Highly accurate discovery of terpene synthases powered by machine learning reveals functional terpene cyclization in Archaea

Raman Samusevich[1,2], Téo Hebra[1], Roman Bushuiev[1,2], Anton Bushuiev[2], Tereza Čalounová[1], Helena Smrčková[1], Ratthachat Chatpatanasiri[2], Jonáš Kulhánek[2], Milana Perković[1], Martin Engst[1], Adéla Tajovská[1], Josef Sivic[2], Tomáš Pluskal[1*]

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

[2]Czech Institute of Informatics, Robotics and Cybernetics (CIIRC), Czech Technical University in Prague, Czech Republic

[*]Corresponding author: tomas.pluskal@uochb.cas.cz

Abstract

Terpene synthases (TPSs) generate the scaffolds of the largest class of natural products, including several first-line medicines. The amount of available protein sequences is increasing exponentially, and accurate computational characterization of their function remains an unsolved challenge. We assembled a curated dataset of one thousand characterized TPS reactions and developed a method to devise highly accurate machine-learning models for functional annotation in a low-data regime. Our models significantly outperform existing methods for TPS detection and substrate prediction. By applying the models to large protein sequence databases, we discovered seven TPS enzymes previously undetected by state-of-the-art protein signatures and experimentally confirmed their activity, including the first reported TPSs in the major domain of life Archaea. Furthermore, we discovered a new TPS structural domain and distinct subtypes of previously known domains. This work demonstrates the potential of machine learning to speed up the discovery and characterization of novel TPSs.

Introduction

Terpene synthases (TPSs) are ubiquitous enzymes that produce the hydrocarbon scaffolds for the largest and the most diverse class of natural products called terpenoids, which include widely used flavors, fragrances, and first-line medicine. The most natural scents a human has ever experienced are terpenoids[1]. Some sesquiterpenes treat malaria (e.g., a Nobel-prize winning artemisinin[2] with a market size projected to reach USD 697.9 million by 2025[3]), bacterial infections, and migraines[4]. Some diterpenes possess anticancer or anti-inflammatory properties and treat cardiovascular diseases[5,6], e.g., taxol is the first-line anticancer medicine with

billion-dollar pick annual sales[7]. Triterpenes can improve wound healing and blood circulation[8]. Other complex terpenes exhibit anticancer properties by acting on different stages of tumor development[8,9]. Furthermore, terpenes are being used as biofuels[10,11].

Yet terpenes and terpenoids are too complex to be efficiently synthesized industrially[12], and are typically extracted from plants, which is resource-intensive. For instance, each patient requires 3-10 Pacific yew trees for the mentioned anticancer treatment taxol[13]. A more sustainable and efficient production of terpenoids may be achieved via synthetic biology [14].

However, the experimental discovery of TPSs responsible for the biosynthesis of these invaluable natural products is not trivial and time-consuming. At the same time, the number of available protein sequences is increasing exponentially, making the experimental characterization of candidate TPSs infeasible in practice. As a possible solution, preprocessing the results of high-throughput DNA sequencing by detecting the most likely candidates for TPS function can significantly accelerate the progress in bioprospecting, natural product biosynthesis, and synthetic biology.

Recent machine learning (ML) advances enable high-quality automatic data annotation employing models trained on large datasets. Breakthroughs in biological applications of machine learning are best exemplified by protein language models and protein structure prediction models like AlphaFold2[15–17]. Protein language models (PLM) enabled state-of-the-art prediction of enzymatic activity when trained on all available annotated data from Uniprot[18], and the analysis of AlphaFold2 predictions at scale led to new biological insights about protein structure space[19,20]. However, when studying a specific enzymatic family, researchers do not have the luxury of working with labeled data at scale but rather deal with small datasets of characterized enzymes[21]. A low-data regime makes the application of the recent deep-learning techniques challenging. As a result, established computational approaches for *in-silico* TPS characterization are based on techniques such as PSI-BLAST[22] and Profile Hidden Markov Models[23].

In the case of TPSs, existing datasets of characterized sequences are sparse[21]. Mere dozens of characterized proteins cover TPS classes such as sesterTPSs or tetraTPSs. As a result, there are no tools to predict these understudied classes. The state-of-the-art TPS detection and characterization method[24] predicts only the three most abundant substrate classes (monoTPS, diTPS, and sesquiTPS).

Here, we present a curated dataset of over a thousand characterized TPS sequences (Extended data Fig 1) and use it to train a machine-learning pipeline for automated TPS detection and substrate prediction. A protein sequence is the only required input to the pipeline.

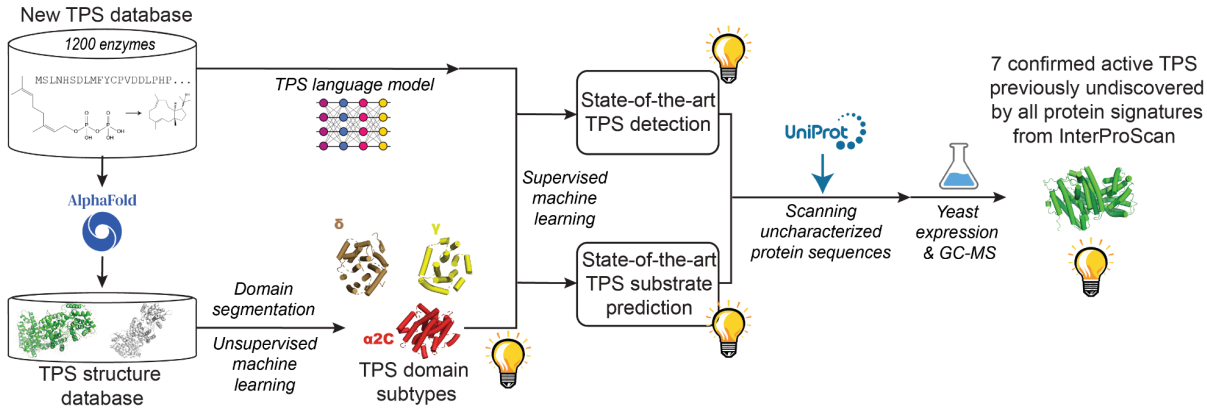


Fig. 1 | An overview of our solution: from curating a TPS database of characterized reactions to the experimental characterization of seven novel TPS enzymes by heterologous expression in *Saccharomyces cerevisiae*. The 7 discovered TPSs were missed as proteins by all computational tools integrated into InterProScan for protein function, domains, and families annotation[25]. We devised robust TPS domain segmentation. It enabled the discovery of TPS domain subtypes utilizing unsupervised machine learning. We finetuned a protein language model Ankh[26] into a TPS language model. Combined with our domain segmentation, it led to novel TPS detection and substrate prediction approaches, significantly improving the state-of-the-art.

Our method leverages and builds on recent PLM[26] and AlphaFold2[17] models and established methods for structural alignment, unsupervised, and supervised ML. Fig.1 overviews our workflows. Final models are highly accurate, even for rare classes like sesterTPSs or tetraTPSs. By increasing the mean average precision of TPS substrate classification from 0.69 to 0.89, our solution greatly outperforms existing methods based on PSI-BLAST[22], Profile Hidden Markov Models[23], and recent state-of-the-art deep-learning models Foldseek[20] and CLEAN[18]. We applied our method to the UniProt[27] and UniRef50[28] databases and selected challenging detections for experimental validation by heterologous expression in *Saccharomyces cerevisiae*. The selected detections were not recognized as proteins by traditional bioinformatic approaches like Pfam[29], SUPFAM[30]. Specifically, in the challenging detections, no known protein signature was detected by any out of all 53.000 tools integrated into InterProScan[25]. We experimentally validated the detections by expressing newly detected TPSs in *Saccharomyces cerevisiae* JWY501, an engineered strain for sesqui and diterpene production. Using this approach, we confirmed the predicted activity for seven TPS sequences, including Archaea's first reported active TPSs. Before our work, it was believed that Archaea can form prenyl monomers but cannot perform cyclization of FPP, the substrate of sesquiTPS, or GGPP, the substrate of diTPS[31]. By leveraging our predictive pipeline, we are the first to report three experimentally confirmed active TPSs in Archaea acting upon FPP (A0A5E4I9B1, A0A0E3NXY0) and GGPP (A0A537EJD0). In addition, we performed experimental validation of substrate prediction accuracy using uncharacterized sequences. The experimental validation via heterologous expression in Saccharomyces cerevisiae demonstrated high accuracy of substrate classification. Furthermore, our computational analysis revealed subtypes of structural TPS domains, with one subtype possessing distinct properties to be considered a separate structural

domain. We derived a discriminative sequence motif for the discovered domain using explainable-AI (XAI) techniques.

Results

Computational characterization of terpene synthases

The goal is to develop an in-silico method for accurate TPS detection in large datasets of protein sequences that are available thanks to high-throughput DNA sequencing. Moreover, we aim to estimate substrate(s) for each detected TPS, including substrate types underrepresented in experimental data. For this task, the main challenge is the sparsity of the TPS sequence space due to the low-data regime. Here, we describe a three-step strategy to overcome this challenge.

Our first step in tackling the sparsity of the TPS sequence space is to gather more characterized examples. To that end, we mined over 700 TPS sequences from Swiss-Prot, the expertly curated UniProtKB component produced by the UniProt consortium[32]. In addition, we gathered characterized TPS sequences from academic literature. Manual curation of the related academic literature resulted in more than 417 proteins not included in Swiss-Prot. We enriched the dataset with over 100 isoprenyl diphosphate synthases that produce TPS substrates to enable a more comprehensive view of TPS-related activity. Details about the curated TPS dataset are in Fig.1a-b and Online Methods.

Next, to overcome the sparsity of small data, we leverage recent Protein language models (PLMs) pre-trained on all available uncharacterized protein sequences[16,26,33]. For small enzymatic datasets, a pre-trained PLM can serve as a starting point for model training or as a feature extractor. Recently, PLMs enabled a performance boost in the sequence-based prediction of enzymatic activity, best exemplified by a state-of-the-art EC number predictor CLEAN[18]. Nevertheless, we found that the prediction of TPS substrates remains challenging for the CLEAN model, see Fig. 3c. We build our models on top of pre-trained PLMs, tailored for highly accurate characterization of TPSs.

Our third step to overcome the sparsity of TPS sequence space is the inclusion of structural information. TPS structure evolves slower than sequence[21], and TPSs with low sequence similarity can share common 3D folds[24,34]. In related work, it was demonstrated that handcrafted features derived from homology-based structure modeling enable binary classification of sesquiTPSs into two broader groups of products[21]. Since then, AlphaFold2 has enabled highly accurate structure prediction from sequence[17]. A recent deep-learning model Foldseek enables the related protein search in a structure space[20], analogously to BLAST-based search in a sequence space. However, similar global structural TPS folds might differ in catalytic activity[24]. Extracting information about fine local differences in structural folds holds promise to significantly improve the modeling of enzymatic function[35]. Yet, for small datasets, the curse of dimensionality challenges data-driven learning of subtle differences in the 3D representation of proteins. We leverage the domain knowledge about TPS structural composition to include

structural information in modeling while escaping the curse of dimensionality. Different terpene synthases share similar local structural modules called α, β, γ domains[24,36,37], and we developed algorithms for their detection.

Discovery of structural domain subtypes

The goal is to extract structural information at the resolution of individual TPS-specific domains. Due to the conservation of TPS folds[24,34], the structural information can help predict TPS activity[21]. Moreover, a computational method for segmenting a TPS structure into family-specific domains would enable us to analyze the TPS domains at the scale of all characterized TPSs we curated.

The main challenge is the inability to leverage state-of-the-art protein structure segmentation, as it fails to detect TPS domains correctly (see Extended data Fig. 2b/i). Neither can a general structure segmentation assign TPS-specific α/β/γ domain types.

To tackle these challenges, we devised a TPS domain segmentation method based on aligning examples of known TPS domains to the AlphaFold2 structures, see Extended data Fig. 2a and Online Methods section for details of our segmentation method. Our domain segmentation results in a set of substructures labeled with a TPS-specific domain type (α/β/γ).

We compared the detected domains among each other via structural alignment[38]. Leveraging the computed structural alignment similarity, we used a state-of-the-art unsupervised machine learning approach HDSCAN[39], and clustered the detected domains. Each cluster defines a group with a distinct fold. We consider the discovered groups to be subtypes of the established α, β, γ structural domains in TPSs.

When analyzing the clusters, it is apparent that there are four/five larger groups of the α TPS domain (Extended data Fig. 3b,e). Nevertheless, we uncovered that each larger group has its subclusters. In addition to the hierarchical organization of the α domain subtypes, it is apparent that the space of possible α subtypes represents a spectrum of gradual structural variations (Extended data Fig. 4a). The space of β/γ domain subtypes, on the other hand, has four distinct subtypes (Extended data Fig. 3c, Extended data Fig. 4b). One insight enabled by the developed domain-level analysis sheds light on TPSs with αα architecture. We discovered that the second α domain differs from all other α TPS domains and is structurally close to the domain found in isoprenyl diphosphate synthases (Fig. 2b). Analysis of the domain subtypes led to several further insights. First, there is a subtype related to γ TPS domain, which has a distinct structure, in Fig. 2c it is denoted as δ. In Fig. 2c/ii, it is shown that structures grouped in the δ cluster align with each other almost perfectly. Fig. 2c/iii-v shows that the δ cluster structure differs from any other subtypes within classical β and γ TPS domains. Folds aligned in Fig. 2c/v are different but used to be considered the same γ TPS domain before this work. The next insight highlights the distinct biochemistry of the δ-domain cluster. Fig. 2d enlists all substrates that at least one TPS with the δ domain can accept. These substrates are accepted exclusively by TPSs with the δ domain. Moreover, there is a discriminative sequence motif for the δ domain, see Fig. 2e.

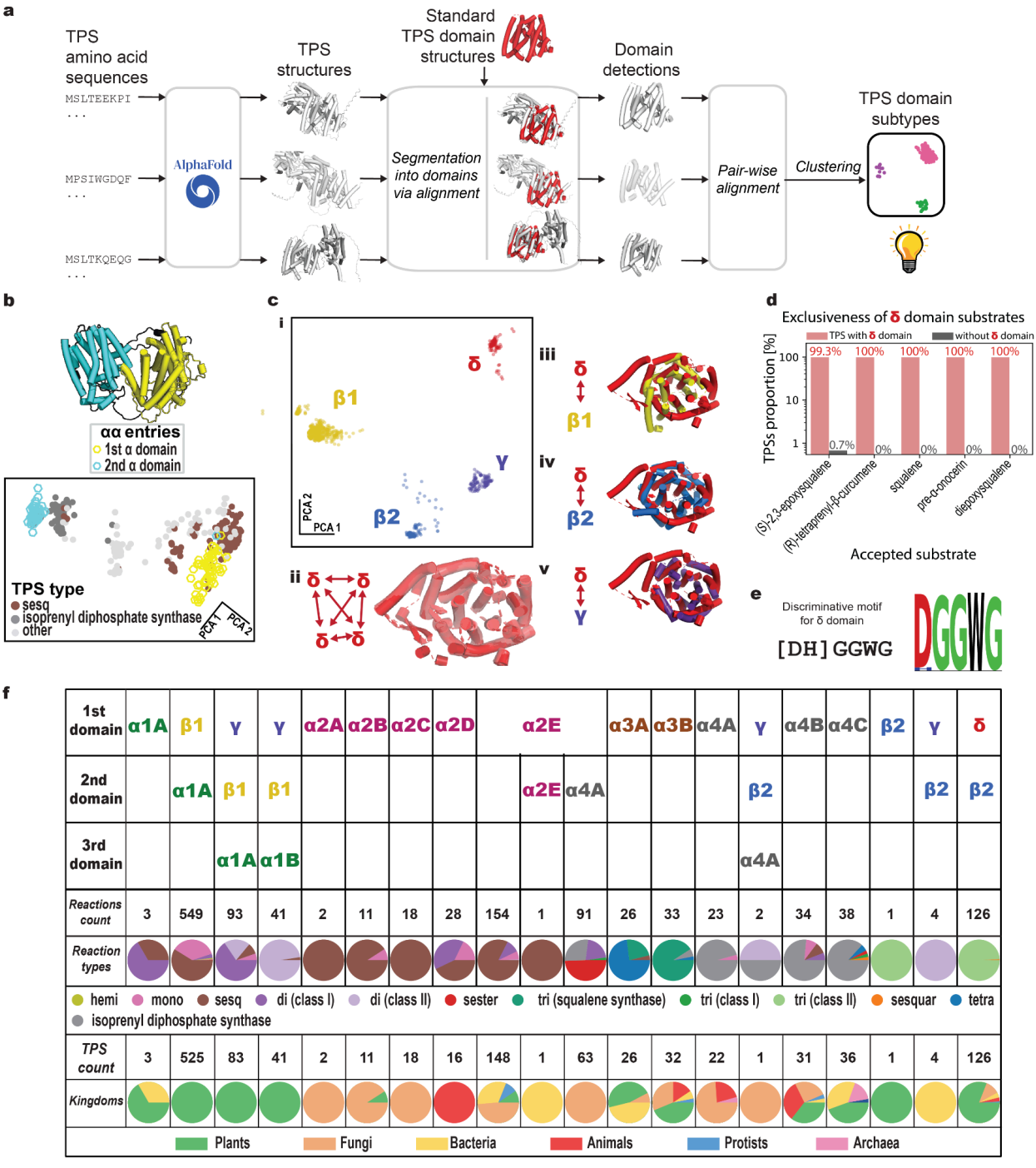


Fig. 2 | Discovery and analysis of structural TPS domains. **a,** An overview of our structure-based workflow based on segmenting predicted TPS structures into TPS-specific domains via structure alignment and clustering. **b,** Domain-level resolution of our analysis sheds light on TPSs with αα architecture. Their α domains differ from each other substantially. The first domain (in yellow) is structurally similar to the domain found in sesquiTPSs (sesq in legend, in brown color). In contrast, the fold of another α domain (in cyan) resembles isoprenyl diphosphate synthase (in grey). **c/i,** Pairwise

comparisons of detected β and γ domains reveal four clusters with distinct folds. We used Principal Components Analysis (PCA), which preserves the global structure of the space of structural similarities. **c/ii,** Domains grouped in the δ cluster align almost perfectly on each other. **c/iii-v,** There are substantial differences in folds when aligning representants of different clusters. **d,** Analysis of substrates accepted by TPSs with the δ domain. All possible substrates TPSs with a δ domain can accept are enumerated on the x-axis. Each substrate is almost exclusively accepted by δ-cluster TPSs. This finding highlights that TPSs from the δ structural domain cluster have a distinct catalytic function. **e,** Discriminative sequence motif for the δ structural domain **f,** Domain configurations of natural TPS enzymes. Sequences of domain subtypes in a TPS sequence and corresponding TPS types and kingdoms profiles.

Furthermore, we described 20 different domain configurations of natural TPS enzymes that can be distinguished based on the developed structural analysis (Fig. 2f). The order of domains in the TPS sequence confirms that β1 and β2 clusters from Fig. 2c/i correspond to subtypes of the established β domain, while γ and δ clusters used to be considered the same γ domain. Please note in Fig. 2f that the δ cluster is essential for triTPSs of Class II, being present in 126 out of 127 Class-II triTPSs. Moreover, we have shown that the δ cluster has a unique biochemical profile (Fig. 2d) and is distinct structurally (Fig. 2c, Extended data Fig. 4a). Considering all these, we propose to consider the δ cluster a separate structural domain and denote it as a *δ domain*.

Next, we performed a phylogenetic analysis of TPS sequences corresponding to different domain configurations from Fig. 2f. The phylogenetic tree is in Extended data Fig. 5. The identified configurations of domain subtypes tend to cluster together in subtrees of the phylogenetic tree, which is another indicator of the biological relevance of the discovered subtypes. In addition, it is fascinating to attempt reading the course of evolution from the constructed tree in Extended data Fig. 5. For instance, one can observe that some of TPSs with configuration γ+β+α1A likely lost their γ domain throughout evolution, while others got their α domain mutated from the α1A subtype to α1B, and it changed the corresponding biochemical activity.

The main contribution of our accurate TPS domain detection is the possibility of extracting information and analyzing TPSs at the resolution of individual domains. At the same time, the devised domain-based workflow leads to a better quality of whole-protein clustering compared to state-of-the-art sequence-based and structure-based methods (See FigSI_Whole_protein_clustering).

We publish all detected domains with their corresponding types so anyone can assign domain types to their TPS structure via tools like Foldseek[20]. Furthermore, our domain segmentation and type assignment are available in the accompanying GitHub repository and webserver.

Analysis of the new structural δ domain

The proposed δ domain has a distinct structural fold and corresponds to a unique biochemical profile. The domain is a major determinant of triTPS activity of Class II. TriTPSs are responsible for the biosynthesis of steroids and have important medical applications, e.g., in treating diabetes or cancer[40–42]. Motivated by the uniqueness of the δ domain and its biochemical

importance, we conducted additional analysis of amino acid sequences of TPSs containing the δ domain.

The goal of our analysis was to identify the δ-domain sequence motifs. The challenge was to derive a discriminative motif that is not only contained in all δ-domain sequences (high recall of δ-domain detection) but also does not appear in other sequences (high precision).

For this, we defined the following procedure that leverages and builds on explainable-AI (XAI) techniques and multiple sequence alignment (MSA). The overview of our XAI-based procedure for sequence motif derivation is depicted in Extended data Fig. 9a. Inputs to the procedure are TPS sequences labeled by the presence or absence of the δ domain in the corresponding TPS structures. The procedure output is a discriminative sequence motif reliably detecting the presence of the δ domain in a TPS.

Our intuition to apply XAI for the δ-domain motif identification arises from the observation that sequence-based TPS detectors, like PSI-BLAST, perform almost perfectly on triTPSs, see the Fig. 3e. At the same time, roughly 80% of triTPSs are defined by the distinct δ domain. We hypothesized that triTPSs might be easy to detect and reliably distinguish from other TPS types thanks to a discriminative δ-domain-specific sequence motif. If that was the case, then the ML-based TPS classifier must have learned to pay attention to that discriminative motif. XAI techniques can shed light on regions in input data that an ML model used for predictions[43], and, therefore, can help identify a δ domain discriminative motif if there is one.

We first trained a separate sequence-based detector of proteins containing a structural δ domain. Then we applied XAI to the trained model. This way, for each δ-domain-containing TPS, we obtain discriminative sequence regions the model leveraged to predict the presence of the δ-domain. Next, all sequences of the δ-domain-containing TPSs were aligned into a δ-domain-specific MSA, and the XAI-based discriminative regions were mapped onto the MSA. Next, we leave only the most discriminative regions for each sequence in the MSA by applying a cut-off on the XAI scores. Finally, from these discriminatory parts of the MSA, we created a sequence logo with a corresponding regular expression.

To access the performance of the computed discriminative sequence logo, we used a corresponding motif to retrieve δ-domain-containing sequences in the test set, which was hidden during model training. The quality of the corresponding hold-out retrieval is reported in Extended data Fig. 9b, with a mean F1-score of 0.88. We also compared the performance of the derived motif against randomly picked motifs of conserved regions from the MSA. We selected regions with sequence coverage similar to the coverage of the derived motif. The conserved regions had a similar or better recall as our motif but had poor precision, see Extended data Fig. 9c. It means that even though most δ-domain-containing TPSs have those conserved motifs from their MSA, other TPS classes also have it.

There is only a single δ-domain-containing triTPS with a corresponding characterized structure in the Protein Data Bank[44]. The triTPS is a Human lanosterol synthase (LSS), which is crucial in cholesterol biosynthesis[45]. We checked that our sequence motif is in a loop system of LSS, see Extended data Fig. 9d. Interestingly, in the related medical literature, the motif region was

reported to be crucial for the lanosterol synthase structural stability[46], and mutations in the motif region are associated with cataract and a rare genetic condition hypotrichosis simplex, when scalp hair is either absent or sparse[47]. This provides a biological indication of our motif importance for the correct function of a δ-domain-containing triTPSs.

In practice, discovering the discriminative motif for the δ domain enables researchers to detect triTPS of Class II directly from a sequence, even when the protein structure or predictive pipelines are unavailable.

Predictive models for TPS detection and substrate prediction

Given a protein sequence, the goal is to detect TPS activity and predict a substrate for each TPS detection. The main challenge is underrepresented TPS classes with less than a dozen characterized sequences. The conservation of structural folds across TPS with the same function[24,34] suggests that the structural information can help predict TPS function for the underrepresented TPS classes. However, training a machine-learning model with a protein structure as an input is very challenging for such small data because of the curse of dimensionality. To address this issue, we design an approach that leverages our segmented structural domains and compares each domain to the known domains of the same type from characterized TPSs instead of comparing the whole structures all at once. In detail, we obtain AlphaFold2-predicted structures, detect TPS-specific domains, and compare the domains of an unknown protein to the stored domain structures of characterized TPSs. Specifically, for a detected domain in the unknown protein, we store its alignment-based similarity to training domains of the same type and use the vector of similarities as input features for a predictive model. We use a Random Forest[48] as our model for its resilience to overfitting and ability to automatically perform feature selection[49]. We take sequence embeddings from a TPS language model as additional input features to the Random Forest model. We started with a protein language model (PLM)[26] pre-trained on the whole UniProt[27]. To tailor the PLM to our task, we mined sequence databases using TPS-specific Pfam[29] and SUPFAM[30] domains and retrieved sequences resembling TPSs. Using this mined dataset of putative TPS sequences, we fine-tuned the pre-trained PLM using a masked language modeling objective.

An overview of our predictive pipeline is provided in Fig. 3a. In Fig. 3c-e, we report an evaluation of TPS characterization performance. We benchmark our method against well-established Pfam[29], SUPFAM[30] domains, PSI-BLAST[22], Profile Hidden Markov Models[23], the recent deep-learning model Foldseek[20] for protein search in the structure space of AlphaFold2 predictions[20], and a state-of-the-art EC number predictor CLEAN[18]. We found classical methods to have strong performance in the task of TPS detection. Still, our pipeline further improved the detection quality by a large margin. Additionally, Fig. 3d illustrates that the existing approaches struggle to distinguish different TPS types. Our pipeline improves the performance of substrate prediction from the best baseline mean average precision of 0.69 to 0.89.

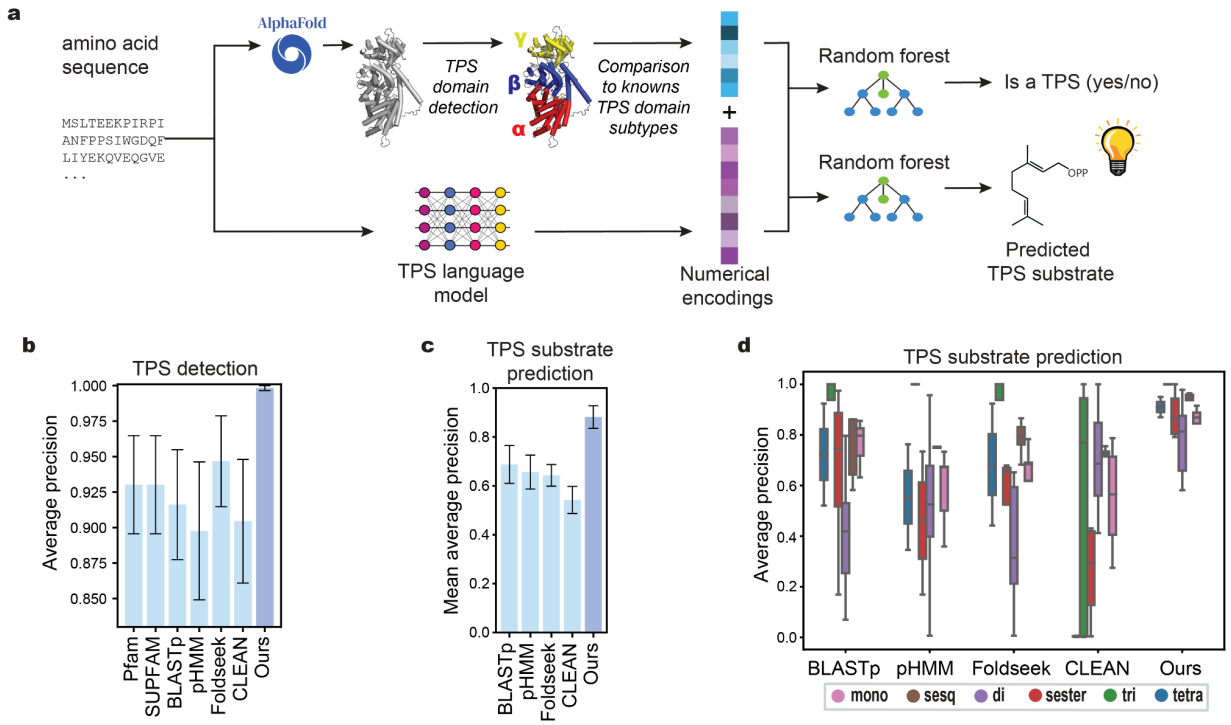


Fig. 3 | Our predictive pipeline outperforms existing methods on tasks of TPS clustering, TPS detection, and TPS substrate classification. a, An overview of our predictive pipeline, extracting numerical protein representations from the TPS language model and from the segmentation of a predicted structure into TPS-specific domains. **b,** In-silico evaluation of TPS detection performance. We use average precision to summarize an area under a precision-recall curve[1]. Error bars display sample standard deviation of results on different folds[2]. **c,** In-silico evaluation of TPS substrate classification performance. **d,** In-silico TPS substrate classification performance for different TPS classes.

Notably, our TPS classification across all TPS types with the strongest performance improvement for underrepresented classes of sesterTPSs and tetraTPSs, see Fig. 3d. Technical details of our in-silico evaluation protocol and modeling are described in Online Methods.

Experimental validation

Experimental validation aimed to assess the practical value of the proposed approach for prioritization of laboratory experiments aiming to discover and characterize novel TPSs. The goal was to validate our model performance in challenging setups. As the first setup, we evaluated the accuracy of our substrate classification for uncharacterized TPS sequences that

[1] Using the precision-recall curve is a recommended evaluation protocol for comparing models capable of outputting a continuous confidence score [50]. The comparison using the area under receiver operating characteristic curve (ROC-AUC) and a recently proposed MCC-F1 score based on the MCC-F1 curve[51] can be found in Extended data Fig. 6. Our method has the best performance when comparing by any metric.

[2] We used a stratified group k-fold with groups defined based on clades from a phylogenetic tree constructed from our dataset. See Online Methods for more details.

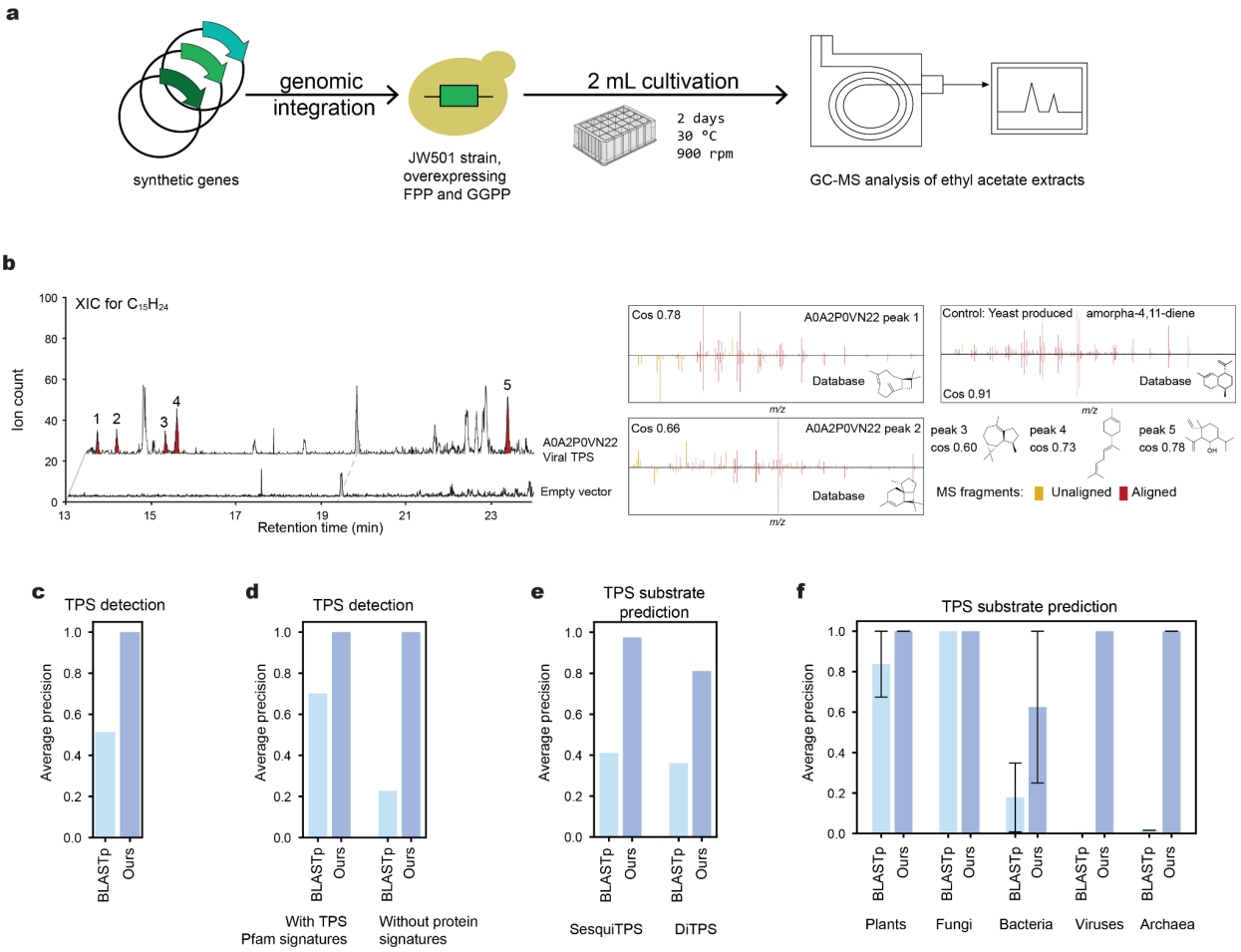


Fig. 4 | a, Experimental validation pipeline: we express predicted TPS in yeast optimized for sesqui- and diterpene production. Ethyl acetate extracts of 2 days of fermentation were therefore analyzed using GC-MS. **b,** GC-EI-MS extracted ion chromatogram for C15H24 molecular ion from A0A2P0VN22 protein displaying 5 terpene-like peaks with their closest hit in NIST-EI library. For reference, amorpha-4,11-diene has been produced in yeast to control NIST EI library retrieval from our analyses. The sesquiterpenes produced by A0A2P0VN22 range from 0.60 to 0.78 cosine similarity, whereas scores> 0.9 are expected for compounds present in the library. **c,** Experimental evaluation of TPS detection. We compared our pipeline against BLASTp matches in our TPS database, which showed the second-best mean average precision in in-silico evaluation. **d,** Experimental TPS detection performance separately on entries with TPS-specific Pfam/SUPFAM domains versus entries without any InterPro-scan integrated protein signature. Our method generalizes to novel entries without known bioinformatic signatures. **e,** Experimental validation of substrate prediction performance. **f,** Experimental evaluation of substrate prediction per different kingdoms. Our method generalizes to non-typical kingdoms.

were phylogenetically the most distant from any characterized TPS. For the second challenging evaluation setup, we accessed the capabilities of our models to detect TPS activity and predict substrates for sequences without any known protein signature and coming from non-plant organisms.

For the first evaluation setup, we mined large sequence databases for proteins matching the TPS-specific Pfam and SUPFAM domains[29,30]. From the mined putative sequences and characterized TPS sequences, we constructed a phylogenetic tree and picked uncharacterized sequences that were phylogenetically the most distant to any characterized TPS. See Extended data Fig. 10a and Online Methods for more details.

The selected top nine evolutionary distant uncharacterized sequences were then scored with our predictive pipeline. We experimentally validated the obtained TPS substrate predictions via heterologous expression in *Saccharomyces cerevisiae* JWY501. The experimental validation pipeline is summarized in Fig. 4a, and more details are described in the following sections. Due to the availability of *Saccharomyces cerevisiae* strains overexpressing FPP, the substrate of sesquiTPS, and GGPP, the substrate of diTPS, we validated our method performance on these two major TPS classes by searching molecular ions corresponding to either C15H24 or C20H32 scaffolds in gas chromatography or liquid chromatography hyphenated to mass spectrometer detector. We confirmed the practical value of our predictive approach, see Fig. 4c-f.

As for the second hard evaluation setup, we demonstrated the utility of our predictive models in discovering novel active TPS, which would be impossible to spot in sequence databases without our method. We report the first ever experimentally confirmed TPSs in Archea. To achieve that, we used the developed predictive pipeline to screen UniRef50[28]. From the results of our UniRef50 screening, we selected hits without any known protein function, domain, or family. Specifically, we filtered all hits containing any out of more than 50.000 protein signatures integrated into InterProScan[25], including domains of unknown function. Moreover, we made the evaluation setup even more challenging by focusing on hits from underrepresented kingdoms or kingdoms missing completely in the curated dataset of the characterized TPS.

Notably, our method is the first to reveal functional terpene cyclization in the Archaea, one of the major domains of life[52]. Before our work, it was believed that Archaea could form prenyl monomers but could not perform terpene cyclization [31]. Thanks to the cyclization, terpenoids are the largest and most diverse class of natural products. The history of TPS biosynthesis origins is crucial in establishing biochemistry in its current form[31]. Yet, no archeal TPS has been reported before. Our predictive pipeline sheds light on the ancient history of TPS biosynthesis. Powered by our predictive models, we are the first to discover and report three experimentally confirmed active TPSs in Archaea cyclizing FPP (A0A5E4I9B1, A0A0E3NXY0) and GGPP (A0A537EJD0). More details on the discovered archaeal TPSs can be found in Fig. 5.
Next, we detected and experimentally confirmed a viral sesquiTPS (A0A2P0VN22). At the time of experimental confirmation of its activity, no viral TPS has been reported. During the preparation of this manuscript, a detailed experimental characterization of another viral TPS was published[53]. The first published viral TPS and our viral TPS have a BLAST percentage identity of 20.27%. From the standpoint of computational TPS characterization, the first viral TPS reported in the related work[53] contains TPS-specific Pfam and SUPFAM domains. In contrast, the viral TPS detected by our predictive method has no known domains. It demonstrates the generalization capability of our models on the understudied TPSs.

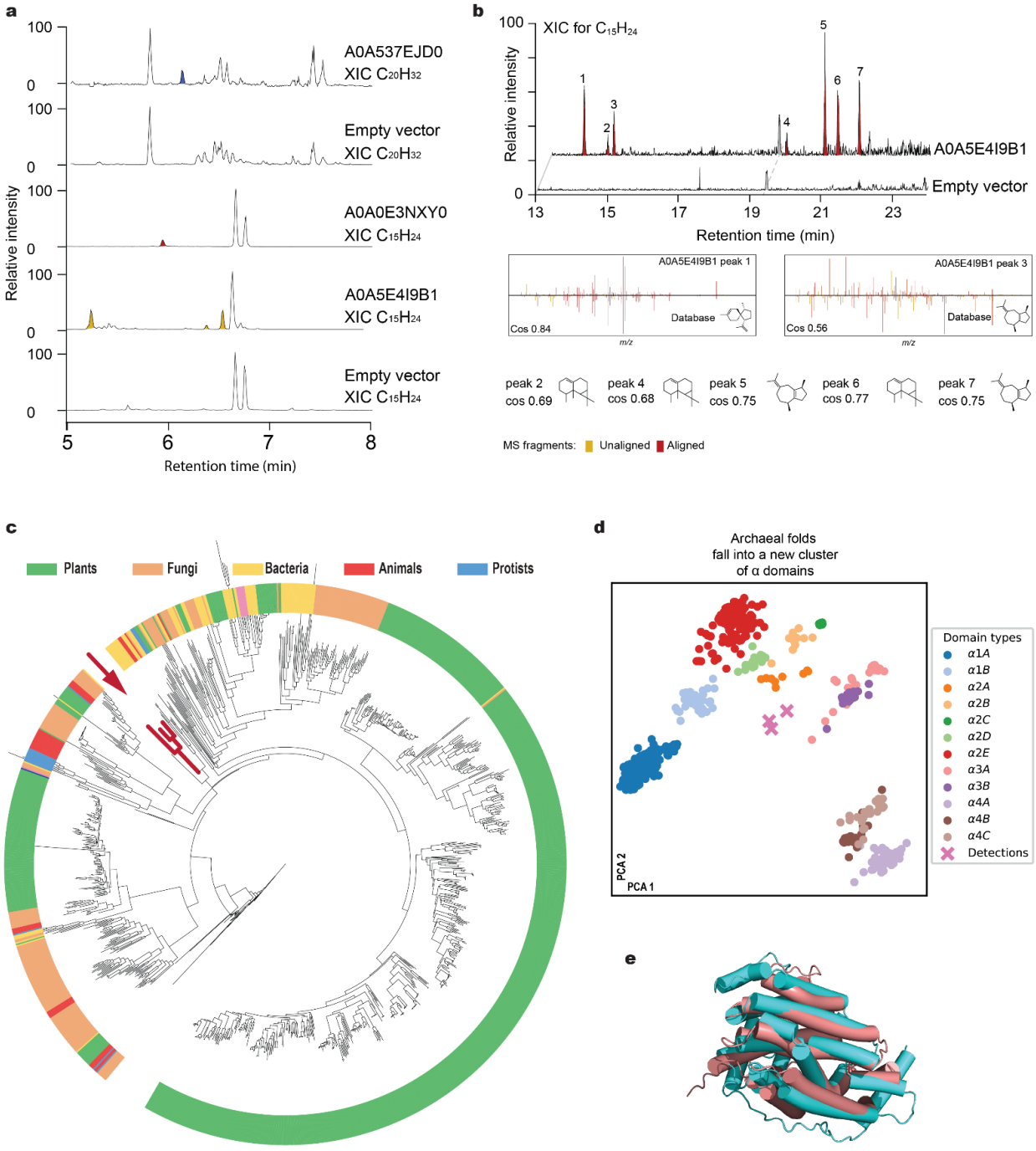


Fig. 5 | First reported Archaeal functional terpene cyclization a, LC-ESI-HRMS extracted ion chromatograms for molecular ion C15H24 (sesquiterpene) or C20H32 (diterpene). The archaeal protein expression in yeast leads to the production of terpene-like peaks compared to control yeast with an empty

plasmid. **b,** GC-EI-MS extracted ion chromatogram for C15H24 molecular ion from A0A5E4I9B1 protein displaying 7 terpene-like peaks with their closest hit in the NIST-EI library. For reference, amorpha-4,11-diene has been produced in yeast to control NIST EI library retrieval from our analyses. The sesquiterpenes produced by A0A5E4I9B1 range from 0.56 to 0.77 cosine similarity, whereas scores> 0.9 are expected for compounds present in the library. **c,** Phylogenetic analysis of our TPS dataset with the discovered archaeal TPS cyclases. It reveals that the isolated Archaeal clade is the closest to bacterial TPSs. **d,** PCA-based visualization of clustering all TPS α domains together with Archaeal α domains (PCA1 and PCA2 capture 91% of variance). The discovered archaeal TPSs have α architecture. One can see that archaeal α domains fall into a new cluster suggesting a previously unobserved α domain subtype. **e,** The closest structural match of archaeal α domains to known TPS α domains. The matched archaeal protein has UniProt accession A0A0E3NXY0, and the closest match from known TPSs has assession A0A385AJM7. A0A385AJM7 has an architecture of the α2E subtype. TM-score of the match is 0.76. For reference, TM-scores of pairwise structural alignment of archaeal TPSs range from 0.85 to 0.98. **f,** The closest sequence match of archaeal TPSs against known TPSs using BLASTp. The percentage identity of 20.49% shows that we discovered archaeal TPSs thanks to our model's ability to generalize to unexplored parts of the sequence space.

Finally, we experimentally confirmed the predicted activity for three bacterial TPS, two of which are acting upon FPP (A0A2H0W7N1, A0A450SFA8), and one acts upon GGPP (A0A5S9IQ85). In experimental validation, we focused on difficult bacterial TPSs, as the results of our UniRef50 screening suggest that bacterial TPSs were underrepresented in the dataset of characterized TPSs. We compared distributions of TPS kingdom proportions between the curated dataset and the results of the screening, see Extended data Fig. 7. As a result of Pearson's chi-square test[3], the null hypothesis was rejected in favor of the alternative that there is a discrepancy in kingdom proportions between the curated dataset of characterized TPSs and our TPS detections from UniRef50. The most notable difference in kingdom proportions suggested that bacterial TPSs are significantly understudied and, as a result, underrepresented in our database. This observation is aligned with recent remarkable findings about TPSs in bacteria[54,55]. To sum up, we validated the capability of our model to generalize to underrepresented TPS by focusing on archaeal, viral, and bacterial hits from our model, which had no known protein signatures.

For each such ML-detected sequence, we cloned the sequence downstream of tdh3 promoter and upstream of ssa1 terminators into the genome *S. cerevisiae* JWY501 and used the pESC plasmid (restoring *ura3* prototrophy) as a control. (Table S1, S2). Then, we cultivated the strains bearing the terpene synthase candidate in selective media with 10 % glucose and 90 % galactose to induce overexpression of GGPP. We proceeded to the analysis of ethyl acetate extract of the 48h culture by GC-EI-MS and LC-ESI-MS.

For GC-EI-MS data, we generated the extracted ion chromatogram for *m/z* 204.2 for sesquiterpenes and *m/z* 272.2 for diterpene. Then, we compared the extracted ion chromatogram of each *S. cerevisiae* JWY501 strain expressing a terpene synthase candidate against the extracted ion chromatogram of *S. cerevisiae* JWY501 with the empty ESC plasmid.

[3] Please note that we also tested a hypothesis that there is no discrepancy in TPS types proportions between the dataset of characterized TPS and the detections in UniRef50. To countercount the effect of multiple comparisons we performed repeated Pearson's chi-square tests with Holm correction.

We detected a new signal at *m/z* 272.2 for the strain A0A5S9IQ85 (FigSI_WetLab1_a). For four enzymes, A0A2POVN22, A0A5E4I9B1, A0A2H0W7N1, A0A450SFA8 we detected new unique signals at *m/z* 204.2 (FigSI_WetLab1_b).

For LC-ESI-MS data, we generated the extracted ion chromatogram for m/z 205.1951 (with 5 ppm error). Within the 5 ppm accuracy, [C15H24 +H]+ is the only possible molecular formula for *m/z* 205.1951. We also generated the extracted ion chromatogram for *m/z* 273.2577 (with a 5 ppm error). Within the 5 ppm accuracy, [C20H32 +H]+ is the only possible molecular formula for *m/z* 273.2577. Then, we compared the extracted ion chromatogram of each *S. cerevisiae* JWY501 strain expressing a terpene synthase candidate against the extracted ion chromatogram of *S. cerevisiae* JWY501 with the empty ESC plasmid. We detected a new signal at *m/z* 273.2577 for enzyme A0A537EJD0 (FigSI_WetLab2_a). Then for *m/z* 205.1951 we detected new signals for four enzymes A0A2H0W7N1, A0A2P0VN22, A0A0E3NXY0, A0A5E4I9B1 (FigSI_WetLab2_b).

Overall, out of 17 tested putative protein sequences, we detected sesquiterpenes or diterpenes for 7 of them. Two signals were only detected using GC-EI-MS (diterpene from A0A5S9IQ85 , sesquiterpene from A0A450SFA8), two signals were only detected by LC-ESI-HRMS (diterpene from A0A537EJD0, sesquiterpene from A0A0E3NXY0) and 3 signals were detected in both GC-EI-MS and LC-ESI-HRMS.

The results of experimental validation of our models' performance on the most difficult UniRef50 entries without any protein signature integrated into InterProScan and coming from kingdoms without previously reported TPS activity indicate that our model can discover new terpene synthases currently not captured by existing databases nor bioinformatical tools.

Discussion

Characterization of terpene synthases is of high practical importance, as TPSs are responsible for the largest class of natural products, including first-line medicines or prevalent flavoring agents[4–6,8,9]. While existing computational tools[29,30,56] assist in detecting the most abundant TPS classes (monoTPSs, sesquiTPSs, and diTPSs), in-silico characterization of underrepresented TPS types has remained an unsolved challenge. Yet, rare TPS classes like sesterTPSs have biomedically important activities, including cancer cell growth suppression, modulation of receptor signaling, antimicrobial effects, or analgesic properties[57,58].

Here, we presented a method for accurately detecting substrates corresponding to six TPS classes: monoTPSs, sesquiTPSs, diTPSs, sesterTPSs, triTPSs, and tetraTPSs. This result is based on several pillars of our work. Firstly, we curated a rich dataset of characterized TPSs. The dataset comprehensively covers different TPS types, including underrepresented sesterTPSs and tetraTPSs. While the detection of triTPSs was impossible with state-of-the-art methods, triTPSs are so well represented in our dataset that established approaches re-trained on the new data have high accuracy of triTPS detection out of the box. Next, we further improved the classification of TPSs by developing computational techniques to efficiently leverage the latest machine learning advancements, AlphaFold2[17] and protein language models

(PLMs)[26]. Our carefully designed procedure for the segmentation of a protein structure into TPS-specific domains enabled the comparison of corresponding structural modules between each other instead of working with the whole enzyme at once. That particularly boosted the performance of sesterTPSs and tetraTPSs detection. SesterTPSs and tetraTPSs are represented with too few sequences in the available databases, which makes modeling difficult for sequence-based methods.

We used our accurate predictive models to screen a large protein sequence space given by the UniRef50[28] set produced by the UniProt consortium[32]. A comparison of the screening results with our comprehensive dataset of characterized TPS sequences suggests that bacterial TPSs are significantly understudied. This supports recent findings about TPSs in bacteria, signifying the importance of their deeper exploration[54,55].
Next, we focused on our predictions for challenging UniRef50 entries having no protein signature, function, domain, or family integrated into InterProScan[25]. Experimentally, by heterologous expression in Saccharomyces cerevisiae, we confirmed the predicted activity for 6 out of 9 TPS hits containing Pfam domains and for 7 out of 17 challenging TPS hits without any protein sequence, including archeal TPSs. Before, it was believed that Archaea could not perform FPP or GGPP cyclization[31]. We are the first to report three experimentally confirmed active TPSs in Archaea acting upon FPP (A0A5E4I9B1, A0A0E3NXY0) and GGPP (A0A537EJD0).

Furthermore, the developed procedure for automatic and reliable detection of TPS domains enabled novel biological insights about TPSs. We discovered subtypes of known structural domains in TPSs using unsupervised machine learning. Rigorous analysis of the discovered subtypes revealed the uniqueness of one domain subgroup. As the subgroup is a major determinant of triTPS activity, as it has a distinct structural fold and a unique biochemical profile, we propose to differentiate it from other domain types by naming it the δ domain. By leveraging techniques of explainable AI, we derived a sequence motif uniquely identifying the new δ domain. The discovered motif was reported to be responsible for the stability of a Human lanosterol synthase[46], and mutations in the motif regions are associated with genetic disorders such as hypotrichosis simplex or cataract[47]. This finding provides a verification of the motif's biological importance for the function of triTPSs containing the δ domain.
Our analysis of the TPS domain configurations enables novel insights into TPS biochemistry in a broader context. We observe that all sesterTPSs characterized to date have αα architecture, and the first α domain is analogous to sesquiTPS α architecture, while the second sesterTPS α domain is very distinct from all TPS α domains, resembling the isoprenyl diphosphate synthase domain.

The curated dataset, representative structures of discovered TPS domain subtypes, and our models are freely available via a web server, and we anticipate it will speed up the discovery of TPSs with novel properties. Likewise, we make the source code of the developed methods publicly available with the potential to be further expanded to other enzyme families and thus accelerate biological discoveries. Furthermore, we expect the development of ML pipelines on

top of the published TPS dataset and the proposed method to enable full computational characterization of the whole TPS enzyme family.

Code availability

The source code of this study is freely available on GitHub (https://github.com/SamusRam/TPS_ML_Discovery).

References

1. Caputi, L. & Aprea, E. Use of terpenoids as natural flavouring compounds in food industry. *Recent Pat. Food Nutr. Agric.* **3**, 9–16 (2011).
2. Su, X.-Z. & Miller, L. H. The discovery of artemisinin and the Nobel Prize in Physiology or Medicine. *Sci. China Life Sci.* **58**, 1175–1179 (2015).
3. Liu, K., Zuo, H., Li, G., Yu, H. & Hu, Y. Global research on artemisinin and its derivatives: Perspectives from patents. *Pharmacol. Res.* **159**, 105048 (2020).
4. Chadwick, M., Trewin, H., Gawthrop, F. & Wagstaff, C. Sesquiterpenoids lactones: benefits to plants and people. *Int. J. Mol. Sci.* **14**, 12780–12805 (2013).
5. Vasas, A. & Hohmann, J. Euphorbia Diterpenes: Isolation, Structure, Biological Activity, and Synthesis (2008–2012). *Chem. Rev.* **114**, 8579–8612 (2014).
6. Zhang, Y. *et al.* Tanshinones: sources, pharmacokinetics and anti-cancer activities. *Int. J. Mol. Sci.* **13**, 13621–13666 (2012).
7. Gallego-Jara, J., Lozano-Terol, G., Sola-Martínez, R. A., Cánovas-Díaz, M. & de Diego Puente, T. A Compressive Review about Taxol®: History and Future Challenges. *Molecules* **25**, (2020).
8. James, J. T. & Dubery, I. A. Pentacyclic triterpenoids from the medicinal herb, Centella asiatica (L.) Urban. *Molecules* **14**, 3922–3941 (2009).
9. Tomko, A. M., Whynot, E. G., Ellis, L. D. & Dupré, D. J. Anti-Cancer Potential of Cannabinoids, Terpenes, and Flavonoids Present in Cannabis. *Cancers* **12**, (2020).
10. Mewalal, R. *et al.* Plant-Derived Terpenes: A Feedstock for Specialty Biofuels. *Trends*

Biotechnol. **35**, 227–240 (2017).

11. Pahima, E., Hoz, S., Ben-Tzion, M. & Major, D. T. Computational design of biofuels from terpenes and terpenoids. *Sustainable Energy Fuels* **3**, 457–466 (2019).
12. Quílez del Moral, J. F., Pérez, Á. & Barrero, A. F. Chemical synthesis of terpenoids with participation of cyclizations plus rearrangements of carbocations: a current overview. *Phytochem. Rev.* **19**, 559–576 (2020).
13. Kathiravan, G., Sureban, S. M., Sree, H. N., Bhuvaneshwari, V. & Kramony, E. Isolation of anticancer drug TAXOL from Pestalotiopsis breviseta with apoptosis and B-Cell lymphoma protein docking studies. *J. Basic Clin. Physiol. Pharmacol.* **4**, 14–19 (2012).
14. Zhang, C. & Hong, K. Production of Terpenoids by Synthetic Biology Approaches. *Front Bioeng Biotechnol* **8**, 347 (2020).
15. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
16. Elnaggar, A. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
17. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
18. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
19. Bordin, N. *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol* **6**, 160 (2023).
20. Barrio-Hernandez, I. *et al.* Clustering predicted structures at the scale of the known protein universe. *Nature* (2023) doi:10.1038/s41586-023-06510-w.
21. Durairaj, J. *et al.* Integrating structure-based machine learning and co-evolution to investigate specificity in plant sesquiterpene synthases. *PLoS Comput. Biol.* **17**, e1008197 (2021).

22. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
23. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
24. Christianson, D. W. *Structural and Chemical Biology of Terpenoid Cyclases*. (2017).
25. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
26. Elnaggar, A. *et al.* Ankh ☥: Optimized protein language model unlocks general-purpose modelling. *bioRxiv* (2023) doi:10.1101/2023.01.16.524265.
27. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
28. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
29. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–41 (2004).
30. Pandit, S. B. *et al.* SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinformatics* **5**, 28 (2004).
31. Correction to: four billion years of microbial terpenome evolution. *FEMS Microbiol. Rev.* **47**, (2023).
32. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
33. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* 2022.07.20.500902 (2022) doi:10.1101/2022.07.20.500902.
34. Gao, Y., Honzatko, R. B. & Peters, R. J. Terpenoid synthase structures: a so far incomplete view of complex catalysis. *Nat. Prod. Rep.* **29**, 1153–1175 (2012).
35. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional

networks. *Nat. Commun.* **12**, 1–14 (2021).

36. Köksal, M., Jin, Y., Coates, R. M., Croteau, R. & Christianson, D. W. Taxadiene synthase structure and evolution of modular architecture in terpene biosynthesis. *Nature* **469**, 116–120 (2010).
37. Karunanithi, P. S. & Zerbe, P. Terpene Synthases as Metabolic Gatekeepers in the Evolution of Plant Terpenoid Chemical Diversity. *Front. Plant Sci.* **10**, 1166 (2019).
38. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
39. McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017).
40. Nazaruk, J. & Borzym-Kluczyk, M. The role of triterpenes in the management of diabetes mellitus and its complications. *Phytochem. Rev.* **14**, 675–690 (2015).
41. Gill, B. S., Kumar, S. & Navgeet. Triterpenes in cancer: significance and their influence. *Mol. Biol. Rep.* **43**, 881–896 (2016).
42. Garg, A., Sharma, R., Dey, P., Kundu, A. & Kim, H. S. Analysis of triterpenes and triterpenoids. *Recent advances in* (2020).
43. Xu, F. *et al.* Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. in *Natural Language Processing and Chinese Computing* 563–574 (Springer International Publishing, 2019).
44. Burley, S. K. *et al.* Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol.* **1607**, 627–641 (2017).
45. Cerqueira, N. M. F. S. A. *et al.* Cholesterol Biosynthesis: A Mechanistic Overview. *Biochemistry* **55**, 5483–5506 (2016).
46. Chen, X. & Liu, L. Congenital cataract with LSS gene mutations: a new case report. *J. Pediatr. Endocrinol. Metab.* **30**, 1231–1235 (2017).
47. Zhao, B. *et al.* A novel homozygous mutation in LSS gene possibly causes hypotrichosis

simplex in two siblings of a Tibetan family from the western Sichuan province of China. *Front. Physiol.* **13**, 992190 (2022).

48. Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).
49. Hasan, M. A. M., Nasser, M., Ahmad, S. & Molla, K. I. Feature selection for intrusion detection using random forest. *Journal of information security* **7**, 129–140 (2016).
50. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
51. Cao, C., Chicco, D. & Hoffman, M. M. The MCC-F1 curve: a performance evaluation technique for binary classification. *arXiv [stat.ML]* (2020).
52. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **16**, 120 (2018).
53. Jung, Y. *et al.* Function and Structure of a Terpene Synthase Encoded in a Giant Virus Genome. *J. Am. Chem. Soc.* **145**, 25966–25970 (2023).
54. Yamada, Y. *et al.* Terpene synthases are widely distributed in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 857–862 (2015).
55. Duan, Y.-T. *et al.* Widespread biosynthesis of 16-carbon terpenoids in bacteria. *Nat. Chem. Biol.* **19**, 1532–1539 (2023).
56. Priya, P., Yadav, A., Chand, J. & Yadav, G. Terzyme: a tool for identification and analysis of the plant terpenome. *Plant Methods* **14**, 4 (2018).
57. Li, K. & Gustafson, K. R. Sesterterpenoids: chemistry, biology, and biosynthesis. *Nat. Prod. Rep.* **38**, 1251–1281 (2021).
58. Ebada, S. S., Lin, W. & Proksch, P. Bioactive sesterterpenes and triterpenes from marine sponges: occurrence and pharmacological significance. *Mar. Drugs* **8**, 313–346 (2010).
59. Bansal, P. *et al.* Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **50**, D693–D700 (2022).

60. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
61. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).
62. Lesburg, C. A., Zhai, G., Cane, D. E. & Christianson, D. W. Crystal structure of pentalenene synthase: mechanistic insights on terpenoid cyclization reactions in biology. *Science* **277**, 1820–1824 (1997).
63. Starks, C. M., Back, K., Chappell, J. & Noel, J. P. Structural basis for cyclic terpene biosynthesis by tobacco 5-epi-aristolochene synthase. *Science* **277**, 1815–1820 (1997).
64. DeLano, W. L. & Bromberg, S. PyMOL user's guide. *DeLano Scientific LLC* **629**, (2004).
65. Barozet, A., Chacón, P. & Cortés, J. Current approaches to flexible loop modeling. *Curr Res Struct Biol* **3**, 187–191 (2021).
66. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
67. Park, H.-S. & Jun, C.-H. A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.* **36**, 3336–3341 (2009).
68. Shahapure, K. R. & Nicholas, C. Cluster Quality Analysis Using Silhouette Score. in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* 747–748 (IEEE, 2020).
69. BFD. https://bfd.mmseqs.com/.
70. Pundir, S., Martin, M. J. & O'Donovan, C. UniProt Protein Knowledgebase. *Methods Mol. Biol.* **1558**, 41–55 (2017).
71. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
72. Matasci, N. *et al.* Data access for the 1,000 Plants (1KP) project. *Gigascience* **3**, 17 (2014).
73. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics.

Nucleic Acids Res. **40**, D1178–86 (2012).

74. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
75. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–37 (2011).
76. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
77. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* (SciPy, 2010). doi:10.25080/majora-92bf1922-011.
78. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
79. Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**, 2272–2274 (2020).
80. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
81. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
82. Lee, M. E., DeLoache, W. C., Cervantes, B. & Dueber, J. E. A Highly Characterized Yeast Toolkit for Modular, Multipart Assembly. *ACS Synth. Biol.* **4**, 975–986 (2015).
83. Lõoke, M., Kristjuhan, K. & Kristjuhan, A. Extraction of genomic DNA from yeasts for PCR-based applications. *Biotechniques* **50**, 325–328 (2011).
84. Gietz, R. D. & Schiestl, R. H. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.* **2**, 31–34 (2007).
85. Lau, A. M., Kandathil, S. M. & Jones, D. T. Merizo: a rapid and accurate protein domain

segmentation method using invariant point attention. *Nat. Commun.* **14**, 1–11 (2023).

86. Mehta, V., Bawa, S. & Singh, J. Analytical review of clustering techniques and proximity measures. *Artificial Intelligence Review* **53**, 5995–6023 (2020).

87. Wong, J. *et al.* High-titer production of lathyrane diterpenoids from sugar by engineered Saccharomyces cerevisiae. *Metab. Eng.* **45**, 142–148 (2018).

Online Methods

Dataset

We mined Swiss-Prot, the curated UniProtKB component produced by the UniProt consortium[32] for protein sequences that match the TPS-specific Pfam domains PF01397, PF03936, PF19086, PF13249, PF13243, PF01397 and SUPFAM domain combinations for Terpenoid synthases sf48239 and sf48576. Furthermore, we mined the Rhea[59] knowledgebase of biochemical reactions to retrieve TPSs and isoprenyl diphosphate synthases from Swiss-Prot. It resulted in 669 TPS sequences and 109 isoprenyl diphosphate synthase sequences. Finally, we conducted a manual literature search to retrieve 416 characterized TPSs not included in Swiss-Prot. We have also included 6 hard negatives, which are single-point mutants of TPSs without TPS activity. The final dataset contains 1194 sequences and is available online at 10.5281/zenodo.10567437.

Computational evaluation of predictive models

We use the novel TPS dataset as a source of positive TPS examples and the Swiss-Prot knowledgebase[32] as a source of negative samples (excluding TPSs, we randomly sampled 10000 proteins from Swiss-Prot). For TPSs catalyzing multiple reactions, we used only reactions corresponding to major products. In the dataset, we still included reactions for which we did not find reliable information about the product being major, but we did not use those for model training or evaluation, as this information can introduce noise into our models and evaluation. For each evaluated model, we perform multiple experiments when part of the data is hidden as a test set. We use 5-fold cross-validation, i.e., during a train/test experiment, a single fold serves as a hold-out dataset untouched in the process of training the predictive model, while the remaining four folds serve for training the predictive model.

Motivated by the pursuit of discoveries, our goal for the in-silico evaluation is to assess the ability of models to generalize learned patterns to proteins far from the training set. To that end, we split proteins into folds, minimizing the overlap of sequences between folds while preserving the proportions of different classes across folds. We used a stratified group k-fold with groups defined based on clades from a phylogenetic tree. We used IQ-Tree software[60] to create the phylogenetic tree of TPS sequences on top of MSA computed with the MAFFT program[61]. We require at most 70% of TPSs that accept a particular substrate to belong to a single clade. Otherwise, the substrate is excluded from evaluation, as the absolute majority of representatives of the TPS substrate class belong to a single fold. We still include such entries in the model creation, but they are flagged in the column “ignore_in_evaluation” in the published dataset. The folds are in the “validation_split” column in the dataset. Entries, for which we were not able to reliably identify the major product, have the value “-1” in the column “validation_split”.

We assess the performance of the models using mean average precision (mAP). The output of our models is confidence in detection/classification. In other words, we create score-based classifiers. Average precision (AP), closely related to the area under the precision-recall curve,

summarizes the performance of score-based classifiers over different detection thresholds. For score-based models and imbalanced problems, it is a recommended evaluation metric[50].

Additionally, we evaluate our models using the area under the ROC curve (ROC-AUC) and recently proposed MCC-F1 score based on the MCC-F1 curve[51]. Our pipeline also has dominant performance when using these metrics; see Extended data Fig. 6. The error bars in the evaluation plots represent standard deviation over five folds.

Discovery of structural domain subtypes

The goal is to develop a reliable method for the segmentation of a TPS structure into domains. Furthermore, once all TPS structures are segmented into domains, we aim to detect groups of similar folds within domains. The input to our computation procedure is a dataset of TPS structures. The results of the devised procedure are subtypes of established TPS domains and computation tools for analysis of novel TPSs at resolution of individual domains.

We use structural domains of characterized TPSs from related work[36] as our standards for TPS-specific structural domains. Namely, the fold of pentalenene synthase[62] represents a standard for the α domain. Next, we aligned the α domain standard to the characterized structure of 5-*epi*-aristolochene synthase[63] and used the remaining, unaligned part of the 5-*epi*-aristolochene synthase fold as a standard for the β domain. Finally, after aligning both α and β standards to the structure of taxadiene synthase[36], we defined a standard for the γ domain as the remaining unaligned part.

For sequence-independent alignment, we use PyMOL[64] and TM-align[38]. It is known that the modeling of protein loops is challenging[65]. Therefore, to align a new structure to the defined standards of the structural modules, we use only α helixes and β sheets. The TM-align algorithm does not map all residues of the target standard to the novel TPS structure. In order to overcome this issue, we post-process the mapping of residues between a structure-module standard and an unknown TPS structure by pairing all residues in the standard to corresponding residues in the novel structure. To do so, for each residue in the TPS domain standard, we find the closest domain residues with mappings reported by the TM-align algorithm and derive the local shift between the domain sequence and the novel-structure sequence based on the shift of the mapped neighboring residues. In order to have reliable domain detections, we require at least 60% of the structure-module standard to be mapped. Furthermore, we only consider residues predicted by AlphaFold2 with high confidence (pLLDT higher than 90). As a domain detection threshold, we used 0.4, i.e. structures with a TM-score above 0.4 are deemed matched. The threshold selection was selected based on the fact that with a TM-score of pairwise similarity under 0.4, there are almost no pairs of structures with the same fold according to the consensus definition of SCOP and CATH[66].

After detecting all TPS-specific domains, we process the unassigned parts of secondary structures. For each such unassigned part, we find the closest helix from any detected domain and assign the unmapped α helix to the corresponding domain if the secondary structure is not

too far from the domain in terms of physical proximity. The threshold on the maximum allowed distance from the detected domain is computed based on the domain statistics. For each helix in a particular domain, we check how far it is from the closest helix in the same domain and use the maximum distance among all domain helices as a threshold for inclusion into this domain.

After all TPS structures were segmented into domains, we used the same method for pairwise alignment of the domain detections. The TM-score of pairwise alignment was then used as a precomputed metric for K-medoids[67] and HDBSCAN[39] clustering algorithms with default parameters. To determine an optimal number of clusters for the K-medoids method, we performed Silhouette analysis[68]. Considering how structurally different β/γ domains are from α domains, for computational efficiency, we analyzed the space of α domains separately from β/γ domains. Based on Silhouette analysis, the space of β/γ domains has four apparent clusters.

On the other hand, the space of α domains does not have as straightforward clustering patterns as β/γ domains. For α domains, the optimal value of the silhouette score corresponded to two clusters. This fact is due to the uniqueness of domains found in isoprenyl diphosphate synthases. The score for four clusters is comparably high, suggesting three major groups in non-isoprenyl-diphosphate-synthase α domains. Therefore, we use K-medoids with k=4 to define four global types of α domain. We notice an increase in the silhouette score as the number of α-domain medoids goes up to fourteen, see Extended Fig. 2. It suggests the presence of smaller local subgroups of α domains. The HDBSCAN automatically determines an optimal number of clusters, excluding noise. For the space of α domains, HDBSCAN found thirteen clusters and a set of noisy points. As both methods independently suggested a similar number of α domain subclusters, we defined thirteen subtypes of α domains based on clusters automatically found by HDBSCAN.

Mining TPS-like sequences

We mined the BFD,[69] UniParc,[70] Mgnify,[71] 1KP,[72] Phytozome,[73] and NCBI Transcriptome Shotgun Assembly databases for protein sequences that match the TPS-specific Pfam domains PF01397, PF03936, PF19086, PF13249, PF13243, PF01397 and SUPFAM domain combinations for Terpenoid synthases sf48239 and sf48576. This mining procedure resulted in roughly two hundred thousand TPS-like sequences.

Predictive modeling

Using the mined TPS-like sequences (see the “Mining TPS-like sequences” section of Online Methods), we finetuned the complete encoder-decoder Ankh base model[26]. As a masking strategy, we employ unigram T5 span masking (Experiment 4 in the Ankh paper[26]) and dynamically sample new masking tokens every epoch. We set the maximum sequence length to 512 and randomly select 20% of tokens for masking, following the approach described in the Ankh paper. For optimization, we use the AdamW optimizer, a batch size of 256, and a learning rate of 5e-5. The finetuning is performed with early stopping until the training loss convergence, corresponding to approximately seven training epochs.

TPS-specific domains are detected using the procedures described in the Methods section "Discovering domain subtypes."

For the Random Forest model[48], we use Scikit-learn implementation with default parameters[74]. For Foldseek[20] and CLEAN[18] models, we use implementations from the official GitHub repositories as per instructions, with recommended or default parameters. For profile Hidden Markov models, we have followed the procedure from the Terzyme paper[56] and trained models on top of our currated dataset using the HMMER library[75]. The PSI-BLAST[22] was used with default parameters.

We conducted an ablation study of our approach, see Extended data Fig. 6. It turns out that the finetuning of the protein language model (PLM) has almost no effect on the performance of the pipeline both for TPS detection and TPS substrate prediction. Nevertheless, both PLM and structural-domain features contribute to the superior performance of our pipeline. Interestingly, PLM features are more important for TPS detection, while structural domains are more critical for TPS substrate prediction. Structural domains have a notable effect on the pipeline performance for minor TPS types like tetraTPSs or sesterTPSs, see Extended data Fig. 6c. PLM features shine for major classes like sesqTPSs and contribute to the dominant performance of the pipeline for minor classes.

UniRef50 screening was performed using PLM features only for the sake of computational efficiency for large-scale analysis. Based on the above ablation, this simplification does not corrupt the performance of our approach to the TPS detection task. However, it likely decreases the quality of TPS substrate prediction for underrepresented classes like sesterTPSs or tetraTPSs.
For the screening, we used a detection threshold on the predicted probability of TPS activity of 0.3. This threshold was selected by cross-validation, as described in the Methods section "Computational evaluation of models." The threshold corresponds to both a precision and a recall higher than 0.9.

For checking the consistency between class and kingdom proportions in the curated TPS dataset vs. in Uniref50 screening results, we performed Pearson's chi-square tests with a significance level of α=5%. We omitted all categories with less than five samples. We corrected p-values with Holm correction as we tested multiple hypotheses on our data. For statistical analysis, we used standard Python libraries SciPy[76] and Statsmodels[77].

δ domain sequence motif derivation

We detected the δ domain in the protein structures of our dataset using the developed procedure for TPS domain detection described in the section "Discovery of structural domain subtypes". The presence of the δ domain is a binary label we aim to predict.

As described above, we used the same pipeline with a TPS language model and a Random Forest to build a predictive model for detecting sequences corresponding to structures with the δ domain.

MSA was created with the MAFFT program[61]. As a model explainability technique, we used SHAP[78]. Sequence logos were created with the library Logomaker[79]. Visualization of Human lanosterol synthase and motif position was performed in PyMOL[64]. The cut-off threshold on SHAP values was optimized using the classifier's training data, and the derived motif's retrieval performance was evaluated on the held-out set.

As training data for the classifier, we used our curated TPS dataset combined with randomly sampled 10000 proteins from the Swiss-Prot knowledgebase[32].

Selecting sequences phylogenetically distant from characterizes TPSs

Using the mined TPS-like sequences (see the “Mining TPS-like sequences” section of Online Methods), an MSA was created using the MAFFT program[61]. The MSA trimming was performed using trimAl[80], and the phylogenetic tree was constructed using Fasttree2[81]. For all uncharacterized sequences, the distance to the closest characterized sequence in the tree was calculated as a sum of the lengths of the branches, i.e. the closest characterized sequence has the shortest distance.

Experimental material and methods

Strains, Growth Media.

Chemicals used for media preparation were purchased from either Sigma-Aldrich (St. Louis, Missouri, United States), Duchefa Biochemie (Haarlem, Netherlands), Lach:ner (Neratovice, Czech Republic) or Penta chemicals (Prague, Czech Republic). Solvent for metabolic sample preparation, LC and GC MS were purchased from Fisher Chemical (Waltham, Massachusetts, United States) and were LC-MS grade.
The list of all strains used in the study is available in Table S2
S. cerevisiae JWY501 derivative strains were used to express terpene synthases and produce terpenes. Selective medium (SCE) used to grow transformants contained 1.92 g.L-1 (w/v) Yeast Synthetic Drop-out Medium Supplements without Uracil (Sigma-Aldrich, Y1501), 6.7 g.L-1 Yeast nitrogen base without amino acid (Sigma-Aldrich, Y0626), and 20 g.L-1 glucose.
DH10β electrocompetent E. 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.

Growth condition.

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, Hamburg, Germany, Netherlands) sealed with AeraSeal™ (Excel Scientific, Victorville, California), at 30 °C, 800 RPM on an Eppendorf ThermoMixer® C (Eppendorf, Hamburg, Germany).
For plasmid amplification, E. coli strains were cultivated in LB medium in 24 deep well plates (CR1426, Enzyscreen, Hamburg, Germany, Netherlands) sealed with AeraSeal™ (Excel Scientific, Victorville, California), at 37 °C, 800 RPM on an Eppendorf ThermoMixer® C (Eppendorf, Hamburg, Germany).

Plasmids

The list of all plasmids used in the study is available in Table S1
All pTP plasmids were generated using GoldenGate assembly based on the Lee et al. toolkit and overhangs.[82]
Part plasmids (Table S1) use pYTK001 as a backbone. Terpene synthases DNA parts were synthesized by TwistBioscience (South San Francisco, California, United States) as gene fragments. Cassette plasmids for genomic integration (Table S4) use pYTK096 as a backbone. Plasmid extraction was achieved from 2 mL of LB of E. coli harboring plasmid overnight culture. Plasmid extraction was achieved using the QIACube robot (Quiagen, Hilden, Germany) and QIAprep Spin Miniprep Kit (27104, Quiagen) with the QIAprep miniprep “rapid” protocol. Plasmids were eluted in 50 µL ddH2O.

Polymerase chain reactions (PCR)

For colony PCR and yeast genotyping we used Phire Green Hot Start II PCR Master Mix (Thermo Fisher Scientific, Waltham, Massachusetts, United States) with the following conditions:
In 10 µL final, 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 ddH2O, and boiled for 10 minutes before being used as a DNA template.
Yeast genotyping was adapted from Lõoke et al.[83], 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 ddH2O and cell debris spined down 15 000 × g for 1 min.

GoldenGate 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, Ipswich, Massachusetts, United States), BsmBI-v2 0.5 µL (R0739L, New England Biolabs, Ipswich, Massachusetts, United States), 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 (Waltham, Massachusetts, United States). 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, Ipswich, Massachusetts, United States), BsaI-HF®v2 0.5 µL (R3733L, New England Biolabs, Ipswich, Massachusetts, United States), DNA parts 0.5 µL (20 fm), ddH2O up to 10 µL. Reactions were conducted in ProFlex™ 3 × 32-well PCR System thermocycler (Waltham, Massachusetts, United States). 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.

S. cerevisiae transformation

Budding yeast strains were generated by standard lithium acetate transformation protocol[84] 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 K2HPO4). 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 media in 96 deep well plate (CR1496, Enzyscreen, Hamburg, Germany, Netherlands) sealed with AeraSeal™ (Excel Scientific, Victorville, California), at 28 °C, 1500 RPM on an Eppendorf ThermoMixer® C (Eppendorf, Hamburg, Germany) 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, Victorville, California), 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, Hamburg, Germany) centrifuged for 5 mins, 14 100 × g and the organic phase was carefully collected and dried under N2 flow.

LC-ESI-MS

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 (H-ESI). 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.

GC-MS

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, Santa Clara, CA, USA). The samples (1 µl) were injected into 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.

Data analysis

LC-MS .raw data files were directly imported into MZmine 3.4.16. Extracted ion chromatograms for compounds of interest were generated using the raw data overview feature and exported as .pdf.
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.

Acknowledgments

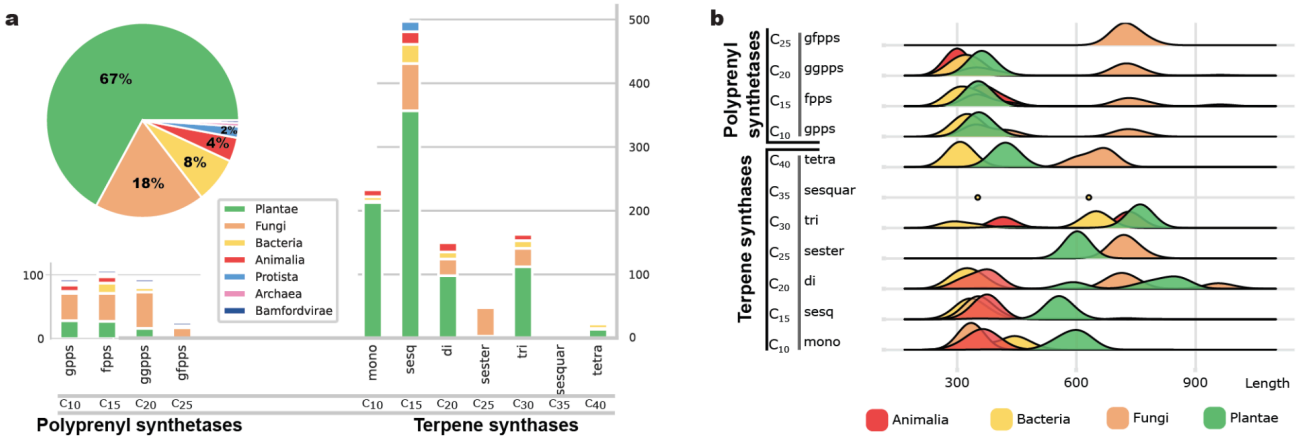
We thank Peter G Mikhael, Itamar Chinn, Sotirios C. Kampranis, Chi Zhang, Jitka Štáfková, Petr Kouba, and Regina Barzilay for their constant friendly support and inspiring, insightful discussions about this work.

R.S. 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 IOCB Fellowship program. T.P. is supported by the Czech Science Foundation (GA CR) grant 21-11563M and by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No. 891397. J.S. was supported by the European Union (ERC, FRONTIER, 101097822) and (EXA4MIND, 101092944). 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.

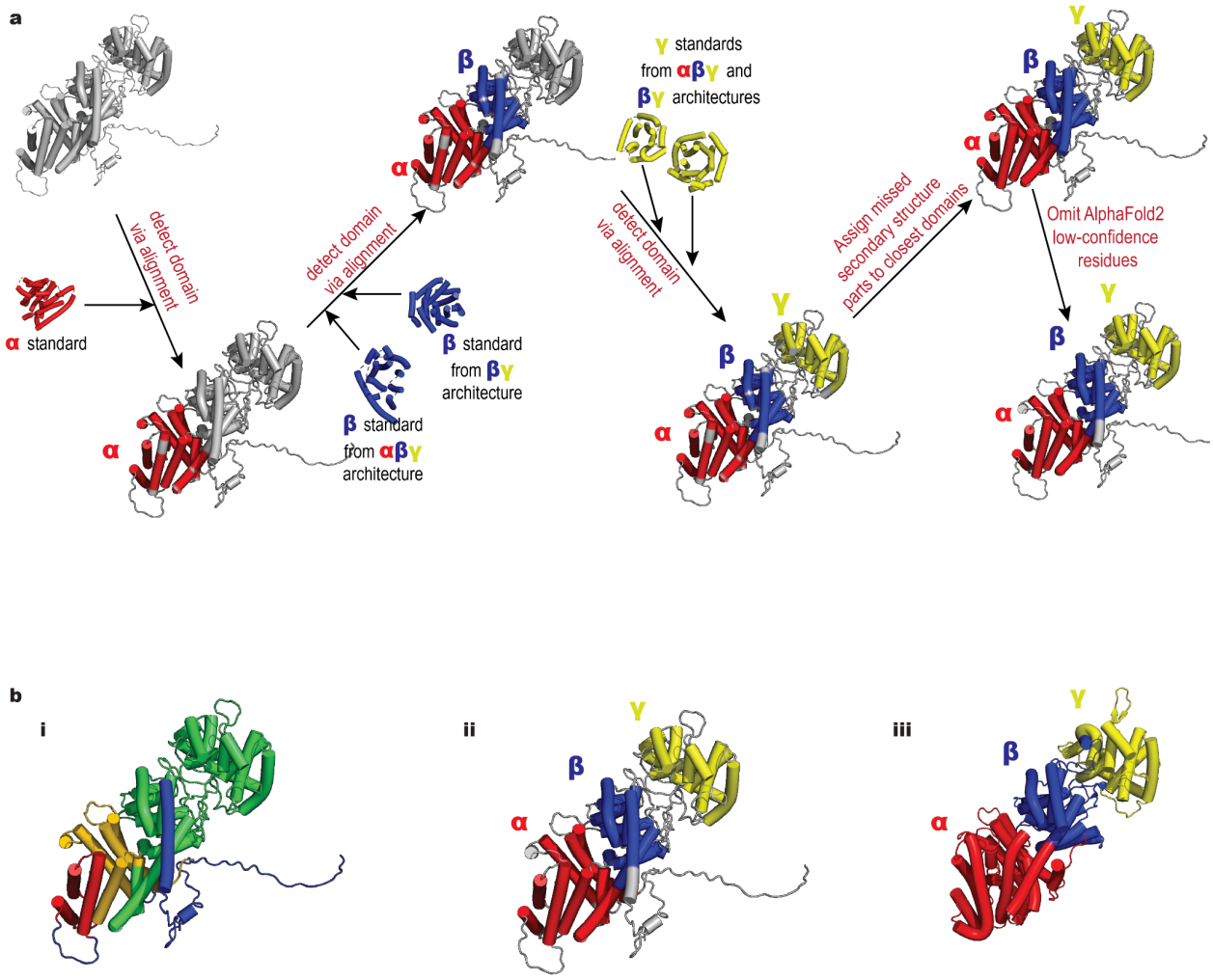
Author contributions

T.P. conceptualized the project. R.S. performed exploratory data analysis, developed structural domain segmentation, clustering; derived biological insights from domain analysis; built, trained, and evaluated predictive models; screened UniRef50 and UniProt; developed XAI-based procedure for sequence motif derivation; automatically mined characterized TPSs; wrote the manuscript with inputs from all authors. T.H. performed metabolic engineering, LC-ESI-MS, and GC-MS analysis; processed laboratory experimental data and wrote corresponding parts of the manuscript. T.Č. curated the TPS database and mined putative TPSs. H.S. 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. A.T., R.S., R.Ch., M.P., and M.E. curated the TPS database. J.S. and T.P. supervised the project.

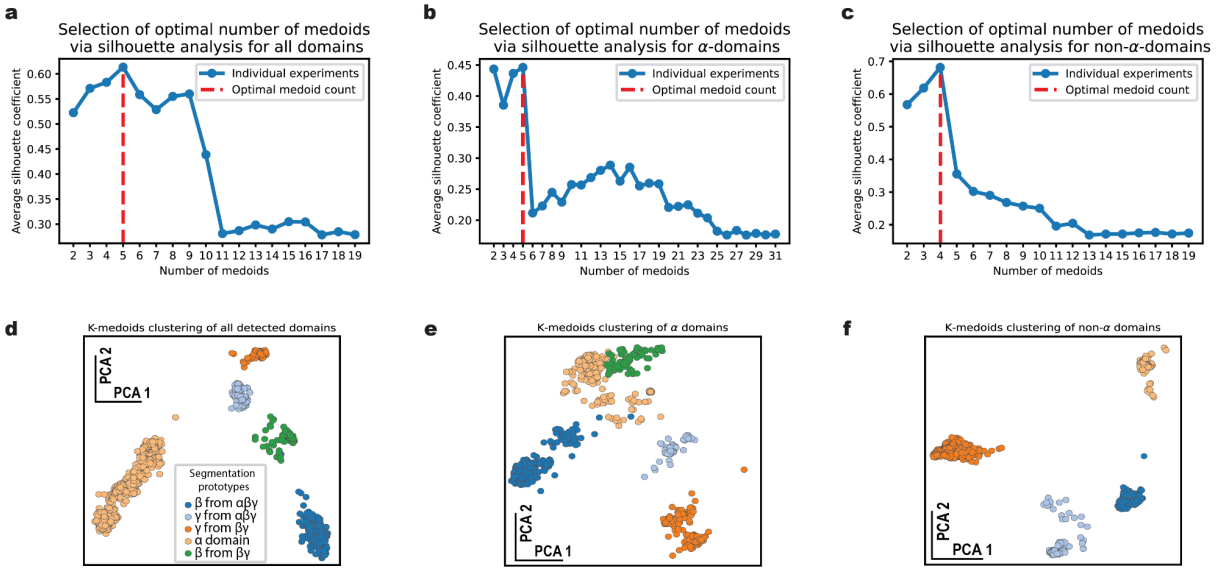
Extended Data



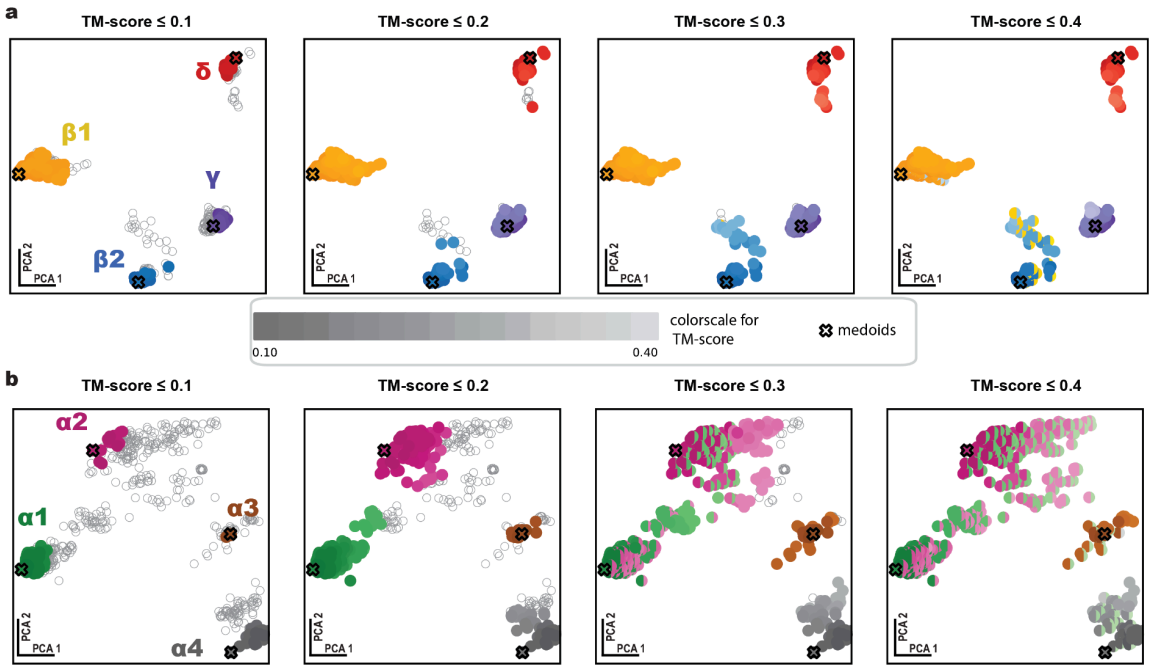
Extended Data Fig. 1 | Curated dataset and workflow overview. a, The dataset comprises more than 1,000 TPSs, three times more than the state-of-the-art TPS dataset[56]. Our dataset covers a larger variety of TPS classes, notably a significant amount of triterpene synthases. In addition, less than 70% of the TPSs come from plants, while related work [56] was based solely on plant TPSs. **b,** The novel comprehensive dataset enables population-wide insights, such as revealing differences and similarities of protein length distributions per TPS classes and kingdoms.



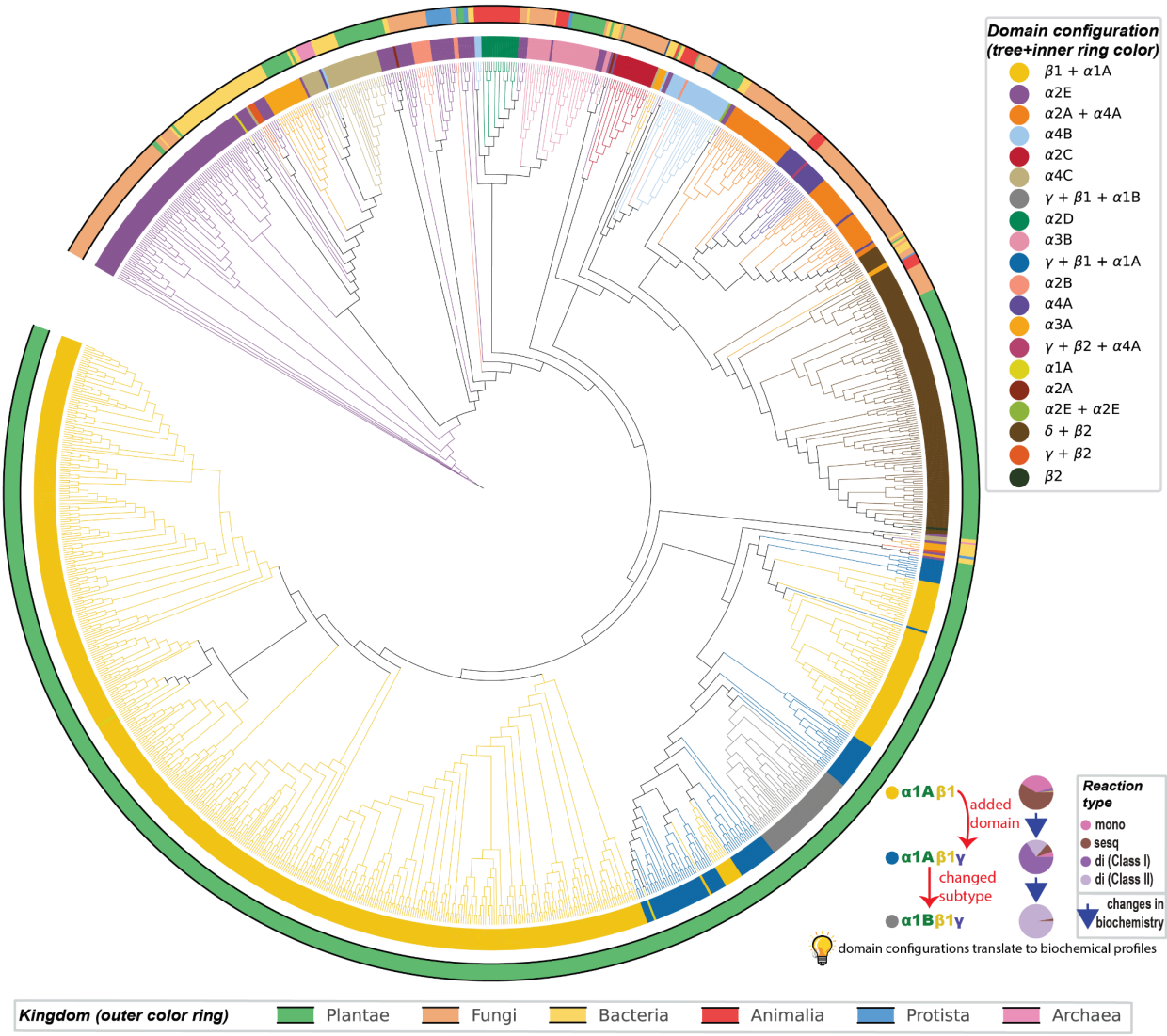
Extended Data Fig. 2 | Segmentation into domains. a, High-level overview of our pipeline for TPS structure segmentation into domains. **b,i,** State-of-the-art (SOTA) general segmentation[85] fails to partition the structure of a TPS. This is demonstrated on a randomly selected TPS with UniProt accession B9GSM9. The SOTA method does not assign domain types to individual segmented domains. **b/ii,** The result of our segmentation algorithm for the same UniProt accession B9GSM9. The TPS with αβγ architecture is segmented correctly, with assignments of corresponding domain types to each domain α, β, and γ. **b/iii,** ground truth segmentation of TPS with the αβγ architecture (UniProt accession Q41594) into CATH domains: 3p5rA01, 3p5rA02, and 3p5rA03.



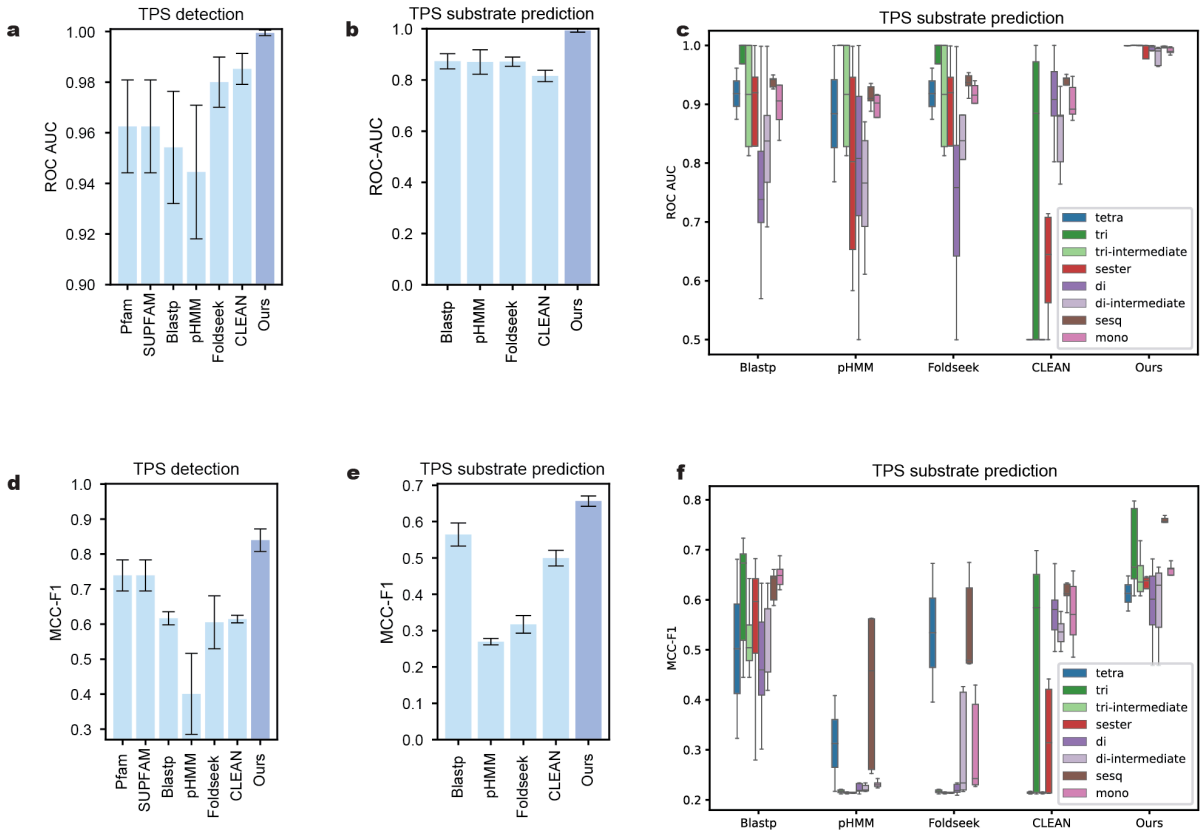
Extended Data Fig. 3 | Determining domain clusters. a-c, Silhouette analysis for determining an optimal number of global clusters **(a),** α domain clusters **(b)**, and β/γ domain clusters **(c)**. **d-f,** Principal-Components-Analysis visualization of corresponding K-medoid clusters. One can see that the global space of all TPS-specific domains has 5 distinct groups. The α domain cluster is the largest group with a complex internal structure of subclusters, while other groups did not exhibit it.



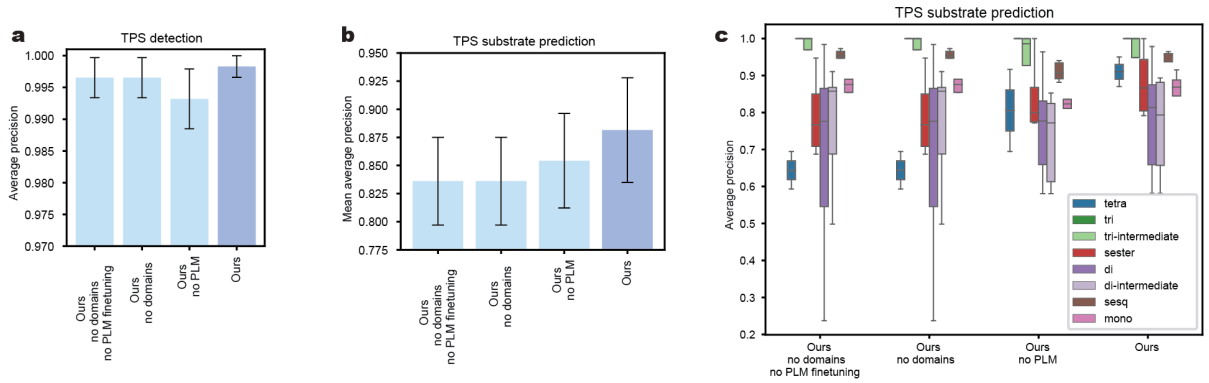
Extended Data Fig. 4 | Clustering TPSs by structural domains. a, Cluster membership under different thresholds on the maximum allowed TM-score distance from the cluster medoid. If a point can be associated with two clusters, then it has two colors, with the left one corresponding to a closer cluster. **b,** The space of possible α subtypes represents a spectrum of gradual structural variations, e.g. under threshold on TM-score of 0.4, there is a significant intermixing of individual global α subtypes shown in green, pink, brown, and gray colors. The space of β/γ domain subtypes, on the other hand, has four distinct subtypes depicted in yellow and blue for β-domain clusters, and in red and violet for γ and δ domains.



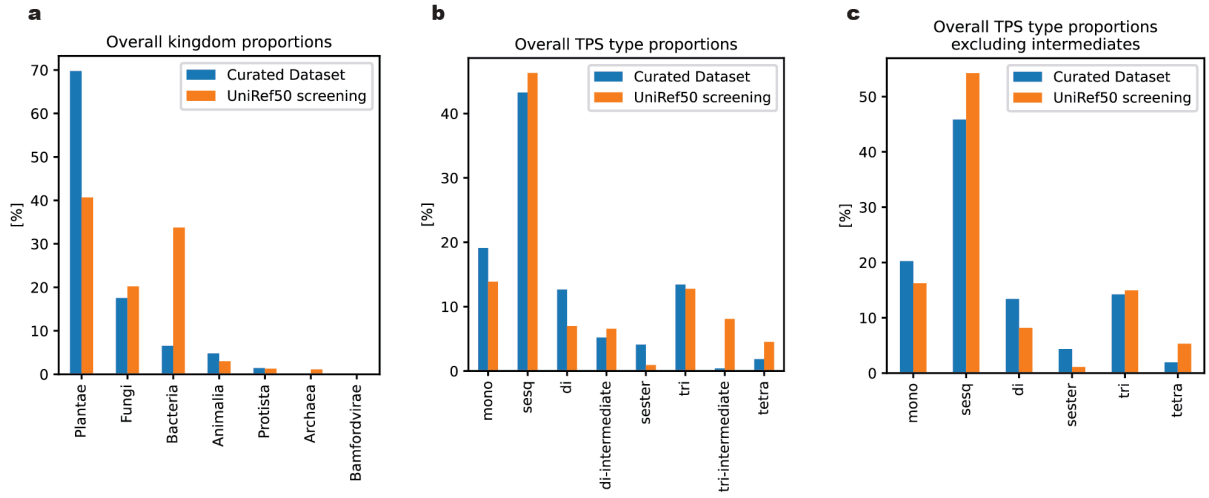
Extended Data Fig. 5 | Domain subtypes order on top of a phylogenetic tree of characterized TPSs. One can see the consistent clustering of the detected domain configurations and evolutionary relationships between the discovered domain subtypes. By comparing with Fig. 2e one can observe how changes in the domain configurations translate into modification of biochemical profile.



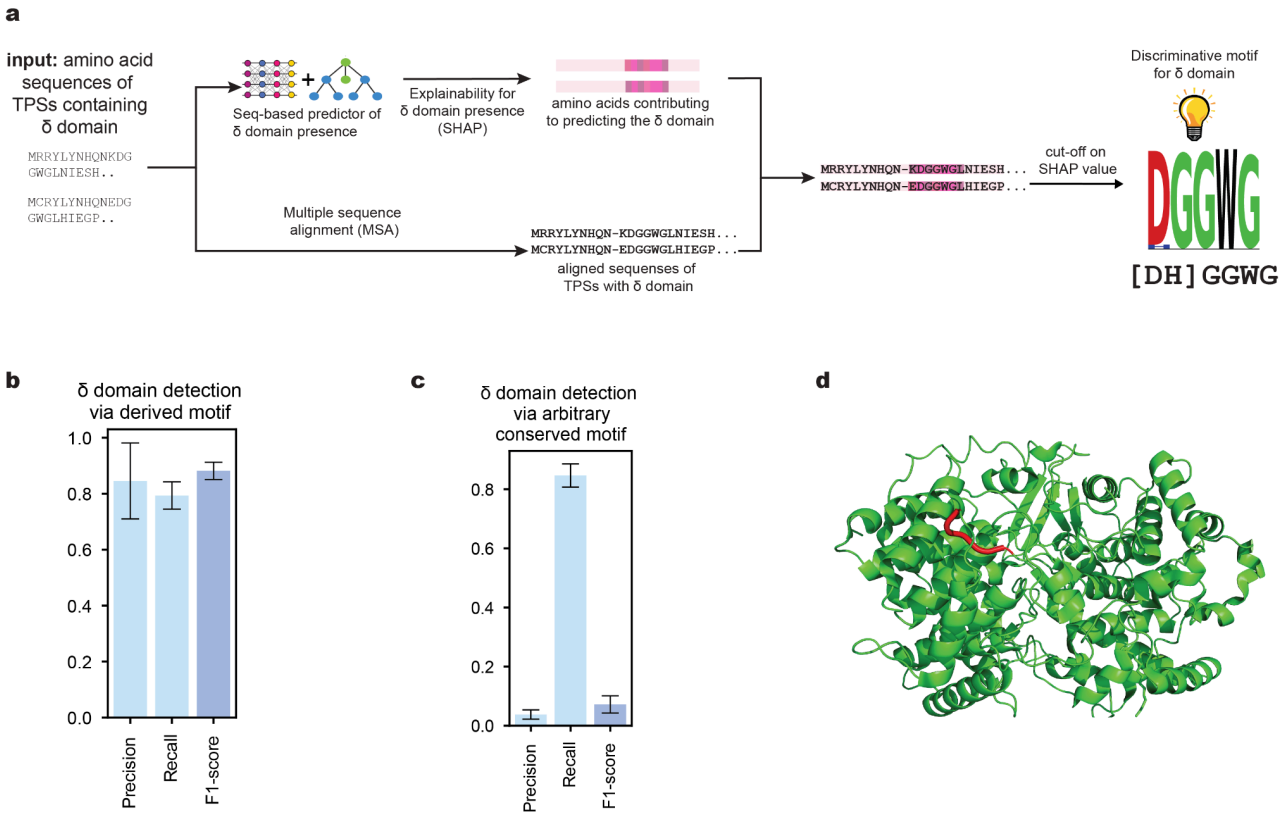
Extended Data Fig. 6 | Benchmarking performance of our TPS detection and substrate prediction methods against existing approaches using alternative metrics. a, area under the ROC curve (ROC-AUC) for TPS detection, **b,** ROC-AUC for substrate prediction, **c,** ROC-AUC for substrate prediction computed separately per different TPS types, **d-f,** analogous experiments reported using the MCC-F1 score[51]



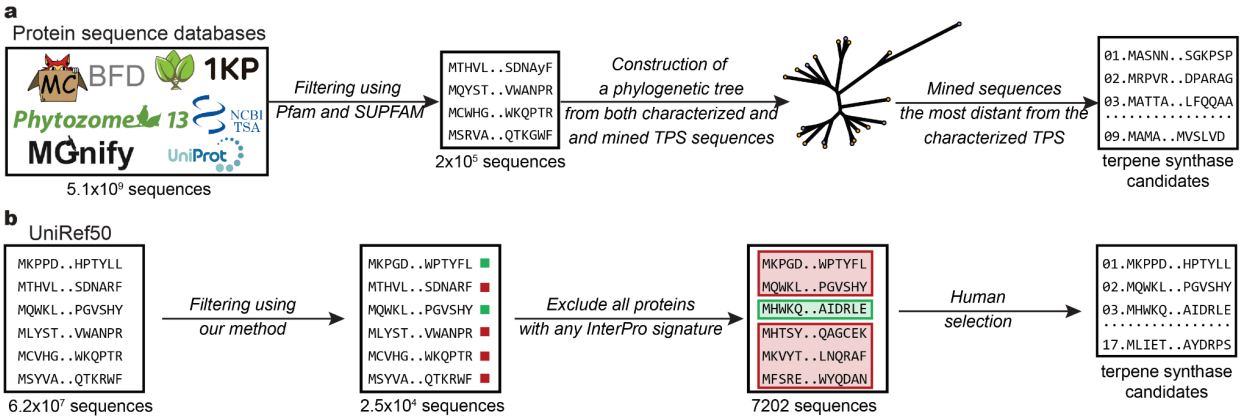
Extended Data Fig. 7 | Ablation study. a, TPS detection task, **b,** overall performance in the TPS substrate prediction task, **c,** performance in TPS substrate prediction task per TPS type. See Online Methods for the discussion.



Extended Data Fig. 8 | Differences in categorical distributions between the curated TPS dataset and results of the UniRef50 screening. a, Kingdom proportions comparison, **b,** TPS type proportion comparison, **c**, Comparison of TPS type proportions excluding intermediates.

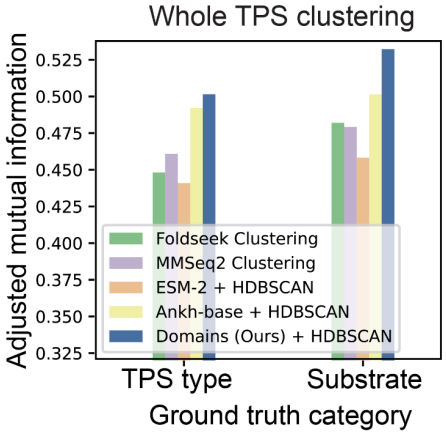


Extended Data Fig. 9 | Discovery of the δ-domain motif using predictive model explainability. a, An approach based on multiple sequence alignment (MSA) and explainable AI for the derivation of δ domain motif. **b,** The performance of δ domain retrieval with the automatically derived motif. **c,** The performance of δ domain retrieval with the arbitrarily selected conserved region of the MSA. **d,** Localization of the derived motif in Human lanosterol synthase, depicted in red.

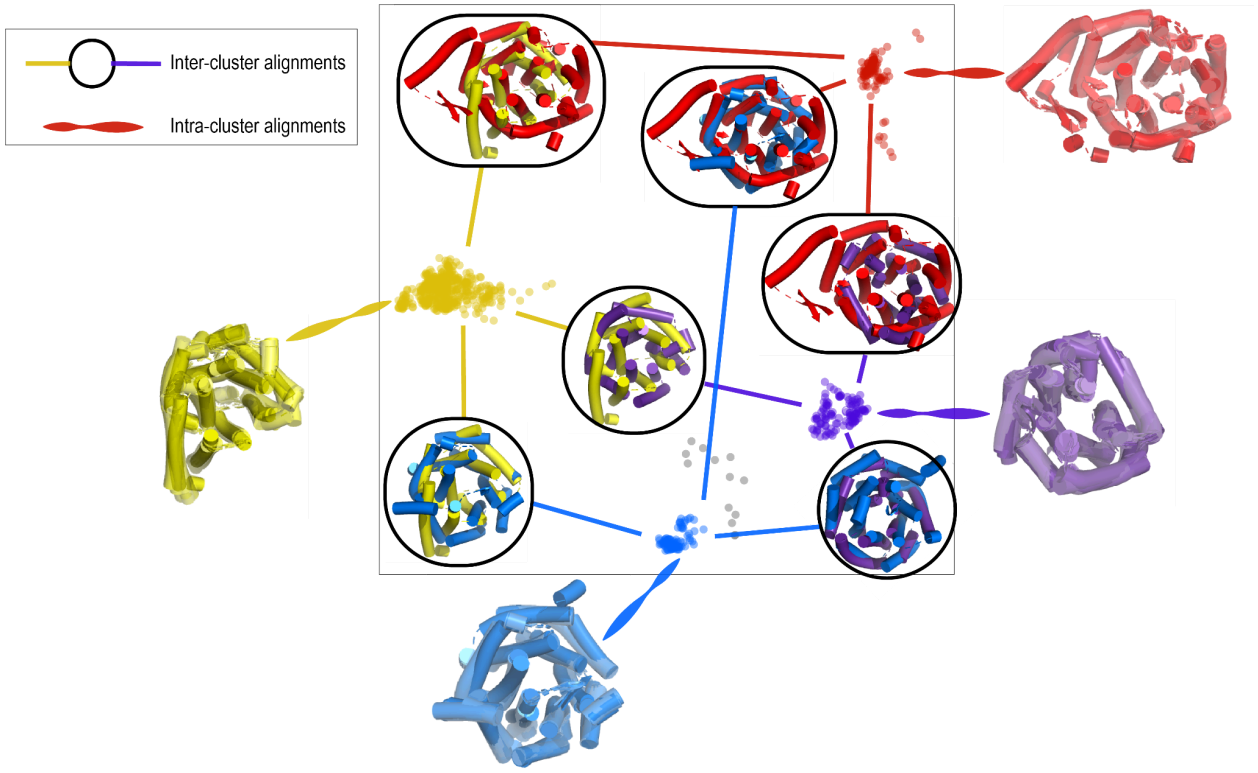


Extended Data Fig. 10 | Screening sequence databases. a, In-silico screening workflow for selecting evolutionary distant uncharacterized TPS with TPS-specific Pfam/SUPFAM domains. **b,** In-silico screening workflow for selecting TPS without any protein signature: from $6.2x10^7$ sequences in UniRef50, $2.5x10^4$ have been detected as terpene synthases. For our model validation, all proteins with an InterPro signature have been excluded, reducing to 7202 putative proteins without annotation. We selected 17 enzymes from non-plant organisms for protein expression in yeast, optimized for sesqui- and diterpene production. Ethyl acetate extracts of 2 days fermentation were therefore analyzed using GC-MS.

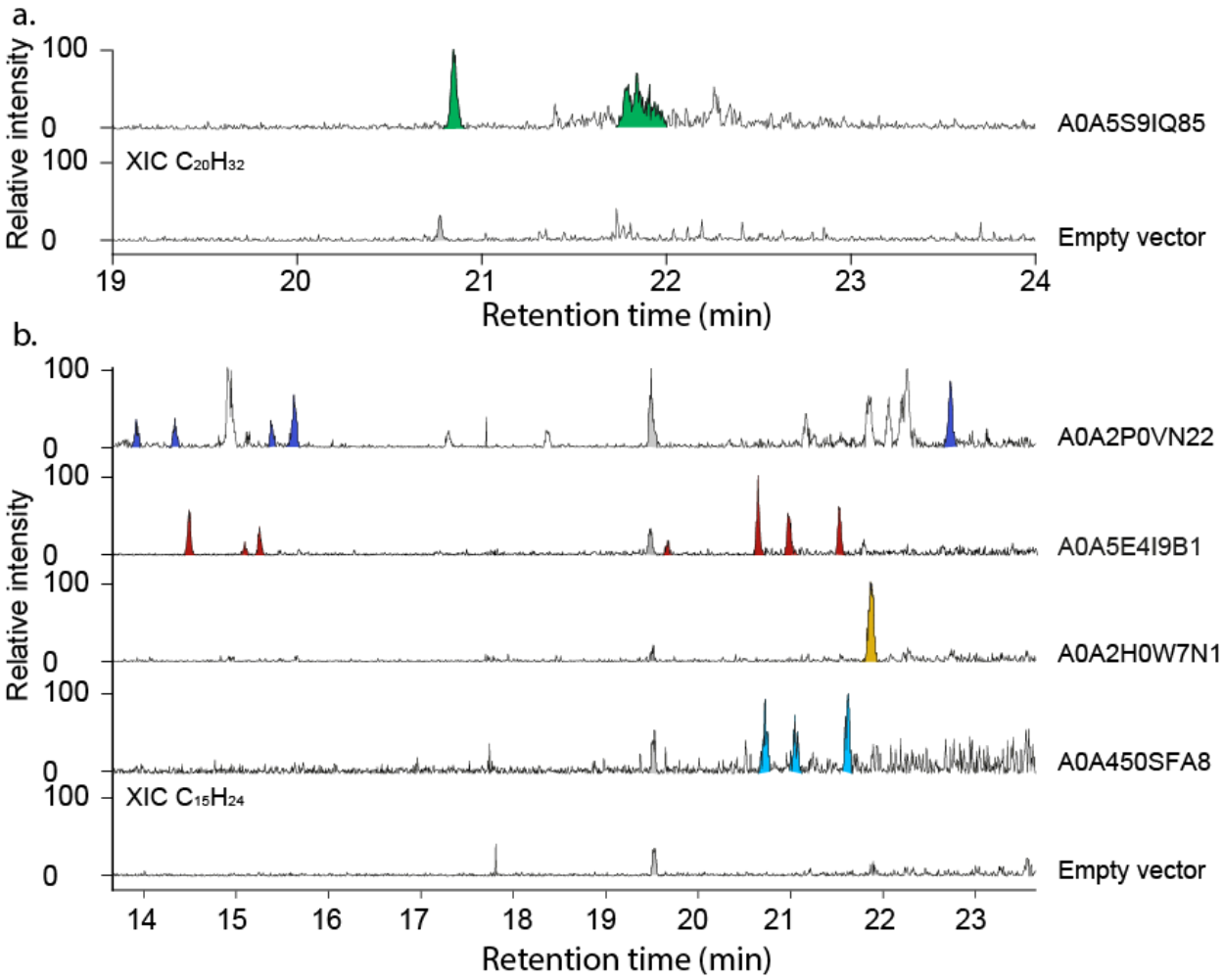
Supplementary



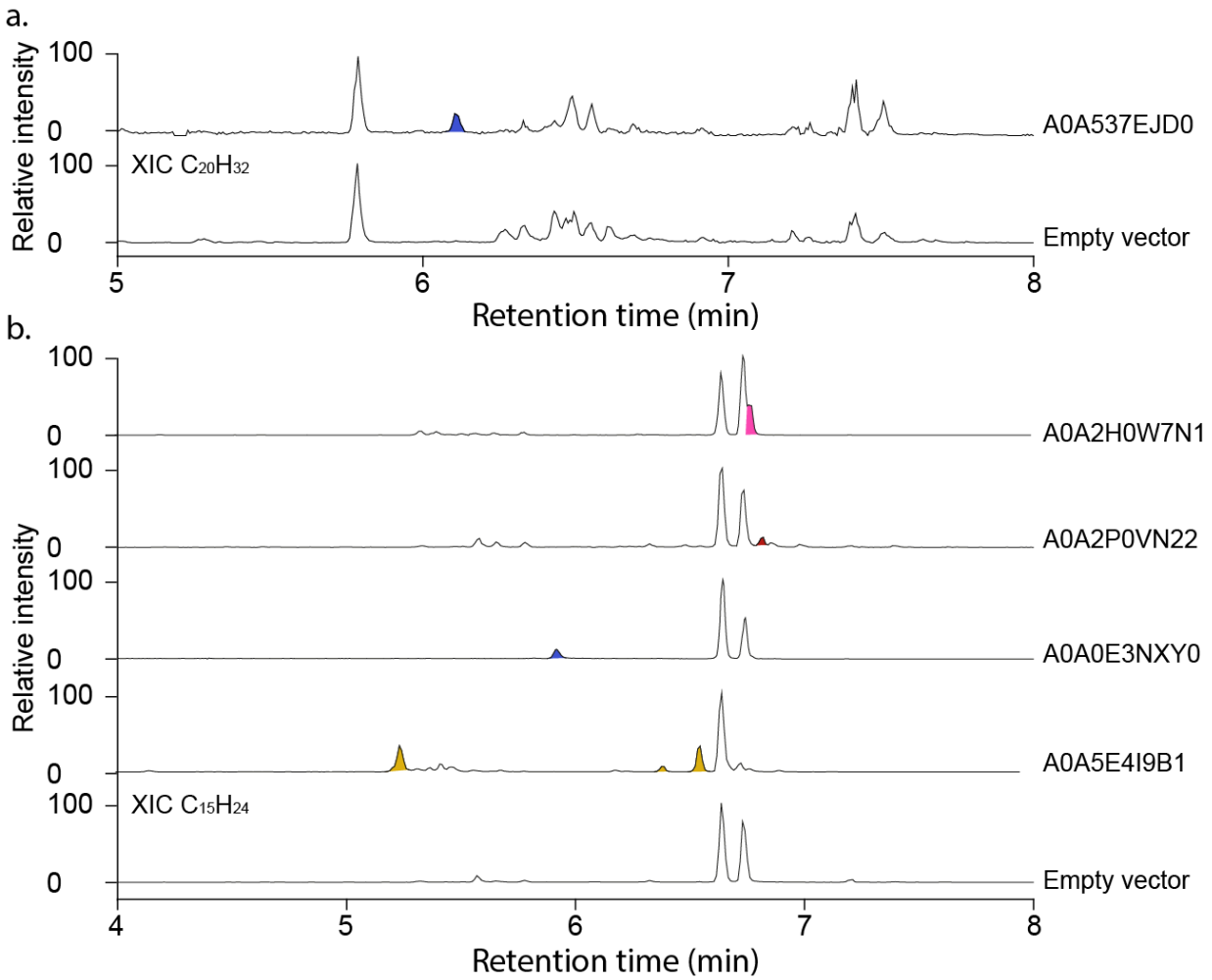
FigSI_Whole_protein_clustering. Assessment of the quality of whole-protein TPS clustering. We compared how well automatically derived clusters correspond to ground truth categories defined by TPS type or substrate. Adjusted mutual information is a standard metric for assessing clustering quality when ground truth labels are available[86].



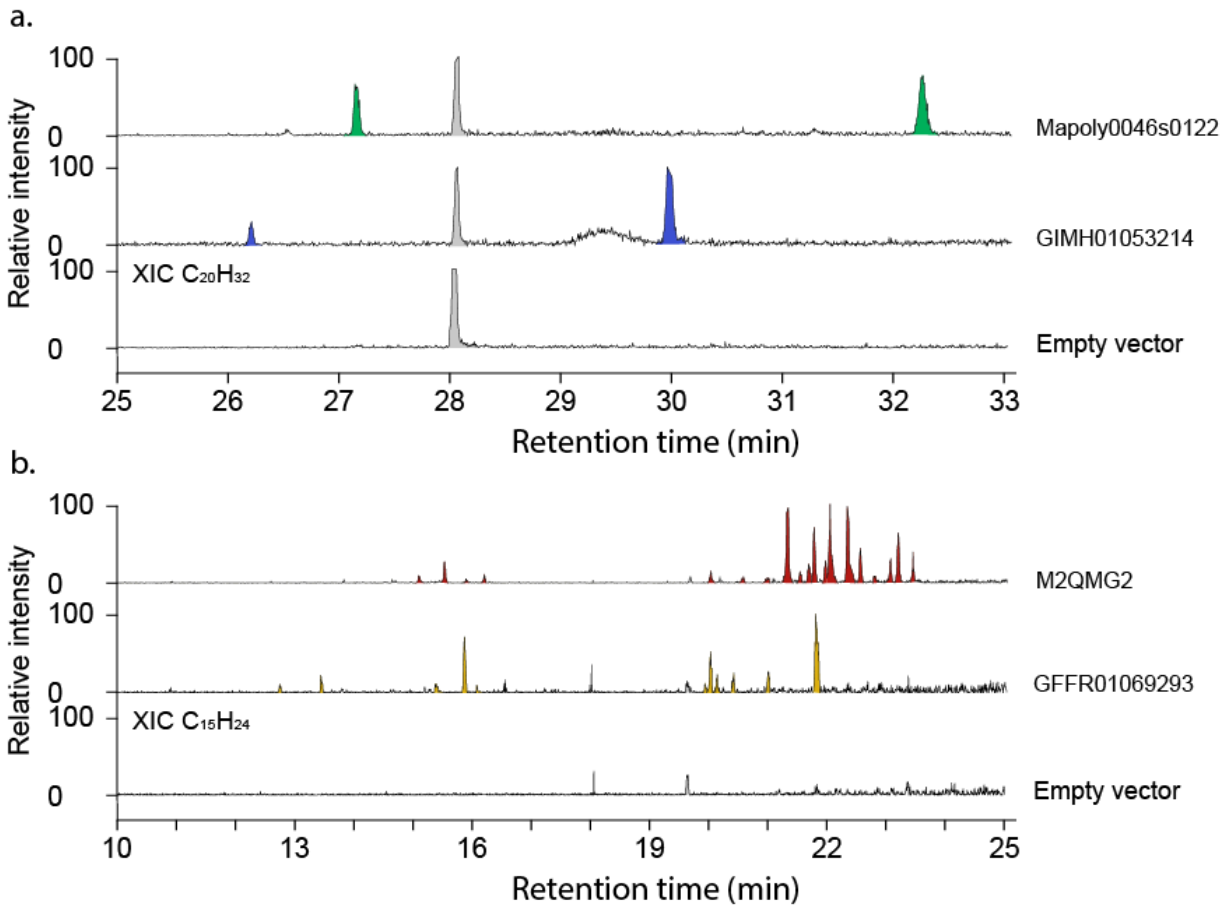
FigSI_Domain_subtypes_comparison. All-to-all pairwise comparisons for β/γ domain subtypes.



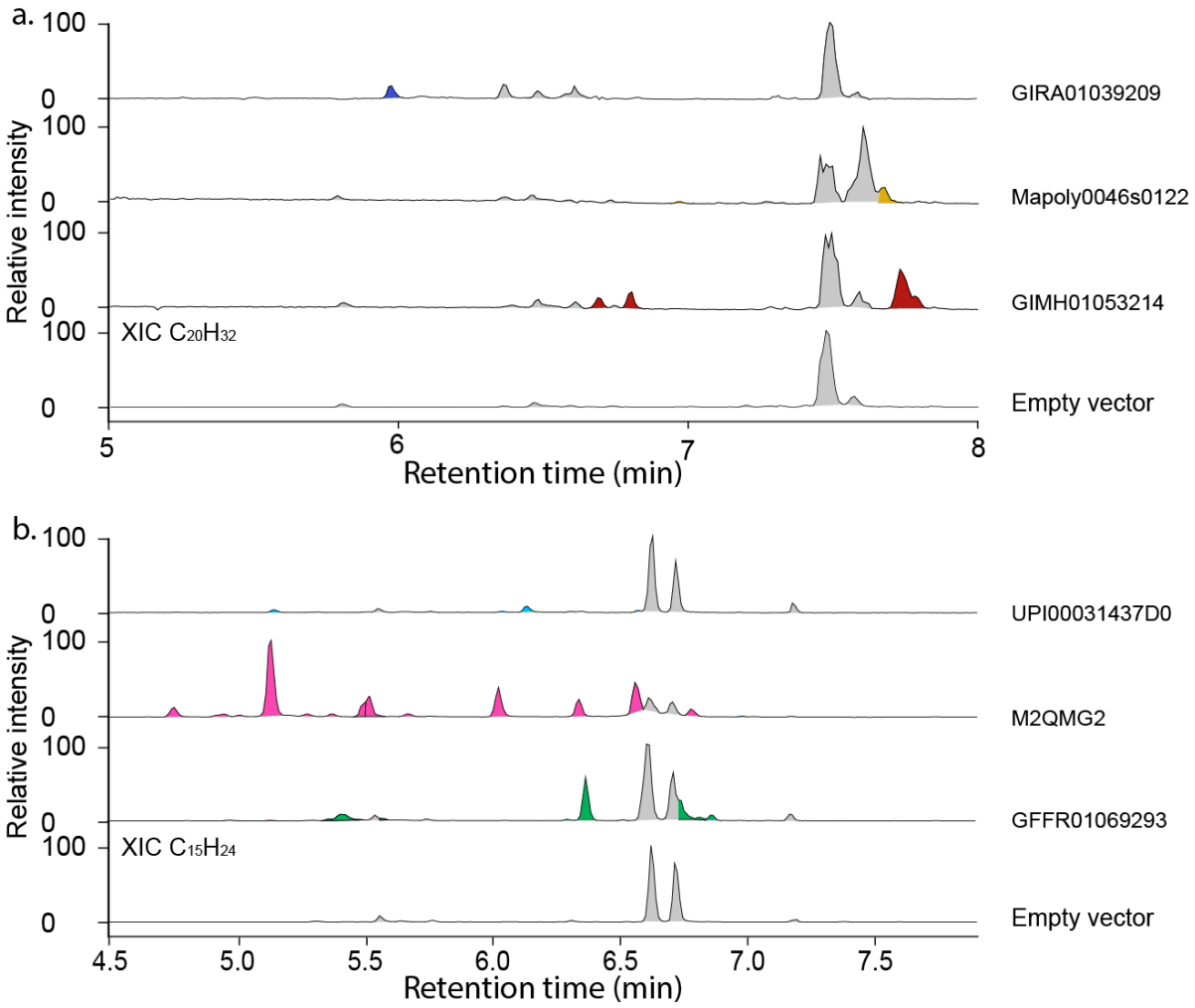
FigSI_WetLab1. Analysis of Yeast extract by GC-EI-MS. a. Extracted ion chromatogram for diterpene signal, *m/z* 272.2. b. Extracted ion chromatogram for sesquiterpene signal at *m/z* 204.2.



FigSI_WetLab2. Analysis of Yeast extract by LC-ESI-MS. a. Extracted ion chromatogram for diterpene signal at *m/z* 273.2577 (+/- 5 ppm). b. Extracted ion chromatogram for sesquiterpene signal at *m/z* 205.1951 (+/- 5 ppm).



FigSI_WetLab3. Analysis of "easy candidates" Yeast extract by GC-EI-MS. a. Extracted ion chromatogram for diterpene signal, *m/z* 272.2. b. Extracted ion chromatogram for sesquiterpene signal at *m/z* 204.2.



FigSI_WetLab4. Analysis of “easy candidates” Yeast extract by LC-ESI-MS. a. Extracted ion chromatogram for diterpene signal at *m/z* 273.2577 (+/- 5 ppm). b. Extracted ion chromatogram for sesquiterpene signal at *m/z* 205.1951 (+/- 5 ppm).

SI Wet Lab

Table S1: List of plasmids

Plasmid name	Plasmid type	Backbone&Selection marker	Description	Reference
pYTK001	Backbone	pYTK001, Cm	Storage plasmid	Lee et al.[82]

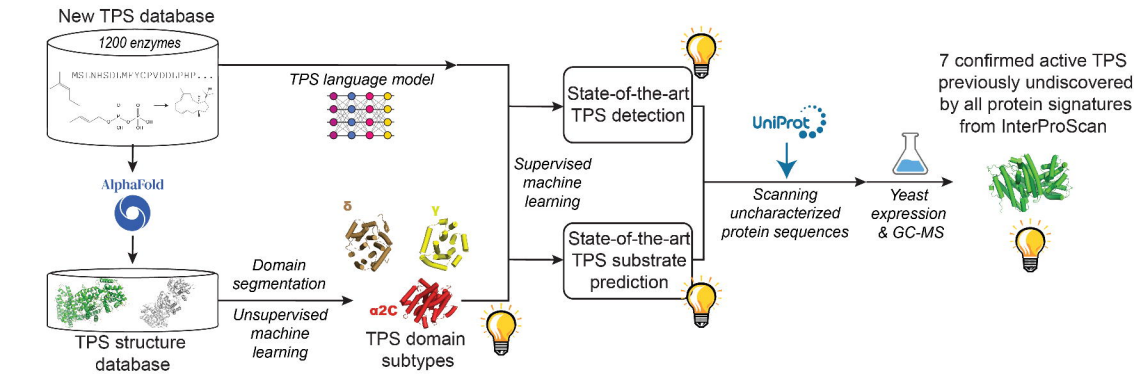
pYTK002	Part type 1	pYTK001, Cm	ConLS	Lee et al.[82]
pYTK009	Part type 2	pYTK001, Cm	P*tdh3*	Lee et al.[82]
pYTK047	Part type 234	pYTK001, Cm	GFP dropout	Lee et al.[82]
pYTK052	Part type 4	pYTK001, Cm	T*ssa1*	Lee et al.[82]
pYTK067	Part type 5	pYTK001, Cm	ConR1	Lee et al.[82]
pYTK074	Part type 6	pYTK001, Cm	URA3	Lee et al.[82]
pYTK082	Part type 7	pYTK001, m	2µ origin of replication	Lee et al.[82]
pYTK083	Part type 8	pYTK083, Amp	Backbone, bacterial module	Lee et al.[82]
pYTK096	Backbone	pYTK096, Amp	Ura3 targeting integration vector	Lee et al.[82]
pESC-Ura	Backbone	Backbone, Amp	Ura3 resistance marker	Agilent
pTP0027	Backbone	pYTK083, Amp	Transient expression, Ura3 prototrophy	This study
pTP0100	Part type 3	pYTK001, Cm	GIMH01053214	This study
pTP0101	Part type 3	pYTK001, Cm	GIRA01039209	This study
pTP0103	Part type 3	pYTK001, Cm	Mapoly0046s0122	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
pTP0197	Transient expression	pTP0027, Amp	P*thd3,* GIMH01053214, T*ssa1*	This study
pTP0198	Transient expression	pTP0027, Amp	P*thd3,* GIRA01039209, T*ssa1*	This study
pTP0200	Transient expression	pTP0027, Amp	P*thd3,* Mapoly0046s0122, 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
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
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
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

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
pTP0256	Integration vector	pYTK096, amp	P*tdh3*, A0A537EJD0, T*ssa1*	This study

Table S2: List of strains

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*.[87]
JWY501_ESC-Ura	JWY501 with pESC-Ura	This study
JWY501_pTP0197	JWY501 with pTP0197	This study
JWY501_pTP0198	JWY501 with pTP0198	This study
JWY501_pTP0200	JWY501 with pTP0200	This study
JWY501_pTP0201	JWY501 with pTP0201	This study
JWY501_pTP0202	JWY501 with pTP0202	This study
JWY501_pTP0203	JWY501 with pTP0203	This study
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_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_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


New TPS database
1200 enzymes
TPS language model
AlphaFold
TPS structure database
Domain segmentation
Unsupervised machine learning
δ
γ
α2C
TPS domain subtypes
Supervised machine learning
State-of-the-art TPS detection
State-of-the-art TPS substrate prediction
UniProt
Scanning uncharacterized protein sequences
Yeast expression & GC-MS
7 confirmed active TPS previously undiscovered by all protein signatures from InterProScan