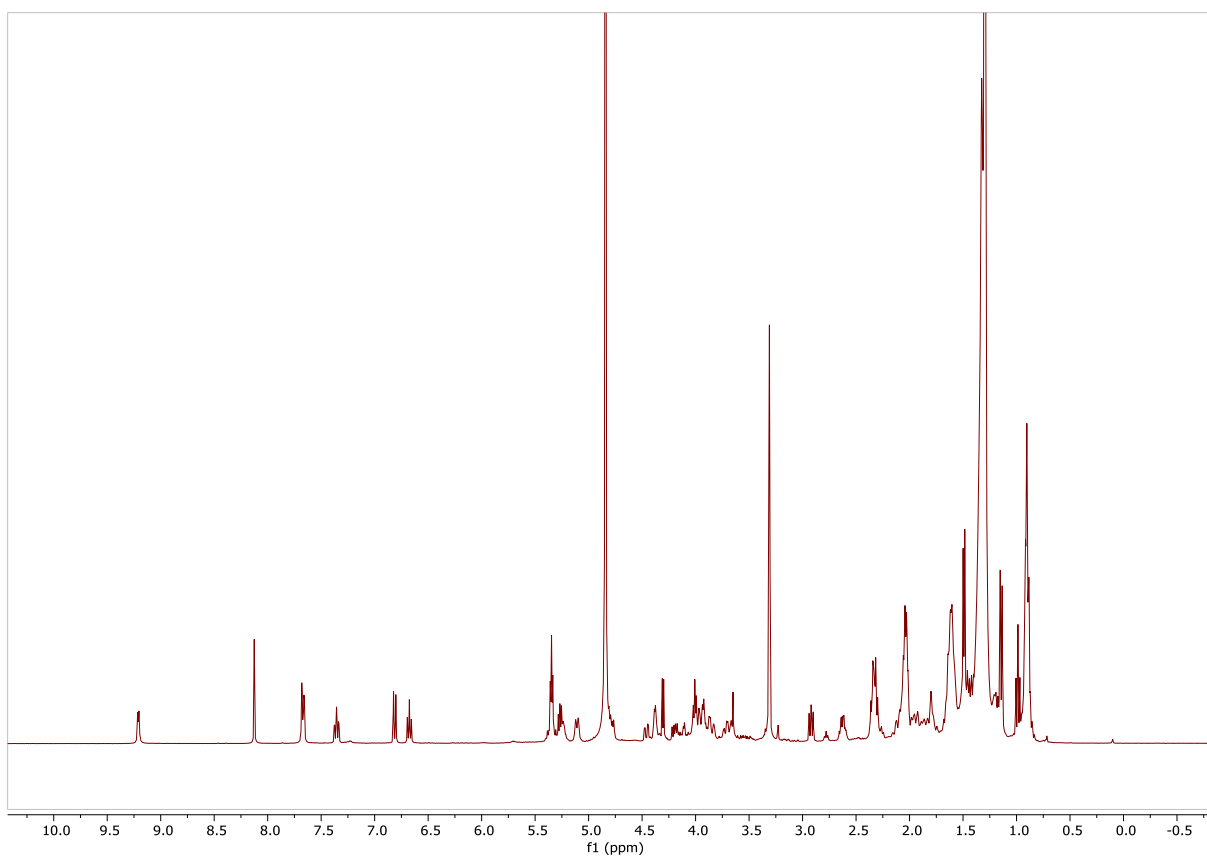


Figure S09. ^1H NMR spectrum (400 MHz) of **1** in d_4 -MeOH

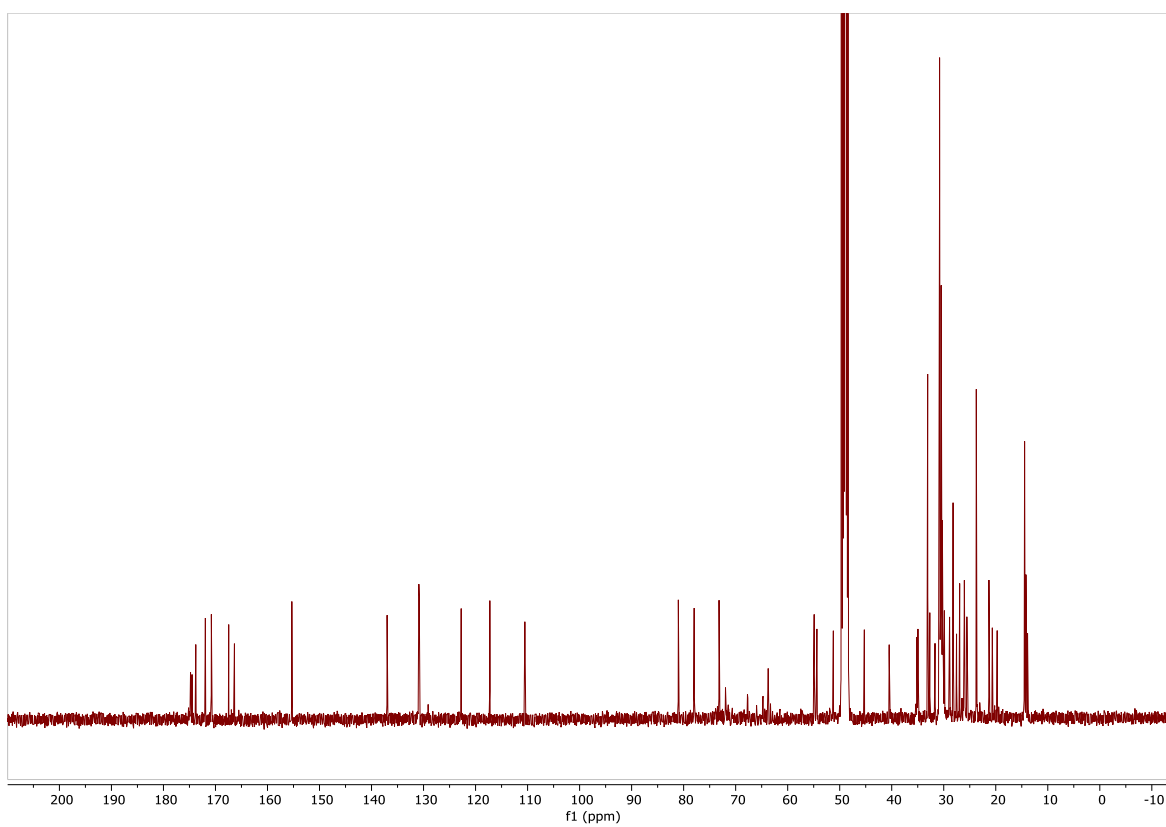


Figure S10. ^{13}C NMR spectrum (100 MHz) of **1** in d_4 -MeOH

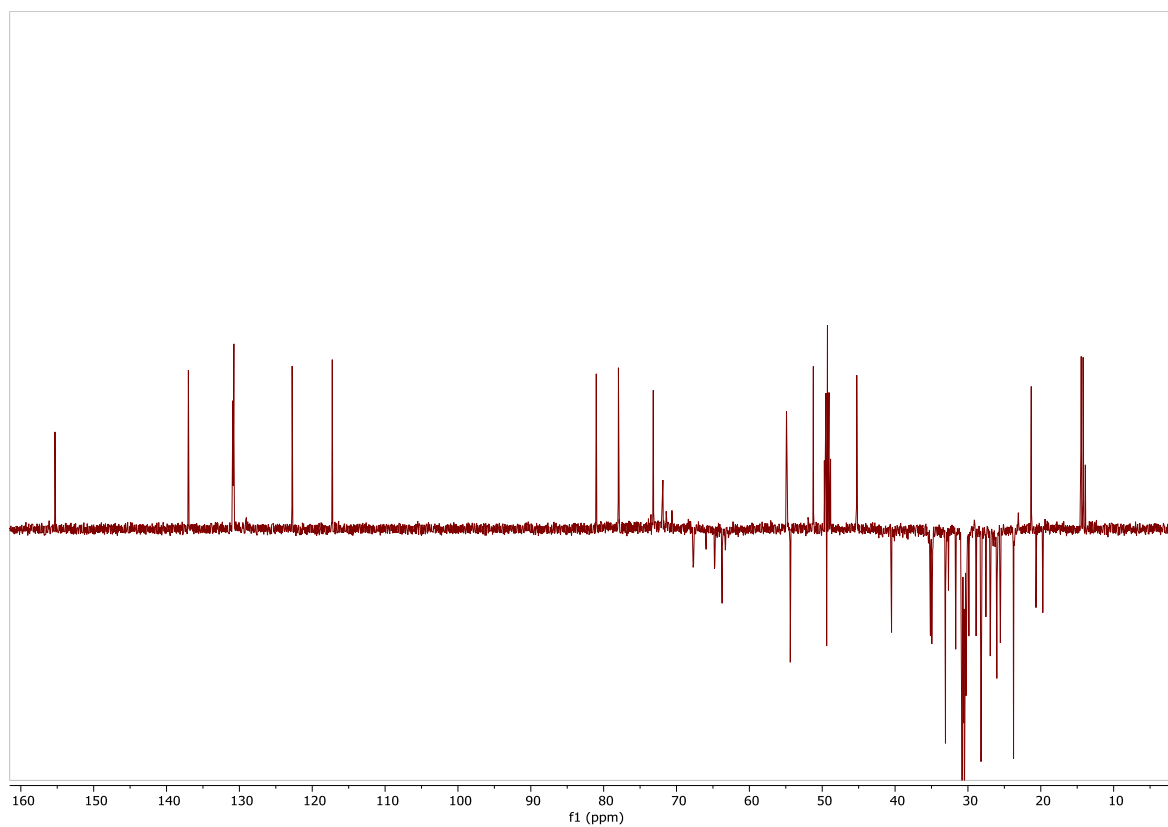


Figure S11. DEPT135 NMR spectrum (100 MHz) of **1** in d_4 -MeOH

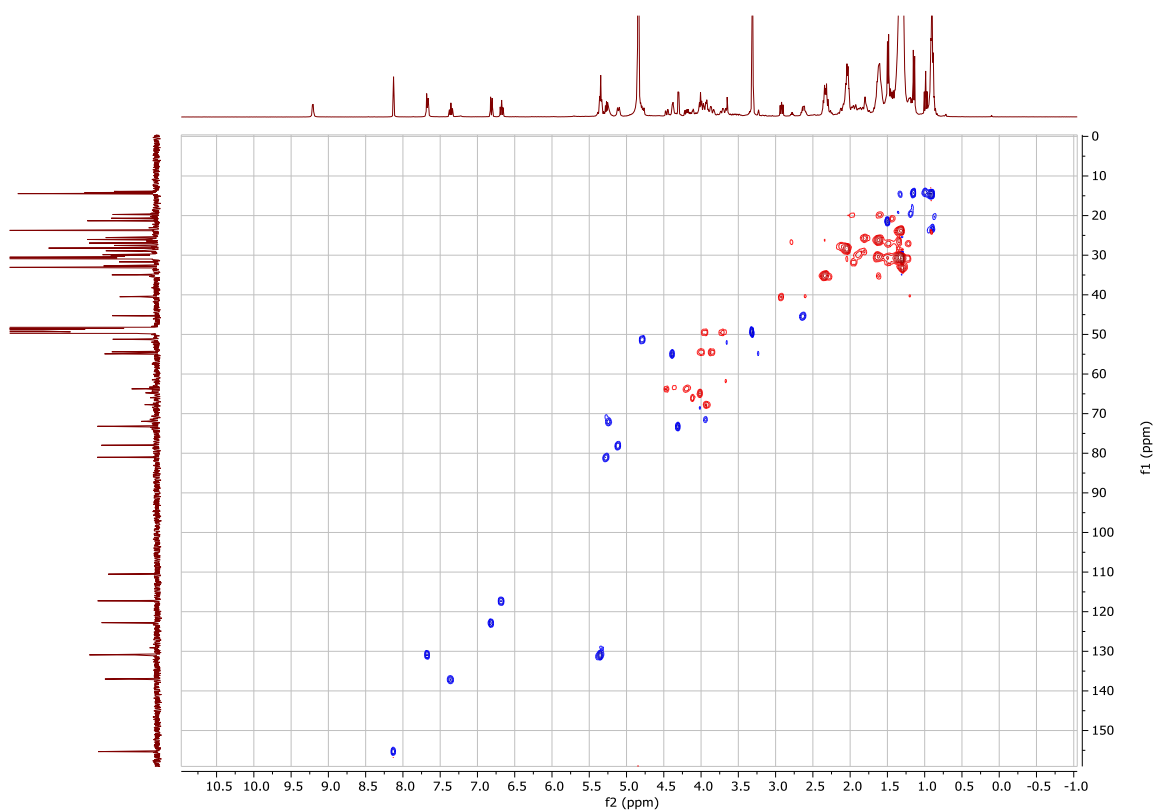


Figure S12. ^1H - ^{13}C HSQC NMR spectrum (400 MHz) of **1** in d_4 -MeOH

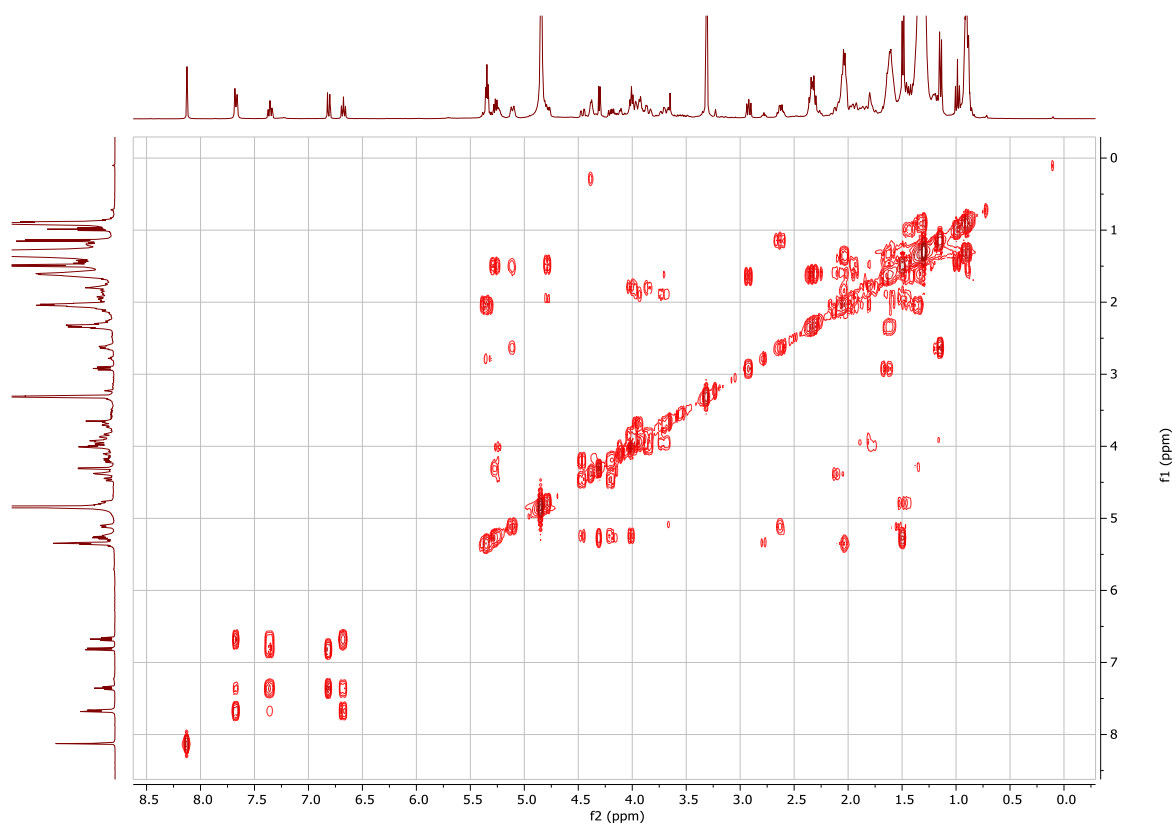


Figure S13. ^1H - ^1H COSY NMR spectrum (400 MHz) of **1** in d_4 -MeOH

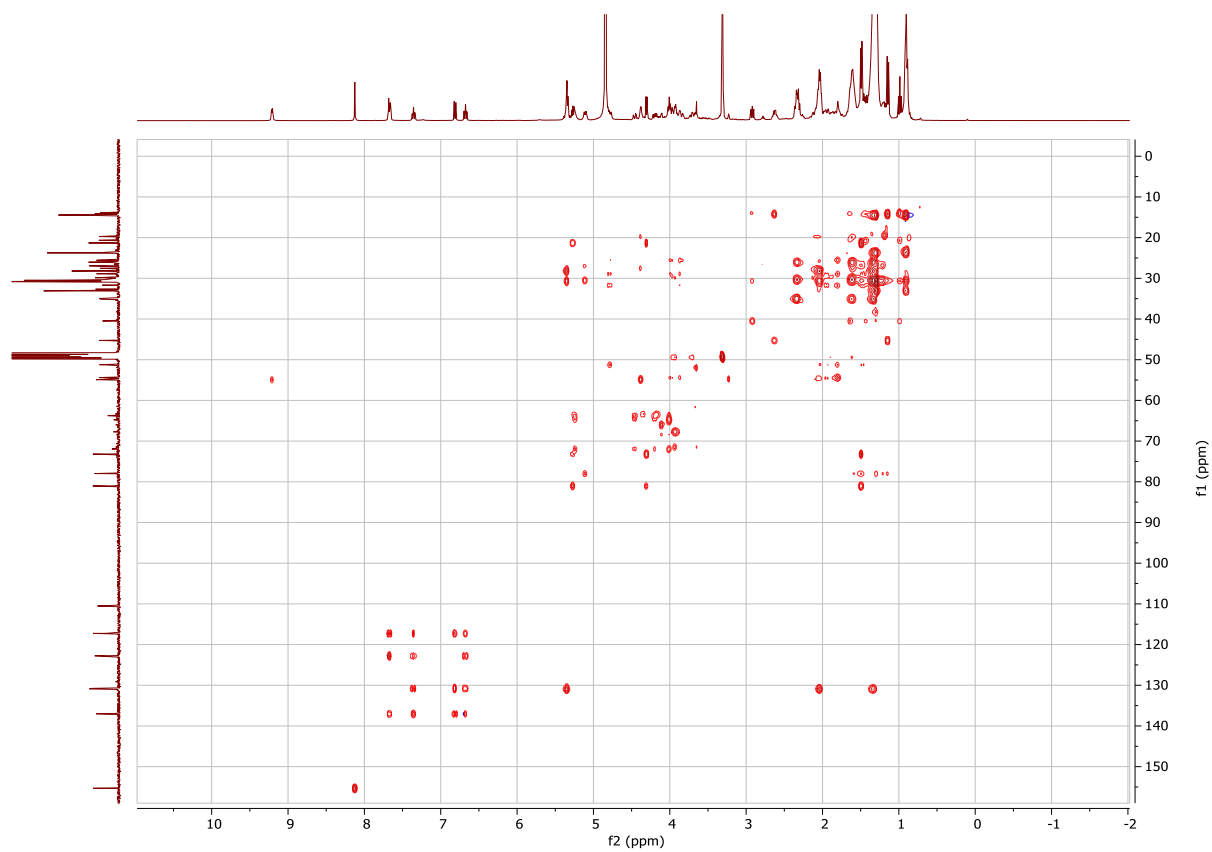


Figure S14. ^1H - ^{13}C HSQC-TOCSY NMR spectrum (400 MHz) of **1** in d_4 -MeOH

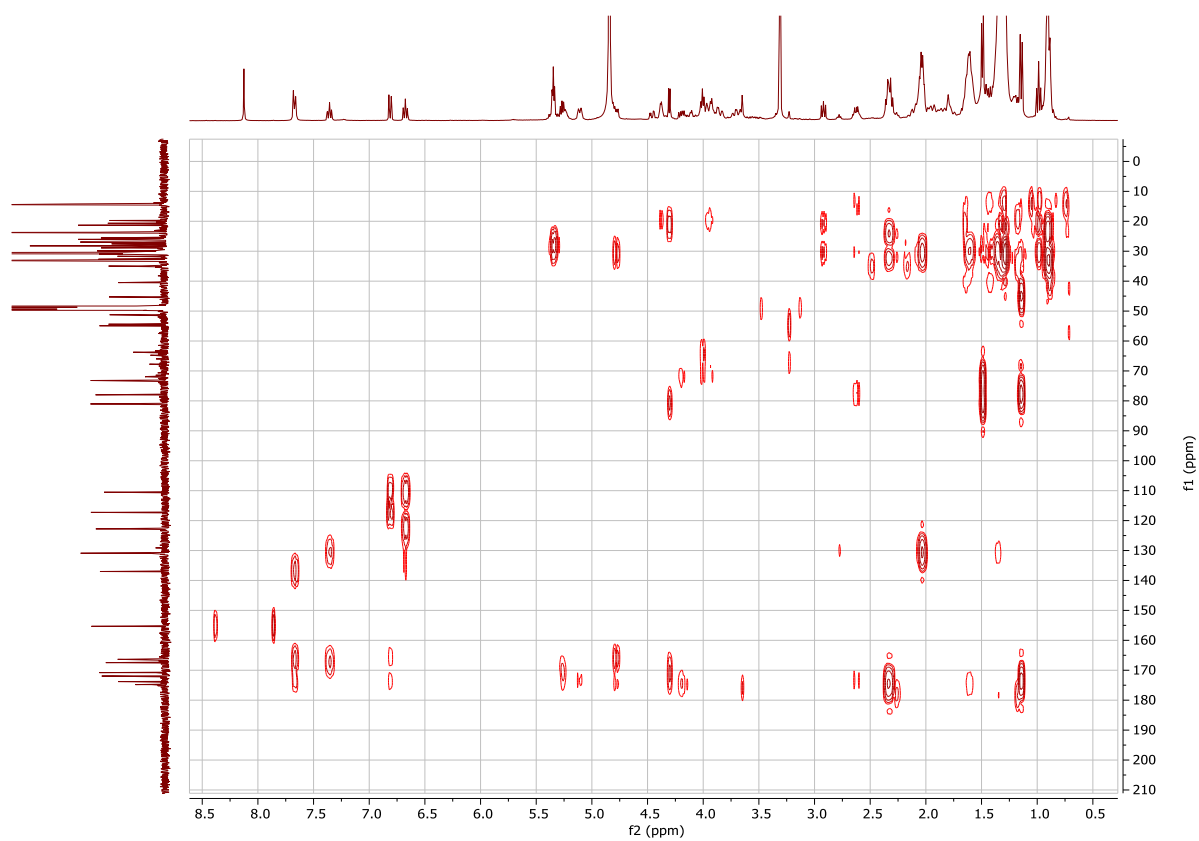


Figure S15. ^1H - ^{13}C HMBC NMR spectrum (400 MHz) of **1** in d_4 -MeOH

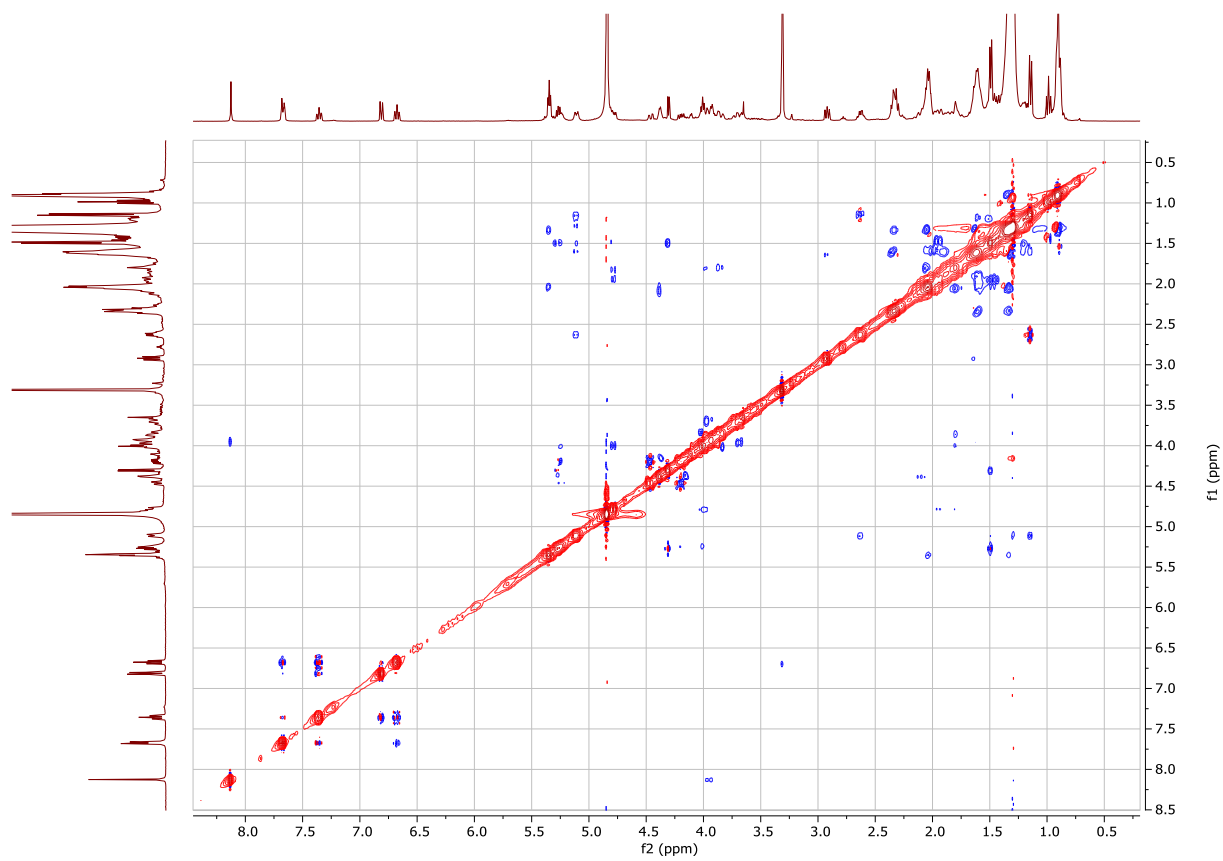


Figure S16. ^1H - ^1H NOESY NMR spectrum (400 MHz) of **1** in d_4 -MeOH

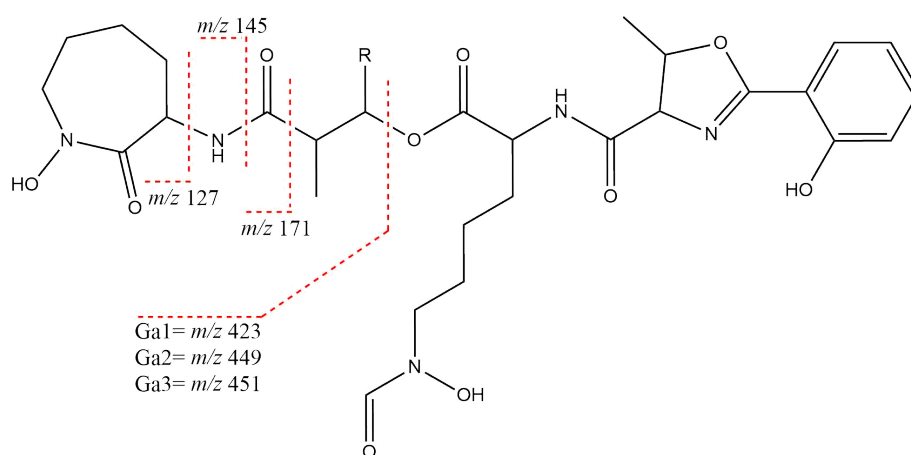
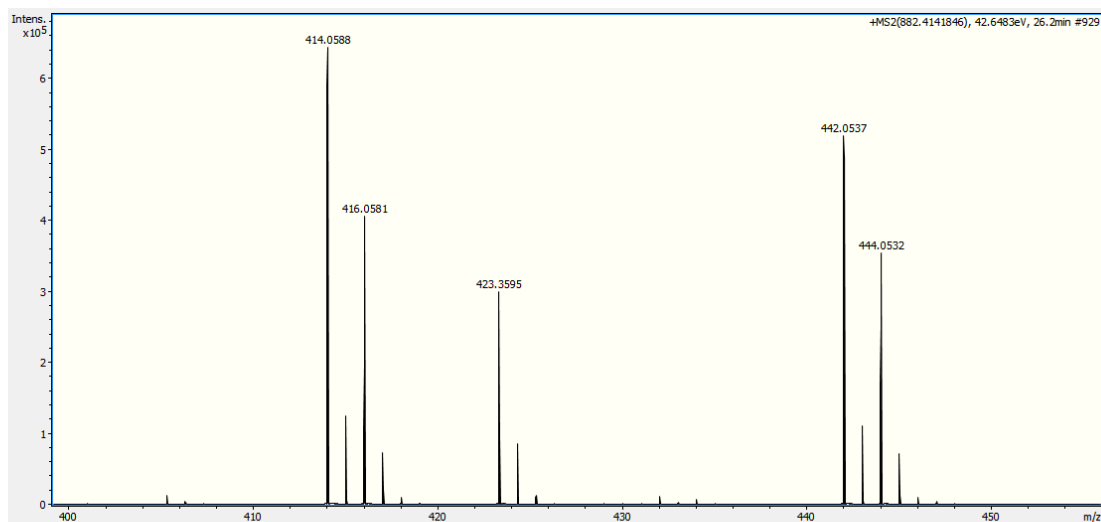


Figure S17. Top: Enlargement of HR-ESI-MS of terpenibactin A showing a fragment at m/z 423 and rationalization of MS/MS fragmentation of terpenibactins A-C (bottom).

Optical rotation value of **2**: $[\alpha]_{23}^D +57.6^\circ$ (c 0.33, MeOH)

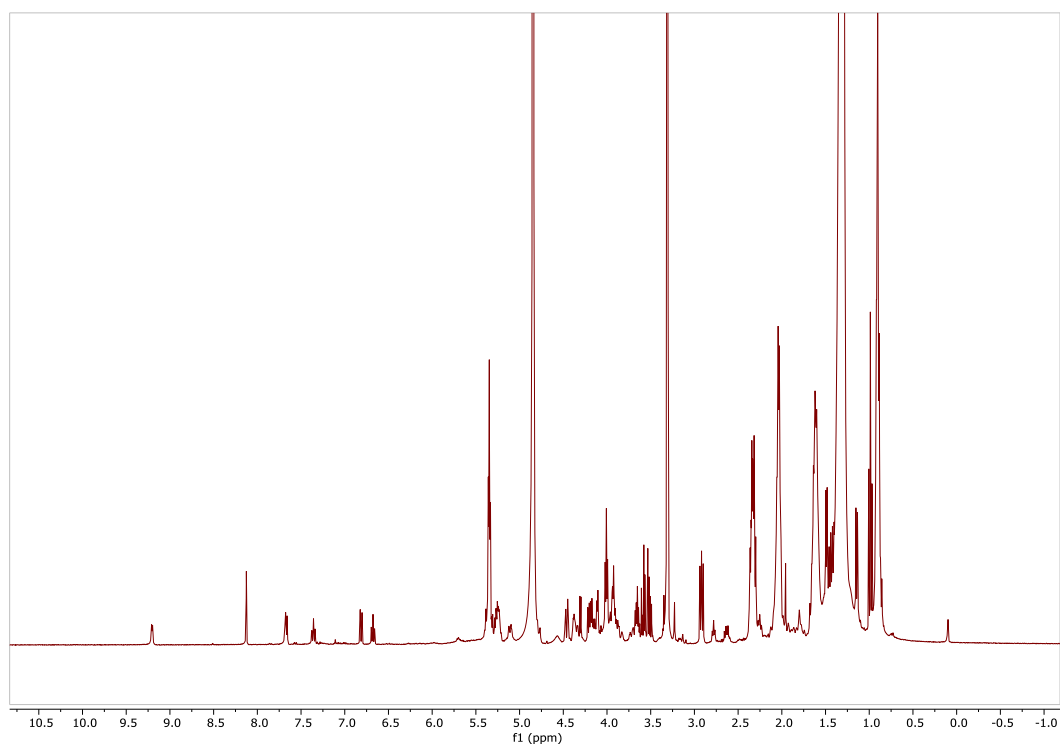


Figure S18. ¹H-NMR spectrum (400 MHz) of **2** in *d*₄-MeOH

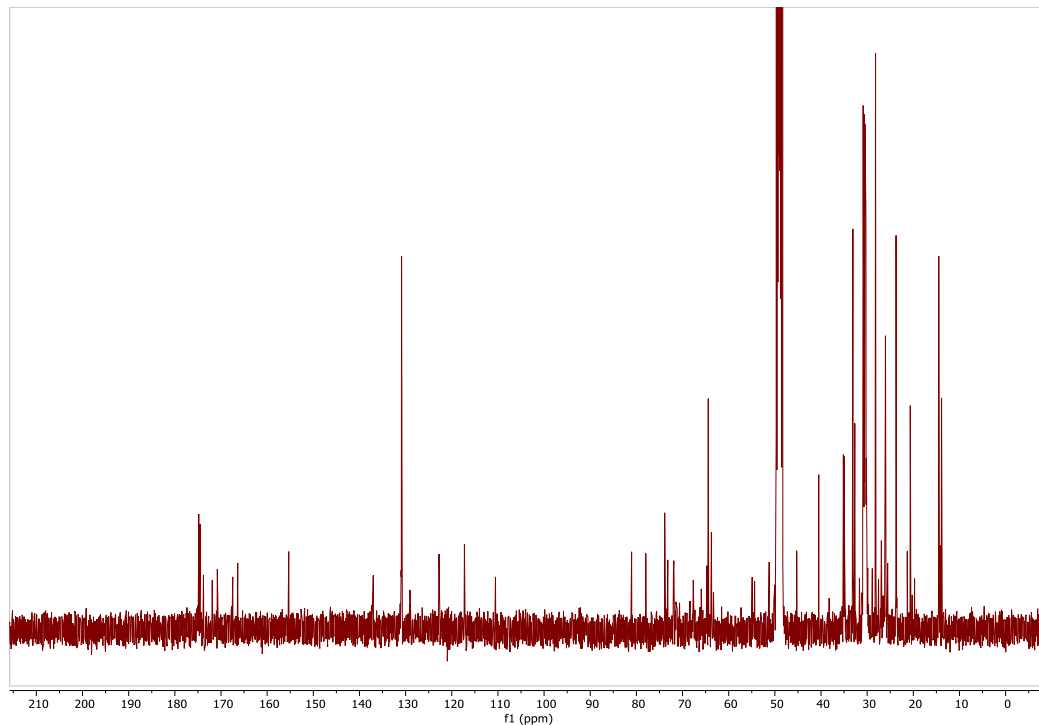


Figure S19. ¹³C-NMR spectrum (100 MHz) of **2** in *d*₄-MeOH

Optical rotation value of **3**: $[\alpha]_{23}^D +75.8^\circ$ (c 0.33, MeOH)

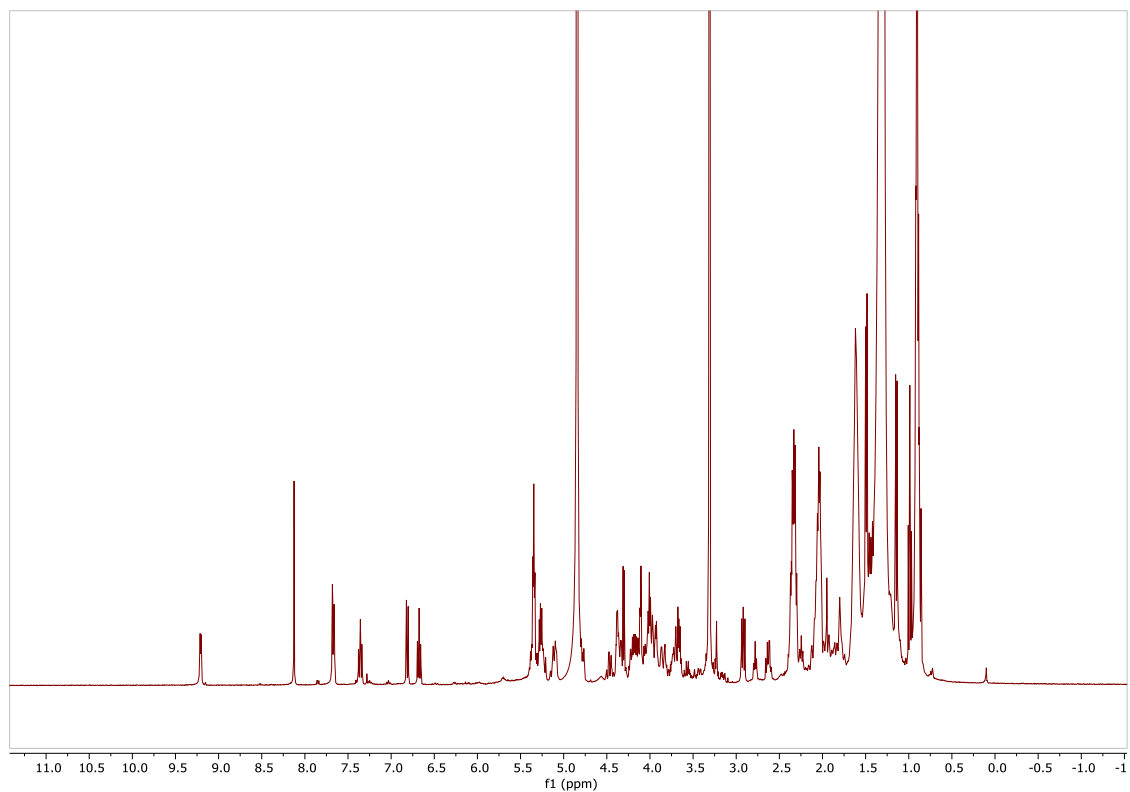


Figure S20. ¹H-NMR spectrum (400 MHz) of **3** in *d*₄-MeOH

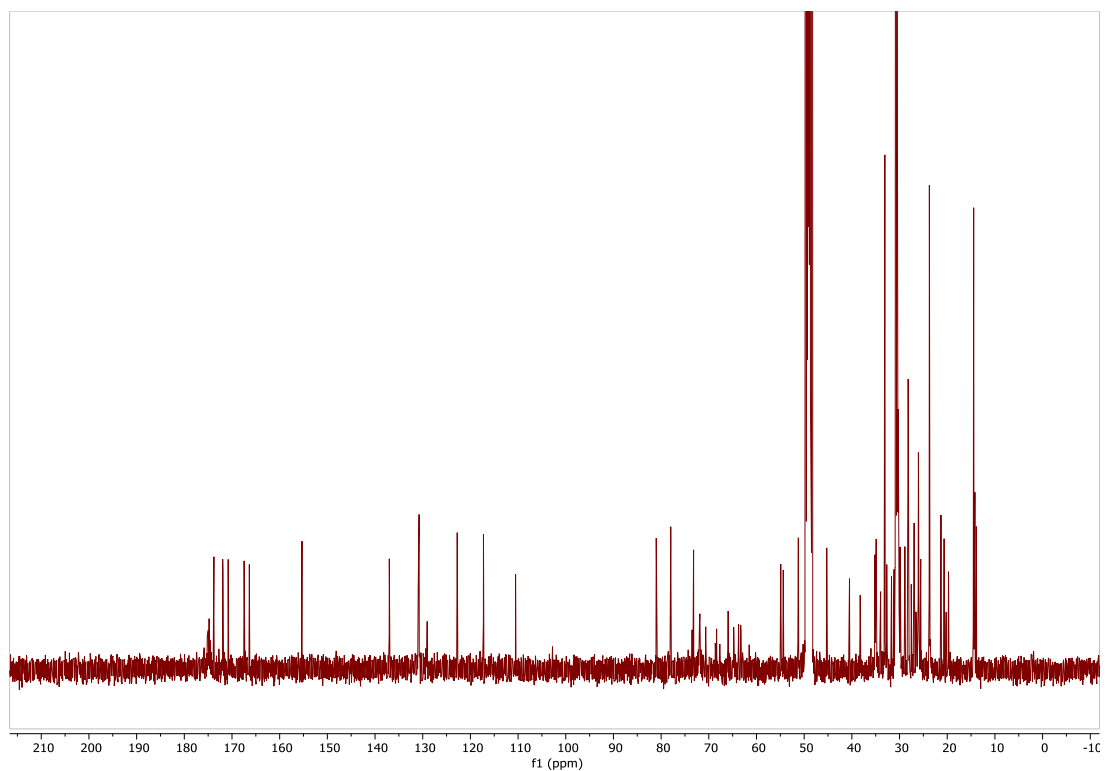


Figure S21. ¹³C-NMR spectrum (100 MHz) of **3** in *d*₄-MeOH

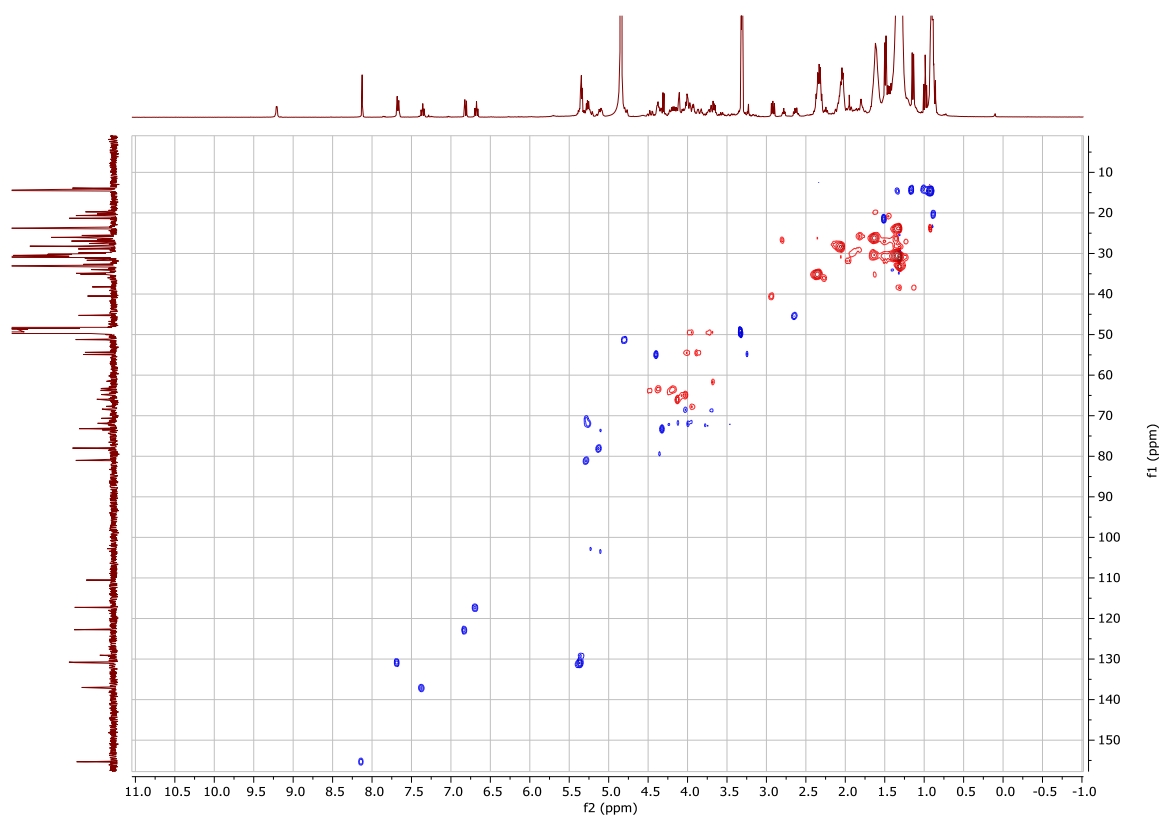


Figure S22. ^1H - ^{13}C HSQC NMR spectrum (400 MHz) of **3** in d_4 -MeOH

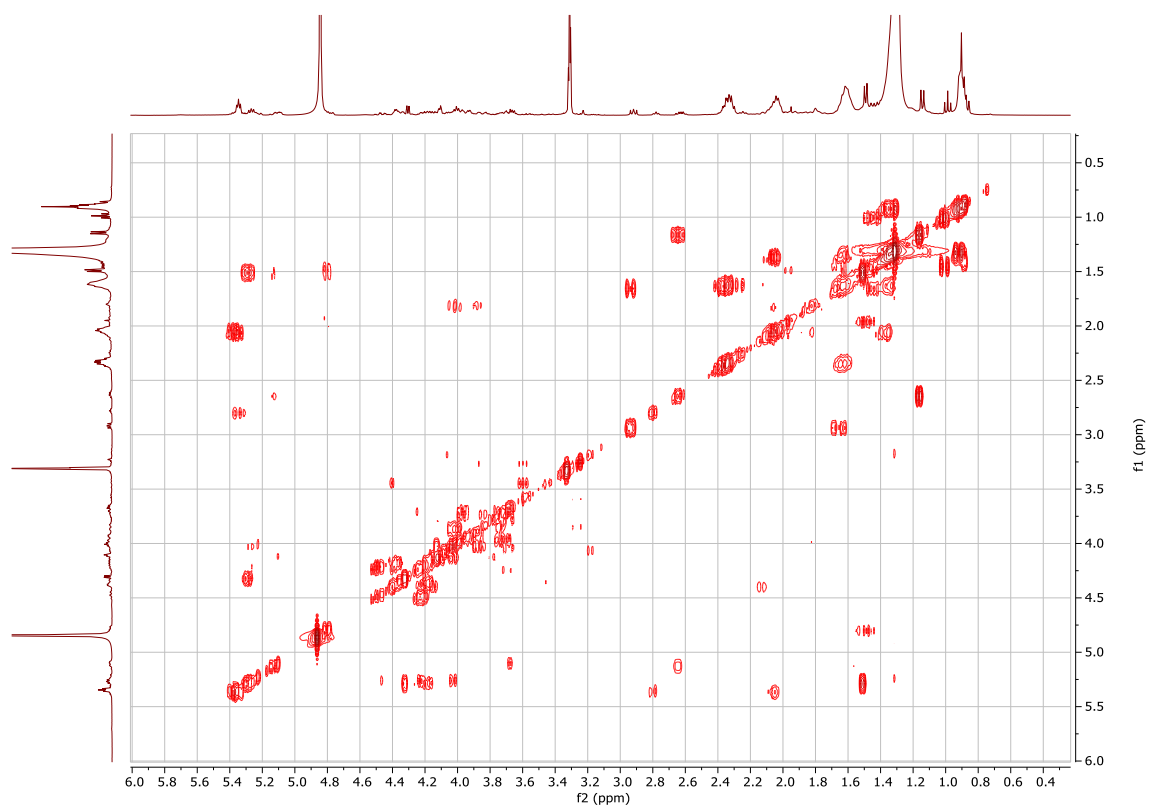


Figure S23. ^1H - ^1H COSY NMR spectrum (400 MHz) of **3** in d_4 -MeOH



Figure S24. ^1H - ^{13}C HSQC-TOCSY NMR spectrum (400 MHz) of **3** in d_4 -MeOH

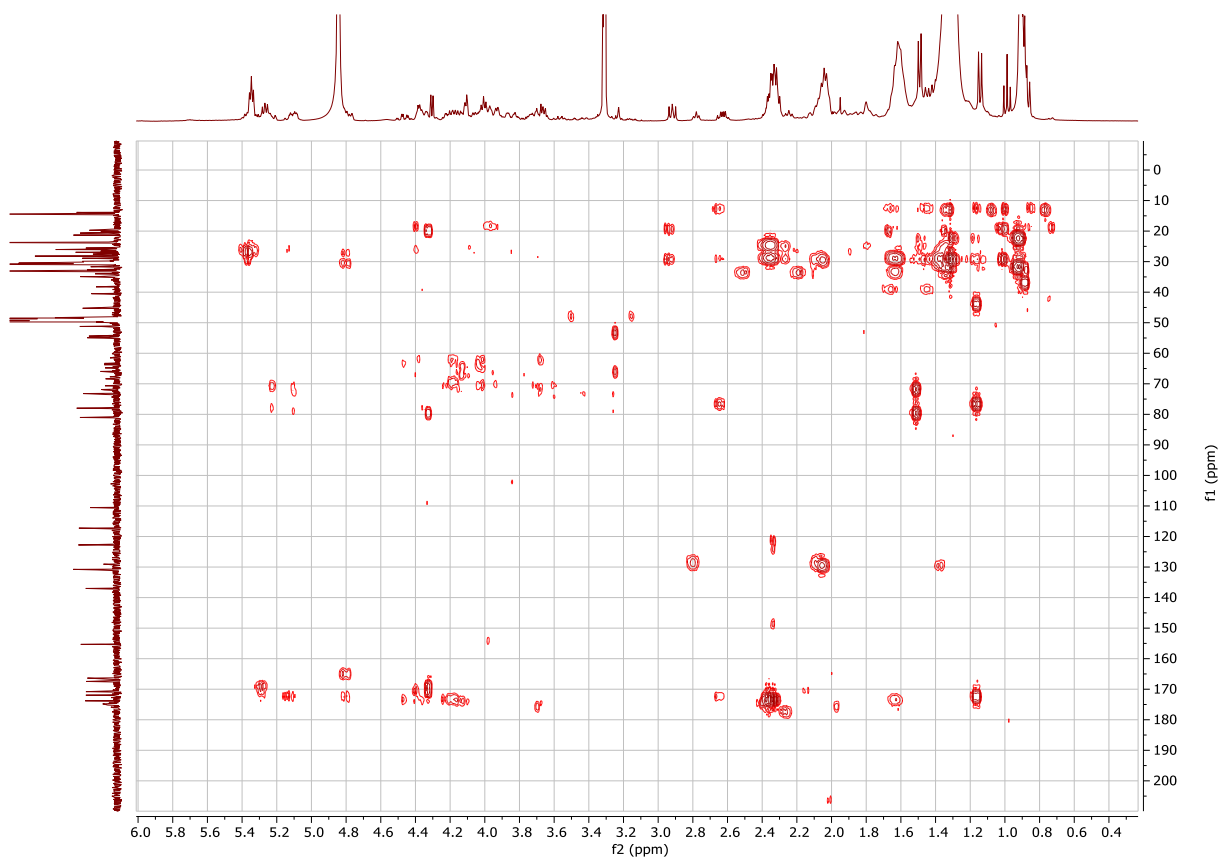


Figure S25. ^1H - ^{13}C HMBC NMR spectrum (400 MHz) of **3** in d_4 -MeOH

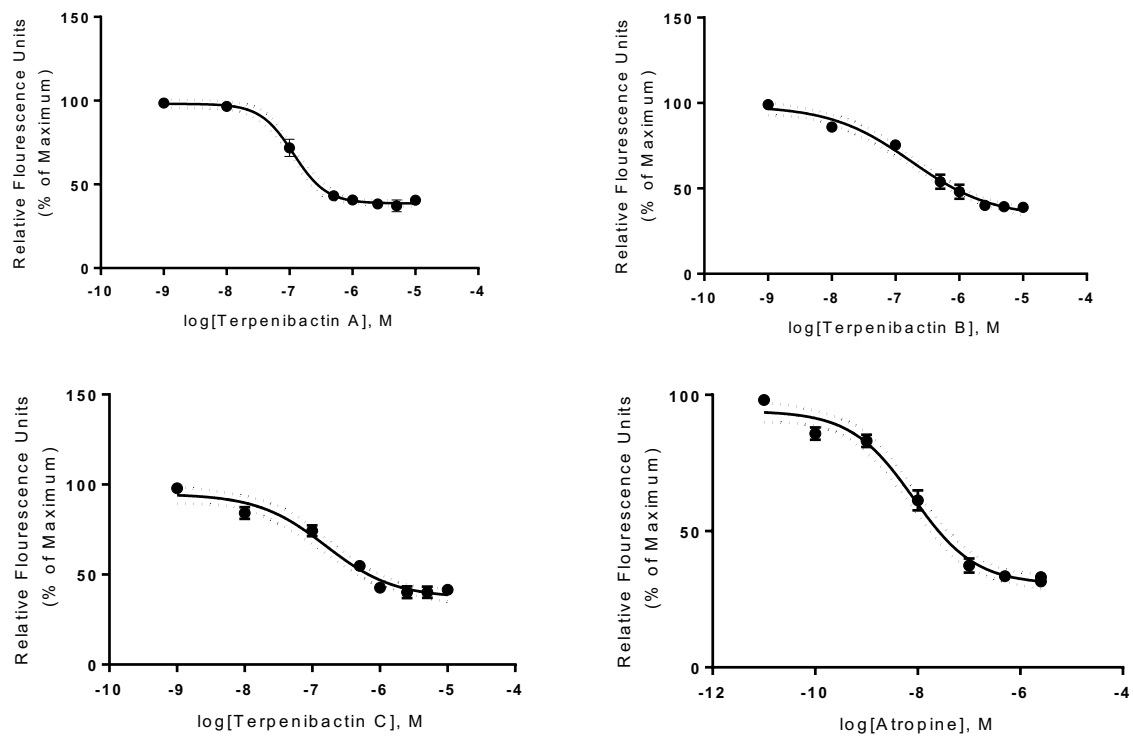


Figure S26. Calcium mobilisation assay in M3-expressing Chem-1 cells (data represent the mean of three independent experiments) using antagonist test compounds (terpenibactins A, B, and C) and positive antagonist control, atropine.