MARTS-DB: A Database of Mechanisms And Reactions of Terpene Synthases

Martin Engst[1,2], Martin Brokeš[1,2], Tereza Čalounová[1], Adéla Tajovská[1], Milana Perković[1], Raman Samusevich[1], Ratthachat Chatpatanasiri[3], Tomáš Pluskal[1]

[1]Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo náměstí 2, Prague, Czech Republic

[2]Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technická 5, Prague, Czech Republic

[3]Czech Institute of Informatics, Robotics and Cybernetics (CIIRC), Czech Technical University, Jugoslávských partyzánů 1580/3, 160 00 Prague, Czech Republic

Abstract

Terpene synthases (TPS) are enzymes that catalyze some of the most complex reactions in nature, the cyclizations of terpenes, the carbon backbones to the largest group of natural products, terpenoids. On average, more than half of the carbon atoms in a terpene scaffold undergo a change in connectivity or configuration during these enzymatic cascades. Understanding TPS reaction mechanisms remains challenging, often requiring intricate computational modeling and isotopic labeling studies. Moreover, the relationship between TPS sequence and catalytic function is difficult to decipher, and data-driven approaches remain limited due to the lack of comprehensive, high-quality data sources. To address this gap, we introduce the **Mechanisms And Reactions of Terpene Synthases DataBase (MARTS-DB)**—a manually curated, structured, and searchable database that integrates TPS enzymes, the terpenes they produce, and their detailed reaction mechanisms. MARTS-DB includes over 2,600 reactions catalyzed by 1,334 annotated enzymes from across all domains of life. Where available, reaction mechanisms are mapped as stepwise cascades. Accessible at https://www.marts-db.org, the database provides advanced search functionality and supports full dataset download in machine-readable formats. It also encourages community contributions to promote continuous growth. By enabling systematic exploration of TPS catalysis, MARTS-DB opens new avenues for computational analysis and machine learning, as recently demonstrated in the prediction of novel terpene synthases.

Introduction

Terpene synthases (TPS) are enzymes that catalyze some of the most complex reactions in nature, the cyclizations of terpenes. Often volatile compounds, with diverse functions, terpenes are most notable for being the precursors to terpenoids, the largest class of natural products, with more than 100000 known compounds.[1,2] Their function in nature is as varied as their structure, ranging from specialized plant metabolites, to insect pheromones and steroids.[3–6] This remarkable chemical diversity originates from a relatively small pool of linear isoprenoid diphosphate precursors composed of connected five-carbon (C5) isoprene units and a diphosphate group.[1,7]

The biosynthesis of these precursors is carried out by isoprenyl diphosphate synthases (IDSs), enzymes that are structurally similar to TPSs.[1,7,8] TPS reactions are initiated by the formation of a reactive carbocation, either through diphosphate abstraction or alkene/epoxide protonation. The active site environment guides this carbocation through a sequence of transformations including cyclizations, hydride/methyl shifts, and proton transfers, concluding with deprotonation or hydroxylation.[2,9] While early hypotheses relied heavily on chemical intuition, recent experimental and computational techniques have elucidated complete mechanistic cascades. However, new mechanistic studies still rely on analogies to known reactions, and currently, no unified resource systematically maps these mechanisms.[10–13]

Currently machine learning methods are being successfully applied to a variety of biological tasks as exemplified by protein language models or protein structure prediction like AlphaFold.[14] Such approaches critically depend on well-annotated high-quality datasets. Although resources exist for various protein families, there is no specialized database focused on TPSs.[15,16]

To bridge this gap, we present MARTS-DB (Mechanisms And Reactions of Terpene Synthases DataBase): a manually curated, continuously updated database of characterized TPSs and their detailed reaction mechanisms. MARTS-DB is accessible through a user-friendly web interface that supports community data contributions. Designed for a variety of applications, special care is being taken to ensure its applicability in data-driven machine learning studies.

Data collection

The largest dataset of plant sesquiterpene synthases to date has been collected in (Durairaj, J. et al. 2019).[17,18] We have expanded this dataset by mining TPS sequences from UniProtKB/Swiss-Prot and other available resources based on their matches to TPS-specific Pfam domains.[19–21] The resulting set was broadened by the addition of hundreds of manually curated enzymes from published literature. Currently, MARTS-DB includes over 1300 characterized TPSs and IDSs that collectively catalyze the biosynthesis of more than 500 terpenes. Over 400 reaction mechanisms are annotated.

Enzymes in MARTS-DB span all domains of life, including rare examples from viruses and archaea.[22,23] Plant-derived TPSs comprise approximately two-thirds of the dataset, followed by fungal and bacterial sequences (**Fig. 1)**.[24]

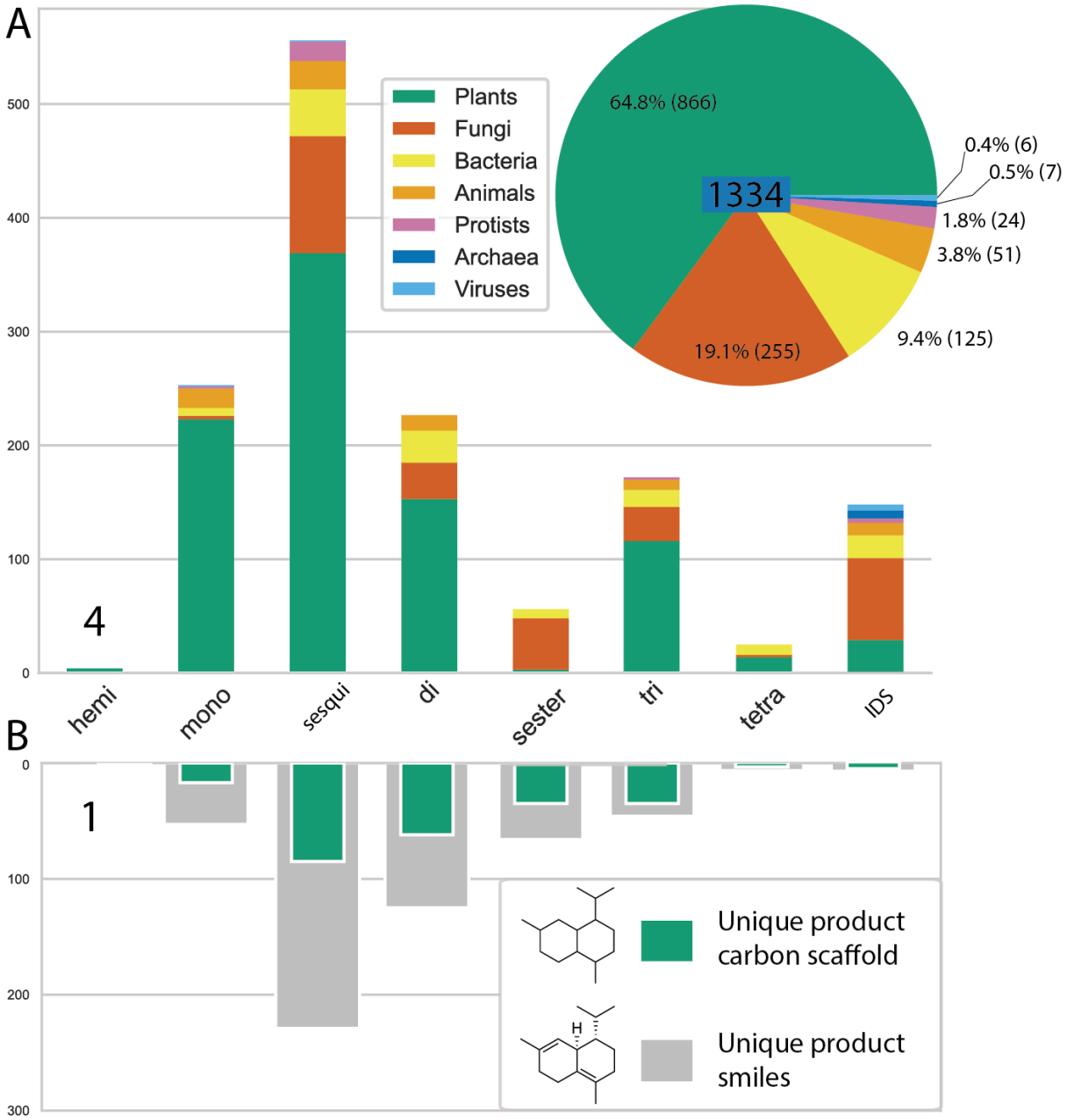

Fig. 1. The number of terpenes and terpene synthases by type and taxonomical kingdom (A) The amount of TPS catalyzing the production of a terpene by the type of terpene, colored by the taxon of their source species, along with a pie chart of the taxonomical origin of all the TPS. If a TPS catalyzes reactions of multiple types, it is counted in each relevant bar. (B) The number of unique terpene products and of unique carbon scaffolds of those products by terpene type.

Generally terpenes are categorized based on their carbon backbone: hemiterpenes (C5), monoterpenes (C10), sesquiterpenes (C15), diterpenes (C20), sesterterpenes (C25), triterpenes (C30), and beyond.[1,7] Sesquiterpenes are the most represented group in MARTS-DB, followed by diterpenes, sesterterpenes, monoterpenes, and triterpenes. Even when considering only carbon scaffolds—ignoring stereochemistry and heteroatoms—sesquiterpenes remain the most abundant. SesquiTPSs are the most common type of enzymes within MARTS, followed by monoTPSs and diTPSs

Structure of MARTS

MARTS is structured around three core entities: enzymes, molecules and reaction mechanisms. Each entity has a unique persistent identifier (MARTS ID) with the format: marts_[EMR] followed by a five digit number (E = Enzyme, M = Molecule, R = Reaction Mechanism). Even when the record is removed its identifier can be found and referenced (**Fig. 2**).

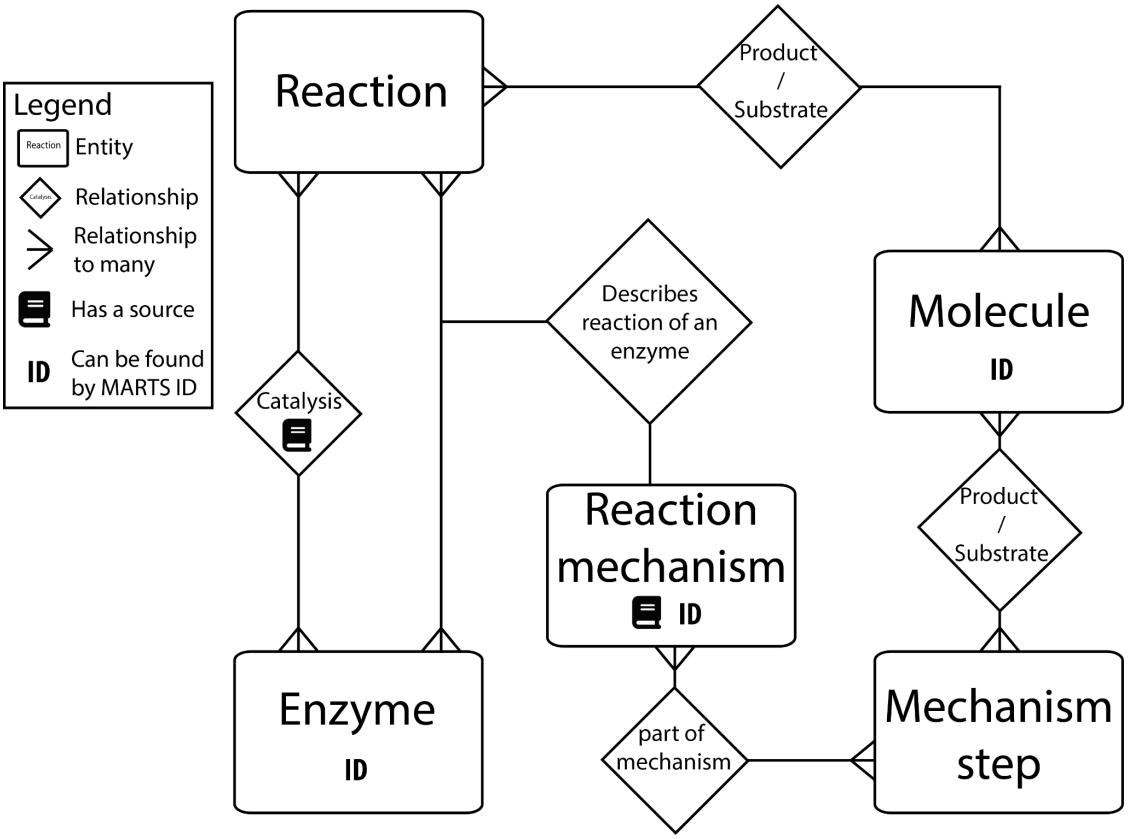

Fig. 2: The schematic structure of MARTS-DB: The core of the database are Enzymes, molecules and reaction mechanisms, which can all be found by their persistent identifier called MARTS ID. Enzymes catalyze reactions, where Molecules serve as either products or substrates. Mechanisms are composed of steps, whereres molecules again serve as either “products” or “substrates” of those steps. A pair of an enzyme and a reaction can have an assigned mechanism. A publication that is the source of this record can be found for every assignment of a reaction to an enzyme as well as every reaction mechanism.

Molecules include both terpenes and reaction intermediates. Reactions are linked to enzymes that catalyze them, each such connection has a cross reference to the original source. This allows for correct sourcing in cases where a TPS was a subject of multiple investigations. Mechanisms are represented as sequences of atomic steps, each describing a single bond change, and are also linked to their literature sources. Every reaction step is assigned its type out of 18 types found in literature. Additionally mechanisms are assigned their evidence, be it an experiment, calculation or only theoretical proposal. Any enzyme-reaction pair may have one associated mechanism. While it is unlikely this allows for theoretical cases where two enzymes utilize a different mechanism for the same reaction.[25,26]

The web interface

MARTS-DB is freely accessible at https://www.marts-db.org/. The homepage provides basic statistics and a quick search function that compares a simple query to all the fields in the database. (**Fig. 3)** Search results are displayed as enzyme cards, which include information on catalyzed reactions. Each reaction and mechanism is linked to its source publication. Alternatively records can be directly searched for by their MARTS ID. Advanced search capabilities allow for filtering by multiple parameters, including mechanism features. The statistics page provides an overview of the database contents, and a Help page includes documentation and usage instructions.

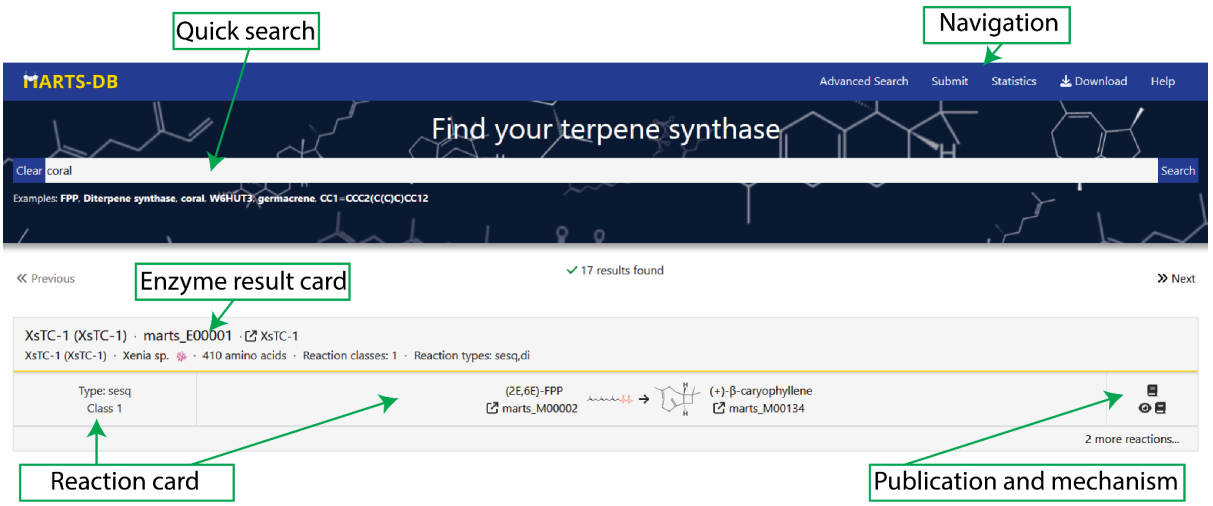

Fig. 3. The front page of MARTS: The front page of MARTS also serves as the quick search results page. Results of search queries are displayed as cards for each enzyme, with the reactions they catalyze.

Users can download the entire database through the dedicated Download page. This page provides not only the most recent version of MARTS-DB but also the complete archive of all previous releases. This comprehensive availability allows researchers to maintain backward compatibility in long-term computational studies, where data consistency across training and evaluation phases is essential. The database is distributed in CSV format, which is compatible with a wide range of analysis tools and programming environments. Details on the file format and its usage are thoroughly described in the Help section to support reproducibility and ease of integration.

Detailed record pages are available for enzymes and molecules. Enzyme records include taxonomic origin, reaction details, and cross-references to UniProt and NCBI.[20,27] Conserved TPS motifs are highlighted in the sequence view. Molecule records include SMILES and InChI identifiers, molecular properties, ChEBI cross-references (where available), and a list of associated reactions and enzymes. For both molecules and enzymes, their complete record history, tied to their MARTS ID, is described.[20,27,28]

Mechanisms are presented as sequences of individual steps, currently viewable in popup windows from reaction cards. A dedicated pathway visualization module is planned for future releases. (**Fig. 4)**

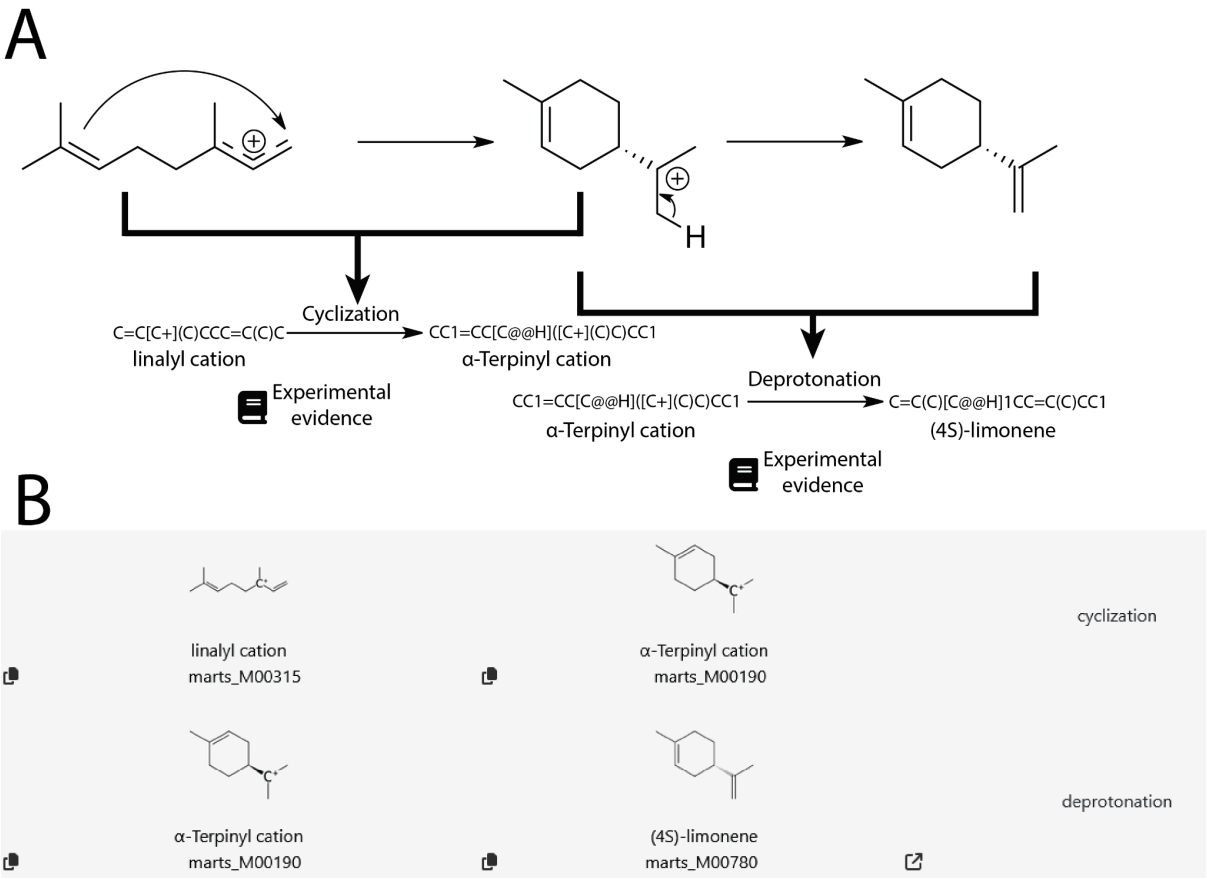

Fig. 4. **Mechanism representation in MARTS:** A) Mechanisms are represented as individual steps of the reaction cascade. Every step approximates a change in single bonding. Each step is assigned a type and evidence sourced from literature. B) The current display of mechanisms within MARTS. In the future a dedicated interface will be developed.

The database will be curated and expanded by our team in the future, additionally registered users can submit new data via the Submit page. All submitted entries will undergo a review and will be added in the next release of the database with appropriate credits.

Conclusion

We have developed MARTS-DB, a manually curated and publicly accessible database of terpene synthases and their reaction mechanisms. With over 1300 enzymes and detailed mechanistic annotations, MARTS-DB is a valuable resource for both experimental and computational studies. Early versions have already supported machine learning applications, highlighting the database's utility and potential for future research.[29]

MARTS-DB not only fills a critical gap in terpene synthase data resources but also sets the stage for interdisciplinary research by offering a structured, searchable, and evolving platform that can be adapted to emerging analytical and computational approaches. Furthermore, the meticulous curation and persistent identifier system ensure that researchers can trace every entry back to its source, maintaining the transparency and reproducibility required for robust scientific investigation.

As MARTS-DB continues to expand, we anticipate it will become an indispensable tool for the study of terpene biosynthesis, aiding in the annotation of novel enzymes, the discovery of new natural products, and the development of predictive models for enzyme function. The community-driven aspect of the platform also ensures that it will remain up to date and relevant, reflecting the collective knowledge and ongoing contributions of researchers in the field.

Methods

Dataset construction

At the time of MARTS conception, the largest dataset of TPS was constructed for plant sesquiterpene synthases with more than 250 enzymes. Additional sequences were identified using TPS-specific Pfam domains (PF01397, PF03936, PF19086, PF13249, PF13243) and extracted from UniProt/Swiss-Prot and other resources.[19–21] Manual literature curation of hundred of sources expanded the dataset to 1334 enzymes and 2606 associated reactions.

Database

MARTS-DB is built on a MySQL database with backup control handled via binary logs and scheduled mysqldump backups stored on a separate server. The internal versioning is handled by a custom system of change logs.

Web page

The website is implemented using JavaScript-based technologies. The backend uses Node.js and Axios, while the frontend uses React and CSS. Data visualizations are generated using Recharts (v2.15.1) and react-google-charts (v5.2.1).

Chemoinformatics

Utility scripts are written in Python. Chemical structure processing, including SMILES canonicalization, is performed using RDKit (v2024.03.1). Structures are visualized using RDKit and ChemDraw (v23.0.1) via its Python API. Molecules are imported from SMILES, converted to MOL files using ChemDraw, and then rendered using RDKit.

Data analysis

Statistical analyses and plots are performed using Python libraries including Pandas (v2.1.4), NumPy (v1.26.4), Matplotlib (v3.7.3), and Seaborn (v0.12.2).

References

1. Christianson, D. W. Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).
2. Whitehead, J. N., Leferink, N. G. H., Johannissen, L. O., Hay, S. & Scrutton, N. S. Decoding Catalysis by Terpene Synthases. *ACS Catal.* **13**, 12774–12802 (2023).
3. Beran, F. *et al.* Novel family of terpene synthases evolved from trans-isoprenyl diphosphate synthases in a flea beetle. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2922–2927 (2016).
4. Baker, C. H., Matsuda, S. P., Liu, D. R. & Corey, E. J. Molecular cloning of the human gene encoding lanosterol synthase from a liver cDNA library. *Biochem. Biophys. Res. Commun.* **213**, 154–160 (1995).
5. Karunanithi, P. S. & Zerbe, P. Terpene synthases as metabolic gatekeepers in the evolution of plant terpenoid chemical diversity. *Front. Plant Sci.* **10**, 1166 (2019).
6. Burkhardt, I., de Rond, T., Chen, P. Y.-T. & Moore, B. S. Ancient plant-like terpene biosynthesis in corals. *Nat. Chem. Biol.* **18**, 664–669 (2022).
7. Chen, C.-C. *et al.* Terpene cyclases and prenyltransferases: structures and mechanisms of action. *ACS Catal.*
8. Tarshis, L. C., Yan, M., Poulter, C. D. & Sacchettini, J. C. Crystal structure of recombinant farnesyl diphosphate synthase at 2.6-A resolution. *Biochemistry* **33**, 10871–10877 (1994).
9. Pan, X., Rudolf, J. D. & Dong, L.-B. Class II terpene cyclases: structures, mechanisms, and engineering. *Nat. Prod. Rep.* **41**, 402–433 (2024).
10. Rabe, P. *et al.* Conformational Analysis, Thermal Rearrangement, and EI-MS Fragmentation Mechanism of (1 (10) E, 4E, 6S, 7R)-Germacradien-6-ol by 13C-Labeling Experiments. *Angew. Chem. Int. Ed.*
11. Rabe, P., Rinkel, J., Klapschinski, T. A., Barra, L. & Dickschat, J. S. A method for investigating the stereochemical course of terpene cyclisations. *Org. Biomol. Chem.* **14**, 158–164 (2016).
12. Rinkel, J. & Dickschat, J. S. Recent highlights in biosynthesis research using stable isotopes. *Beilstein J. Org. Chem.* **11**, 2493–2508 (2015).
13. Tantillo, D. J. Biosynthesis via carbocations: theoretical studies on terpene formation. *Nat. Prod. Rep.* **28**, 1035–1053 (2011).
14. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
15. Wang, H. *et al.* PCPD: Plant cytochrome P450 database and web-based tools for structural construction and ligand docking. *Synth Syst Biotechnol* **6**, 102–109 (2021).
16. Bretaudeau, A. *et al.* CyanoLyase: a database of phycobilin lyase sequences, motifs and functions. *Nucleic Acids Res* **41**, D396–401 (2013).
17. Durairaj, J. *et al.* An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).
18. Durairaj, J. *et al.* Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comput. Biol.* **17**, e1008197 (2021).
19. Priya, P., Yadav, A., Chand, J. & Yadav, G. Terzyme: a tool for identification and analysis of the plant terpenome. *Plant Methods* **14**, 4 (2018).

20. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).
21. Paysan-Lafosse, T. *et al.* The Pfam protein families database: embracing AI/ML. *Nucleic Acids Res* **53**, D523–D534 (2025).
22. Jung, Y. *et al.* Function and Structure of a Terpene Synthase Encoded in a Giant Virus Genome. *J Am Chem Soc* **145**, 25966–25970 (2023).
23. Tachibana, A. *et al.* Novel prenyltransferase gene encoding farnesylgeranyl diphosphate synthase from a hyperthermophilic archaeon, Aeropyrum pernix. Molecularevolution with alteration in product specificity. *Eur J Biochem* **267**, 321–328 (2000).
24. Pichersky, E. & Raguso, R. A. Why do plants produce so many terpenoid compounds? *New Phytol* **220**, 692–702 (2018).
25. Pan, X. *et al.* Discovery, structure, and mechanism of a class II sesquiterpene cyclase. *J. Am. Chem. Soc.* **144**, 22067–22074 (2022).
26. Garms, S., Köllner, T. G. & Boland, W. A multiproduct terpene synthase from Medicago truncatula generates cadalane sesquiterpenes via two different mechanisms. *J. Org. Chem.* **75**, 5590–5600 (2010).
27. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res* **53**, D20–D29 (2025).
28. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res* **44**, D1214–9 (2016).
29. Samusevich, R. *et al.* Highly accurate discovery of terpene synthases powered by machine learning reveals functional terpene cyclization in Archaea. *bioRxiv* 2024.01.29.577750 (2024) doi:10.1101/2024.01.29.577750.

Funding

M.E. was supported from the grant of Specific university research – grant No. A2_FCHT_2025_011. T.P. was supported by the Czech Science Foundation (GA CR) grant 21-11563M and by the European Union's HORIZON program (ERC, TerpenCode, 101170268). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Author contribution

T.P. conceptualized the project. M.E, A.T., T.Č., R.Sam., R.Ch., and M.P. collected and curated the dataset. M.E. and M.B. created the web interface. M.E. wrote the manuscript. T.P. supervised the project.

Data availability

The MARTS-DB dataset is provided under CC BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).