Structure-enabled enzyme function prediction unveils elusive terpenoid biosynthesis in archaea

Raman Samusevich[1,2&], Téo Hebra[1&], Roman Bushuiev[1,2], Martin Engst[1], Jonáš Kulhánek[2], Anton Bushuiev[2], Joshua D. Smith[1,3], Tereza Čalounová[1], Helena Smrčková[1], Marina Molineris[1], Renana Schwartz[4], Adéla Tajovská[1], Milana Perković[1], Ratthachat Chatpatanasiri[2], Sotirios C. Kampranis[5], Dan Thomas Major[4], Josef Sivic[2*], Tomáš Pluskal[1*]

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

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

[3]First Faculty of Medicine, Charles University, Kateřinská 1660/32, 121 08 Prague, Czech Republic

[4]Department of Chemistry and Institute of Nanotechnology and Advanced Materials and Israel National Institute of Energy Storage, Bar-Ilan University, Ramat-Gan 5290002, Israel

[5]Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg C, 1871, Denmark

[*]Corresponding authors: josef.sivic@cvut.cz, tomas.pluskal@uochb.cas.cz

[&]These authors contributed equally to this work.

Abstract

The exponential growth of uncharacterized enzyme sequences in genomic repositories demands novel tools for functional annotation. Here, we combined alignment-driven structural domain analysis with protein language models to create EnzymeExplorer, a machine-learning pipeline for enzyme function prediction. We applied this approach to terpene synthases, which present an ideal model case because they catalyze complex carbocationic rearrangements whose product outcomes cannot be predicted from active site residues alone. We detected new structural domains and achieved state-of-the-art performance on function prediction. By analyzing the UniRef90 database, we identified terpene synthases overlooked by current computational methods. Remarkably, we uncovered and experimentally validated the widespread biosynthesis of terpenoids in archaea. Our approach offers a powerful framework for characterizing enzyme “dark matter” in the rapidly expanding genomic and metagenomic datasets.

Main text

The number of protein sequences in genomic repositories is increasing exponentially, making experimental characterization of all discovered proteins unfeasible. Bioinformatic signature scanning tools such as InterProScan (*1*) can provide partial sequence annotations, including conserved domains, sequence motifs, and binding sites. However, for enzymes, these annotations typically offer only broad functional classifications (e.g., “O-methyltransferase”) without revealing additional biochemical context, such as substrates or reaction products. Moreover, over 40 million protein sequences in the UniProt database (*2*) currently lack InterPro annotations entirely. This unannotated “dark matter” of protein sequence space likely harbors millions of uncharacterized enzymes, presenting both a barrier to biological research and an untapped resource for biotechnology (*3*). While current state-of-the-art methods for functional annotation of enzymes (*4*, *5*) perform well at identifying previously reported enzymatic activities, it is difficult for them to annotate protein sequences distant from training data. Thus, enhanced, contextually-rich predictions for uncharacterized enzyme functions are needed to facilitate efficient exploration of enzyme diversity within the uncharted protein sequence space.

Computational prediction of enzyme functions without recognizable sequence signatures remains a formidable scientific challenge (*6*). To address this challenge, we developed a new predictive approach that leverages protein structure prediction to provide informative features for machine learning. Structural information can help overcome the sparsity of sequence space, as structure is known to evolve slower than sequence (*7*), and catalytically similar enzymes with low sequence similarity share common folds (*8*). Thus, we set out to build a predictive pipeline that integrates structure-based and language-based representations of protein sequences. To implement our approach, we focused on a model enzyme family that exemplifies the current challenges in sequence-based function prediction.

Terpene synthases (TPSs) are widely distributed enzymes responsible for producing the core hydrocarbon scaffolds of terpenoids, the largest class of natural products (*9*). With over 100,000 known molecular structures, terpenoids include widely used flavors, fragrances, and first-line medicines, including the antimalarial drug artemisinin and the chemotherapy agent paclitaxel (*10*, *11*). TPSs catalyze complex cyclization reactions on isoprenoid diphosphate substrates composed of C_5 isoprene units (e.g., geranyl pyrophosphate C_{10} or farnesyl pyrophosphate C_{15}, see overview in **Fig. S1**), often yielding multiple products with fused rings and intricate stereochemistry (*9*). The intricate reaction mechanism of a TPS involves a series of carbocationic rearrangements whose outcome is dictated primarily by the surface contour and dynamics of their active sites rather than the direct involvement of active site residues in the enzymatic mechanism. Terpene synthase active sites are mostly lined by chemically “inert” residues and substitution of even a single residue has been shown to disrupt the precise choreography of the carbocationic interconversion to result in synthases with drastically different product profiles (*12*–*14*). These minor substitutions, only changing the active site contour by a few angstroms, have profound outcomes that are difficult to predict by sequence information alone. This mechanistic complexity positions TPSs as an ideal model enzyme class for applying machine learning approaches (*15*).

Here, we present a novel machine learning pipeline that combines alignment-driven structural analyses with a fine-tuned protein language model. The pipeline was trained on a manually curated

dataset of 2,246 known TPS reactions to predict the TPS enzymatic activity and substrate preference. We demonstrate that this approach outperforms existing tools developed for enzyme detection and classification and generalizes well to unseen protein sequences. As a result, we are able to identify previously undetected structural domain subtypes and to discover and experimentally characterize several new functional TPSs overlooked by current annotation methods. Remarkably, we reveal that TPSs are widespread in archaea and functionally characterize the first TPSs ever described from this kingdom.

Structure-based approach reveals undetected structural domain subtypes

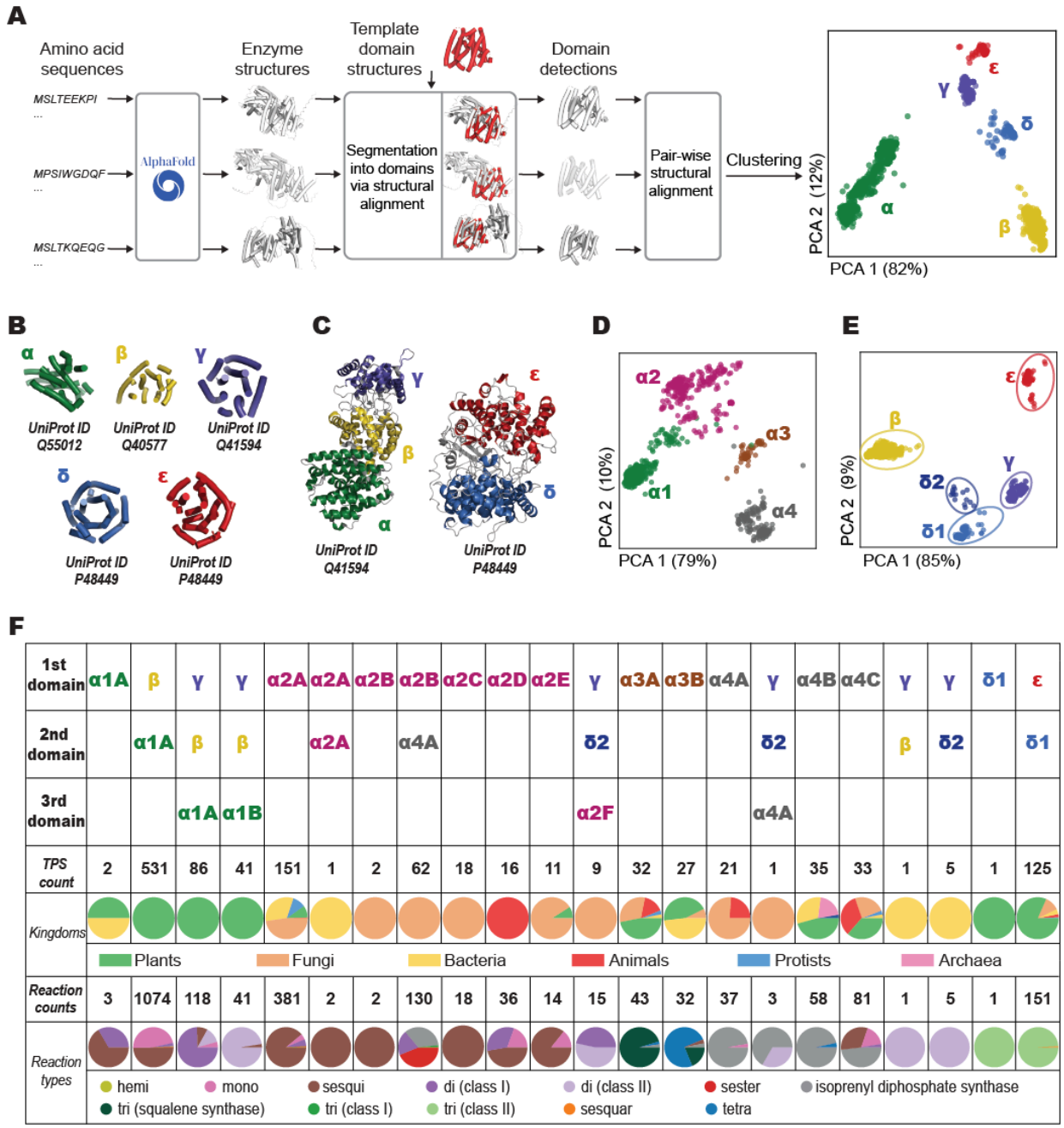
The goal of predictive discovery of family-specific enzymatic activity is to utilize a small dataset of previously characterized enzymes from that family to detect the activity in larger datasets of uncharacterized protein sequences. The quality of the training dataset is critical for the effectiveness of the prediction. Thus, to ensure reliable prediction, we began by manually curating a dataset of TPSs and their associated reactions, including only experimentally validated data. This resulted in a database of 1,211 characterized TPS enzymes that catalyze 2,246 corresponding reactions (**Supplementary File 1**). This dataset provides robust coverage of common TPS classes, such as monoterpene synthases (monoTPSs), sesquiterpene synthases (sesquiTPSs) and diterpene synthases (diTPSs; as defined in **Fig. S1**), but includes relatively fewer examples of rarer TPS classes, such as sesterTPSs or tetraTPSs, which are less well studied (**Fig. S2A** and **Fig. S2B**).

To construct our predictive pipeline, we set out to develop structure-based representations of the input protein sequence and combine them with language-based representations (**Fig. 1A**). Since enzymes in general comprise one or more structural domains (*16*) and the organization of these domains largely determines enzymatic function, we studied the domain configurations of the enzymes in our curated dataset to provide informative features for machine learning. TPS enzymes feature a modular protein architecture, composed of structural domains known as α, β, and γ (*9*, *15*). Thus, for the structure-based component of our pipeline, we leveraged AlphaFold 2 (*17*) to predict the protein structure and designed an approach that segments AlphaFold 2-predicted structures into structural domains (**Fig. S3**). This process generates a domain configuration of each input protein sequence.

We applied this approach to our dataset of 1,211 characterized TPS enzymes and performed a clustering analysis of pairwise structural alignments of all detected domains (**Fig. 1A**, right panel). Remarkably, this analysis revealed two clusters of previously unreported domains, in addition to the established α, β, and γ domains. We designated these novel domain types as δ and ε (**Fig. S4A**, **Fig. S4B**). All five domain types are structurally distinct, both when they are aligned separately (**Fig. 1B**) and when they are compared in the context of the full-length TPS proteins (**Fig. 1C**). Furthermore, additional clustering analysis revealed distinct subtypes of these five main structural domains. Specifically, the α domain structures can be further separated into 13 distinct subtypes, which we designated α1A to α4C (**Fig. S4C**, **Fig. S4D,** and **Fig. 1D**), while the newly identified δ domain structures contain two distinct subtypes, designated δ1 and δ2 (**Fig. S4E**, **Fig. S4F**, and **Fig. 1E**).

Overall, we were able to distinguish 22 existing TPS domain configurations (**Fig. 1F** and **Supplementary File 2**), further supported by phylogenetic analysis (**Fig. S5**). This analysis

provides insights into the evolutionary diversity of TPS structures: for example, all known sesterTPS enzymes have a two-domain α2B-α4A architecture, which is also shared with the recently discovered class I triTPSs that cyclize hexaprenyl pyrophosphate (*18*). In addition, the novel δ and ε domains are primarily associated with class II triTPSs, being the only structural domains synthesizing lanosterol or cycloartenol, the precursors of steroidal compounds.



1st domain	α1A	β	γ	γ	α2A	α2A	α2B	α2B	α2C	α2D	α2E	γ	α3A	α3B	α4A	γ	α4B	α4C	γ	γ	δ1	ε
2nd domain		α1A	β	β		α2A		α4A				δ2				δ2			β	δ2		δ1
3rd domain			α1A	α1B								α2F				α4A						
TPS count	2	531	86	41	151	1	2	62	18	16	11	9	32	27	21	1	35	33	1	5	1	125
Reaction counts	3	1074	118	41	381	2	2	130	18	36	14	15	43	32	37	3	58	81	1	5	1	151

Fig. 1. Discovery and analysis of structural domain types and subtypes using a segmentation-clustering pipeline. (A) Overview of our structure-based workflow. AlphaFold2-predicted protein structures are segmented into family-specific domains via structural alignment. These detected domains are then clustered, revealing distinct structural domain groups. **(B)** Typical representatives of α, β, γ, δ, and ε TPS domains, structurally aligned. **(C)** The highlighted structures of typical α, β, γ domains in the context of the diTPS taxadiene synthase (UniProt ID Q41594, PDB ID 3P5R) and the typical δ and ε domains in the context of a triTPS lanosterol synthase (UniProt ID P48449, AlphaFold-generated structure). **(D)** Principal component analysis of clustering of all detected α domains. **(E)** Principal component analysis of clustering of all detected non-α domains, highlighting the two distinct subtypes of the δ domain. **(F)** Structural domain configurations of 1,211 validated natural TPS enzymes (**Supplementary File 1**), including the taxonomy distribution and TPS reaction types for each domain configuration. Please note that if two different

enzymes counted in the ***TPS count*** row catalyze the same reaction (substrate/product pair), these are counted independently as two reactions in the ***Reaction counts*** row.

Predictive model for enzyme functional annotation

The key challenge in predicting enzymatic activities lies in designing a predictor that (i) generalizes well to unseen protein sequences, especially with limited training data (some TPS classes have only a handful of characterized sequences), and (ii) can distinguish fine-grained structural differences between the different TPS classes. To address this challenge, we have developed a pipeline that combines alignment-driven structural analysis with a fine-tuned protein language model (**Fig. 2A**). In detail, the similarity of detected structural domains to each recognized TPS domain template (**Fig. 1F**) is analyzed using the Foldseek (*19*) algorithm. This set of similarities, quantified as TM-scores, is then encoded into a numerical vector representing the structural information. This representation enables encoding fine-grained structural information specific to the given enzyme class (here TPS). In the language-based component, we employed the large protein language model ESM-1v *(20)* fine-tuned on a dataset of 80 thousand putative (uncharacterized) TPS sequences (**Supplementary File 3**), which provided an additional numerical vector encoding the sequence information. Leveraging a strong pre-trained protein language model enables generalization to new sequences and finetuning specializes the model to represent well the sequence variation in the given enzyme class (here TPS). We used the concatenated structural and sequence vector representations of the input protein as an input feature for three random forest classifiers (*21*) that predict the structural domain configuration of the given protein, the presence or absence of TPS enzymatic activity, and the substrate of the enzyme (**Fig. 2A**, right side)**.** The random forest classifiers were trained in a supervised manner using our curated dataset of 2,246 TPS reactions (**Supplementary File 1**) and 10,000 negative examples (non-TPS protein sequences). The complete dataset was divided into five folds for cross-validation, based on the phylogenetic distribution of the sequences (**Supplementary File 4).** To put functional predictions into context of known TPSs, domain subtypes of functionally characterized TPSs were used as targets to train a random forest classifier predicting domain types for novel TPS detections (**Fig. S6**).

Having established our pipeline, we investigated its performance in predicting enzyme function by comparing it to well-established profile hidden Markov models (pHMMs) in the Pfam (*22*) and SUPFAM (*23*) databases as well as to TPS-specific pHMMs retrained on our own TPS dataset (**Supplementary File 1**) (*24*). As additional baselines, we used the protein sequence search algorithm BLASTp (*25*), the protein structure search algorithm Foldseek (*19*), and the state-of-the-art EC number predictor CLEAN (*5*). Compared to all these methods, our approach consistently achieved the highest average precision for TPS detection across diverse taxa (**Fig. 2B and 2C**) and the most accurate multi-label substrate predictions across different TPS types (**Fig. 2D and 2E**), including underrepresented TPS classes such as sesterTPSs or tetraTPSs. The detailed methodology, baselines, validation scheme, evaluation metrics, and ablation studies supporting these results are outlined in the Materials and Methods and in Supplementary Information (**Fig. S7, S8, S9**).

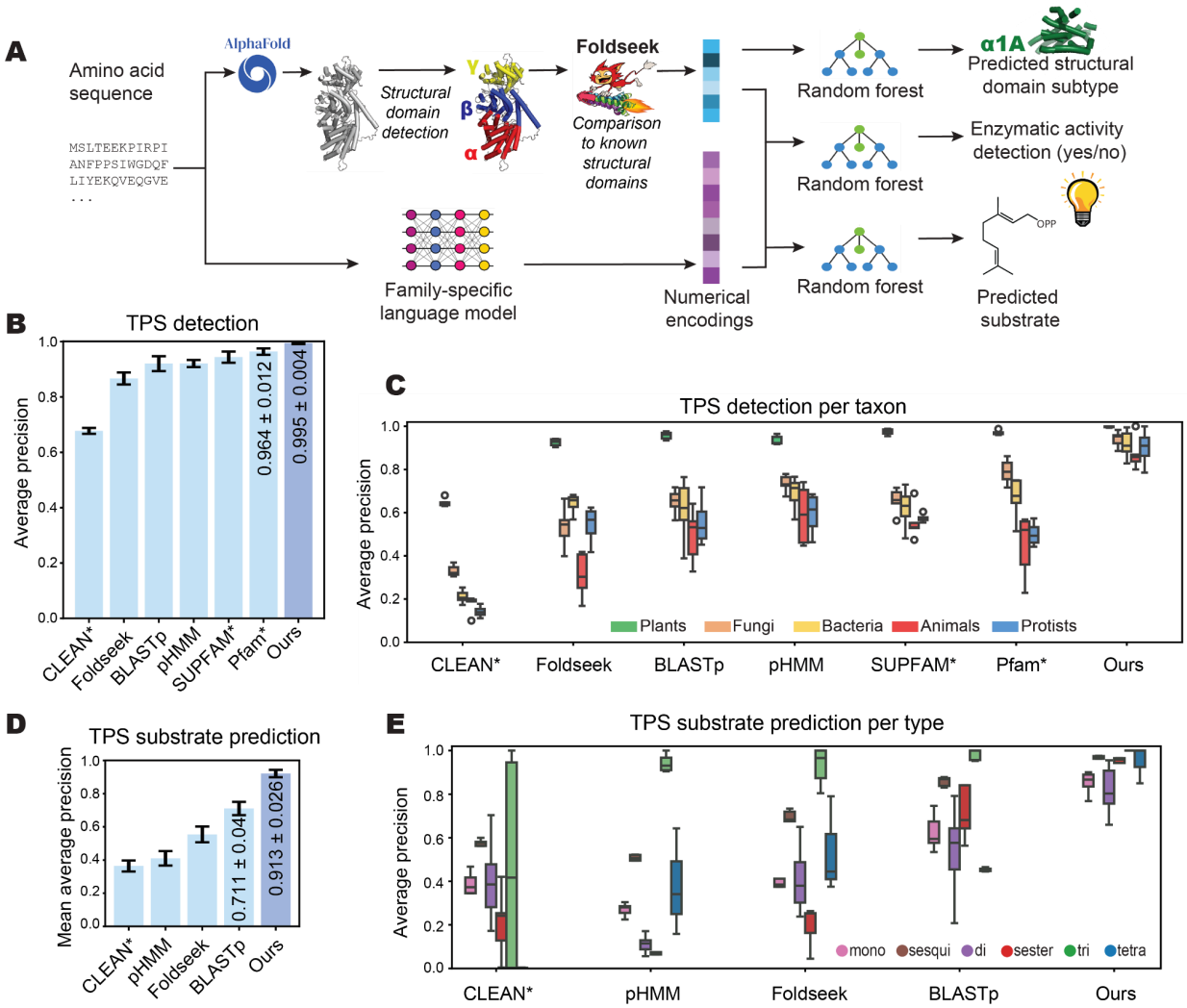


Fig. 2. Accurate enzyme function prediction using EnzymeExplorer (A) Overview of our predictive pipeline. Protein structures predicted by AlphaFold are segmented into enzyme-specific domains, which are compared to domains of the same type within previously characterized enzymes from the same family. Similarities to characterized domains, quantified as TM-scores, are recorded as features for downstream analysis. Protein sequences are also processed by a family-specific language model to extract embeddings. Finally, the concatenated TM-score features and embeddings are fed into random forest classifiers to predict new structural domains, identify the presence of enzymatic activity, and classify substrates. **(B)** In-silico evaluation of TPS detection performance was measured using average precision (area under the precision-recall curve). Individual precision-recall curves are in **Fig. S7**. CLEAN, Pfam, and SUPFAM were used as off-the-shelf methods without retraining them on our dataset; these methods are indicated by asterisks next to their names. **(C)** Taxon-specific performance of our approach for TPS detection compared to other methods. **(D-E)** Substrate prediction performance, highlighting the efficacy of our approach compared to existing methods across all types, including underrepresented minority TPS classes like tetraTPS. Additional comparisons using other metrics are provided in **Fig. S7**. Error bars in the bar plots represent the standard error of the mean (SEM), computed from the aggregated variance across all folds.

Experimental validation of model predictions

To validate the performance of our machine learning approach for the discovery and characterization of novel TPSs, we experimentally characterized the enzymatic activity of selected candidate synthases. To functionally characterize these enzymes, we heterologously expressed the corresponding genes in the budding yeast *Saccharomyces cerevisiae* strain JWY501 (*26*). This strain has been engineered for elevated production of the sesquiTPS and diTPS substrates farnesyl diphosphate (FPP) and geranyl geranyl diphosphate (GGPP), respectively.

In the first validation experiment, we mined uncharacterized protein sequences that carry PFAM or SUPFAM signatures characteristic for TPSs (**Supplementary File 3**). Through phylogenetic analysis, we selected nine candidates phylogenetically most distant from any previously characterized TPSs in our dataset (**Fig. S10A, Table S1**). Using gas chromatography mass spectrometry (GC-MS) and liquid chromatography mass spectrometry (LC-MS) we confirmed that out of the nine tested enzymes, six exhibited production of either sesquiterpene or diterpene compounds, confirming that our algorithm can provide reliable predictions on phylogenetically distant biological sequences (**Fig. S11-S14**).

To further assess our method's ability to facilitate novel discoveries in unexplored parts of protein sequence space, we aimed to identify "*impossible to discover"* TPSs, focusing on cases where existing bioinformatic methods would completely miss these proteins. We applied EnzymeExplorer to sequences from the UniRef90 database that lack any known InterPro signatures, including any "domain of unknown function" (**Fig. S10B**, full dataset available on Zenodo). For experimental validation, we manually selected enzymes originating from organisms where terpenoid biosynthesis has been less studied or unreported, such as bacteria, animals, amoebae, viruses, and archaea, and selected hits with predicted domain types recognizable as TPS structural domains (**Table S2**). Following the heterologous expression of 17 candidates in *S. cerevisiae*, we confirmed that seven candidates exhibited TPS activity (**Fig. 3A, Fig. 3B, Fig. S15-S18**). Notably, our algorithm predicted the correct substrate with the highest probability in six out of seven confirmed candidates (**Table S2**). Among these highly unconventional protein sequences, we identified the second viral TPS (A0A2P0VN22) reported to date (*27*). Notably, we identified three active TPS enzymes in archaea (A0A0E3NXY0, A0A537EJD0, A0A5E4I9B1), a kingdom of organism considered with no previous TPS reports until now (*28*).

To further validate the finding of functional TPSs in archaea, we selected the most active of the identified TPSs, A0A5E4I9B1 (**Fig. 3B**), to structurally characterize its main product. To enhance terpene production in yeast, we transitioned from genomic integration to plasmid-based expression of the A0A5E4I9B1 gene. This enabled the isolation of the main product and its identification as *β*-bisabolol by nuclear magnetic resonance (NMR) spectroscopy (**Fig. 3C**, **Fig. S19**).

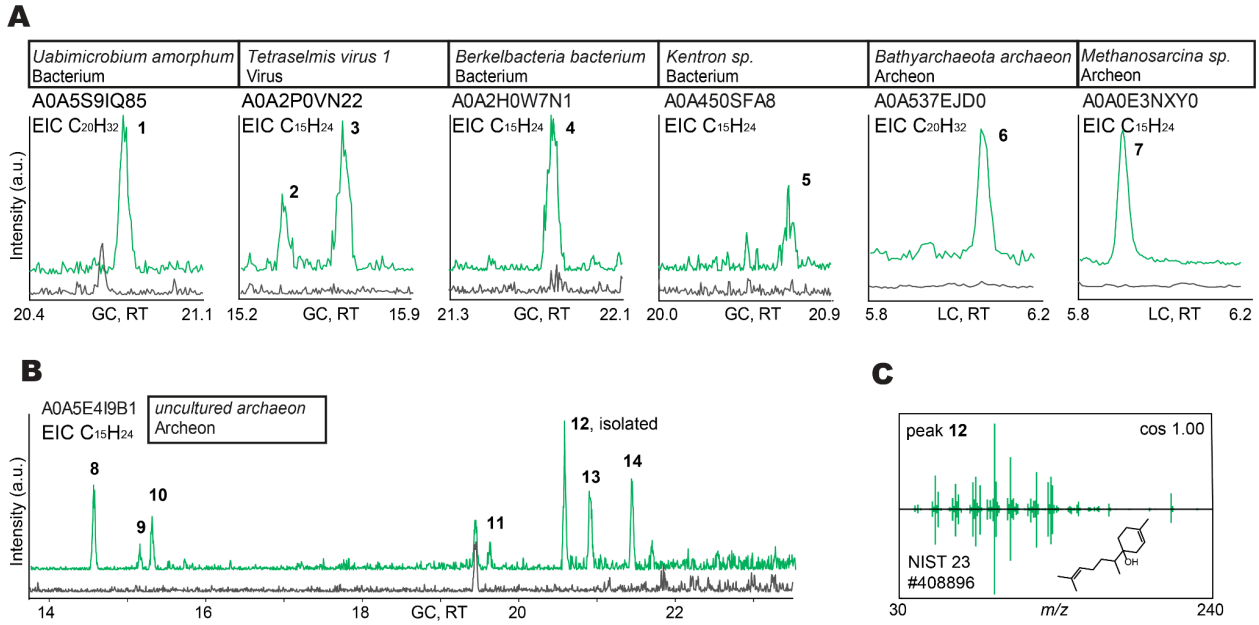


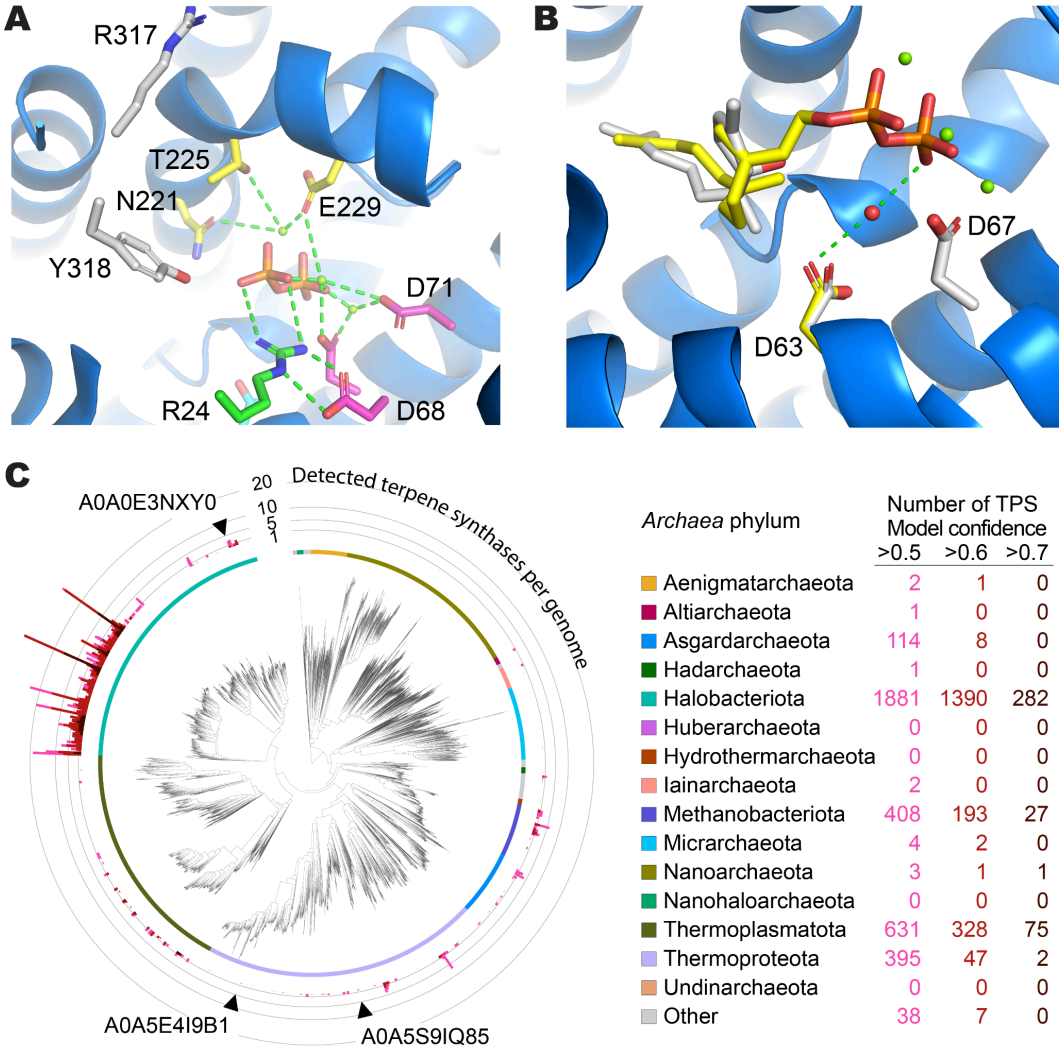
Fig. 3. Selected results of the experimental validation of EnzymeExplorer predictions for sequences that lack InterPro signatures. (A) GC-MS and LC-MS extracted ion chromatograms (EICs) for sesquiterpene ($C_{15}H_{24}$) and diterpene ($C_{20}H_{32}$) molecular ions detected in the culture fluid of yeast strain JWY501 transformed with plasmids containing selected indicated enzymes (green). Empty vector control in JWY501 shown in black. **(B)** A GC-MS extracted ion chromatogram for sesquiterpene ($C_{15}H_{24}$) molecular ion detected in the culture fluid of yeast strain JWY501 transformed with plasmids containing selected indicated enzymes (green). Empty vector control in JWY501 shown in black. **(C)** The GC-MS fragmentation spectrum of peak 12 (top) shown as a mirror plot against the closest NIST23 EI library match (bottom). Matching peaks are in green; non-matching peaks in grey; cosine similarity score indicated (cos).

Structural analysis of the archaeal terpene synthases

Although the detected archaeal TPSs have no typical PFAM or SUPFAM domains, we were able to identify several known motifs in their amino acid sequences (**Fig. S20**). To further examine their structure, we generated an AlphaFold 3 (*29*) model of A0A5E4I9B1 including three Mg^{2+} metal ions and docked pyrophosphate in it (**Fig. S21**). The AlphaFold 3 model correctly assigns the ^{67}**D**DXX**D** motif in helix D (*9*), wherein D67 and D71 coordinate Mg_A and Mg_C metal ions in bi- and monodentate modes, respectively (bold letters indicate metal binding residues; naming of ions in **Fig. S22**) (**Fig. 4A**). D68 coordinates R24, which is a conserved residue observed in plant TPSs towards the N-terminus sequence region (*30*). An additional well-known metal binding motif, which binds Mg_B and Mg_C, is (**N,D**)D(L,I,V)X(**S,T**)XXX**E** in helix H (*9*), and in A0A5E4I9B1 this motif is ^{221}**N**TLN**T**WPR**E**.

The active site contains an unusual D63 which protrudes into the active site (**Fig. 4B**). Considering that this residue is a full helical turn downstream from the highly conserved D67, this is unlikely an artefact of the structure model. A possible role of this Asp residue is to bind and deprotonate an active-site water, which is involved in forming *β*-bisabolol. To assess the catalytic mechanism in A0A5E4I9B1, we performed reaction modeling. A water molecule possibly binds between D67 and the diphosphate moiety, which resembles a conserved water binding site in plant TPS (*30*) (**Fig. 4B**). To model the reaction, we performed EnzyDock simulations, which allows rapid multistate docking that can provide mechanistic insight (*30*, *31*). We docked the substrate FPP and

various diastereomers of *β*-bisabolol using EnzyDock (**Fig. 4B**). These docking simulations suggest that the isoprenoid chain in FPP connects to the O1*α* oxygen of the diphosphate as observed for plant TPS (*30*) and is bound in an extended form (**Table S4**), and the tail of *β*-bisabolol occupies the same binding pocket as in FPP (**Fig. 4B**). In our docking simulations, the aforementioned water molecule is within 4.2 Å of the C6 position in FPP and is close to the position of the hydroxyl group in *β*-bisabolol (4.1-4.5 Å) (**Table S3**).



Archaea phylum	Number of TPS Model confidence >0.5	>0.6	>0.7
Aenigmatarchaeota	2	1	0
Altiarchaeota	1	0	0
Asgardarchaeota	114	8	0
Hadarchaeota	1	0	0
Halobacteriota	1881	1390	282
Huberarchaeota	0	0	0
Hydrothermarchaeota	0	0	0
Iainarchaeota	2	0	0
Methanobacteriota	408	193	27
Micrarchaeota	4	2	0
Nanoarchaeota	3	1	1
Nanohaloarchaeota	0	0	0
Thermoplasmatota	631	328	75
Thermoproteota	395	47	2
Undinarchaeota	0	0	0
Other	38	7	0

Fig. 4. Structural and evolutionary analysis of the archaeal TPSs. (A) Refined TPS model of A0A5E4I9B1. Conserved interactions (≤ 4 Å) marked with green dashed lines are detailed in **Table S3**. **(B)** Superposition of the lowest energy pose of the substrate FPP and matching low energy pose of *β*-bisabolol in A0A5E4I9B1 obtained from EnzyDock docking. Position of the catalytic water molecule is shown in between the diphosphate moiety and D63. **(C)** A Genome Taxonomy Database (*32*) phylogenetic tree of the archaeal domain with the distribution of terpene synthases depending on the confidence of the EnzymeExplorer model. The tree is annotated with the number of predicted TPS detections, based on the selected EnzymeExplorer detection threshold (>0.5 = pink, >0.6 = light brown, 0>0.7 = dark brown; higher value indicates higher prediction confidence for TPS activity).

Terpenoid biosynthesis is widespread in archaea

The discovery of functional TPSs in archaea marks a significant advancement in our understanding of the distribution and evolutionary origins of TPSs, as these enzymes were thought to exist in archaea, but their discovery remained elusive (*28*). To shed more light into this unique finding, we investigated the prevalence of TPSs across the phylogeny of archaea. Analysis of archaeal genomes from the National Center for Biotechnology Information (NCBI) revealed TPS presence in 678 archaeal taxa at a 0.5 prediction confidence level and in 125 taxa at a 0.7 confidence level (**Supplementary File 5**). Notably, some genomes, such as Haloarcula sp. H-GB4, encode up to 16

different TPSs. TPSs were detected across 12 of the 16 archaeal phyla, with the highest representation in *Halobacteriota*, *Thermoproteota*, and *Methanobacteriota* (**Fig. 4C**).

Discussion

Compared to state-of-the-art methods for predicting enzyme function (*5*), the EnzymeExplorer approach comprises several innovations. By focusing on a single enzyme class, we were able to incorporate class-specific biochemical knowledge on the characteristic structural domains into our machine learning workflow. To achieve that, we developed a structural alignment-based method for precise segmentation of protein structures into structural domains and an unsupervised clustering approach for the identification of subtypes within the established domain types. This annotation of structural domains supports functional predictions, and is therefore important for building trust in the computational annotations, especially in the absence of other recognizable protein signatures. Domain structures tend to be more conserved than sequences, and are linked to enzymatic reaction mechanisms (*7*, *16*).

By combining the structural analyses with a fine-tuned protein language model, we developed a machine-learning approach that significantly outperforms existing methods for enzyme sequence annotation, particularly for sequences from minority classes for which only few examples exist (e.g., tetraTPS). The enhanced performance of our predictive model allowed us to examine the "dark matter" protein sequences that lack any of the 53,700 existing InterPro signatures (*33*). Existing annotated enzyme datasets are relatively small and sparse, and machine learning methods have not yet demonstrated the capability to discover entirely novel enzymatic sequences within such unexplored regions of protein sequence space. In contrast, EnzymeExplorer successfully predicted seven such enzymes, which we experimentally validated. Furthermore, our automatic annotation of the refined structural domain subtypes provided additional contextual evidence and interpretable family-specific insights for the discoveries.

The broad generalization capabilities of EnzymeExplorer in annotating the unexplored protein sequence space led to the discovery of the first archaeal TPSs. While it has been well established that cellular membranes of archaea are composed of terpenoid-based phospholipids (*28*), and several isoprenyl diphosphate synthase genes have been previously found in archaea (*34*, *35*), no functional terpene cyclization has been observed in archaea until now. This absence of evidence even contributed to the hypothesis that archaea might lack secondary metabolite production (*28*). Using NMR spectroscopy, we identified the major product of one archaeal TPS as the sesquiterpene alcohol *β*-bisabolol. Although *β*-bisabolol has been previously reported in various plant species (*36*), no TPS enzyme responsible for producing this compound has been previously identified. Our method has thus not only uncovered the first *β*-bisabolol synthase but also demonstrated its potential to discover elusive enzyme sequences from underexplored organisms or metagenomic datasets.

References

1. P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A. F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, S. Hunter, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

2. UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).

3. R. Buller, S. Lutz, R. J. Kazlauskas, R. Snajdrova, J. C. Moore, U. T. Bornscheuer, From nature to industry: Harnessing enzymes for biocatalysis. *Science* **382**, eadh8615 (2023).

4. G. B. Kim, J. Y. Kim, J. A. Lee, C. J. Norsigian, B. O. Palsson, S. Y. Lee, Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat. Commun.* **14**, 7370 (2023).

5. T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang, H. Zhao, Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).

6. V. de Crécy-Lagard, R. Dias, I. Friedberg, Y. Yuan, M. A. Swairjo, Limitations of current machine-learning models in predicting enzymatic functions for uncharacterized proteins, *bioRxiv* (2024). https://doi.org/10.1101/2024.07.01.601547.

7. J. Durairaj, E. Melillo, H. J. Bouwmeester, J. Beekwilder, D. de Ridder, A. D. J. van Dijk, Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comput. Biol.* **17**, e1008197 (2021).

8. Y. Gao, R. B. Honzatko, R. J. Peters, Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat. Prod. Rep.* **29**, 1153–1175 (2012).

9. D. W. Christianson, Structural and Chemical Biology of Terpenoid Cyclases. *Chem. Rev.* **117**, 11570–11648 (2017).

10. N. J. White, Qinghaosu (artemisinin): the price of success. *Science* **320**, 330–334 (2008).

11. M. Köksal, Y. Jin, R. M. Coates, R. Croteau, D. W. Christianson, Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* **469**, 116–120 (2010).

12. Y. Yoshikuni, T. E. Ferrin, J. D. Keasling, Designed divergent evolution of enzyme function. *Nature* **440**, 1078–1082 (2006).

13. B. T. Greenhagen, P. E. O'Maille, J. P. Noel, J. Chappell, Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9826–9831 (2006).

14. S. C. Kampranis, D. Ioannidis, A. Purvis, W. Mahrez, E. Ninga, N. A. Katerelos, S. Anssour, J. M. Dunwell, J. Degenhardt, A. M. Makris, P. W. Goodenough, C. B. Johnson, Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: structural insights into the evolution of terpene synthase function. *Plant Cell* **19**, 1994–2005 (2007).

15. J. N. Whitehead, N. G. H. Leferink, L. O. Johannissen, S. Hay, N. S. Scrutton, Decoding Catalysis by Terpene Synthases. *ACS Catal.* **13**, 12774–12802 (2023).

16. G. L. Holliday, J. D. Fischer, J. B. O. Mitchell, J. M. Thornton, Characterizing the complexity of enzymes on the basis of their mechanisms and structures with a bio-computational analysis: The complexity of enzymes. *FEBS J.* **278**, 3835–3845 (2011).

17. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

18. H. Tao, L. Lauterbach, G. Bian, R. Chen, A. Hou, T. Mori, S. Cheng, B. Hu, L. Lu, X. Mu, M. Li, N. Adachi, M. Kawasaki, T. Moriya, T. Senda, X. Wang, Z. Deng, I. Abe, J. S. Dickschat, T. Liu, Discovery of non-squalene triterpenes. *Nature* **606**, 414–419 (2022).

19. I. Barrio-Hernandez, J. Yeo, J. Jänes, M. Mirdita, C. L. M. Gilchrist, T. Wein, M. Varadi, S. Velankar, P. Beltrao, M. Steinegger, Clustering predicted structures at the scale of the known protein universe. *Nature*, doi: 10.1038/s41586-023-06510-w (2023).

20. J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, A. Rives, Language models enable zero-shot prediction of the effects of mutations on protein function, *bioRxiv* (2021). https://doi.org/10.1101/2021.07.09.450648.

21. L. Breiman, Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

22. A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, S. R. Eddy, The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–41 (2004).

23. S. B. Pandit, R. Bhadra, V. S. Gowri, S. Balaji, B. Anand, N. Srinivasan, SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinformatics* **5**, 28 (2004).

24. P. Priya, A. Yadav, J. Chand, G. Yadav, Terzyme: a tool for identification and analysis of the plant terpenome. *Plant Methods* **14**, 4 (2018).

25. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

26. J. Wong, T. de Rond, L. d'Espaux, C. van der Horst, I. Dev, L. Rios-Solis, J. Kirby, H. Scheller, J. Keasling, High-titer production of lathyrane diterpenoids from sugar by engineered Saccharomyces cerevisiae. *Metab. Eng.* **45**, 142–148 (2018).

27. Y. Jung, T. Mitsuhashi, S. Sato, M. Senda, T. Senda, M. Fujita, Function and Structure of a Terpene Synthase Encoded in a Giant Virus Genome. *J. Am. Chem. Soc.* **145**, 25966–25970 (2023).

28. Y. Hoshino, L. Villanueva, Four billion years of microbial terpenome evolution. *FEMS Microbiol. Rev.* **47** (2023).

29. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H.

Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, J. M. Jumper, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

30. R. Schwartz, S. Zev, D. T. Major, Differential substrate sensing in terpene synthases from plants and microorganisms: Insight from structural, bioinformatic, and EnzyDock analyses. *Angew. Chem. Int. Ed Engl.* **63**, e202400743 (2024).

31. S. Das, M. Shimshi, K. Raz, N. Nitoker Eliaz, A. R. Mhashal, T. Ansbacher, D. T. Major, EnzyDock: Protein-ligand docking of multiple reactive states along a reaction coordinate in enzymes. *J. Chem. Theory Comput.* **15**, 5116–5134 (2019).

32. C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, A standardized archaeal taxonomy for the Genome Taxonomy Database. *Nat. Microbiol.* **6**, 946–959 (2021).

33. T. Paysan-Lafosse, M. Blum, S. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunić, A. Marchler-Bauer, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, C. Rivoire, C. J. A. Sigrist, I. Sillitoe, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, A. Bateman, InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).

34. A. Tachibana, Y. Yano, S. Otani, N. Nomura, Y. Sako, M. Taniguchi, 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).

35. T. E. Petrova, K. M. Boyko, A. Y. Nikolaeva, T. N. Stekhanova, E. V. Gruzdev, A. V. Mardanov, V. S. Stroilov, J. A. Littlechild, V. O. Popov, E. Y. Bezsudnova, Structural characterization of geranylgeranyl pyrophosphate synthase GACE1337 from the hyperthermophilic archaeon Geoglobus acetivorans. *Extremophiles* **22**, 877–888 (2018).

36. M. A. Egbuta, S. McIntosh, D. L. E. Waters, T. Vancov, L. Liu, In vitro anti-inflammatory activity of essential oil and β-bisabolol derived from Cotton Gin trash. *Molecules* **27**, 526 (2022).

37. J. Durairaj, A. Di Girolamo, H. J. Bouwmeester, D. de Ridder, J. Beekwilder, A. D. van Dijk, An analysis of characterized plant sesquiterpene synthases. *Phytochemistry* **158**, 157–165 (2019).

38. P. Bansal, A. Morgat, K. B. Axelsen, V. Muthukrishnan, E. Coudert, L. Aimo, N. Hyka-Nouspikel, E. Gasteiger, A. Kerhornou, T. B. Neto, M. Pozzato, M.-C. Blatter, A. Ignatchenko, N. Redaschi, A. Bridge, Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **50**, D693–D700 (2022).

39. E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, A. Bairoch, “UniProtKB/Swiss-Prot” in *Plant Bioinformatics: Methods and Protocols*, D. Edwards, Ed. (Humana Press, Totowa, NJ, 2007), pp. 89–112.

40. BFD. https://bfd.mmseqs.com/.

41. S. Pundir, M. J. Martin, C. O’Donovan, UniProt Protein Knowledgebase. *Methods Mol. Biol.* **1558**, 41–55 (2017).

42. L. Richardson, B. Allen, G. Baldi, M. Beracochea, M. L. Bileschi, T. Burdett, J. Burgin, J. Caballero-Pérez, G. Cochrane, L. J. Colwell, T. Curtis, A. Escobar-Zepeda, T. A. Gurbich, V. Kale, A. Korobeynikov, S. Raj, A. B. Rogers, E. Sakharova, S. Sanchez, D. J. Wilkinson, R. D. Finn, MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).

43. N. Matasci, L.-H. Hung, Z. Yan, E. J. Carpenter, N. J. Wickett, S. Mirarab, N. Nguyen, T. Warnow, S. Ayyampalayam, M. Barker, J. G. Burleigh, M. A. Gitzendanner, E. Wafula, J. P. Der, C. W. dePamphilis, B. Roure, H. Philippe, B. R. Ruhfel, N. W. Miles, S. W. Graham, S. Mathews, B. Surek, M. Melkonian, D. E. Soltis, P. S. Soltis, C. Rothfels, L. Pokorny, J. A. Shaw, L. DeGironimo, D. W. Stevenson, J. C. Villarreal, T. Chen, T. M. Kutchan, M. Rolf, R. S. Baucom, M. K. Deyholos, R. Samudrala, Z. Tian, X. Wu, X. Sun, Y. Zhang, J. Wang, J. Leebens-Mack, G. K.-S. Wong, Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17 (2014).

44. D. M. Goodstein, S. Shu, R. Howson, R. Neupane, R. D. Hayes, J. Fazo, T. Mitros, W. Dirks, U. Hellsten, N. Putnam, D. S. Rokhsar, Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178–86 (2012).

45. C. A. Lesburg, G. Zhai, D. E. Cane, D. W. Christianson, Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. *Science* **277**, 1820–1824 (1997).

46. C. M. Starks, K. Back, J. Chappell, J. P. Noel, Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science* **277**, 1815–1820 (1997).

47. R. Thoma, T. Schulz-Gasch, B. D'Arcy, J. Benz, J. Aebi, H. Dehmlow, M. Hennig, M. Stihle, A. Ruf, Insight into steroid scaffold formation from the structure of human oxidosqualene cyclase. *Nature* **432**, 118–122 (2004).

48. W. L. DeLano, S. Bromberg, PyMOL user's guide. *DeLano Scientific LLC* **629** (2004).

49. Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

50. A. Barozet, P. Chacón, J. Cortés, Current approaches to flexible loop modeling. *Curr Res Struct Biol* **3**, 187–191 (2021).

51. J. Xu, Y. Zhang, How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).

52. H.-S. Park, C.-H. Jun, A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36**, 3336–3341 (2009).

53. K. R. Shahapure, C. Nicholas, "Cluster Quality Analysis Using Silhouette Score" in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (IEEE, 2020), pp. 747–748.

54. L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).

55. A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, R. Fergus, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).

56. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniProt Consortium, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

57. B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).

58. B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, R. Lanfear, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

59. K. Katoh, H. Toh, Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).

60. D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).

61. C. Cao, D. Chicco, M. M. Hoffman, The MCC-F1 curve: a performance evaluation technique for binary classification, *arXiv [stat.ML]* (2020). http://arxiv.org/abs/2006.11278.

62. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

63. M. E. Lee, W. C. DeLoache, B. Cervantes, J. E. Dueber, A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).

64. M. Lõoke, K. Kristjuhan, A. Kristjuhan, Extraction of genomic DNA from yeasts for PCR-based applications. *Biotechniques* **50**, 325–328 (2011).

65. R. D. Gietz, R. H. Schiestl, High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).

66. R. Schmid, S. Heuckeroth, A. Korf, A. Smirnov, O. Myers, T. S. Dyrlund, R. Bushuiev, K. J. Murray, N. Hoffmann, M. Lu, A. Sarvepalli, Z. Zhang, M. Fleischauer, K. Dührkop, M. Wesner, S. J. Hoogstra, E. Rudt, O. Mokshyna, C. Brungs, K. Ponomarov, L. Mutabdžija, T. Damiani, C. J. Pudney, M. Earll, P. O. Helmer, T. R. Fallon, T. Schulze, A. Rivas-Ubach, A. Bilbao, H. Richter, L.-F. Nothias, M. Wang, M. Orešič, J.-K. Weng, S. Böcker, A. Jeibmann, H. Hayen, U. Karst, P. C. Dorrestein, D. Petras, X. Du, T. Pluskal, Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.*, doi: 10.1038/s41587-023-01690-2 (2023).

67. I. Letunic, P. Bork, Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.* **52**, W78–W82 (2024).

68. X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–4 (2014).

69. W. Hwang, S. L. Austin, A. Blondel, E. D. Boittier, S. Boresch, M. Buck, J. Buckner, A. Caflisch, H.-T. Chang, X. Cheng, Y. K. Choi, J.-W. Chu, M. F. Crowley, Q. Cui, A. Damjanovic, Y. Deng, M. Devereux, X. Ding, M. F. Feig, J. Gao, D. R. Glowacki, J. E. Gonzales 2nd, M. B. Hamaneh, E. D. Harder, R. L. Hayes, J. Huang, Y. Huang, P. S. Hudson, W. Im, S. M. Islam, W. Jiang, M. R. Jones, S. Käser, F. L. Kearns, N. R. Kern, J. B. Klauda, T. Lazaridis, J. Lee, J. A. Lemkul, X. Liu, Y. Luo, A. D. MacKerell Jr, D. T. Major, M. Meuwly, K. Nam, L. Nilsson, V. Ovchinnikov, E. Paci, S. Park, R. W. Pastor, A. R. Pittman, C. B. Post, S. Prasad, J. Pu, Y. Qi, T. Rathinavelan,

D. R. Roe, B. Roux, C. N. Rowley, J. Shen, A. C. Simmonett, A. J. Sodt, K. Töpfer, M. Upadhyay, A. van der Vaart, L. I. Vazquez-Salazar, R. M. Venable, L. C. Warrensford, H. L. Woodcock, Y. Wu, C. L. Brooks 3rd, B. R. Brooks, M. Karplus, CHARMM at 45: Enhancements in accessibility, functionality, and speed. *J. Phys. Chem. B* **128**, 9976–10042 (2024).

70. W. Im, M. S. Lee, C. L. Brooks 3rd, Generalized born model with a simple smoothing function. *J. Comput. Chem.* **24**, 1691–1702 (2003).

71. J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, A. D. MacKerell Jr, CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).

72. K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, A. D. Mackerell Jr, CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31**, 671–690 (2010).

73. K. Raz, R. Driller, N. Dimos, M. Ringel, T. Brück, B. Loll, D. T. Major, The impression of a nonexisting catalytic effect: The role of CotB2 in guiding the complex biosynthesis of cyclooctat-9-en-7-ol. *J. Am. Chem. Soc.* **142**, 21562–21574 (2020).

74. A. M. Lau, S. M. Kandathil, D. T. Jones, Merizo: a rapid and accurate protein domain segmentation method using invariant point attention. *Nat. Commun.* **14**, 1–11 (2023).

Acknowledgments

We thank Peter G Mikhael, Itamar Chinn, Chi Zhang, Jitka Štáfková, Petr Kouba, and Regina Barzilay for their constant friendly support and inspiring, insightful discussions about this work.

Funding: This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90140). T.H. was supported by the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101130799 and the IOCB Postdoctoral Fellowship program. T.P. was supported by the Czech Science Foundation (GA CR) grant 21-11563M. J.S. was supported by the European Union (ERC, FRONTIER, 101097822) and (ELIAS, 101120237). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Author contributions: T.P. conceptualized the project. R.Sam. performed exploratory data analysis, developed structural domain segmentation, clustering; derived biological insights from domain analysis; built, trained, and evaluated predictive models; screened UniRef90 and UniProt; automatically mined characterized TPSs; wrote the manuscript with inputs from all authors. T.H. performed metabolic engineering, LC-MS, and GC-MS analyses; processed laboratory experimental data and wrote corresponding parts of the manuscript. T.H. and J.D.S. conducted compound purification and NMR analyses. S.C.K. wrote the manuscript and provided critical feedback to the project. T.Č. mined putative TPSs for model fine-tuning. H.S. and M.M. performed metabolic engineering. R.B., A.B., and J.K. performed fine-tuning of protein language models. R.Ch. performed error analysis. R.Ch., T.Č., R.B., and A.B. performed exploratory data analysis. D.T.M and R.Sch. performed TPS structural model refinement and docking. A.T., T.Č., R.Sam., R.Ch., M.P., and M.E. curated the TPS database. J.S. and T.P. supervised the project.

Code and data availability

The source code developed for this study is available on GitHub under the MIT license (https://github.com/pluskal-lab/EnzymeExplorer).

The predictions of EnzymeExplorer for UniRef90 clusters with sequences longer than 230 amino acid residues and without InterPro annotations are available on Zenodo under CC by 4.0 (DOI: 10.5281/zenodo.14892332).

Supplementary Materials

Materials and Methods

Dataset construction

The largest dataset of TPS reactions to date has been assembled for plant sesquiterpene (C_{15}) synthases, counting over 250 different enzymes (*37*). We expanded this collection by mining over 700 annotated TPS sequences from UniProtKB/Swiss-Prot, Rhea, and other available resources (*24*, *38*, *39*). Protein sequences were selected based on matches to TPS-specific Pfam domains (PF01397, PF03936, PF19086, PF13249, PF13243, PF01397) and SUPFAM domain combinations (sf48239 and sf48576). Hundreds of additional enzymes were manually curated from the primary literature. The final curated dataset contains 1,132 characterized TPS sequences catalyzing 2,066 reactions. It is supplemented with 79 additional sequences of isoprenoid diphosphate synthases (IDSs), which produce the precursors for terpene synthases. The total dataset size is 1,211 sequences and 2,246 reactions (**Supplementary File 1**).

To gather a dataset of putative TPS-like sequences, we mined large-scale protein sequence resources, including BFD (*40*), UniParc (*41*), Mgnify (*42*), 1KP (*43*), Phytozome (*44*) and NCBI Transcriptome Shotgun Assembly (TSA) databases. Protein sequences were selected based on matches to the same TPS-specific Pfam and SUPFAM domain combinations. This mining effort resulted in approximately 80 thousand TPS-like sequences (**Supplementary File 3**).

Structural domain segmentation and annotation

We implemented a three-step methodology to segment predicted protein structures into enzymatic domains and to annotate them with refined structural subtypes. The first step employs an alignment-based approach to partition predicted structures into family-specific enzymatic domains using established structural templates (**Fig. S6**). Next, an unsupervised clustering method based on structural similarity metrics (TM-score) identifies subtypes within these domains, leveraging characterized enzymes as reference examples. Finally, a supervised classifier trained on the curated dataset predicts domain subtypes in novel proteins, providing robust and interpretable annotations.

Structural domain segmentation

We used structural domains of characterized TPSs from prior studies (*11*) as templates for TPS-specific structural domains. The fold of pentalenene synthase (*45*) served as the template for the α domain. To define the β domain, we aligned the α-domain template to the characterized structure of 5-epi-aristolochene synthase (*46*) and used the unaligned portion of its fold as the β domain standard. Similarly, after aligning both α and β templates to taxadiene synthase (*11*), the γ domain standard was defined as the remaining unaligned region. Additionally, domains from lanosterol synthase with βγ architecture (*47*) were used as alternative templates for β and γ domains.

To segment predicted structures, we aligned them to these family-specific domain templates using sequence-independent alignment tools, PyMOL v3.0.2 (*48*) and TM-align v20170708 (*49*). Due to the challenges of modeling protein loops (*50*), alignments used only α helices and β sheets. The

TM-align algorithm outputs partial residue mappings between the template and query structure. Hence, we developed an approach (applied as a post-processing step) to pair all residues in the domain template to corresponding residues in the query structure. For each unmapped residue in the template, the nearest mapped residue pair was used to derive local sequence shifts, extending mappings to include all residues in the template. We required mapping at least 60% of the structural template for the detection and only considered residues with high confidence from AlphaFold2 predictions (pLLDT > 90). Matches were deemed valid if the TM-score exceeded 0.4, a threshold ensuring structural similarity consistent with the SCOP and CATH consensus for domains sharing the same fold (*51*). For each unassigned part, we again searched for matches with domain templates. This way αα architectures were segmented. Afterwards, unassigned helices were mapped to the closest detected domain based on physical proximity. Distance thresholds were computed as the maximum observed distances between helices within each domain template.

Domain subtype identification

After segmenting all protein structures into domains, we performed pairwise comparisons of the detected domains using TM-align (*49*) to compute TM-scores to measure structural similarity. Unsupervised machine learning was then employed to group structurally consistent domains within each established domain type. The TM-scores served as a precomputed metric for clustering using K-medoids (*52*) to define major domain types. Silhouette analysis (*53*) was performed to determine the optimal number of clusters for K-medoids. Clustering results were visualized using Principal Component Analysis (PCA), which preserves the global structure of the space of structural similarities.

To explore finer granularity within the major domain types, we applied clustering method HDBSCAN (*54*) for subtype identification. HDBSCAN automatically determines the number of subclusters and excludes outliers as noise. This combination of K-medoids for major types and HDBSCAN for subtypes enables a hierarchical view of structural domain diversity.

Domain subtype prediction

We trained a Random Forest Classifier with 500 trees to predict domain subtypes for novel protein structures. The classifier used TM-scores between a detected domain and known reference examples as input features. The dataset was divided into five folds for cross-validation (see **Performance Evaluation of Predictive Models**). Five domain subtype classifiers were trained following cross-validation. Novelty detection was performed by computing a novelty score, defined as one minus the highest predicted probability of a known domain subtype. The classifier was implemented in Python 3.10.0, using the scikit-learn library (v1.5.1).

Protein language model fine-tuning

Using the mined TPS-like sequences (**Supplementary File 3**, see **Dataset construction**), we performed an end-to-end fine-tuning of the ESM-1b model (*55*), where the original checkpoint (ESM-1v) was trained on the UniRef90 dataset (*56*). We adopted the original BERT-like masking strategy and selected 15% of the input protein sequence tokens for masking. Of these tokens, 80% were replaced with the <mask> token, 10% were replaced with a random token, and the remaining 10% were left unchanged. Unlike conventional token-wise masking, we applied masking to contiguous subsequences of randomly sampled lengths ranging from 2 to 6 tokens. We limited the input length to 1024 tokens, and for sequences exceeding this length, we randomly sampled a

subsequence of the desired length. Model was trained for 100,000 steps with Adam optimizer with inverse square root learning rate scheduler, with an initial learning rate of 1e-4 and 16,000 warm-up steps.

Predictive modeling

We extracted numerical representations of input sequences from the finetuned protein language model (PLM) and the predicted structure. Predicted protein structures were segmented into family-specific domains (see **Structural Domain Segmentation and Annotation**). We computed each detected domain's structural similarity (TM-score) against all known domains of the same type in the training dataset. The TM-scores served as a structural similarity profile for each domain. We concatenated the structural profile with PLM embeddings to obtain an input vector for predictive models. Using this input vector, we trained a binary random forest classifier to detect family-specific enzymatic activity. A multi-label random forest classifier was trained to predict substrates of detected enzymes. The pipeline was implemented in Python 3.10.0. We used the random forest model (*21*) from the scikit-learn library (v1.5.1). Feature extraction based on PLM representations was implemented using PyTorch.

Hyperparameter tuning was performed on training data via cross-validation within the training folds. Additionally, we implemented a per-class hyperparameter tuning approach to optimize model performance for each substrate class individually. Hyperparameter optimization was conducted using the scikit-optimize library (v0.10.2), leveraging Gaussian Processes (gp_minimize function).

Performance evaluation of predictive models

We evaluated the performance of predictive models using our TPS dataset (**Supplementary File 1**) as a source of positive TPS examples and the Swiss-Prot knowledgebase (*57*) as a source of negative examples. Negative samples were obtained by randomly sampling 10,000 non-TPS proteins from Swiss-Prot. The full evaluation dataset including negative examples is in **Supplementary File 3**. For each evaluated model, we conducted multiple experiments where part of the data was held out as a test set. We employed 5-fold cross-validation, in which one fold served as a hold-out dataset, untouched during model training and hyperparameter tuning, while the remaining four folds were used for training.

Stratified Group K-Fold Splitting

To ensure that models generalize to proteins distant from the training set, we used a stratified group k-fold. Groups were defined based on clades in a phylogenetic tree constructed with IQ-TREE v2.3.0 (*58*) using multiple sequence alignment (MSA) computed by MAFFT v7.526 (*59*). The average maximum BLAST identity between the hold-out sample and training set ranged from 44% to 48%. Any test sequences with a maximum train sequence identity above 60% were excluded from the evaluation.

The default implementation of stratified group k-fold in scikit-learn produced folds with unstable class proportions. To address this, we iterated over random seeds and selected the fold split that minimized the Jensen-Shannon divergence between the class distributions in each fold and the overall dataset. This ensured more balanced class distributions across folds.

Evaluation Metrics

Model performance was evaluated using mean average precision (mAP) as the primary metric. mAP summarizes the performance of score-based classifiers across different detection thresholds and is well-suited for imbalanced datasets (*60*). Additionally, we computed the area under the ROC curve (ROC-AUC) and the recently proposed MCC-F1 score, which is based on the MCC-F1 curve (*61*).

Baselines

We compared the performance of our approach against several well-established baseline models: Pfam (15) and SUPFAM (16) domain-based annotation tools for identifying conserved protein families and functional domains, a state-of-the-art deep-learning-based tool Foldseek (12) designed for protein search within the structure space of AlphaFold 2 predictions, a state-of-the-art deep-learning model CLEAN (5) for enzyme annotation based on Enzyme Commission (EC) number prediction, Profile Hidden Markov Models (pHMMs) trained using the HMMER library (48) on our curated dataset following the procedure outlined in the Terzyme paper (17), a sequence homology search tool BLASTp (18). We used official implementations with default or recommended parameters for all baselines.

Ablation study

To evaluate the contribution of input features and classifier algorithms to the overall predictive performance, we conducted two ablation studies. We assessed the role of protein language model (PLM) embeddings, structural domain features, and their combination in predicting TPS functions and substrate specificities. Models were trained using either PLM embeddings only, or structural domain features only, or a concatenation of PLM embeddings and structural features. Results demonstrated that combining PLM embeddings with structural domain features consistently achieved the highest performance for both TPS detection and substrate prediction, highlighting the complementary strengths of these feature types (**Fig. S8**). Next, we evaluated the performance of different supervised learning algorithms for TPS detection and substrate prediction. The classifiers tested included random forest, logistic regression, and feed-forward neural network. For multi-label substrate prediction, random forest outperformed other classifiers across most substrate classes (**Fig. S9)**. Logistic regression achieved the highest performance for TPS detection, with random forest showing comparable results.

Selecting sequences for experimental validation

Mining and Prioritizing Phylogenetically Distant TPS Sequences with Pfam/SUPFAM domains

We screened large-scale protein sequence databases BFD, UniParc, MGnify, 1KP, Phytozome, and NCBI Transcriptome Shotgun Assembly (TSA), encompassing 5.1 billion sequences. Protein sequences were filtered based on TPS-specific Pfam domains (e.g., PF01397, PF03936) and SUPFAM domain combinations (sf48239 and sf48576), resulting in a reduced set of ~200,000 TPS-like sequences. Then, an MSA of the filtered putative and characterized TPS sequences was created using MAFFT v7.526 (*59*). The MSA was trimmed using trimAl to remove poorly aligned regions. A phylogenetic tree was constructed using FastTree2 (*62*). For each uncharacterized TPS-like sequence, the phylogenetic distance to the closest characterized TPS sequence was calculated as the sum of branch lengths in the tree. Sequences with the greatest phylogenetic

distance from known TPSs were selected as high-priority candidates, assuming they represent novel functions or types..

Mining TPS-like Sequences from UniRef90 lacking InterPro signatures

We scanned UniRef90 (*56*), containing 19.9 × 10^7 representative protein sequences. We have excluded sequences already annotated with InterPro signatures to focus on uncharacterized sequences. Sequences shorter than 230 amino acids were also excluded, leading to 8.9 × 10^6 sequences selected for screening. Sequences were filtered using our TPS detection pipeline, incorporating PLM embeddings and structural predictions for proteins with an existing predicted structure in the AlphaFold database, to identify TPS-like candidates. With the threshold for EnzymeExplorer prediction at 0.35 the screening resulted in 1,773 TPS-like sequences. The final set of candidates was manually curated based on two criteria: high detection confidence scores generated by the predictive pipeline and significant phylogenetic distance from known TPS sequences, ensuring novelty and diversity among selected candidates.

Yeast strains and cultivation

Chemicals used for media preparation were purchased from either Sigma-Aldrich, Duchefa Biochemie, Lach:ner or Penta chemicals. Solvent for metabolite extract preparation, LC- and GC-MS were purchased from Fisher Chemical and were of LC-MS grade.

The list of all strains used in the study is available in **Table S5**. *Saccharomyces cerevisiae* JWY501 derivative strains were used to express terpene synthases and produce terpenes.(*26*) Selective medium (SCE) used to grow transformants contained 1.92 g/L Yeast Synthetic Drop-out Medium Supplements without Uracil (Sigma-Aldrich, Y1501), 6.7 g/L Yeast nitrogen base without amino acid (Sigma-Aldrich, Y0626), and 20 g/L glucose.

DH10β electrocompetent *Escherichia coli* cells were used for all cloning experiments. Transformed cells were selected on Lysogeny Broth (LB) with the appropriate antibiotics (ampicillin or chloramphenicol). SOC was used for recovery after electroporation.

For maintenance *S. cerevisiae* strains were cultivated in solid selective medium at 30 °C. For general preculture, *S. cerevisiae* strains were cultivated in a liquid selective medium (SCE) at 30 °C, 200 RPM in an orbital shaker. For the production run, *S. cerevisiae* strains were cultivated in selective medium SCE in 24 deep well plates (CR1426, Enzyscreen) sealed with AeraSeal (Excel Scientific), at 30 °C, 800 RPM on an Eppendorf ThermoMixer C (Eppendorf).

For plasmid amplification, *E. coli* cultures were cultivated in LB medium in 24 deep well plates (CR1426, Enzyscreen) sealed with AeraSeal (Excel Scientific), at 37 °C, 800 RPM on an ThermoMixer C (Eppendorf).

Yeast plasmids

The list of all plasmids used in the study is available in **Table S6.** All pTP plasmids were generated using GoldenGate assembly based on the MoClo-YTK toolkit and overhangs (*63*). Part plasmids use pYTK001 as a backbone. Terpene synthases DNA parts were synthesized by Twist Bioscience as gene fragments. Cassette plasmids for genomic integration use pYTK096 as a backbone.

Polymerase chain reactions

For colony PCR and yeast genotyping we used Phire Green Hot Start II PCR Master Mix (Thermo Fisher Scientific) with the following conditions:

In 10 µL final volume, Phire Green Hot Start II PCR Master Mix 5 µL, 25 µM forward primer 0.2 µL, 25 µM reverse primer 0.2 µL, DNA template 4.6 µL. Reactions were conducted in ProFlex 3 × 32-well PCR System thermocycler (Waltham, Massachusetts, United States). Thermocycling conditions used the following template: 98°C for 2 min as initial denaturation, for 30 cycles: 98 °C for 10 s, annealing temperature for 10 s, 72 °C for 10 s.kb-1 and a final extension 72 °C for 2 minutes. PCR products were separated on agarose gel (0.8 % w/v), 130 V, 30 min.

E. coli colonies were selected with a toothpick and spotted 4 times on selective media. The remaining bacteria were thus resuspended in 10 µL ddH_2O, and boiled for 10 minutes before being used as a DNA template.

Yeast genotyping was adapted from Lõoke et al. (*64*). *S. cerevisiae* colonies were selected with a toothpick and resuspended in 100 µL 200 mM LiOAc, 1 % SDS solution, and boiled for 10 minutes. Then 300 µL of EtOH 96 % were added and the solution was a vortex and centrifuge 15 000 × g for 3 minutes. The supernatant was discarded and the pellet washed with 500 µL EtOH 70% before centrifugation 15 000 × g for 1 min. The supernatant was discarded and the pellet dried for 1 minute at room temperature. The precipitated DNA was dissolved in 100 µL ddH_2O and cell debris spined down 15 000 × g for 1 min.

Golden Gate assembly reaction

Part plasmids were generated in 10 µL reaction volume. T4 ligase buffer 1 µL, T4 ligase 0.5 µL (M0202L, New England Biolabs), BsmBI-v2 0.5 µL (R0739L, New England Biolabs), pYTK001 0.5 µL, DNA part 0.5 µL (20 fm), ddH2O up to 10 µL. Reactions were conducted in ProFlex 3 x 32-well PCR System thermocycler (Thermo Fisher Scientific). Thermocycling conditions used the following template: for 25 cycles, 42 °C for 2 min, 16 °C for 2 min, then 60 °C for 30 min and 80 °C for 10 min.

Cassette plasmids were generated in 10 µL reaction volume. T4 ligase buffer 1 µL, T4 ligase 0.5 µL (M0202L, New England Biolabs), BsaI-HFv2 0.5 µL (R3733L, New England Biolabs), DNA parts 0.5 µL (20 fm), ddH2O up to 10 µL. Reactions were conducted in ProFlex 3 × 32-well PCR System thermocycler (Thermo Fisher Scientific). Thermocycling conditions used the following template: for 25 cycles, 37 °C for 5 min, 16 °C for 5 min, then 60 °C for 30 min and 80 °C for 10 min.

***E. coli* transformation**

Electrocompetent cuvettes were cooled down at 4°C 30 min before the experiment. Electrocompetent 20 µL of E. coli cells were thawed at 4°C 10 min before the experiment. Once thawed, 0.5 µL of plasmid was added to the cell with gentle mixing. The electroporator was set to 1700 V. For chloramphenicol, spectinomycin and kanamycin selective markers, cells recovered in 1 mL of SOC for 1 h at 37 °C, 200 RPM. Thus, cells were concentrated to 100 µL through centrifugation 5000 × g for 3 min and plated to their respective LB + selection marker. In case of

ampicillin, cells were resuspended into 100 µL of SOC after electroporation and directly plated in LB + ampicillin. Cells were then grown overnight at 37 °C.

Yeast transformation

Budding yeast strains (**Table S5**) were generated by standard lithium acetate transformation protocol (*65*) and selected using auxotrophy. For genomic integration, 500-1500 ng of Plasmid were digested using NotI-HF in 10 µL final volume. For plasmid integration, 500-1500 ng of plasmid were used. Cells were precultured overnight in YPD medium, 30 °C, 200 RPM. Then, cells were diluted to OD 0.1 and cultivated in YPD medium, 30 °C, 200 RPM until they reached OD 0.5. For 5 mL of cells at OD 0.5: the cells were pelleted, 2500 × g, 5 min and washed in 5mL sterile sorbitol wash buffer (600 mM sorbitol, 100 mM K_2HPO_4). Cells were pelleted again 2500 × g, 5 min and resuspended in 1 mL sterile sorbitol wash buffer, and pelleted again 16 000 × g, 1 min and resuspended in 1 mL TE/LiAc. Then, cells were concentrated in 100 µL TE/LiAc and 10 µL of digested plasmid and 10 µL of Salmon sperm were added, mixed gently and incubated for 10 min at room temperature. Then, 260 µL of TE/LiAc/PEG (40 % w/v) was added and the cells were incubated for 45 min at 42 °C. Cells were then centrifuged 6000 × g for 1 min and washed in water before plating on selective medium. Cells were grown for 3-5 days at 30 °C.

Metabolite extraction

Cells were maintained in solid selective medium and preculture in 1 mL selective medium in 96 deep well plate (CR1496, Enzyscreen) sealed with AeraSeal (Excel Scientific), at 28 °C, 1500 RPM on an ThermoMixer C (Eppendorf) for 1 day. Cells were seeded at OD 0.05 in 2.2 mL in inducible selective medium SCE (10 % glucose, 90 % galactose) in 24 deep well plates (CR1426, Enzyscreen, Hamburg, Germany, Netherlands) sealed with AeraSeal (Excel Scientific), at 28 °C, 300 RPM for 48 hours. Then, 1 mL of ethyl acetate (E196-4, Fisher Scientific) was added to the culture medium and mixed for 1 hour at 28 °C, 250 RPM. Then 1 mL of ethyl acetate was added and thoroughly mixed by pipetting. The sample was collected in a 2 mL round bottom tube (Eppendorf) centrifuged for 5 mins, 14 100 × g and the organic phase was carefully collected and dried under N2 flow.

LC-MS analyses

For LC-MS analysis, samples were resuspended in 100 µL ethyl acetate. LC-MS analyses were performed using Vanquish Flex UHPLC System interfaced to an Orbitrap ID-X Tribrid mass spectrometer, equipped with heated electrospray ionization (Thermo Fisher Scientific). The LC conditions were as follows: column, Waters BEH (Ethylene Bridget Hybrid) C18 50 × 2.1 mm, 1.7 µm; mobile phase, (A) water with 0.1 % formic acid; (B) acetonitrile with 0.1% formic acid; flow rate, 350 µL.min-1; column oven temperature, 40 °C, injection volume, 1 µL, linear gradient of 5 to 100 % B over 5 min and isocratic at 100 % B for 2 min. Electrospray ionization was achieved in positive mode and mass spectrometer parameters were as follows: ion transfer tube temperature, 325 °C, auxiliary gas flow rate 10 L.min-1, vaporizer temperature 350 °C; sheath gas flow rate, 50 L.min-1; capillary voltage, 3000 V, MS resolution 60 000, quadrupole isolation, scan range from m/z 100-1000, RF Lens 45 %, maximum injection time 118 ms. LC-MS .raw data files were directly imported into mzmine 4.10.0 (*66*). Extracted ion chromatograms for compounds of interest were generated using the raw data overview feature and exported as .pdf.

GC-MS analyses

For GC-MS analysis, samples were resuspended into 100 µL ethyl acetate. GC-MS analyses were performed using a 7890A gas chromatograph coupled with a 5975C mass spectrometer, equipped with electron ionization (EI) and quadrupole analyzer (Agilent Technologies). The samples (1 µl) were injected into a split/splitless inlet in split mode (split ratio 10:1). The injector temperature was 250 °C. A DB-1ms fused silica capillary column (30 m × 250 µm; a film thickness of 0.25 µm, J&W Scientific) was used for separation. The carrier gas was helium at a constant flow rate of 1.0 ml/min. The temperature program was: 40 °C (1 min), then 5 °C.min-1 to 100 °C, followed by 15 °C.min-1 to 230 °C . The temperatures of the transfer line, ion source and quadrupole were 320 °C, 230 °C and 150 °C, respectively. EI spectra (70 eV) were recorded from 25 to 500 m/z. GC-MS .dx files were analyzed using OpenLab CDS 2.4. Extracted ion chromatograms for compounds of interest were exported as .csv files and built using the ggplot2 package in R. Figures were generated using Rstudio and the following packages: ggplot2, gridExtra, patchwork, dplyr, forcats, ggthemes, ggprism, DescTools, tidyverse, scales and Adobe Illustrator CS6. Peaks for sesquiterpene (m/z 204) and diterpene (m/z 272) were retrieved from extracted ion chromatograms using OpenLab CDS 2.4. The MS spectra from each peak were therefore queried to NIST EI search software, with NIST 2023 EI database. Both spectrum from the experimental data and its best match on the NIST EI database were exported as .MSPEC files. MSPEC files were converted to mgf files using a python script and imported into mzmine 4.10.0 and compared as a mirror plot (*66*).

Isolation of compound 12

Yeast strain JWY501 (*26*) carrying plasmid TP0255µ was cultivated in 8 L of selective medium with 10 % glucose and 90 % galactose for autoinduction. After 3 days at 30 °C, 200 RPM, supernatant was gathered after 30 min centrifugation, 5000 × g at 4 °C and extracted using cyclohexane at room temperature and the organic layer was dried *in vacuo*. Extract was first fractionated using normal phase flash chromatography with silica gel using n-hexane with increasing amounts of EtOAc as the mobile phase. Fractions containing terpenes were then further fractionated on normal phase semi preparative chromatography with Hexane:EthylAcetate solvent to yield pure terpene **12** (1 mg).

NMR spectroscopy

NMR spectra of compound **12**, 1H and 13C APT, have been acquired on a Bruker Avance III 600 MHz spectrometer equipped with 1.7 mm micro-cryoprobe, using CDCL3 as solvent.

Taxonomic analysis of archaea

Assemblies of Archaea genomes were retrieved from the NCBI Genome dataset, filtered for Archaea, as `protein.faa.gz` (accession date June 14, 2024). Then we screened them using our predictive pipeline. Then, the hits were clustered using CD-HIT with a threshold of 0.9 and mapped into ar53.tree (version 220) using Interactive Tree of Life (*67*).

Multiple sequence alignment

Multiple sequence alignment was done using MAFFT v7.526 (*59*) with default settings. Multiple sequence alignment was visualized using espript 3.0 (*68*).

Archaeal TPS structure analysis and ligand docking

To create a structure model of A0A5E4I9B1, we employed AlphaFold 3 (*29*). The model was generated using the A0A5E4I9B1 sequence and three Mg^{2+} ions as cofactors. Attempts to include the diphosphate moiety during modelling resulted in poor models. In A0A5E4I9B1, the well-known Mg_B and Mg_C metal binding motif is ^{221}**N**TLN**T**WPR**E**. In the original AlphaFold 3 model, the N/T/E residues of this motif do not interact with the Mg_B and Mg_C ions, and hence we refined the model. To this end, we first docked the diphosphate moiety to complete the polar region of the binding site prior to modelling the AlphaFold 3 model. The diphosphate moiety was docked into its standard binding pocket (*9*) using Nuclear Overhauser Effect (NOE) restraints between the Mg^{2+} ions and the diphosphate oxygens, as implemented in EnzyDock (*30*, *31*). EnzyDock is a CHARMM-based multistate, multiscale docking program for enzymes (*69*). Subsequently, we refined the AlphaFold 3 model using a CHARMM-based minimization protocol designed to create an active holo-form of TPS. The protein was fixed except for the helix hosting the ^{221}**N**TLN**T**WPR**E** motif and neighboring loop regions (i.e., the flexible residues were 195-237) and the diphosphate-$(Mg^{2+})_3$ cluster were flexible. We applied a series of TPS-specific NOE restraints during stepwise minimization of the AlphaFold 3 model to form the conserved TPS interactions between A0A5E4I9B1 and the diphosphate-$(Mg^{2+})_3$ cluster. The protocol included 1,000 steps of steepest descent and 500 steps of adopted basis Newton-Raphson minimization. Hydration was included via the continuum GBSW model (*70*).

To model reaction states in A0A5E4I9B1, we employed EnzyDock. These docking simulations were performed using the above-mentioned refined AlphaFold 3 model, which includes the protein, the diphosphate moiety, and three Mg^{2+} ions. The cationic isoprenoid tail of FPP and the main product β-bisabolol were docked into the active site. The substrate was docked as an (E,E)-farnesyl cation and a covalent bond was formed between its C1 position and either the O1α or O2α oxygens of the diphosphate on the fly during docking. Binding to O1α was favorable (**Table S4**), as has been observed for plant TPS (*30*). *β*-bisabolol was docked in all four diastereomeric forms. When docking FPP, a water molecule was first docked in a position between D63 and the diphosphate moiety. The native protein, cofactors, and ligands were described by the CHARMM 36 (*71*) and CGenFF force fields (*72*). The nonbonded interactions of the enzyme-ligand complexes were treated using a grid potential generated around the centre of the active site of the protein for both van der Waals and electrostatic interactions. The centre of the grid was taken as the centre of the active site. In the present study, the grid was calculated with a grid spacing of 0.25 Å and dimensions of 15×15×15 $Å^3$. The size of the grid was chosen so that it encompasses the active site.

To compute the matching ligand states along the reaction path (*73*), we applied Pathfinder that comes with EnzyDock, and matches neighbouring reaction states (e.g., FPP → *β*-bisabolol) using RMSD criteria.

Supplementary Figures

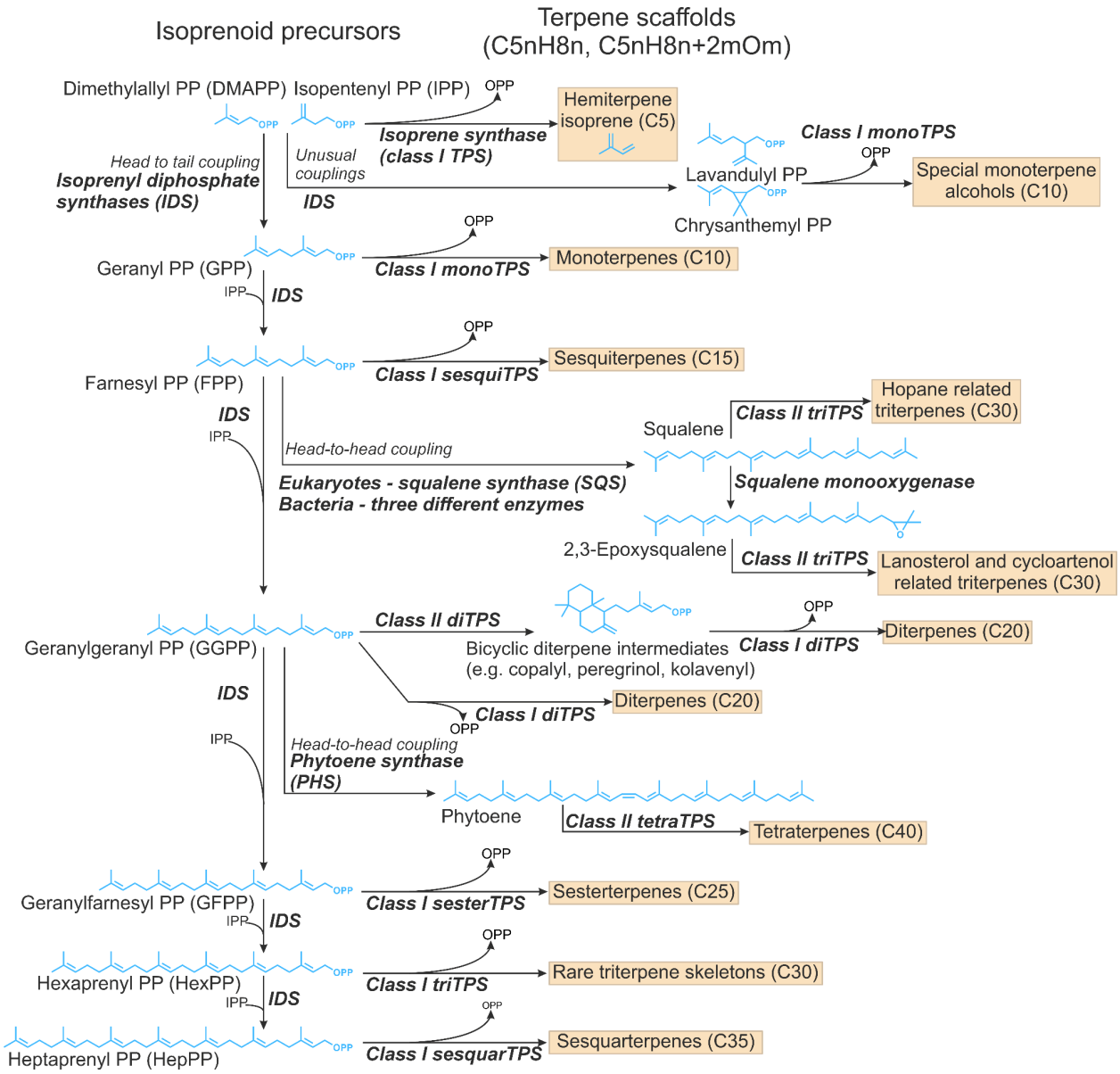


Fig. S1. Summary of TPS reaction types, substrates and product classes. TPS substrates are synthesized by isoprenyl diphosphate synthases, enzymes that catalyze the sequential head-to-tail condensation of isoprene units (C5) in the form of isopentenyl diphosphate. By the number of isoprene units, terpenes are divided into types (hemi, mono, sesqui, di, etc.) where monoterpenes are composed of two isoprene units. Similar enzymes also catalyze other manners of condensations of isoprenyl diphosphates into products like squalene. The cyclization of those substrates into terpene products is guided by terpene synthases. Class I terpene synthases initiate reactions by abstracting the diphosphate group while class II terpene synthases initiate their reactions by protonation of a terminal alkene or of an epoxide group.

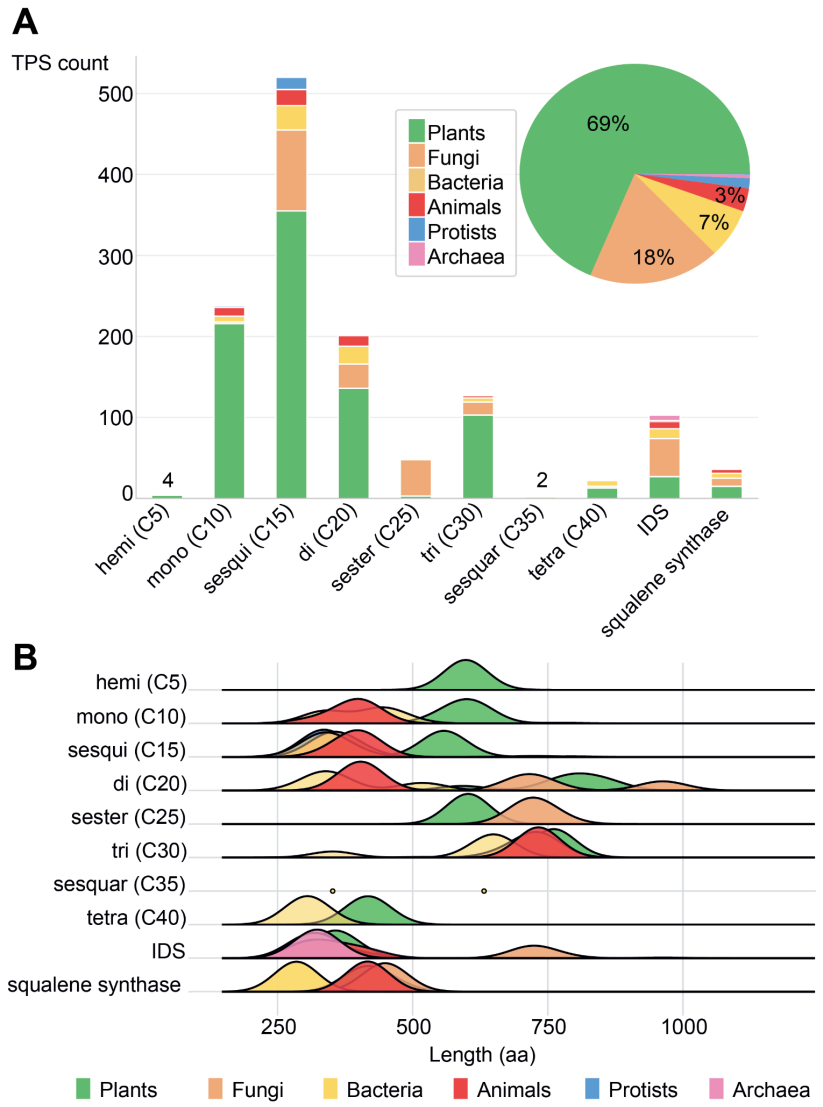


Fig. S2. Descriptive statistics of the curated TPS dataset. (A) Plot showing the number of reactions catalyzed by TPSs in our dataset (**Supplementary File 1**) divided into TPS types and taxonomic kingdoms of the source organisms. Note that in cases where a single enzyme catalyzes multiple types of reactions (e.g., sesquiTPS and diTPS), such an enzyme is counted multiple times in corresponding columns. **(B)** The distribution of protein sequence length (amino acid residues) per TPS types and kingdoms.

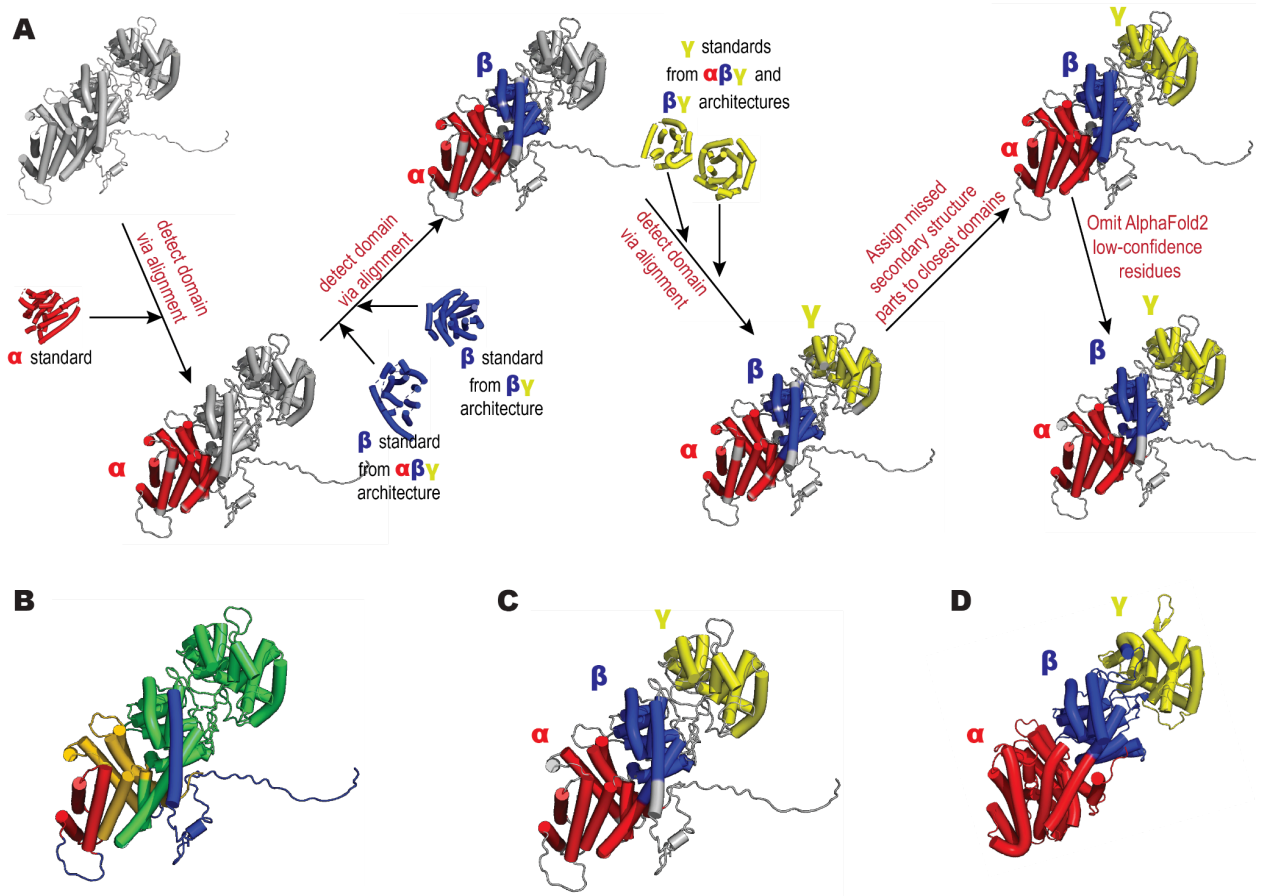


Fig. S3. The workflow for protein structure segmentation into domains. (A) A high-level overview of our pipeline for structure segmentation into domains. The pipeline uses alignment to known structural templates to partition predicted protein structures into distinct domains. **(B)** An example of state-of-the-art (SOTA) general segmentation (*74*) failing to partition the structure of a TPS, demonstrated on a randomly selected TPS with UniProt accession B9GSM9. Moreover, the SOTA method does not assign domain types to the segmented regions, resulting in incomplete domain characterization. **(C)** Segmentation result using our algorithm for the same UniProt accession B9GSM9. The TPS with an αβγ architecture is correctly segmented, with domains assigned to their corresponding types: α, β, and γ. **(D)** Ground truth segmentation of a TPS with αβγ architecture (UniProt accession Q41594) into CATH domains: 3p5rA01, 3p5rA02, and 3p5rA03.

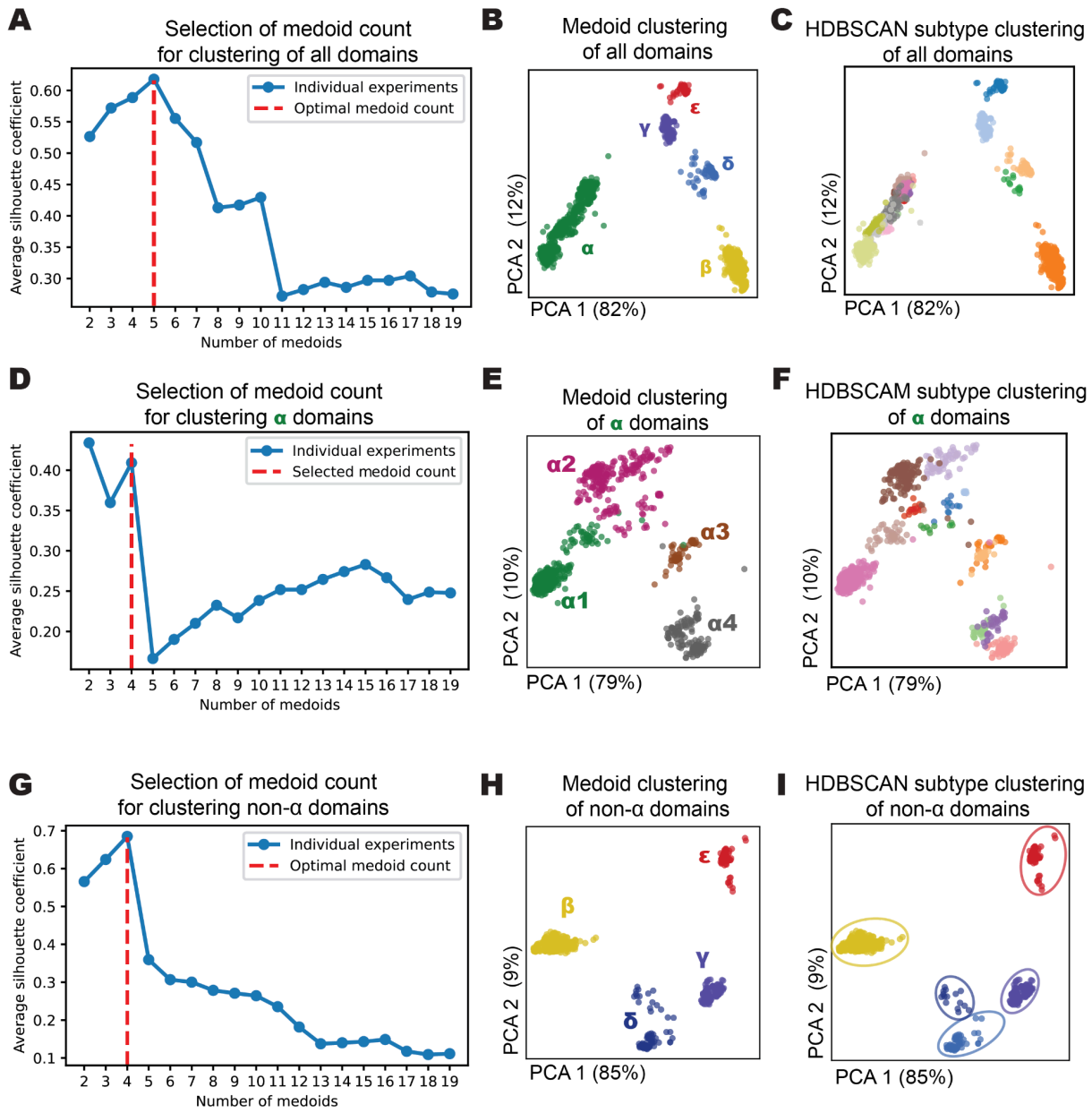


Fig. S4. Determining domain clusters and subtypes. (A, D, G) Silhouette analysis for determining the number of clusters for global domain clustering **(A)**, α domain clustering **(D)**, and β/γ domain clustering **(G)** using the K-medoids algorithm. The red dashed line indicates the selected medoid count based on silhouette scores, identifying major domain clusters.

(B, E, H) Principal component analysis (PCA) visualization of the K-medoids-derived clusters using the TM-score matrix of pairwise domain comparisons. **(B)** Global clustering reveals five major types of TPS-specific domains (α, β, γ, ε, δ). **(E)** The α domain cluster, the largest group, displays a complex internal structure with distinct types (α1, α2, α3, α4), derived using the K-medoids algorithm. **(H)** β and γ domains form well-separated clusters with minimal internal structure.

(C, F, I) Visualization of subtypes derived using HDBSCAN. **(C)** The detected subtypes provide additional granularity to the medoid clusters. **(F)** Subtypes within the α cluster reveal the complex hierarchical organization of α domains. HDBSCAN detected 13 distinct subtypes. **(I)** Subtypes for non-α domains indicate less structural diversity than the α domain.

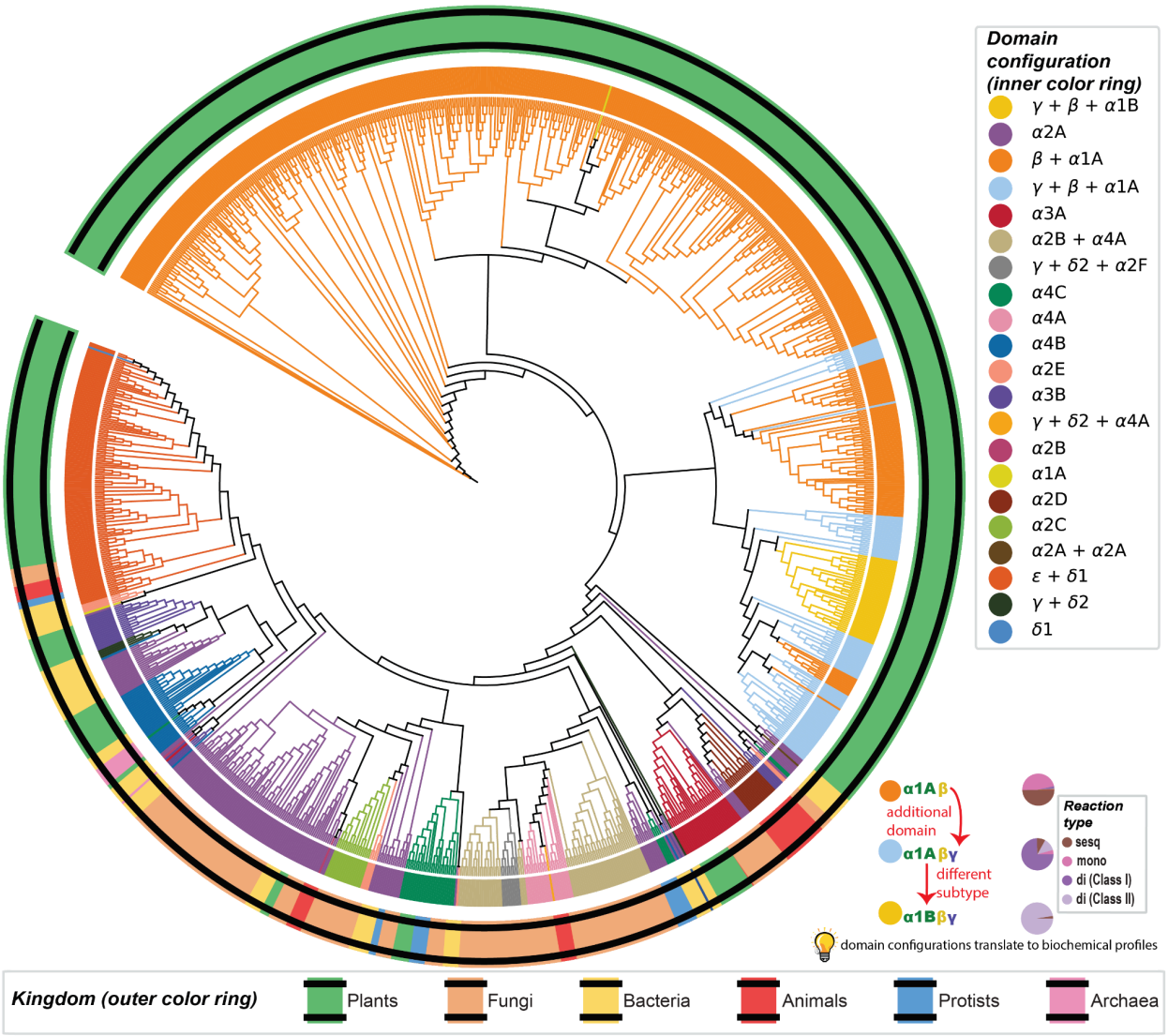


Fig. S5. Domain subtypes mapped onto the phylogenetic tree of characterized TPSs.
The figure highlights the consistency of discovered domain configurations with the evolutionary relationships of corresponding enzymes. The inner color ring represents the identified domain configurations for the TPSs in our dataset (**Supplementary File 1**), while the outer color ring denotes the corresponding kingdoms. Comparison with **Fig 1F** illustrates how changes in domain configurations align with modifications in the biochemical activity profiles of TPSs. This analysis emphasizes the evolutionary and functional significance of the discovered domain architectures. For instance, plant TPSs with the configuration α1A-β-γ are capable of diTPS synthesis, unlike those with the domain configuration α1A-β. Yet, enzymes with the configuration α1A-β-γ predominantly catalyze di class I TPS reactions, while α1B-β-γ catalyze di class II TPS reactions (see **Supplementary Fig. 1** for the definitions of the TPS classes).

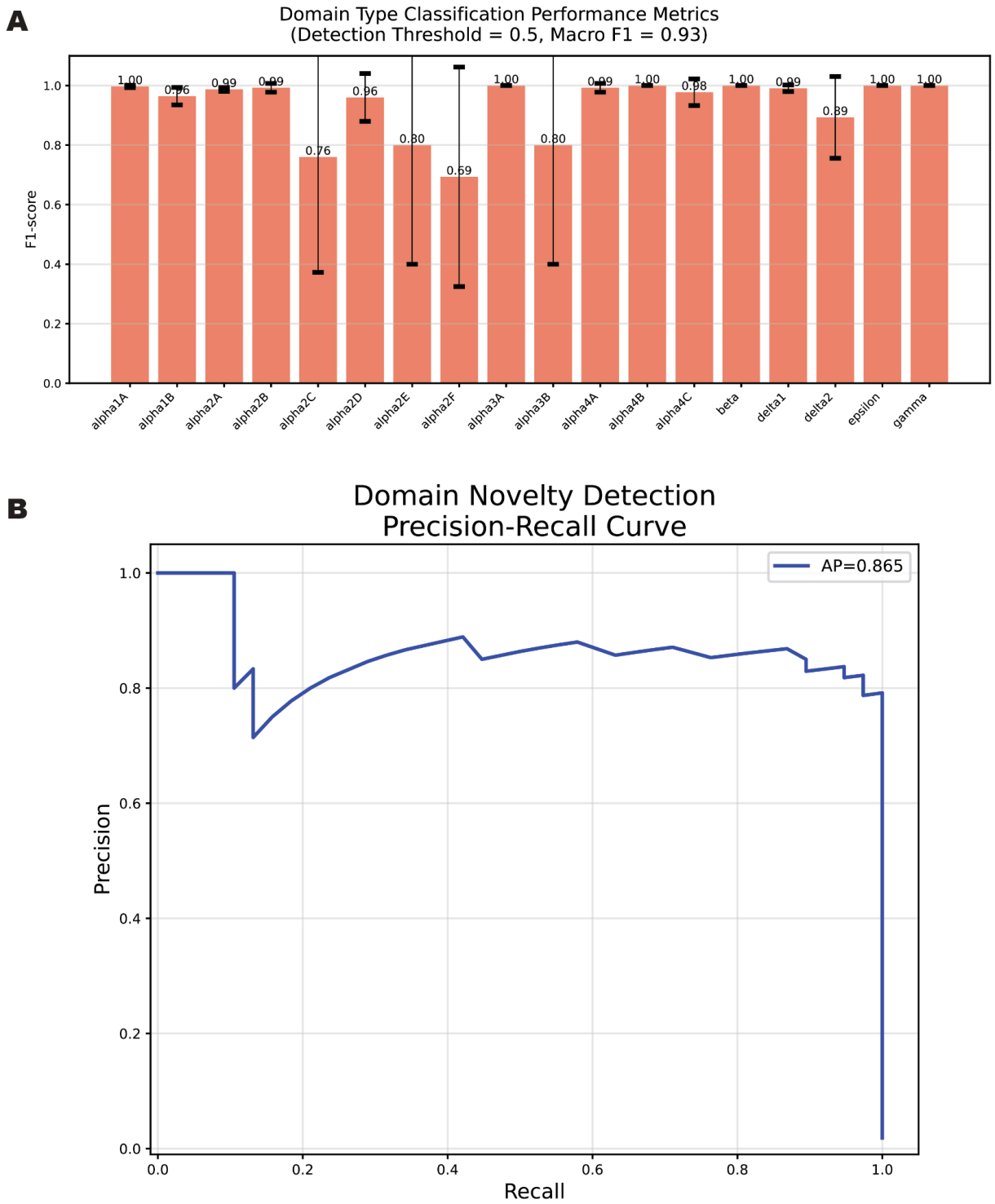


Fig. S6. Performance analysis of domain subtype classification and novel domain subtype detection. **(A)** Per-class F1-scores with standard deviations for classifying newly discovered domains into known subtypes. **(B)** Precision-recall curve for detecting novel domain subtypes, with an average precision (AP) of 0.865.

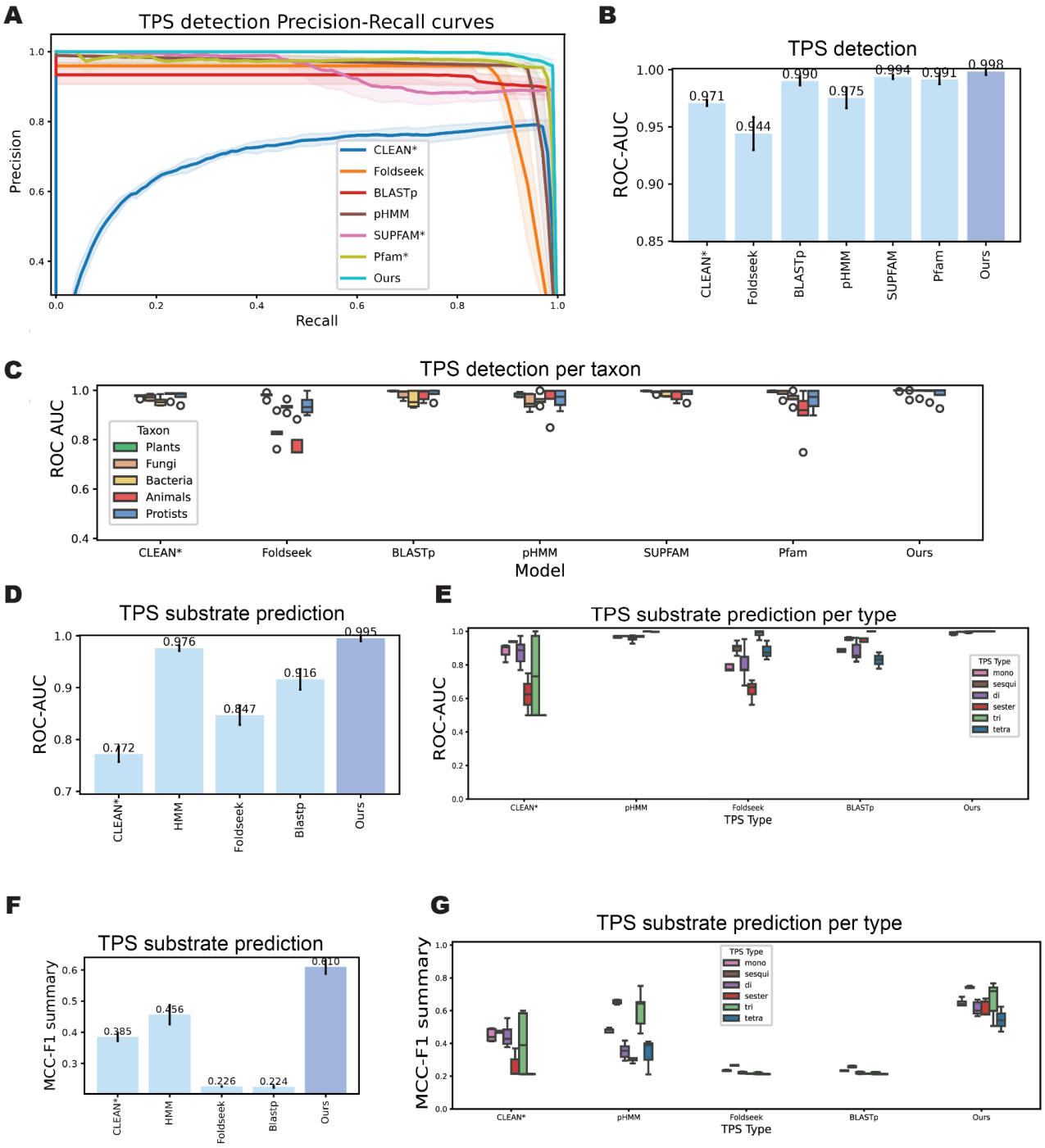


Fig. S7. Comprehensive evaluation of TPS detection and substrate prediction models.
(A) Precision-recall (PR) curves for TPS detection models, comparing CLEAN*, Foldseek, BLASTp, pHMM, SUPFAM*, Pfam*, and our approach. The shaded areas represent variability across folds.
(B) Average ROC-AUC values summarizing overall TPS detection performance across models, with our approach achieving the highest scores. **(C)** Taxon-specific TPS detection performance evaluated using ROC-AUC. Our approach consistently outperforms other methods across all taxonomic groups, including underrepresented groups such as Animals and Protists. **(D)** Overall substrate prediction performance measured using ROC-AUC. **(E)** TPS substrate prediction performance across types (mono-, sesqui-, di-,

sester-, tri-, and tetraterpenes) using ROC-AUC. **(F)** TPS substrate prediction performance summarized using the MCC-F1 score. **(G)** Substrate prediction performance per TPS type using the MCC-F1 score. Our approach consistently achieves superior performance across all metrics and TPS types, particularly excelling in underrepresented TPS types like tetraterpenes.

Panels **D** and **E** include error bars representing the standard error of the mean (SEM), calculated from aggregated variance across folds. In panels **B, D,** and **F,** methods CLEAN, SUPFAM, and Pfam were used as off-the-shelf methods (as they would be typically used) without training on our dataset and hence are marked with *.

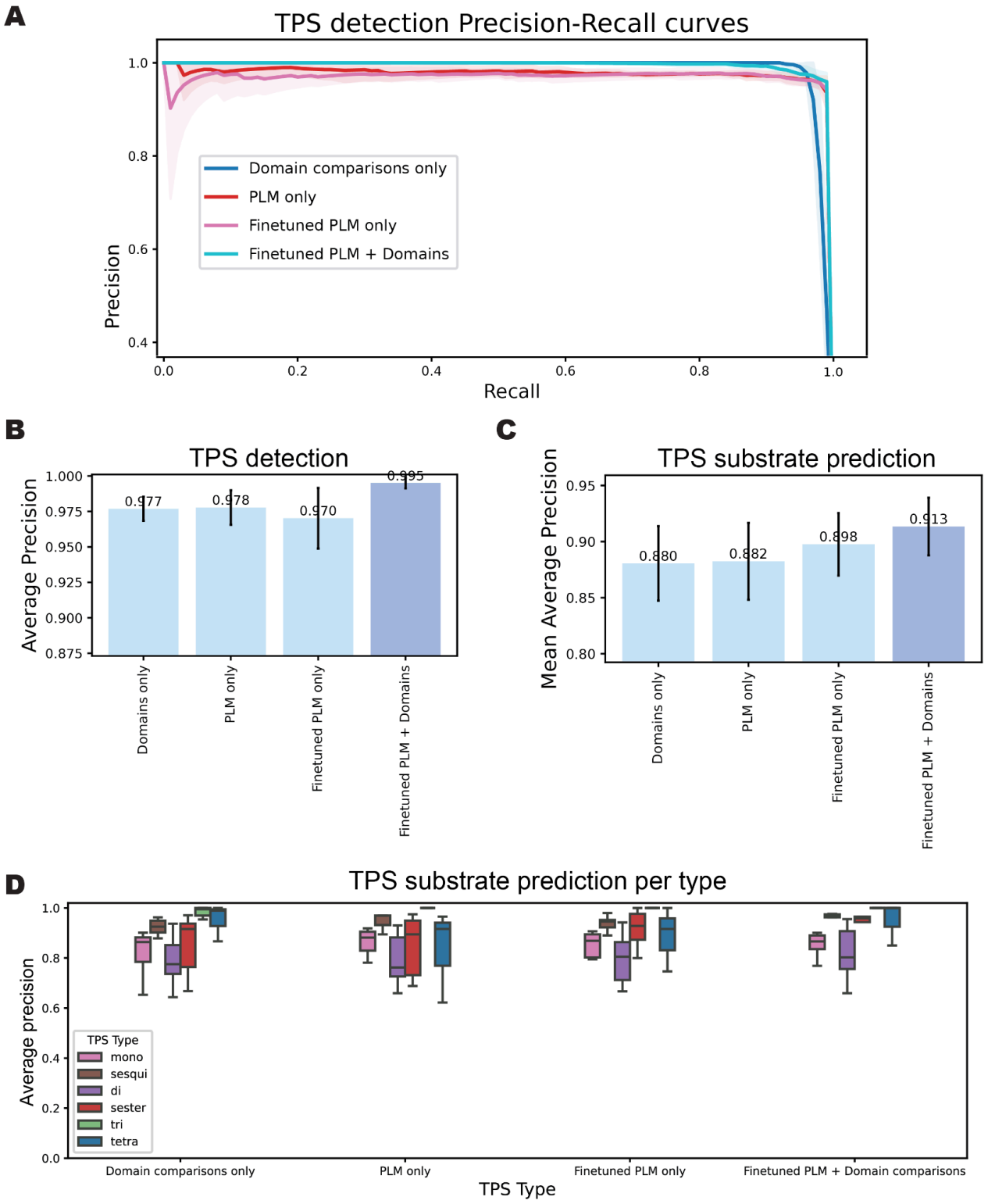


Fig. S8. Ablation study evaluating the impact of different input features on TPS detection and substrate prediction.

(A) Precision-recall (PR) curves comparing the performance of models using different input features: structural domain comparisons only ("Domains only"), protein language model (PLM) embeddings ("PLM only"), fine-tuned PLM embeddings ("Finetuned PLM only"), and the combination of fine-tuned PLM

embeddings with structural domain comparisons (“Finetuned PLM + Domains”). Shaded areas represent variability across folds.

(B) Average precision (AP) scores for TPS detection across different feature sets. Models combining fine-tuned PLM embeddings with structural domain comparisons (“Finetuned PLM + Domains”) achieve the highest AP scores. Error bars represent the standard error of the mean (SEM), computed from aggregated variance across folds.

(C) Mean average precision (mAP) scores for TPS substrate prediction, showing similar trends with the best performance achieved when fine-tuned PLM embeddings are combined with structural domain comparisons (“Finetuned PLM + Domains”). Error bars represent SEM.

(D) TPS substrate prediction performance across TPS types using different input features. The combination of fine-tuned PLM embeddings with structural domain comparisons (“Finetuned PLM + Domains”) achieves consistently high performance across all TPS types, particularly excelling in underrepresented classes such as tetraTPS.

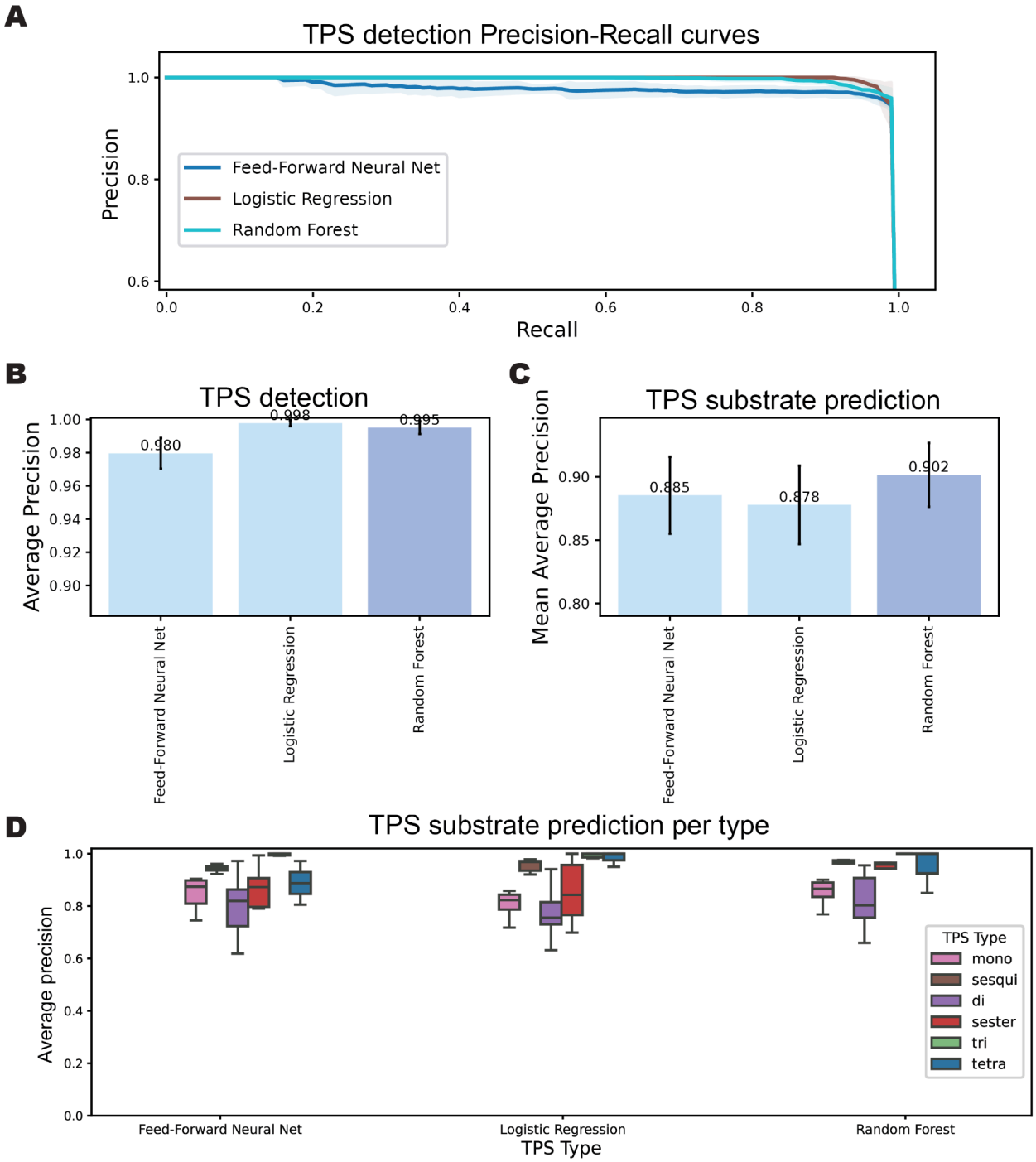


Fig. S9. Ablation study evaluating the impact of different classifier algorithms for TPS detection and substrate prediction. (A) Precision-recall (PR) curves comparing the performance of different classifiers: feed-forward neural network, logistic regression, and random forest. All models use the combined input features of structural domain comparisons and fine-tuned protein language model (PLM) embeddings. Shaded areas represent variability across folds.

(B) Average precision (AP) scores for TPS detection across classifier algorithms. The random forest classifier achieves the highest AP scores. Error bars represent the standard error of the mean (SEM), computed from aggregated variance across folds.

(C) Mean average precision (mAP) scores for TPS substrate prediction, highlighting the superior performance of the random forest classifier compared to other approaches. Error bars represent SEM.

(D) TPS substrate prediction performance across TPS types using different classifiers. The random forest classifier consistently demonstrates strong performance across all TPS types

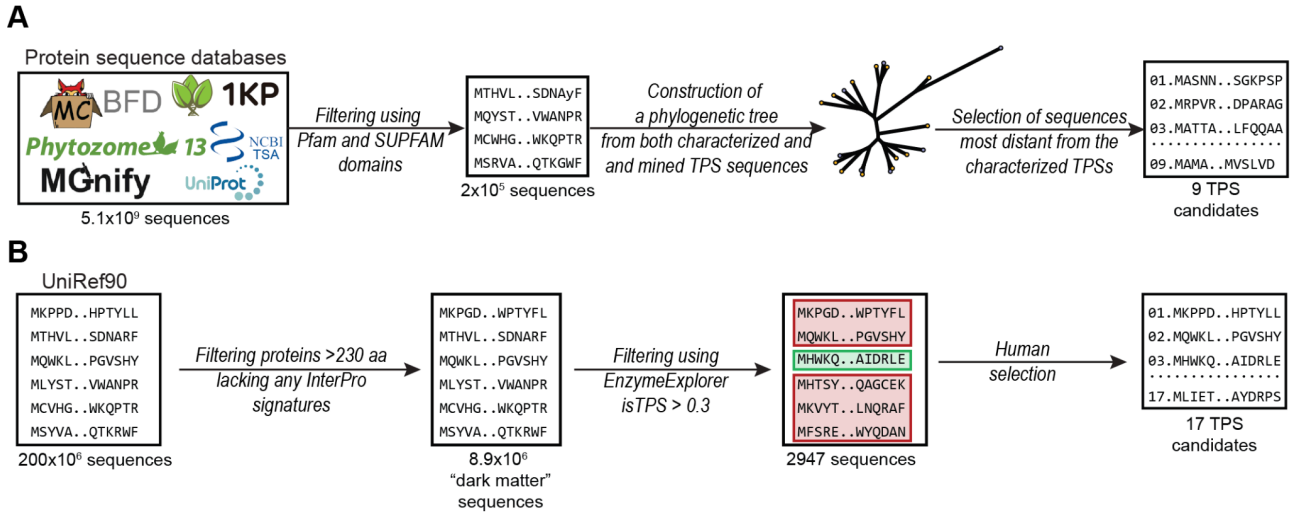


Fig. S10. Selection of candidate TPS sequences for experimental expression. (A) In silico screening workflow for selecting evolutionary distant uncharacterized TPS with TPS-specific Pfam/SUPFAM domains. **(B)** In silico screening workflow for selecting TPS that lack any InterPro signatures.

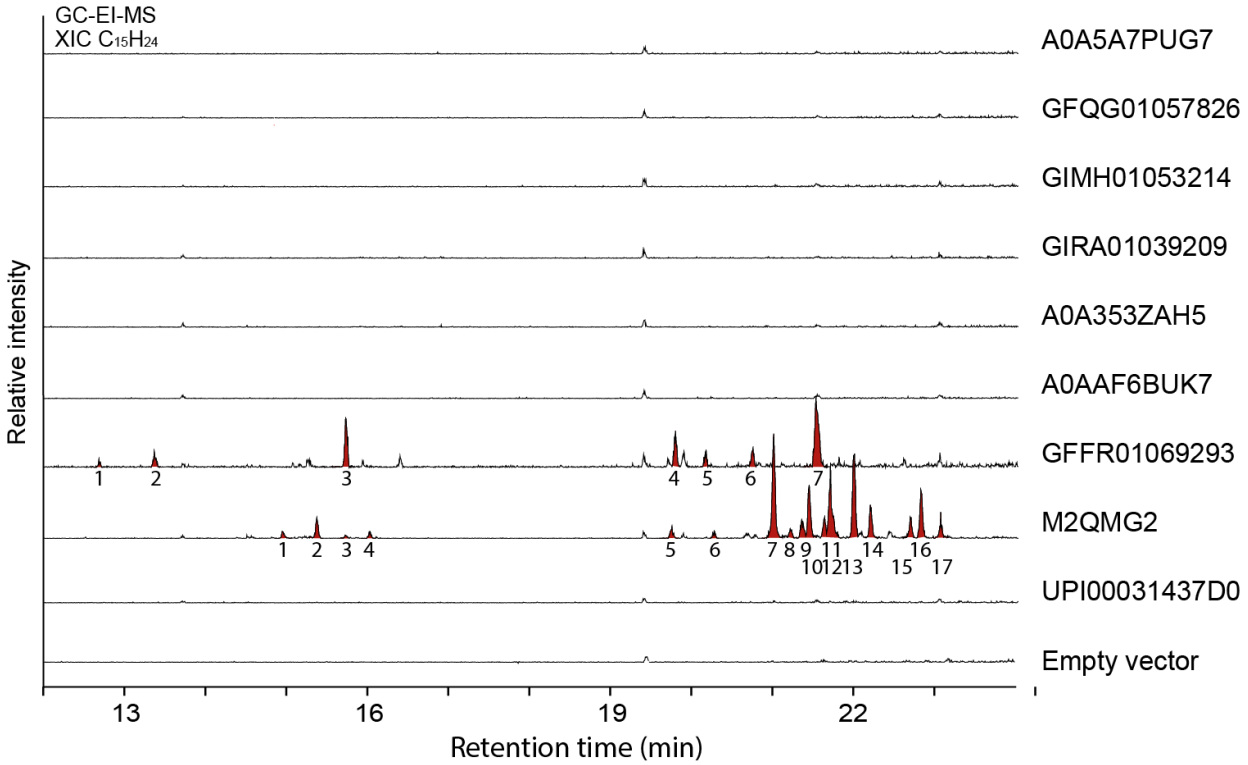


Fig. S11. GC-MS extracted ion chromatograms of the molecular mass of $C_{15}H_{24}$. Results of yeast expressions of proteins shown in **Table S1** (protein accession IDs shown on the right side). Detected sesquiterpene ions are highlighted in red.

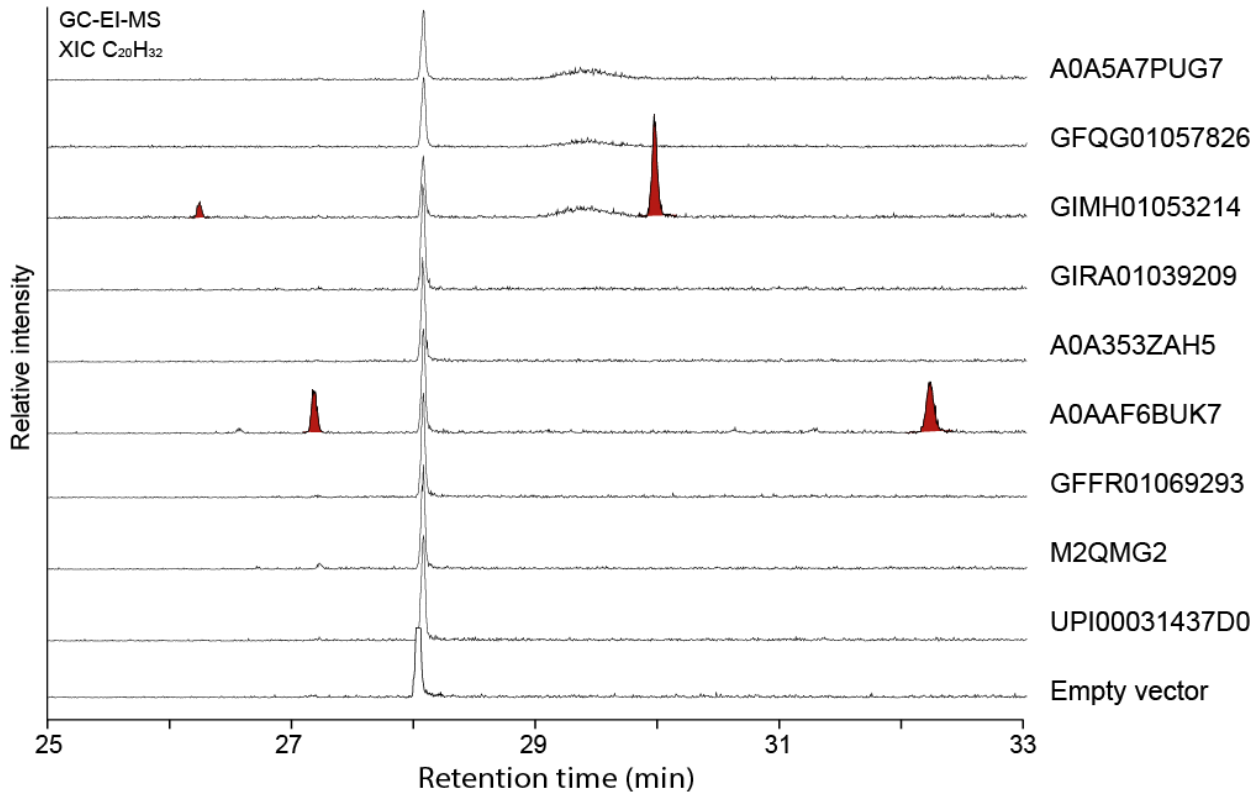


Fig. S12. GC-MS extracted ion chromatograms of the molecular mass of $C_{20}H_{32}$. Results of yeast expressions of proteins shown in **Table S1** (protein accession IDs shown on the right side). Detected diterpene ions are highlighted in red.

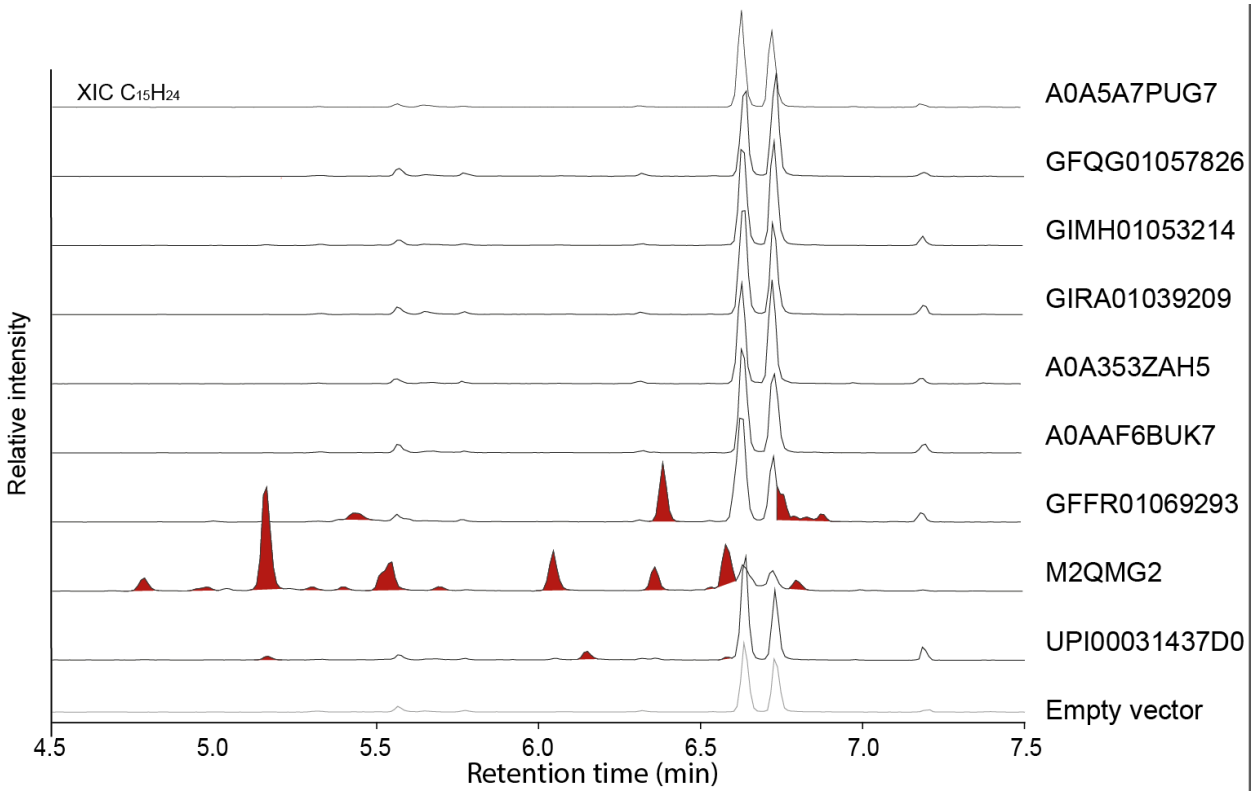


Fig. S13. LC-MS extracted ion chromatograms of the molecular mass of $[C_{15}H_{24}+H]^+$ ions at *m/z* 205.1951 (+/- 5 ppm). Results of yeast expressions of proteins shown in **Table S1** (protein accession IDs shown on the right side). Detected sesquiterpene ions are highlighted in red.

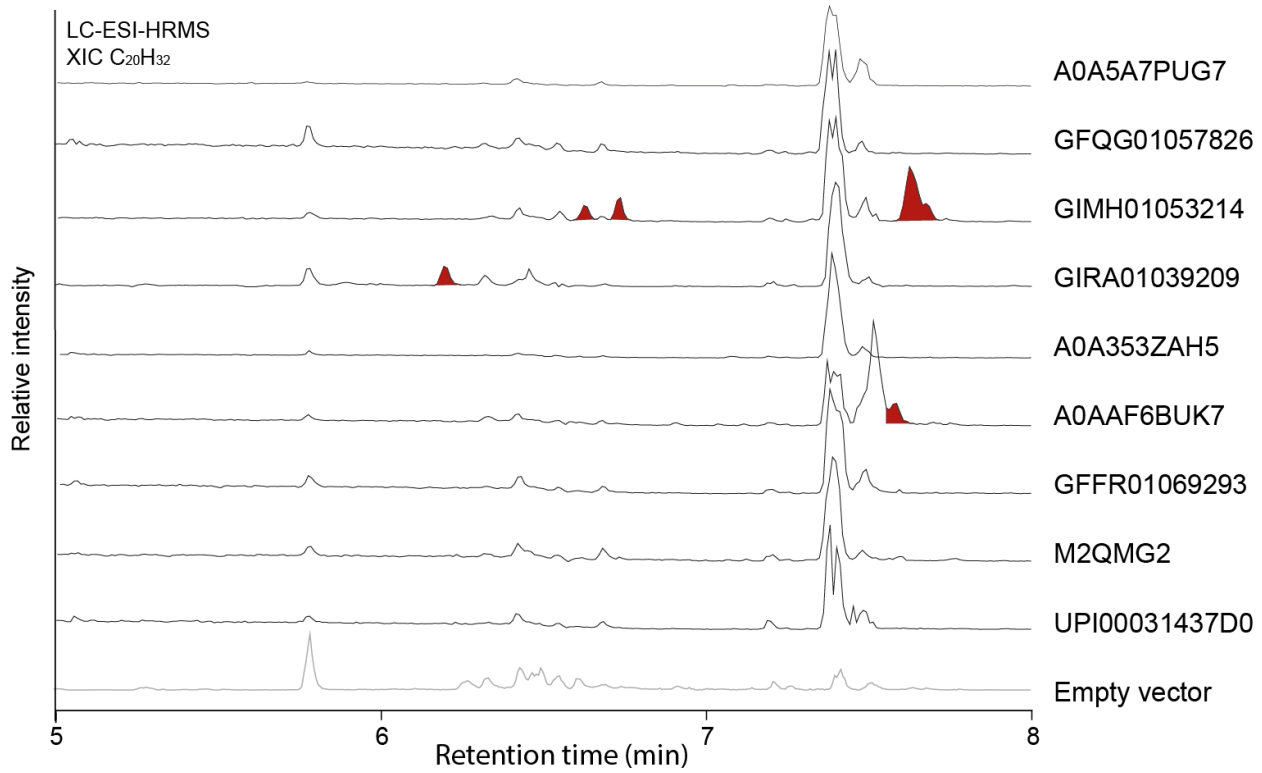


Fig. S14. LC-MS extracted ion chromatograms of the molecular mass of $[C_{20}H_{32}+H]^+$ ions at *m/z* 273.2577 (+/- 5 ppm). Results of yeast expressions of proteins shown in **Table S1** (protein accession IDs shown on the right side). Detected diterpene ions are highlighted in red.

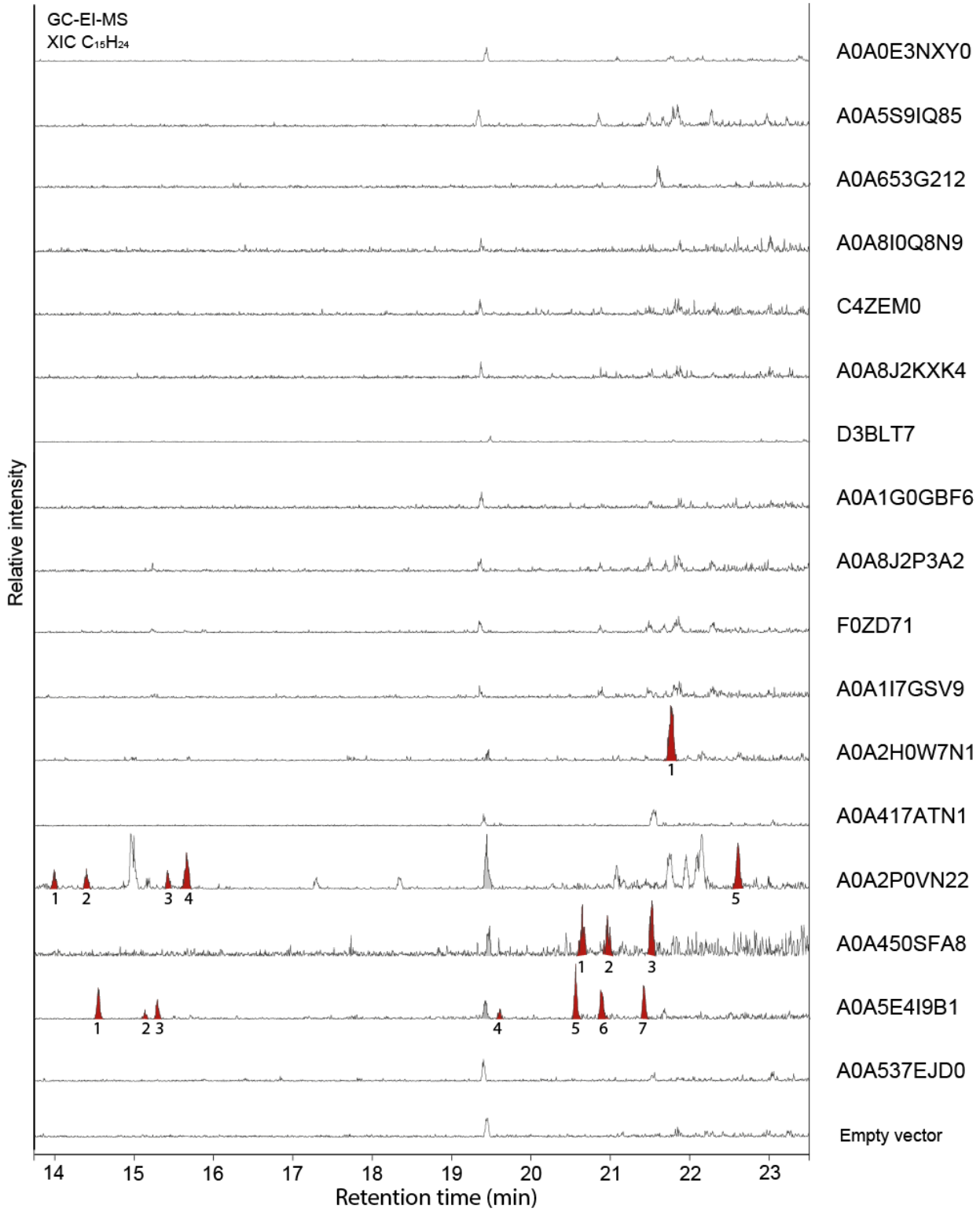


Fig. S15. GC-MS extracted ion chromatograms of the molecular mass of $C_{15}H_{24}$. Results of yeast expressions of proteins shown in **Table S2** (protein accession IDs shown on the right side). Detected sesquiterpene ions are highlighted in red.

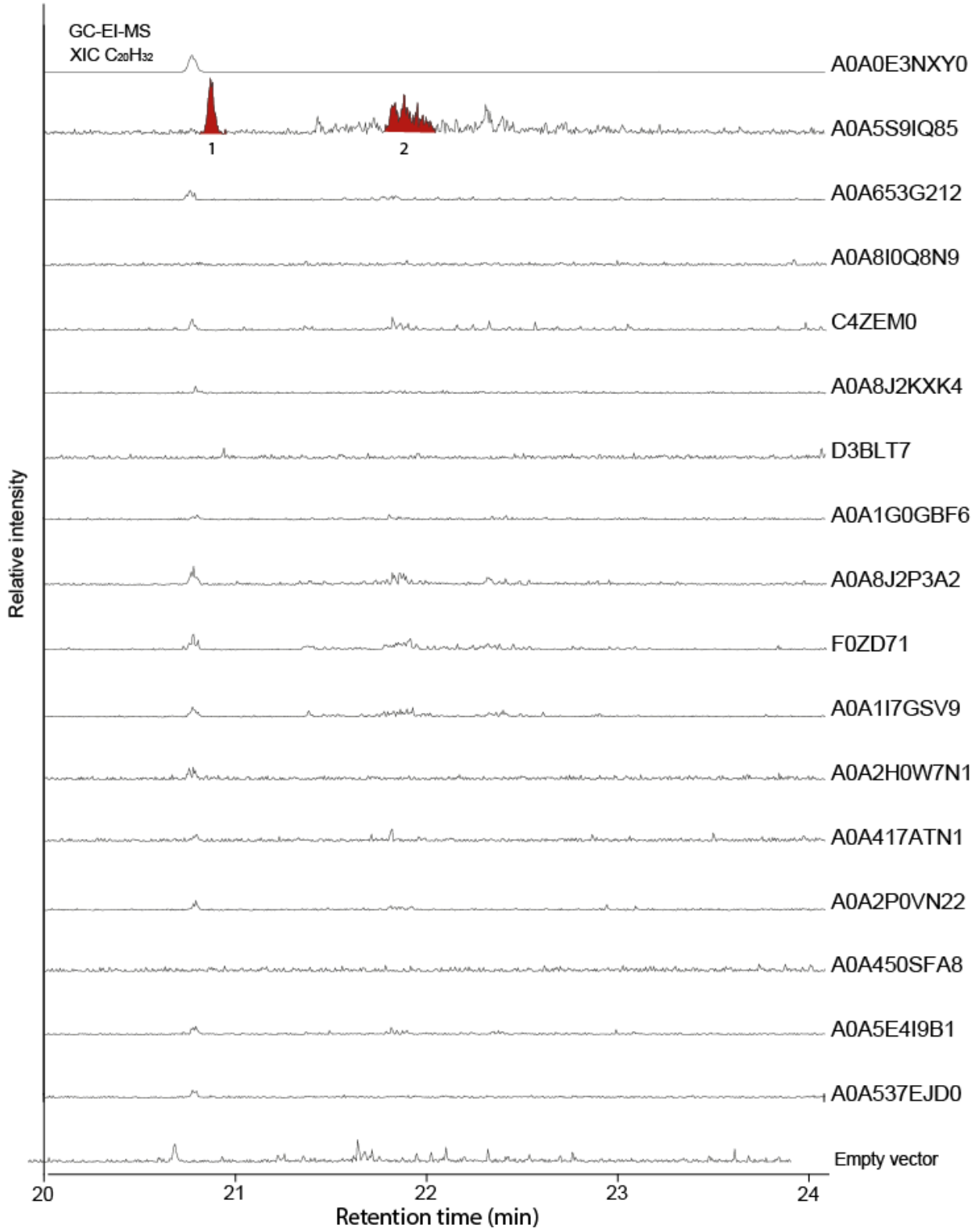


Fig. S16. GC-MS extracted ion chromatograms of the molecular mass of $C_{20}H_{32}$. Results of yeast expressions of proteins shown in **Table S2** (protein accession IDs shown on the right side). Detected diterpene ions are highlighted in red.

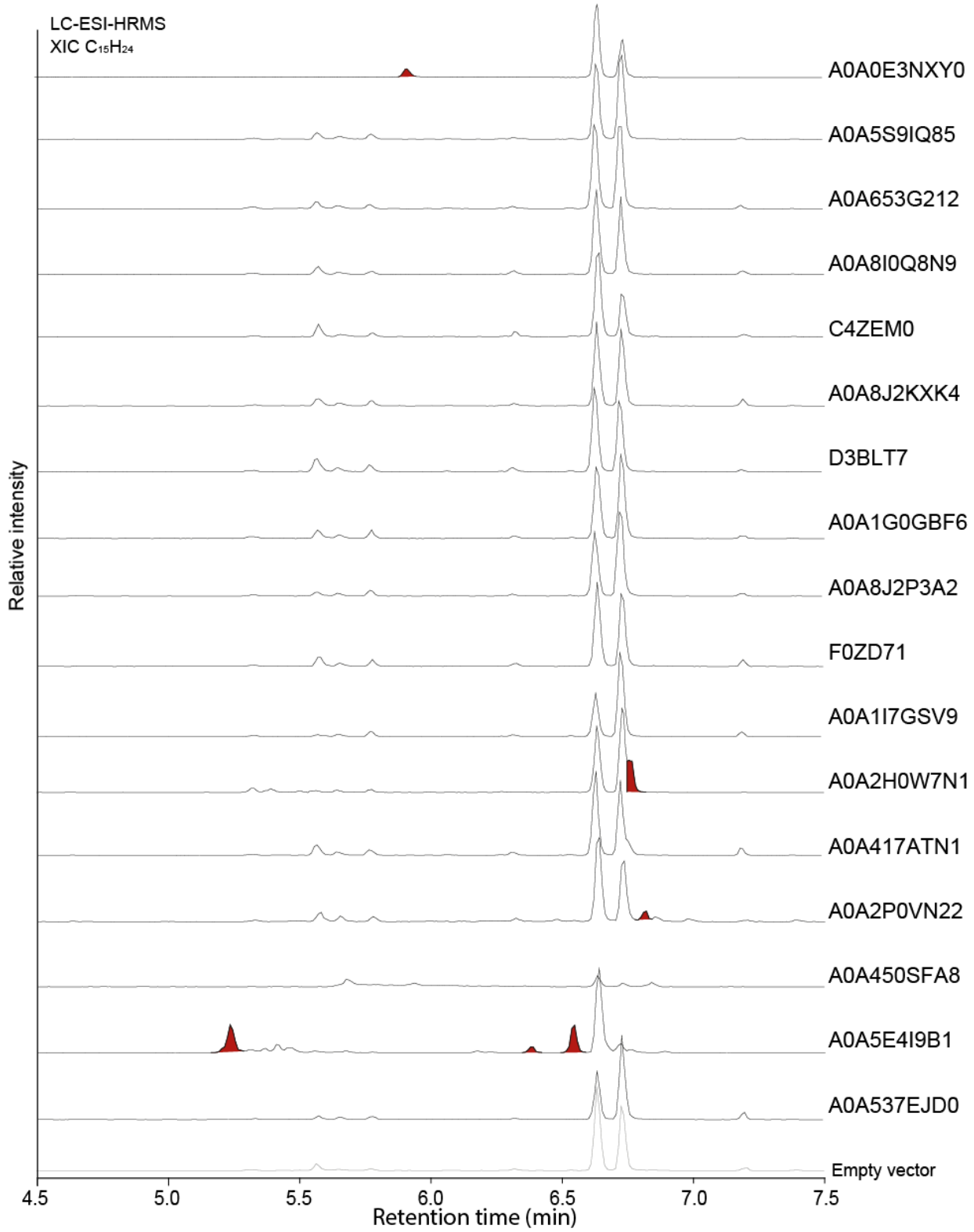


Fig. S17. LC-MS extracted ion chromatograms of the molecular mass of $[C_{15}H_{24}+H]^+$ ions at *m/z* 205.1951 (+/- 5 ppm). Results of yeast expressions of proteins shown in **Table S2** (protein accession IDs shown on the right side). Detected sesquiterpene ions are highlighted in red.

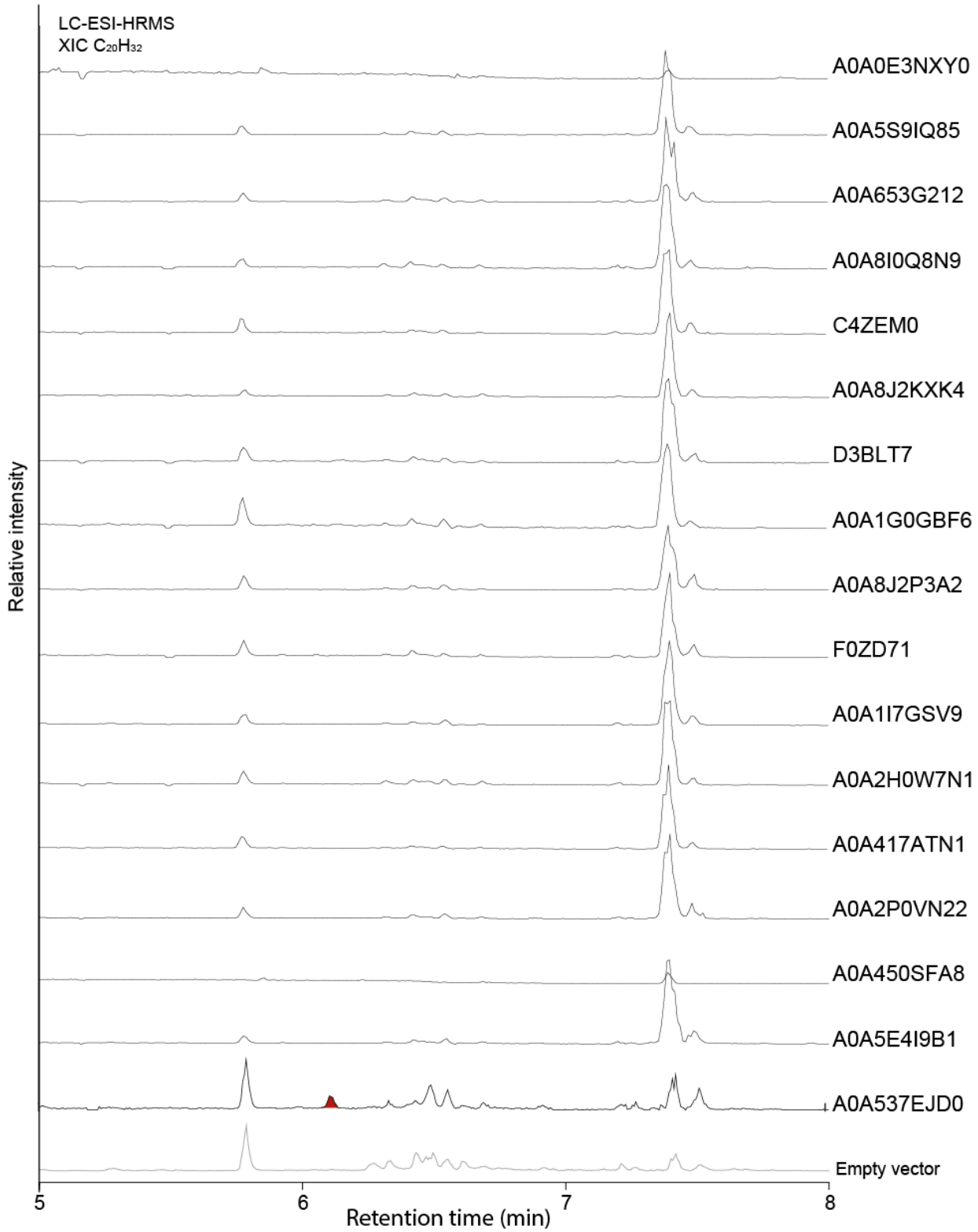


Fig. S18. LC-MS extracted ion chromatograms of the molecular mass of $[C_{20}H_{32}+H]^+$ ions at *m/z* 273.2577 (+/- 5 ppm). Results of yeast expressions of proteins shown in **Table S2** (protein accession IDs shown on the right side). Detected diterpene ions are highlighted in red.

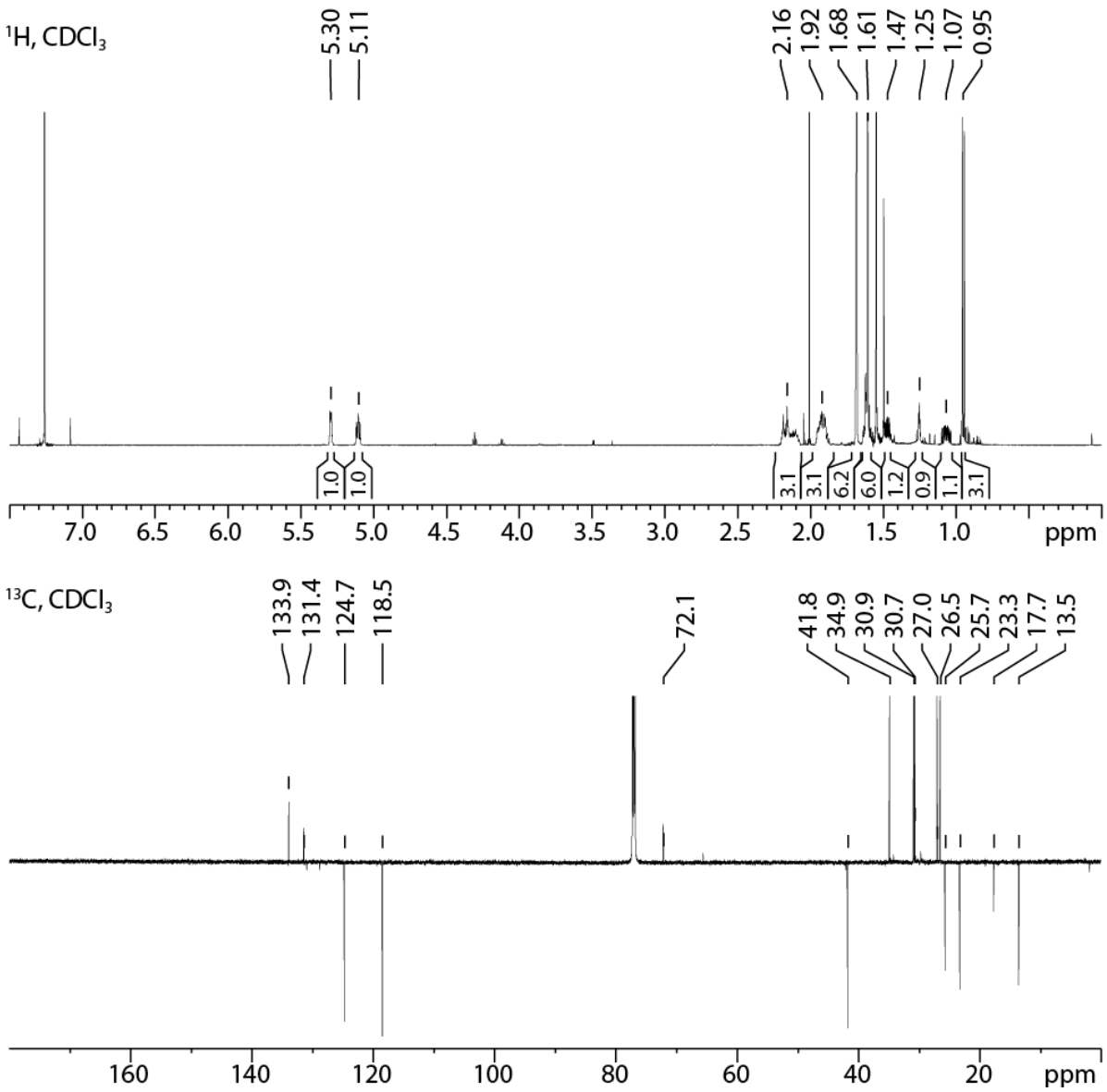


Fig. S19. NMR spectra of isolated terpene **12** from the cultivation of yeast strain JWY501 expressing the archaeal enzyme A0A5E4I9B1.

SMILES: CC1=CCC(C(C)CC/C=C(C)/C)(O)CC1
INCHIKEY:WTVHAMTYZJGJLJ-UHFFFAOYSA-N

δH: 5.30 (1H, m), 5.11 (1H, m), 2.16 (3H, m), 1.92 (3H, m), 1.68 (6H, m), 1.61 (6H, m),1.47 (1H, m), 1.25 (1H, m), 1.07 (1H, m), 0.95 (3H, d).

δC: 133.9 (C), 131.4 (C), 124.7 (CH), 118.5 (CH), 72.1 (C), 41.8 (C), 34.9 (CH), 30.9 (CH2), 30.7 (CH2), 27.0 (CH2), 26.5 (CH2), 25.7 (CH3), 23.3 (CH3), 17.7 (CH3), 13.5 (CH3).

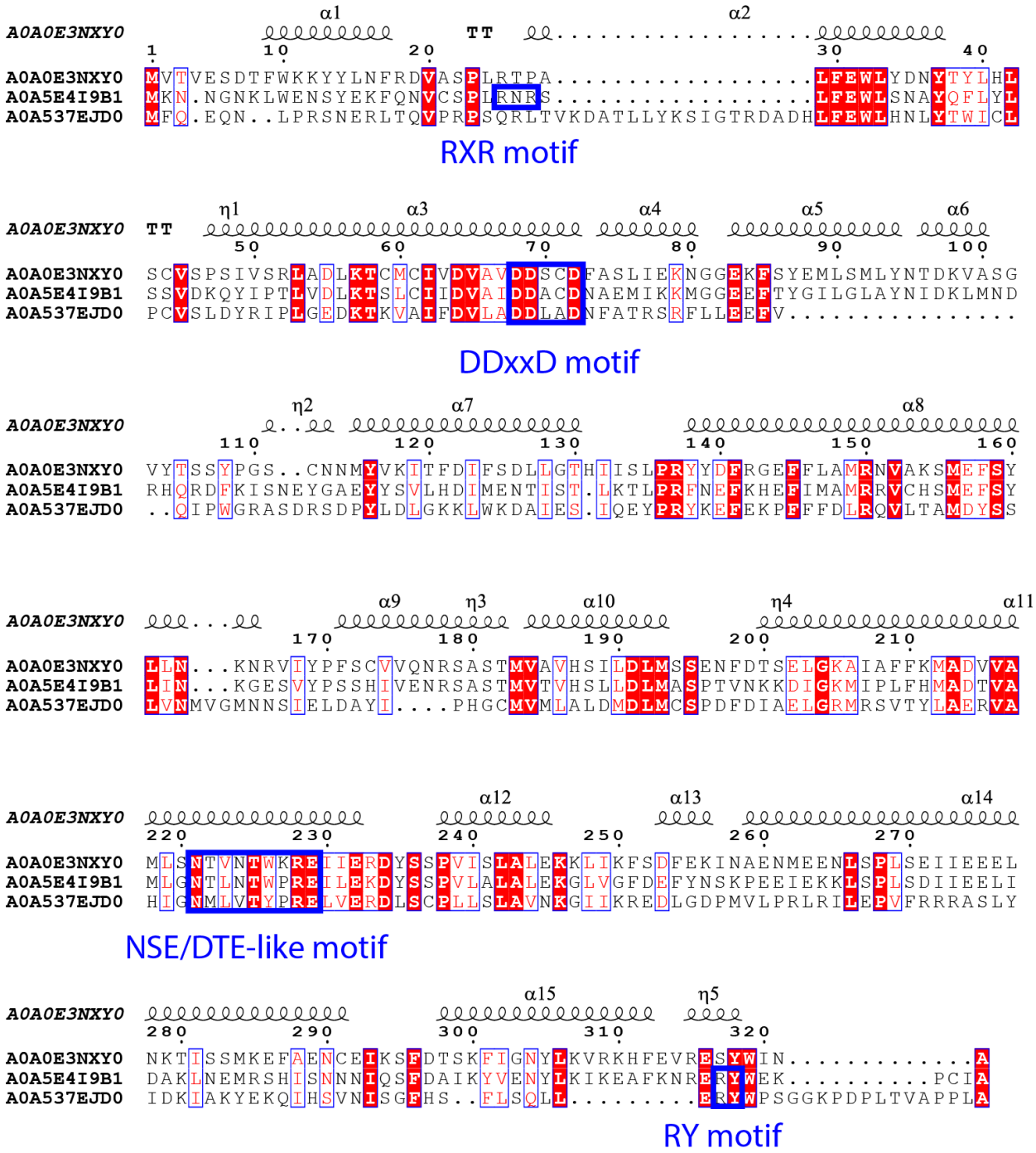


Fig. S20. Multiple sequence alignment of the three newly identified archeal TPSs. A0A5E4I9B1 and A0A0E3NXY0 were both identified as sesquiterpene synthases, while A0A537EJD0 is a diterpene synthase. Predicted secondary structure is shown for A0A0E3NXY0. Of the typical terpene synthase motifs, the aspartate rich motif DDXXD, typical for class I terpene synthases, is conserved across the three enzymes. A conserved sequence similar to the NSE/DTE motif can also be identified.

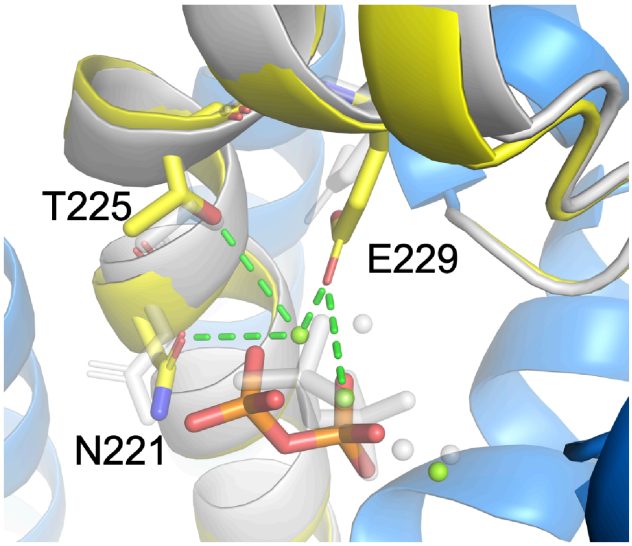


Fig. S21. The structure of the archaeal TPS A0A5E4I9B1 active site as predicted by AlphaFold. The initial AlphaFold 3 model of A0A5E4I9B1 (grey) and activated model obtained from restrained energy minimization of the initial AlphaFold 3 model, which enforces TPS-specific interactions of conserved amino acid motifs (e.g., ^{221}**N**TLN**T**WPR**E**) with the diphosphate and Mg^{2+} cluster (yellow carbon color code). Blue color represents the region fixed during minimization.

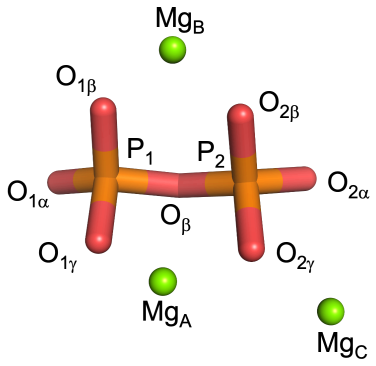


Fig. S22. Labeling of the diphosphate atoms and Mg^{2+} ions in the TPS structure.

Supplementary Tables

Table S1. List of candidate TPS sequences expressed in yeast. Sequences were selected based on phylogenetic analysis of mined TPS-like sequences (**Fig. S10A**). We predict substrates only for sequences for which we estimate the probability of TPS activity above 30%. A0A353ZAH5 had predicted TPS activity with 25% confidence; therefore, we predicted no substrates.

Accession ID	Origin	Species of origin	Length (aa)	InterPro annotations	Predicted substrate	Observed activity
A0A5A7PUG7	Plant	*Striga asiatica*	721	PF01397, PF03936	FPP	
GFQG01057826	Plant	*Amaranthus tricolor*	831	PF01397, PF03936	FPP	
GIMH01053214	Plant	*Selaginella wallacei*	803	PF01397, PF03936	GGPP	**Di**
GIRA01039209	Plant	*Beta vulgaris*	822	PF01397, PF03936	GGPP	**Di**
A0A353ZAH5	Bacterium	*Planctomycetaceae bacterium*	749	SSF48239	None	
A0AAF6BUK7	Plant	*Marchantia polymorpha*	849	PF01397, PF03936	GGPP	**Di**
GFFR01069293	Plant	*Camellia sinensis*	570	PF01397, PF03936	FPP	**Sesqui**
M2QMG2	Fungus	*Gelatoporia subvermispora*	354	PF19086	FPP	**Sesqui**
UPI00031437D0	Bacterium	*Pseudomonas agarici*	355	PF19086	FPP	**Sesqui**

Table S2. List of candidate TPS sequences expressed in yeast. Sequences were selected based on EnzymeExplorer predictions from UniRef90 sequences that lack any InterPro annotations ("dark matter" sequences, see **Fig. S10B**).

Accession ID	Origin	Species of origin	Length (aa)	InterPro annotations	Predicted substrate	Observed activity
A0A0E3NXY0	Archeon	*Methanosarcina sp.*	322	none	FPP	**Sesqui**
A0A1G0GBF6	Bacterium	*Gammaproteobacteria bacterium*	284	none	FPP	
A0A1I7GSV9	Bacterium	*Clostridium sp.*	390	none	FPP	
A0A2P0VN22	Virus	*Tetraselmis virus 1*	319	none	FPP	**Sesqui**
A0A2H0W7N1	Bacterium	*Candidatus Berkelbacteria bacterium*	326	none	FPP	**Sesqui**
A0A417ATN1	Bacterium	*Blautia wexlerae*	316	none	FPP	
A0A450SFA8	Bacterium	*Candidatus Kentron sp.*	340	none	FPP	**Sesqui**
A0A537EJD0	Archeon	*Candidatus Bathyarchaeota archeon*	321	none	FPP	**Di**
A0A5E4I9B1	Archeon	*uncultured archaeon*	325	none	FPP	**Sesqui**
A0A5S9IQ85	Bacterium	*Candidatus Uabimicrobium amorphum*	341	none	GGPP	**Di**
A0A653G212	Bacterium	*Candidatus Roizmanbacteria bacterium*	273	none	FPP	
A0AAP1RP55	Bacterium	*Mammaliicoccus lentus*	291	none	FPP	
A0A8J2KXK4	Animal	*Allacma fusca*	287	none	FPP	
A0A8J2P3A2	Animal	*Allacma fusca*	287	none	FPP	
C4ZEM0	Bacterium	*Agathobacter rectalis*	327	none	FPP	
D3BLT7	Amoeba	*Heterostelium album*	258	none	FPP	
F0ZD71	Amoeba	*Dictyostelium purpureum*	283	none	FPP	

Table S3. Conserved interactions in the polar region of the active site in the refined AlphaFold 3 model of A0A5E4I9B1.

Atom 1	Atom 2	Distance (Å)
Asp63(OD1)	W_{cat}	3.3
PP(O2a)	W_{cat}	2.9
Asn221(OD1)	MG_B	3.9
Glu229(OE1)	MG_B	1.9
Glu229(OE1)	MG_A	4.0
Thr225(OG1)	MG_B	4.0
Asp68(OD2)	Arg24(NE)	3.0
Asp68(OD1)	Arg24(NH2)	2.6
Arg24(NH1)	PP(O1a)	4.0
Arg24(NH2)	PP(O1g)	4.0
Asp67(OD2)	MG_A	3.1
Asp67(OD1)	MG_C	1.8
PP(O1g)	MG_A	1.8
PP(O2g)	MG_A	1.9
PP(O2g)	MG_C	1.9
Asp71(OD2)	MG_C	2.0
Asp71(OD2)	MG_A	3.6

Table S4. Top ten poses for the substrate FPP in A0A5E4I9B1 bound to either the O1*α* or the O2*α* oxygens. EnzyDock energies are total binding energies (kcal/mol).

Cluster	E_{tot}[a]	Diphosphate oxygen
1	0.27	O1*α*
2	0.62	O1*α*
3	0.94	O1*α*
4	14.36	O1*α*
5	16.36	O1*α*
6	16.61	O1*α*
7	22.84	O1*α*
8	27.96	O1*α*
9	30.22	O1*α*
10	32.39	O1*α*
1	14.50	O2*α*
2	20.38	O2*α*
3	22.11	O2*α*
4	29.37	O2*α*
5	31.07	O2*α*
6	35.78	O2*α*
7	39.13	O2*α*
8	73.88	O2*α*
9	80.10	O2*α*
10	89.00	O2*α*

[a] Total force field energy of the ligand.

Table S5. List of yeast strains used in this study.

Name	Genotype	Ref.
JWY501	MATα *leu2*–3112::*His3MX6*_P_{GAL1}-*ERG19*/P_{GAL10}-*ERG8* *ura3–52::URA3*_P_{GAL1}-*mvaS(A110G)*/P_{GAL10}-*mvaE(CO) his3Δ1::hphMX4*_P_{GAL1}-*ERG12*/P_{GAL10}-*IDI1* *trp1–289::TRP1*_P_{GAL1}-*crtE(X.den)*/P_{GAL10}-*ERG20* *yprcδ15::natMX*_P_{GAL1}-*crtE(opt)*/P_{GAL10}-*crtE* (*ura3–52* prototrophy removed for use of Cas9 system)	Wong *et al*. (*26*)
JWY501_TP0234	JWY501 *ura3-52::ura3::*P*tdh3*::A0A0E3NXY0::T*ssa1*	This study
JWY501_TP0235	JWY501 ura3-52::ura3::P*tdh3*::A0A5S9IQ85::T*ssa1*	This study
JWY501_TP0236	JWY501 *ura3-52::ura3::*P*tdh3*::A0A653G212::T*ssa1*	This study
JWY501_TP0239	JWY501 ura3-52::ura3::P*tdh3*::A0A8I0Q8N9::T*ssa1*	This study
JWY501_TP0240	JWY501 ura3-52::ura3::P*tdh3*::C4ZEM0::T*ssa1*	This study
JWY501_TP0242	JWY501 ura3-52::ura3::P*tdh3*::A0A8J2KXK4::T*ssa1*	This study
JWY501_TP0243	JWY501 *ura3-52::ura3::*P*tdh3*::D3BLT7::T*ssa1*	This study
JWY501_TP0244	JWY501 ura3-52::ura3::P*tdh3*::A0A1G0GBF6::T*ssa1*	This study
JWY501_TP0245	JWY501 ura3-52::ura3::P*tdh3*::A0A8J2P3A2::T*ssa1*	This study
JWY501_TP0246	JWY501 ura3-52::ura3::P*tdh3*::F0ZD71::T*ssa1*	This study
JWY501_TP0247	JWY501 ura3-52::ura3::P*tdh3*::A0A1I7GSV9::T*ssa1*	This study
JWY501_TP0250	JWY501 ura3-52::ura3::P*tdh3*::A0A2H0W7N1::T*ssa1*	This study
JWY501_TP0251	JWY501 ura3-52::ura3::P*tdh3*::A0A417ATN1::T*ssa1*	This study
JWY501_TP0253	JWY501 ura3-52::ura3::P*tdh3*::A0A2P0VN22::T*ssa1*	This study
JWY501_TP0254	JWY501 ura3-52::ura3::P*tdh3*::A0A450SFA8::T*ssa1*	This study
JWY501_TP0255	JWY501 ura3-52::ura3::P*tdh3*::A0A5E4I9B1::T*ssa1*	This study
JWY501_TP0256	JWY501 ura3-52::ura3::P*tdh3*::A0A537EJD0::T*ssa1*	This study

Table S6. List of yeast plasmids used in this study.

Plasmid name	Plasmid type	Backbone & selection marker	Description	Ref.
pYTK001	Backbone	pYTK001, Cm	Storage plasmid	Lee et al. (*63*)
pYTK002	Part type 1	pYTK001, Cm	ConLS	Lee et al. (*63*)
pYTK009	Part type 2	pYTK001, Cm	P*tdh3*	Lee et al. (*63*)
pYTK030	Part type 2	pYTK001, Cm	P*gal1*	Lee et al. (*63*)
pYTK047	Part type 234	pYTK001, Cm	GFP dropout	Lee et al. (*63*)
pYTK052	Part type 4	pYTK001, Cm	T*ssa1*	Lee et al. (*63*)
pYTK067	Part type 5	pYTK001, Cm	ConR1	Lee et al. (*63*)
pYTK074	Part type 6	pYTK001, Cm	URA3	Lee et al. (*63*)
pYTK082	Part type 7	pYTK001, m	2μ origin of replication	Lee et al. (*63*)
pYTK083	Part type 8	pYTK083, Amp	Backbone, bacterial module	Lee et al. (*63*)
pYTK096	Backbone	pYTK096, Amp	Ura3 targeting integration vector	Lee et al. (*63*)
pESC-Ura	Backbone	Backbone, Amp	Ura3 resistance marker	Agilent, cat. #217454
pTP0027	Backbone	pYTK083, Amp	Transient expression, Ura3 prototrophy	This study
pTP0098	Part type 3	pYTK001, Cm	A0A5A7PUG7	This study
pTP0099	Part type 3	pYTK001, Cm	GFQG01057826	This study
pTP0100	Part type 3	pYTK001, Cm	GIMH01053214	This study
pTP0101	Part type 3	pYTK001, Cm	GIRA01039209	This study
pTP0102	Part type 3	pYTK001, Cm	A0A353ZAH5	This study
pTP0103	Part type 3	pYTK001, Cm	A0AAF6BUK7	This study
pTP0104	Part type 3	pYTK001, Cm	GFFR01069293	This study
pTP0105	Part type 3	pYTK001, Cm	M2QMG2	This study
pTP0106	Part type 3	pYTK001, Cm	UPI00031437D0	This study
pTP0195	Transient expression	pTP0027, Amp	P*thd3,* A0A5A7PUG7, T*ssa1*	This study
pTP0196	Transient expression	pTP0027, Amp	P*thd3,* GFQG01057826, T*ssa1*	This study
pTP0197	Transient expression	pTP0027, Amp	P*thd3,* GIMH01053214, T*ssa1*	This study
pTP0198	Transient expression	pTP0027, Amp	P*thd3,* GIRA01039209, T*ssa1*	This study
pTP0199	Transient expression	pTP0027, Amp	P*thd3,* A0A353ZAH5, T*ssa1*	This study
pTP0200	Transient expression	pTP0027, Amp	P*thd3,* A0AAF6BUK7, T*ssa1*	This study
pTP0201	Transient expression	pTP0027, Amp	P*thd3,* GFFR01069293, T*ssa1*	This study
pTP0202	Transient expression	pTP0027, Amp	P*thd3,* M2QMG2, T*ssa1*	This study
pTP0203	Transient expression	pTP0027, Amp	P*thd3,* UPI00031437D0, T*ssa1*	This study
pTP0209	Part type 3	pYTK001, Cm	A0A0E3NXY0	This study
pTP0210	Part type 3	pYTK001, Cm	A0A5S9IQ85	This study
pTP0211	Part type 3	pYTK001, Cm	A0A653G212	This study
pTP0214	Part type 3	pYTK001, Cm	A0A8I0Q8N9	This study

pTP0215	Part type 3	pYTK001, Cm	C4ZEM0	This study
pTP0217	Part type 3	pYTK001, Cm	A0A8J2KXK4	This study
pTP0218	Part type 3	pYTK001, Cm	D3BLT7	This study
pTP0219	Part type 3	pYTK001, Cm	A0A1G0GBF6	This study
pTP0220	Part type 3	pYTK001, Cm	A0A8J2P3A2	This study
pTP0221	Part type 3	pYTK001, Cm	F0ZD71	This study
pTP0222	Part type 3	pYTK001, Cm	A0A1I7GSV9	This study
pTP0225	Part type 3	pYTK001, Cm	A0A2H0W7N1	This study
pTP0226	Part type 3	pYTK001, Cm	A0A417ATN1	This study
pTP0228	Part type 3	pYTK001, Cm	A0A2P0VN22	This study
pTP0229	Part type 3	pYTK001, Cm	A0A450SFA8	This study
pTP0230	Part type 3	pYTK001, Cm	A0A5E4I9B1	This study
pTP0231	Part type 3	pYTK001, Cm	A0A537EJD0	This study
pTP0234	Integration vector	pYTK096, amp	P*tdh3*, A0A0E3NXY0, T*ssa1*	This study
pTP0235	Integration vector	pYTK096, amp	P*tdh3*, A0A5S9IQ85, T*ssa1*	This study
pTP0236	Integration vector	pYTK096, amp	P*tdh3*, A0A653G212, T*ssa1*	This study
pTP0239	Integration vector	pYTK096, amp	P*tdh3*, A0A8I0Q8N9, T*ssa1*	This study
pTP0240	Integration vector	pYTK096, amp	P*tdh3*, C4ZEM0, T*ssa1*	This study
pTP0242	Integration vector	pYTK096, amp	P*tdh3*, A0A8J2KXK4, T*ssa1*	This study
pTP0243	Integration vector	pYTK096, amp	P*tdh3*, D3BLT7, T*ssa1*	This study
pTP0244	Integration vector	pYTK096, amp	P*tdh3*, A0A1G0GBF6, T*ssa1*	This study
pTP0245	Integration vector	pYTK096, amp	P*tdh3*, A0A8J2P3A2, T*ssa1*	This study
pTP0246	Integration vector	pYTK096, amp	P*tdh3*, F0ZD71, T*ssa1*	This study
pTP0247	Integration vector	pYTK096, amp	P*tdh3*, A0A1I7GSV9, T*ssa1*	This study
pTP0250	Integration vector	pYTK096, amp	P*tdh3*, A0A2H0W7N1, T*ssa1*	This study
pTP0251	Integration vector	pYTK096, amp	P*tdh3*, A0A417ATN1, T*ssa1*	This study
pTP0253	Integration vector	pYTK096, amp	P*tdh3*, A0A2P0VN22, T*ssa1*	This study
pTP0254	Integration vector	pYTK096, amp	P*tdh3*, A0A450SFA8, T*ssa1*	This study
pTP0255	Integration vector	pYTK096, amp	P*tdh3*, A0A5E4I9B1, T*ssa1*	This study
pTP0255μ	Transient expression	pTP0027, Amp	P*gal1*, A0A5E4I9B1, T*ssa1*	This study
pTP0256	Integration vector	pYTK096, amp	P*tdh3*, A0A537EJD0, T*ssa1*	This study