Supplementary Tables

Supplementary Table 1. Logic for internal class separation.

antiSMASH 4/5 annotation	BiG-SCAPE class
t1pks, T1PKS	PKS I
transatpks, t2pks, t3pks, otherks, hglks, transAT-PKS, transAT-PKS-like, T2PKS, T3PKS, PKS-like, hglE-KS and combinations of these with {t1pks, T1PKS} or themselves	PKS other
nrps, NRPS, NRPS-like, thioamide-NRP	NRPS
lantipeptide, thiopeptide, bacteriocin, linaridin, cyanobactin, glycocin, LAP, lassopeptide, sactipeptide, bottromycin, head_to_tail, microcin, microviridin, proteusin, lanthipeptide, lipolanthine, RaS-RiPP, fungal-RiPP and combinations of these	RiPPs
amglyccycl, oligosaccharide, cf_saccharide, saccharide and combinations of these	Saccharides
terpene	Terpene
any of {t1pks, T1PKS, transatpks, t2pks, t3pks, otherks, hglks, transAT-PKS, T2PKS, T3PKS, PKS-like, hglE-KS} + any of {nrps, NRPS, NRPS-like, thioamide-NRP}	PKS/NRPS Hybrids
acyl_amino_acids, arylpolyene, aminocoumarin, ectoine, butyrolactone, nucleoside, melanin, phosphoglycolipid, phenazine, phosphonate, other, cf_putative, resorcinol, indole, ladderane, PUFA, furan, hserlactone, fused, cf_fatty_acid, siderophore, blactam, fatty_acid, PpyS-KS, CDPS, betalactone, PBDE, tropodithietic-acid, NAGGN, halogenated and any combined annotation	Others
<any> or <none>	< mix >

Internal BGC classification used by BiG-SCAPE to separate the analysis using antiSMASH 4 annotations. If using the hybrids mode (default), BiG-SCAPE will also assign BGCs with combined annotations from antiSMASH to each individual class (e.g. a BGC annotated as "t1pks-terpene" will go to the Others, PKS I and Terpene BiG-SCAPE classes). GenBank files without antiSMASH annotations will be classified as Others.

Supplementary Table 2. Size of each correlation round resulting from various BiG-SCAPE outputs.

Correlation round	Correlatable GCFs	Ion-GCF hypotheses
glocal 0.30	4,474	26,056,576
glocal 0.50	5,067	29,510,208
global 0.30	4,237	24,676,288
global 0.50	5,176	30,145,024

Supplementary Table 3. Correlation scores for known ion-GCF pairs in the dataset across the four correlation rounds.

Global 0.30

Natural Product	Known Ion-GCF Score
Benarthin	-177
CE-108	453
Chlortetracycline	374
Desertomycin	319
Enterocin	-
Oxytetracycline	248
Rimosamide	297
Tambromycin	370
Tyrobetaine	271

Global 0.50

Natural Product	Known Ion-GCF Score
Benarthin	-162
CE-108	362
Chlortetracycline	381
Desertomycin	339
Enterocin	381
Oxytetracycline	254
Rimosamide	383
Tambromycin	376
Tyrobetaine	278

Glocal 0.30

Natural Product	Known Ion-GCF Score
Benarthin	-141
CE-108	142
Chlortetracycline	378
Desertomycin	324
Enterocin	359
Oxytetracycline	451
Rimosamide	401
Tambromycin	353
Tyrobetaine	375

Glocal 0.50

Natural Product	Known Ion-GCF Score
Benarthin	-138
CE-108	400
Chlortetracycline	381
Desertomycin	338
Enterocin	362
Oxytetracycline	434
Rimosamide	400
Tambromycin	356
Tyrobetaine	378

Supplementary Table 4. Occurrences of *tauD* homologues in secondary metabolism BGCs as reported in MIBiG.

No	MIBiG	Compound	Class	Producer Organism
1	BGC0000163_ACR50790	tetronasin	Polyketide	*Streptomyces longisporoflavus*
2	BGC0000287_AAG05698	2-amino-4-methoxy-trans-3-butenoic acid	NRP	*Pseudomonas aeruginosa* PAO1
3	BGC0000653_ADO85576	pentalenolactone	Terpene	*Streptomyces arenae*
4	BGC0000654_ABB69741	phenalinolactone	Saccharide / Terpene	*Streptomyces* sp. Tu6071
5	BGC0000678_BAC70706	pentalenolactone	Terpene	*Streptomyces avermitilis* MA-4680 = NBRC 14893
6	BGC0000715_ABW87795	spectinomycin	Saccharide	*Streptomyces spectabilis*
7	BGC0000846_ctg1_orf9	tabtoxin	Other	*Pseudomonas syringae*
8	BGC0000961_ABC36162	bactobolin	NRP / Polyketide	*Burkholderia thailandensis* E264
9	BGC0001070_CAN89617	kirromycin	NRP / Polyketide	*Streptomyces collinus* Tu 365
10	BGC0001140_ACO31277	platensimycin / platencin	Terpene	*Streptomyces platensis*
11	BGC0001140_ACO31282	platensimycin / platencin	Terpene	*Streptomyces platensis*
12	BGC0001156_ADD83004	platencin	Terpene	*Streptomyces platensis*
13	BGC0001183_AGC09525	lobophorin	Polyketide	*Streptomyces* sp. FXJ7.023
14	BGC0001205_KGO40482	communesin	Polyketide	*Penicillium expansum*
15	BGC0001205_KGO40485	communesin	Polyketide	*Penicillium expansum*
16	BGC0001760.1 WP_004571774.1	rimosamide	NRP	*Streptomyces rimosus subsp. rimosus* ATCC 10970

All *tauD* homologues from the same EvoMining expansion that include the *tauD* from rimosamide (annotated with circles in Supplementary Figure 8.

Supplementary Table 5. Clustering methods used to identify potential BGC families in the training networks.

Clustering method	Reference	Notes
Affinity propagation (AP)	(Frey & Dueck, 2007)[1]	scikit-learn (Pedregosa et al., 2011)[2]
Fast greedy (FG)	(Clauset, Newman, & Moore, 2004)[3]	igraph
Label propagation (LP)	(Raghavan, Albert, & Kumara, 2007)[4]	igraph
Walktrap (WT)	(Pons & Latapy, 2005)[5]	igraph
Infomap (IM)	(Martin Rosvall & Bergstrom, 2008; M. Rosvall, Axelsson, & Bergstrom, 2009)[6,7]	igraph and original implementation
Louvain community (LV)	(Blondel, Guillaume, & Lambiotte, 2008)[8]	igraph and original implementation
Markov Clustering (MCL)	(Van Dongen, 2000)[9]	original implementation
Topological Overlap Matrix (TOM)	(Zhang & Horvath, 2005)[10]	WGCNA package (Langfelder & Horvath, 2008)[11]

Supplementary Table 6. Compute times for BiG-SCAPE calculations for datasets of multiple sizes.

BGCs	Time in seconds (HH:MM:SS)
10	62.2 (00:01:02)
100	102.8 (00:01:42)
1000	897.5 (00:14:57)
10000	15244.6 (04:14:04)
11618	16052.1 (4:27:32)

Compute times for different numbers of randomly picked BGCs from the Expanded Actinobacteria set and the assembled set from the comparison with the Doroghazi et al. 2014 approach (last row). BGCs were chosen as complete (i.e. not flagged by antiSMASH as 'contig_edge'. BiG-SCAPE was run with default parameters using 16 threads (`--c 16`) Each run was calculated from scratch.

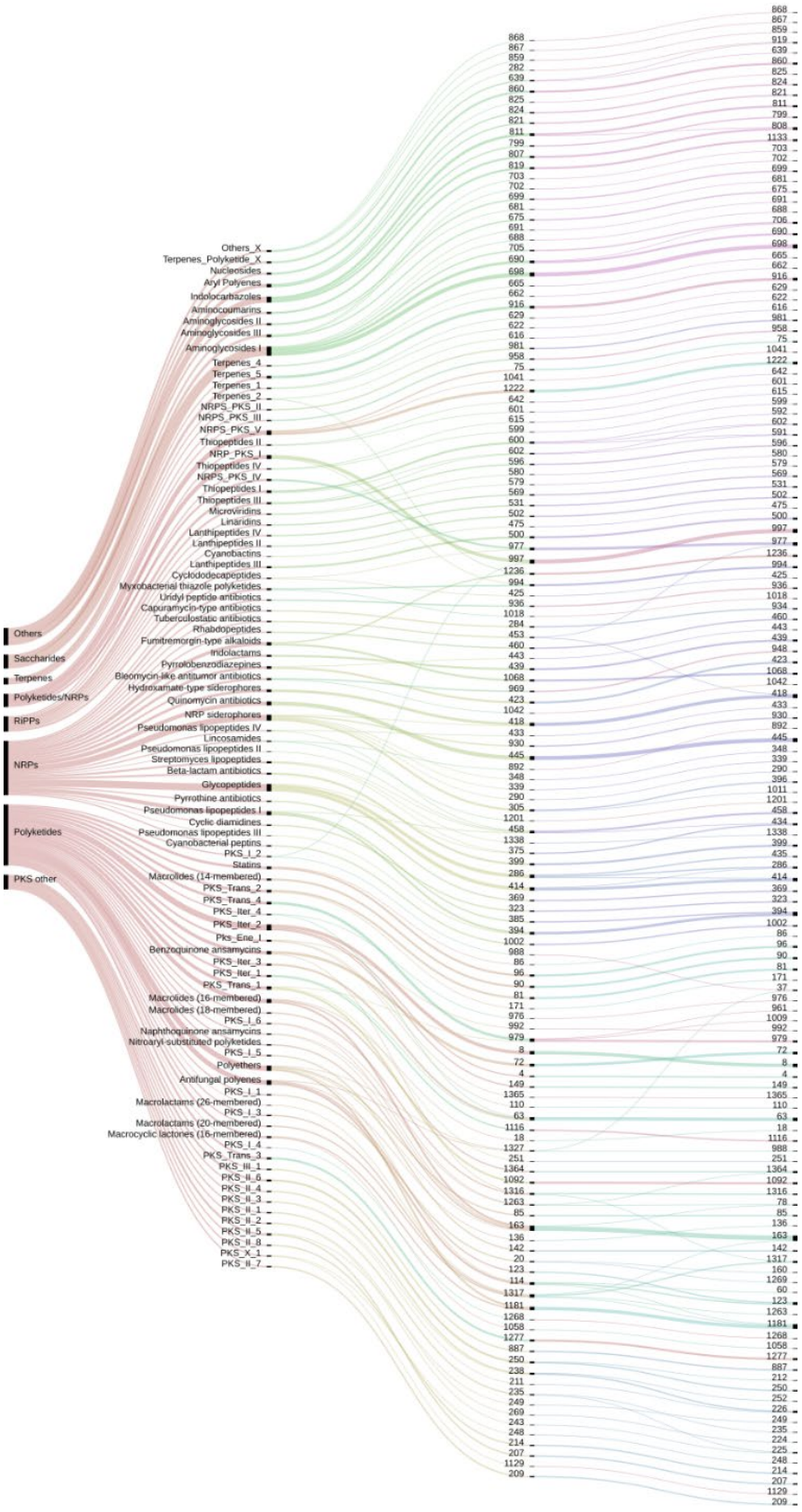
Supplementary Table 7. Default Anchor Domain List used in the DSS index.

PF00668	Condensation domain [NRPS]
PF00501	AMP-binding enzyme [NRPS]
PF00109	Beta-ketoacyl synthase N-terminal [PKS]
PF02801	Beta-ketoacyl synthase C-terminal [PKS]
PF01397	Terpene synthase, N-terminal domain (Terpene_synth) [Terpene]
PF03936	Terpene synthase family, metal binding domain (Terpene_synth_C) [Terpene]
PF00195	Chalcone and stilbene synthases, N-terminal domain (Chal_sti_synt_N)
PF02797	Chalcone and stilbene synthases, C-terminal domain (Chal_sti_synt_C)
PF05147	Lanthionine synthetase C-like protein (LANC_like) [lantipeptide/RiPP]
PF00494	Squalene/phytoene synthase (SQS_PSY) [Terpene]
PF00432	Prenyltransferase and squalene oxidase repeat (Prenyltrans) [Indole alkaloids]
PF02624	YcaO cyclodehydratase, ATP-ad MG2+-binding (YcaO) [RiPP]

Supplementary Figures

Supplementary Figure 1. Alluvial plot depicting BiG-SCAPE's assignment of MIBiG BGCs to Gene Cluster Families against the manual curated groups in Supplementary Dataset.

To assess the accuracy of clustering with curated data, we employed purity, a criterion of clustering quality which counts the most frequent label (manually curated group) within each cluster (GCF). A purity score close to 1 means that most clusters have a single label.
BiG-SCAPE was run on the MIBiG database with hybrid mode disabled to prevent BGCs appearing in more than one BiG-SCAPE class, and cutoff value c=0.75 (after the analysis described in Online Methods: Clustering algorithm optimization). First column: Biosynthetic classes as defined in BiG-SCAPE; second column: Compound Groups from curated table; third column: global mode (purity: 0.881); fourth column: glocal mode (purity: 0.880). Colors have been assigned randomly.
For both modes, purity was calculated using the following formula: $P = \frac{1}{N}\sum_k \max_j |\omega_k \cap c_j|$ where $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ is the set of different GCFs. Only GCFs of size 2 and larger were considered to prevent GCFs comprised of only one BGC to increase purity. For global mode, Ω=82 GCFs and for glocal mode, Ω=79 GCFs (from a total of 134 GCFs in both cases). N is the total number of BGCs analyzed (i.e. all from GCFs of size 2 and larger; N=244 for global mode, N=235 for glocal mode). Finally, $CG = \{c_1, c_2, \dots, c_j\}$ is the set of curated groups (89 for both cases). Two manually defined classes (Lanthipeptides_I and Microcins) were not considered because the compounds in those classes do not have fully elucidated chemical structures.


Others
Saccharides
Terpenes
Polyketides/NRPs
RiPPs
NRPs
Polyketides
PKS other
Others_X
Terpenes_Polyketide_X
Nucleosides
Aryl Polyenes
Indolocarbazoles
Aminocoumarins
Aminoglycosides II
Aminoglycosides III
Aminoglycosides I
Terpenes_4
Terpenes_5
Terpenes_1
Terpenes_2
NRPS_PKS_II
NRPS_PKS_III
NRPS_PKS_V
Thiopeptides II
NRP_PKS_I
Thiopeptides IV
NRPS_PKS_IV
Thiopeptides I
Thiopeptides III
Microviridins
Linaridins
Lanthipeptides IV
Lanthipeptides II
Cyanobactins
Lanthipeptides III
Cyclododecapeptides
Myxobacterial thiazole polyketides
Uridyl peptide antibiotics
Capuramycin-type antibiotics
Tuberculostatic antibiotics
Rhabdopeptides
Fumitremorgin-type alkaloids
Indolactams
Pyrrolobenzodiazepines
Bleomycin-like antitumor antibiotics
Hydroxamate-type siderophores
Quinomycin antibiotics
NRP siderophores
Pseudomonas lipopeptides IV
Lincosamides
Pseudomonas lipopeptides II
Streptomyces lipopeptides
Beta-lactam antibiotics
Glycopeptides
Pyrrothine antibiotics
Pseudomonas lipopeptides I
Cyclic diamidines
Pseudomonas lipopeptides III
Cyanobacterial peptins
PKS_I_2
Statins
Macrolides (14-membered)
PKS_Trans_2
PKS_Trans_4
PKS_Iter_4
PKS_Iter_2
Pks_Ene_I
Benzoquinone ansamycins
PKS_Iter_3
PKS_Iter_1
PKS_Trans_1
Macrolides (16-membered)
Macrolides (18-membered)
PKS_I_6
Naphthoquinone ansamycins
Nitroaryl-substituted polyketides
PKS_I_5
Polyethers
Antifungal polyenes
PKS_I_1
Macrolactams (26-membered)
PKS_I_3
Macrolactams (20-membered)
Macrocyclic lactones (16-membered)
PKS_I_4
PKS_Trans_3
PKS_III_1
PKS_II_6
PKS_II_4
PKS_II_3
PKS_II_1
PKS_II_2
PKS_II_5
PKS_II_8
PKS_X_1
PKS_II_7

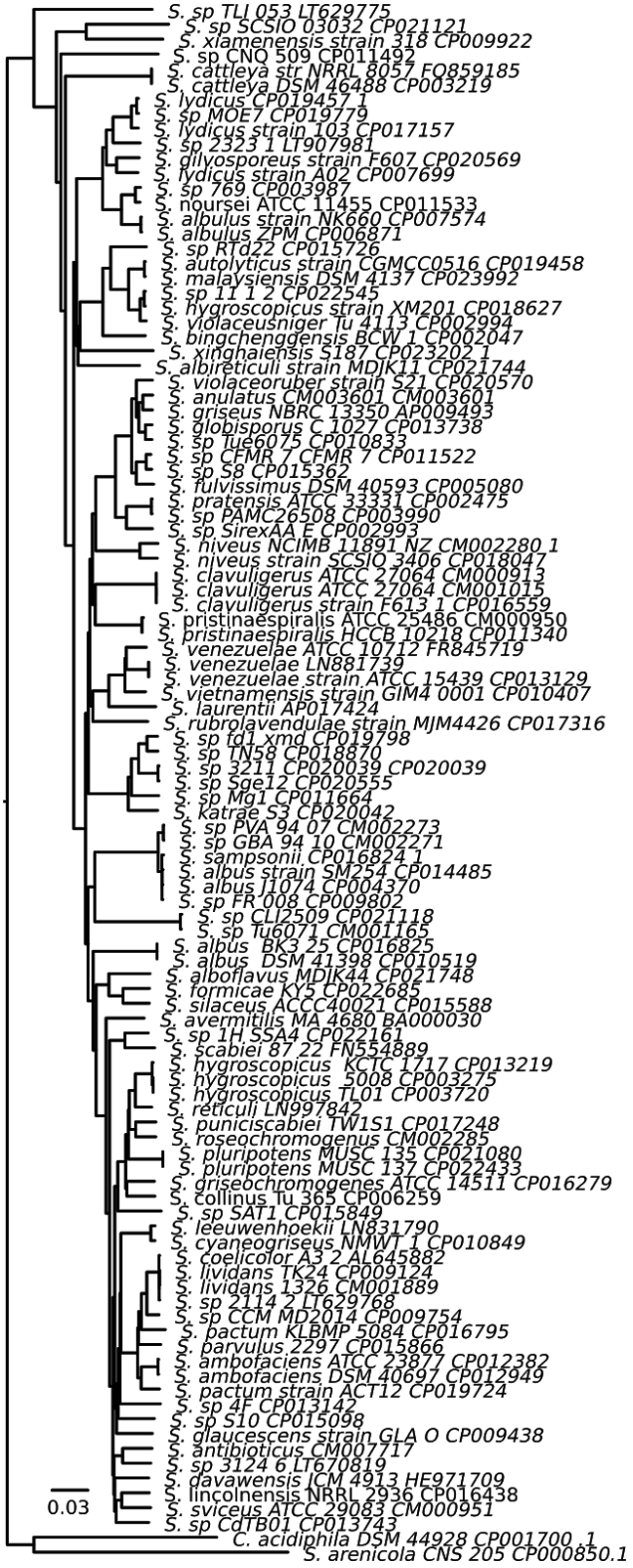
Supplementary Figure 2. Binary correlation score chart between ions and GCFs used for metabologenomics.

	+ Ion	- Ion
+ GCF	+ 10	+ 0
- GCF	- 10	+ 1

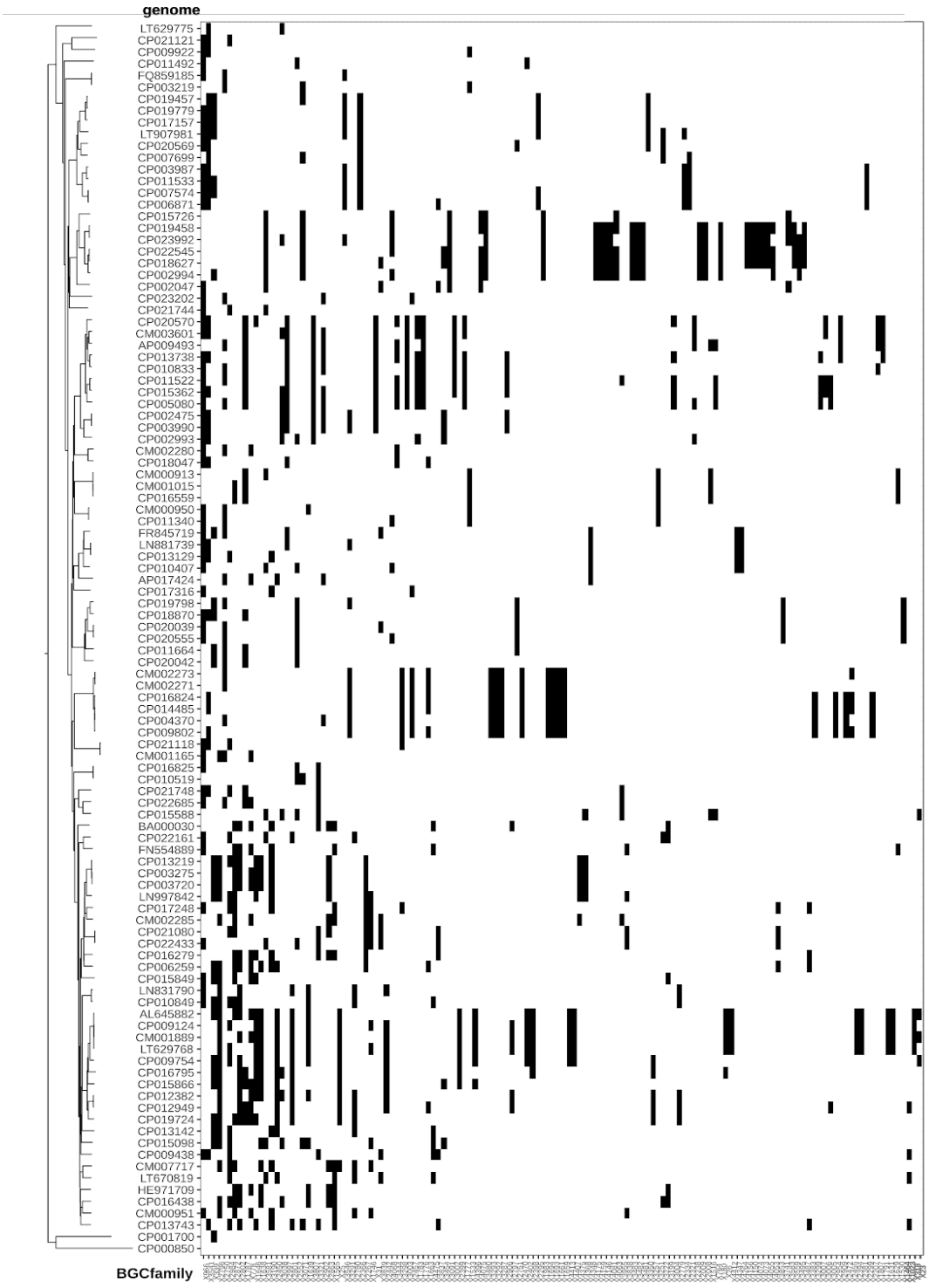
Ion-GCF pairs in correlation rounds were assigned correlation scores by applying a binary scoring method as previously described[12]. Each of the values that comprise the correlation score were selected after considering natural product production patterns in Actinobacteria. The '+10' score was implemented to improve the GCF-ion association score when strains were observed with a "GCF present, ion present" phenotype. A decision to assign a score of '+0' in instances of "GCF present, ion absent" was made to avoid penalizing the GCF-ion association when the GCF was silent in that particular strain in the screened growth conditions. A score of '-10' was given for the observation of "GCF absent, ion present" since this is indicative of either an ion that is associated with a different GCF, an ion associated with growth media, or spectral noise. Furthermore, a score of +1 in cases of "GCF absent, ion absent" was implemented as a metric to measure improvements in correlative power as the dataset grows. The scoring system was verified experimentally using a series of known compound-BGC pairs.

Supplementary Figure 3. Verified clusters encoding benarthin biosynthesis. Signature gene(s) highlighted.

Supplementary Figure 4. Phylogenomic reconstruction of *103* complete *Streptomyces* genomes with outgroups *Catenulispora acidiphila* CP001700.1 and *Salinispora arenicola* CP000850.1.

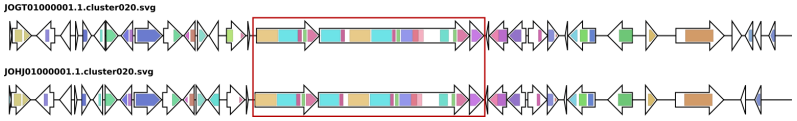


Supplementary Figure 5. BiG-SCAPE / CORASON GCFs of the closed *Streptomyces* genomes.

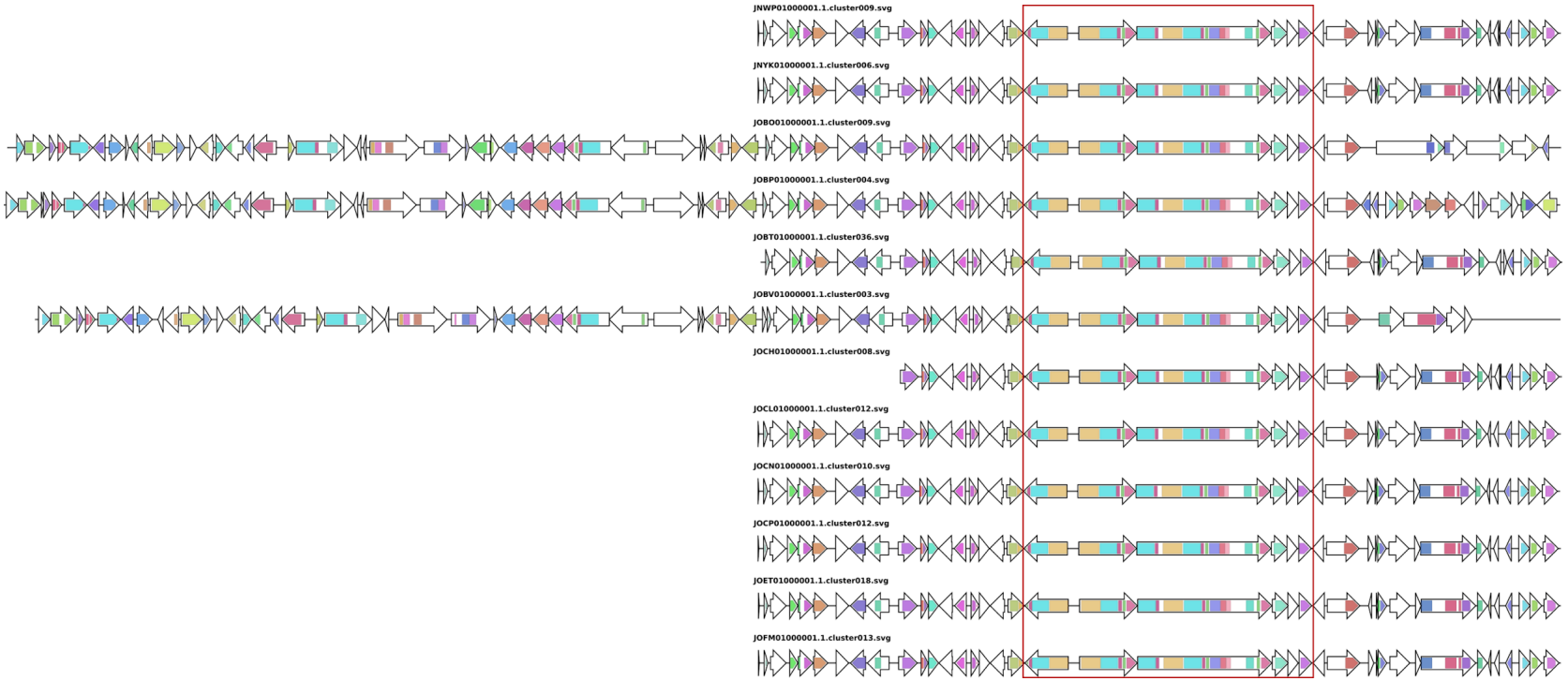


BiG-SCAPE/CORASON GCFs (columns) sorted by frequency of appearance dispersed across the phylogenetic reconstruction of 103 *Streptomyces* closed genomes. Species names can be found clearer in the large phylogenetic tree in Supplementary Figure 4. BGCs present in less than 5 species are not shown in the heatplot (2036 out of 3184 total BGCs). Only species were used to count the size of each GCF (i.e. BGCs from the same species that were clustered in the same GCF only incremented the size of that GCF by one). BiG-SCAPE network was calculated using parameters: c=0.3, hybrid mode: off, distance mode: glocal and --mibig (1.3). MIBiG BGCs are not shown in this analysis.

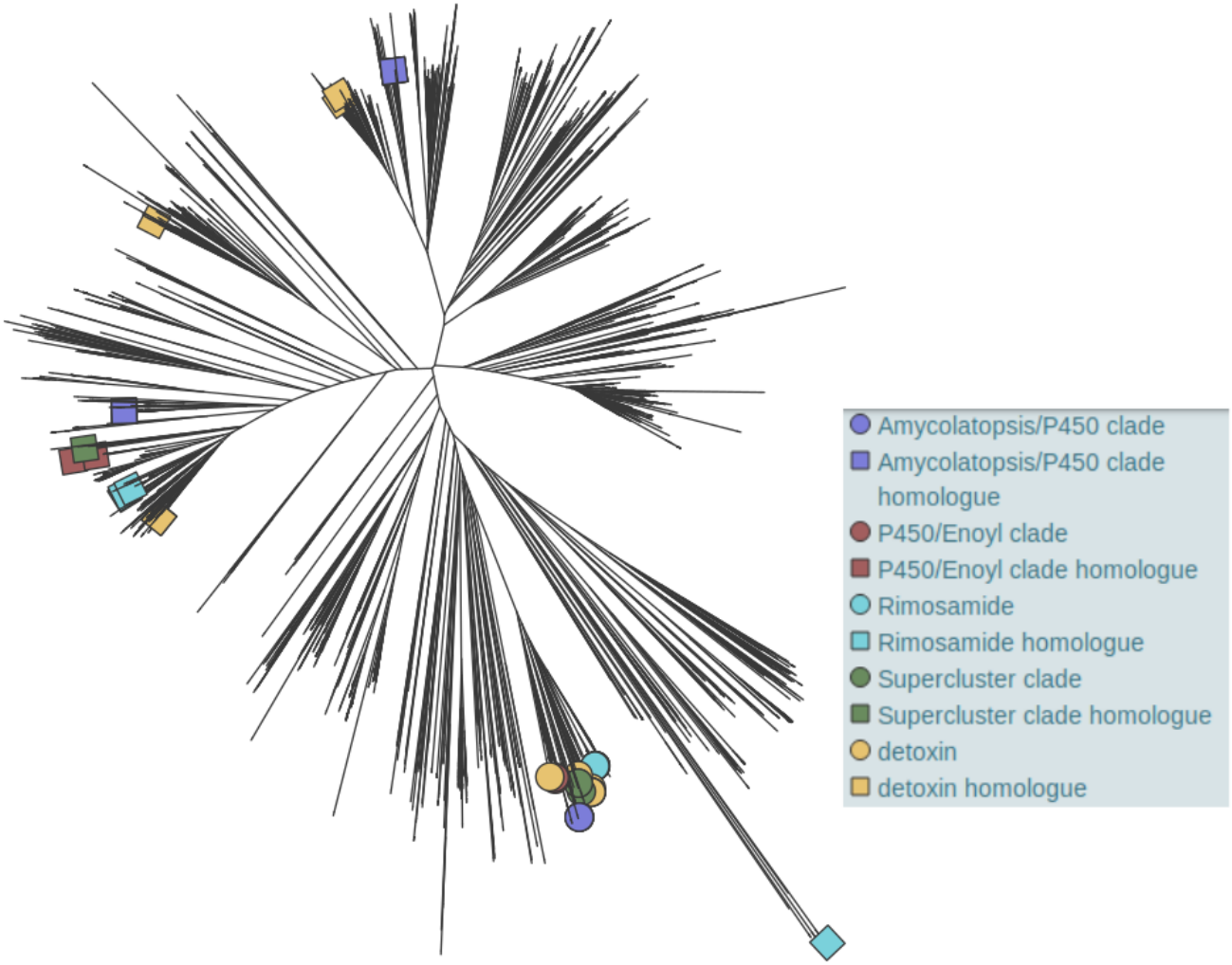
Supplementary Figure 6. Verified clusters encoding detoxin biosynthesis. Signature gene(s) highlighted.



Supplementary Figure 7. Verified clusters encoding rimosamide biosynthesis. Signature gene(s) highlighted.

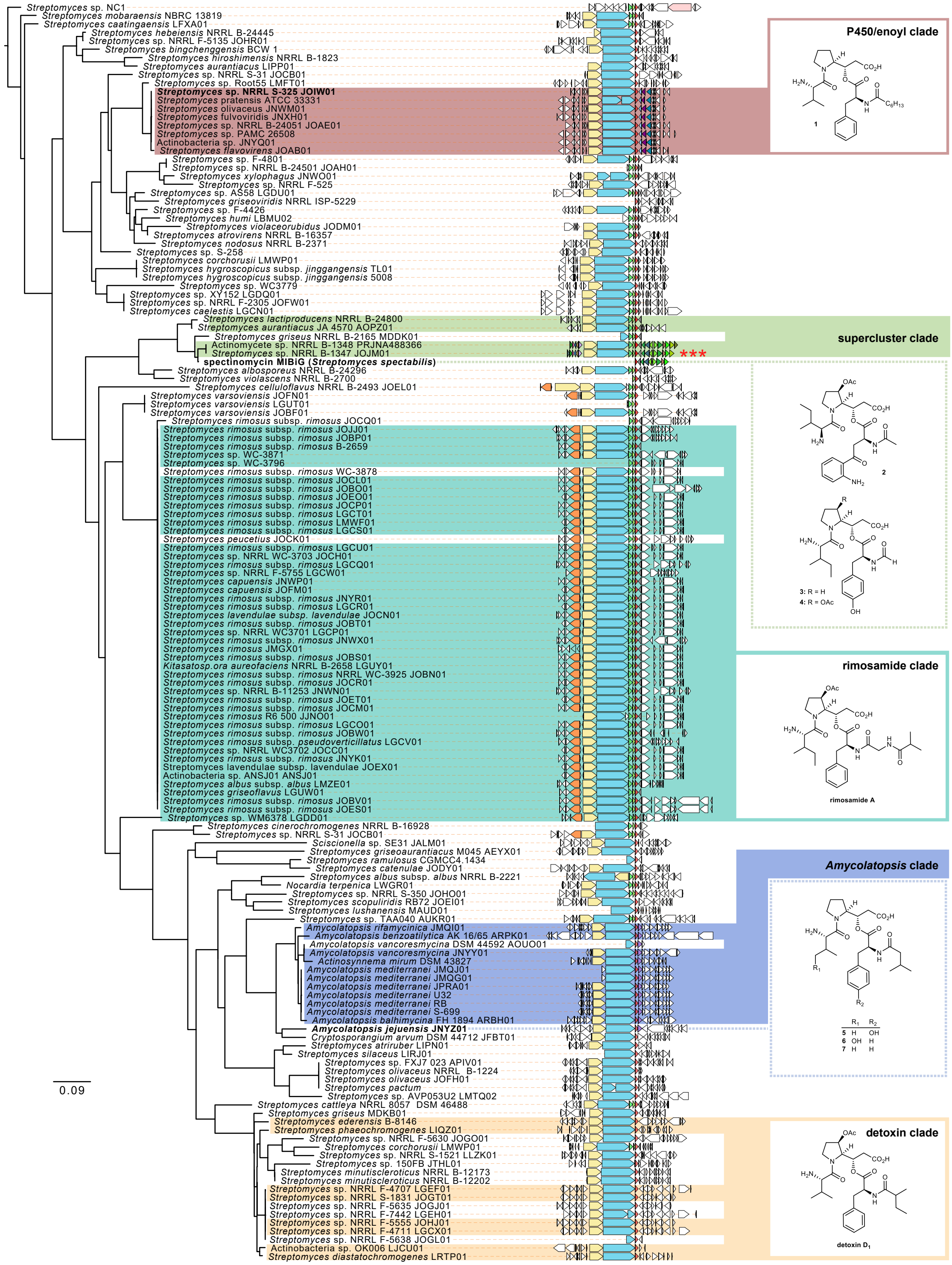


Supplementary Figure 8. *tauD* Actinobacteria EvoMining expansions tree.

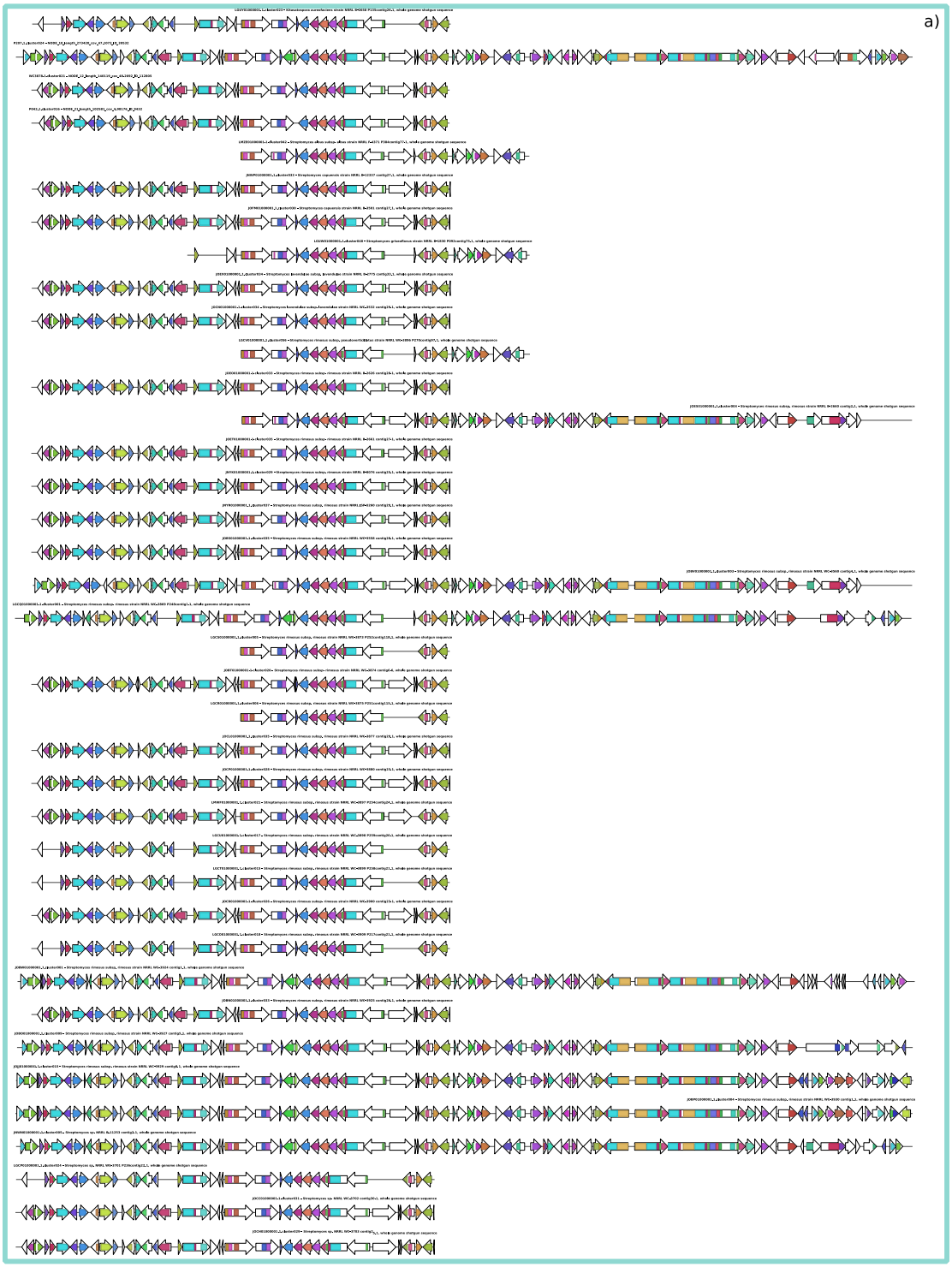


A branch containing several *tauD* homologues identified as part of MIBiG BGCs is shown in dark gray and includes *tauD* homologues from the known rimosamide and detoxin BGCs, indicated by turquoise and beige circles. The tree was generated from our Actinobacterial genome database using *tauD* from *E.coli*. The same *tauD* was used as the query gene for CORASON analysis. Colors in the tree match the BiG-SCAPE-defined GCFs shown in Figure. 5. This tree and metadata are available for further exploration at microreact[13] in the site https://microreact.org/project/H1UuQE0qm

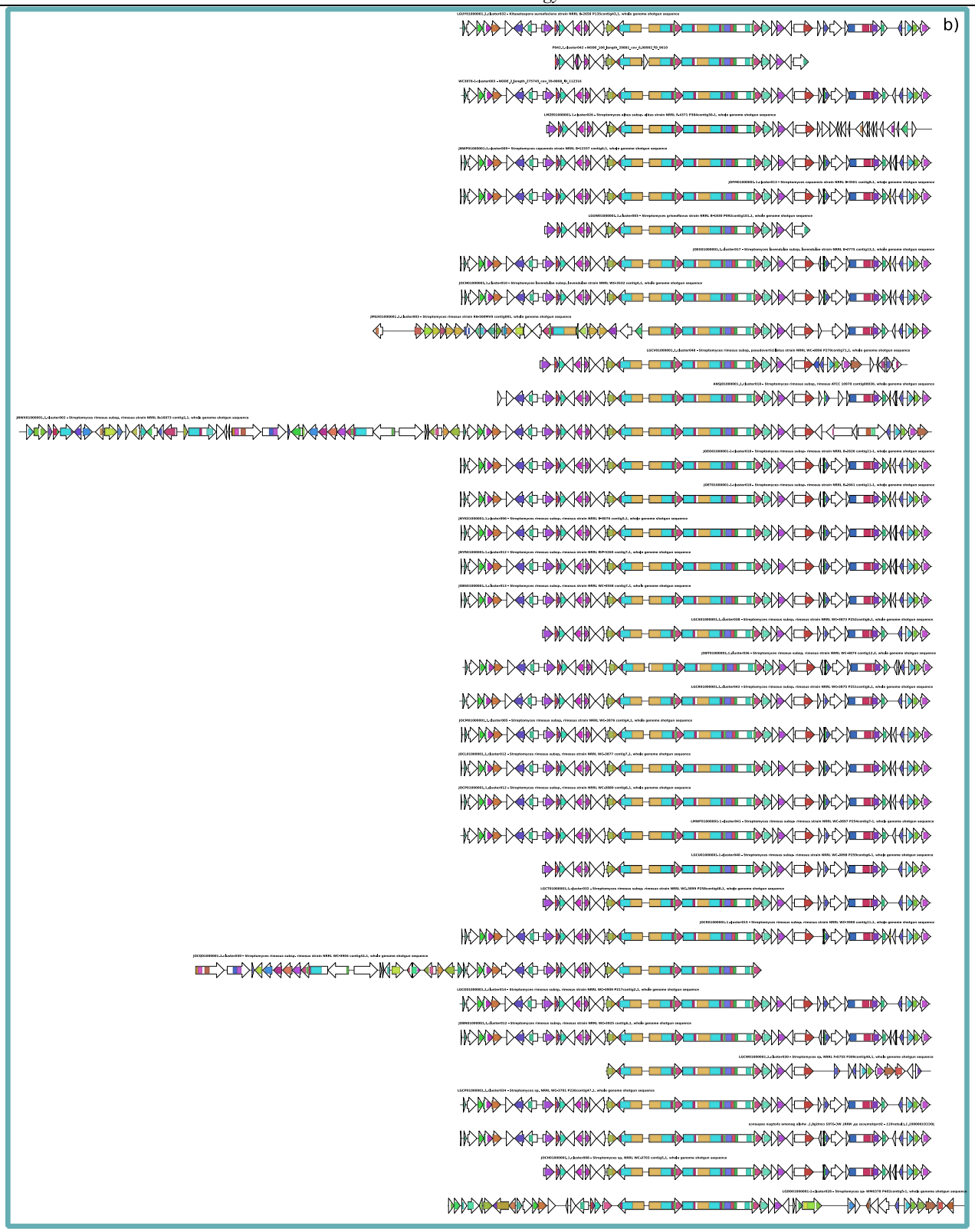
Supplementary Figure 9. CORASON phylogeny of detoxin/rimosamide-related BGCs.



Supplementary Figure 10. Rimosamide-related GCFs in Figure 3a.

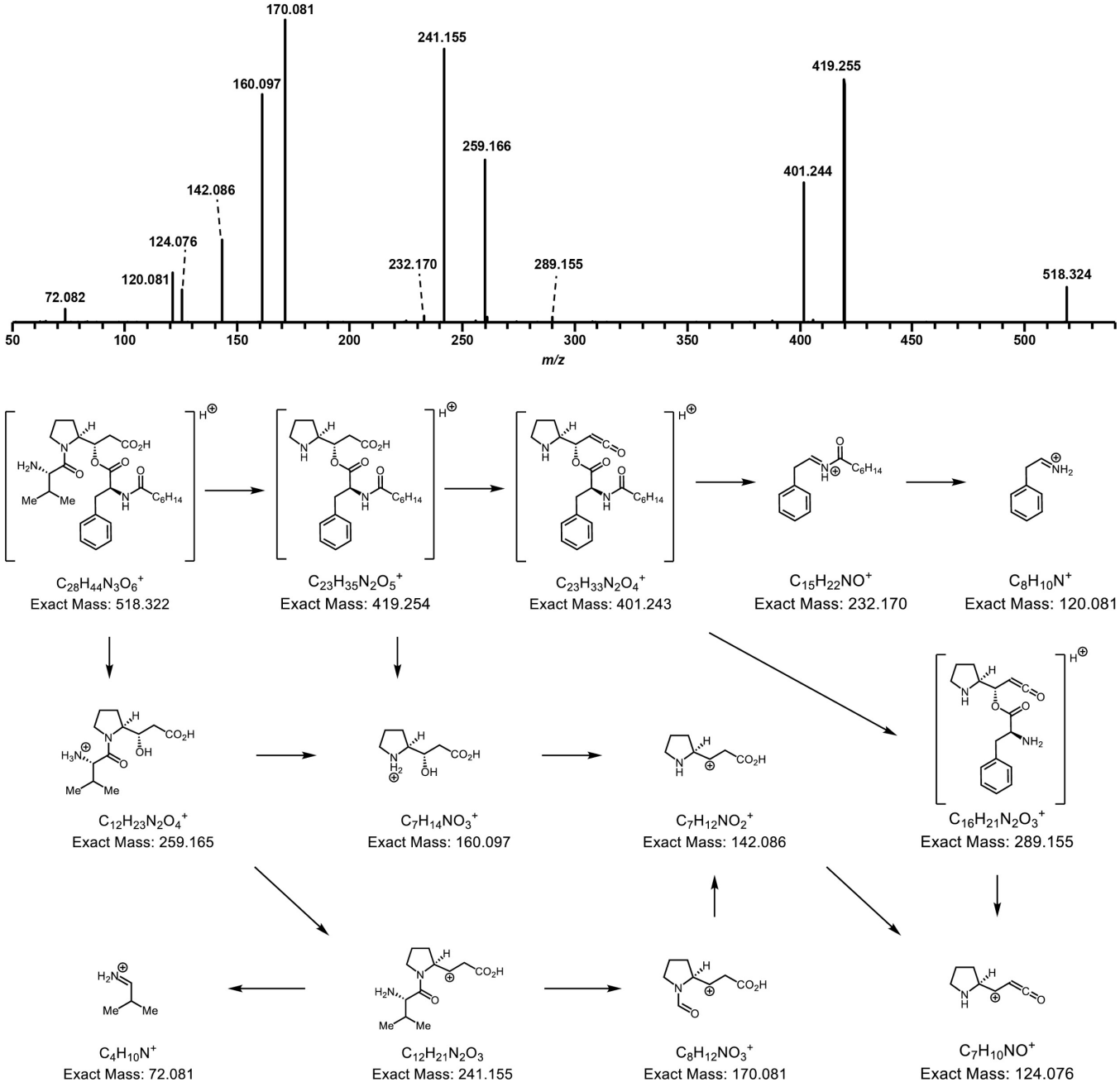


a. The light-turquoise GCF contains clusters not related to the target rimosamide BGCs due to them being clustered with true rimosamide BGCs through 'linking' regions (facilitated by the glocal mode). The large clusters arise when antiSMASH predicts a very large 'merged' cluster in cases when more Core Biosynthetic Genes are found during the stage of border expansion (in this case, more NRPS genes; light blue). Moreover, even though this network was filtered to only include BGCs that contained the TauD domain, some non-rimosamide BGCs also contain a second copy of this domain and therefore were added to the network.



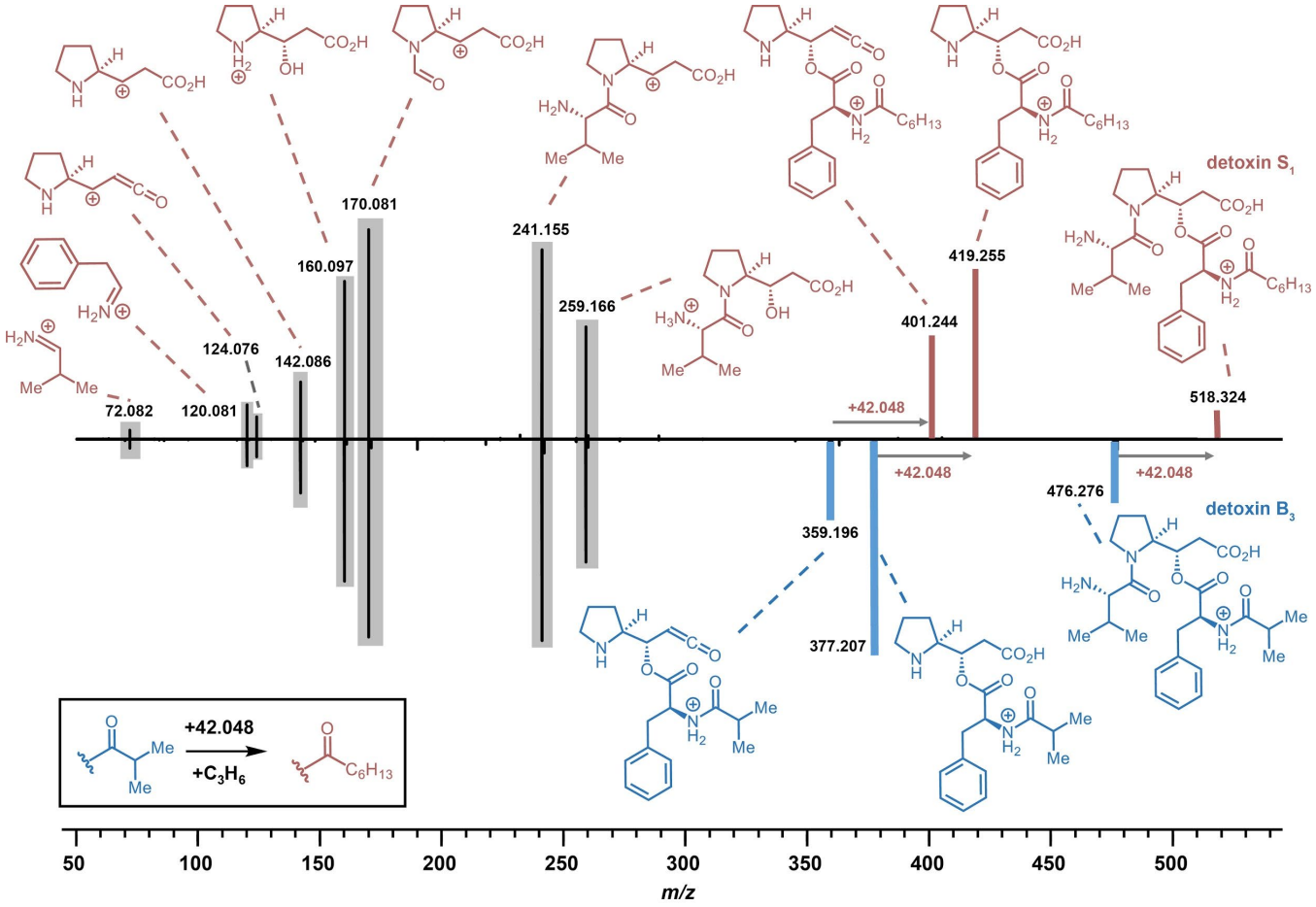
b. BGCs in the dark-turquoise GCF contain only clusters with the signature NRPS, PKS/NRPS Hybrid, *tauD* gene cluster architecture.

Supplementary Figure 11. Tandem MS spectrum of detoxin S_1 (**1**, *m/z* 518.324) from *Streptomyces* species NRRL S-325.



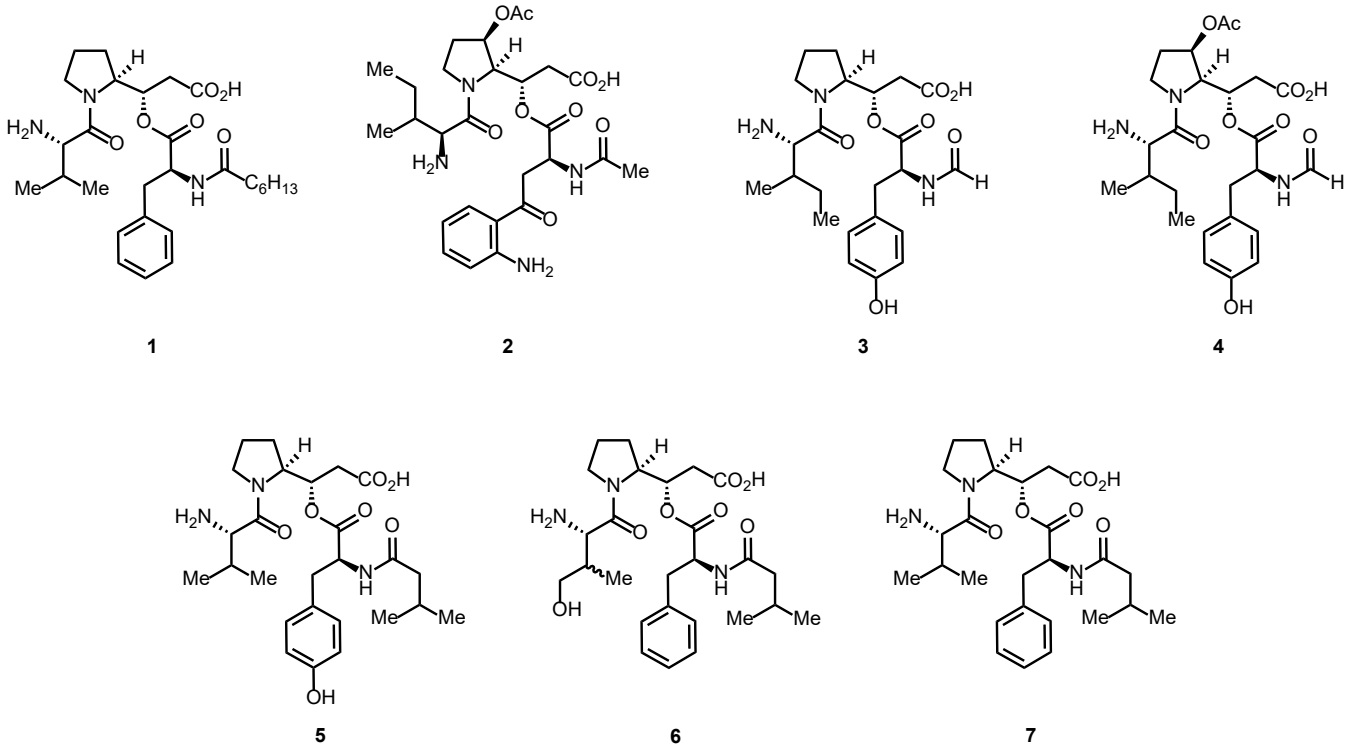
Tandem MS fragmentation supports the assignment of valine, phenylalanine, heptanoyl, and modified proline residues in the proposed structure of detoxin S_1.

Supplementary Figure 12. Comparison of the tandem MS spectrum of detoxin S_1 (**1**, *m/z* 518.324) from *Streptomyces* species NRRL S-325 with the tandem MS spectrum of known detoxin B_3[14].

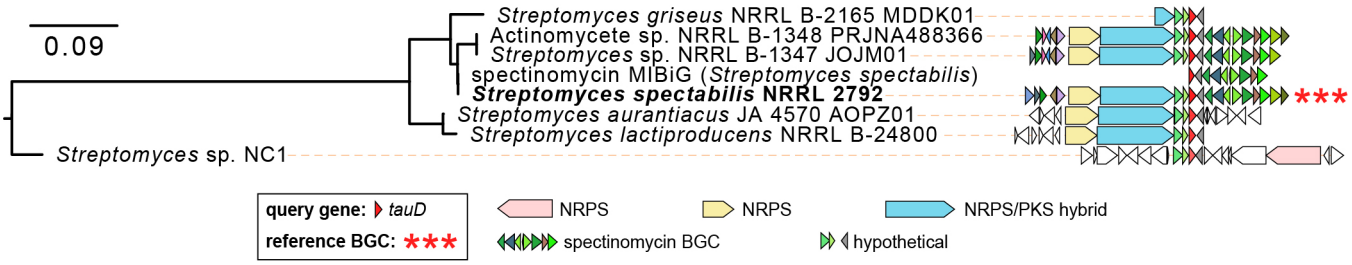


Matching fragments between detoxin S_1 and detoxin B_3 in the low *m/z* range support assignment of valine, phenylalanine, and the modified proline common among this class. An increase in *m/z* 42.048 from fragments of detoxin B_3 to those of detoxin S_1 in the higher *m/z* range support an increase in C_3H_6 for the detoxin S_1 side chain, suggesting a hydrocarbon chain.

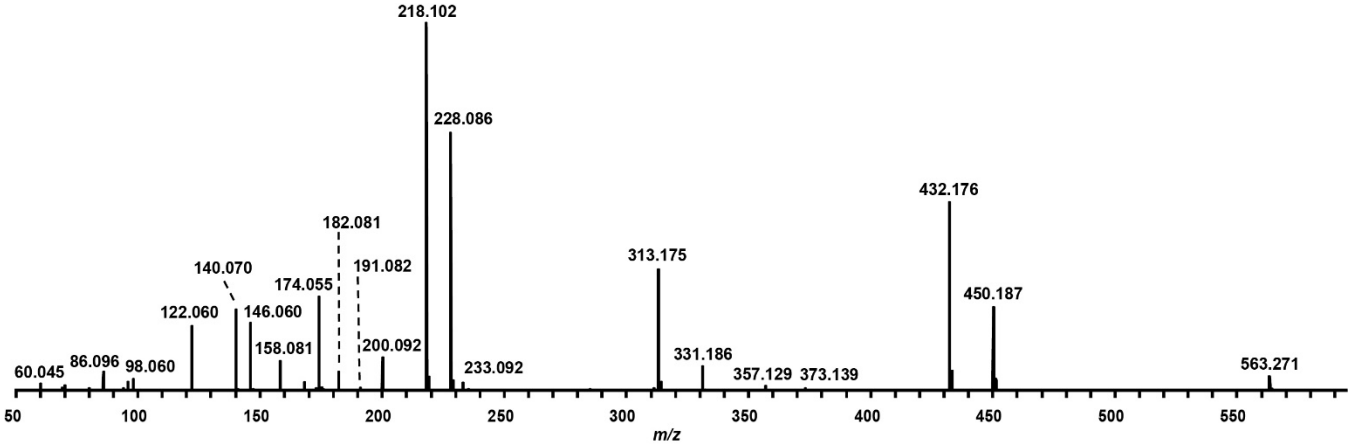
Supplementary Figure 13. Structures of detoxins S_1 (**1**), N_1 (**2**), N_2 (**3**), N_3 (**4**), P_1 (**5**), P_2 (**6**), and P_3 (**7**).

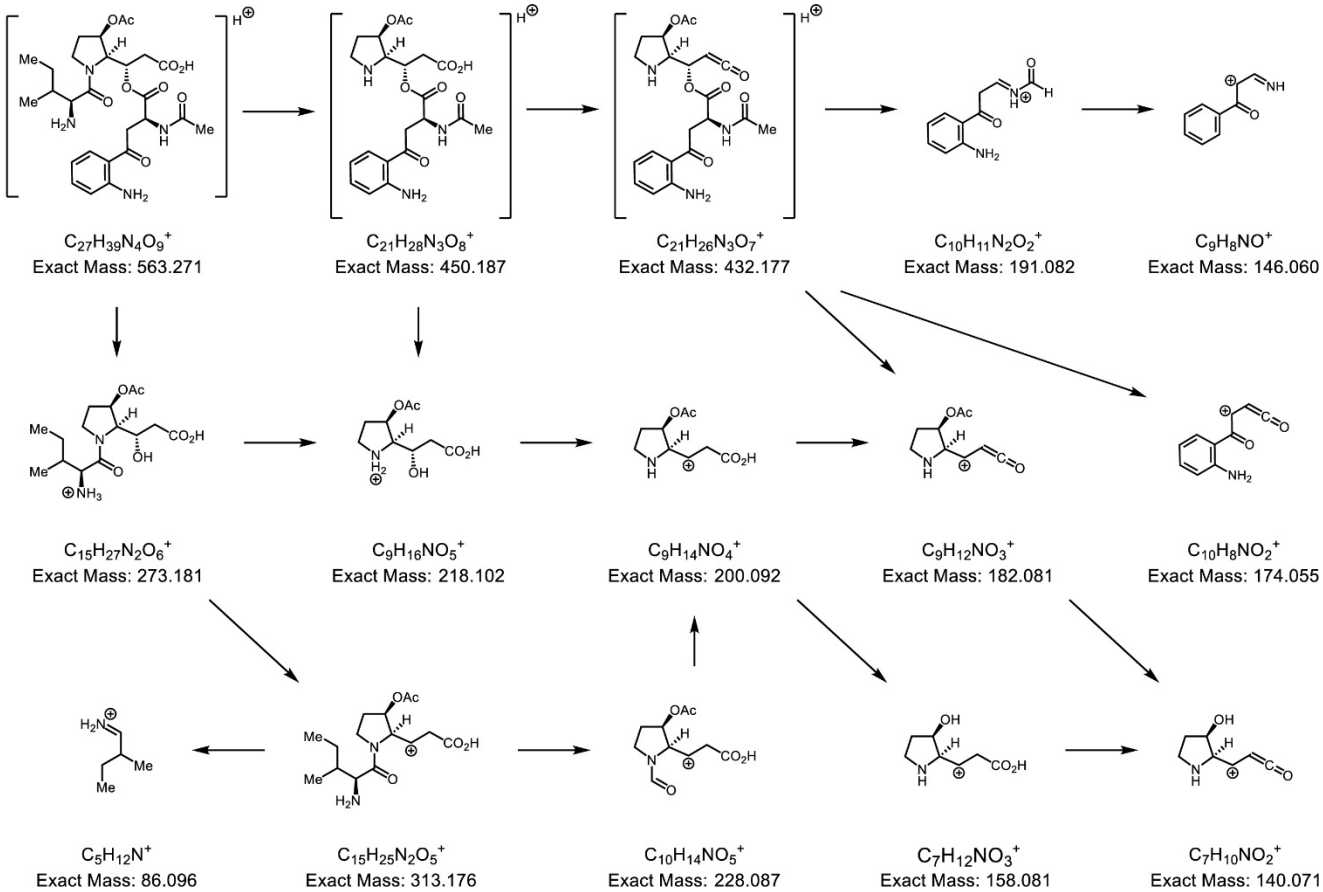


Supplementary Figure 14. Predicted detoxin cluster in the newly sequenced genome of *Streptomyces spectabilis* NRRL 2792.



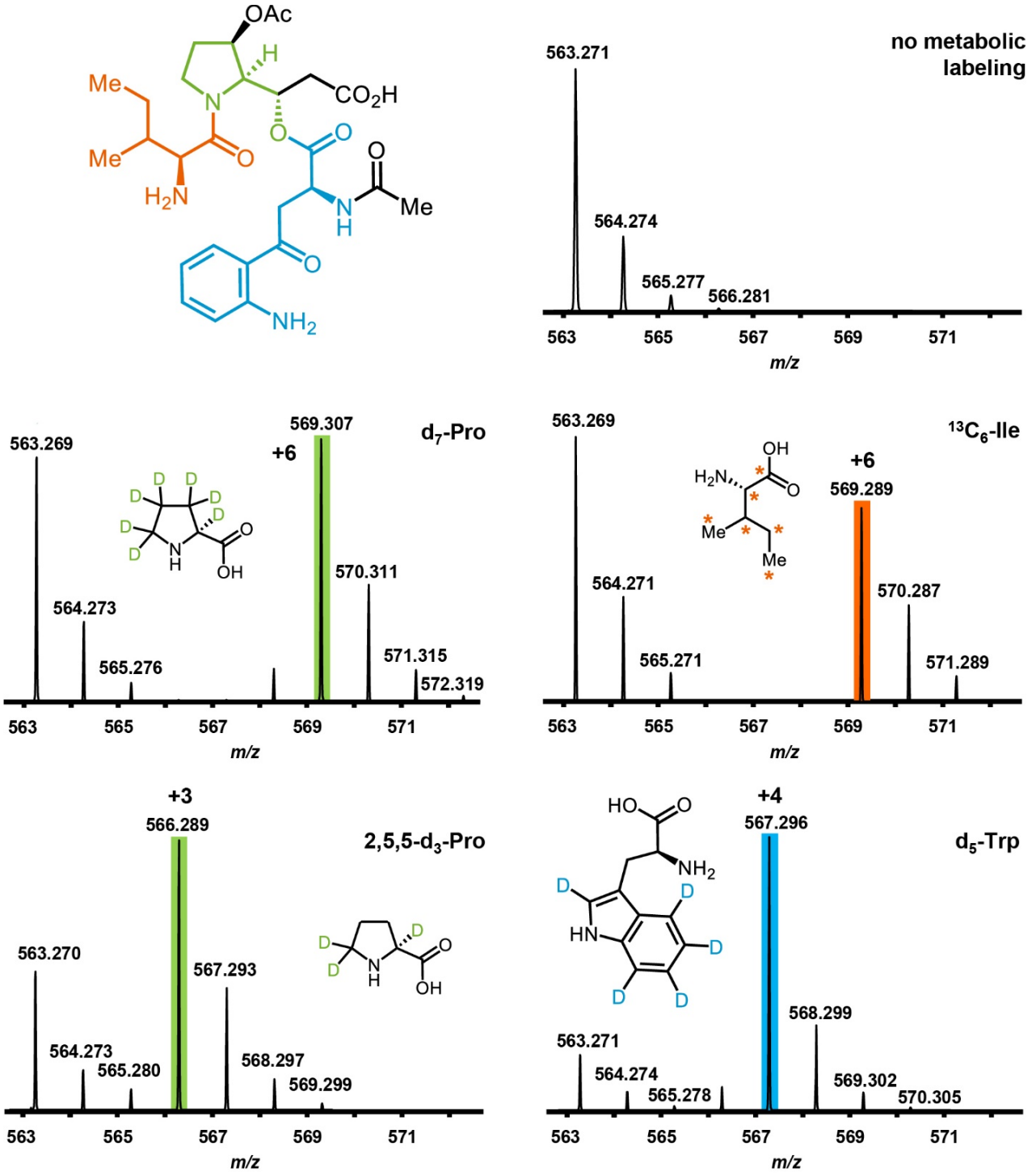
Supplementary Figure 15. Tandem MS spectrum of detoxin N_1 (**2**, *m/z* 563.271) from *Streptomyces spectabilis* NRRL 2792.





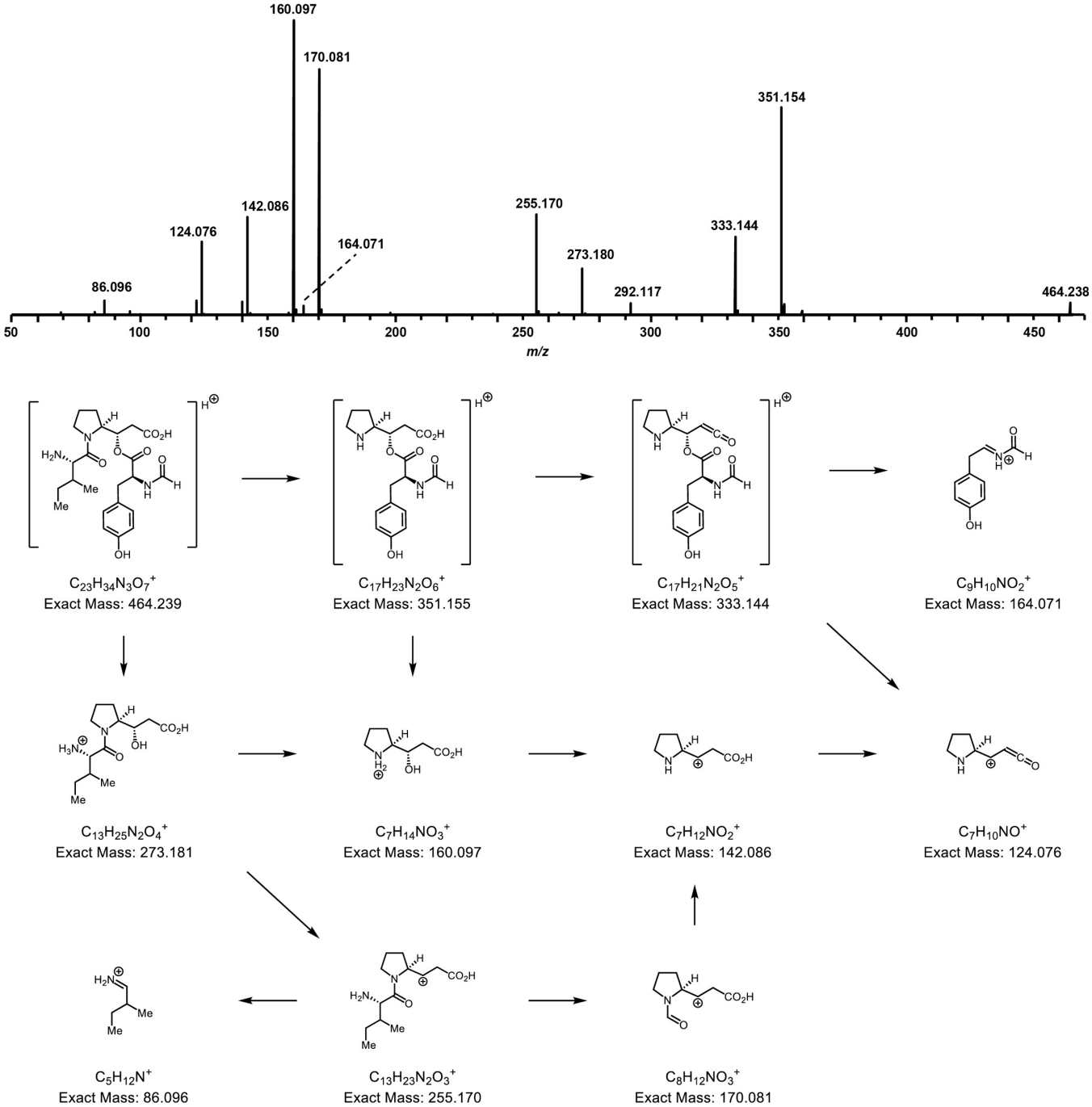
Tandem MS fragmentation supports the assignment of isoleucine, kynurenine, *N*-acetyl, and modified proline residues in the detoxin N_1 chemical structure.

Supplementary Figure 16. Stable isotope labeled-amino acid incorporation of d_7-proline, 2,5,5-d_3-proline. $^{13}C_6$-isoleucine, and indole-d_5-tryptophan in detoxin N_1 (**2**, *m/z* 563.271) from *Streptomyces spectabilis* NRRL 2792.



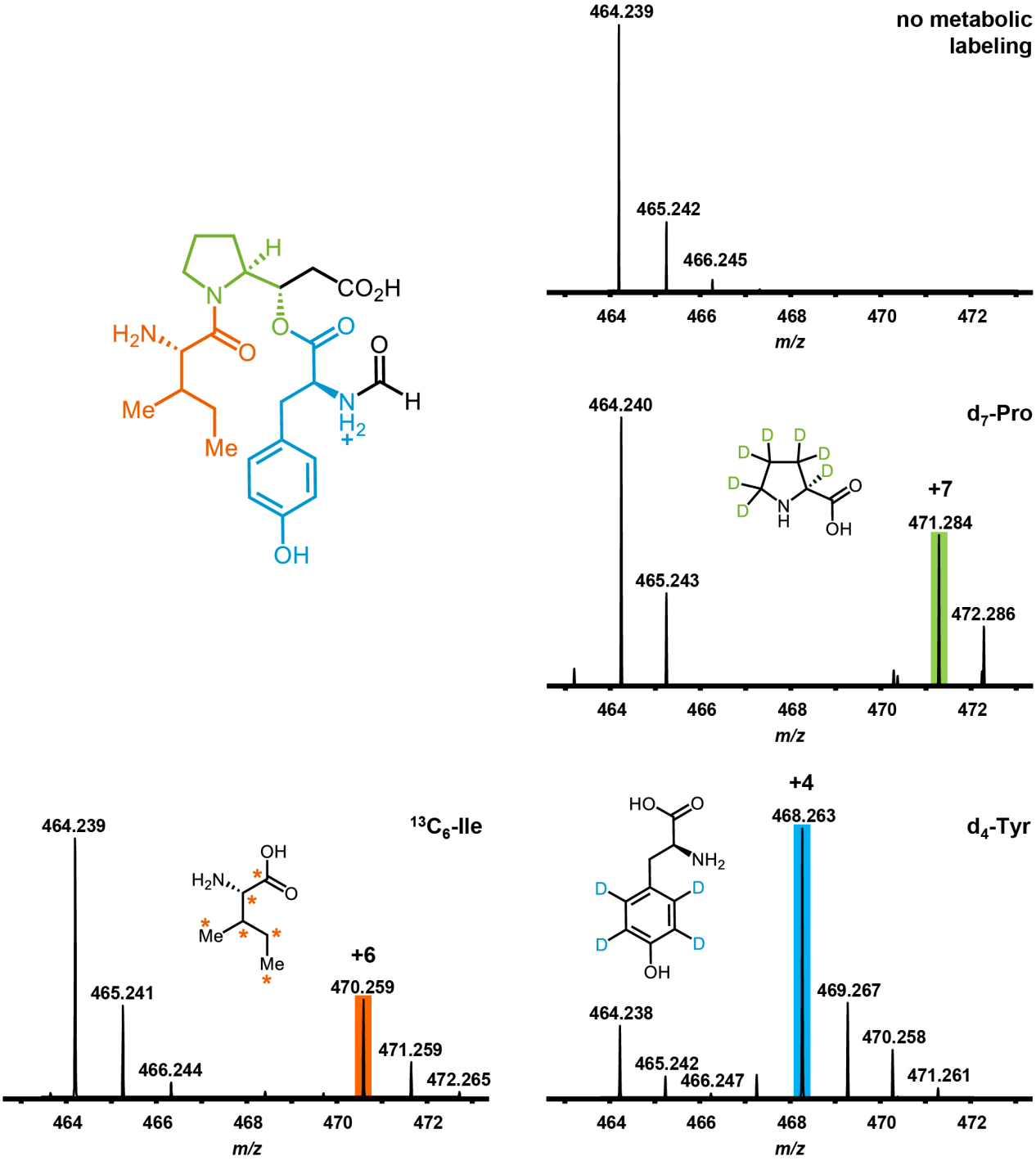
All ions indicative of labeled amino acid incorporation in MS^1 spectra were confirmed by analysis of fragmentation patterns in the corresponding tandem MS data. Replicates for labeling experiments were as follows: d_7-Pro incorporation was verified with 76 repeat MS^1 scan measurements and 7 repeat MS^2 scan measurements in one experiment; d_3-Pro incorporation was verified with 207 repeat MS^1 scan measurements and 58 repeat MS^2 scan measurements over two replicates; $^{13}C_6$-Ile incorporation was verified with 567 repeat MS^1 scan measurements and 174 repeat MS^2 scan measurements over three replicates; and d_5-Trp incorporation was verified with 241 repeat MS^1 scan measurements and 45 repeat MS^2 scan measurements over one experiment.

Supplementary Figure 17. Tandem MS spectrum of detoxin N_2 (**3**, *m/z* 464.239) from *Streptomyces spectabilis* NRRL 2792.



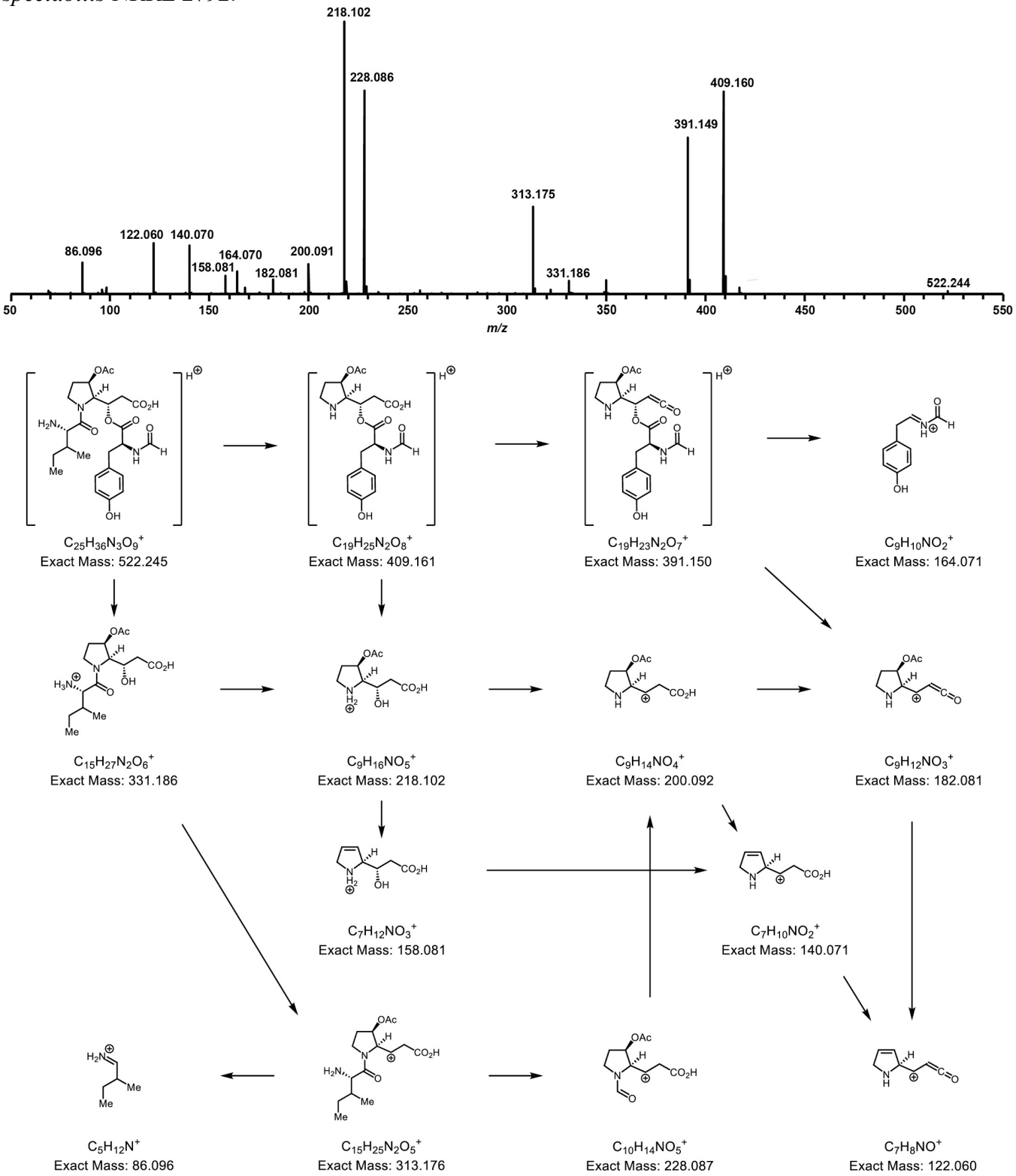
Tandem MS fragmentation supports the assignment of isoleucine, tyrosine, formyl, and modified proline residues in the detoxin N_1 chemical structure.

Supplementary Figure 18. Stable isotope labeled-amino acid incorporation of d_7-proline, phenyl-d_4-tyrosine, and $^{13}C_6$-isoleucine in detoxin N_2 (**3**, *m/z* 464.239) from *Streptomyces spectabilis* NRRL 2792.



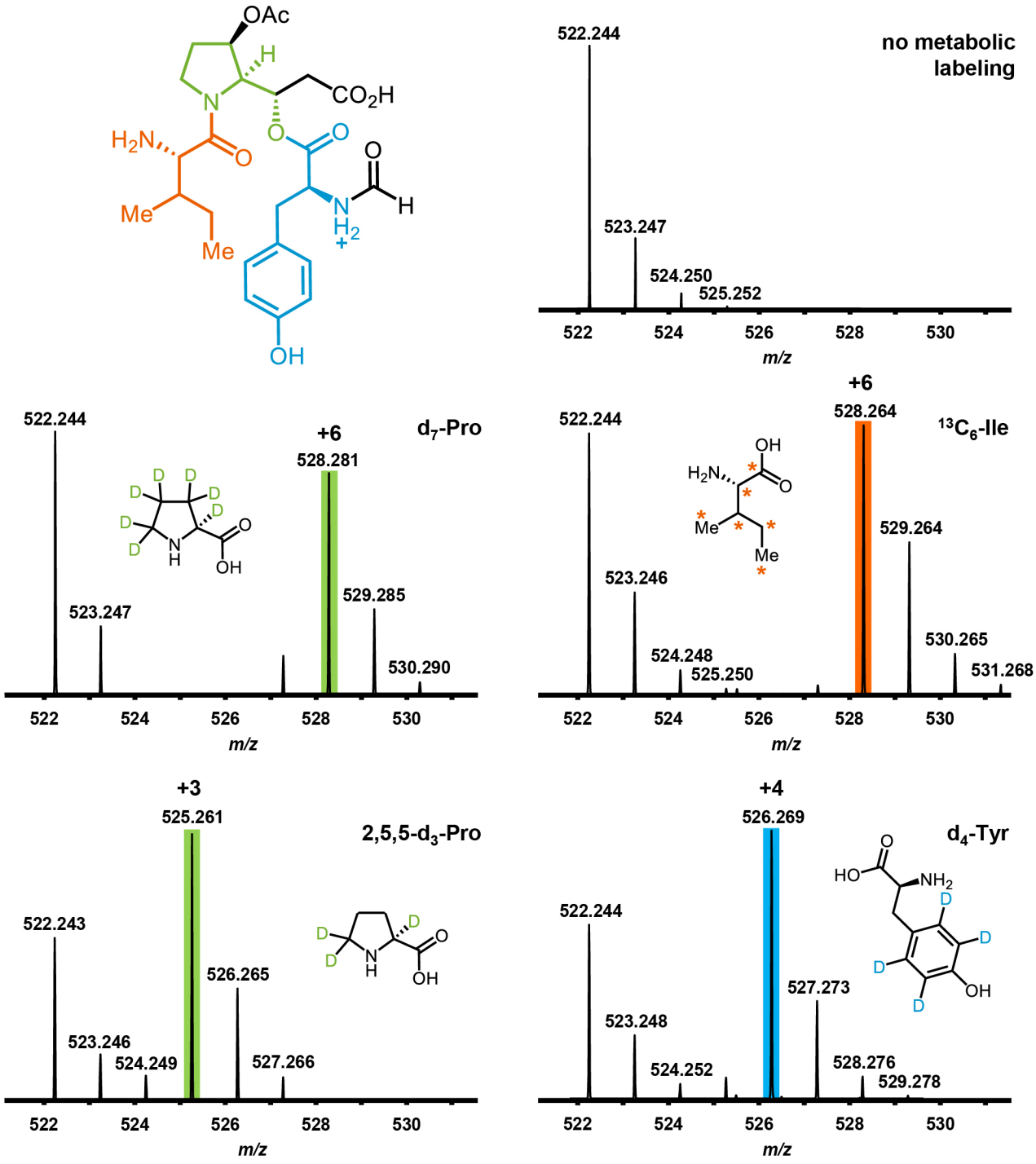
All ions indicative of labeled amino acid incorporation in MS^1 spectra were confirmed by analysis of fragmentation patterns in the corresponding tandem MS data. Replicates for labeling experiments were as follows: d_7-Pro incorporation was verified with 208 repeat MS^1 scan measurements and 12 repeat MS^2 scan measurements in one experiment; $^{13}C_6$-Ile incorporation was verified with 167 repeat MS^1 scan measurements and 35 repeat MS^2 scan measurements over three replicates; and d_4-Tyr incorporation was verified with 1143 repeat MS^1 scan measurements and 39 repeat MS^2 scan measurements over three replicates.

Supplementary Figure 19. Tandem MS spectrum of detoxin N_3 (**4**, *m/z* 522.244) from *Streptomyces spectabilis* NRRL 2792.



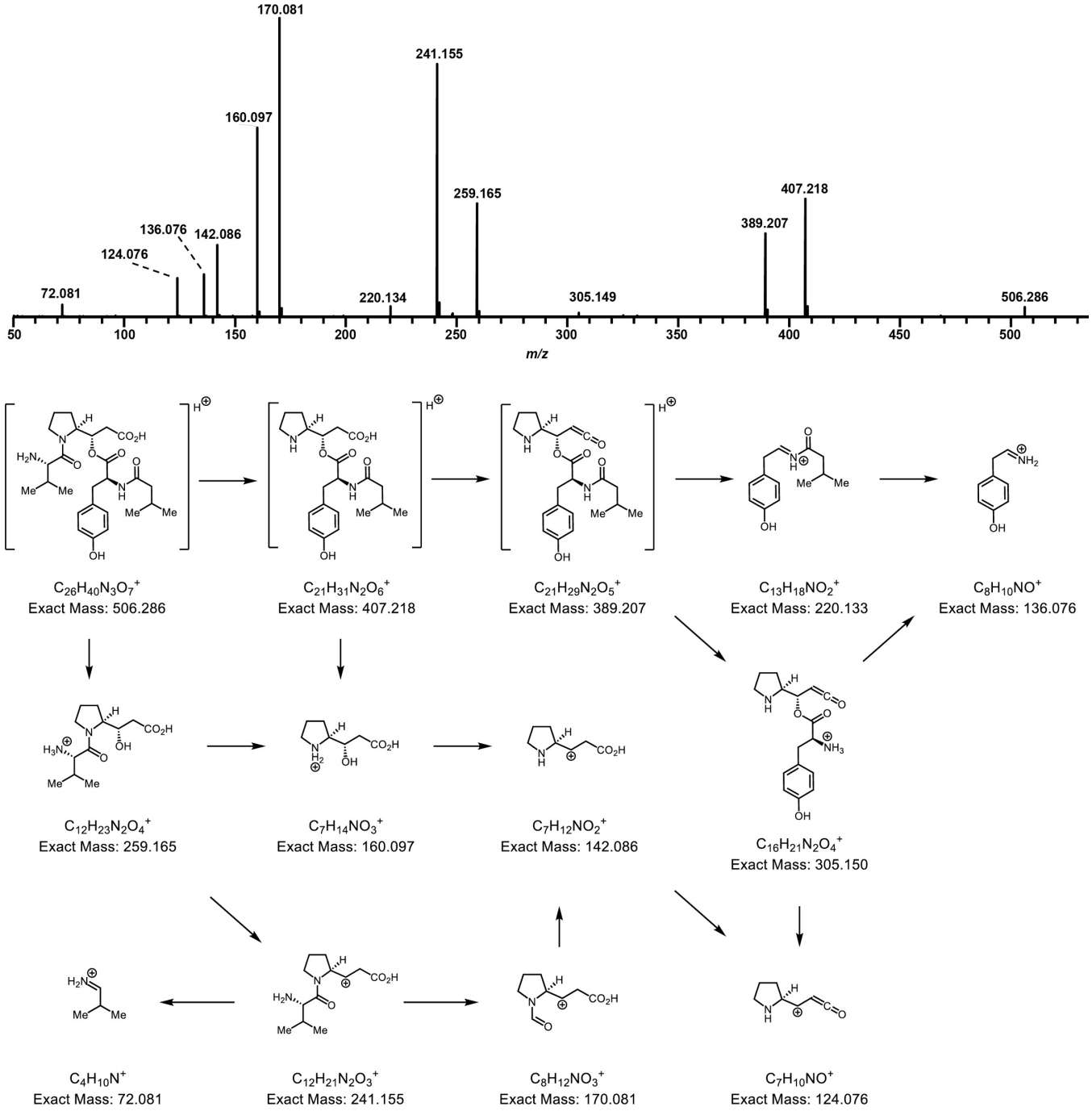
Tandem MS fragmentation supports the assignment of isoleucine, tyrosine, formyl, and acetoxy-modified proline residues in the detoxin N_2 chemical structure.

Supplementary Figure 20. Stable isotope labeled-amino acid incorporation of d_7-proline, phenyl-d_4-tyrosine, and $^{13}C_6$-isoleucine in detoxin N_3 (**4**, *m/z* 522.244) from *Streptomyces spectabilis* NRRL 2792.



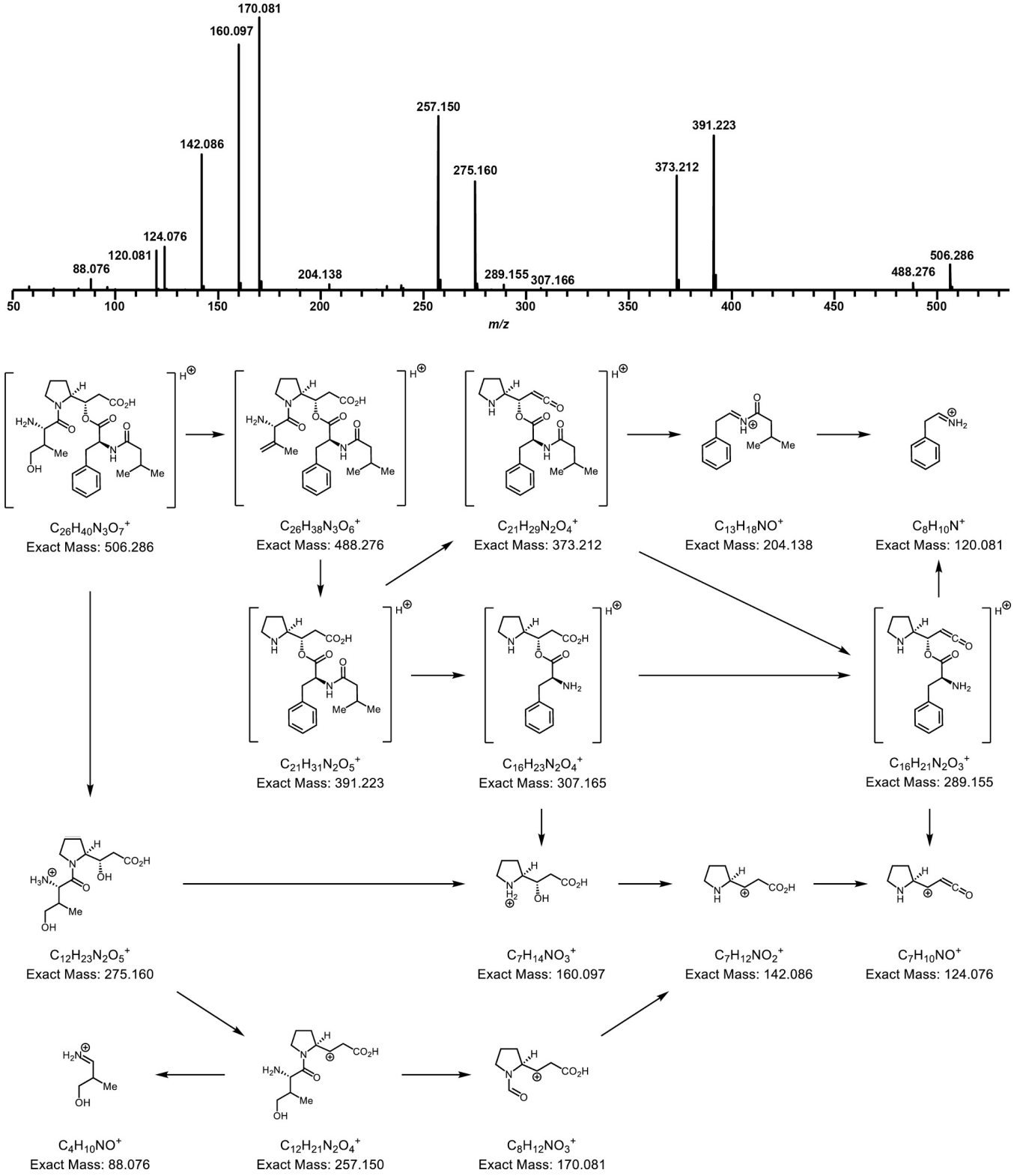
Combined with molecular formula analysis and comparison to known detoxins and rimosamides, the loss of one deuteron from d_7-proline, retention of all deuterons in 2,5,5-d_3-proline in metabolic labeling experiments, and comparison to known analogs supports assignment of the acetoxy group at the 3 position on proline. All ions indicative of labeled amino acid incorporation in MS^1 spectra were confirmed by analysis of fragmentation patterns in the corresponding tandem MS data. Replicates for labeling experiments were as follows: d_7-Pro incorporation was verified with 186 repeat MS^1 scan measurements and 5 repeat MS^2 scan measurements in one experiment; d_3-Pro incorporation was verified with 201 repeat MS^1 scan measurements and 3 repeat MS^2 scan measurements over two replicates; $^{13}C_6$-Ile incorporation was verified with 132 repeat MS^1 scan measurements over three replicates; and d_4-Tyr incorporation was verified with 298 repeat MS1 scan measurements and 2 repeat MS^2 scan measurements over three replicates.

Supplementary Figure 21. Tandem MS spectrum of detoxin P_1 (**5**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



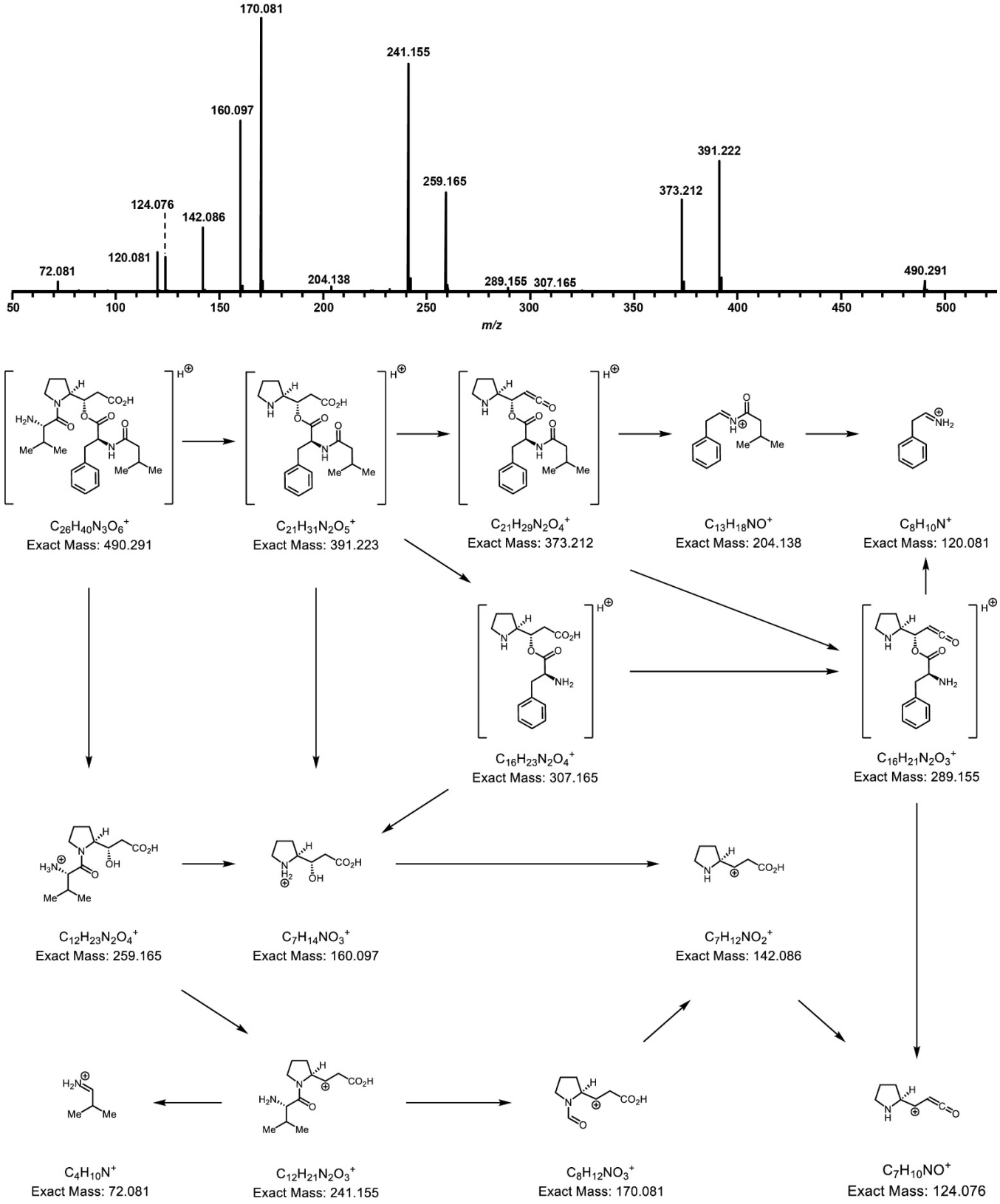
Tandem MS fragmentation supports the assignment of valine, tyrosine, isovaleryl, and modified proline residues in the detoxin P_1 chemical structure. The *m/z* 136.076 and *m/z* 305.150 ions are key in assignment of hydroxylation on the aryl ring and not on the isovaleryl residue.

Supplementary Figure 22. Tandem MS spectrum of detoxin P_2 (**6**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



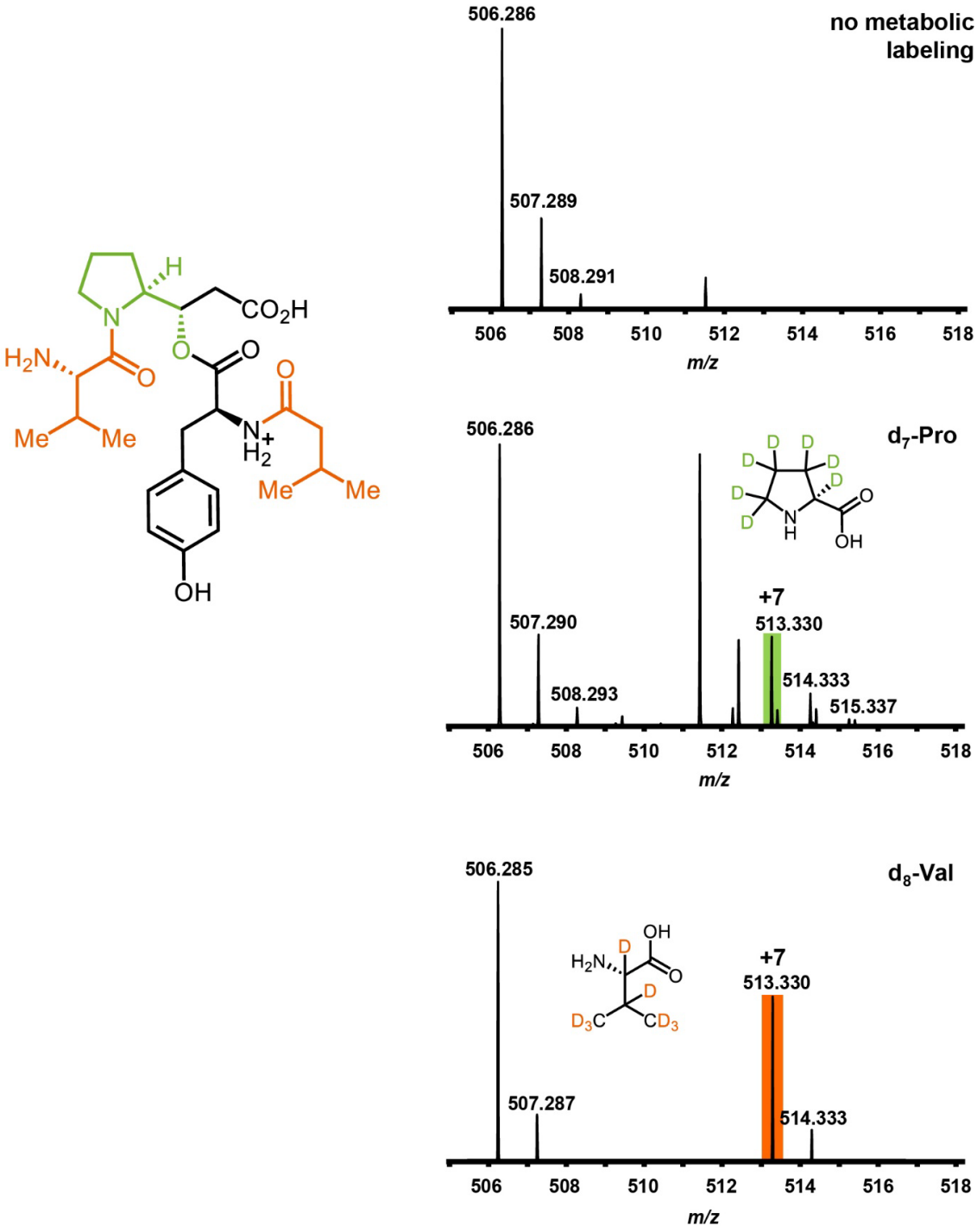
Tandem MS fragmentation supports the assignment of hydroxyvaline, phenylalanine, isovaleryl, and modified proline residues in the detoxin P_2 chemical structure.

Supplementary Figure 23. Tandem MS spectrum of detoxin P_3 (**7**, *m/z* 490.291) from *Amycolatopsis jejuensis* NRRL B-24427.



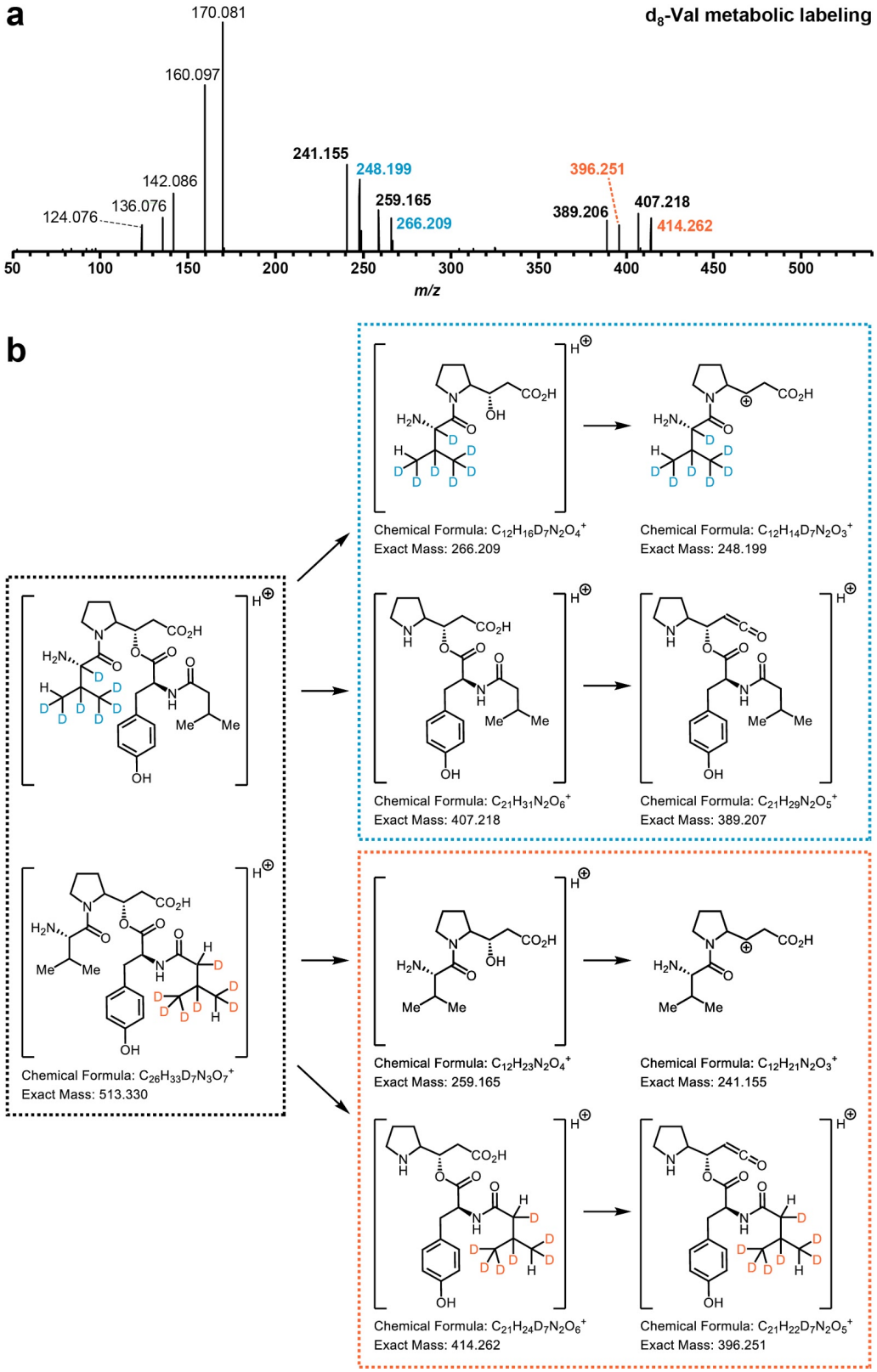
Tandem MS fragmentation supports the assignment of valine, phenylalanine, isovaleryl, and modified proline residues in the detoxin P_3 chemical structure.

Supplementary Figure 24. Stable isotope labeled-amino acid incorporation of d_7-proline and d_8-valine with loss of one deuteron in detoxin P_1 (**5**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



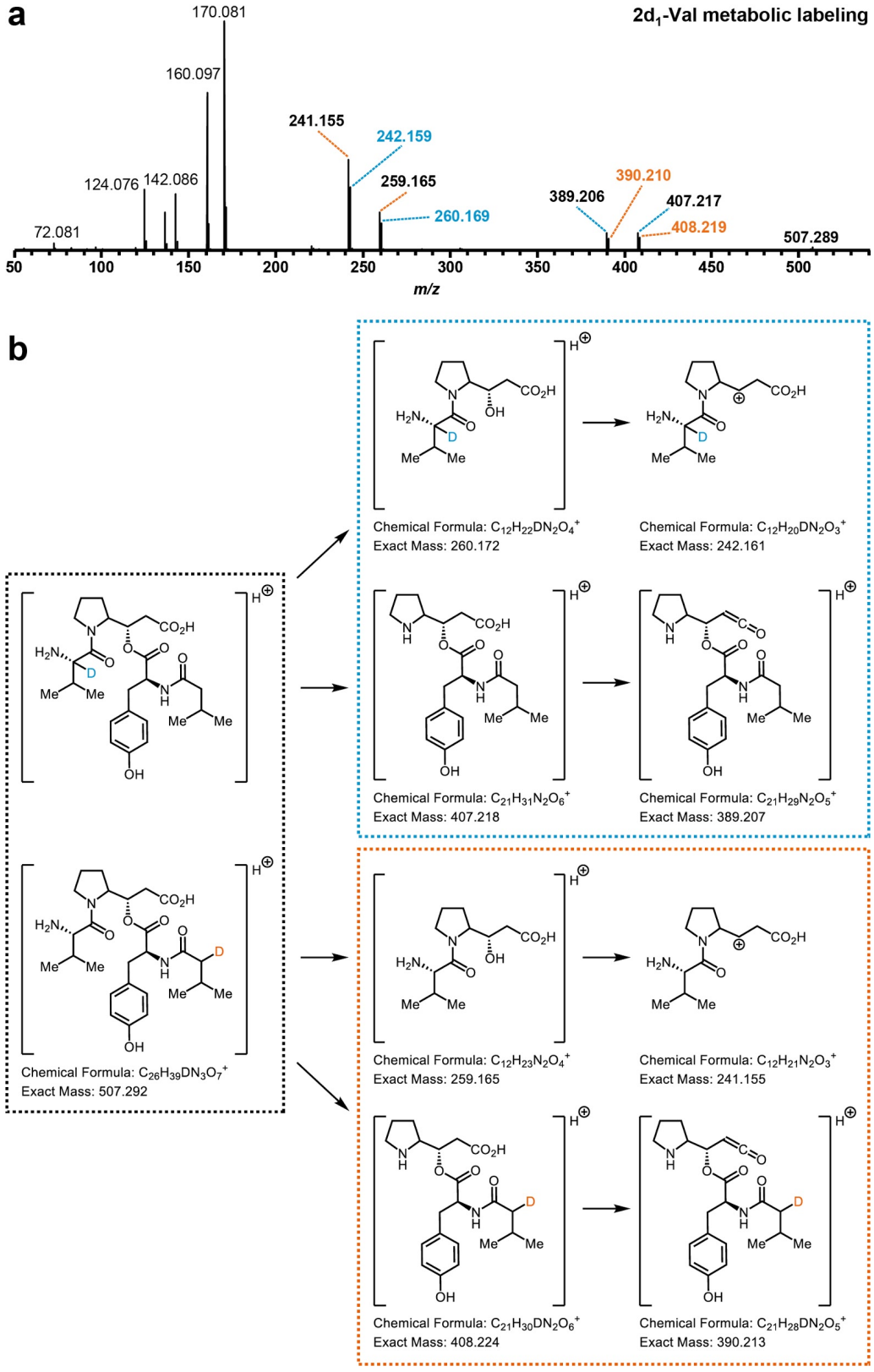
All ions indicative of labeled amino acid incorporation in MS^1 spectra were confirmed by analysis of fragmentation patterns in the corresponding tandem MS data. Though phenyl-d_4-tyrosine incorporation was not observed, incorporation of tyrosine is proposed based on molecular formula and tandem MS analysis as well as comparison to known detoxins and rimosamides. The loss of one deuteron from d_8-valine in the metabolic labeling experiment is likely due to oxidation and reduction reactions carried out by a P450 enzyme in the *Amycolatopsis jejuensis* NRRL B-24427 detoxin BGC. The oxidation is retained in the valine of detoxin P_2, but these are fully reduced in detoxins P_1 and P_3. Replicates for labeling experiments were as follows: d_7-Pro incorporation was verified with 49 repeat MS^1 scan measurements and 5 repeat MS^2 scan measurements in one experiment; and d_8-Val incorporation was verified with 207 repeat MS^1 scan measurements and 4 repeat MS^2 scan measurements in one experiment.

Supplementary Figure 25. Tandem MS fragmentation data for stable isotope labeled-amino acid incorporation of d_8-valine in detoxin P_1 (**5**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



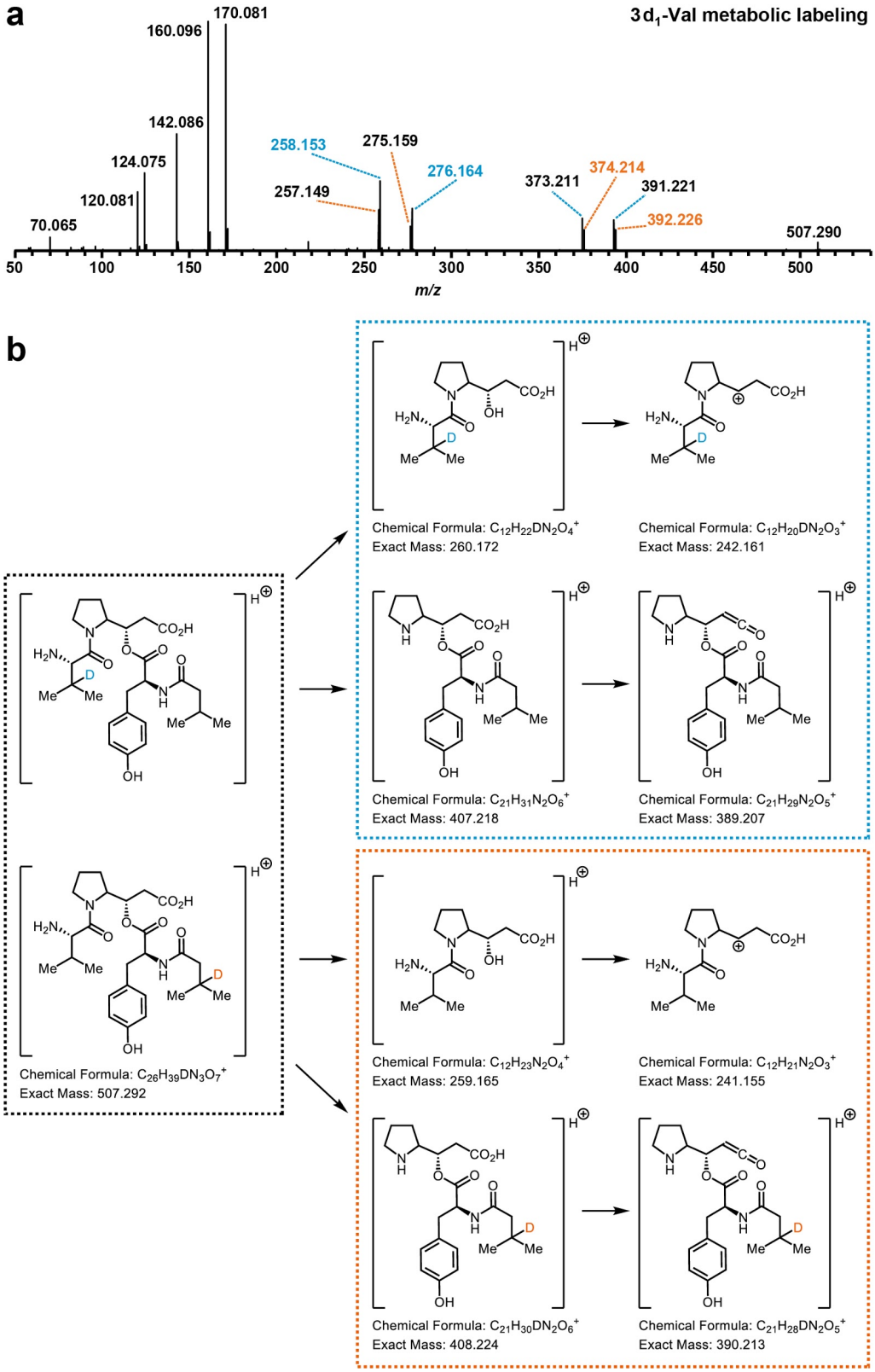
a, MS/MS spectrum of d_8-valine-labeled detoxin P_1. **b**, Predicted fragmentation indicates labeled and unlabeled valine and isovaleryl residue incorporation with loss of one deuteron likely due to oxidation and reduction by a P450 enzyme.

Supplementary Figure 26. Tandem MS fragmentation data for stable isotope labeled-amino acid incorporation of $2d_1$-valine in detoxin P_1 (**5**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



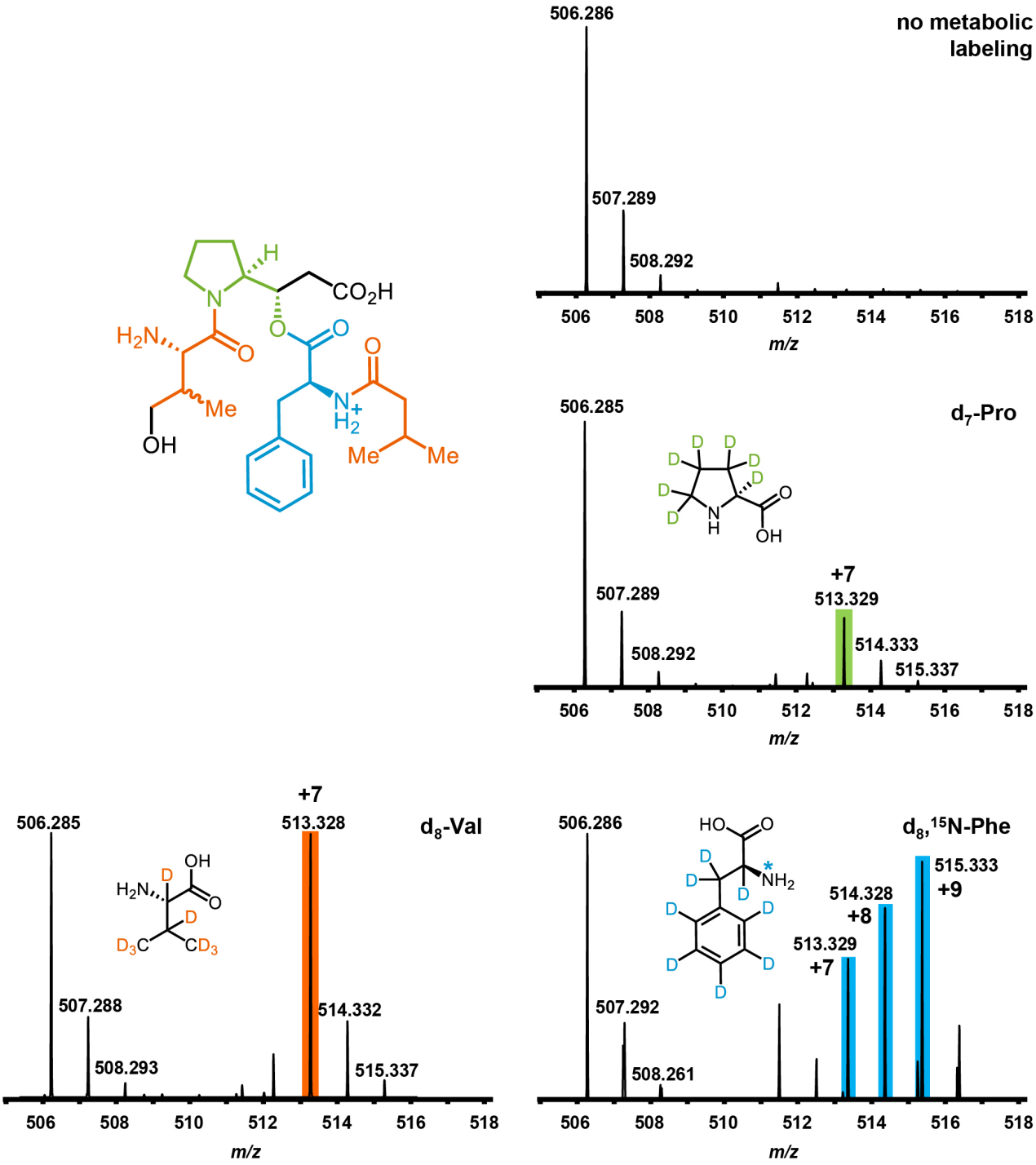
a, MS/MS spectrum of $2d_1$-valine-labeled detoxin P_1. **b**, Fragmentation with retention of deuterium at the 2 position of valine support direct incorporated of valine and valine incorporated as an isovaleryl residue, both without oxidation at position 2.

Supplementary Figure 27. Tandem MS fragmentation data for stable isotope labeled-amino acid incorporation of $3d_1$-valine in detoxin P_1 (**5**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



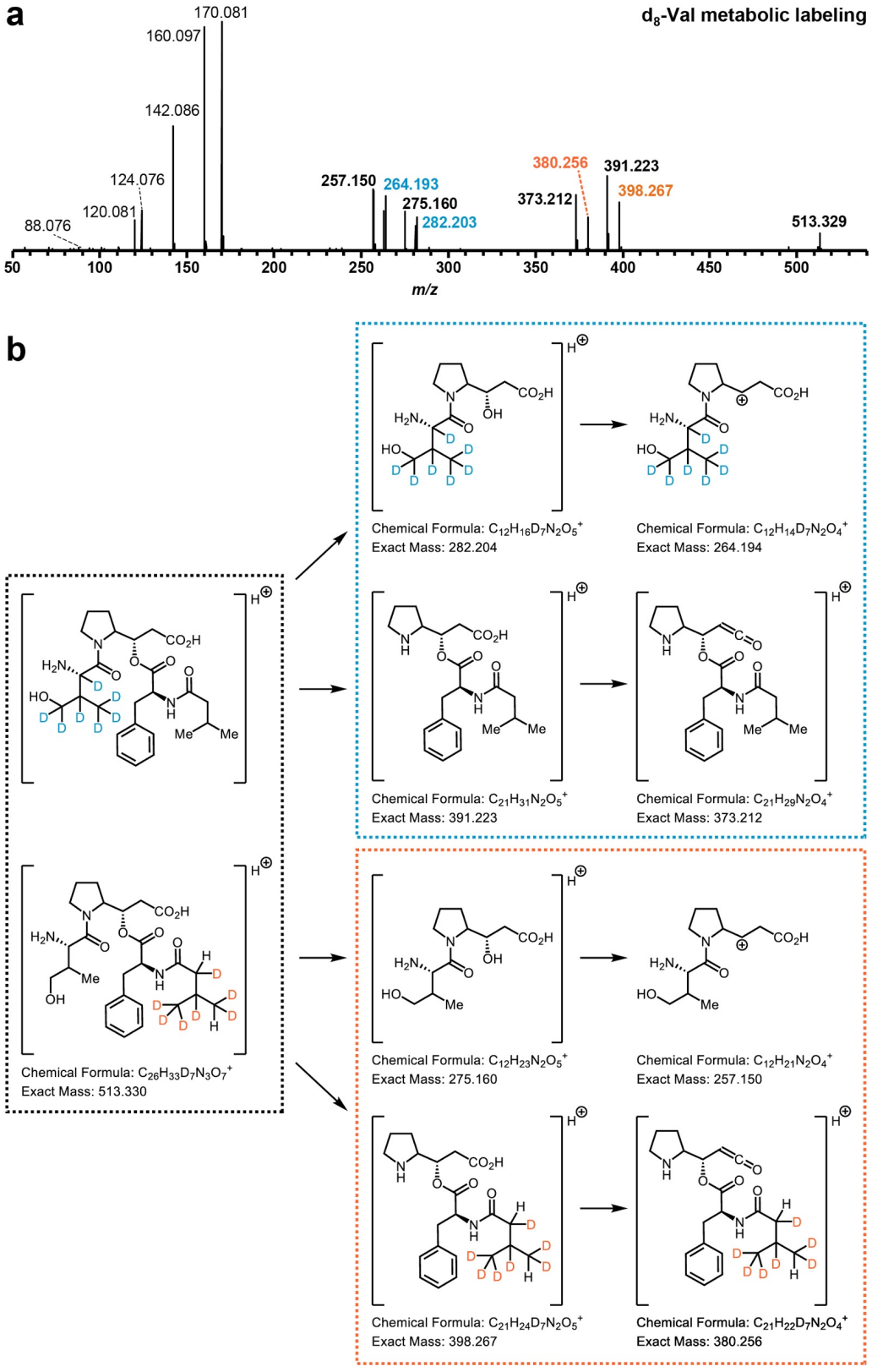
a, MS/MS spectrum of $3d_1$-valine-labeled detoxin P_1. **b**, Fragmentation with retention of deuterium at the 3 position of valine support direct incorporated of valine and valine incorporated as an isovaleryl residue, both without oxidation at position 3.

Supplementary Figure 28. Stable isotope labeled-amino acid incorporation of d_7-proline, d_8,^{15}N-phenylalanine, and d_8-valine with loss of one deuteron in detoxin P_2 (**6**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



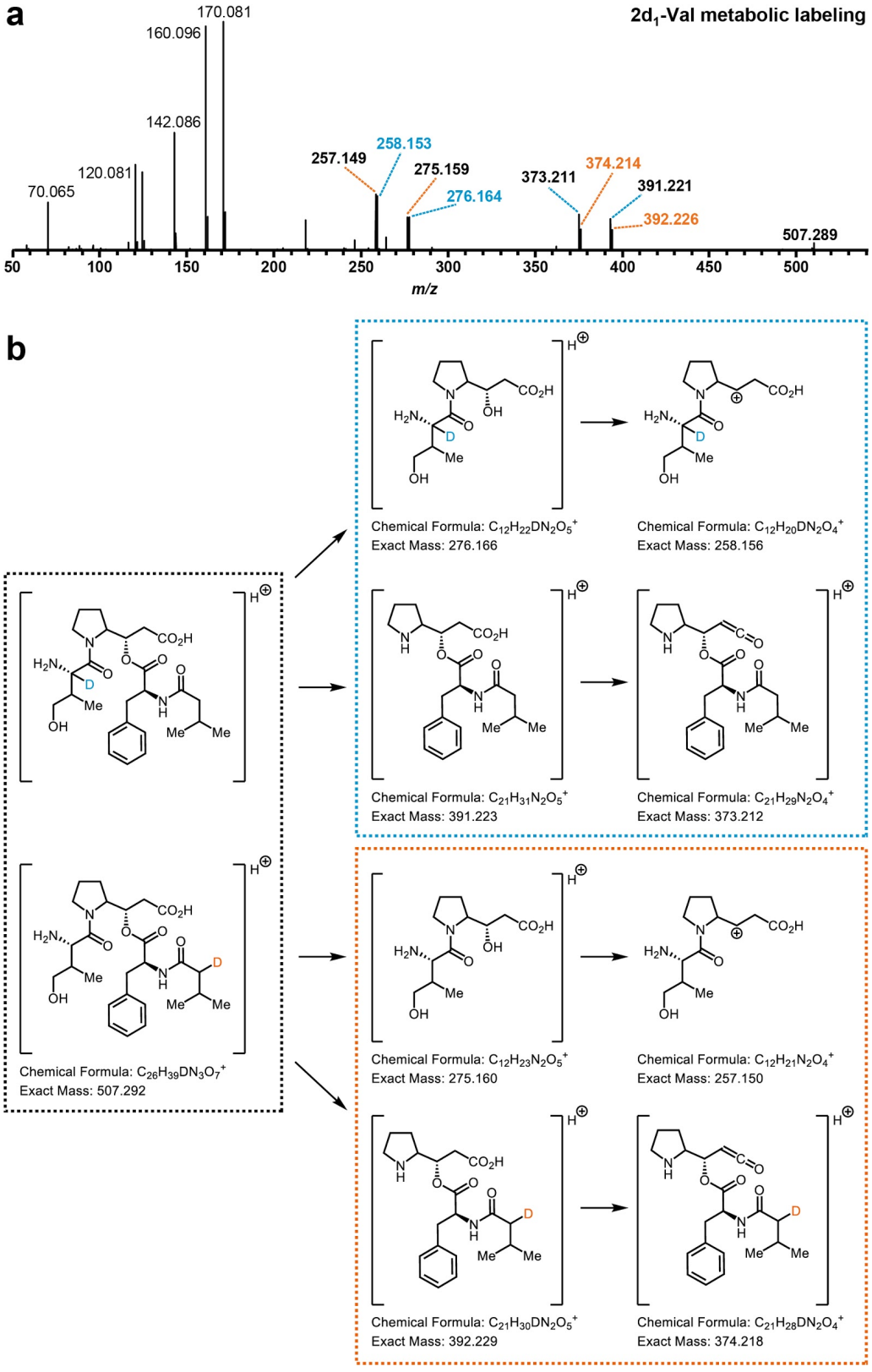
All ions indicative of labeled amino acid incorporation in MS^1 spectra were confirmed by analysis of fragmentation patterns in the corresponding tandem MS data. The loss of one deuteron from d_8-valine in the metabolic labeling experiment is due to hydroxylation of the valine methyl group and a putative oxidation and subsequent reduction of the methyl group in the isovaleryl residue. These transformations are further elucidated through other metabolic labeling experiments presented here. Replicates for labeling experiments were as follows: d_7-Pro incorporation was verified with 34 repeat MS^1 scan measurements and 4 repeat MS^2 scan measurements in one experiment; d_8-Val incorporation was verified with 118 repeat MS^1 scan measurements and 2 repeat MS^2 scan measurements in one experiment; and d_8,^{15}N-Phe incorporation was verified with 26 repeat MS^1 scan measurements and 3 repeat MS^2 scan measurements in one experiment.

Supplementary Figure 29. Tandem MS fragmentation data for stable isotope labeled-amino acid incorporation of d_8-valine in detoxin P_2 (**6**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



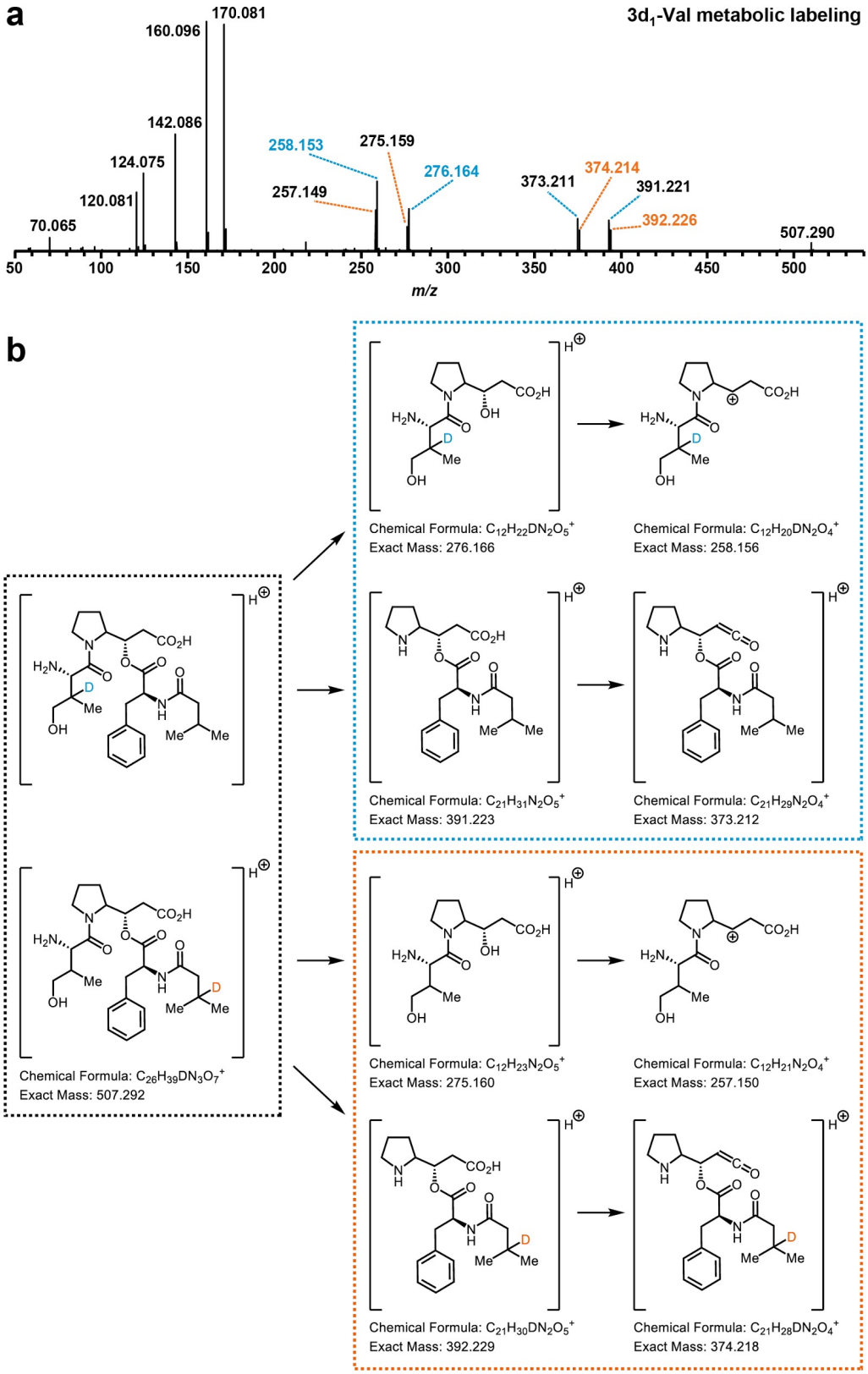
a, MS/MS spectrum of d_8-valine-labeled detoxin P_2. **b**, Predicted fragmentation indicates labeled and unlabeled hydroxyvaline and isovaleryl residue incorporation with loss of one deuteron likely due to oxidation and reduction by a P450 enzyme.

Supplementary Figure 30. Tandem MS fragmentation data for stable isotope labeled-amino acid incorporation of $2d_1$-valine in detoxin P_2 (**6**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



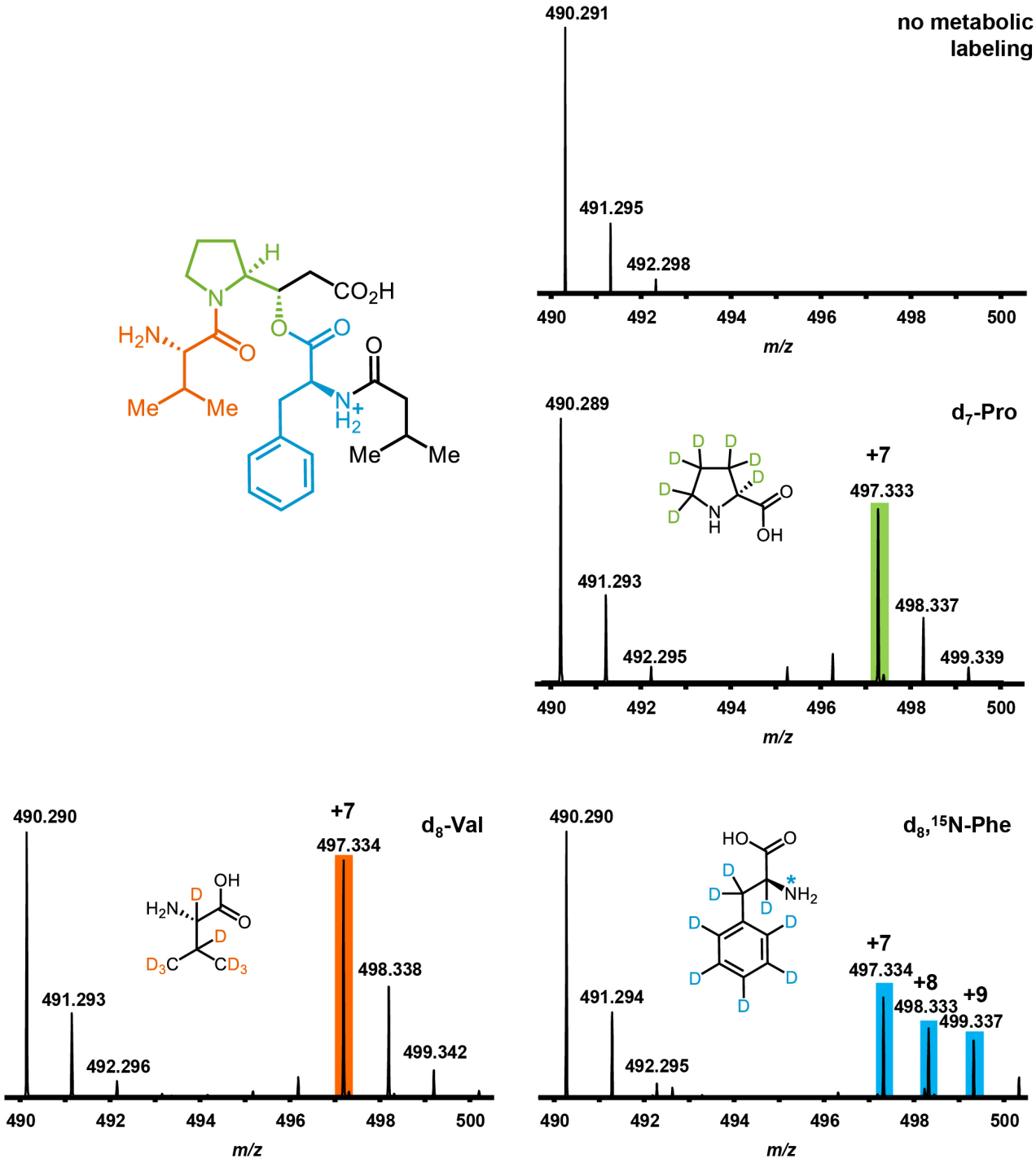
a, MS/MS spectrum of $2d_1$-valine-labeled detoxin P_2. **b**, Fragmentation with retention of deuterium at the 2 position of valine support incorporation of valine as 4-hydroxyvaline and isovaleryl residues, both without oxidation at position 2.

Supplementary Figure 31. Tandem MS fragmentation data for stable isotope labeled-amino acid incorporation of $3d_1$-valine in detoxin P_2 (**6**, *m/z* 506.286) from *Amycolatopsis jejuensis* NRRL B-24427.



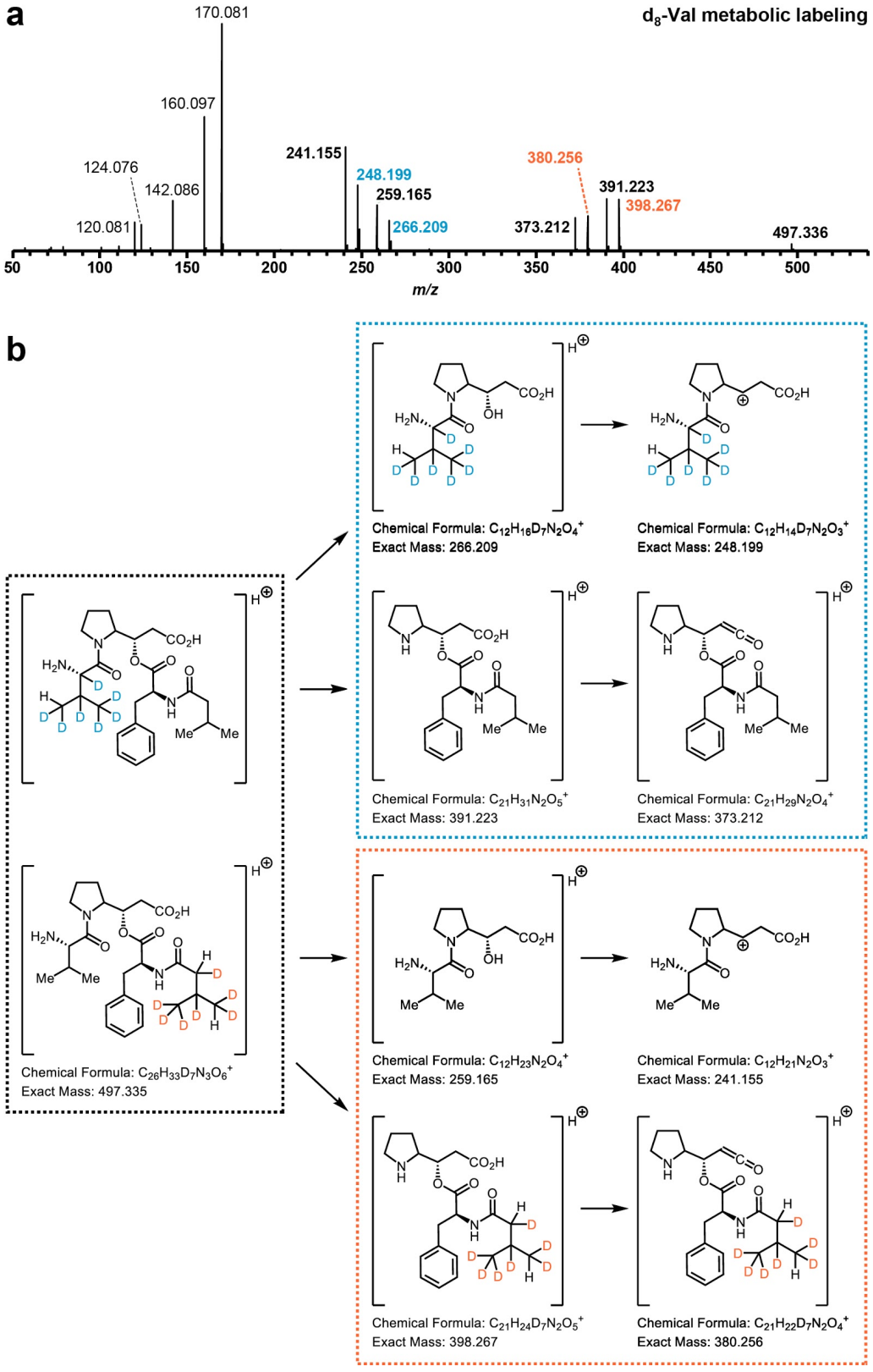
a, MS/MS spectrum of $3d_1$-valine-labeled detoxin P_2. **b**, Fragmentation with retention of deuterium at the 3 position of valine support incorporation of valine as 4-hydroxyvaline and isovaleryl residues, both without oxidation at position 3.

Supplementary Figure 32. Stable isotope labeled-amino acid incorporation of d_7-proline, d_8-phenylalanine, and d_8-valine with loss of one deuteron in detoxin P_3 (**7**, *m/z* 490.291) from *Amycolatopsis jejuensis* NRRL B-24427.



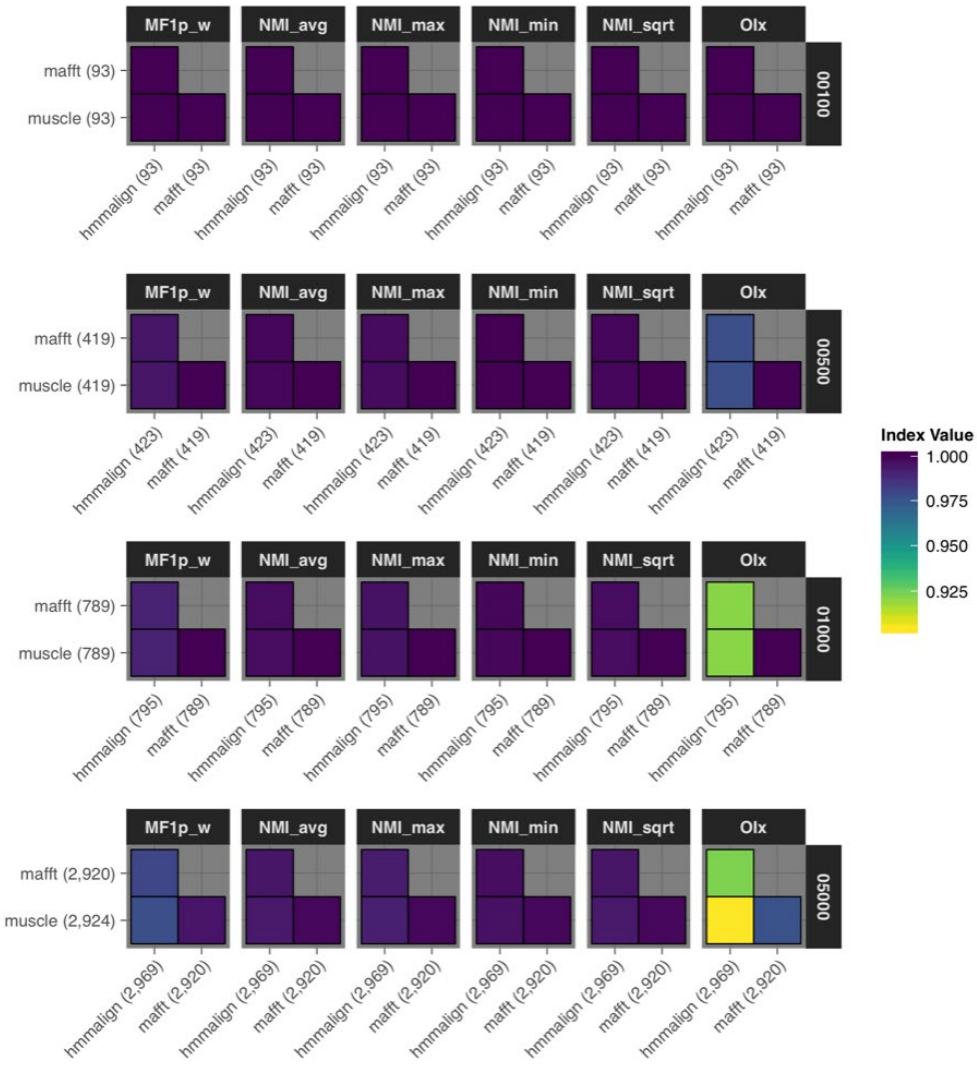
All ions indicative of labeled amino acid incorporation in MS^1 spectra were confirmed by analysis of fragmentation patterns in the corresponding tandem MS data. The loss of one deuteron from d_8-valine in the metabolic labeling experiment is likely due to oxidation and reduction reactions carried out by a P450 enzyme in the *Amycolatopsis jejuensis* NRRL B-24427 detoxin BGC. The oxidations are retained in detoxin P_2, but are fully reduced in detoxins P_1 and P_3. Replicates for labeling experiments were as follows: d_7-Pro incorporation was verified with 65 repeat MS^1 scan measurements and 5 repeat MS^2 scan measurements in one experiment; d_8-Val incorporation was verified with 203 repeat MS^1 scan measurements in one experiment; and d_8,^{15}N-Phe incorporation was verified with 40 repeat MS^1 scan measurements and 3 repeat MS^2 scan measurements in one experiment.

Supplementary Figure 33. Tandem MS fragmentation data for stable isotope labeled-amino acid incorporation of d_8-valine in detoxin P_3 (**7**, *m/z* 490.291) from *Amycolatopsis jejuensis* NRRL B-24427.



a, MS/MS spectrum of d_8-valine-labeled detoxin P_3. **b**, Predicted fragmentation indicates labeled and unlabeled valine and isovaleryl residue incorporation with loss of one deuteron likely due to oxidation and reduction by a P450 enzyme.

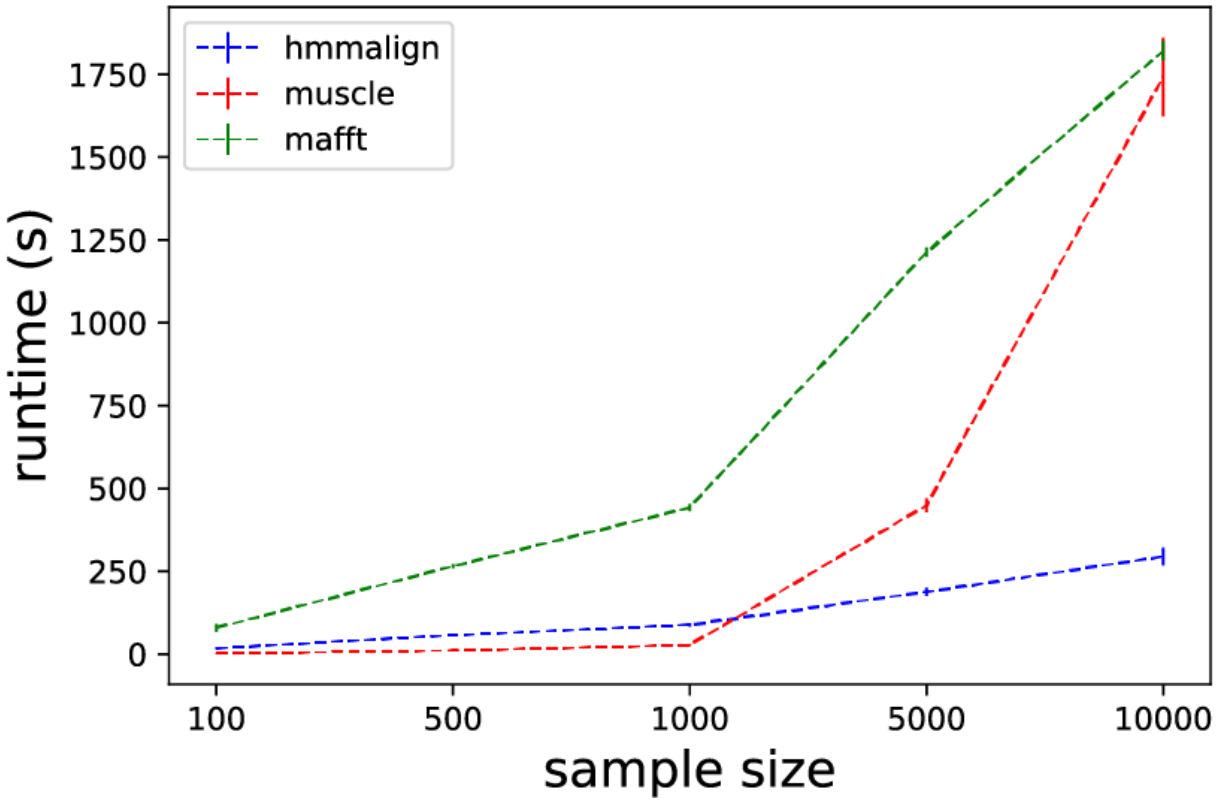
Supplementary Figure 34. Gene cluster family comparison when using other multiple sequence alignment methods



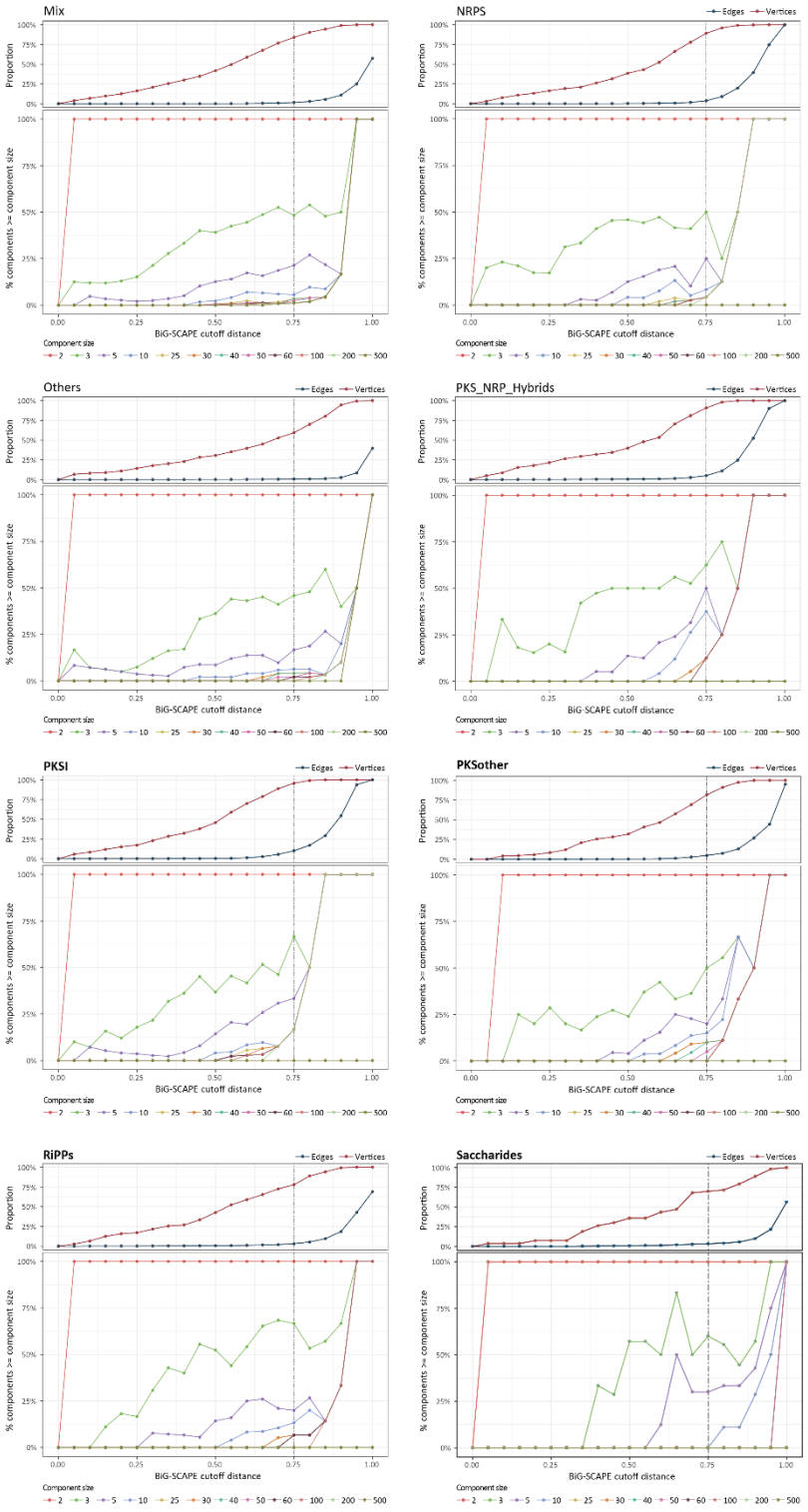
Clustering score comparison while using different multiple sequence alignment approaches. Numbers in parenthesis are the number of gene cluster families obtained with each method. Numbers at the right of each comparison row indicate the number of (randomly chosen) input BGCs. Score ranges and methods are as described in Supplementary Figure 42.

Supplementary Figure 35. Compute times for BiG-SCAPE using different multiple sequence alignment methods.

Average alignment step runtimes of different aligners considered for BiG-SCAPE on samples of 100, 500, 1000, 5000, and 10000 randomly selected BGCs, in three replicates. The runs were performed on a 3.6 GHz Intel i7-4790 CPU (8 cores) with 8GB RAM and 150MB/s SSD.

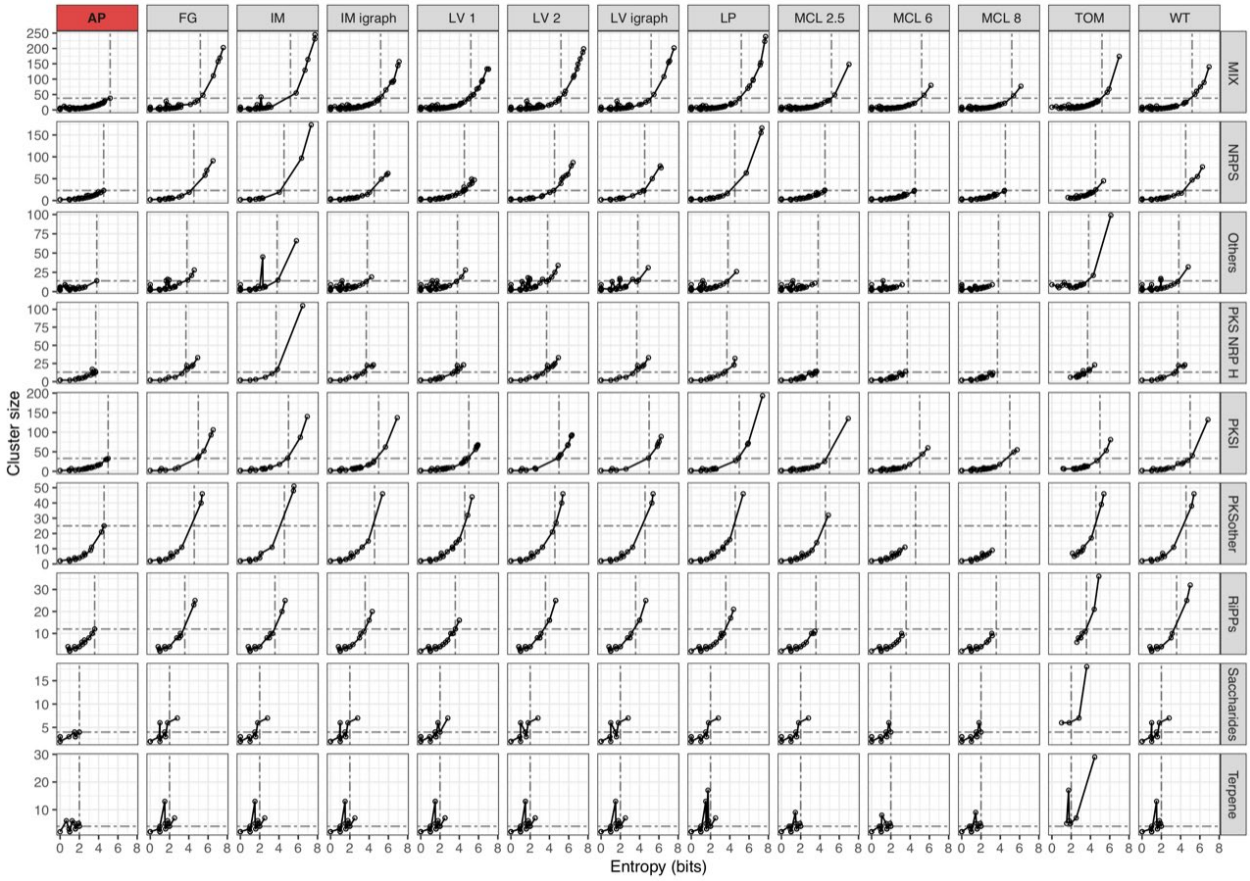


Supplementary Figure 36. Targeted attack of the MIBiG network



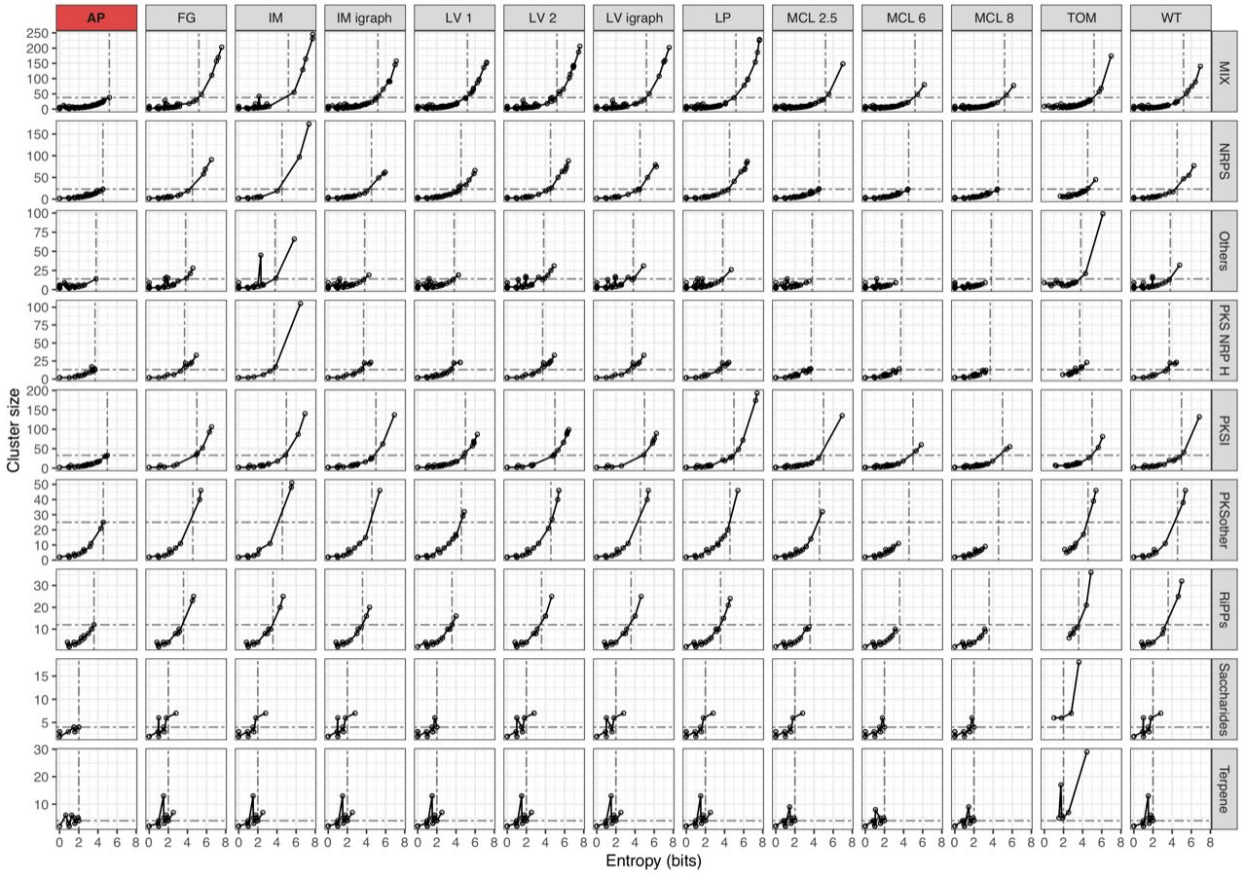
Targeted attack to the different training networks. Panel on top represents the proportion of nodes and edges present in the resulting network after applying the different thresholds. Panel on bottom shows the proportion of components with a number of members larger than the size of the defined component size. Vertical lines represent the threshold selected (0.75)

Supplementary Figure 37. Clustering analysis on the MIBiG network using glocal mode



Overview of the clustering results for the glocal aligned training networks. Each facet shows the relationship between the entropy and the size of the clusters for each clustering method and training network. Grey horizontal and vertical lines are the intersection of the maximum values for the Affinity Propagation (highlighted in red) clustering method for each training network. A good clustering result should be restricted to the third quadrant. AP: affinity propagation; FG: fast greedy; IM: infomap; LV: Louvain; LP; label propagation; MCL: markov clustering; TOM: topological overlap matrix; WT: walktrap.

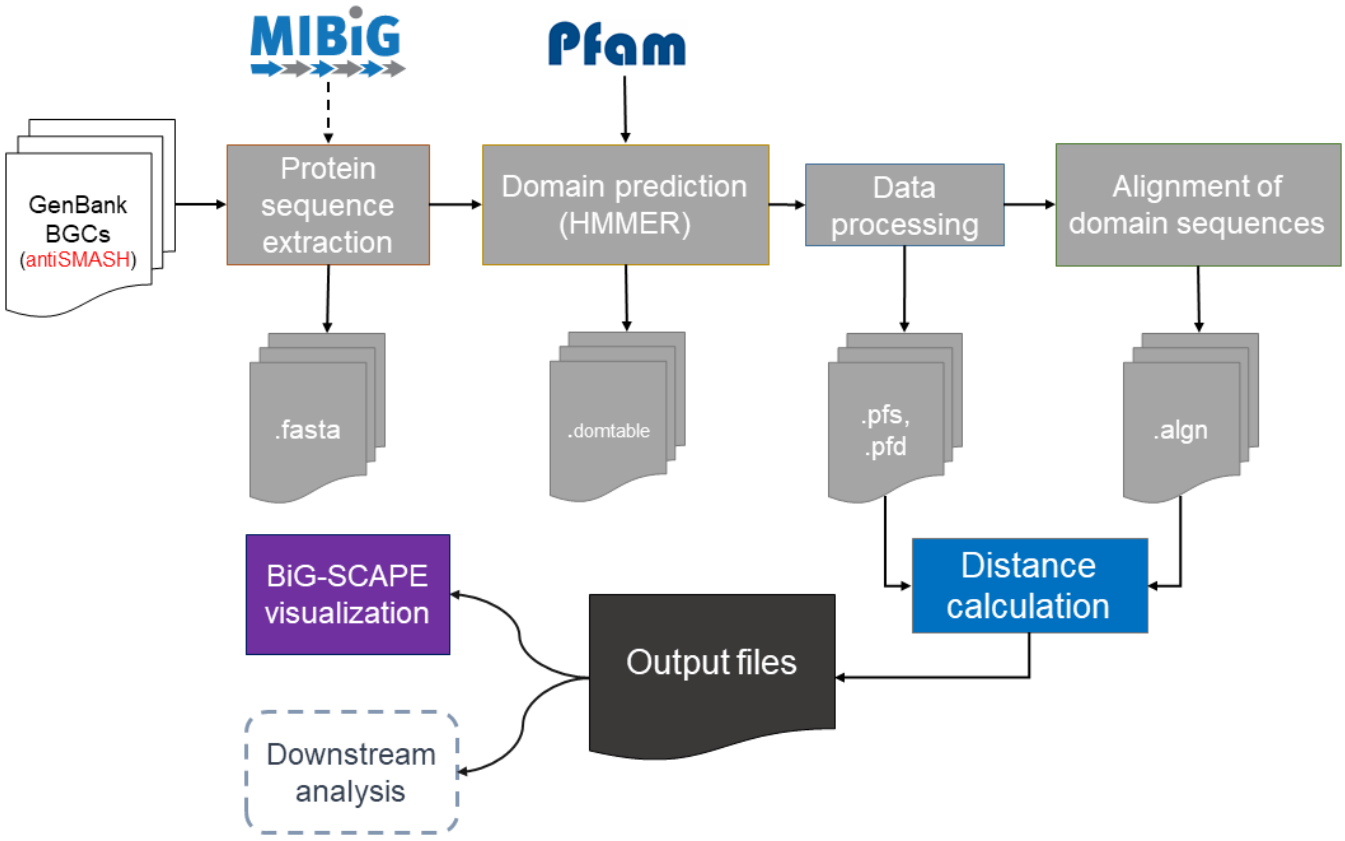
Supplementary Figure 38. Clustering analysis on the MIBiG network using global mode



Overview of the clustering results for the global aligned training networks. Each facet shows the relationship between the entropy and the size of the clusters for each clustering method and training network. Grey horizontal and vertical lines are the intersection of the maximum values for the Affinity Propagation (highlighted in red) clustering method for each training network. A good clustering result should be restricted to the third quadrant. AP: affinity propagation; FG: fast greedy; IM: infomap; LV: Louvain; LP; label propagation; MCL: markov clustering; TOM: topological overlap matrix; WT: walktrap.

Supplementary Figure 39. Flowchart of BiG-SCAPE components.

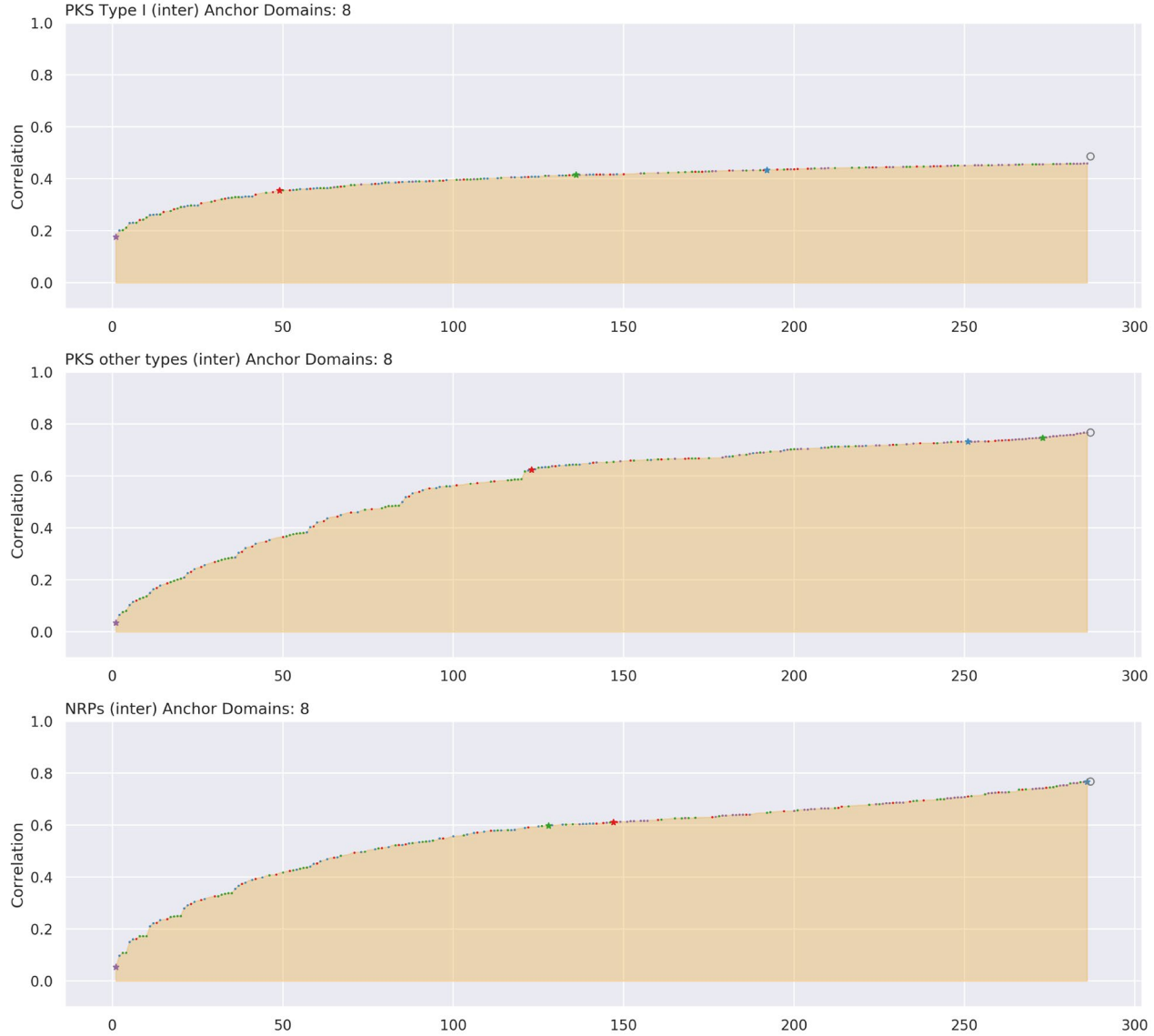
Downstream analysis includes: SVG figures for every cluster; text "network files" that include pairwise distance; a file with annotations for every cluster analyzed (one file per BiG-SCAPE class) and Gene Cluster Family (GCF) labeling.



Supplementary Figure 40. Weight optimization plots

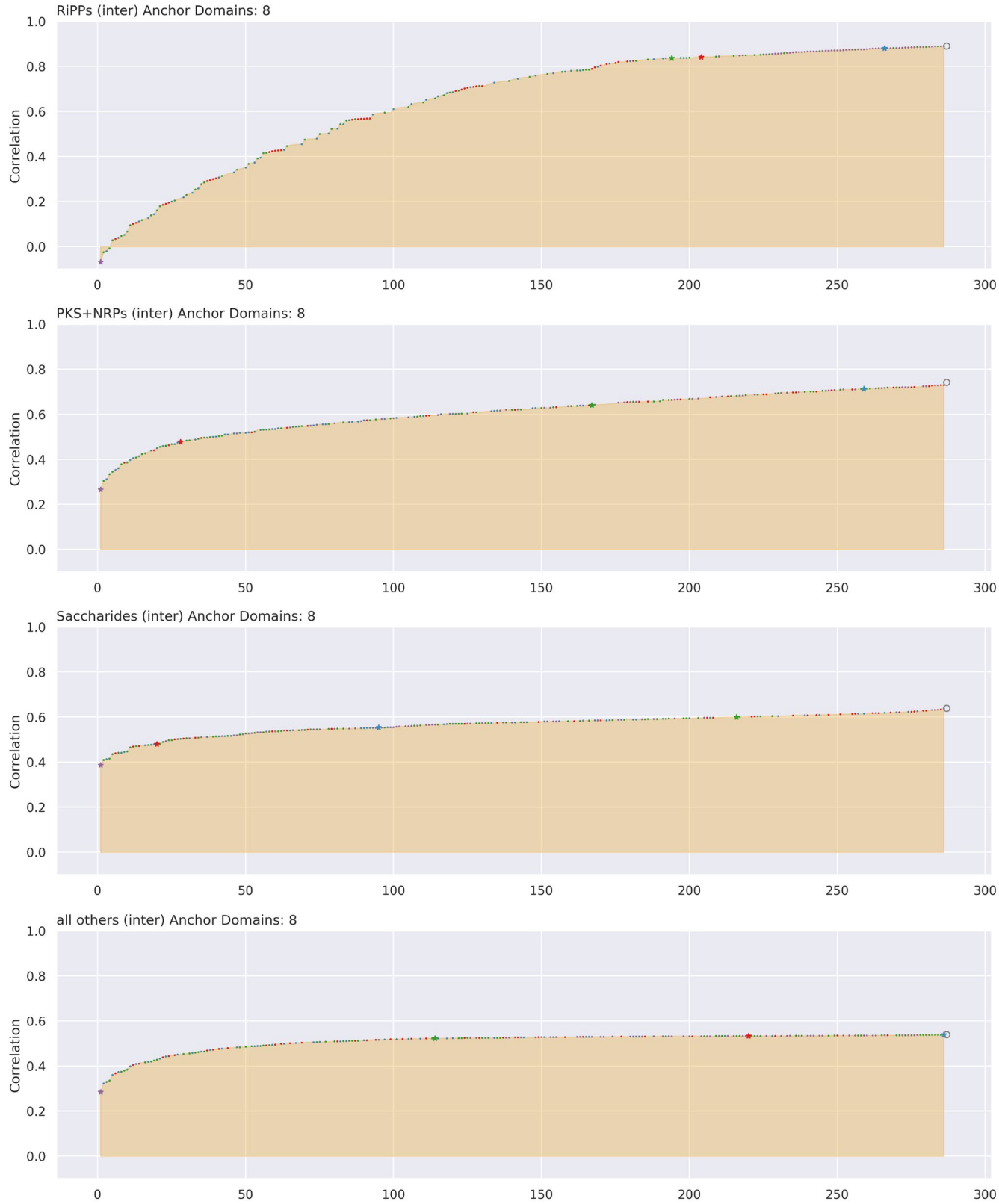
Example plots of weight combinations measuring Pearson's correlation between chemical and BiG-SCAPE distance as described in detail in Supplementary Note 2. These graphs depict a fixed value of the `anchorboost` parameter (`anchorboost`=4), inter group pairings and extended set of anchor domains. Selected combinations of the four indices (J, DSS, GK and AI) are shown (only combinations where all indices' second decimal were zero). Stars indicate combinations where one of the indices is 1.0; colored dots indicate combinations where one of the indices is 0.0 and dark grey hollow circles at the end of the plot indicate the best correlation from the complete set of values.
As can be seen, a single index does not always provide the highest correlation, and when it does (e.g. NRPs, Others), it's not the same index. This indicates that a carefully chosen combination is needed for each case.
Colors: J=red, DSS=blue, GK=purple, AI=green
A larger set of example plots that include intra group pairings and the base set of anchor domains is available in Online Data: Weights optimization plots.

a)



Supplementary Figure 40a. Example correlation plots from the optimization procedure described in Supplementary Note 2 for the PKS I, PKS other and NRPs classes.

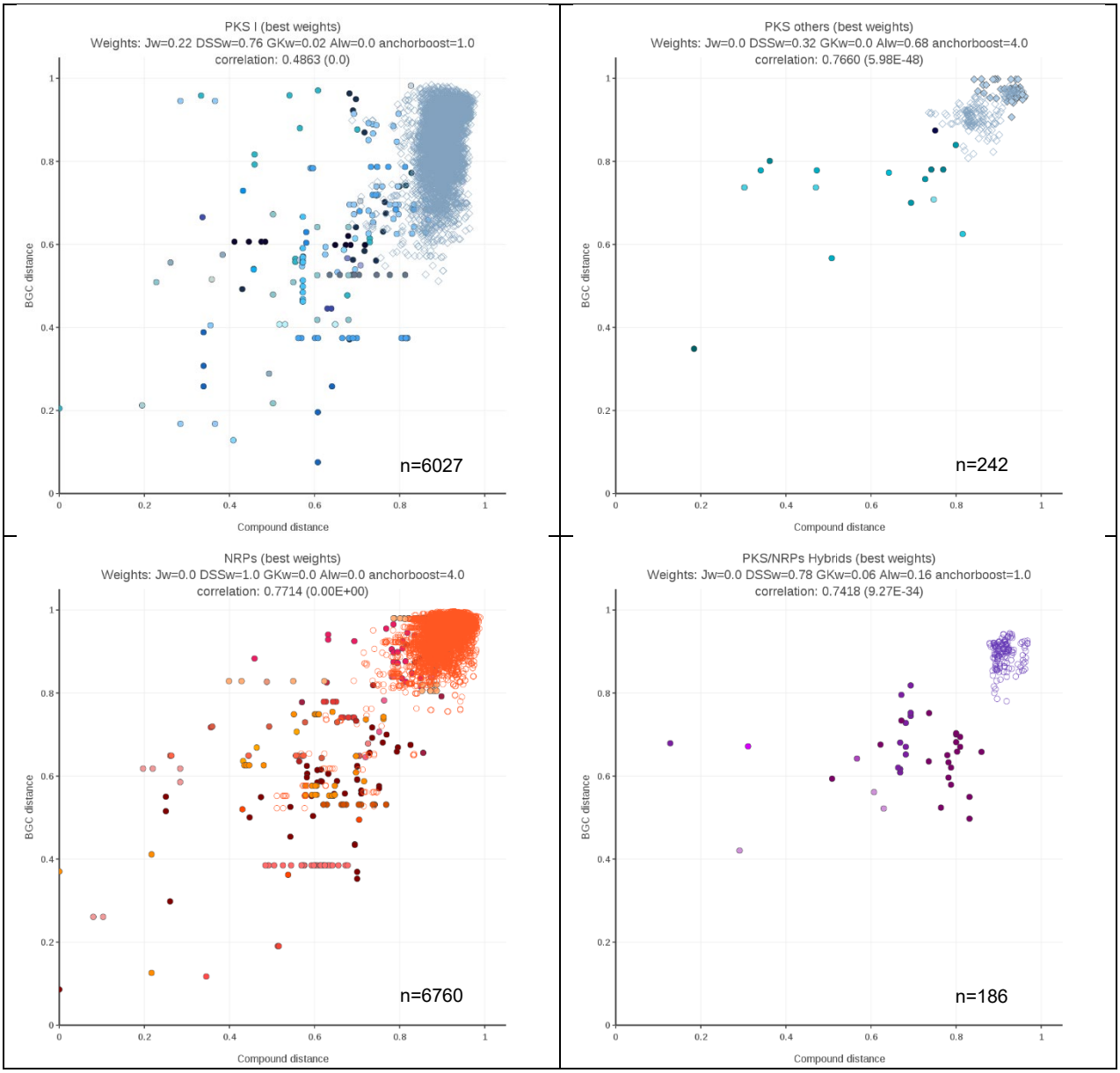
b)

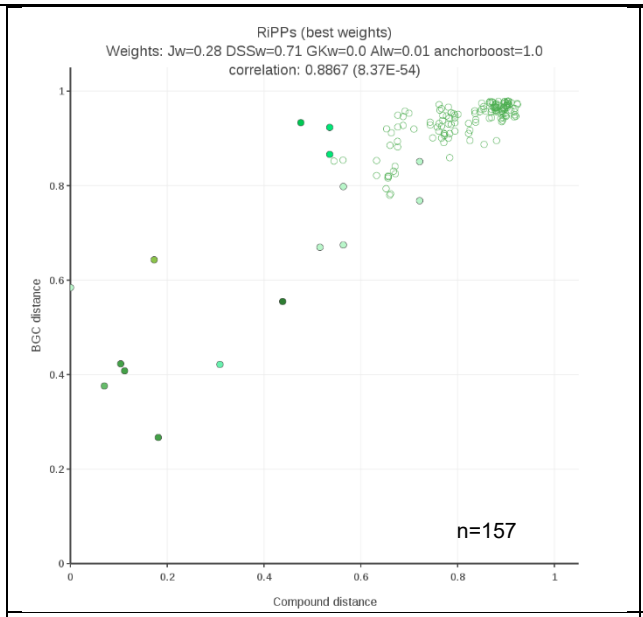


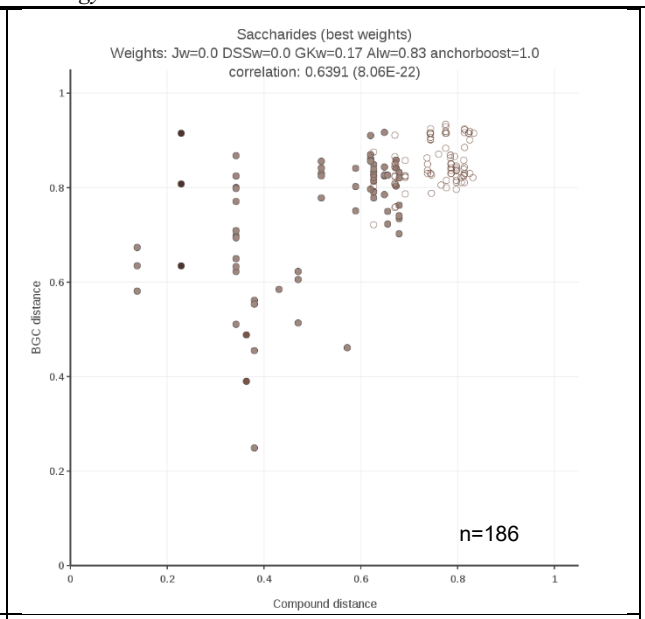
Supplementary Figure 40b. Example correlation plots from the optimization procedure described in Supplementary Note 2 for the RiPPs, PKS/NRPS hybrids, Saccharides and Others classes.

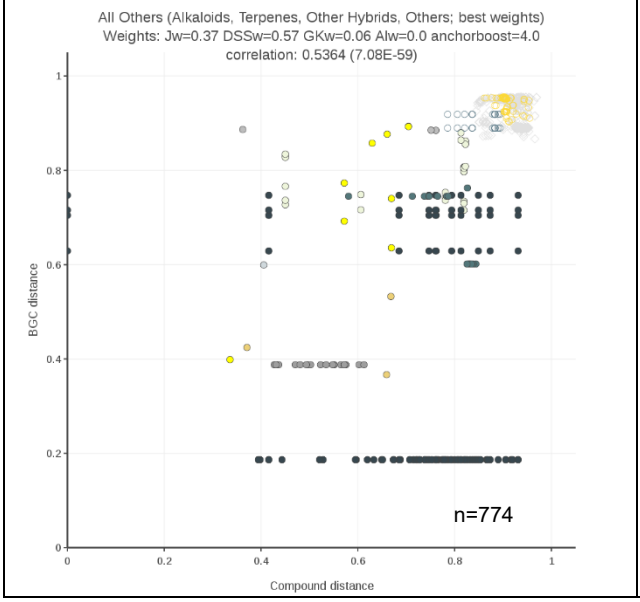
Supplementary Figure 41. Scatterplots with weight optimization results.

Compound distance versus distance as defined by BiG-SCAPE for combinations of weights that maximized Pearson's correlation (see Supplementary Note 2). Filled colored points represent pairs of BGCs within the same curated Compound Group from Supplementary Dataset while hollow points represent pairs of BGCs belonging from different groups but the same Curated Compound Class.



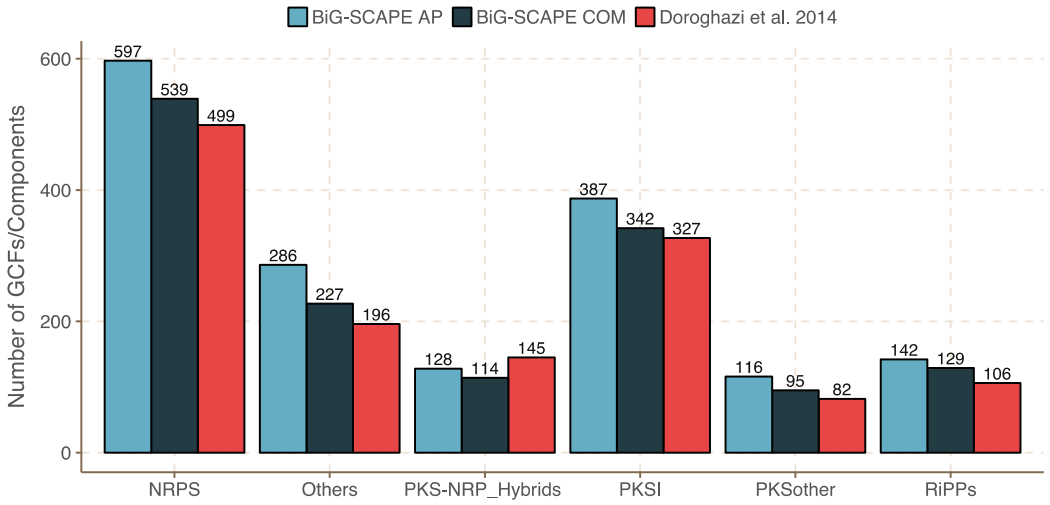

RiPPs (best weights)
Weights: Jw=0.28 DSSw=0.71 GKw=0.0 Alw=0.01 anchorboost=1.0
correlation: 0.8867 (8.37E-54)
BGC distance
Compound distance
n=157


Saccharides (best weights)
Weights: Jw=0.0 DSSw=0.0 GKw=0.17 Alw=0.83 anchorboost=1.0
correlation: 0.6391 (8.06E-22)
BGC distance
Compound distance
n=186

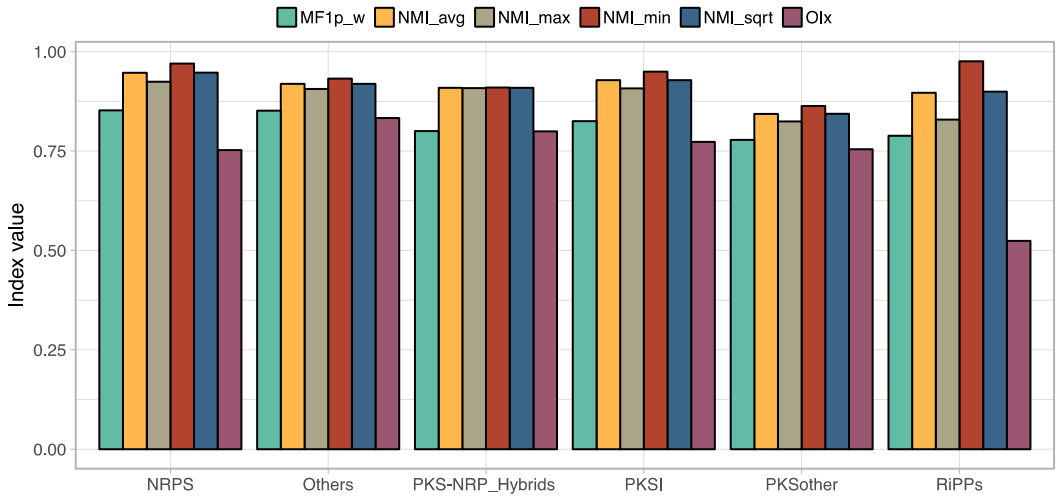

All Others (Alkaloids, Terpenes, Other Hybrids, Others; best weights)
Weights: Jw=0.37 DSSw=0.57 GKw=0.06 Alw=0.0 anchorboost=4.0
correlation: 0.5364 (7.08E-59)
BGC distance
Compound distance
n=774

Supplementary Figure 42. Comparison with results of Doroghazi et al. 2014

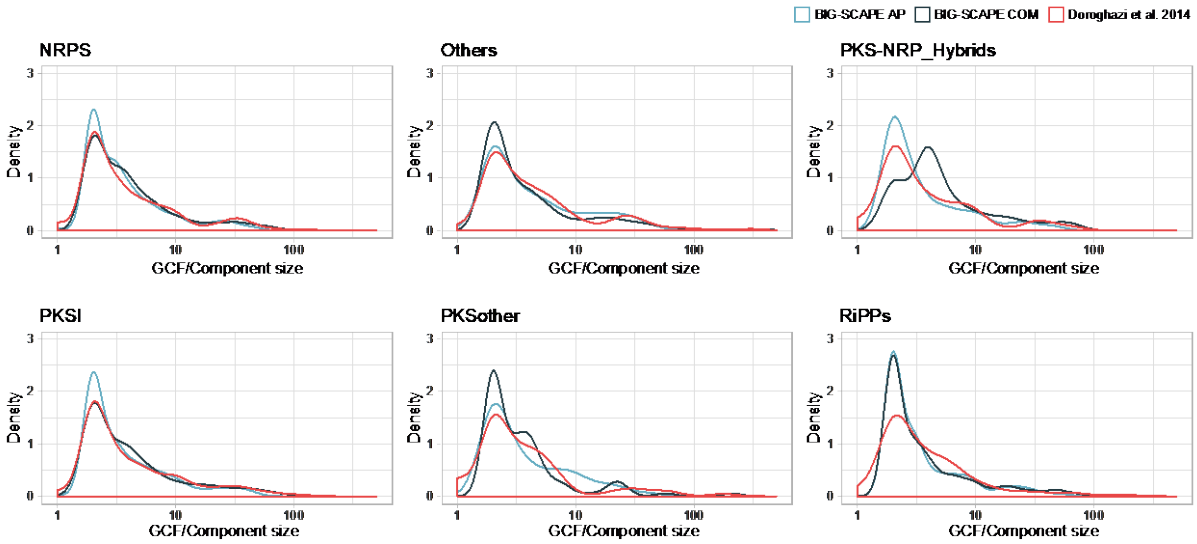
BiG-SCAPE follows a two-step process to infer GCFs, by performing affinity-propagation clustering on the connected components obtained in the initial network; as the Doroghazi et al. method uses no such clustering, the BiG-SCAPE connected components are conceptually (as well as in practice) more similar to the GCFs from Doroghazi et al. (Supplementary Figure 42a). The method from Doroghazi et al. 2014[12] can produce overlapping clusters, meaning that a BGC can be assigned to two different GCFs. We used the Soft Omega Index, Mean F1 Score (F1p) and the Generalized NMI as implemented in xmeasures[15] to evaluate the cluster agreement between BiG-SCAPE and Doroghazi et al. 2014 GCFs using default parameters and specifying `-o` to treat the cluster results as overlapping (Supplementary Figure 42b). All these methods have been designed to evaluate overlapping clustering results. In addition to the cluster evaluation, we also calculated intrinsic cluster properties like the size distribution of the GCFs/components (Supplementary Figure 42c) and the 1:1 ratio between BiG-SCAPE components and the GCFs from Doroghazi et al. 2014 (Supplementary Figure 42d). The GCF data from both data sets have been analyzed with the igraph package v1.1.0[16] and plotted with ggplot2 v3.1.0[17] under the R statistical language v3.5.0[18].



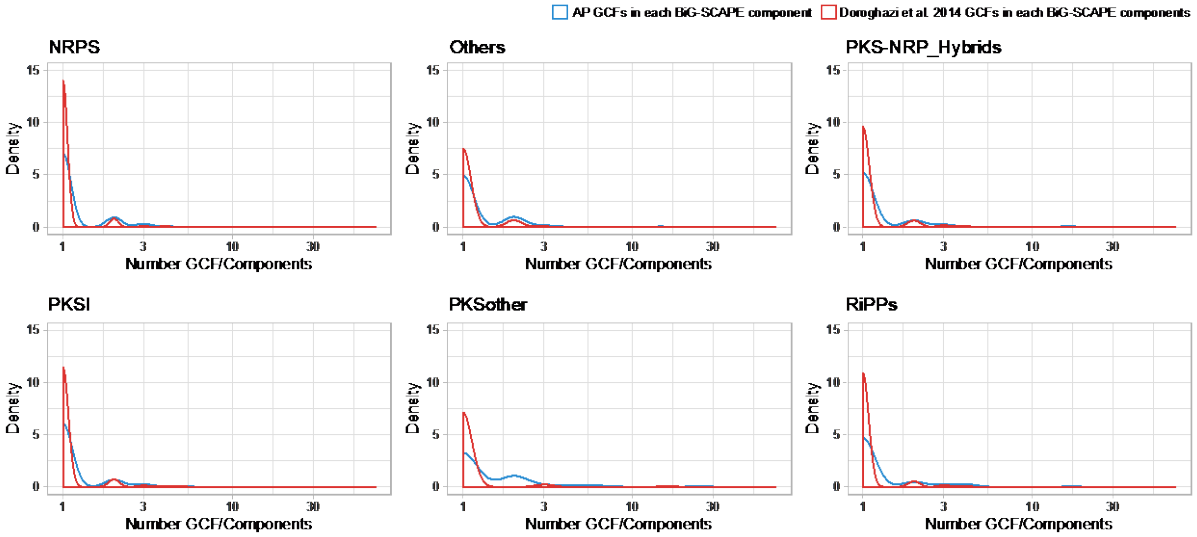
Supplementary Figure 42a. Bar plot with the number GCFs/Components inferred by the Affinity Propagation clustering algorithm (AP) applied in the BiG-SCAPE network, the natural components emerging from the filtering of BiG-SCAPE network (COM), and the GCFs inferred by Doroghazi et al. 2014



Supplementary Figure 42b. Comparison between the natural components emerging from the filtering of BiG-SCAPE network and the Doroghazi et al. 2014 GCFs. Several approaches implemented in xmeasures have been applied: an extension of the Average F1 score (MF1p_w, [0,1]), the Generalized Normalized Mutual Information (NMI, [0,1]) with different normalizations and the Soft Omega Index (OIx, (-1,1]). All comparisons show a good agreement between the BiG-SCAPE components and the Doroghazi et al. 2014 GCFs.

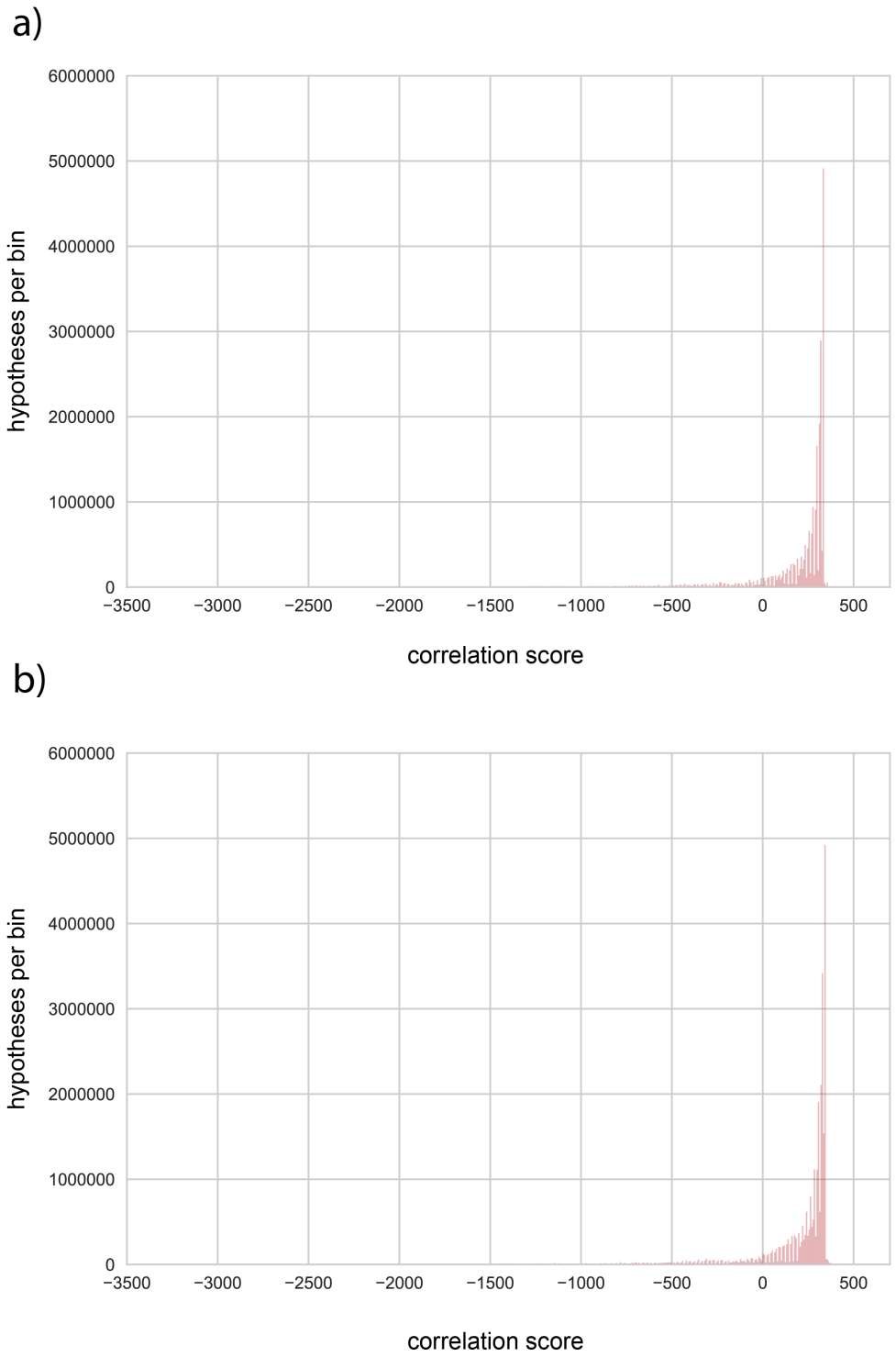


Supplementary Figure 42c. GCF/Component size distribution for the BiG-SCAPE Affinity Propagation GCFs, BiG-SCAPE components and Doroghazi et al. 2014 GCFs.

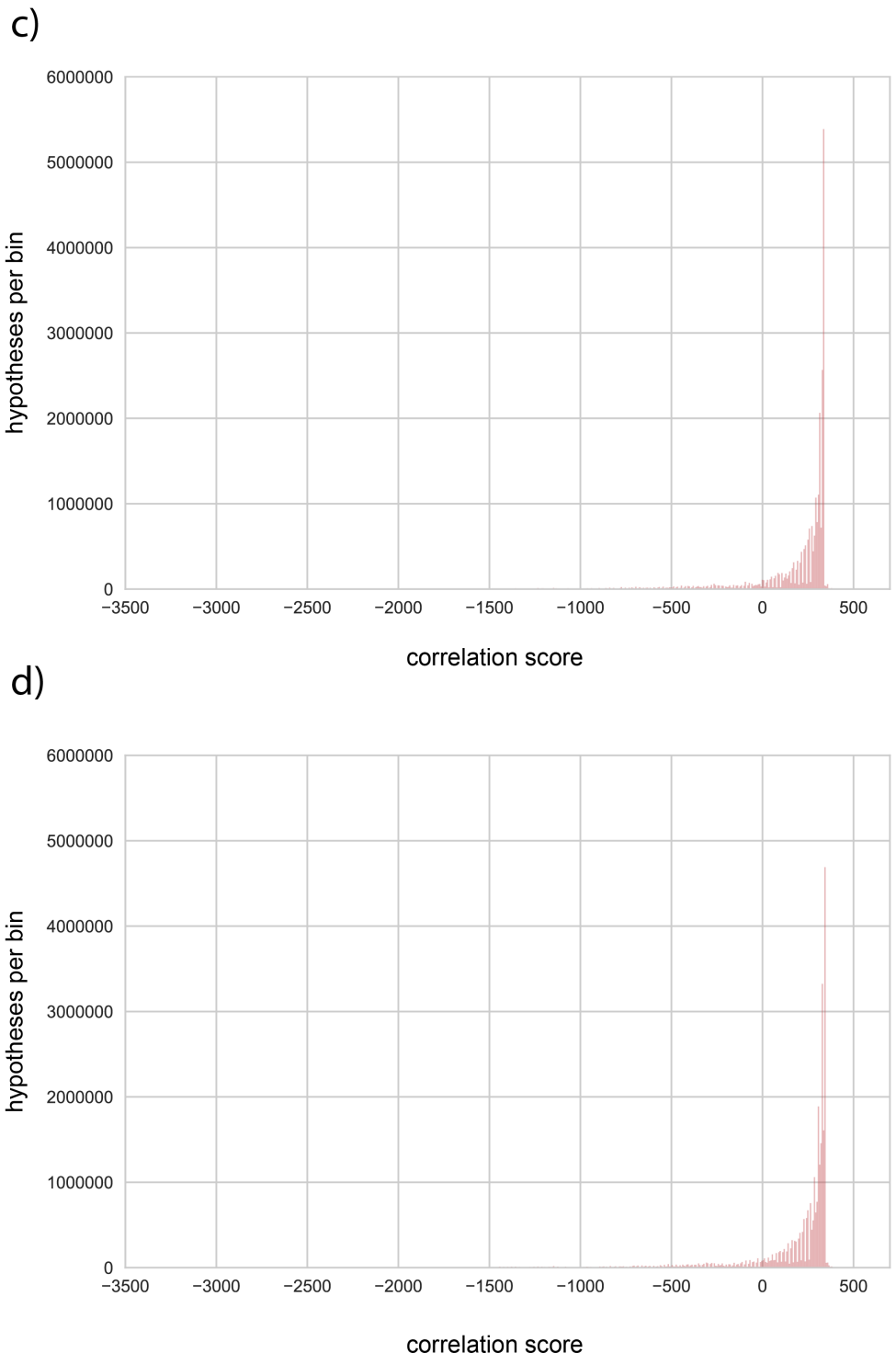


Supplementary Figure 42d. Correspondence between BiG-SCAPE components to BiG-SCAPE Affinity Propagation GCFs and Doroghazi et al. 2014 GCFs.

Supplementary Figure 43. Full metabologenomic correlation histograms



a) BiG-SCAPE global mode, c=0.3; b) BiG-SCAPE global mode, c=0.5. x-axis: correlation score; y-axis: hypotheses per bin.



c) BiG-SCAPE glocal mode, c=0.3; d) BiG-SCAPE glocal mode, c=0.5.

Supplementary Notes

Supplementary Note 1. Index Information

The distance between two Biosynthetic Gene Clusters (BGCs) A and B is calculated by combining three similarity scores:

Jaccard Index: a coefficient of all distinct shared types of domains divided by the total number of distinct domain types:

$$JI_{AB} = \frac{N'_A \cap N'_B}{N'_A \cup N'_B}$$

where $N_{A(B)}$ is the total number of domains in BGC A (B) and $N'_{A(B)}$ is the total number of distinct domains in BGC A (B)

Adjacency Index: a coefficient of all distinct shared pairs of domains divided by the total number of distinct pairs of domain types:

$$AI_{AB} = \frac{P'_A \cap P'_B}{P'_A \cup P'_B}$$

where $P'_{A(B)}$ are all the different adjacent domain pairs types in BGC A (B), e.g.:

$$P'_A = \{(D_A[n], D_A[n+1]) \mid n \in \{0, 1, 2, \dots, |D_A| - 1\}\}$$

where $|D_A|$ is the length of the list of domains predicted in BGC A. For genes in multi-record files (e.g. with information from different loci), the pairs will be formed in the order in which the domains appear in the original file.

Domain Sequence Similarity: a score that considers the sequence similarities for every domain type.

$$DSS = 1 - DSS_d$$

Where DSS_d is the Domain Sequence Dissimilarity. This is divided further into two subcomponents, $DSS_d = c_1 DSS_a + c_2 DSS_{n\alpha}$. The first accounts for so-called *anchor domains*, a list of domains which can be given a special weight (for a list of default anchor domains α, which contains well-known domains for e.g. NRPS or PKS BGCs, see Supplementary Table 7) while the second one accounts for the rest of domains $n\alpha$. Each (non)anchor subcomponent is calculated in the same manner:

$$DSS_\alpha = \frac{1}{S_a} sd(d_A, d_B), \qquad DSS_{n\alpha} = \frac{1}{S_{n\alpha}} sd(d_A, d_B)$$

Where $S_\alpha = \sum_{d \in \{\alpha\}} \max(N_A^d, N_B^d)$; d are all distinct domain types in the pair that belong to the list of anchor domains and $N_{A(B)}^d$ are the number of copies of domain *d* in BGC A(B). *sd* is a function that takes all copies of domain *d* in A and B, and returns the sum of the complement of the *sequence similarity*, $(1 - ss$, the latter calculated with domain sequences aligned against their hmm profile using `hmmalign`) of the best matching copies of the same domain type (using the Munkres algorithm). For extra copies that don't have a match or unshared domains, the function returns 1 (a complete dissimilarity).
Finally, if there exist domains of each kind in the pair, both subcomponents are weighted first proportionally to the total number of domains of each type (including copies):

$$w_a = \frac{S_a}{S_a + S_{na}}, \qquad w_{na} = \frac{S_{na}}{S_a + S_{na}}$$

and then re-weighted to increase the perceived amount of anchor domains through the "`anchorboost`" parameter:

$$c_1 = \frac{w_a \times anchorboost}{w_a \times anchorboost + w_{na}}, \qquad c_2 = \frac{w_{na}}{w_a \times anchorboost + w_{na}}$$

Supplementary Note 2. BiG-SCAPE classes weight optimization

Weight optimization for BiG-SCAPE was performed as described in the Online Methods with the following characteristics.

Within DSS, two sets of "anchor domains" were used: an initial set with Condensation Domain, PF00668; Beta-ketoacyl synthase, N-terminal domain (PF00109), Beta-ketoacyl synthase, C-terminal domain, (PF02801) and Terpene synthase, N-terminal domain (PF01397) and an extended set that included AMP-binding enzyme (PF00501), Lanthionine synthase C-like protein (PF05147), Chalcone and stilbene synthases, N-terminal domain (PF00195) and Chalcone and stilbene synthases, C-terminal domain (PF02797).

Base correlation uses manually assigned default BiG-SCAPE weights: Jw = 0.2, DSSw = 0.75, GKw = 0.05, AIw = 0.0, as well as anchorboost = 2.0. A value of 1 for anchorboost means no change in perceived proportion of anchor domains.
Best correlation: Ranges used for the optimization are: Jw, DSSw, Gkw, AIw $\in$ [0, 1] with the condition that Jw + DSSw + GKw + AIw = 1. Anchorboost (ab) $\in$ [1,4]. Optimization step: 0.01 except for anchorboost (0.5).
Best weights are in the format Jw, DSSw, GKw, AIw. Numbers in parenthesis are the P-values calculated by the Pearson function from Python.

Intra-groups: Each data point is comprised of pairs of nodes where both belong to the same group, for each group within the class of the Curated Compound Groups table (Supplementary Dataset).
Inter groups: Data points comprise pairs of nodes where each belong to the same class (but might belong to different groups).

Additionally, only data points with at least two predicted domains were considered. All calculated scores are available in Online Data: Weights optimization results. See also Supplementary Figures 40 and 41.

BiG-SCAPE Classes / Curated Compound classes:

- PKS type I (Family: Polyketides, Subfamily: Type I Mechanism)
- PKS Other Types (Family: Polyketides, Subfamily: Type II Mechanism, Type III Mechanism, Other Mechanism)
- NRPs
- RiPPs
- Polyketides/NRP hybrids
- Saccharides
- All other families (Alkaloids, Terpenes, Other Hybrids, Others)

Polyketides Type I Mechanism

Intra groups (188 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.2647 (0.0002)	0.2660 (0.0002)
Best correlation	0.3056 (1.99e-05)	0.3056 (1.99e-05)
Best weights	Jw:0.63, DSSw:0.31, GKw:0.06, AIw:0.0, ab: 1.0	Jw:0.63, DSSw:0.31, GKw:0.06, AI:0.0, ab: 1.0

Inter groups (6027 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.4676 (0.0)	0.4698 (0.0)
Best correlation	0.4863 (0.0)	0.4863 (0.0)
Best weights	Jw:0.22, DSSw:0.76, GKw:0.02, AI:0.0, ab: 1.0	Jw:0.22, DSSw:0.76, GKw:0.02, 0.0, AI:ab: 1.0

Chosen weights: Jw:0.22, DSSw:0.76, AI:0.02, anchorboost:1.0

Polyketides Type II Mechanism, Type III Mechanism, Other Mechanism

Intra groups (17 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.2623 (0.3089)	0.2495 (0.3340)
Best correlation	0.4552 (0.0663)	0.4552 (0.0663)
Best weights	Jw:0.0, DSSw:0.0, GKw:0.0, AI:1.0, ab: 1.0	Jw:0.0, DSSw:0.0, GKw:0.0, AI:1.0, ab: 1.0

Inter groups (242 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.7294 (1.81e-41)	0.7296 (1.73e-41)
Best correlation	0.7659 (5.97e-48)	0.7674 (3.14e-48)
Best weights	Jw:0.0, DSSw:0.32, GKw:0.0, AI:0.68, ab: 4.0	Jw:0.0, DSSw:0.33, GKw:0.0, AI:0.67, ab: 4.0

Chosen weights: Jw:0.0, DSSw:0.32, AI:0.68, anchorboost:4.0

NRPs

Intra groups (286 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.6074 (3.08e-30)	0.6110 (1.16e-30)
Best correlation	0.6606 (3.02e-37)	0.6556 (1.59e-36)
Best weights	Jw:0.0, DSSw:1.0, GKw:0.0, AI:0.0, ab: 4.0	Jw:0.0, DSSw:1.0, GKw:0.0, AI:0.0, ab: 4.0

Inter groups (6760 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.7469 (0.0)	0.7473 (0.0)
Best correlation	0.7714 (0.0)	0.7678 (0.0)
Best weights	Jw:**0.0,** DSSw:**1.0,** GKw:**0.0,** AI:**0.0,** ab**: 4.0**	Jw:0.01, DSSw:0.98, GKw:0.0, AI:0.01, ab: 3.5

Chosen weights: Jw:0.0, DSSw:1.0, AI:0.0, anchorboost:4.0

RiPPs

Intra groups (16 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.7812 (0.0003)	0.7832 (0.0003)
Best correlation	0.8846 (5.34e-06)	0.8845 (5.37e-06)
Best weights	Jw:0.04, DSSw:0.43, GKw:0.53, AI:0.0, ab: 4.0	Jw:0.04, DSSw:0.43, GKw:0.53, AI:0.0, ab: 1.5

Inter groups (157 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.8696 (2.17e-49)	0.8726 (4.02e-50)
Best correlation	0.8867 (8.36e-54)	0.8906 (6.29e-55)
Best weights	Jw:0.28, DSSw:0.71, GKw:0.0, AI:0.01, ab: 1.0	Jw:0.23, DSSw:0.74, GKw:0.0, AI:0.03, ab: 4.0

Chosen weights: Jw:0.28, DSSw:0.71, AI:0.01, anchorboost:1.0

PKS/NRPs hybrids

Intra groups (37 points)

	Anchor domains:4	Anchor domains:8
Base correlation	-0.0209 (0.9019)	0.0108 (0.9490)
Best correlation	0.3507 (0.0332)	0.3507 (0.0332)
Best weights	Jw:0.0, DSSw:0.0, GKw:0.33, AI:0.67, ab: 1.0	Jw:0.0, DSSw:0.0, GKw:0.33, AI:0.67, ab: 1.0

Inter groups (186 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.7165 (1.33e-30)	0.7192 (6.40e-31)
Best correlation	0.7418 (9.26e-34)	0.7418 (9.26e-34)
Best weights	Jw:0.0, DSSw:0.78, GKw:0.06, AI:0.16, ab: 1.0	Jw:0.0, DSSw:0.78, GKw:0.06, AI:0.16, ab: 1.0

Chosen weights: Jw:0.0, DSSw:0.78, AI:0.22, anchorboost:1.0
These weights were chosen due to the Goodman-Kruskal index being dropped in the final version of BiG-SCAPE.

Saccharides

Intra groups (80 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.3860 (0.0004)	0.3841 (0.0004)
Best correlation	0.4848 (5.16e-06)	0.4848 (5.16e-06)
Best weights	Jw:0.0, DSSw:0.0, GKw:0.21, AI:0.79, ab: 1.0	Jw:0.0, DSSw:0.0, GKw:0.21, AI:0.79, ab: 1.0

Inter groups (186 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.5703 (9.57e-17)	0.5689 (1.16e-16)
Best correlation	0.6390 (8.05e-22)	0.6390 (8.05e-22)
Best weights	Jw:0.0, DSSw:0.0, GKw:0.17, AI:0.83, ab: 1.0	Jw:0.0, DSSw:0.0, GKw:0.17, AI:0.83, ab: 1.0

Chosen weights: Jw:0.0, DSSw:0.0, AI:1.0, anchorboost:1.0. These weights were chosen due to the Goodman-Kruskal index being dropped in the final version of BiG-SCAPE

All other groups (Alkaloids, Terpenes, Other Hybrids, Others)

Intra groups (262 points)

	Anchor domains:4	Anchor domains:8
Base correlation	-0.1182 (0.0559)	-0.1163 (0.0599)
Best correlation	0.0751 (0.2252)	0.0751 (0.2252)
Best weights	Jw:0.0, DSSw:0.0, GKw:1.0, AI:0.0, ab: 1.0	Jw:0.0, DSSw:0.0, GKw:1.0, AI:0.0, ab: 1.0

Inter groups (774 points)

	Anchor domains:4	Anchor domains:8
Base correlation	0.5355 (1.17e-58)	0.5367 (5.90e-59)
Best correlation	0.5363 (7.07e-59)	0.5392 (1.32e-59)
Best weights	Jw:0.37, DSSw:0.57, GKw:0.06, AI:0.0, ab: 4.0	Jw:0.01, DSSw:0.97, GKw:0.02, AI:0.0, ab: 4.0

Chosen weights: Jw:0.01, DSSw:0.97, AI:0.02, anchorboost:4.0

Due to the lack of BGCs related to the curated Terpene class, the base set of weights were chosen for this BiG-SCAPE class: Jw:0.2, DSSw:0.75, AI:0.05, anchorboost:2.0

REFERENCES

1. Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. *Science (80-)*. 2007;315(5814):972-976. doi:10.1126/science.1136800
2. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12(Oct):2825-2830.
3. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Phys Rev E*. 2004;70(6):066111. doi:10.1103/PhysRevE.70.066111
4. Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. *Phys Rev E*. 2007;76(3):036106. doi:10.1103/PhysRevE.76.036106
5. Pons P, Latapy M. Computing Communities in Large Networks Using Random Walks. In: Springer, Berlin, Heidelberg; 2005:284-293. doi:10.1007/11569596_31
6. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci*. 2008;105(4):1118-1123.
7. Rosvall M, Axelsson D, Bergstrom CT. The map equation. *Eur Phys J Spec Top*. 2009;178(1):13-23.
8. Blondel V, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech*. 2008;10:P10008.
9. Van Dongen SM. Graph clustering by flow simulation. 2000.
10. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4(1).
11. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9(1):559. doi:10.1186/1471-2105-9-559
12. Doroghazi JR, Albright JC, Goering AW, et al. A roadmap for natural product discovery based on large-scale genomics and metabolomics. *Nat Chem Biol*. 2014;10(11):963-968. doi:10.1038/nchembio.1659
13. Argimón S, Abudahab K, Goater RJE, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genomics*. 2016;2(11).
14. Ogita T, Seto H, Otake N, Yonehara H. The Structures of Minor Congeners of the Detoxin Complex. *Agric Biol Chem*. 1981;45(11):2605-2611.
15. Lutov A, Khayati M, Cudré-Mauroux P. Accuracy Evaluation of Overlapping and Multi-resolution Clustering Algorithms on Large Datasets. February 2019. http://arxiv.org/abs/1902.01691. Accessed May 20, 2019.
16. Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal*. 2006;Complex Sy:1695. http://igraph.org.
17. Wickham H, Chang W, others. ggplot2: An implementation of the Grammar of Graphics. *R Packag version 07, URL http//CRAN R-project org/package= ggplot2*. 2008.
18. Team RC, others. R: A language and environment for statistical computing. 2013.