

MITE: the Minimum Information about a Tailoring Enzyme database for capturing specialized metabolite biosynthesis

Adriano Rutz¹, Daniel Probst², César Aguilar³, Daniel Y. Akiyama⁴, Fabrizio Alberti⁵, Hannah E. Augustijn^{2,6}, Nicole E. Avalon^{7,8}, Christine Beemelmans^{9,10}, Hellen Bertoletti Barbieri⁴, Friederike Biermann^{2,11,12}, Alan J. Bridge¹³, Esteban Charria Girón^{2,14}, Russell Cox¹⁵, Max Crüsemann^{16,17}, Paul M. D'Agostino⁹, Marc Feuermann¹³, Jennifer Gerke¹⁸, Karina Gutiérrez García^{19,20}, Jonathan E. Holme²¹, Ji-Yeon Hwang²², Riccardo Iacovelli²³, Júlio César Jeronimo Barbosa⁴, Navneet Kaur²⁴, Martin Klapper²⁵, Anna M. Köhler¹⁵, Aleksandra Korenskaia²⁶, Noel Kubach²⁶, Byung T. Lee²⁷, Catarina Loureiro², Shrikant Mantri^{24,28}, Simran Narula²⁴, David Meijer², Jorge C. Navarro-Muñoz², Giang-Son Nguyen²¹, Sunaina Paliyal²⁴, Mohit Panghal^{24,28}, Latika Rao²⁴, Simon Sieber²⁹, Nika Sokolova³⁰, Sven T. Sowa³¹, Judit Szenei³², Barbara R. Terlouw², Heiner G. Weddeling³¹, Jingwei Yu³³, Nadine Ziemert^{26,34}, Tilmann Weber³², Kai Blin³², Justin J.J. van der Hooff^{2,35}, Marnix H. Medema^{2,*}, Mitja M. Zdouc^{2,*}

¹Institute for Molecular Systems Biology, ETH Zürich, Otto-Stern-Weg 3, Zürich 8093, Switzerland

²Bioinformatics Group, Wageningen University & Research, Droevendaalsesteeg 1, Wageningen 6708PB, The Netherlands

³Industrial Genomics Laboratory, Centro de Biotecnología FEMSA, Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Nuevo Leon 64700, Mexico

⁴Department of Organic Chemistry, Institute of Chemistry, University of Campinas (UNICAMP), Rua Monteiro Lobato 270, Campinas, São Paulo 13.083-862, Brazil

⁵School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, United Kingdom

⁶Institute of Biology, Leiden University, Sylviusweg 72, Leiden 2333BE, The Netherlands

⁷Department of Pharmaceutical Sciences, University of California, 856 Health Sciences Road, Irvine, CA 92697, United States

⁸Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0212, United States

⁹Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI), Campus E8.1, Saarbrücken 66123, Germany

¹⁰Department Anti-infectives from Microbiota, Saarland University, Campus E8.1, Saarbrücken 66123, Germany

¹¹Institute for Molecular Biosciences, Goethe University Frankfurt, Max-von-Laue-Straße 9, Frankfurt am Main 60438, Germany

¹²LOEWE Center for Translational Biodiversity Genomics (TBG), Senckenberganlage 25, Frankfurt am Main 60325, Germany

¹³SIB Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland

¹⁴Department of Microbial Drugs, Helmholtz Centre for Infection Research (HZI), Inhoffenstraße 7, Braunschweig 38124, Germany

¹⁵Institute for Organic Chemistry and BMWZ, Leibniz Universität Hannover, Schneiderberg 38, Hannover 30167, Germany

¹⁶Institute of Pharmaceutical Biology, University of Bonn, Nussallee 6, Bonn 53115, Germany

¹⁷Institute for Pharmaceutical Biology, Goethe University Frankfurt, Max-von-Laue-Straße 9, Frankfurt am Main 60438, Germany

¹⁸Institute for Organic Chemistry, Leibniz Universität Hannover, Schneiderberg 38, Hannover 30167, Germany

¹⁹Biosphere Sciences and Engineering Division, Carnegie Institution for Science, 3520 San Martin Dr, Baltimore, MD 21218, United States

²⁰Department of Ecology and Evolutionary Biology, University of Arizona, 1041 E. Lowell St., Tucson, AZ 85721, United States

²¹Department of Biotechnology and Nanomedicine, SINTEF Industry, P.O. Box 4760 Torgard, Trondheim N-7465, Norway

²²Molecular Targets Program, Center for Cancer Research, National Cancer Institute, Frederick, MD 21702-1201, United States

²³Production Host Engineering Team, VTT Technical Research Centre of Finland Ltd, Maarintie 3, Espoo 02150, Finland

²⁴Computational Biology Lab, National Agri-Food and Biomanufacturing Institute (NABI), Sector 81, S.A.S. Nagar, Mohali, Punjab 140306, India

²⁵Department of Paleobiotechnology, Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute, Beutenbergstraße 11A, Jena 07745, Germany

²⁶Translational Genome Mining for Natural Products, Interfaculty Institute of Microbiology and Infection Medicine Tübingen (IMIT), Interfaculty Institute for Biomedical Informatics (IBMI), University of Tübingen, Auf der Morgenstelle 24, Tübingen 72076, Germany

²⁷Institute of Applied Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

²⁸Regional Centre for Biotechnology, NCR Biotech Science Cluster, 3rd Milestone, Faridabad–Gurugram Expressway, Faridabad, Haryana (NCR Delhi) 121001, India

²⁹Department of Chemistry, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland

³⁰Department of Chemical and Pharmaceutical Biology, University of Groningen, Antonius Deusinglaan 1, Groningen 9713AV, The Netherlands

Received: August 13, 2025. Revised: August 29, 2025. Accepted: September 2, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

³¹Department of Pharmaceutical Sciences, University of Basel, Klingelbergstrasse 50, Basel 4056, Switzerland

³²The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Building 220, Søtofts Plads, Kongens Lyngby 2800, Denmark

³³Institute of Plant and Food Science, Department of Biology, School of Life Sciences, Southern University of Science and Technology, 1088 Xueyuan Avenue, Shenzhen 518055, P.R. China

³⁴German Center for Infection Research (DZIF), Partner Site Tübingen, 72076 Tübingen, Germany

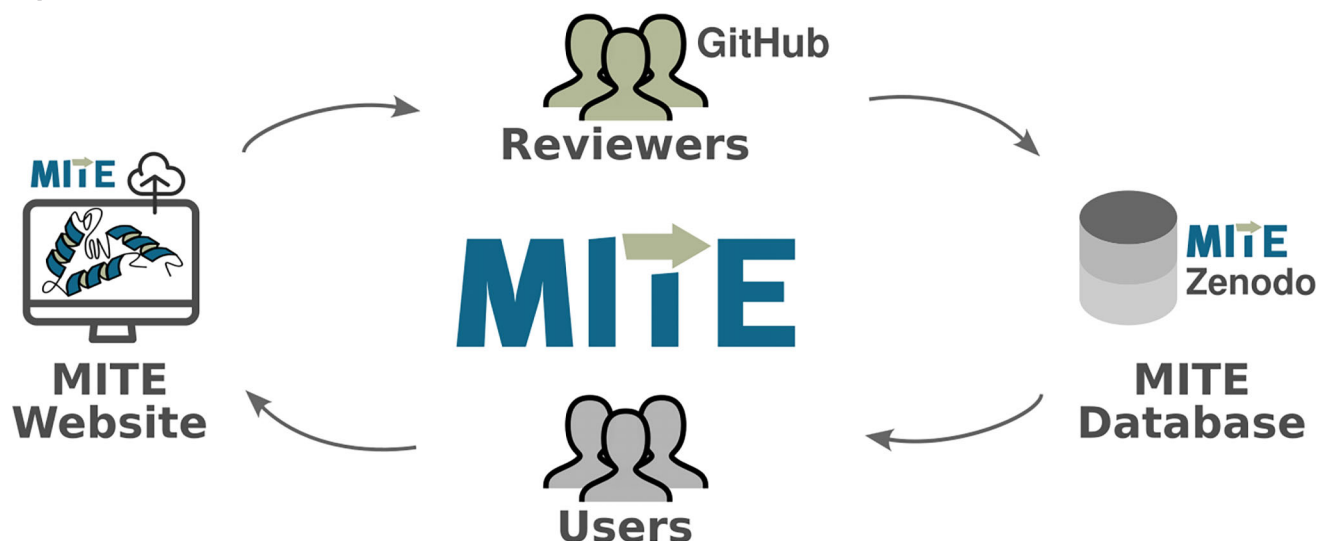
³⁵Department of Biochemistry, University of Johannesburg, C2 Lab Building 224, Kingsway Campus, Cnr University & Kingsway Road, Auckland Park, Johannesburg 2006, South Africa

*To whom correspondence should be addressed. Email: marnix.medema@wur.nl
Correspondence may also be addressed to Mitja M. Zdouc. Email: mitja.zdouc@wur.nl

Abstract

Secondary or specialized metabolites show extraordinary structural diversity and potent biological activities relevant for clinical and industrial applications. The biosynthesis of these metabolites usually starts with the assembly of a core 'scaffold', which is subsequently modified by tailoring enzymes to define the molecule's final structure and, in turn, its biological activity profile. Knowledge about reaction and substrate specificity of tailoring enzymes is essential for understanding and computationally predicting metabolite biosynthesis, but this information is usually scattered in the literature. Here, we present MITE, the Minimum Information about a Tailoring Enzyme database. MITE employs a comprehensive set of parameters to annotate tailoring enzymes, defining substrate and reaction specificity by the expressive reaction SMARTS (Simplified Molecular Input Line Entry System Arbitrary Target Specification) chemical pattern language. Both human and machine readable, MITE can be used as a knowledge base, for *in silico* biosynthesis, or to train machine-learning applications, and tightly integrates with existing resources. Designed as a community-driven and open resource, MITE employs a rolling release model of data curation and expert review. MITE is freely accessible at <https://mite.bioinformatics.nl/>.

Graphical abstract



Introduction

Many organisms can produce small molecules with intricate chemical structures and potent biological activities, known as specialized or secondary metabolites (SMs). Besides their widespread therapeutic and industrial application [1], SMs have high environmental relevance and are generally considered to grant an evolutionary advantage to the producing organism [2]. In microorganisms, their biosynthesis is often genetically organized in biosynthetic gene clusters (BGCs), sets of physically clustered genes encoding enzymes, transporters, and regulators that are collectively responsible for the controlled production of SMs [3]. Following the definitions of Walsh [4], the enzymes directly involved in the biosynthesis can be generally separated into gatekeeper and tailoring enzymes. Firstly, gatekeeper enzymes (also known as core enzymes) redirect primary metabolism building blocks into SM biosynthesis, forming the scaffold of the molecules [4]. Next,

tailoring enzymes modify the nascent metabolite, eventually leading to its mature structure. These chemical modifications are often essential for bioactivity and target affinity [5, 6] and may have additional effects, such as enhanced chemical stability or solubility [4]. Therefore, detailed information about the reaction specificity and substrate promiscuity of tailoring enzymes, which can vary considerably even for closely related enzymes, is essential to understand and direct SM biosynthesis. Furthermore, it constitutes the foundation of algorithms for the prediction of chemical structures and biological activities directly from genomic data [7, 8], the investigation of quantitative structure–activity relationships, and the engineering of biosynthetic pathways [9].

However, details on tailoring enzymes, their reactions, and substrate specificities are usually deposited in narrative scientific articles, hampering computational access. A few databases exist that contain information on tailoring enzymes, but they are limited in their level of detail. While the Mini-

Information about a Biosynthetic Gene Cluster (MIBiG) database provides information on BGCs including tailoring enzymes [10], their functional characterization is more focused on the scaffold-forming enzymes (e.g. non-ribosomal peptide synthetases), and tailoring enzyme descriptions are limited to generic terms. Some of the reactions contained in the Rhea database are associated with tailoring enzymes [11], but Rhea's focus on stoichiometrically balanced and generalized reactions using a defined set of reactants is not always applicable to partially characterized and highly specific SM pathways, in which the precise order of reactions and therefore the precise substrate and product per reaction is often difficult to resolve. Databases such as ExPasy ENZYME [12], KEGG [13], or BRENDA [14] also characterize tailoring enzymes, but are limited by the rigidity of the used EC (Enzyme Commission) number reaction classification system [15]. Hence, there is a need for a resource that provides flexible, comprehensive descriptions of tailoring enzymes and their substrate and reaction specificities linked to their genomic context, without relying on predefined reactants.

Here, we present the Minimum Information about a Tailoring Enzyme (MITE) database, dedicated to the characterization of SM-acting tailoring enzymes. MITE summarizes experimental data on the reactions and substrate specificities of these tailoring enzymes in both human- and machine-readable forms, using the reaction SMARTS notation (<https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>). This well-established chemical transformation language allows encoding enzymatic substrate recognition and modification as atom-bond patterns, allowing for flexible and concise description of substrate and reaction specificities. MITE is permissive towards unknown intermediates and ambiguous reaction order, and provides genomic context by specifying co-acting enzymes. MITE tightly integrates with established resources such as UniProt [16], NCBI GenBank [17], MIBiG [10], and Rhea [11], and can be used as a knowledge base, as a reference database for enzyme annotations (already in use by the genome mining tool antiSMASH [18]), and as a dataset for machine learning. With applications in pathway annotation, phylogenetic analyses, synthetic biology, metabolic engineering, and drug discovery, we expect MITE to be a highly beneficial database for both computational scientists and experimentalists alike. MITE is freely available at <https://mite.bioinformatics.nl/>.

MITE database outline and infrastructure

Data collection and database content

As an expert-curated database, MITE exclusively contains information on tailoring enzymes sourced from primary literature. Enzymes are within the scope of MITE if they directly partake in SM biosynthesis but are not 'core' or 'gatekeeper' enzymes (for a detailed discussion, see Walsh [4]). Therefore, transporter or resistance-conferring enzymes are excluded, as are multi-domain synthases and other classical scaffold-forming enzymes. Inspired by the MIBiG database [10], the MITE data model specifies a compact set of mandatory and optional parameters (Fig. 1A). Briefly, each MITE entry represents a single tailoring enzyme encoded by a single gene (one entry per protein isoform, if applicable) and is assigned a permanent identifier for consistent referencing. Protein sequences are referenced from either UniProt [16] or NCBI Gen-

Bank [17], and optional database cross-links to MIBiG [10] or Wikidata [19] can be specified. Additional 'auxiliary' enzymes may be referenced if required for the proper function of the described tailoring enzyme (e.g. macrolactam formation by McjC in the biosynthetic pathway of microcin J25 requires the presence of protease McjB [5]). Each MITE entry must also contain at least one reaction description in the form of a reaction SMARTS as an abstract representation of the substrate specificity and the reaction of the enzyme. This atom-bond chemical transformation pattern may represent discrete molecules or chemical substructures (e.g. the chlorination of an indole functional group or a tryptophan-containing peptide; see [Supplementary Fig. S1](#) and [Supplementary Table S1](#)). Since reaction SMARTS may also represent non-viable substructures and contain wildcard characters or Boolean logic, they must be accompanied by an example reaction of discrete reactants in SMILES format, used to validate the reaction SMARTS. This reaction example is flexible and may also represent an R-group attached to a (partially known or unknown) core molecule (the latter represented as an asterisk), to be permissive to pathways where the exact order of reactions is not fully clear (e.g. [Supplementary Fig. S1](#), see reaction example 1c). Reactions may be cross-referenced with a Rhea [11] reaction identifier or an EC number, and automation is in place to check for missing identifiers. Each entry must include at least one digital object identifier to primary literature (including preprints).

In its current version (1.16 [20]), the MITE database contains 202 'active' entries, amounting to a total of 2639 data points (defined as key-value pairs, excluding metadata specified in Fig. 1A), including 283 reaction SMARTS representations and 291 example reactions. Entries show a wide taxonomic distribution (Fig. 1B), with most enzymes being associated with phyla Actinomycetota (66%) and Dikarya (19%). While 89% of MITE entries can be cross-referenced with a BGC from the MIBiG database, only 15% of enzymes are also covered by Rhea, and 7% could not be cross-referenced to either of the two databases ([Supplementary Fig. S2](#)). To functionally characterize the MITE dataset, we first investigated the sequence diversity of the covered enzymes. A comparison against the NCBI non-redundant protein database at a 70% sequence similarity cut-off (analogous to antiSMASH's cut-off for comparison against the MITE database [18]) matched 77 461 protein sequences (Fig. 1C; see [Supplementary Table S2](#)). A sequence similarity network (SSN) of MITE entries ([Supplementary Fig. S3](#)) showed the formation of a few subnetworks, with the largest annotated as cytochrome P450, radical S-adenosyl-L-methionine enzymes, and flavin-dependent halogenases, although 46% of enzymes were 'orphans' with no next neighbours. When the SSN was annotated with tailoring function terms, sequence-related enzymes typically shared the same functional label, except for cytochrome P450 enzymes, which are known for their reaction diversity [21]. To get further insight into the represented biosynthetic reaction space, we sampled an example reaction from each MITE entry, generated bit fingerprints using the differential reaction fingerprint DRFP [22], clustered fingerprints using a *k*-nearest neighbour algorithm, visualized them using TMAP [23], and annotated them with tailoring reaction terms ([Supplementary Fig. S4](#)). The resulting graph (Fig. 1D) represents the diversity of newly generated substructures [i.e. the substructure difference between reactant(s) and product(s)], with next neighbours showing reaction similarity.

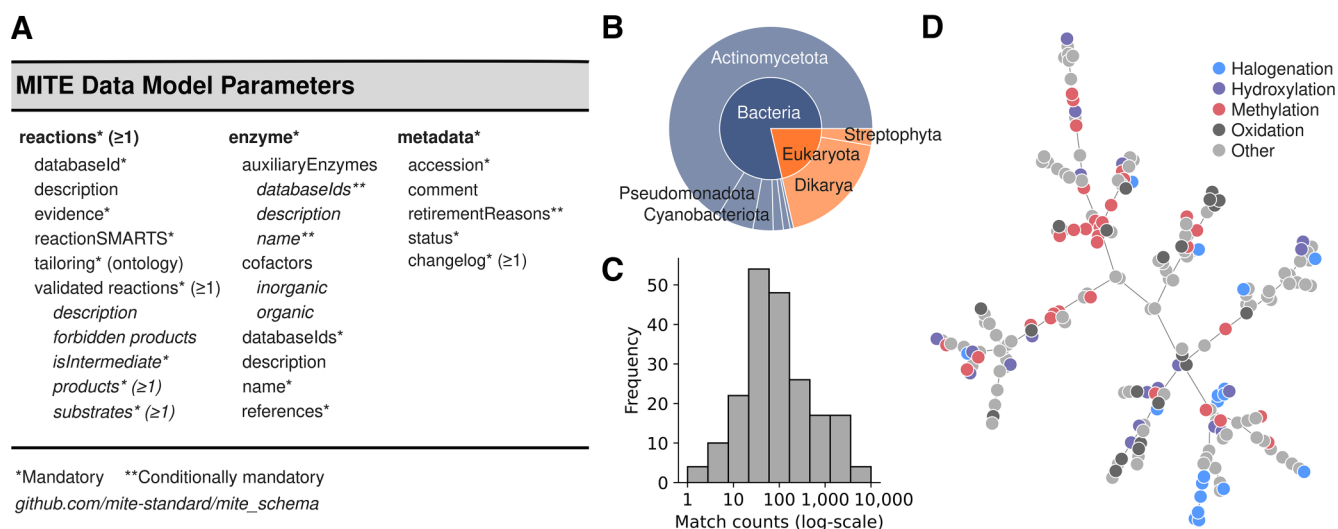


Figure 1. MITE data standard and database content. Summary of the MITE data model in panel (A), indicating mandatory and optional parameters. Panel (B) shows a sunburst plot of MITE entries, for which taxonomic information could be retrieved from NCBI Entrez ($n = 187$), with kingdom (inner circle) and phylum (outer circle) information. Panel (C) shows a histogram of distribution of BLASTp matches of MITE entries ($n = 202$) against the NCBI non-redundant protein database (at 70% sequence similarity cut-off, maximum 5000 matches per entry). Panel (D) shows a TMAP plot, with nodes representing DRFP-encoded example reactions from MITE entries ($n = 202$), annotated with tailoring function labels.

While some aggregations of nodes indicating reaction similarity could be observed (e.g. halogenations), their clustering was not as distinct as in the SSN, with a wider dispersion of nodes indicating a diverse reaction space. Manual inspection of halogenation-labelled entries (Supplementary Fig. S5) confirmed reaction similarity of next neighbours exemplified by tryptophan halogenases, even though their substrate specificity could still vary considerably (e.g. MibH requires a full peptidic substrate, while PyrH acts on free tryptophan, even though they are both tryptophan 5-halogenases and thus next neighbours in the TMAP plot).

Web interface and infrastructure

The MITE database is available as a collection of JavaScript Object Notation (JSON) files following a bespoke JSON Schema (<https://json-schema.org/>) data model. The JSON files are stored on Zenodo (<https://zenodo.org/>) [20], serving as the single source of truth. To facilitate interaction with the data, we developed the MITE web server implemented as a Flask (<https://flask.palletsprojects.com/>) web application, running within a Docker container (<https://www.docker.com/>). On startup, the web server populates the displayed data by downloading the most recent release of the MITE dataset from its Zenodo record. A non-persistent PostgreSQL (<https://www.postgresql.org/>) database is implemented to allow for complex search queries. The front end is implemented using Bootstrap (<https://getbootstrap.com/>), and the MITE website can be also viewed using mobile or tablet devices.

Central to the web server are the entry pages, which display the tailoring enzyme information in a dashboard-like format (Fig. 2A). The overview page allows for browsing entries and a range of search operations discussed below. All data can be downloaded as flat (text) files, and an application programming interface (API) following OpenAPI specifications (<https://swagger.io/>) is available for computational interactions. Besides data display and querying, the web server allows for the submission of new entries or the modification of existing ones, using integrated infrastructure (Fig. 2B). Online documenta-

tion and (video) tutorials are available to facilitate adoption by the community. Each contribution undergoes extensive automated validation (e.g. verifying the reaction SMARTS with the example reaction) and creates a new pull request on the GitHub page managing the MITE dataset, where it is reviewed by one of the expert reviewers. All contributions are automatically released to the public domain under the Creative Commons ‘No Rights Reserved’ license (<https://creativecommons.org/public-domain/cc0/>), which encourages and facilitates its reuse in other resources. The MITE database is updated regularly and new releases are automatically archived on the Zenodo data repository. From there, the MITE web server and interoperable tools retrieve the MITE entries in JSON format. The MITE database and accompanying infrastructure are governed by the MITE Standard GitHub organization (<https://github.com/mite-standard>), which also contains a forum for discussions and news items (e.g. web server downtime schedule). To ensure sustainability of the MITE database and minimize maintenance burden, social and technical workflows are in place, following O3 guidelines [24]. As a community-driven project, MITE welcomes participation in the form of data contributions, review, and governance participation. All contributors who have made a significant contribution [~ 6 h of time investment, as specified in the MITE governance documentation (<https://github.com/mite-standard/github/blob/main/GOVERNANCE.md>)] qualify for co-authorship.

Applications of the MITE database

The MITE web server comes with a variety of search operations to query and subset data. Simple searches can be performed by querying the interactive overview table, while more complex queries can be constructed with the query builder, using Boolean logic. Queries can also be combined with substructure and reaction search using SMILES, SMARTS, or reaction SMARTS, and BLASTp searches for protein sequence similarity. For instance, a user could query entries annotated with ‘Methylation’ as tailoring reaction term and restrict the search to peptidic substrates using the SMARTS expression

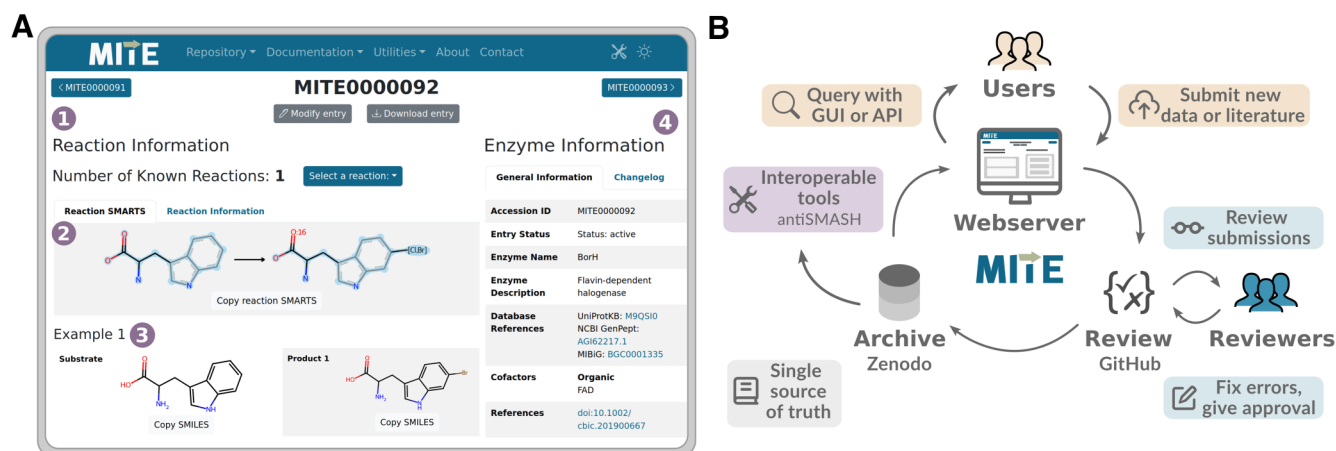


Figure 2. MITE web interface and infrastructure. Panel (A) shows an example entry page with (i) navigation bar, (ii) abstract substrate and reaction specificity visualized from reaction SMARTS, (iii) concrete example of substrate–product pair, and (iv) general enzyme information including changelog. Panel (B) shows a schema of MITE’s infrastructure, with circular data flow. Users can query data using the graphical user interface web server or API. Users can also create new or modify existing entries, which are automatically validated and undergo expert review. Approved entries are archived as releases on Zenodo, which acts as the single source of truth. From Zenodo, data are retrieved by the MITE web server or interoperable tools.

‘[C](=[O]-[N-C-C](=[O])’ (Supplementary Fig. S6). As of 12 August 2025, using MITE data version 1.16 [20], this results in seven entries, which can be then browsed, further subsetted, or downloaded in tabular format for downstream analysis.

Another use case for the MITE data repository is for *in silico* simulation of SM biosynthesis. Since each MITE entry has one or more associated reaction SMARTS, these can be rapidly applied to substrate SMILES as ‘reaction rules’ with ‘built-in’ substrate recognition, acting as an abstraction of the enzymatic reaction. Combining multiple MITE entries thus allows to approximate biosynthetic pathways. As a proof of concept, we investigated the biosynthetic pathway of bottromycin A2, a thoroughly experimentally characterized metabolite modified by no fewer than 10 tailoring enzymes [25]. Starting from the peptidic precursor BotA (MIBiG cluster BGC0000469), biosynthesis can be simulated stepwise in an automated fashion, using MITE reaction SMARTS retrieved via the API (Fig. 3). As expected, the resulting final product matched the structure of the literature-reported bottromycin A2 [25]. To facilitate such explorative *in silico* biosynthesis, we provide a reaction planner on the MITE website (Supplementary Fig. S7).

Discussion and future directions

The MITE database significantly advances SM research by providing a dedicated resource for the detailed characterization of tailoring enzymes. Representing reaction and substrate specificity in the reaction SMARTS format allows for compact descriptions of both discrete compounds as well as atom–bond patterns, exceeding the capabilities of traditional reactant-based representations of biological reactions. Chemical transformations can be expressed abstractly, accommodating arbitrary levels of substrate promiscuity and specificity. This includes the description of partial or full structures, the use of wildcard characters for atoms and bonds, and Boolean logic. The resulting expressions are easily searchable, compatible with a variety of cheminformatics and synthetic biology tools, and suited for machine-learning-based applications. Reaction SMARTS can be created and modified by various chemistry drawing programs [e.g. Ketcher (<https://lifecycle.opensource.epam.com/ketcher/>), Marvin (<https://chemaxon.com/marvin/>)] and invite gradual refinement over time, allowing to incorporate new knowledge on tailoring enzyme reaction specificity as it becomes available. MITE is designed following FAIR (findable, accessible, interoperable, reusable) data principles and facilitates access to machine-readable information on tailoring enzymes substrate promiscuity, previously only available through laborious manual search of scientific articles. MITE emphasized accessibility and interoperability by its use of standardized data formats and data exchange protocols, and is registered on Bioregistry [26] and FAIRsharing [27] platforms. MITE is already used as a reference database by version 8 of the popular genome mining tool antiSMASH [18].

Among existing databases, MITE shows the greatest similarity with the Rhea database. This resource also annotates tailoring enzymes, but it is mainly a biological reaction database, focusing on the representation of standardized, balanced ‘master’ reactions with clearly defined substrates and products intended for pathway mapping and metabolic modelling. In contrast, MITE is an enzyme database, allowing higher flexibility in representing catalytic functions. MITE is permissive to partially resolved structures without ChEBI identifiers [28], unknown co-factors and intermediates, and unclear reaction order characteristic of poorly investigated SM biosynthetic pathways. The enzyme coverage of the MITE database is clearly distinct from Rhea, and only 15% of enzymes characterized in MITE can be annotated with a Rhea entry. Despite the different scope of the resources, we believe that MITE and Rhea are actually highly complementary: for well-characterized tailoring enzymes, MITE can reference master reactions from Rhea, while MITE can provide chemical information on enzymes that are currently difficult to annotate by the standards of Rhea. MITE entries may be also suitable to serve as a resource for new Rhea entries, and annotation efforts from both resources are planned to reinforce each other, creating a virtuous circle of data enhancement. MITE also integrates with other resources, especially the BGC repository MIBiG, which uses MITE as a resource to annotate tailoring enzyme-encoding genes. Additional cross-links can be established by providing an EC number (e.g. <https://lifecycle.opensource.epam.com/ketcher/>), Marvin (<https://chemaxon.com/marvin/>) and invite gradual refinement over time, allowing to incorporate new knowledge on tailoring enzyme reaction specificity as it becomes available. MITE is designed following FAIR (findable, accessible, interoperable, reusable) data principles and facilitates access to machine-readable information on tailoring enzymes substrate promiscuity, previously only available through laborious manual search of scientific articles. MITE emphasized accessibility and interoperability by its use of standardized data formats and data exchange protocols, and is registered on Bioregistry [26] and FAIRsharing [27] platforms. MITE is already used as a reference database by version 8 of the popular genome mining tool antiSMASH [18].

Among existing databases, MITE shows the greatest similarity with the Rhea database. This resource also annotates tailoring enzymes, but it is mainly a biological reaction database, focusing on the representation of standardized, balanced ‘master’ reactions with clearly defined substrates and products intended for pathway mapping and metabolic modelling. In contrast, MITE is an enzyme database, allowing higher flexibility in representing catalytic functions. MITE is permissive to partially resolved structures without ChEBI identifiers [28], unknown co-factors and intermediates, and unclear reaction order characteristic of poorly investigated SM biosynthetic pathways. The enzyme coverage of the MITE database is clearly distinct from Rhea, and only 15% of enzymes characterized in MITE can be annotated with a Rhea entry. Despite the different scope of the resources, we believe that MITE and Rhea are actually highly complementary: for well-characterized tailoring enzymes, MITE can reference master reactions from Rhea, while MITE can provide chemical information on enzymes that are currently difficult to annotate by the standards of Rhea. MITE entries may be also suitable to serve as a resource for new Rhea entries, and annotation efforts from both resources are planned to reinforce each other, creating a virtuous circle of data enhancement. MITE also integrates with other resources, especially the BGC repository MIBiG, which uses MITE as a resource to annotate tailoring enzyme-encoding genes. Additional cross-links can be established by providing an EC number (e.g. <https://lifecycle.opensource.epam.com/ketcher/>), Marvin (<https://chemaxon.com/marvin/>) and invite gradual refinement over time, allowing to incorporate new knowledge on tailoring enzyme reaction specificity as it becomes available. MITE is designed following FAIR (findable, accessible, interoperable, reusable) data principles and facilitates access to machine-readable information on tailoring enzymes substrate promiscuity, previously only available through laborious manual search of scientific articles. MITE emphasized accessibility and interoperability by its use of standardized data formats and data exchange protocols, and is registered on Bioregistry [26] and FAIRsharing [27] platforms. MITE is already used as a reference database by version 8 of the popular genome mining tool antiSMASH [18].

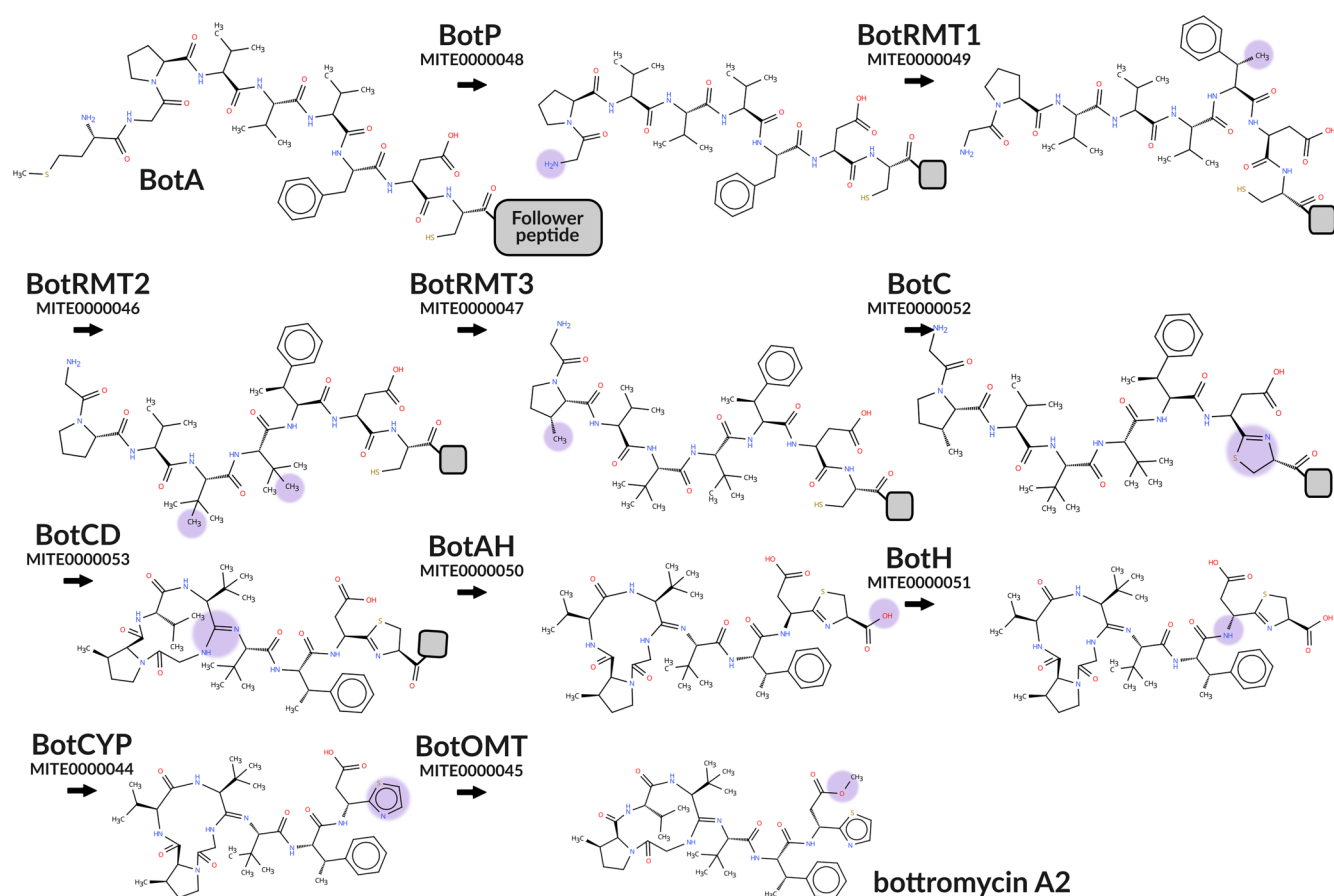


Figure 3. *In silico* biosynthesis of bottromycin A2. Enzymes involved in bottromycin A2 biosynthesis (MIBiG cluster BGC0000469) are represented by MITE entries. Starting from the precursor peptide substrate BotA, reaction SMARTS are applied consecutively, each including one or more tailoring reactions performed by the respective enzyme. Purple circles indicate the area of modification following the previous enzymatic biotransformation (e.g. epimerization by BotH). Eventually, this biosynthetic cascade results in the mature, literature-reported bottromycin A2 metabolite.

//bioregistry.io/mite:MITE0000029), or linking to a Wikidata enzyme entry, and automation during data submission is in place to support manual data curation with linking to external resources.

The MITE database also comes with some limitations, including the type of reaction SMARTS it accepts. While the Daylight company (<https://www.daylight.com/>) remains the central authority in defining reaction SMARTS, there is a proliferation of sometimes conflicting variants or ‘flavors’ employed by various tools, including proprietary extensions such as Chemaxon’s CXSMARTS. To provide a common ground compatible with a majority of downstream applications without potential licensing restrictions, MITE only accepts generic reaction SMARTS, and ‘standardizes’ them using the cheminformatics tool RDKit (<https://www.rdkit.org>). This validation takes place during community submission, and only reaction SMARTS passing this initial step are accepted, ensuring an interoperable and high-quality dataset.

Moving forward, we intend to further develop the coverage of MITE by participating in the upcoming MIBiG 5.0 hackathons and further encouraging community participation. Additionally, we will work on adding BGCs of the 7% of MITE entries that currently do not have an entry in MIBiG, further improving both resources. We will continue periodic cross-referencing with MIBiG, introduce automation that synchronizes with the Wikidata knowledge graph, and explore

semiautomated data exchange with the Rhea database. We intend to continue developing the data submission workflow and extend the capability of the MITE’s API. We will also proceed exploring ways to make the MITE database compatible and accessible to machine-learning tools as training and/or test dataset.

In conclusion, the MITE database is a novel resource to characterize SM-acting tailoring enzymes. Following publication, MITE is expected to rapidly expand in coverage through its community-driven curation and rolling release system and is designed to be a sustainable, open, and interoperable source of knowledge. The researchers represented by the present publication commit to submitting MITE-compliant data when publishing experimental results on tailoring enzymes, and we encourage the broader research community to join the initiative. With applications as a knowledge base, parts catalog for synthetic biology, and a resource for phylogenetic analyses, genome mining, and pathway annotation, we expect MITE to be a highly beneficial database to further enhance knowledge on the enzymology and chemistry connected to SMs. The MITE database is available at <https://mite.bioinformatics.nl/>.

Acknowledgements

We thank the anonymous MITE contributors for their data submissions.

Author contributions: Conceptualization (A.J.B., N.Z., T.W., K.B., J.J.J.vdH., M.H.M. [equal], M.M.Z. [equal]), Data curation (A.R., C.A., D.Y.A., F.A., N.E.A., C.B., H.B.B., F.B., E.C.G., M.C., P.M.D., M.F., J.G., K.G.G., J.E.H., J-Y.H., R.I., J.C.J., N.K., M.K., A.M.K., A.K., N.Ku., B.T.L., C.L., S.M., S.N., D.M., J.C.N-M., G-S.N., S.P., M.P., L.R., S.S., N.S., S.T.S., J.S., B.R.T., H.G.W., J.Y., M.M.Z. [lead]), Funding acquisition (J.J.J.vdH., M.H.M.), Methodology (A.R., D.P., A.J.B., R.C., K.B., M.H.M., M.M.Z. [lead]), Software (A.R., D.P., H.E.A., M.M.Z. [lead]), Supervision (J.J.J.vdH., M.H.M.), Writing—original draft (M.M.Z.), Writing—review & editing (A.R., D.P., N.E.A., C.B., A.J.B., E.C.G., M.C., R.I., J.C.N-M., B.R.T., T.W., J.J.J.vdH., M.H.M.).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

J.J.J.vdH. is a member of the Scientific Advisory Board of NAICONS Srl., Milano, Italy and consults for Corteva Agriscience, Indianapolis, IN, USA. M.H.M. is a member of the Scientific Advisory Boards of Hexagon Bio and Hothouse Therapeutics Ltd. All other authors declare to have no competing interests.

Funding

D.Y.A. was supported by the São Paulo Research Foundation (FAPESP) research scholarship (Grant 21/07038-0); F.A. was supported by the UKRI Future Leaders Fellowship (MR/V022334/1); N.E.A. was supported by the National Center for Complementary and Integrative Health of the NIH under Award number F32AT011475; C.B. was supported by the European Union Horizon 2020 research and innovation program (ERC Grant No. 802736, MORPHEUS) and by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) Project-ID 239748522—CRC 1127 (project A6); H.B.B. was supported by the São Paulo Research Foundation (FAPESP) research scholarship (Grant 2021/08947-3); M.C. was supported by the German Research Foundation (DFG) Grant No. 495740318; P.M.D. was supported by the Hans-Fischer-Gesellschaft, Germany; M.F. was supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation SERI; J.E.H. was supported by the SINTEF internal projects, POP-SEP BiocatDB (102 022 750), SEP AGREE (102 029 187), POS BIOINFO 2024 (102024676-14), and SIP IN SAMS—Systematic Workflows for AI in Chemistry and Materials Research (102 029 407) and by the European Union's Horizon 2020 research and innovation programme under Grant Agreement Nos. 101 000 392 (MARBLEs), 101 081 957 (BLUETOOLS), and 862 923 (AtlantECO); R.I. was supported by the Novo Nordisk Foundation, NNF23OC0086472; J.C.J. was supported by the São Paulo Research Foundation (FAPESP) research scholarship (Grant 2023/06874-4); N.K. was supported by the Department of Biotechnology (DBT), Government of India; M.K. was supported by the Werner Siemens Foundation; A.M.K. was supported by the DFG Research Unit 5170 CytoLabs—Systematic Investigation and Exploitation of Cytochalasans; A.K. was supported by the Horizon Europe Marie Skłodowska-Curie Grant Agreement No.

101 072 485; S.M. was supported by the Department of Biotechnology (DBT), Government of India and by the National Agri-Food and Biomanufacturing Institute (NABI); D.M. was supported by the European Research Council (ERC Starting Grant 948770-DECIPHER); G-S.N. was supported by the SINTEF internal projects, POP-SEP BiocatDB (102 022 750), SEP AGREE (102 029 187), POS BIOINFO 2024 (102024676-14), and SIP IN SAMS—Systematic Workflows for AI in Chemistry and Materials Research (102 029 407) and by the European Union's Horizon 2020 research and innovation programme under Grant Agreement Nos. 101 000 392 (MARBLEs), 101 081 957 (BLUETOOLS), and 862 923 (AtlantECO); M.P. was supported by the Department of Biotechnology (DBT), Government of India and by the University Grants Commission (UGC), Ministry of Education, Government of India; J.S. was supported by the European Union's Horizon Europe programme under the Marie Skłodowska-Curie Grant Agreement No. 101 072 485 (MAGic-MOLFUN); N.Z. was supported by the TTU09.716/German Center for Infection Research (DZIF) and by the European Union's Horizon Europe programme under the Marie Skłodowska-Curie Grant Agreement No. 101 072 485 (MAGic-MOLFUN); T.W. was supported by the Novo Nordisk Foundation (NNF20CC0035580), the Danish National Research Foundation CeMiSt, DNRF137, and by the European Union's Horizon Europe programme under the Marie Skłodowska-Curie Grant Agreement No. 101 072 485 (MAGic-MOLFUN); K.B. was supported by the Novo Nordisk Foundation, NNF20CC0035580; J.J.J.vdH. was supported by the Netherlands Organisation for Scientific Research (NWO) KIC Microbe Mission Grant KICH1LWV04.21.013 and by the European Union's Horizon Europe programme under the Marie Skłodowska-Curie Grant Agreement No. 101 072 485 (MAGic-MOLFUN); M.H.M. and M.M.Z. were supported by the Netherlands Organisation for Scientific Research (NWO) KIC Grant KICH1LWV04.21.013.

Data availability

The MITE database is released to the public domain under the Creative Commons 'No Rights Reserved' license (<https://creativecommons.org/public-domain/cc0/>) and available on <https://mite.bioinformatics.nl/> and Zenodo [20]. The source code used for the web interface [29], the data model [30], and data validation [31] is available at <https://github.com/mite-standard>, licensed under the MIT License (<https://opensource.org/licenses/mit>). Source code for the generation of figures for this manuscript is available at https://github.com/mite-standard/mite_ms and also licensed under the MIT License.

References

1. Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J Nat Prod* 2020;83:770–803. <https://doi.org/10.1021/acs.jnatprod.9b01285>
2. Traxler MF, Kolter R. Natural products in soil microbe interactions and evolution. *Nat Prod Rep* 2015;32:956–70. <https://doi.org/10.1039/C5NP00013K>
3. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: how natural selection drives chemical innovation. *Proc Natl Acad Sci USA* 2008;105:4601–8. <https://doi.org/10.1073/pnas.0709132105>

4. Walsh CT. Tailoring enzyme strategies and functional groups in biosynthetic pathways. *Nat Prod Rep* 2023;40:326–86. <https://doi.org/10.1039/D2NP00048B>
5. Yan K-P, Li Y, Zirah S et al. Dissecting the maturation steps of the lasso peptide microcin J25 *in vitro*. *ChemBioChem* 2012;13:1046–52. <https://doi.org/10.1002/cbic.201200016>
6. Nagarajan R. Structure–activity relationships of vancomycin-type glycopeptide antibiotics. *J Antibiot* 1993;46:1181–95. <https://doi.org/10.7164/antibiotics.46.1181>
7. Lee Y-Y, Guler M, Chigumba DN et al. HypoRiPPAtlas as an Atlas of hypothetical natural products for mass spectrometry database search. *Nat Commun* 2023;14:4219. <https://doi.org/10.1038/s41467-023-39905-4>
8. Yuan Y, Shi C, Zhao H. Machine learning-enabled genome mining and bioactivity prediction of natural products. *ACS Synth Biol* 2023;12:2650–62. <https://doi.org/10.1021/acssynbio.3c00234>
9. Hudson GA, Mitchell et al. RiPP antibiotics: biosynthesis and engineering potential. *Curr Opin Microbiol* 2018;45:61–9. <https://doi.org/10.1016/j.mib.2018.02.010>
10. Zdouc MM, Blin K, Louwen NLL et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res* 2025;53:D678–D690. <https://doi.org/10.1093/nar/gkae1115>
11. Bansal P, Morgat A, Axelsen KB et al. Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res* 2022;50:D693–D700. <https://doi.org/10.1093/nar/gkab1016>
12. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;28:304–5. <https://doi.org/10.1093/nar/28.1.304>
13. Kanehisa M, Furumichi M, Sato Y et al. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res* 2025;53:D672–D677. <https://doi.org/10.1093/nar/gkae909>
14. Chang A, Jeske L, Ulbrich S et al. The ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res* 2021;49:D498–D508. <https://doi.org/10.1093/nar/gkaa1025>
15. Dönertaş HM, Martínez Cuesta S, Rahman SA et al. Characterising complex enzyme reaction data. *PLoS One* 2016;11:e0147952.
16. Consortium UP. UniProt: the Universal Protein knowledgebase in 2025. *Nucleic Acids Res* 2025;53:D609–D617. <https://doi.org/10.1093/nar/gkae1010>
17. Sayers EW, Beck J, Bolton EE et al. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* 2025;53:D20–D29. <https://doi.org/10.1093/nar/gkae979>
18. Blin K, Shaw S, Vader L et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res* 2025;53:W32–W38. <https://doi.org/10.1093/nar/gkaf334>
19. Vrandečić D, Krötzsch MW. Wikidata: a Free Collaborative Knowledgebase. *Commun ACM* 2014;57:78–85.
20. Zdouc MM, Rutz A. MITE dataset. Version 1.16. Zenodo, <https://doi.org/10.5281/zenodo.16759748>, 8 August 2025.
21. Malhotra K, Franke J. Cytochrome P450 monooxygenase-mediated tailoring of triterpenoids and steroids in plants. *Beilstein J Org Chem* 2022;18:1289–310. <https://doi.org/10.3762/bjoc.18.135>
22. Probst D, Schwaller P, Reymond J-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit Discov* 2022;1:91–7. <https://doi.org/10.1039/D1DD00006C>
23. Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminform* 2020;12:12. <https://doi.org/10.1186/s13321-020-0416-x>
24. Hoyt CT, Gyori BM. The O3 guidelines: open data, open code, and open infrastructure for sustainable curated scientific resources. *Sci Data* 2024;11:547. <https://doi.org/10.1038/s41597-024-03406-w>
25. Franz L, Kazmaier U, Truman AW et al. Bottromycins—biosynthesis, synthesis and activity. *Nat Prod Rep* 2021;38:1659–83. <https://doi.org/10.1039/D0NP00097C>
26. Hoyt CT, Balk M, Callahan TJ et al. Unifying the identification of biomedical entities with the Bioregistry. *Sci Data* 2022;9:714. <https://doi.org/10.1038/s41597-022-01807-3>
27. Sansone S-A, McQuilton P, Rocca-Serra P et al. FAIRsharing as a community approach to standards, repositories and policies. *Nat Biotechnol* 2019;37:358–67. <https://doi.org/10.1038/s41587-019-0080-8>
28. Hastings J, Owen G, Dekker A et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;44:D1214–D1219. <https://doi.org/10.1093/nar/gkv1031>
29. Zdouc MM. MITE webserver. Zenodo. Version 1.5.0. <https://doi.org/10.5281/zenodo.16789434>, 10 August 2025.
30. Zdouc MM, Rutz A. MITE schema. Zenodo. Version 1.8.0, <https://doi.org/10.5281/zenodo.15681189>, 17 June 2025.
31. Zdouc MM, Rutz A. MITE extras. Zenodo. Version 1.5.1, <https://doi.org/10.5281/zenodo.16439200>, 26 July 2025.