MIBiG 3.0: A community-driven effort to annotate experimentally validated biosynthetic gene clusters

Abstract

With an ever-increasing amount of (meta)genomic data being deposited in sequence databases, (meta)genome mining for natural product biosynthetic pathways occupies a critical role in the discovery of novel pharmaceutical drugs, crop protection agents and biomaterials. The genes that encode these pathways are often organised into biosynthetic gene clusters (BGCs).

In 2015, we defined the Minimum Information about a Biosynthetic Gene cluster (MIBiG): a standardised data format that describes the minimally required information to uniquely characterise a BGC. We simultaneously constructed an accompanying online database of BGC entries in this format, which has since been widely used by the community as reference dataset for BGCs. Here, we describe MIBiG 3.0, a database update comprising large-scale validation and re-annotation of existing entries, extensive cross-linking to the Natural Product Atlas database, and 669 new entries. Particular attention was paid to the annotation of compound structures and biological activities, as well as protein domain selectivities. Together, these new features keep the database up-to-date, and will provide new opportunities for the scientific community to use its freely available data, e.g. for the training of new machine learning models to predict sequence-structure-function relationships for diverse natural products.

Introduction

Across all kingdoms of life, organisms produce specialised metabolites: molecules that are produced by bacteria, fungi and plants to gain an advantage over their competitors in challenging environments. Specialised metabolites exhibit a wide variety of biological activities, including many that are useful for pharmaceutical and agricultural applications, e.g. antibiotics, anti-cancer drugs, pesticides and herbicides. The production of specialised metabolites, also referred to as secondary metabolites, is typically encoded by biosynthetic gene clusters (BGCs): groups of co-localised genes that jointly encode a biosynthetic pathway. Therefore, novel specialised metabolites can be discovered by detecting BGCs in microbial and plant genomes and predicting their encoded products and functions. Similar to how the relationship between DNA, mRNA

and protein describes the flow of information in cells, we can define a ‘central dogma’ of specialised metabolism: a BGC sequence encodes a set of enzymes, which together assemble a compound structure, and this compound structure dictates specialised metabolite function. Understanding how information is translated from sequence to structure to function is key to natural product discovery. To address the first stage, sequence information, various tools have been developed that automatically detect BGCs from DNA sequence, including antiSMASH and its siblings fungiSMASH and plantiSMASH (Blin et al. 2021; Kautsar et al. 2017), GECCO (Carroll et al. 2021), DeepBGC (Hannigan et al. 2019), RiPPMiner (Agrawal et al. 2017), and PRISM 4 (Skinnider et al. 2020).

To facilitate dereplication and comparative analysis of predicted BGCs with known BGCs, and to characterise the interplay between sequence, structure and function, standardised data annotation and storage are essential. To this purpose, we developed the Minimum Information about a Biosynthetic Gene cluster (MIBiG) standard, and built a database which contains standardised entries for experimentally validated BGCs of known function (Kautsar et al. 2020; Medema et al. 2015). Each entry minimally contains information about the nucleotide entry and coordinates of the genomic locus involved, the producing organism’s taxonomy, biosynthetic class, name of the produced compound(s), and literature reference(s). There are also various optional fields for non-minimal entries, including fields for gene function, product structure and bioactivity, crosslinks to chemical structure databases such as NP Atlas and PubChem, and monomer identity. With MIBiG 2.0 containing over 2000 entries, the database has become an important reference for many researchers that mine genomes for natural products. For example, it has been used to estimate the potential for biosynthetic novelty in large-scale microbiome studies (Paoli et al. 2022; Nayfach 2020), to identify conserved amino acids playing key roles in catalytic activities across enzyme families (Izoré et al. 2021), to help guide natural product discovery efforts towards high-potential taxa (Gavriilidou et al. 2022), and to train machine-learning algorithms for natural product activity prediction (Walker and Clardy 2021).

Here, we present MIBiG 3.0: an update designed around increasing the number of non-minimal entries in our database and adding new data entries through a large-scale community annotation effort. We focused on three features: the characterisation and cross-linking of 918 chemical structures; the annotation of 1002 bioactivities of BGC products; and the validation and annotation of 2027 protein domain substrates of nonribosomal peptide synthetases (NRPSs). In addition, we added 669 novel BGCs to the MIBiG database which were published since the last database update, and removed 63 duplicate and low-quality entries (Figure 1). Together, these additions keep the database current,

and provide unique opportunities for exploring complex sequence-structure-function relationships in diverse natural product domains.

Methods and implementation

Manual curation through crowdsourcing and mass online 'annotathons'

As authors themselves typically have the best understanding of the BGC they have studied, we greatly encourage natural product researchers to submit their BGCs to MIBiG during the process of publishing their work. To this purpose, MIBiG supplies an online form through which researchers can request a unique MIBiG identifier and submit their experimentally verified BGCs, pre- or post publication. Since MIBiG version 2.0, this has yielded 97 manually submitted, high-quality entries which have now been incorporated into MIBiG 3.0. Still, there are far more published BGCs that are not manually submitted to MIBiG.

With more papers containing novel BGCs being published every year, manually annotating, validating, and adding BGCs to MIBiG has become a mammoth task. Therefore, we took to social media to gauge the community's interest in participating in an online annotation event. We received many positive responses, with 86 people from four different continents volunteering to participate in our 'annotathons'. We organised eight three-hour online sessions, accommodating different time-zones, with various breakout rooms dedicated to specific annotation tasks: annotating new clusters, annotating and cross-linking compound structures, annotating compound bioactivities, and assigning substrate selectivities to NRPS protein domains. We prepared multiple instruction videos and assigned an expert to each of the breakout rooms who could be directly approached with questions from annotators to ensure that annotation quality was consistent. In addition, one of our annotators at the CINVESTAV research institute mobilised fourteen MSc students of their 2021 Bacterial Genomics class of Integrative Biology to annotate compound bioactivities under supervision. Finally, we resolved 125 database issues that were raised by users on our Github page, redefining BGC boundaries, correcting biosynthetic classes, adding and removing literature references, fixing compound structures, and removing duplicate entries.

Annotating and cross-linking compound structures

Since version 2.0, compound structures in MIBiG have been cross-linked to the NP Atlas database: a database of natural product structures isolated from bacteria and fungi. For version 3.0, we collaborated with the NP Atlas team to 1) add structures for compounds in SMILES format, including stereochemical information where possible, and 2) cross-link them to five databases of chemical structures: NP Atlas (van Santen et al. 2022), PubChem, ChemSpider, LOTUS (Rutz et al. 2022), and ChEMBL (Gaulton et al. 2016). If compound entries were found in multiple databases, SMILES strings from NP Atlas were prioritised. SMILES strings were also collected for existing entries that were already cross-linked to a database but did not report a SMILES string. Correctness of SMILES syntax was validated with PIKAChU (Terlouw, Vromans, and Medema 2022).

Annotating compound bioactivities

To improve MIBiG as a resource for machine learning models predicting sequence-structure-function relationships, we added bioactivity data for 1002 compounds and chemical target data for 95 compounds. 708 of these annotations were transferred from the dataset assembled by Walker and Clardy, who designed a machine learning model to predict BGC function from sequence (Walker and Clardy 2021). To accommodate these annotations, we added 40 functional categories in addition to the eight categories previously described in MIBiG (Supplementary Table 1).

Annotating NRPS protein domains

To concretise the relationship between NRPS sequence and the structure of its produced nonribosomal peptide (NRP), we annotated and validated the substrate selectivities of 2782 NRPS adenylation (A) domains. A-domains dictate which amino acid monomers are incorporated into (hybrid) NRP scaffolds. Substrate annotation can be performed at different levels: we can define the pre-tailored substrate precursor (e.g. L-aspartic acid); the substrate as recognised by the A-domain (e.g. (3R)-3-hydroxy-L-aspartic acid); or the post-tailored integrated monomer that ends up in the final NRP scaffold (e.g. (3R)-3-hydroxy-D-aspartic acid). We chose to annotate the substrates as recognised by the A-domain, as this best reflects the biological relationship between A-domain and monomer. In addition to substrate identity, we also recorded evidence for substrate selectivity in the form of an evidence code and literature references. To this purpose, we added 13 evidence codes to the JSON schema which is used to standardise MIBiG entries (Table 1).

Table 1. Evidence codes for adenylation domain substrate annotations.

Evidence code	**Accepted as standalone evidence code**	**New in MIBiG 3.0**
Activity assay	X	
ACVS assay	X	X
ATP-PPi exchange assay	X	X
Enzyme-coupled assay	X	X
Feeding study	X	
Heterologous expression	X	X
Homology		X
HPLC	X	X
In-vitro experiments	X	X
Knock-out studies	X	X
Mass spectrometry	X	X
NMR	X	X
Radio labelling	X	X
Sequence-based prediction		
Steady-state kinetics	X	X
Structure-based inference	X	
X-ray crystallography	X	X

After community annotation, substrate naming was homogenised and each stereochemically ambiguous substrate was manually curated by an expert. Where stereochemistry could be inferred from structure, this is reflected in the substrate name for each stereocenter. Exceptions are amino acid names, which are assumed to be in their L-configuration. To avoid any ambiguity in substrate naming, we also linked each of our 274 unique substrate names to an isomeric SMILES string representing the substrate structure (Figure 2; Supplementary Table 2). SMILES validation and deduplication were handled using PIKAChU (Terlouw, Vromans, and Medema 2022).

Results and discussion

Taking the 'minimal' out of MIBiG

While MIBiG 2.0 serves an important role in the community as a reference database to quickly identify whether a BGC is similar to any known BGCs, its utility as a resource for exploring sequence-structure-function relationships could be improved. This can mainly be explained by the high number of minimal entries in the database: entries that only contain sequence and compound information that could be augmented by adding further standardised annotations. For MIBiG 3.0, we aimed to promote as many existing and novel entries as possible to non-minimal entries by annotating compound structures (918), bioactivities (1002), and NRPS substrates (2027). In total, we added 669 novel BGCs and 4553 separate data entries to our database, increasing our number of non-minimal entries to minimal entries from 486 to 930 (Figure 1).

Streamlining research into the central dogma of specialised metabolism

With 1540 NRPS and modular Type I PKS BGCs in MIBiG 3.0, modular BGCs constitute a substantial part of our database. Modular systems are characterised by enzyme complexes comprising repeating domain architectures, which collectively assemble a natural product scaffold. When the substrate selectivities of the recognition domains are known (acyl-transferase (AT) domains for PKS and A-domains for NRPS), these consistent architectures make it possible to predict the structure of these scaffolds with reasonable accuracy. The majority of AT domains in PKS systems recognise one of two substrates, malonyl-CoA or methylmalonyl-CoA, and excellent bioinformatics tools exist to distinguish between the two. However, for A-domains in NRPS systems, which recognise over 500 known substrates, substrate prediction is a greater challenge, which will require substantially more data to obtain models of comparably predictive power. Therefore, we decided to make the annotation of the substrate selectivity of NRPS A-domains a major focus of MIBiG 3.0. MIBiG 3.0 now contains annotations for 2782 A-domains (compared to 755 annotations in MIBiG 2.0; Figure 1B), covering 274 unique substrates which are identified by stereochemically curated isomeric SMILES strings (Figure 2; Supplementary Table 2). This makes MIBiG the largest resource for A-domain substrate data, containing 3-4 times as many labelled data points as the training sets used for the A-domain selectivity predictors SANDPUMA (Chevrette et al. 2017) and NRPSPredictor2 (Röttig et al. 2011). We hope that eventually this

dataset will be leveraged to train an improved A-domain substrate predictor, which can in turn be integrated into tools like antiSMASH to improve NRP scaffold structure prediction.

Since version 2.0, we have added 918 compound structures to our database in SMILES format, increasing the number of BGCs with structural data from 1347 to 1769 (Figure 1). By pulling SMILES strings directly from cross-linked databases where possible, we avoid conflicts caused by versioning and SMILES formatting. Additionally, we linked 1002 additional compounds to 51 unique bioactivities, creating opportunities for computationally predicting compound bioactivity from structure. For a further 95 compounds, we were also able to annotate their molecular targets (Figure 1B).

By centering MIBiG 3.0 around the annotation of substrate building blocks, compound structures, and bioactivities, we aspired to streamline future research into all aspects of sequence-structure-function relationships that lie at the heart of natural product research. All data can be easily downloaded and parsed in bulk from our database in JSON and GenBank format, or accessed on an entry-by-entry basis through our searchable online repository (Figure 3). As such, we hope that MIBiG 3.0 will prove an important resource for future machine learning endeavours that aim to decode the central dogma of specialised metabolism.

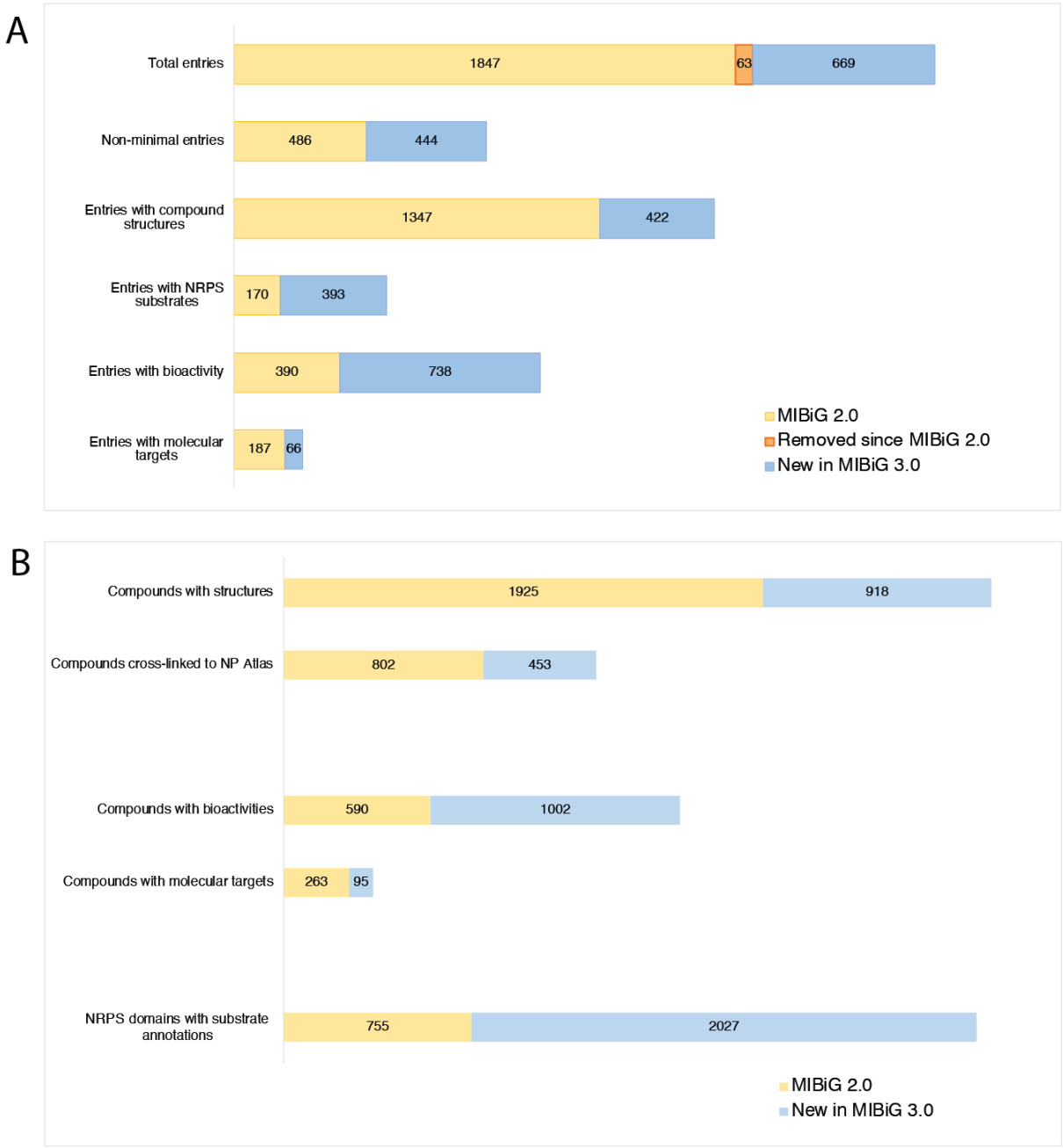

Figure 1: Overview of MIBiG 3.0. A. Added, removed, and updated entries since MIBiG 2.0. B. Improvements in the annotation of compounds, bioactivities, molecular targets and NRPS domain substrates.

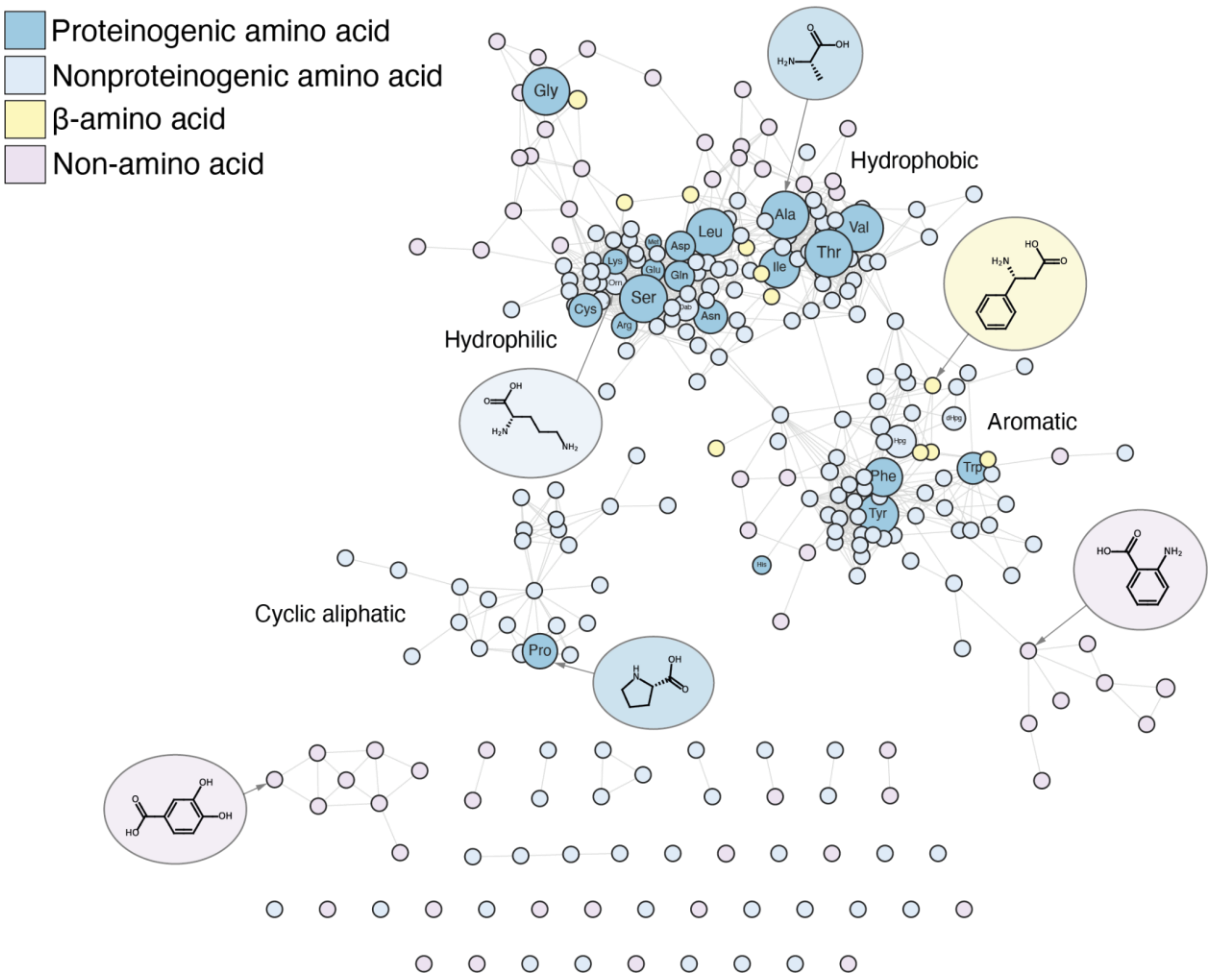

Figure 2. Similarity network of annotated NRPS substrates. Each node represents one of 275 unique NRPS substrate structures in MIBiG 3.0. Colours indicate substrate categories, and node size correlates with the number of annotations for that substrate in the MIBiG database. Substrates were clustered based on Tanimoto similarity (edge cut-off=0.46).

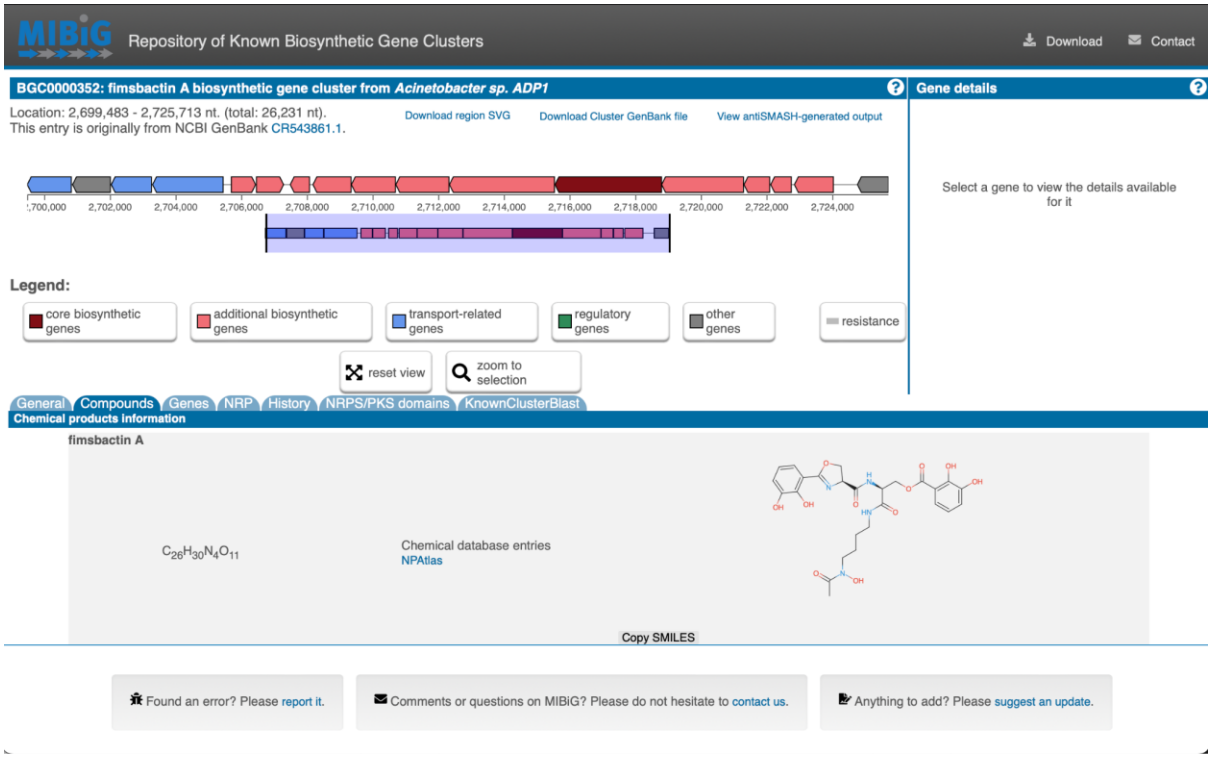

Figure 3. Example of a MIBiG overview page. Users can navigate to different tabs to access data such as compound structure, domain architecture and NRPS substrates.

Data availability

The MIBiG Repository is available at https://mibig.secondarymetabolites.org/. There is no access restriction for academic or commercial use of the repository and its data. The source code components, JSON-formatted data standard, and SQL schema for the MIBiG Repository are available on GitHub (https://github.com/mibig-secmet) under an OSI-approved Open Source licence.

References

Agrawal, Priyesh, Shradha Khater, Money Gupta, Neetu Sain, and Debasisa Mohanty. 2017. “RiPPMiner: A Bioinformatics Resource for Deciphering Chemical Structures of RiPPs Based on Prediction of Cleavage and Cross-Links.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkx408.

Blin, Kai, Simon Shaw, Alexander M. Kloosterman, Zach Charlop-Powers, Gilles P. van Wezel, Marnix H. Medema, and Tilmann Weber. 2021.

"antiSMASH 6.0: Improving Cluster Detection and Comparison Capabilities." *Nucleic Acids Research* 49 (W1): W29–35.
Carroll, Laura M., Martin Larralde, Jonas Simon Fleck, Ruby Ponnudurai, Alessio Milanese, Elisa Cappio, and Georg Zeller. 2021. "Accurate *de Novo* Identification of Biosynthetic Gene Clusters with GECCO." https://doi.org/10.1101/2021.05.03.442509.
Chevrette, Marc G., Fabian Aicheler, Oliver Kohlbacher, Cameron R. Currie, and Marnix H. Medema. 2017. "SANDPUMA: Ensemble Predictions of Nonribosomal Peptide Chemistry Reveal Biosynthetic Diversity across Actinobacteria." *Bioinformatics* 33 (20): 3202–10.
Gaulton, Anna, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, et al. 2016. "The ChEMBL Database in 2017." *Nucleic Acids Research* 45 (D1): D945–54.
Gavriilidou, Athina, Satria A. Kautsar, Nestor Zaburannyi, Daniel Krug, Rolf Müller, Marnix H. Medema, and Nadine Ziemert. 2022. "Compendium of Specialized Metabolite Biosynthetic Diversity Encoded in Bacterial Genomes." *Nature Microbiology* 7 (5): 726–35.
Hannigan, Geoffrey D., David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, et al. 2019. "A Deep Learning Genome-Mining Strategy for Biosynthetic Gene Cluster Prediction." *Nucleic Acids Research* 47 (18): e110–e110.
Izoré, Thierry, Y. T. Candace Ho, Joe A. Kaczmarski, Athina Gavriilidou, Ka Ho Chow, David L. Steer, Robert J. A. Goode, et al. 2021. "Structures of a Non-Ribosomal Peptide Synthetase Condensation Domain Suggest the Basis of Substrate Selectivity." *Nature Communications* 12 (1): 2511.
Kautsar, Satria A., Kai Blin, Simon Shaw, Jorge C. Navarro-Muñoz, Barbara R. Terlouw, Justin J. J. van der Hooft, Jeffrey A. van Santen, et al. 2020. "MIBiG 2.0: A Repository for Biosynthetic Gene Clusters of Known Function." *Nucleic Acids Research* 48 (D1): D454–58.
Kautsar, Satria A., Hernando G. Suarez Duran, Kai Blin, Anne Osbourn, and Marnix H. Medema. 2017. "plantiSMASH: Automated Identification, Annotation and Expression Analysis of Plant Biosynthetic Gene Clusters." *Nucleic Acids Research* 45 (W1): W55–63.
Medema, Marnix H., Anja Greule, Andreas Bechthold, and Frank Oliver Glöckner. 2015. *Minimum Information about a Biosynthetic Gene Cluster*.
Nayfach, Stephen. 2020. *A Genomic Catalog of Earth's Microbiomes*.
Paoli, Lucas, Hans-Joachim Ruscheweyh, Clarissa C. Forneris, Florian Hubrich, Satria Kautsar, Agneya Bhushan, Alessandro Lotti, et al. 2022. "Biosynthetic Potential of the Global Ocean Microbiome." *Nature* 607 (7917): 111–18.
Röttig, Marc, Marnix H. Medema, Kai Blin, Tilmann Weber, Christian Rausch, and Oliver Kohlbacher. 2011. "NRPSpredictor2--a Web Server for Predicting NRPS Adenylation Domain Specificity." *Nucleic Acids Research* 39 (Web Server issue): W362–67.
Rutz, Adriano, Maria Sorokina, Jakub Galgonek, Daniel Mietchen, Egon Willighagen, Arnaud Gaudry, James G. Graham, et al. 2022. "The LOTUS

Initiative for Open Knowledge Management in Natural Products Research." *eLife* 11 (May). https://doi.org/10.7554/eLife.70780.
Santen, Jeffrey A. van, Ella F. Poynton, Dasha Iskakova, Emily McMann, Tyler A. Alsup, Trevor N. Clark, Claire H. Fergusson, et al. 2022. "The Natural Products Atlas 2.0: A Database of Microbially-Derived Natural Products." *Nucleic Acids Research* 50 (D1): D1317–23.
Skinnider, Michael A., Chad W. Johnston, Mathusan Gunabalasingam, Nishanth J. Merwin, Agata M. Kieliszek, Robyn J. MacLellan, Haoxin Li, et al. 2020. "Comprehensive Prediction of Secondary Metabolite Structure and Biological Activity from Microbial Genome Sequences." *Nature Communications* 11 (1): 6058.
Terlouw, Barbara R., Sophie P. J. M. Vromans, and Marnix H. Medema. 2022. "PIKAChU: A Python-Based Informatics Kit for Analysing Chemical Units." *Journal of Cheminformatics* 14 (1): 34.
Walker, Allison S., and Jon Clardy. 2021. "A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters." *Journal of Chemical Information and Modeling* 61 (6): 2560–71.

Supplementary Table 1. Bioactivity annotations in MIBiG 2.0 and MIBiG 3.0.

Bioactivity	Count MIBiG 2.0	Count MIBiG 3.0
antibacterial	262	843
cytotoxic	176	549
inhibitor	124	175
antifungal	80	309
signalling	21	30
neurotoxic	17	17
siderophore	15	74
surfactant	10	50
antimalarial	7	11
antioxidant	7	11
antiviral	7	65
antiproliferative	4	5
immunosuppressant	3	9
insecticidal	3	11
sunscreen	3	3
anti-tubulin	2	2
antitumor	2	2
dna-interfering	2	2

induction of apoptosis	2	2
odorous metabolite	2	4
phytotoxic	2	10
pigment	2	6
swarming/biofilm induction	2	3
agonist of CD1D-restricted natural killer T cells	1	1
anti-algal	1	1
anticancer agent	1	1
anticoccidial	1	1
antiparasitic	1	6
antiplasmodial	1	3
dermatotoxic	1	1
fluorescence	1	1
hemolytic	1	1
hepatotoxin	1	0
herbicidal	1	5
inhibits neutrophil recruitment, immunosuppressive	1	1
iron reducing	1	1
modulator of apoptosis	1	1
neuroprotectant	1	3
p-glycoprotein-mediated multiple drug resistance	1	1
sodium channel blocking	1	1
uv protection	1	2
virulence	1	6
adhesion	0	1
anti-hiv	0	4
anti-infective	0	1
antihelmentic	0	1
antihelmintic	0	11
antineoplastic	0	2
antiparasidal	0	2
antiprotozoal	0	37
arginine metabolism	0	1
beta-lactamse induction	0	1
biofilm formation	0	4
cancer cell migration inhibitor	0	1
carbon storage	0	1
cell envelope	0	1

cell protectant	0	1
cell wall	0	1
cellular envelope	0	1
cold stress	0	2
cyst formation	0	1
denitrification	0	1
differentiation	0	1
emulsifier	0	3
enterotoxic	0	1
exopolysaccharide	0	3
extracelluar capsule	0	2
extracellular	0	1
hepatotoxic	0	1
immunomodulatory	0	1
ionophore	0	3
membrane	0	4
morphogenesis	0	2
nitrogen reduction	0	1
osmolytic	0	1
plant growth regulation	0	1
predation	0	1
protection against o2 penetration	0	1
protective layer	0	1
radical scavenging	0	1
regulatory	0	2
toxic	0	3

Supplementary Table 2. SMILES strings of NRPS substrates

Substrate name	SMILES string
(2S,3R)-2-amino-3-hydroxy-4-(4-nitrophenyl)butanoic acid	C1=CC(=CC=C1C[C@H]([C@@H](C(=O)O)N)O)[N+](=O)[O-]
(2S,6R)-diamino-(5R,7)-dihydroxy-heptanoic acid	C(C[C@@H](C(=O)O)N)[C@H]([C@@H](CO)N)O
(4S)-5,5,5-trichloroleucine	CCC(=O)CCCCC[C@@H](C(=O)O)N
(E)-4-methylhex-2-enoic acid	CCC(C)/C=C/C(=O)O
(S,E)-2-amino-4-decenoic acid	CCCCC/C=C/C[C@@H](C(=O)O)N

1-(1,1-dimethylallyl)-tryptophan	CC(C)(C=C)N1C=C(C2=CC=CC=C21)C[C@@H](C(=O)O)N
1-aminocyclopropane-1-carboxylic acid	C(O)(=O)C1(CC1)(N)
1-pyrroline-5-carboxylic acid	O=C(O)C1/N=C\CC1
2-(1-methylcyclopropyl)-D-glycine	CC1(CC1)[C@H](C(=O)O)N
2-amino-3,5-dimethyl-4-hexenoic Acid	CC(C=C(C)C)C(C(=O)O)N
2-amino-6-hydroxy-4-methyl-8-oxodecanoic acid	CCC(=O)CC(CC(C)CC(C(=O)O)N)O
2-aminoadipic acid	C(C[C@@H](C(=O)O)N)CC(=O)O
2-aminobutyric acid	CC[C@@H](C(=O)O)N
2-aminoisobutyric acid	O=C(O)C(N)(C)C
2-carboxy-6-hydroxyoctahydroindole	N1[C@H](C(=O)O)C[C@@H]2CC[C@@H](O)C[C@H]12
2-chloro-3,5-dihydroxy-4-methylphenylglycine	CC1=C(O)C(Cl)=C(C=C(O)1)[C@@H](C(=O)O)N
2-chlorobenzoic acid	C1=CC=C(C(=C1)C(=O)O)Cl
2-chlorophenylalanine	C1=CC=C(C(=C1)C[C@@H](C(=O)O)N)Cl
2-fluorophenylalanine	C1=CC=C(C(=C1)C[C@@H](C(=O)O)N)F
2-hexenoic acid	CCCC=CC(=O)O
2-hydroxy-4-methylpentanedioic acid	CC(C)CC(C(=O)O)O
2-hydroxy-4-methylpentanoic acid	CC(C)CC(C(=O)O)O
2-hydroxypent-4-enoic acid	C=CCC(C(=O)O)O
2-ketobutyric acid	CCC(=O)C(=O)O
2-ketoisocaproic acid	O=C(C(=O)O)CC(C)C
2-ketoisovaleric acid	O=C(C(=O)O)C(C)C
2-methylserine	C[C@](CO)(C(=O)O)N
2-octenoic acid	CCCCCC=CC(=O)O
2,3-diamino-3-methylpropanoic acid	NC(C)[C@@H](C(=O)O)N
2,3-diaminopropionic acid	C([C@@H](C(=O)O)N)N

2,3-dihydroxy-para-aminobenzoic acid	C1=CC(=C(C(=C(N)1)O)O)C(=O)O
2,3-dihydroxybenzoic acid	C1=CC(=C(C(=C1)O)O)C(=O)O
2,3-dihydroxyhexadecanoic acid	CCCCCCCCCCCCCC(C(C(=O)O)O)O
2,4-diaminobutyric acid	C(CN)[C@@H](C(=O)O)N
2,4-dihydroxypentanoic acid	CC(CC(C(=O)O)O)O
2,5-dihydroxy-4-methylpentanoic acid	CC(CC(C(=O)O)O)CO
2R-hydroxy-3-methylpentanoic acid	CCC(C)[C@H](C(=O)O)O
2R-hydroxyisovaleric acid	CC(C)[C@H](C(=O)O)O
2S-amino decanoic acid	CCCCCCCC[C@H](N)C(=O)O
2S-amino-4-hexenoic acid	C/C=C/CC(C(=O)O)N
2S-amino-8-oxodecanoic acid	CCC(=O)CCCCC[C@@H](C(=O)O)N
2S-amino-8S-hydroxydecanoic acid	CC[C@@H](CCCCC[C@@H](C(=O)O)N)O
2S-amino-9,10-epoxy-8-oxodecanoic acid	C1C(O1)C(=O)CCCCC[C@@H](C(=O)O)N
2S-amino-decanoic acid	CCCCCCCC[C@@H](C(=O)O)N
2S-amino-dodecanoic acid	CCCCCCCCCC[C@@H](C(=O)O)N
2S-amino-octanoic-acid	CCCCCC[C@@H](C(=O)O)N
2S-aminodecanoic acid	CCCCCCCC[C@@H](C(=O)O)N
2S-hydroxyisocaproic acid	CC(C)C[C@@H](C(=O)O)O
2S-hydroxyisovaleric acid	CC(C)[C@@H](C(=O)O)O
2S-methyl-3-oxobutyrine	CC(=O)[C@](C)(N)C(=O)O
2S,3S-diaminobutyric acid	C[C@@H]([C@@H](C(=O)O)N)N
3-(2-nitrocyclopropylalanine)	C1[C@H]([C@@H]1[N+](=O)[O-])C[C@@H](C(=O)O)N
3-(3-pyridyl)-alanine	C1=CC(=CN=C1)C[C@@H](C(=O)O)N
3-amino-6-hydroxy-2-piperidone	C1CC(NC(=O)C1N)O
3-aminobenzoic acid	C1=CC(=CC(=C1)N)C(=O)O

3-aminoisobutyric acid	CC(CN)C(=O)O
3-chloro-N-methylalanine	C([C@@H](C(=O)O)NC)Cl
3-chlorophenylalanine	C1=CC(=CC(=C1)Cl)C[C@@H](C(=O)O)N
3-chlorotyrosine	C1=C(Cl)C(=CC=C1C[C@@H](C(=O)O)N)O
3-fluorophenylalanine	C1=CC(=CC(=C1)F)C[C@@H](C(=O)O)N
3-hydroxy-4-methoxyphenylalanine	OC1=C(C=CC(=C1)C[C@@H](C(=O)O)N)OC
3-hydroxy-4-methylproline	CC1C(O)[C@H](NC1)C(=O)O
3-hydroxy-O-methyl-5-methyltyrosine	C1=C(O)C(=C(C)C=C1C[C@@H](C(=O)O)N)OC
3-hydroxy-O-methyltyrosine	C1=C(O)C(=CC=C1C[C@@H](C(=O)O)N)OC
3-hydroxy-para-aminobenzoic acid	C1=CC(=C(C=C1C(=O)O)O)N
3-hydroxyasparagine	N[C@H](C(O)=O)C(O)C(N)=O
3-hydroxyaspartic acid	N[C@@H](C(C(=O)O)O)(C(=O)O)
3-hydroxybenzoic acid	C1=CC(=CC(=C1)O)C(=O)O
3-hydroxyglutamine	C(C([C@@H](C(=O)O)N)O)C(=O)N
3-hydroxykynurenine	C1=CC(=C(C(=C1)O)N)C(=O)C[C@@H](C(=O)O)N
3-hydroxyleucine	CC(C)C([C@@H](C(=O)O)N)O
3-hydroxypicolinic acid	C1=CC(=C(N=C1)C(=O)O)O
3-hydroxyquinaldic acid	c1ccc2c(c1)cc(c(n2)C(=O)O)O
3-hydroxytyrosine	C1=CC(=C(C=C1C[C@@H](C(=O)O)N)O)O
3-hydroxyvaline	CC(O)(C)[C@@H](C(=O)O)N
3-methoxyanthranilic acid	COC1=CC=CC(=C1N)C(=O)O
3-methoxyaspartic acid	N[C@H](C(C(=O)O)OC)(C(=O)O)
3-methylasparagine	CC([C@@H](C(=O)O)N)C(=O)N
3-methylaspartic acid	CC([C@@H](C(=O)O)N)C(=O)O
3-methylleucine	CC(C)C(C)[C@@H](C(=O)O)N

3-nitrotyrosine	C1=CC(=C(C=C1C[C@@H](C(=O)O)N)[N+](=O)[O-])O
3,4-dihydroxybenzoic acid	C1=CC(=C(C=C1C(=O)O)O)O
3,5-dichloro-4-hydroxyphenylglycine	C1=C(Cl)C(=C(Cl)C=C1[C@@H](C(=O)O)N)O
3,5-dihydroxyphenylglycine	N[C@H](C(=O)O)c1cc(O)cc(O)c1
3R-amino-2S,5R-dihydroxy-8-phenyloctanoic acid	N[C@H](C[C@H](O)CCCc1ccccc1)[C@H](O)C(O)=O
3R-aminoisobutyric acid	C[C@H](CN)C(=O)O
3R-chloroproline	C1[C@@H](Cl)[C@H](NC1)C(=O)O
3R-hydroxy-2,4-diaminobutyric acid	NC[C@@H](O)[C@@H](C(=O)O)N
3R-hydroxyasparagine	N[C@H](C(O)=O)[C@@H](O)C(N)=O
3R-hydroxyaspartic acid	N[C@@H]([C@H](C(=O)O)O)(C(=O)O)
3R-hydroxyhomotyrosine	C1=CC(=CC=C1C[C@H]([C@@H](C(=O)O)N)O)O
3R-hydroxyleucine	CC(C)[C@H]([C@@H](C(=O)O)N)O
3R-methyl-D-aspartic acid branched	N[C@H]([C@@H](O)C(O)=O)C(O)=O
3R-methylbeta-alanine	NC[C@@H](C)C(=O)O
3R-methylglutamic acid	C[C@H](CC(=O)O)[C@@H](C(=O)O)N
3S-carboxypiperazine	C1NN[C@H](C(=O)O)CC1
3S-cyclohex-2-enylalanine	C1C=C[C@H](CC1)C[C@@H](C(=O)O)N
3S-hydroxy-4S-methylproline	C[C@@H]1[C@H](O)[C@H](NC1)C(=O)O
3S-hydroxy-6-chlorohistidine	C1=C(NC(Cl)=N1)[C@H]([C@@H](C(=O)O)N)O
3S-hydroxyasparagine	N[C@H](C(O)=O)[C@H](O)C(N)=O
3S-hydroxyleucine	CC(C)[C@@H]([C@@H](C(=O)O)N)O
3S-hydroxypipecolic acid	C1C[C@@H]([C@H](NC1)C(=O)O)O
3S-hydroxyproline	C1CN[C@@H]([C@H]1O)C(=O)O
3S-methyl-D-aspartic acid branched	N[C@H]([C@H](O)C(O)=O)C(O)=O
3S-methylaspartic acid	C[C@@H]([C@@H](C(=O)O)N)C(=O)O

3S-methylproline	C1[C@H](C)[C@H](NC1)C(=O)O
3S,4R-dichloroproline	Cl[C@H]1[C@@H](Cl)[C@H](NC1)C(=O)O
3S,4S-dihydroxyhomotyrosine	C1=CC(=CC=C1[C@H](O)[C@H]([C@@H](C(=O)O)N)O)O
4-acetamidopyrrole-2-carboxylic acid	CC(=O)NC1=CNC(=C1)C(=O)O
4-amino-2-hydroxy-3-isopropoxybenzoic acid	CC(C)OC1=C(C=CC(=C1O)C(=O)O)N
4-aminobutyric acid	NCCCC(=O)O
4-aminophenylalanine	C1=CC(=CC=C1C[C@@H](C(=O)O)N)N
4-bromophenylalanine	C1=CC(=CC=C1C[C@@H](C(=O)O)N)Br
4-chlorobenzoic acid	C1=CC(=CC=C1C(=O)O)Cl
4-chlorophenylalanine	C1=CC(=CC=C1C[C@@H](C(=O)O)N)Cl
4-chlorothreonine	ClC[C@H]([C@@H](C(=O)O)N)O
4-fluorophenylalanine	C1=CC(=CC=C1C[C@@H](C(=O)O)N)F
4-hydroxy-D-kynurenine	C1=C(O)C=C(C(=C1)C(=O)C[C@H](C(=O)O)N)N
4-hydroxyglutamine	C(C(O)C(=O)N)[C@@H](C(=O)O)N
4-hydroxyindole-3-carboxylic acid	c1cc2c(c(c1)O)c(c[nH]2)C(=O)O
4-hydroxyphenylglycine	C1=CC(=CC=C1[C@@H](C(=O)O)N)O
4-hydroxyphenylpyruvic acid	C1=CC(=CC=C1CC(=O)C(=O)O)O
4-hydroxyproline	C1[C@H](NCC1O)C(=O)O
4-methoxytryptophan	C1=CC=C2C(=C1OC)C(=CN2)C[C@@H](C(=O)O)N
4-methylphenylalanine	CC1=CC=C(C=C1)C[C@@H](C(=O)O)N
4-methylproline	CC1C[C@H](NC1)C(=O)O
4-nitrotryptophan	C1=CC=C2C(=C1[N+](=O)[O-])C(=CN2)C[C@@H](C(=O)O)N
4-oxoproline	C1[C@H](NCC1=O)C(=O)O
4,5-dehydroarginine	O=C(O)[C@@H](N)C/C=C/NC(N)=N
4,5-dehydroleucine	C=C(C)C[C@@H](C(=O)O)N

4,5-dihydroxyornithine	C([C@@H](C(=O)O)N)C(C(N)O)O
4R-butenyl-4R-methylthreonine	C[C@H](C/C=C/C)[C@@H](O)[C@H](N)C(=O)O
4R-hydroxyproline	C1[C@H](NC[C@@H]1O)C(=O)O
4R-methylproline	C[C@@H]1C[C@H](NC1)C(=O)O
4R-propylproline	CCC[C@@H]1C[C@H](NC1)C(=O)O
4S-acetyl-5S-methylproline	CC(=O)O[C@H]1C[C@H](N[C@H](C)1)C(=O)O
4S-hydroxylysine	NCC[C@H](O)C[C@@H](C(=O)O)N
4S-methylazetidine-2S-carboxylic acid	C[C@H]1C[C@H](N1)C(=O)O
4S-methylproline	C[C@H]1C[C@H](NC1)C(=O)O
4S-propenylproline	C/C=C\[C@H]1C[C@H](NC1)C(=O)O
5-chlorotryptophan	C1=CC2=C(C=C1Cl)C(=CN2)C[C@@H](C(=O)O)N
5-methoxytyrosine	C1=C(OC)C(=CC=C1C[C@@H](C(=O)O)N)O
5,5-dimethylpipecolic acid	C1C(C)(C)CN[C@@H](C1)C(=O)O
5R-hydroxy-2S-aminopentanoic acid	O[C@@H](C[C@@H](C(=O)O)N)CO
6-chloro-4-hydroxy-1-methyl-indole-3-carboxylic acid	C(O)1=C(Cl)C=C2C(=C1)C(=CN(C)2)C(=O)O
6-chlorotryptophan	C1=C(Cl)C=C2C(=C1)C(=CN2)C[C@@H](C(=O)O)N
6,7-dichlorotryptophan	C1=C(Cl)C(Cl)=C2C(=C1)C(=CN2)C[C@@H](C(=O)O)N
alanine	C[C@@H](C(=O)O)N
alaninol	C[C@@H](CO)N
allo-isoleucine	CC[C@@H](C)[C@@H](C(=O)O)N
allo-threonine	C[C@@H]([C@@H](C(=O)O)N)O
alpha-methylphenylalanine	C[C@](CC1=CC=CC=C1)(C(=O)O)N
anthranilic acid	C1=CC=C(C(=C1)C(=O)O)N
arginine	C(C[C@@H](C(=O)O)N)CN=C(N)N
asparagine	C([C@@H](C(=O)O)N)C(=O)N

aspartic acid	C([C@@H](C(=O)O)N)C(=O)O
aspartic acid branched	C([C@@H](C(=O)O)N)C(=O)O
azetidine-2-carboxylic acid	O=C(O)[C@H]1NCC1
benzoic acid	C1=CC=C(C=C1)C(=O)O
benzoxazolinate	c1ccc2c(c1)nc(o2)C(=O)O
beta-alanine	NCCC(=O)O
beta-hydroxy-3-hydroxy-O-methyl-5-methyltyrosine	C1=C(C)C(=C(O)C=C1C(O)[C@@H](C(=O)O)N)OC
beta-hydroxyarginine	C(C(O)[C@@H](C(=O)O)N)CN=C(N)N
beta-hydroxyphenylalanine	OC(C1=CC=CC=C1)[C@@H](C(=O)O)N
beta-hydroxytyrosine	C1=CC(=CC=C1C([C@@H](C(=O)O)N)O)O
beta-lysine	C(C[C@@H](CC(=O)O)N)CN
betaine	C[N+](C)(C)CC(=O)O
butyric acid	CCCC(=O)O
capreomycidine	C1CN=C(N[C@H]1[C@@H](C(=O)O)N)N
cinnamic acid	C1=CC=C(C=C1)/C=C/C(=O)O
citrulline	C(C[C@@H](C(=O)O)N)CNC(=O)N
coumaric acid	C1=CC(=CC=C1/C=C/C(=O)O)O
cysteic acid	C([C@@H](C(=O)O)N)S(=O)(=O)O
cysteine	C([C@@H](C(=O)O)N)S
D-2-aminobutyric acid	CC[C@H](C(=O)O)N
D-2,3-diaminopropionic acid	C([C@H](C(=O)O)N)N
D-alanine	C[C@H](C(=O)O)N
D-allo-threonine	C[C@H]([C@H](C(=O)O)N)O
D-arginine	C(C[C@H](C(=O)O)N)CN=C(N)N
D-aspartic acid branched	C([C@H](C(=O)O)N)C(=O)O

D-glutamic acid branched	C(CC(=O)O)[C@H](C(=O)O)N
D-glutamine	C(CC(=O)N)[C@H](C(=O)O)N
D-leucine	CC(C)C[C@H](C(=O)O)N
D-phenylalanine	C1=CC=C(C=C1)C[C@H](C(=O)O)N
D-phenyllactic acid	C1=CC=C(C=C1)C[C@H](C(=O)O)O
D-pipecolic acid	C1CCN[C@H](C1)C(=O)O
D-tyrosine	C1=CC(=CC=C1C[C@H](C(=O)O)N)O
dehydroarginine	C(CN=C(N)N)/C=C(/C(=O)O)\N
dehydrolysine	C(CCN)=C[C@@H](C(=O)O)N
dehydrotryptophan	C1=CC=C2C(=C1)C(=CN2)/C=C(/C(=O)O)\N
dehydrovaline	CC(=C(C(=O)O)N)C
E-butenyl-4R-methylthreonine	C/C=C/C[C@@H](C)[C@H]([C@@H](C(=O)O)N)O
enduracididine	C1[C@H](NC(=N1)N)C[C@@H](C(=O)O)N
fatty acid	*CC(=O)O
glutamic acid	C(CC(=O)O)[C@@H](C(=O)O)N
glutamine	C(CC(=O)N)[C@@H](C(=O)O)N
glycine	NCC(=O)O
glycocyamine	C(C(=O)O)N=C(N)N
glycolic acid	C(C(=O)O)O
graminine	O=NN(O)CCC[C@H](N)(C(=O)O
guanidinoacetic acid	C(C(=O)O)N=C(N)N
hexanoic acid	CCCCCC(=O)O
histidine	C1=C(NC=N1)C[C@@H](C(=O)O)N
homoleucine	CC(C)C[C@@H](CC(=O)O)N
homophenylalanine	C1=CC=C(C=C1)CC[C@@H](C(=O)O)N

homoserine	C(CO)[C@@H](C(=O)O)N
homotyrosine	C1=CC(=CC=C1CC[C@@H](C(=O)O)N)O
hydrocinnamic acid	C1=CC=C(C=C1)CCC(=O)O
hydroxyproline	C(*)1C[C@H](NC(*)1)C(=O)O
indole-3-pyruvic acid	C1=CC=C2C(=C1)C(=CN2)CC(=O)C(=O)O
isoleucine	CC[C@H](C)[C@@H](C(=O)O)N
isovaleric acid	CC(C)CC(=O)O
kynurenine	C1=CC=C(C(=C1)C(=O)C[C@@H](C(=O)O)N)N
lactic acid	C[C@@H](C(=O)O)O
leucine	CC(C)C[C@@H](C(=O)O)N
linoleic acid	CCCCC/C=C\C/C=C\CCCCCCCC(=O)O
lysine	C(CCN)C[C@@H](C(=O)O)N
malic acid	C(C(C(=O)O)O)C(=O)O
meta-tyrosine	C1=CC(=CC(=C1)O)C[C@@H](C(=O)O)N
methionine	CSCC[C@@H](C(=O)O)N
N-(1-methyl)-tryptophan	C1=CC=C2C(=C1)C(=CN(C)2)C[C@@H](C(=O)O)N
N-(1-propargyl)-tryptophan	C1=CC=C2C(=C1)C(=CN(CC#C)2)C[C@@H](C(=O)O)N
N-dimethylglycine	CN(C)CC(=O)O
N-formylalanine	C[C@@H](C(=O)O)NC=O
N-formylglycine	C(C(=O)O)NC=O
N-hydroxyvaline	CC(C)[C@@H](C(=O)O)NO
N-methylserine	CN[C@@H](CO)C(=O)O
N1-methoxytryptophan	C1=CC=C2C(=C1)C(=CN(OC)2)C[C@@H](C(=O)O)N
N5-acetyl-N5-hydroxyornithine	CC(=O)N(CCC[C@@H](C(=O)O)N)O
N5-formyl-N5-hydroxyornithine	C(C[C@@H](C(=O)O)N)CN(C=O)O

N5-hydroxyornithine	C(C[C@@H](C(=O)O)N)CNO
N5-nitroso-N5-hydroxyornithine	O=NN(CCC[C@@H](C(=O)O)N)O
N5-trans-anhydromevalonyl-N5-hydroxyornithine	C(C[C@@H](C(=O)O)N)CN(O)C(=O)/C=C(C)/CCO
N6-hydroxylysine	C(CCNO)C[C@@H](C(=O)O)N
nicotinic acid	C1=CC(=CN=C1)C(=O)O
norcoronamic acid	C[C@H]1C[C@]1(C(=O)O)N
O-methylthreonine	C[C@H]([C@@H](C(=O)O)N)OC
O-methyltyrosine	COC1=CC=C(C=C1)C[C@@H](C(=O)O)N
octanoic acid	CCCCCCCC(=O)O
ornithine	C(C[C@@H](C(=O)O)N)CN
p-hydroxybenzoylformic acid	C1=CC(=CC=C1C(=O)C(=O)O)O
para-aminobenzoic acid	O=C(O)c1ccc(N)cc1
pentanoic acid	CCCCC(=O)O
phenazine-1-carboxylic acid	C1=CC=C2C(=C1)N=C3C=CC=C(C3=N2)C(=O)O
phenazine-1,6-dicarboxylic acid	C1=CC(=C2C(=C1)N=C3C(=N2)C=CC=C3C(=O)O)C(=O)O
phenylalanine	C1=CC=C(C=C1)C[C@@H](C(=O)O)N
phenylglycine	C1=CC=C(C=C1)[C@@H](C(=O)O)N
phenylpyruvic acid	C1=CC=C(C=C1)CC(=O)C(=O)O
pipecolic acid	C1CCN[C@@H](C1)C(=O)O
piperazic acid	C1C[C@H](NNC1)C(=O)O
proline	C1C[C@H](NC1)C(=O)O
propionic acid	CCC(=O)O
pyrrole-2-carboxylic acid	C1=CNC(=C1)C(=O)O
pyruvic acid	CC(=O)C(=O)O
quinoxaline-2-carboxylic acid	C1=CC=C2C(=C1)N=CC(=N2)C(=O)O

R-aza-beta-tyrosine	C1=CC(=NC=C1O)[C@@H](CC(=O)O)N
R-beta-hydroxyphenylalanine	O[C@H](C1=CC=CC=C1)[C@@H](C(=O)O)N
R-beta-hydroxytyrosine	C1=CC(=CC=C1[C@H]([C@@H](C(=O)O)N)O)O
R-beta-methylphenylalanine	C[C@H](C1=CC=CC=C1)[C@@H](C(=O)O)N
R-beta-methyltryptophan	C[C@H](C1=CNC2=CC=CC=C21)[C@@H](C(=O)O)N
R-beta-phenylalanine	C1=CC=C(C=C1)[C@@H](CC(=O)O)N
R-beta-tyrosine	C1=CC(=CC=C1[C@@H](CC(=O)O)N)O
S-adenosylmethionine	C[S+](CC[C@@H](C(=O)[O-])N)C[C@@H]1[C@H]([C@H]([C@@H](O1)N2C=NC3=C(N=CN=C32)N)O)O
S-beta-hydroxycyclohex-2S-enylalanine	C1C=C[C@H](CC1)[C@H](O)[C@@H](C(=O)O)N
S-beta-hydroxyenduracididine	C1[C@H](NC(=N1)N)[C@H](O)[C@@H](C(=O)O)N
S-beta-hydroxyphenylalanine	O[C@@H](C1=CC=CC=C1)[C@@H](C(=O)O)N
S-beta-tyrosine	C1=CC(=CC=C1[C@H](CC(=O)O)N)O
salicylic acid	C1=CC=C(C(=C1)C(=O)O)O
serine	C([C@@H](C(=O)O)N)O
threonine	C[C@H]([C@@H](C(=O)O)N)O
trans-2-crotylglycine	C/C=C/C[C@@H](C(=O)O)N
tryptophan	C1=CC=C2C(=C1)C(=CN2)C[C@@H](C(=O)O)N
tyrosine	C1=CC(=CC=C1C[C@@H](C(=O)O)N)O
valine	CC(C)[C@@H](C(=O)O)N