Chapter 3 Ancestral Sequence Reconstruction to Enable Biocatalytic Synthesis

3.1 Summary

Biocatalysis can be powerful in organic synthesis but is often limited by enzymes' intrinsic substrate scope and selectivity. Developing a biocatalytic step typically begins with identifying an initial enzyme for the targeted reaction, followed by optimization through protein engineering using rational design, directed evolution, or both. These steps are time-consuming, resource-intensive, and require expertise that can extend beyond what is typical for organic chemists. Thus, an effective strategy to streamline the process from enzyme identification to implementation in a synthetic route is essential to expanding the use of biocatalytic transformations. Here, we present a strategy that incorporates bioinformatics-guided enzyme mining and ancestral sequence reconstruction (ASR) to resurrect enzymes for biocatalytic synthesis. Specifically, we achieve a biocatalytic, enantioselective synthesis of azaphilone natural products using two ancestral enzymes: a flavin-dependent monooxygenase (FDMO) for stereodivergent oxidative dearomatization and a substrate-selective acyltransferase (AT) for the subsequent acylation of the enzymatically installed hydroxyl group. This novel cascade, which is stereocomplementary to established chemoenzymatic routes, expands the accessible chemical space of linear tricyclic azaphilones by providing access to enantiomeric compounds. The initial challenge of discovering an AT candidate was overcome by leveraging the co-occurrence and co-evolution of FDMO and AT in azaphilone biosynthetic pathways. The subsequent bottleneck of engineering the low solubility and

---

Adapted from Chiang, C.-H.; Wang, Y.; Hussain, A.; Brooks III, C. L.; Narayan, A. R. H. Ancestral sequence reconstruction to enable biocatalytic synthesis. *J. Am. Chem. Soc.* **2024,** *146* (44), 30194–30203. Copyright 2024 American Chemical Society.

In this work, Ye Wang synthesized substrates for screening and performed large-scale biocatalytic reactions and Azam Hussain was responsible for molecular modeling, including docking and molecular dynamics simulation.

stability of the AT, CazE, was addressed using ASR to obtain a more soluble, stable, promiscuous, and reactive ancestral AT (AncAT) for our enzymatic scheme. Sequence analysis of AncAT reveals it as a chimeric composition of its descendants, with enhanced reactivity likely due to the promiscuity inherent in ancestral enzymes. The structural basis for AncAT's reactivity was established through flexible receptor docking and molecular dynamics simulations, showing that the most reactive AncAT best promotes a reactive geometry between substrates. We anticipate that our bioinformatics-guided and ASR-based approach can be broadly applied in target-oriented synthesis, providing an avenue to circumvent potentially lengthy development of biocatalytic steps and efficiently access superior biocatalysts.

3.2 Introduction

Biocatalysis is a powerful tool, as enzymes can provide catalyst-controlled selectivity in the synthesis of complex molecules such as the HIV drug islatravir and insulin derivitives.[1-2] However, the implementation of biocatalysis is often limited by the native substrate scope or selectivity of enzymes, necessitating additional effort in protein engineering to achieve different substrate scope, selectivity, or reactivity.[3] Directed evolution has been a cornerstone in optimizing enzymes for various purposes.[4] However, directed evolution is generally time- and resource-intensive, which increases the barrier to its use. Additionally, this approach can face many practical challenges at different stages, including identifying a soluble and stable starting enzyme with the desired activity which can be optimized and navigating decreasing protein stability over the course of an evolution campaign.[5-8] Finding a suitable starting point is essential to a protein engineering approach, and the properties of the initial protein directly affect the evolvability of an enzyme and its compatibility with a high throughput screening workflow, which eventually determines the success of an evolution campaign.[6] As an alternative to directed evolution, de novo enzyme design has shown some success in modifying or extending enzymatic activities, however, it commonly requires further optimization after the original design.[3] Other approaches, such as consensus sequence design[9] and ProteinMPNN,[10] have also been applied to improve enzyme expression, stability, and function. Although these two approaches can achieve excellent outcomes, they currently

are limited to optimizing inherent enzymatic activities and have not been used for generating enzymes with non-native substrate scope or selectivity, which is often required for target-oriented applications. Therefore, more efficient and effective biocatalyst development workflows are necessary to find or create enzymes for target-oriented synthesis (Figure 3.1).

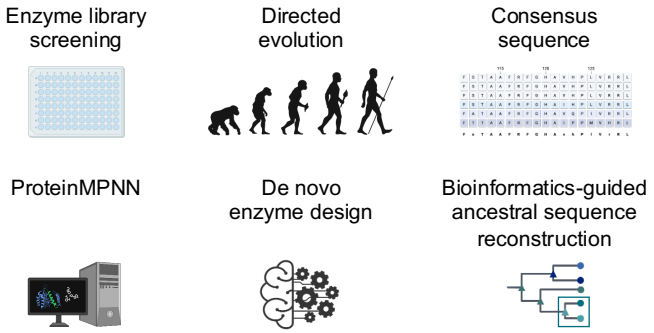


Figure 3.1. Common strategies for biocatalyst development.

With experience in the challenges of applying biocatalysis in target-oriented synthesis, we envisioned using ancestral sequence reconstruction (ASR) and resurrection in combination with bioinformatics to identify and optimize enzymes for synthesis. ASR is commonly employed in studying protein sequence-function relationships.[11-17] As ancestral enzymes often possess attractive features over their extant relatives, such as enhanced solubility, thermostability, and promiscuity,[18-21] ASR shows great potential for the application of biocatalysis. For example, researchers have used ASR to engineer P450 enzymes and ene-reductases with improved stability and solubility,[22-24] enhance the activity of ketol-acid reductoisomerases,[22] and obtain enzymes with greater promiscuity.[15,18,23] As ASR is based on phylogeny, it can be seamlessly combined with other bioinformatic tools such as homology searches,[25] sequence similarity networks,[26-27] and BGC analyses,[26-27] to guide the identification of an initial enzyme candidate. We anticipated that a bioinformatics-guided ASR approach could provide a retrospective path to find enzymes with desired traits and substrate selectivity if an appropriate sequence space was targeted and sampled.

To test our bioinformatics-guided ASR strategy, we chose azaphilones, a class of fungal polyketide natural products, as our synthetic target.[28-29] Azaphilones, which feature a conserved oxygenated pyranoquinone bicyclic core and a tetrasubstituted stereocenter,

are particularly valuable due to their wide range of biological activities.[28-29] In nature, the C7-stereocenter of azaphilones is installed through a flavin-dependent monooxygenase (FDMO) catalyzed oxidative dearomatization.[30] The azaphilone core can be further elaborated by introducing the third ring through acylation by an acyltransferase (AT) and subsequent spontaneous Knoevenagel condensation.[31-32] Inspired by nature, we recently reported chemoenzymatic syntheses of linear tricyclic azaphilones,[32] including rubropunctatin (**3.1**) and monascorubrin (**3.2**),[33] using the enzymatic sequence involving AzaH[34] and MrPigD[31] (Figure 3.2). Specifically, the chemoenzymatic synthesis requires FDMOs which can catalyze stereodivergent oxidative dearomatization reactions to transform **3.3** into (*S*)-**3.4** or (*R*)-**3.4** selectively. This stereodivergent step can be followed by an acylation event with **3.5**, which can be mediated by an AT to form (*S*)-**3.6** or (*R*)-**3.6**, respectively (Figure 3.2). By combining modifications at R^1 and R^2, and also diversification through amination, a number of linear tricyclic azaphilone derivatives can be accessed (Figure 3.2). However, realization of this stereodivergent strategy was stymied by the lack of an AT capable of acylating (*S*)-**3.4** to form (*S*)-**3.6** as no reported biosynthetic gene clusters (BGCs) responsible for synthesis of linear tricyclic azaphilones bearing the (*S*)-configuration at C7. In addition, AfoD[35] is the only extant FDMO reported to act on **3.3**, and it does so with low conversion.[36] Therefore, synthetic access to (*S*)-tricyclic linear azaphilone structures was not possible through this chemoenzymatic approach.

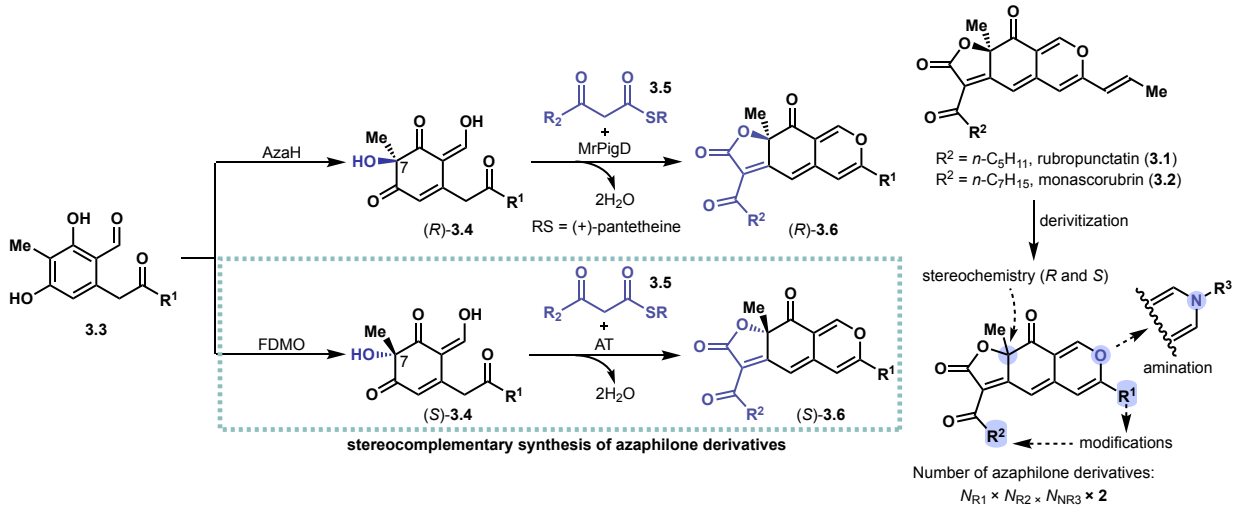


Figure 3.2. Stereocomplementary biocatalytic synthesis of linear tricyclic azaphilones doubles the number of accessible azaphilone derivatives.

To solve this problem and achieve a stereocomplementary biocatalytic synthesis, we used an ancestral flavin-dependent monooxygenase (AncFDMO) reconstructed and engineered for our previous enzyme evolutionary study[37] with a newly reconstructed ancestral acyltransferase (AncAT). The reconstruction and sampling of AncATs was guided by knowledge of FDMO stereoselectivity[16] and co-localization of FDMOs and ATs in biosynthetic pathways.[38] This strategy allowed us to focus on a smaller sequence space most likely to have the substrate selectivity of interest. By bringing together bioinformatic tools and computational molecular evolution, ASR is positioned to guide the generation of biocatalysts tailored for specific synthetic challenges. We anticipated that this type of ASR-based strategy could be more broadly applied in the biocatalytic synthesis of natural products and derivatives.

3.3 Results and Discussion

Guiding acyltransferase discovery through bioinformatics. To identify a candidate AT to achieve a stereocomplementary synthetic strategy, we leveraged knowledge of FDMO enantioselectivity to increase the likelihood of discovering a candidate AT. Since FDMO and AT genes are generally co-localized in the same BGC,[30] our strategy was to find candidate ATs that are part of the same gene cluster as FDMOs that stereoselectively afford *(S)*-**3.4** (Figure 3.2). This assumes that co-localization leads to co-evolution on the related substrate scaffolds, increasing the likelihood of identifying a candidate gene to further explore. We began our AT mining by testing five FDMO co-localizing ATs in our library but did not find any hits (Table S3.1). Therefore, we redirected our attention to reported azaphilone-related BGCs that possess a FDMO-AT pair. The search returned three FDMO-AT pairs for rubropunctatin/monascorubrin (**3.1**/**3.2**),[33] azanigerone A (**3.7**),[34] and chaetoviridin A (**3.8**, Figure 3.3).[39] The AT in the BGC of azanigerone A, AzaD,[34] was excluded as the pyranoquinone core of the product **3.7** has the opposite stereochemical configuration to our targeted substrate (*S*)-**3.4** (R^1 = *n*-Pr). Additionally, the acyl group incorporated by AzaD does not possess the 3-ketoacyl motif required to provide access to the butanolide ring. MrPigD,[31] the AT in the BGCs of rubropunctatin and monascorubrin was tested for activity with the substrate **3.4** (R^1 = *n*-Pr). As expected, MrPigD only

reacted with (*R*)-**3.4** (R^1 = *n*-Pr) and did not show any activity toward (*S*)-**3.4** (R^1 = *n*-Pr), implying that absolute configuration is important for the substrate binding in the active site of these ATs. Chaetoviridin A has the same absolute configuration at C7 as rubropunctatin and monascorubrin,[39] so we omitted its corresponding AT, CazE, in initial testing.

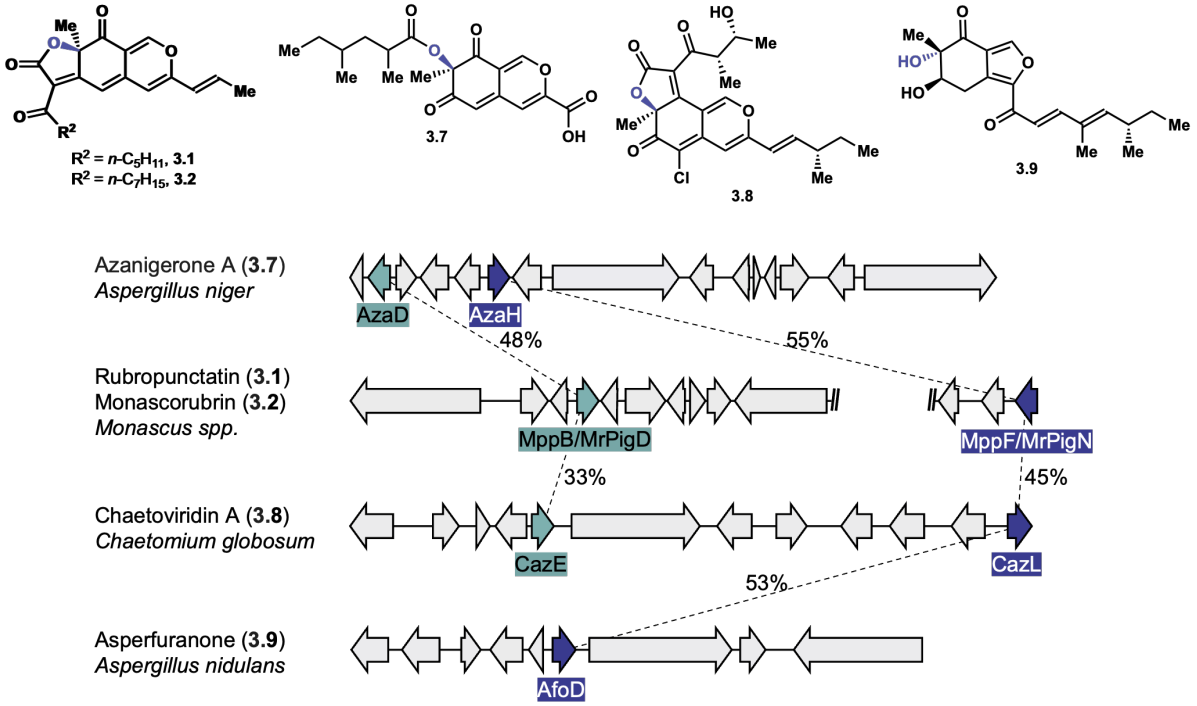


Figure 3.3. Azaphilone-related biosynthetic gene clusters and their natural products.

After not finding any active AT candidates from previously reported BGCs, we turned our focus to the BGC of asperfuranone (**3.9**, Figure 3.3). Although this BGC does not contain an AT-encoding gene, it possesses a FDMO, AfoD,[35] that can afford the (*S*)-dearomatized product akin to (*S*)-**3.4** (Figure 3.3).[36] Considering the frequent co-localization and sequential relationship of FDMO and AT in the azaphilone biosynthetic pathway (Figure 3.4A), we reasoned that identifying the AT's paired with FDMOs which are homologous to AfoD could offer an additional avenue of exploration. The FDMO homologs in PF01494 were searched by performing BLAST using AfoD as the query on EFI-EST (Figure S3),[26-27] and these genes were then filtered to retain those having a co-occurring AT located within a range of 10 genes either upstream or downstream by EFI-GNT.[26-27] Next, we generated separate sequence similarity networks (SSNs) for those FDMOs in PF01494 and ATs in PF02458. Clustering by SSN is commonly used to search for homologous sequences based on the assumption that the sequences in the same

cluster are likely to be “isofunctional.” Built upon clustering data and genome neighborhood analysis of FDMO and AT genes, the agreement between the two clusterings was assessed using the adjusted mutual information (AMI) score,[40] with scores near 1 indicating a high degree of agreement (shared information). We observed an AMI of 0.85, validating our co-localization and co-evolution hypothesis. The high degree of mutual information shared can be visualized in Figure 3.4B by mapping the SSN clusters between the co-localized FDMOs and ATs, which shows that most FDMOs from the same cluster in the SSN of PF01494 have their adjacent AT partners in the same cluster in the SSN of PF02458. In brief, this proof-of-concept lays the foundation for the subsequent AT mining workflow shown in Figure 3.4C that leverages knowledge about the enantioselectivity of FDMOs,[16] given the co-localization of FDMO and AT genes.

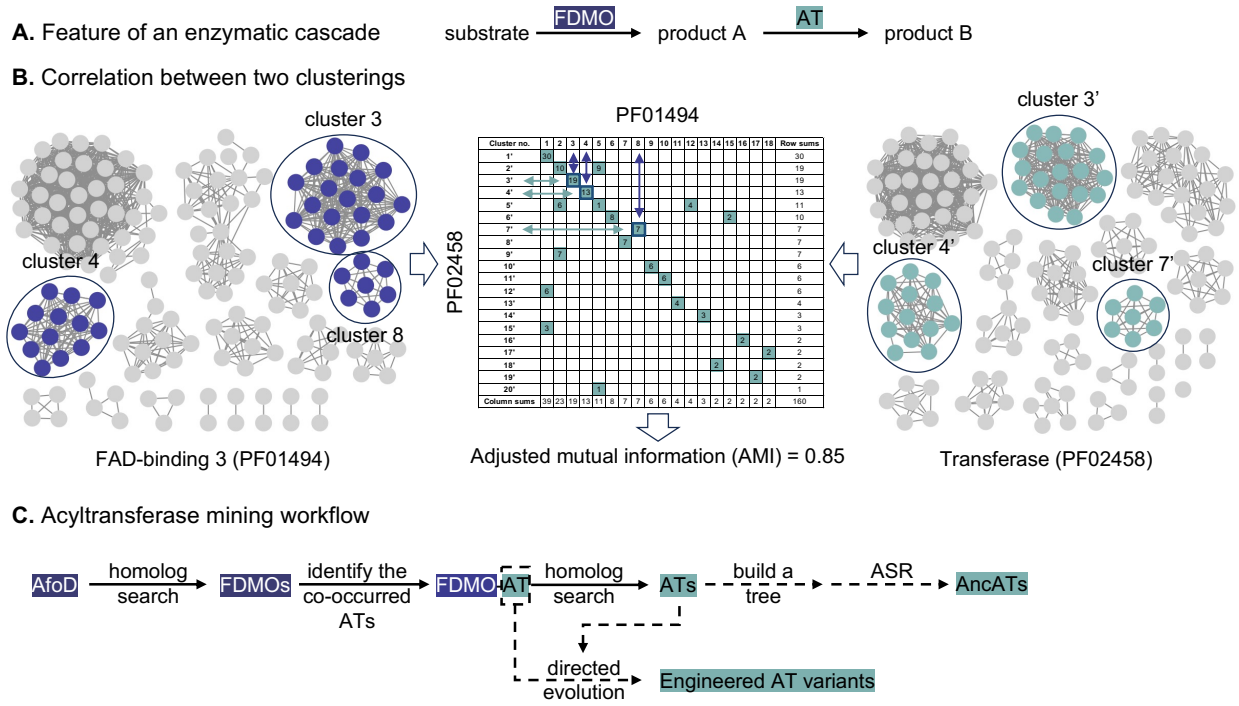


Figure 3.4. Acyltransferase mining based on the co-localization and the upstream-downstream relationship of flavin-dependent monooxygenase (FDMO) and acyltransferase (AT) in azaphilone biosynthetic gene clusters (BGCs). **A.** Biosynthetic feature that upstream products are substrates for downstream enzymes. **B.** Proof-of-concept that the functional information of acyltransferases (PF02458) can be inferred by the functional information of flavin-dependent monooxygenases (PF01494). **C.** Acyltransferase mining workflow using AfoD as our query.

After demonstrating that the substrate selectivity of ATs can be inferred from the stereoselectivity of FDMOs in the same BGC, we followed the workflow depicted in Figure

3.4C to significantly narrow the AT sequence space to target (~39,000 PF02458 sequences in the UniProtKB database as of November 2023). The PSI-BLAST[41] search was performed on the NCBI database using AfoD as our query to potentially cover a more diverse sequence set of FDMOs, and these sequences were then filtered to retain those with adjacent ATs in PF02458. In the top twenty hits of the BLAST search, four FDMOs with 55–60% sequence identity to AfoD were found to coexist with an AT (Table S3.3). Surprisingly, CazL,[39] which is the associated FDMO of CazE (AT), was found to be one of the four hits (Table S3.3). We proceeded to investigate the reactivity of CazE. Interestingly, although CazE was previously reported to react with **3.11** (Figure 3.5A), about 10% substrate conversion was seen when (*S*)-**3.12** was incubated with purified CazE, encouraging us to further explore CazE and its close homologs.

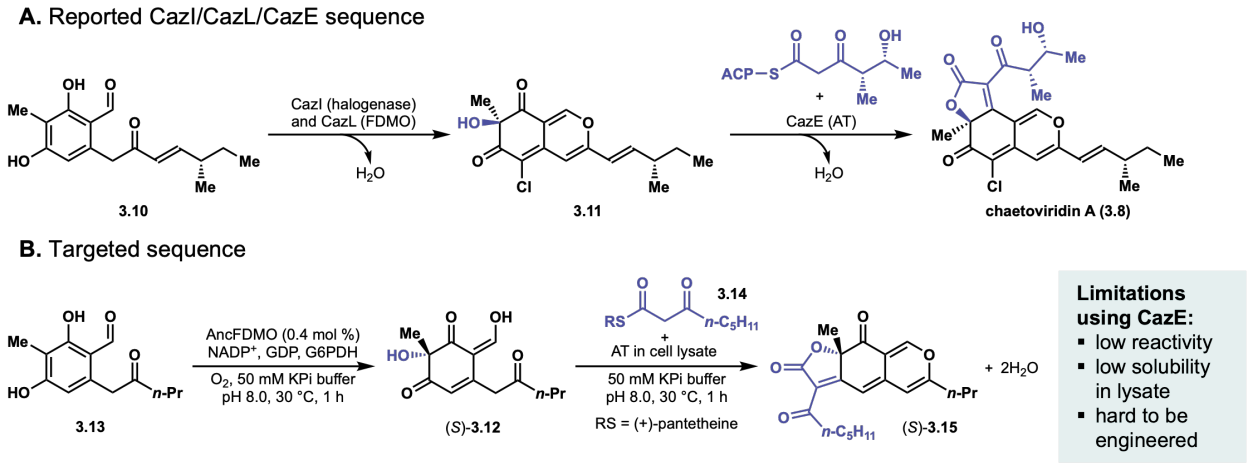


Figure 3.5. Reported CazL/CazE (FDMO/AT) sequence in the biosynthetic gene cluster of chaetoviridin A and the targeted two-enzyme sequence to access a linear tricyclic azaphilone.

Reconstruction and resurrection of ancestral acyltransferases. The acylation reaction catalyzed by CazE possessed too low a yield to make it practical for implementation in a synthetic route. To increase enzyme activity, we considered using directed evolution,[42-43] which relies on iterative rounds of high-throughput screening of CazE variants. A critical prerequisite of fitting an enzyme into a directed evolution campaign is whether the enzyme is sufficiently soluble.[44] However, we were not able to detect any product when performing reactions using the clarified cell lysate of CazE (Figure 3.5B). After careful consideration of the applicability of directed evolution, we recognized that CazE would not be compatible with the high-throughput screening due to

its low level of soluble expression. To overcome this hurdle, we leveraged ASR as an approach to acquire more stable, soluble, and promiscuous enzymes, and to avoid a resource intensive evolution campaign. Although the reactivity of CazE was low, the preliminary result suggested that we may have already found sequence space which captured our desired selectivity. From the ASR, we anticipated that we could access more reactive ancestral enzymes by deliberately introducing promiscuity or at least generate enzymes with higher stability and soluble expression.[45-47] The ASR was started by reconstructing an unrooted maximum likelihood phylogenetic tree built from 240 homologous AT sequences acquired through BLAST search using CazE as the query (Figure 3.6). From this tree, we resurrected seven ancestral ATs (AncAT1–7) that are from the ancestral nodes either on the evolutionary trajectory to CazE or in the close sister clades (Figure 3.7A), and sampled a closely homologous extant AT from *Talaromyces proteolyticus*, TpAT. The ancestral sequences were then aligned with CazE and TpAT, demonstrating that all ATs share over 58% identity with each other (Figure 3.7B). The expression level and reactivity of all ATs were qualitatively characterized in the clarified cell lysate format. All ancestors showed superior expression levels and solubility compared to CazE and TpAT based on SDS-PAGE (Figure 3.8A). Additionally, we observed that the reaction mixture with AncAT4 turned orange compared to the negative control after 1 h in the screening against (*S*)-**3.12** (dearomatized product of **3.13**) and the C5-thioester **3.14**, indicating that some (*S*)-**3.15** was likely formed (Figure 3.8B). Based on these preliminary results, the representative ATs, including AncAT4–6, CazE, and TpAT were further purified, and their melting temperatures were measured using differential scanning fluorometry. This revealed that, although AT4–6 were all recent ancestors, they were still generally more stable than the extant enzymes including CazE and TpAT (43.5 vs. 40.4 °C, Figure 3.8C and Table S3.2), aligning with many previous reports that ancestral proteins tend to be more thermally stable.[20] AlphaFold2 homology models of AncAT4–6 and CazE were generated, and all of them showed high structural alignment with each other (Figure 3.8D). In summary, all ancestral ATs were successfully resurrected.

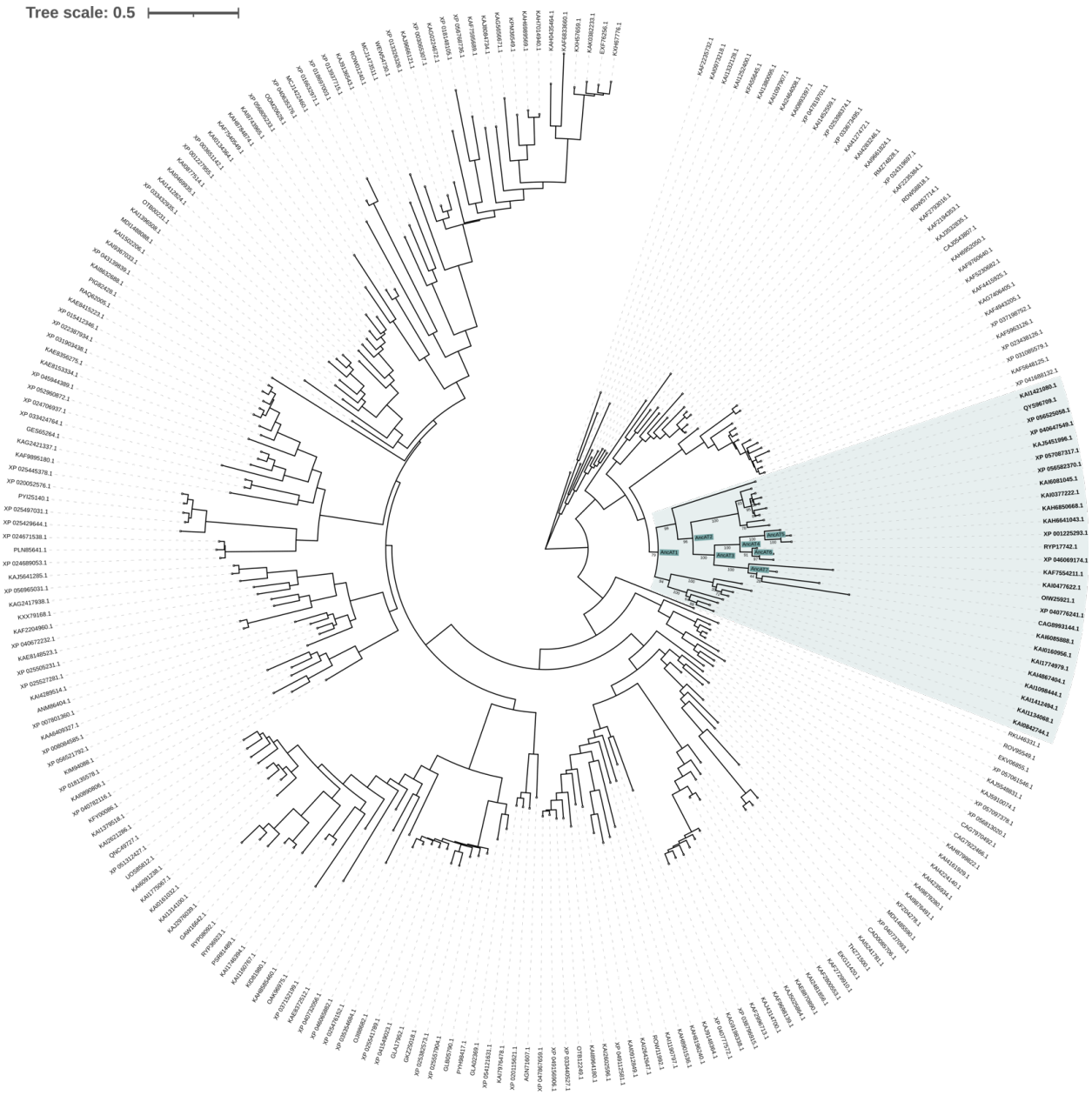


Figure 3.6. Unrooted maximum-likelihood phylogenetic tree of selected acyltransferases in PF02458.

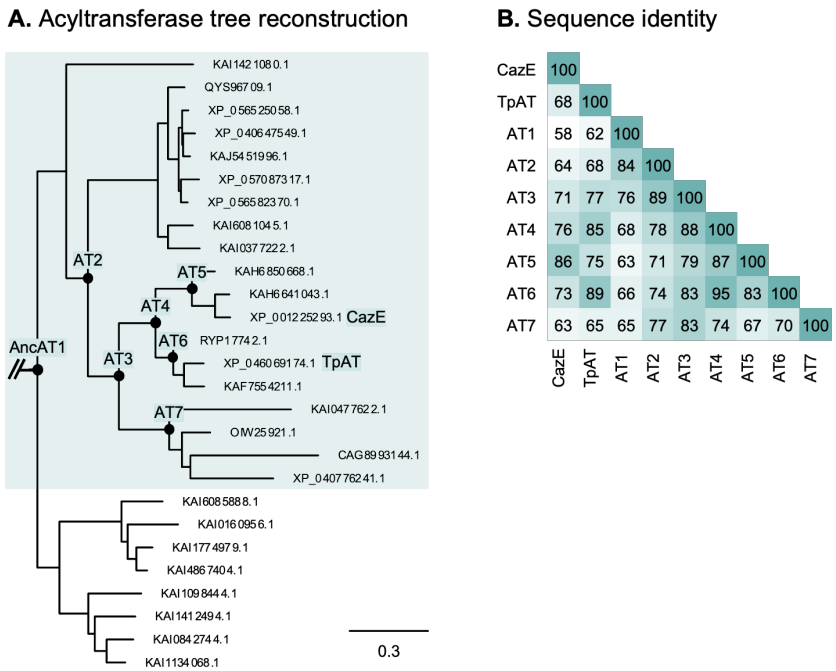


Figure 3.7. Evolutionary relationships of ancestral and extant acyltransferases in the CazE clade.
A. Zoom-in view of the CazE clade including seven ancestral acyltransferases selected for resurrection. **B.** Sequence identity of the ancestral and extant acyltransferases.

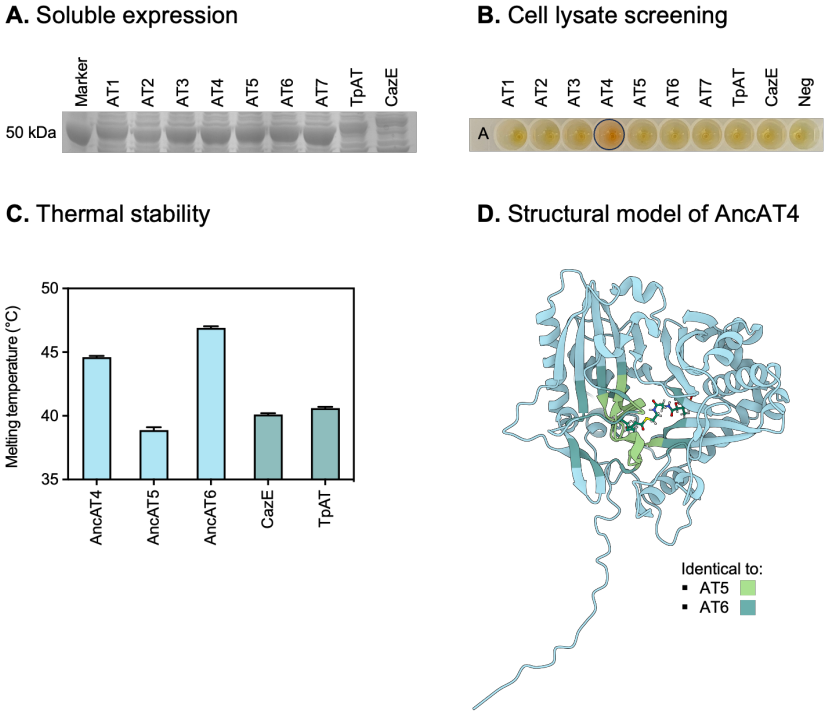


Figure 3.8. Comparison of the extant and ancestral acyltransferases.
A. Soluble expression. **B.** Cell lysate screening. **C.** Thermal stability. **D.** Structural model of AncAT4.

Exploring expanded substrate scope in ancestral acyltransferases. With select purified ATs in hand, we evaluated the activity of each enzyme against substrate (*S*)-**3.12**.

In the two-enzyme sequence, (*S*)-**3.4** was first generated from **3.3** through the oxidative dearomatization reaction catalyzed by the purified AncFDMO (Figure 3.9A). After one hour, thioester acyl donor and acyltransferase were added and incubated for an additional hour. Based on the consumption of (*S*)-**3.12**, the activity of the acyltransferases with the C5-thioester as an acyl donor followed the trend of AT4 > AT5 > CazE > AT6 ~ TpAT, with AT4 producing the highest conversion (40%, Figure 3.9B). In addition to the C5-thioester, thioesters with varying chain lengths (C1, 3, 6, 7, and 9) were tested against (*S*)-**3.12** using AncAT4. Alternative chain lengths were tolerated, but the activity of AncAT4 peaked with the C5-thioester (Figure 3.9C). Notably, we observed much lower enzymatic activity in AT6 and TpAT, which was surprising considering the similarity shared between AT4 and AT6 or TpAT. Additional azaphilone substrates were tested against the panel of purified enzymes with the C5-thioester coupled including (*S*)-**3.16**–**3.18**, which possess modifications of the R^1-substituent. Again, AncAT4 showed the greatest activity among all the enzymes (Figure 3.9D). We were also curious as to whether AncAT4 is more selective for (*S*)-**3.12** over (*R*)-**3.12**. To answer this question, the oxidative dearomatization was catalyzed by AzaH, and the (*R*)-product was incubated with each purified AT to be coupled with the C5-thioester. Interestingly, we found that AT4 was more reactive with (*S*)-**3.12** than (*R*)-**3.12** (40% vs. 9% conversion, Figure 3.9B), implying that the reconstructed enzyme has higher selectivity for the (*S*)-substrate. These observations support the conclusion that our workflow successfully guided the sampling of ancestral ATs to an appropriate sequence space with desired substrate selectivity (enantiodiscrimination). Finally, the enzymatic cascade for the synthesis of (*S*)-**3.15** was performed on preparative scale to showcase its potential synthetic implementation. (*S*)-**3.15** was successfully purified and characterized with a ^{1}H-NMR spectrum matching the previously reported (*R*)-**3.15**, but opposite optical rotation and circular dichroism signals compared to (*R*)-**3.15** (Figures S3.26 and S3.27). In summary, the results demonstrated that we successfully acquired the stereocomplementary product.

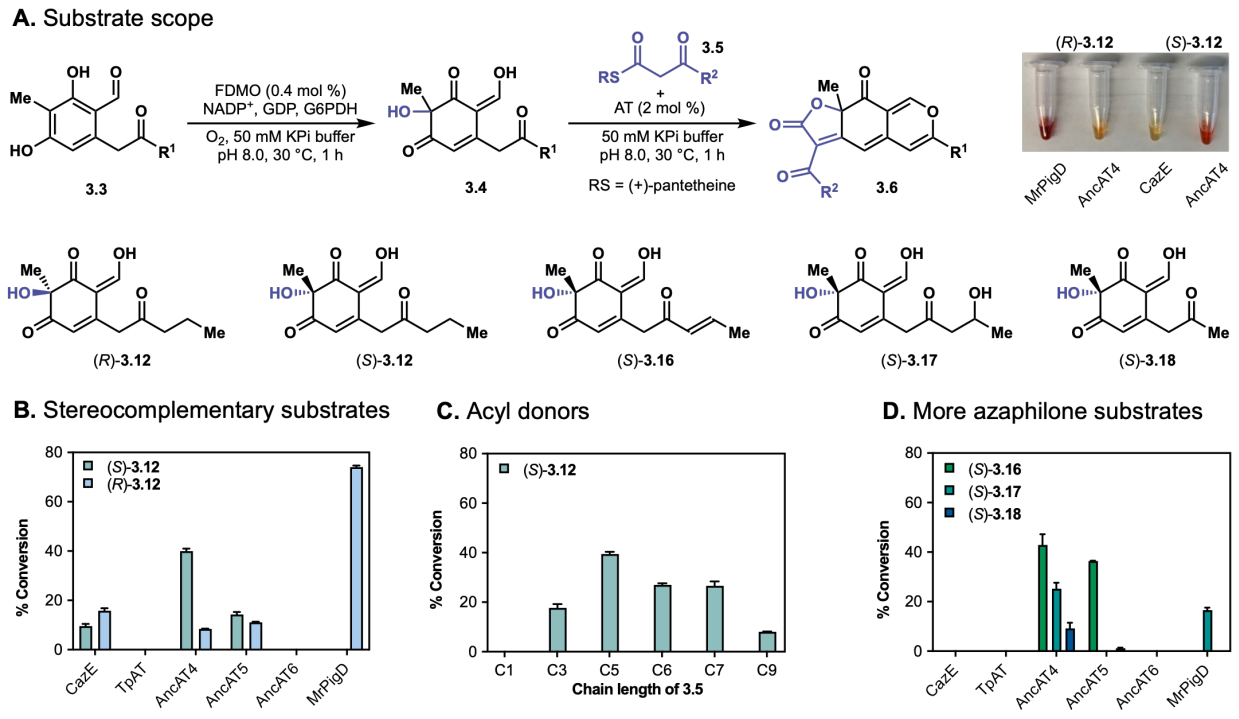


Figure 3.9. Substrate scope of the acyltransferases (ATs) to access linear tricyclic azaphilones.
A. Reaction conditions and substrate scope. **B.** Stereocomplementary substrates. **C.** Acyl donors with different chain lengths. **D.** More azaphilone substrates.

Our initial hypothesis was that the function and the (*S*)-substrate selectivity of AncAT4 could diverge during evolution, leading to that seen either in the clade including AT5 and CazE or that in the clade of AT6 and TpAT (Figure 3.7A). However, the inferior activity shown in all AT4 descendants suggested that the higher reactivity of AT4 might be a different function/selectivity compared with CazE and TpAT and is less likely a function/selectivity kept in any lineage. This is supported by the fact that our targeted reaction was different from the native reaction of CazE and that no activity was seen when running the targeted reaction using TpAT. Sequence reconstruction provides an opportunity to visualize evolutionary trajectories at the molecular level. As a result, all ancestors and extant enzymes in the AncAT4 clade were aligned and the active site center was assigned as the geometric center of His166 in the key HX_3D motif. Based on the AlphaFold2 model of AncAT4, 96 residues were identified within 18 Å of the active site (SI). AncAT4 shares 87% sequence identity to AT5 and 95% identity to AT6 (Figure 3.7B). Interestingly, all 28 closest residues within 10.4 Å are identical between AT4, AT5 and CazE (Table S3.5). From 10.7 to 13.7 Å, 35 residues are identical between AncAT4 and AT6. The remaining 33 more distant residues of AncAT4 from 14.0 to 18.0 Å have

one and four substitutions compared with AT5 and AT6, respectively (Table S3.5). Taken together, the core region of AncAT4 centered around the active site can be considered a chimera of AT5 and AT6 (Figure 3.8D). Based on this analysis, we suggested that the first-sphere identity of AncAT4 to AT5 endows the basal activity of AT4 and the second-sphere identity to AT6 potentially introduces a different substrate selectivity and thus activity further.

Structural basis for observed reactivity in acyltransferases. Considering the high similarity between AncAT4 and other less reactive ATs, we were interested in gaining further insight into what differentiates AncAT4 apart from others. We hypothesized that molecular modeling would provide a structural basis for the enzyme's enhanced activity. The molecular modeling was done by Azam Hussain. Structural models of AT4–6 and CazE were generated with AlphaFold2[48] and the phosphopantetheine C5-thioester was introduced based on the homologous structure of 2E1T[49], a malonyl transferase (Figure 3.10A). The model substrate (*S*)-**3.12** was docked into these models using the CDOCKER module from pyCHARMM[50] (Figure 3.10B). To identify potential reactive poses, we calculated the distance between the oxygen of tertiary alcohol on (*S*)-**3.12** and the carbonyl carbon of phosphopantetheine C5-thioester for the top cluster poses of each docking approach, plotted in Figure 3.10B. We observe that with flexible CDOCKER[51-52] modified to use a flexible phosphopantetheine C5-thioester, AncAT4 was the only system with a putative reactive pose for the majority of the top cluster poses. This suggests that the apo-structure prediction from AlphaFold requires significant reorientation of the binding site and phosphopantetheine C5-thioester, and that flexible receptor docking is a useful tool for future screening of acyltransferases.

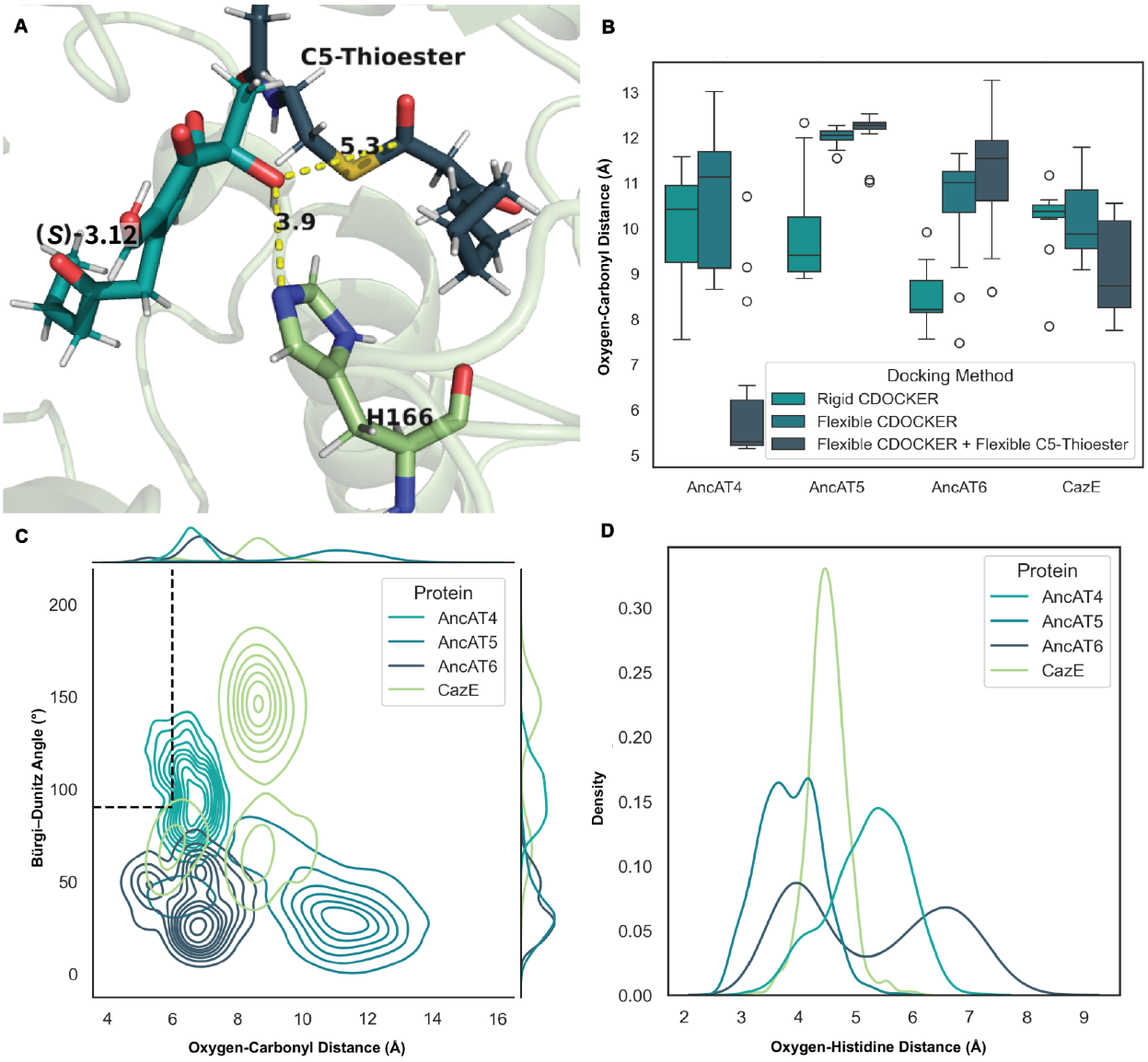


Figure 3.10. Substrate docking and molecular dynamics (MD) simulations.
A. Near attack complex of the substrate (*S*)-**3.12** in the active site of AncAT4. Near attack complex was selected by finding smallest (*S*)-**3.12** and C5-thioester carbonyl carbon distance with a Bürgi-Dunitz angle within 90º and 107.5º from 40 ns simulation. **B.** Substrate OH and C5-thioester carbonyl carbon distance of top cluster poses from rigid receptor docking (Rigid CDOCKER), flexible receptor docking (Flexible CDOCKER), and Flexible CDOCKER with a flexible C5-thioester group. **C.** Joint plot of the OH-carbonyl carbon distance of (*S*)-**3.12** with the Bürgi-Dunitz angle obtained from 40 ns simulation. The black dashed lines indicate an OH-carbonyl carbon distance of less than 6 Å and Bürgi-Dunitz angle of greater than 90º. **D.** Plot of the histidine-OH distance obtained from 40 ns simulation.

The top clustered pose for AncAT4 with a flexible phosphopantetheine C5-thioester was positioned in CazE and AncAT4–6 and simulated for 40 ns to assess the stability of the reactive geometry for each enzyme. In the simulations, AncAT4 maintained the (*S*)- or (*R*)-**3.12** near a reactive geometry most effectively, while other enzymes explored various potentially unreactive conformations. This is depicted in Figure 3.10C

and S3.28, which plots the Bürgi-Dunitz angle[53] of the acylation reaction against the carbonyl carbon-*S*-(**3.12**)-oxygen distance. The Bürgi-Dunitz angle characterizes the angle of attack of nucleophile to the carbonyl plane and ranges from ~90° to 107.5° for effective nucleophilic attack. We adopt a reactivity criterium cutoff of an angle between ~90° and 107.5° and OH-carbonyl distance of less than 6 Å similar to previous work characterizing human acyltransferases.[54] With these criteria, AncAT4 is the only enzyme that occupies a reactive geometry during the course of the simulation, with 2% of frames within this cutoff. The pose with the closest OH-carbonyl distance and reactive angle is depicted in Figure 3.10A. In addition, the majority of the enzymes maintain a histidine-OH distance of less than 6 Å (Figure 3.10D), suggesting that the orientation of the substrate and C5-thioester is key for determining reactivity. The docking and simulation results collectively suggest that the structural basis for AncAT4's increased reactivity is its superior positioning of (*S*)-**3.12** in a reactive conformation, with the identified reactive conformation illustrated in Figure 3.10A. Furthermore, this study demonstrates the potential of molecular modeling as a system to guide the exploration of other ancestral enzymes and the testing of potential binding site and second-sphere mutations to modulate activity.

Enzyme promiscuity is generally defined as the ability of an enzyme to perform reactions beyond their primary physiological functions.[55] This latent trait is crucial since the ability for an enzyme to carry out non-native reactions provides its corresponding organism with some selective advantages, enabling the organism to adapt to changing selection more quickly.[56] For this reason, promiscuity is widely considered as a driving force of evolution.[57] Inspired by nature, if possible, intentionally introducing promiscuity into an extant enzyme within a sequence space of interest by ASR would be a preferable strategy over the random search for an extant enzyme to achieve desired chemical transformations in some circumstances.[58] Additionally, based on a reliable reconstruction, sampling representative ancestral sequences provides a more efficient and economical way than sampling extant sequences to explore the sequence space of interest.[59] In our case, five ancestral ATs (AT3–7) can potentially cover the exploration of the sequence space that includes ten extant sequences. Screening these representative ancestral enzymes can help narrow down the target sequence space and guide future sampling of

extant enzyme homologs if necessary. In summary, our bioinformatics-guided ASR method offers a strategically efficient approach to generate novel enzymes with different selectivity. This is particularly beneficial for designing non-native selectivity or reactivity that is challenging to obtain even through extensive screening of existing enzymes and subsequent engineering.

3.4 Conclusions

Tailoring enzymes toward a specific selectivity or novel reactivity is a goal that biocatalytic practitioners strive to achieve and are often met with challenges. In the present study, we solved the challenge of achieving a stereocomplementary chemoenzymatic synthesis of azaphilones as an example to demonstrate a gene mining approach that can be seamlessly combined with ASR to generate ancestral enzymes fulfilling a designed set of synthetic requirements. Specifically, a chemoenzymatic sequence comprising two ancestral enzymes, AncFDMO and AncAT, was designed to enable the biocatalytic synthesis of tricyclic azaphilones. To acquire a suitable AncAT, a bioinformatics-guided workflow was adopted to find an appropriate sequence space where ancestral enzymes were sampled. Starting from identification of initial hits, the strategy enables users to efficiently access active enzymes by circumventing the time- and resource-intensive process of traditional protein engineering and by overcoming challenges associated with protein starting points with limited solubility and/or thermal stability. The best reconstructed AT demonstrated prominently improved soluble expression, reactivity, selectivity, and significant potential to be further engineered. According to our results from experiments, sequence alignment, docking, and MD simulations, we proposed that the reactivity seen in the best reconstructed AT, which is a chimera of multiple extant enzymes, was likely related to the promiscuity introduced during the ASR. Considering that promiscuity has been leveraged by nature to drive evolution, we envision that our approach to introduce promiscuity into potential ancestral sequence space by ASR will be valuable for tailoring enzymes for specific selectivity or reactivity in biocatalysis.

3.5 Experimental Procedures

3.5.1 Plasmids, DNA and protein sequences

General information

All genes encoding acyltransferases in a pET-28a(+) vector were codon-optimized for expression in *E. coli* and purchased from Twist Bioscience. *E. coli* cloning strain DH5α (Invitrogen) was used for DNA propagation. ddH2O was sourced from a MilliQ Biocel water purification unit from Millipore. The restriction sites were EcoRI and NotI for the *cazE* gene, BamHI and XhoI for the genes from UniProtKB (A0A#######), and NdeI and XhoI for other genes. The DNA and amino acid characters highlighted in red are not from the genes but from the templates instead.

Sequence information

CazE sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGG
CTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCGAATTCATGGGCAGCAACCAGGATGA
CGTGATTACGCAGCAGCCAGGCGCAGGCGAAGACGGTGTTGGTAGCAGCAATGTGAATCTGGAT
CGCCTGCCTGACCTGATGGGTCAGTTCCCGGTTCTCAACGCCTATACCCATATTACCTTTGGCT
TTGAATTGCCGCTGGACGTTGATCGCGATGCCATTACCAGCGCGTTGCAGACCAGTCTCGACCG
CTTGATTGAACTGATTCCGTGGCTGGGCTGGCAAATTGGTCTGGAAAGCGGCGTGCGCACCACC
GTACCGTGGCCGGAAGATACCCCGCGCGTGTTGCTGCATGTGAAAAACTGCGATGATATTGTGG
TGCCAATGGCACAACTGCTGGCTGCGGGCATTCCAATCAACAAACTCGATGCGAAAGTGCTGGC
GCCCTGGCCTGCACTGCCGCAGCCGCATGGGCTGACCGGTCCGGTGCCGGTGATCACGCTGCAG
GCGAACTTTGTTCGCGGCGGCCTGATTCTGAACTTCAGCAGTCATCATACCGTGATGGATGGCA
CCGCCAACCTGCAGGTTCCGAAACTGTTTGCCGCCATCCTCAATGGCGACACCATTCCTGAAGC
GGATCTGCAACAGGCAAACCGCGACCGTCGTGGCTTAATCGAATTAATTCCGCGTAGCGAACCG
ATGAAAGATTACAGCCATCTGCGTGGTCCACCGGGCTATCGCTTCACCCTGCCGGCGTCGGCAC
CGATCTGGTGCCATTTTAACATGCCGATGGCCAGCTTAAGCAAATTAACCAAAAGCGCCCGTGA
TCCGTCGCGCCCGGTAAACGAAGATGATGTTCTGATCGCCTTCTTCTGGCAGCGTCTGTGCGCG
GCGCGTATTGCCCGCGGTATTGCTCGCGATACGGTGAGTAAAATCAGCCGTGCAATTGATGCAC
GCGTTGCCCTCGGTATCCCGCCAACCTATCTGGGTGCCCAGGTGCATATGGCGATCACCCGCTT
ACCGATGGGGCAGGTCGCCAGTCTGACGTTGTTGCAAGTGGCGCAGGTGCTGCGCCGTGCGCTG
GCAGACGCCAATACGCCGTGGGCGATTCGCTCGCTGGTGACGTTTATTGCGCGCGAGCCGGATC
GTTCGCGTTTACTGTATAACGGCCCACACGATGGTAACACCGATGTGGGGGCGACCTCCAGCCT
GAATATGGCGCATGCGCGTCCGCCGCACTGGGGCCCCGTGCTCGGCCCGTCACGTTTTTTTAGC
CGCAGTAACGGTGCGCCGATTCCCGGCAGTTTTACCGTTTATAGCCCGGACGGCGCGGGTTTTG
TGCGTATGTGTATTTGCCTGCCAGTTGACGATCTCGAAGCCCTGAAAAAAGACGCGCTGTGGAA
ACAGCATACGAAATTTGTCGGTGCGGCCGCACTCGAGCACCACCACCACCACCACTGA

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSEFMGSNQDDVITQQPGAGEDGVGSSNVNLD
RLPDLMGQFPVLNAYTHITFGFELPLDVDRDAITSALQTSLDRLIELIPWLGWQIGLESGVRTT
VPWPEDTPRVLLHVKNCDDIVVPMAQLLAAGIPINKLDAKVLAPWPALPQPHGLTGPVPVITLQ
ANFVRGGLILNFSSHHTVMDGTANLQVPKLFAAILNGDTIPEADLQQANRDRRGLIELIPRSEP
MKDYSHLRGPPGYRFTLPASAPIWCHFNMPMASLSKLTKSARDPSRPVNEDDVLIAFFWQRLCA
ARIARGIARDTVSKISRAIDARVALGIPPTYLGAQVHMAITRLPMGQVASLTLLQVAQVLRRAL
ADANTPWAIRSLVTFIAREPDRSRLLYNGPHDGNTDVGATSSLNMAHARPPHWGPVLGPSRFFS
RSNGAPIPGSFTVYSPDGAGFVRMCICLPVDDLEALKKDALWKQHTKFVGAAALEHHHHHH*

TpAT (XP_046069174.1) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGAGTAGCCGTCAAGATGAACTCAAGGCTCAACAACCGGGCGAGGTCGAGAACGTGGATGAAAG
CCTGTATCCTGACTTGATGGGTCAATTTCCTATTCTTGACGCTTACACACAACTGTGCTTTGGG
TTTGAGCTGCCTCTTGACGTAGATCGCGACGCCGTCGTTAGCGCGCTTCGTACTGGCTTTAACC
GTTTAATAGAGCATATTCCGTGGTTAGGTTGGCAAATCGGACAACGGTCAGGCGTTCGTACAGC
AATACCGTGGCCCGAGGACGTAGCGCGCGACTTATTATACGTAAAGTACTGCGACGAAACAACC
CAGCCTATGGCAGACCTGTTGGGTGCTGGCATACCCATTGGTAAGCTGGATGGCAAGGAGCTTA
CGCCGTGGCCAGCCCTGCCTCAGCCGCATGGAATTGTGGGACCTGTTCCTGTCCTCGCCTTACA
GGCGAATTTCATCAAAGGCGGCCTGATCTTAAACCTGTCCTCTCACCACACGATGATGGATGGC
ACCGCTAATTTCCAAGTTTTAAAGCTGCTGGCAACTGTTTTAGCGGGCGACGAGATACCCGCCG
CGGAGTTAGAACAAGCCAACAGAGACCGTCGCGGGTTGATTGACCTGATTCCGCGTTCTGAGGC
CCTGAAAGATCACAGCTACTTGAGACGTCCACCAGGTTACCAGTTCGTATTCCCAGCGTCACCC
CCTACGTGGTGTTATTTCAAGATGCCTCTCACGTCGTTGTCAAAGCTGGCCAAAACTGCCCAAG
ACCCTAGCCGTCCGGTGACGGAAGACGACATTTTACACGCTTTCTATTGGCAGCGTCTGTGCGC
TGTACGCCTTGCCAGAGGTCTGCCCTTCGACACTGTCTCTAAGATCAGCCGCGCGATTGATGGT
CGAATGGCATTGGGTATCCCGGCATCTTACATGGGAGCGCAAGTCTACACCGCAATTACTCGTC
TCCCAATCGGACAAGTGGCAAGTCTGACTTTGCAGCAGACCGCCCAAGTACTGCGACGCGAGTT
CTCACAAGCCAACACCGCATGGGCAATCAGATCTCTTGCCACAATCATCGCCCGTGAGCCGGAC
CGGTCTATGCTTATGTATAACGGAACACACAACGGAAACACGGACATAGGTGCGACGAGCTCTC
TGAATACGAACCAGACTCTTCCGCCTTCGTGGGGACACTTACTGGGTCCATGTCGTTTCTTTAG
AAGATCGATATCGGCCCCCATTCCAGGTGCTTTTGTCGTACAGAGTGCTGAAGGCGGTGCTATT
CCAATCTTAGTTTGCTTACCACAAGACGACTTGCAAGCTCTGAAGAAAGACAGCCAATGGCGCC
AATACACTAGATGTATCGGCTAG

MGSSHHHHHHSSGLVPRGSHMMSSRQDELKAQQPGEVENVDESLYPDLMGQFPILDAYTQLCFG
FELPLDVDRDAVVSALRTGFNRLIEHIPWLGWQIGQRSGVRTAIPWPEDVARDLLYVKYCDETT
QPMADLLGAGIPIGKLDGKELTPWPALPQPHGIVGPVPVLALQANFIKGGLILNLSSHHTMMDG
TANFQVLKLLATVLAGDEIPAAELEQANRDRRGLIDLIPRSEALKDHSYLRRPPGYQFVFPASP
PTWCYFKMPLTSLSKLAKTAQDPSRPVTEDDILHAFYWQRLCAVRLARGLPFDTVSKISRAIDG
RMALGIPASYMGAQVYTAITRLPIGQVASLTLQQTAQVLRREFSQANTAWAIRSLATIIAREPD
RSMLMYNGTHNGNTDIGATSSLNTNQTLPPSWGHLLGPCRFFRRSISAPIPGAFVVQSAEGGAI
PILVCLPQDDLQALKKDSQWRQYTRCIG*

A0A084B9Z3 sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGG
CTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCATGAGCACCATGGCGAAAAGCCCGGA
AGCGAACAACCTGCATCAGGATGTGATTGCGCAGTTTCCGATTCTGAACGGCTATACCCATACC
GTGGGCGCGTTTAGCCAGCCGCTGAACGTGAGCCGCCTGTTTATTATTGATGAAATTCAGACCG
CGTATGATGAACTGCGCGTGCAGATTCCGTGGCTGGCGCATCAGGTGGTGGTGGTGGATGCGGG
CCCGGGCAAAAGCGGCTATATTACCACCGCGCCGTGGCCGAGCAGCGCGCCGCCGAACGATGTG
ACCTATGAAGAAAAAGATGATGCGTTTCCGAGCCTGAACACCCTGATTAAAAGCGGCGGCAGCT
TTCTGGCGACCAAAGATCTGGTGGGCTATCCGGGCCTGCCGGAACCGCATGGCCTGCATCCGAC
CCCGGTGGCGACCATTCGCCTGGTGTTTATTACCGGCGGCGTGCTGGTGGTGCTGAGCACCCAT
CATAACATTGTGGATGGCATTGGCCTGATGCAGATGTGGGATTATCTGGATATTCTGATGGGCG
GCGGCGCGATTAGCCGCCAGGATGCGCGCAGCGCGAACGCGGATCGCGCGCGCGTGCTGCCGCT
GATTGCGCCGGGCGAACCGGTGAAAGATTATAGCCATCTGATTCGCCCGAACCCGTGGCCGCTG
CCGCCGCCGCCGAAAACCGAATGGCGCCTGTTTAAAATGCATCCGTGGGCGCTGGCGGAAATTC
GCAGCCGCGCGCGCGATGGCACCGATCAGCGCGCGAGCGCGCGCCCGGCGAGCAGCGATGATGC
GCTGACCGCGTTTTGCTGGCAGCGCGTGAGCGCGATGCGCCTGGCGAGCGGCCGCGTGACCGGC
GATCAGGTGAGCAAATTTGGCCGCGCGGTGAACGGCCGCAGCGCGATGGGCCTGGATAGCAGCT
ATCTGTTTCATATGATGCTGCATACCGAAACCCGCCTGCCGATTGAACAGATTGCGCGCAGCAC
CCTGGCGGAACTGAGCACCCAGCTGCGCAAAGATCTGGATGCGGCGCGCACCGAATGGAGCGTG
CGCAGCTATGCGACCTTTCTGGCGGGCGTGGCGGATAAAACCCGCCTGCTGTATGGCGGCATTA
CCAACCCGCAGACCGATCTGGGCGGCACCAGCACCATGCATTGGGCGAGCCGCCGCCCGATTCG
CCTGGGCCTGCTGGGCGATTGCCATCTGATTCGCAAACCGGAAGGCATGCCGCTGCCGGGCTGC
CTGTATTTTATGCCGAGCGGCGGCACCAGCGGCGTGGTGCAGCTGCTGCTGTGCCTGCCGAAAG
AAGAACTGGATGCGCTGCAGGAAGATGCGGAATGGAAACATTATACCGAAAGCGGCGGCCGCCG
CGTGGATGGCCCGCGCCTGCTCGAGCACCACCACCACCACCACTGA

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSMSTMAKSPEANNLHQDVIAQFPILNGYTHT
VGAFSQPLNVSRLFIIDEIQTAYDELRVQIPWLAHQVVVVDAGPGKSGYITTAPWPSSAPPNDV
TYEEKDDAFPSLNTLIKSGGSFLATKDLVGYPGLPEPHGLHPTPVATIRLVFITGGVLVVLSTH
HNIVDGIGLMQMWDYLDILMGGGAISRQDARSANADRARVLPLIAPGEPVKDYSHLIRPNPWPL
PPPPKTEWRLFKMHPWALAEIRSRARDGTDQRASARPASSDDALTAFCWQRVSAMRLASGRVTG
DQVSKFGRAVNGRSAMGLDSSYLFHMMLHTETRLPIEQIARSTLAELSTQLRKDLDAARTEWSV
RSYATFLAGVADKTRLLYGGITNPQTDLGGTSTMHWASRRPIRLGLLGDCHLIRKPEGMPLPGC
LYFMPSGGTSGVVQLLLCLPKEELDALQEDAEWKHYTESGGRRVDGPRLLEHHHHHH*

A0A094B1C4 sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGG
CTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCATGGATTTTCGCCCGTATGTGGAAGA
TACCAGCGCGAGCGCGCCGGAACAGCCGGCGCAGGCGGAAAGCAAAACCCAGGTGGATAGCATT
GATGAACTGCATCAGGATGTGGTGAGCCAGTTTCCGTTTCTGAACGGCTATAGCCATATTGTGT
TTCTGTTTCAGCCGGATGCGGATACCAGCCGCGAAACCATTATTGCGACCGTGCATACCGCGCT
GGCGAAAGTGACCGCGCAGGTGCCGTGGCTGGCGGGCCAGGTGCTGCATAACCCGGGCCCGCCG
GGCAACAGCGGCCATTATATGCTGGCGCCGTGGCCGCTGAACGCGCATAAAAACGATATGGTGC
GCGTGGAAGATTGCGAAGATCGCATGCCGCCGATTGCGCAGATTGTGCAGGCGGGCGCGCCGAT
TAGCATGCTGGATGCGAAAATTCTGACCCCGTTTCCGAGCCTGCCGCTGCCGCATGGCCTGGAA
CCGCCGCTGCCGGTGATGCTGCTGCAGGTGAACTTTCTGCGCGGCGGCGTGATTATTAACCTGA
GCACCCATCATATGGCGGTGGATGGCAACGGCATTGTGCAGATTATGCGCCTGCTGGCGACCGT

GCTGCAGGGCCAGGAAATTGCGGCGGCGGATCGCGAACAGGCGCGCCGCGATCGCCGCCGCGTG
GTGCCGCTGATTCCGCGCGGCGAACCGGTGAAAGATCATAGCCATCTGCGCGCGCCGCCGGGCC
ATGTGACCGTGCCGCCGAGCAGCCCGCCGAAATGGTGCTATTTTAAACTGCCGGTGAGCGCGCT
GCCGACCCTGAAAAAACTGGCGAGCCCGCCGAGCCAGCAGCAGCAGCTGAGCGAAAACGATGTG
ATTTGCGCGTTTTGCTGGCAGCGCATTACCGCGATGCGCCTGGCGCGCGGCTTTAGCCCGGATA
CCATGACCAAAATGACCCGCATTATTGATGCGCGCGCGGCGCTGGGCGTGCCGATTACCTATAT
GGGCCATCTGACCTATTATGCGGTGGCGCAGCTGCCGATGCATCTGGTGACCAGCCTGCCGCTG
AGCAGCATTGCGCAGACCCTGCGCCGCGAACTGAACGCGGCGAACACCACCTGGGCGATTCGCA
GCTATGCGACCTTTCTGGCGCGCGAACCGGATAAAAGCCGCCTGGCGTATGCGGGCCTGCGCAA
CCTGGATACCGATCTGGGCACCACCGCGTTTACCGCGAGCCAGAGCACCAGCCATGATGATGGC
AGCGCGATTCCGGAAGATTTTGGCCCGGTGCTGGGCCGCCTGAAATATACCCGCCGCCCGAACT
GCAGCCCGCTGGTGGGCGGCATTACCATTTGCCCGGTGGAAAGCGGCGCGATTCCGATTGTGAT
GTGCCTGCCGGGCGCGGATCTGGAAGGCCTGGAAAAAGATCGCGAATGGCGCCGCTATGCGCGC
TATATTGGCCTCGAGCACCACCACCACCACCACTGA

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSMDFRPYVEDTSASAPEQPAQAESKTQVDSI
DELHQDVVSQFPFLNGYSHIVFLFQPDADTSRETIIATVHTALAKVTAQVPWLAGQVLHNPGPP
GNSGHYMLAPWPLNAHKNDMVRVEDCEDRMPPIAQIVQAGAPISMLDAKILTPFPSLPLPHGLE
PPLPVMLLQVNFLRGGVIINLSTHHMAVDGNGIVQIMRLLATVLQGQEIAAADREQARRDRRRV
VPLIPRGEPVKDHSHLRAPPGHVTVPPSSPPKWCYFKLPVSALPTLKKLASPPSQQQQLSENDV
ICAFCWQRITAMRLARGFSPDTMTKMTRIIDARAALGVPITYMGHLTYYAVAQLPMHLVTSLPL
SSIAQTLRRELNAANTTWAIRSYATFLAREPDKSRLAYAGLRNLDTDLGTTAFTASQSTSHDDG
SAIPEDFGPVLGRLKYTRRPNCSPLVGGITICPVESGAIPIVMCLPGADLEGLEKDREWRRYAR
YIGLEHHHHHH*

A0A094J5P9 sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGG
CTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCATGGCGGAAGAAGCGCTGCTGCAGGA
TGTGATGGGCCAGTTTCCGCAGCTGAAAACCTATAACCATGGCAGCCTGATTTTTAGCATGCCG
GATGGCGTGAGCCGCGAAAGCCTGGTGACCGCGCTGGAAGCGAGCACCGGCAAAATTCTGAGCG
CGATTCCGTGGCTGCATGAACAGGTGGTGCATGTGGGCCAGGCGCCGGGCACCAGCGGCAAATT
TGAACTGGCGCCGTGGCCGGCGGATGCGCCGAAAAACACCCTGATTCGCGTGCGCGATTGCAGC
GATCTGCTGCCGAGCCATGAAGAAATGCTGAAAAGCCAGGGCCCGGCGAGCATGCTGGATGGCA
ACCTGATTTGCGCGGTGCCGGGCTTTCCGCTGCGCTATGATGAAGCGGTGATTGGCCCGGCGCC
GGCGGCGATTATTCAGGTGAACTGGATTAAAGGCGGCGTGATTCTGACCTTTAGCAACCAGCAT
AACGTGATGGATGGCAGCGGCGTGTTTCAGCTGATTGCGCTGCTGAGCACCGCGATGAACGGCG
AAGAAATTAGCAAAAGCGCGATTGAAGAAGGCAACCGCGATCGCAAAACCGTGGTGCCGCTGTA
TGGCCCGGATGAACCGATTCGCGATCATAGCCATATGCTGATTAAAACCCCGCTGAGCGCGGCG
GCGGCGCCGAACAAAGCGCAGAGCCTGAGCAGCTGGGTGTTTCTGCGCTTTTGGAAAAAAAGCG
TGCCGGAAATTAAAGCGATGGCGACCGATCCGGTGGGCTATGATAAAAGCGTGCCGTTTATTAC
CAGCAACGATGCGATTAGCGCGTTTTATTGGAAACGCCTGGCGATTGCGCGCGTGCGCAACGGC
CAGGATGGCCAGGCGCTGAGCAAATTTAGCCGCGCGATTGATGCGCGCGCGGTGATGGGCGTGA
GCCGCGAATATATGGGCCAGATGGTGTATTTTAGCGCGACCTGGCTGACCTATCAGGAACTGGT
GGATCTGCCGCTGAGCACCATTGCGAGCCGCATGCGCAAAAGCATGAACGAAAGCAACAACGAA
TTTAGCGTGCGCAGCTATGCGACCTTTGTGAGCCGCATTCAGGATCGCACCACCATTGCGTATG
GCGGCCCGTTTAACCGCGAAACCGATATTGCGAGCAGCAGCATGGCGCAGGCGGCGGTGGTGCT

GAAATTTGGCATTCTGGGCACCCCGGAACTGGTGCGCCGCCCGAACCTGGCGCCGATTCCGGGC
ACCCTGTATTTTTTTCCGCCGGATGGCAGCGGCGATCTGAACCTGCTGATGTGCCTGACCCAGC
AGGAACTGGAAGCGATGCGCAGCGATCCGGAATGGAGCCAGTGCACCCAGTATATTGGCCTCGA
GCACCACCACCACCACCACTGA

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSMAEEALLQDVMGQFPQLKTYNHGSLIFSMP
DGVSRESLVTALEASTGKILSAIPWLHEQVVHVGQAPGTSGKFELAPWPADAPKNTLIRVRDCS
DLLPSHEEMLKSQGPASMLDGNLICAVPGFPLRYDEAVIGPAPAAIIQVNWIKGGVILTFSNQH
NVMDGSGVFQLIALLSTAMNGEEISKSAIEEGNRDRKTVVPLYGPDEPIRDHSHMLIKTPLSAA
AAPNKAQSLSSWVFLRFWKKSVPEIKAMATDPVGYDKSVPFITSNDAISAFYWKRLAIARVRNG
QDGQALSKFSRAIDARAVMGVSREYMGQMVYFSATWLTYQELVDLPLSTIASRMRKSMNESNNE
FSVRSYATFVSRIQDRTTIAYGGPFNRETDIASSSMAQAAVVLKFGILGTPELVRRPNLAPIPG
TLYFFPPDGSGDLNLLMCLTQQELEAMRSDPEWSQCTQYIGLEHHHHHH*

A0A135LJ73 sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGG
CTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCATGAGCATTGCGATTGAAAACCATCT
GACCAAAGGCGAAACCATTGAAACCGATAGCGAACTGCCGCGCGATATTGTGAGCCAGTTTCCG
ATTCTGAACGCGTATACCCAGCTGCTGTTTGGCTTTAAACTGCCGGCGGATGTGGATCGCGATG
CGATTGTGGCGAGCCTGCAGGATGGCTTTGATAAACTGCGCGCGGAAATTCCGTGGCTGGGCTG
GCAGGTGGCGCGCGAAAGCGGCCCGAACGGCATTCTGAAAGCGATGCCGTGGCCGGCGGATGTG
CCGAAAGAACGCATTCGCGTGAAAAACTGCGATGATCTGGTGGCGCCGATTAGCAAACTGGTGA
GCGCGGGCGTGCCGATTCATATGCTGGATGGCAGCGTGCTGACCCCGTGGCCGGCGCTGCCGCA
GCCGCGCGGCCTGGATAGCCCGGATCCGGTGGTGGCGATGCAGGCGAACTTTGTGCAGGGCGGC
CTGATTCTGAACCTGAGCACCCATCATACCATTATTGATGGCACCGGCATTTATCAGTTTGTGA
ACCTGCTGGCGCTGGTGATGAGCGGCAAAAAAATTCCGGCGGCGGATCTGGAACAGGCGAACCG
CGATCGCAGCCGCGTGATTCCGCTGATTCCGCGCACCGAACCGGTGAAAAGCTATATTCATCTG
CGCCGCCCGCCGGGCTATACCTGGGCGCCGCCGAGCAGCCCGCCGATGTGGTGCTATTTTAAAA
TTCCGGTGAACGCGCTGGCGCGCCTGATTAAAAGCGTGAAAGATGATCCGCTGAGCAACAAACC
GGGCAGCATGATGGTGAGCGATAACGATATTTTTAGCGCGTTTGCGTGGCAGCGCCTGTGCGCG
GTGCGCGTGGCGAACGGCCAGCCGCCGGAACGCAGCAGCAAACTGGGCCGCGCGATTGATGGCC
GCATGGCGCTGGGCGTGCCGCTGACCTATATGGGCCATATGGTGTGCCATGCGCTGGTGCGCCT
GCCGCTGGGCCAGGTGGCGAAACTGCCGCTGCCGCAGCTGGCGCAGGTGCTGCGCCGCGAACTG
AACCAGGCGAACACCCCGTGGAGCATTCGCAGCTATGCGACCTTTATGGCGCGCGAACCGGATA
CCAGCAGCCTGCTGTATGGCGGCACCCATGATGCGCGCGCGGATCTGATGGTGACCGCGGTGGG
CCAGGCGACCCCGCTGCCGGCGAGCTGGGGCCCGCTGCTGGGCCGCAGCTGCTTTTTTCGCCGC
CCGACCGCGGCGCCGATTCCGGGCTGCTTTATTATTAACGAAACCGAAGGCGCGTTTATTCCGC
TGACCCTGTGCATGCCGGAAGAAGATATTATTGGCCTGAAAAAAGATCCGGTGTGGAAACAGTA
TATTCGCTATGTGGGCCTCGAGCACCACCACCACCACCACTGA

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSMSIAIENHLTKGETIETDSELPRDIVSQFP
ILNAYTQLLFGFKLPADVDRDAIVASLQDGFDKLRAEIPWLGWQVARESGPNGILKAMPWPADV
PKERIRVKNCDDLVAPISKLVSAGVPIHMLDGSVLTPWPALPQPRGLDSPDPVVAMQANFVQGG
LILNLSTHHTIIDGTGIYQFVNLLALVMSGKKIPAADLEQANRDRSRVIPLIPRTEPVKSYIHL
RRPPGYTWAPPSSPPMWCYFKIPVNALARLIKSVKDDPLSNKPGSMMVSDNDIFSAFAWQRLCA

VRVANGQPPERSSKLGRAIDGRMALGVPLTYMGHMVCHALVRLPLGQVAKLPLPQLAQVLRREL
NQANTPWSIRSYATFMAREPDTSSLLYGGTHDARADLMVTAVGQATPLPASWGPLLGRSCFFRR
PTAAPIPGCFIINETEGAFIPLTLCMPEEDIIGLKKDPVWKQYIRYVGLEHHHHHH*

A0A194VYW6 sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGG
CTAGCATGACTGGTGGACAGCAAATGGGTCGCGGATCCATGACCATTGATAACACCCAGGCGCA
GGTGGGCGATGATGATCTGTATCAGGATGTGATGAGCCAGTTTCCGTTTCTGAACGGCTATACC
CATTTTATGCTGGGCTTTCAGCTGGATAGCGATGAAAGCCATAACGCGATTGTGGATGCGCTGA
AAACCGGCGTGCAGCGCGTGATTGATAAAGTGCCGTGGCTGGGCGGCCAGGTGGCGTGCGAAAG
CAGCCCGGAAGGCACCGCGACCTTTCGCCCGACCGCGTGGCCGGAAGGCCTGCCGGTGAACGAA
ATTGTGCGCCTGAACATTTGCGATGCGAAACTGCCGCCGATGGCGGAACTGCTGCGCACCGGCA
CCCCGGTGAGCGTGCTGCCGGCGAGCGTGCTGACCCCGTGGCCGGGCCTGCCGCAGCCGCATGG
CATTAGCGGCCCGGTGCCGATTATTGCGATGCAGGCGAACTTTATTAGCGGCGGCCTGATTCTG
ACCATTAGCTGCCATCATAACATTATGGATGCGACCGCGATTTTTCAGTTTATTCGCCTGCTGA
CCGTGGTGATGAAAAACGGCGAAATTATGCCGCCGGAACTGGATCAGGCGAACCGCGATCGCCG
CGGCGTGATTCCGCTGATTCCGCGCGGCGAACCGGTGAAAAACTTTAGCCATCTGCGCCGCGCG
CCGGATTATGTGCAGCGCACCCCGAGCAGCACCAGCCGCTGGTGCTATTTTCGCTTTCCGGTGG
CGGCGATGCCGCGCATGCGCAAACTGGCGAGCCGCCCGCAGACCACCATGAGCGAACAGGGCAG
CCCGGCGCCGCTGCTGAGCGATAACGATATTCTGAGCGCGTTTTGCTGGAACCGCATTAGCGCG
GTGCGCCTGGGCCGCGGCGCGATGACCGGCGGCCAGAAAGTGGGCCCGACCAGCACCACCAAAT
TTAACCGCGCGATTGATGGCCGCGTGGCGCTGGGCGTGCCGCCGAGCTATATGGGCCATATGGT
GTGCCATGCGAGCACCCGCCTGACCCTGCAGCAGGTGGTGACCCAGCCGCTGAGCACCACCGCG
GCGGCGCTGCGCCGCGATCTGAACGCGGCGAACACCCCGTGGAGCATTCGCAGCTATGCGACCT
TTCTGGCGCGCGAACCGGATCGCAGCAAACTGCTGTATGGCGGCAGCTTTGATCCGAACACCGA
TATTGGCGCGAGCAGCCTGGTGACCGATCTGACCGGCGCGAGCAAAGATGATACCGCGAGCGGC
TTTGCGTTTATGAGCCTGGGCGATCTGCTGGGCAAACCGCGCTTTATTCGCCGCCCGGATCTGC
CGCTGATTCCGGGCGTGCTGTATTTTCAGCCGGTGGAAAACGGCGCGATTCCGATTCTGCTGTG
CCTGAGCGCGGAAGATCTGGATGGCCTGATGAAAGATCGCCAGTGGACCCAGTATATGAAATAT
GTGGGCCTCGAGCACCACCACCACCACCACTGA

MGSSHHHHHHSSGLVPRGSHMASMTGGQQMGRGSMTIDNTQAQVGDDDLYQDVMSQFPFLNGYT
HFMLGFQLDSDESHNAIVDALKTGVQRVIDKVPWLGGQVACESSPEGTATFRPTAWPEGLPVNE
IVRLNICDAKLPPMAELLRTGTPVSVLPASVLTPWPGLPQPHGISGPVPIIAMQANFISGGLIL
TISCHHNIMDATAIFQFIRLLTVVMKNGEIMPPELDQANRDRRGVIPLIPRGEPVKNFSHLRRA
PDYVQRTPSSTSRWCYFRFPVAAMPRMRKLASRPQTTMSEQGSPAPLLSDNDILSAFCWNRISA
VRLGRGAMTGGQKVGPTSTTKFNRAIDGRVALGVPPSYMGHMVCHASTRLTLQQVVTQPLSTTA
AALRRDLNAANTPWSIRSYATFLAREPDRSKLLYGGSFDPNTDIGASSLVTDLTGASKDDTASG
FAFMSLGDLLGKPRFIRRPDLPLIPGVLYFQPVENGAIPILLCLSAEDLDGLMKDRQWTQYMKY
VGLEHHHHHH*

AncAT1 (Anc440) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGAGCTCTACGGCCGCCAGTACTGAAAGTGAGGACGATGACCTCTACCAAGATGTAATGTCTCA
ATTTCCATTCTTAAACGGGTATACCCATTTAGTATTCGGCTTCCAGCTGCCTTCGGACACGTCT

CGCGACGCCATTGTATCGGCTTTACAATCTGCGTTGGATAAGCTTACATCGCAAATCCCGTGGC
TTGGGTGGCAGGTGGTTAGAGTTCCGGGTCCTGGGTCTGGTATCCTTAAGCCGGCGCCCTGGCC
AGCAGATGCACCAGCAAATGAGATTGTGCGGGTCAAAGATTGTGACGACTTAATTCCCCCAATG
GCACAATTGCTGCGAGCCGGGGTCCCAATATCGATGCTGGACGGGAAGATTCTTACTCCCTGGC
CGGGATTACCTCAGCCACACGGCCTTGAGGGACCTGTCCCAGTCATTGCTCTTCAAGCAAACTT
TATCCGTGGCGGGTTGATCTTAAACATGTCAGCCCACCACACCATTATCGACGGAACGGGTATC
GTACAATTTATGCGATTGCTGGCTACGGTCCTTAACGGGCGAGAGATACCGGCCGCGGATCTGG
AACAAGCAAACCGTGACAGAAGCAGAGTTGTGCCACTCATACCAAGAGGCGAACCGGTAAAGGA
CTACTCCCACTTGCGCCGGCCACCAGGCTGGACTCCGGCCCCACCTGCTAGCCCACCTACTTGG
TGTTACTTTAAGTTGCCCGTTAGTGCCTTGGCGGCACTGATGAAGACCGCCAGCTCTCCGTCCC
AGCTGGTAAGTGAAAATGATGTATTGTGTGCGTTCTACTGGCAGCGACTGAGTGCTGTGCGTTT
AGCCCGTGGTTTCTCGCCAGACACCATGTCGAAGTTCTCACGTGCGATCGACGGGCGCACCGCT
CTCGGCGTTCCACTTTCTTATATGGGTCACATGGTGTACCACGCGGTGACTCGCCTGCCTATGG
GCCAGGTCGCGTCAATGCCGCTTTCGACGATCGCTCAAGCACTTAGACGTGAGTTAAACGCAGC
CAACACAGCCTGGGCCGTAAGATCCTACGCAACCTTCCTTGCCCGGGAGCCGGACCGCAGCCAG
TTGTTATACGGTGGGATGCACAATCCTAACACCGACCTTGGGGCAACCTCAACAGTGATAAGTA
GTAGCACAATGCCGGACTCCTGGGGACCGCTTCTTGGGCCTTGTCGCTTCCTGCGCCGCCCTAA
TACTGCGCCAATCCCAGGCTCATTTTCTATCAGTCCGGTAGAGGGTGGTGCGATCCCTATCGTA
CTTTGTTTGCCTGAAGAGGACTTAGAAGGCTTAAAGAAGGACCCCGAGTGGCGCCAATACACCA
GATACGTTGGTTGA

MGSSHHHHHHSSGLVPRGSHMMSSTAASTESEDDDLYQDVMSQFPFLNGYTHLVFGFQLPSDTS
RDAIVSALQSALDKLTSQIPWLGWQVVRVPGPGSGILKPAPWPADAPANEIVRVKDCDDLIPPM
AQLLRAGVPISMLDGKILTPWPGLPQPHGLEGPVPVIALQANFIRGGLILNMSAHHTIIDGTGI
VQFMRLLATVLNGREIPAADLEQANRDRSRVVPLIPRGEPVKDYSHLRRPPGWTPAPPASPPTW
CYFKLPVSALAALMKTASSPSQLVSENDVLCAFYWQRLSAVRLARGFSPDTMSKFSRAIDGRTA
LGVPLSYMGHMVYHAVTRLPMGQVASMPLSTIAQALRRELNAANTAWAVRSYATFLAREPDRSQ
LLYGGMHNPNTDLGATSTVISSSTMPDSWGPLLGPCRFLRRPNTAPIPGSFSISPVEGGAIPIV
LCLPEEDLEGLKKDPEWRQYTRYVG*

AncAT2 (Anc442) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGTCCTCTACCGCCGGATCTGCGGGTAATGCTGGGAACGATGATGACTTGTACCAAGACATTAT
GTCCCAATTCCCAGTATTGAACGCTTACACCCACCTGGTTTTCGGCTTTCAACTGCCGTCAGAC
GTAGACCGTGATGCCGTAGTTTCTGCACTTCAAACCGGCTTGGACAAGCTTGTTGAGCAAATCC
CTTGGTTAGGGTGGCAAGTGGCCCGCACCGGAAGCGGCATACTGAAGGCAGTTCCCTGGCCTCA
AGATGTCGCCAAAGAACGGGTGCGTGTGAAGTACTGCGATGATCTCATACCACCTATGGCACAA
CTGTTGGCTGCCGGTGTCCCCATCAATATGCTGGACGGGAAGGTCTTAACCCCATGGCCGGCAC
TGCCACAGCCGCATGGGCTTACGGGTCCGGTCCCGGTCATAGCTTTACAGGCCAATTTCGTCCG
CGGTGGGTTGATATTAAACCTGAGTGCGCACCACACCATCATGGACGGTACAGGTATCTTCCAG
TTCCTTAAATTACTCGCGACCGTCCTTAACGGCAAGGAGATCCCCGCCGCGGACCTCGAACAGG
CAAATCGTGACCGTAGCAGAGTCGTACCGCTCATACCACGTGGTGAACCGGTTAAGGACTATTC
CCATCTTCGCCGCCCGCCCGGCTGGACATTCGCCCCGCCTGCATCTCCGCCAACGTGGTGCTAC
TTTAAAATGCCCGTGTCTGCACTTGCTCGGTTAGTAAAGACTGCCCGTGATTCTTCCAGTCAAA
TGGTATCAGAGAATGACATCCTGTGTGCGTTCTACTGGCAACGGTTGTGCGCTGTTCGCTTAGC
ACGTGGAATGCCGCCTGATACGATGAGCAAGCTTTCGCGGGCAATTGACGGACGGACCGCCCTT

GGTGTTCCGCTTTCTTACATGGGCCACATGGTCTGCCATGCCATTACGCGCCTGCCGATGGGGC
AAGTTGCTTCCCTTCCGCTCCCACAAATTGCCCAGGTGCTTAGACGCGAACTGAATCAAGCTAA
CACCCCTTGGGCGATACGCAGTTATGCTACGTTCATGGCTAGAGAGCCGGACCGCTCCCGATTA
CTTTATGGTGGTACGCACAATCCGAACACAGATCTCGGTGCTACGTCCACAGTGCAGAGTAGTA
CTCTCCCGGGGAGCTGGGGACCTTTACTCGGTCCGTGCCGCTTCTTCCGCCGCCCCAATGCTGC
ACCGATCCCTGGAAGCTTCACAATACAGGAGCCTGAGGGCGGTGCTATTCCTATTGCTTTATGT
TTGCCTGAAGAGGATCTGGAAGGGCTGAAGAAGGACCCCCTGTGGAAACAATACACCCGATATG
TTGGATAG

MGSSHHHHHHSSGLVPRGSHMMSSTAGSAGNAGNDDDLYQDIMSQFPVLNAYTHLVFGFQLPSD
VDRDAVVSALQTGLDKLVEQIPWLGWQVARTGSGILKAVPWPQDVAKERVRVKYCDDLIPPMAQ
LLAAGVPINMLDGKVLTPWPALPQPHGLTGPVPVIALQANFVRGGLILNLSAHHTIMDGTGIFQ
FLKLLATVLNGKEIPAADLEQANRDRSRVVPLIPRGEPVKDYSHLRRPPGWTFAPPASPPTWCY
FKMPVSALARLVKTARDSSSQMVSENDILCAFYWQRLCAVRLARGMPPDTMSKLSRAIDGRTAL
GVPLSYMGHMVCHAITRLPMGQVASLPLPQIAQVLRRELNQANTPWAIRSYATFMAREPDRSRL
LYGGTHNPNTDLGATSTVQSSTLPGSWGPLLGPCRFFRRPNAAPIPGSFTIQEPEGGAIPIALC
LPEEDLEGLKKDPLWKQYTRYVG*

AncAT3 (Anc450) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGGGGAGTCGGCAAGACGAGACGGGTGGCGCGGGAAACGATGAAAACTTGTATCCAGATATCAT
GGGACAGTTTCCGGTCCTGAACGCCTATACCCACCTGGCTTTCGGTTTCGAGCTGCCACCCGAC
GTAGATCGTGACGCAGTGGTGTCAGCACTTCAGACAGGTCTCGACAAGTTAGTTGAGCAAATAC
CGTGGCTGGGTTGGCAGGTTGCCCGCAAATCTGGTGTACGCACCGCGGTACCTTGGCCACAGGA
CGTCGCGCGCGAACGCGTCCGGGTCAAGTATTGTGATGATACTATCCCACCCATGGCCCAATTA
TTAGCGGCGGGCGTACCGATCAATAAGTTGGATGGTAAGGTGCTTACCCCATGGCCTGCGTTAC
CGCAACCACATGGTCTGACTGGGCCAGTGCCCGTCATAGCGTTGCAAGCCAACTTTGTTCGCGG
TGGACTGATCTTAAACCTGTCCGCCCACCACACAATCATGGACGGGACTGCAAACTTTCAGTTC
TTGAAGTTGCTTGCGACGGTACTGAACGGCGACGAAATCCCCGCCGCCGACCTCGAACAGGCTA
ATCGCGACCGTTCCCGTGTTGTTCCGCTGATTCCGCGAGGAGAGCCCGTAAAGGATTACAGTCA
TTTACGTCGCCCACCAGGATGGCAGTTCGTCTTGCCGGCGAGTCCACCCACGTGGTGCTACTTC
AAGATGCCGGTATCTTCGCTTGCCAAGTTAGTGAAAACTGCGCGGGATAGCTCTTCTCAGCGCG
TGTCTGAGGATGACATTCTTTGTGCCTTCTACTGGCAACGCCTCTGTGCAGTTAGACTCGCCCG
CGGGATGCCACCAGACACTGTAAGCAAGATCTCAAGAGCTATAGACGGGCGTACAGCCCTGGGA
ATTCCCGCGTCTTATATGGGAGCTCAAGTGTGCCATGCGATTACACGGCTGCCTATGGGCCAAG
TCGCGAGTCTCACGCTTCCCCAGATAGCGCAAGTGTTACGCCGGGAGCTTTCACAAGCAAACAC
GCCGTGGGCCATCCGCTCGTTCGCAACTTTCATAGCTAGAGAACCGGATCGGAGCCGATTACTC
TATAATGGAACTCACAATGCCAACACGGACGTTGGGGCGACTTCTACCTTAAACTCGCAAACCC
TTCCTGGCTCATGGGGCCCATTGTTAGGTCCGTGCAGATTCTTTCGCCGCCCGAACGCTGCACC
CATTCCCGGATCATTTACGATACAAGAGCCAGAGGGTGGTGCTATACCAATCGCTGTGTGTTTG
CCGGAAGAGGACTTAGAGGCGCTTAAGAAGGATCCGCTGTGGCGCCAGTACACCCGTTATATAG
GGTAA

MGSSHHHHHHSSGLVPRGSHMMGSRQDETGGAGNDENLYPDIMGQFPVLNAYTHLAFGFELPPD
VDRDAVVSALQTGLDKLVEQIPWLGWQVARKSGVRTAVPWPQDVARERVRVKYCDDTIPPMAQL

LAAGVPINKLDGKVLTPWPALPQPHGLTGPVPVIALQANFVRGGLILNLSAHHTIMDGTANFQF
LKLLATVLNGDEIPAADLEQANRDRSRVVPLIPRGEPVKDYSHLRRPPGWQFVLPASPPTWCYF
KMPVSSLAKLVKTARDSSSQRVSEDDILCAFYWQRLCAVRLARGMPPDTVSKISRAIDGRTALG
IPASYMGAQVCHAITRLPMGQVASLTLPQIAQVLRRELSQANTPWAIRSFATFIAREPDRSRLL
YNGTHNANTDVGATSTLNSQTLPGSWGPLLGPCRFFRRPNAAPIPGSFTIQEPEGGAIPIAVCL
PEEDLEALKKDPLWRQYTRYIG*

AncAT4 (Anc451) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGGGTAGTCGGCAAGATGAATCAACCACTCAACAAAGCGGTGGCGCGGAGGGAGGGGACGAGAA
TTTATACCCCGATCTGATGGGCCAATTCCCAGTACTTAACGCCTATACCCACTTAGCTTTTGGT
TTCGAGCTTCCGCTTGATGTTGATCGAGACGCCGTCGTCAGTGCGCTCCAAACCGGTCTCGACC
GCCTGATAGAACAAATACCTTGGTTGGGGTGGCAAATGGGACGCAAATCAGGTGTTCGCACGGC
AGTTCCTTGGCCCGAGGATGTAGCCCGTGAACTTTTGCACGTAAAGTACTGCGACGACACAATA
CTGCCCATGGCCCAATTGCTTGCCGCAGGCGTACCAATAGGGAAGTTAGACGGCAAAGTGCTGA
CGCCCTGGCCAGCGCTGCCTCAGCCCCACGGACTTACAGGGCCTGTACCCGTGATCACATTACA
AGCGAACTTCGTGCGGGGCGGACTTATCCTTAATCTGAGTAGCCACCACACCGTGATGGATGGC
ACCGCCAACTTTCAAGTCTTGAAGCTCTTGGCCACTGTTTTAAACGGTGATGAAATCCCTGCGG
CGGACCTGCAACAAGCAAATCGTGATCGCAGCCGCCTCATCGACCTCATTCCACGCTCAGAGCC
CCTGAAGGACTACAGCCACTTGCGTCGTCCCCCTGGGTACCAATTCGTACTGCCAGCGAGCCCA
CCTACCTGGTGTTACTTTAAAATGCCTGTTACCAGCCTGAGTAAGCTTGCAAAGACTGCACGTG
ACCCGAGCCGTCCAGTGACCGAAGATGATGTTCTGCATGCGTTCTATTGGCAGCGCCTGTGCGC
CGTACGTTTAGCTCGTGGGATGCCGCCGGACACGGTTTCGAAGATCAGCCGGGCAATCGACGGT
CGAATGGCGTTGGGTATACCGGCATCTTACATGGGAGCACAGGTATACACGGCCATAACACGAC
TTCCGCTTGGGCAAGTTGCGTCTTTAACCTTGCCTCAAACAGCACAGGTGTTGCGCAGAGAGCT
GAGTCAGGCAAACACACCCTGGGCAATTCGATCTTTCGCTACTTTCATCGCTCGCGAGCCTGAC
CGCTCGCGCTTATTATACAACGGTACTCATAATGGGAACACCGATATCGGAGCTACATCCTCAC
TGAATACATCGCAGACCTTACCCCCGAGCTGGGGTCCTCTGTTGGGTCCGTGTCGCTTCTTCCG
ACGCTCAAACGGAGCGCCCATACCAGGATCCTTCACAGTCCAATCTCCGGAGGGAGGCGCAATA
CCAATTGCGGTATGCTTACCCCAAGATGACTTGGAAGCTTTAAAGAAGGATTCATTATGGCGCC
AATATACACGATGCATCGGTTGA

MGSSHHHHHHSSGLVPRGSHMMGSRQDESTTQQSGGAEGGDENLYPDLMGQFPVLNAYTHLAFG
FELPLDVDRDAVVSALQTGLDRLIEQIPWLGWQMGRKSGVRTAVPWPEDVARELLHVKYCDDTI
LPMAQLLAAGVPIGKLDGKVLTPWPALPQPHGLTGPVPVITLQANFVRGGLILNLSSHHTVMDG
TANFQVLKLLATVLNGDEIPAADLQQANRDRSRLIDLIPRSEPLKDYSHLRRPPGYQFVLPASP
PTWCYFKMPVTSLSKLAKTARDPSRPVTEDDVLHAFYWQRLCAVRLARGMPPDTVSKISRAIDG
RMALGIPASYMGAQVYTAITRLPLGQVASLTLPQTAQVLRRELSQANTPWAIRSFATFIAREPD
RSRLLYNGTHNGNTDIGATSSLNTSQTLPPSWGPLLGPCRFFRRSNGAPIPGSFTVQSPEGGAI
PIAVCLPQDDLEALKKDSLWRQYTRCIG*

AncAT5 (Anc452) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGGGAAGCAAACAGGACGATCCCACGACTCAACAGCCCGGCGCAGTAGAGGATGGGGAAGGAAA
CAACTTGAACGATCTTCCGGACCTGATGGGCCAATTTCCTGTCTTAAACGCGTACACCCACCTC

GCATTCGGTTTCGAACTGCCTTTAGACGTGGATCGCGATGCTGTGGTGAGTACCCTGCAAACGG
GCCTCGACCGTCTTATCGAGCAAATACCGTGGTTGGGTTGGCAAATGGGTCGGAAGTCAGGAGT
GCTTACAACAGTACCCTGGCCAGAGGACGTTGCACGCGTATTACTGCACGTCAAGAACTGCGAT
GACACCGTGTTACCAATGGCTCAACTGTTGGCGGCCGGAGTCCCGATTGGCAAATTAGACGGGA
AAGTGCTTACCCCGTGGCCGGCGTTGCCGCAACCGCACGGTTTAACAGGGCCGGTGCCAGTGAT
CACGCTCCAAGCAAATTTCGTCCGGGGCGGCCTCATACTGAACTTCTCATCTCATCACACTGTG
ATGGATGGCACCGCGAACTTTCAAGTGCCTAAGCTGCTTGCGGCTGTTTTAAATGGGGATGCCA
TCCCTGAGGCCGACCTGCAACAAGCTAACCGGGACCGATCCAGACTCATCGAACTGATCCCACG
TAGCGAGCCTTTAAAAGACTACTCTCACCTGCGTCGTCCGCCTGGGTACCAATTCGCATTACCT
GCATCGCCCCCTATATGGTGCAATTTTAAAATGCCGTTAGCATCCTTATCAAAGTTAACAAAGA
GCGCGCGGGATCCCTCGCGCCCCGTGACGGAAGATGATGTGCTGATTGCGTTCTTCTGGCAACG
GTTGTGCGCGGCACGGTTAGCTAGAGGGATGGCACCAGACACCGTGAGCAAGATCTCCCGTGCG
ATAGATGCAAGAATGGCATTAGGCATCCCCTCGGCCTATTTAGGTGCGCAGGTTCACATGGCCA
TTACGCGTCTGCCCCTTGGACAGGTTGCTAGCCTGACACTTCTGCAAACAGCGCAGGTTCTCCG
TCGAGAGTTAGCCCAAGCGAACACGCCATGGGCTATTAGAAGCCTCGCTACCTTCATCGCACGG
GAGCCCGATCGTTCTCGCTTGCTGTACAATGGCACACACGATGGTAATACAGACGTCGGCGCTA
CTTCCTCTCTGAACACCTCGCAAGCCCGTCCACCGCATTGGGGTCCGTTGTTAGGACCATGCCG
TTTCTTCCGCCGGTCTAACGGGGCTCCGATCCCGGGGTCTTTTACCGTATACTCTCCTGACGGT
GGAGCAATCCGCATGTGTGTCTGCCTGCCGGCAGACGATCTTGAGGCGCTGAAGAAGGACGCGC
TGTGGAAGCAATATACAAAGTGCGTGGGCTGA

MGSSHHHHHHSSGLVPRGSHMMGSKQDDPTTQQPGAVEDGEGNNLNDLPDLMGQFPVLNAYTHL
AFGFELPLDVDRDAVVSTLQTGLDRLIEQIPWLGWQMGRKSGVLTTVPWPEDVARVLLHVKNCD
DTVLPMAQLLAAGVPIGKLDGKVLTPWPALPQPHGLTGPVPVITLQANFVRGGLILNFSSHHTV
MDGTANFQVPKLLAAVLNGDAIPEADLQQANRDRSRLIELIPRSEPLKDYSHLRRPPGYQFALP
ASPPIWCNFKMPLASLSKLTKSARDPSRPVTEDDVLIAFFWQRLCAARLARGMAPDTVSKISRA
IDARMALGIPSAYLGAQVHMAITRLPLGQVASLTLLQTAQVLRRELAQANTPWAIRSLATFIAR
EPDRSRLLYNGTHDGNTDVGATSSLNTSQARPPHWGPLLGPCRFFRRSNGAPIPGSFTVYSPDG
GAIRMCVCLPADDLEALKKDALWKQYTKCVG*

AncAT6 (Anc454) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGGGATCGCGTCAAGATGAGTCTACTACACAACAATCCGGTGGGGCTGAAGGCGGTGACGAGAA
CCTCTACCCCGACCTGATGGGTCAGTTTCCGATTCTCGATGCCTATACTCATTTATGTTTCGGA
TTTGAGCTTCCGTTGGATGTTGACCGGGACGCGGTTGTAAGCGCCTTACAGACAGGCCTTGATC
GCTTAATAGAACAAATTCCGTGGCTGGGTTGGCAAATGGGGCAAAAGTCGGGCGTGCGCACAGC
GGTCCCGTGGCCGGAAGATGTTGCCCGTGACTTGTTACACGTAAAGTACTGTGACGACACGATA
CTGCCTATGGCCCAACTGTTAGCCGCAGGTGTTCCTATCGGGAAACTTGACGGAAAGGTTTTAA
CACCATGGCCGGCTCTGCCGCAACCGCATGGGCTGACTGGGCCTGTTCCCGTAATAACCCTGCA
AGCAAACTTCGTGAAAGGTGGTTTGATTCTTAATCTCAGCAGCCACCACACCATGATGGACGGA
ACAGCGAATTTCCAAGTGCTTAAACTCCTTGCGACCGTACTGTCGGGCGACGAGATTCCCGCAG
CAGATCTGCAACAGGCTAACCGTGATAGAACTCGACTTATAGATCTGATCCCGCGCTCAGAGCC
ATTGAAAGACCATTCTCACTTGCGCCGACCCCCGGGTTACCAATTCGTGTTCCCTGCAAGCCCG
CCGACGTGGTGTTATTTTAAAATGCCGGTCACATCGTTGAGTAAGCTCGCGAAGACCGCTCATG
ACCCGAGCAGACCGGTCACAGAAGATGACGTGTTACATGCCTTCTACTGGCAACGCTTATGCGC
TGTCAGACTTGCTAGAGGAATGCCAGCTGACACTGTAAGCAAAATTTCGCGGGCGATTGATGGC

CGGATGGCACTCGGTATACCTGCCAGTTACATGGGAGCGCAGGTATACACTGCTATAACACGCC
TGCCGCTCGGACAAGTCGCTTCGCTTACTCTTCCTCAAACGGCCCAAGTGCTCCGGCGCGAATT
CTCTCAAGCCAACACGGCGTGGGCAATTCGTTCATTCGCTACTTTCATTGCACGAGAGCCAGAC
CGGTCCATGTTAATGTACAACGGTACTCACAATGGCAACACGGACATTGGCGCGACTTCTTCCC
TCAACACAAATCAGACACTGCCCCCTTCCTGGGGCCCATTATTGGGTCCATGTCGCTTCTTCAG
ACGTTCAATAGGCGCGCCCATACCCGGTGCTTTTGTCGTCCAATCTGCTGAAGGCGGCGCCATT
CCAATCGCCGTATGTCTCCCACAAGACGACTTAGAAGCGTTAAAGAAGGACAGTTTGTGGCGTC
AATATACCCGCTGCATAGGATGA

MGSSHHHHHHSSGLVPRGSHMMGSRQDESTTQQSGGAEGGDENLYPDLMGQFPILDAYTHLCFG
FELPLDVDRDAVVSALQTGLDRLIEQIPWLGWQMGQKSGVRTAVPWPEDVARDLLHVKYCDDTI
LPMAQLLAAGVPIGKLDGKVLTPWPALPQPHGLTGPVPVITLQANFVKGGLILNLSSHHTMMDG
TANFQVLKLLATVLSGDEIPAADLQQANRDRTRLIDLIPRSEPLKDHSHLRRPPGYQFVFPASP
PTWCYFKMPVTSLSKLAKTAHDPSRPVTEDDVLHAFYWQRLCAVRLARGMPADTVSKISRAIDG
RMALGIPASYMGAQVYTAITRLPLGQVASLTLPQTAQVLRREFSQANTAWAIRSFATFIAREPD
RSMLMYNGTHNGNTDIGATSSLNTNQTLPPSWGPLLGPCRFFRRSIGAPIPGAFVVQSAEGGAI
PIAVCLPQDDLEALKKDSLWRQYTRCIG*

AncAT7 (Anc456) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGGGTTCACAGGACGGCGAAGCTCCAGAGTATGCTGAAGCGGACATTATGGGTCAATTCCCTGT
CCTTAACGCGTATACACAATTAGCCTTCGGCTTCGAGTTACCACCTGATGTGGACAGAGACGCC
GTCGTGTCCGCCCTTCAAACTAGTTTTGACAAGTTAGTTGAGCAAATTCCCTGGTTAGGGTGGC
AGGTGGCGACGAGCGAGTCAGGAGTACGCACAGTCGTTCCATGGCCCCAAGATGTTGCGAAGGA
GCGTGTCCGCGTCAAAATCTGTGATGATTCCATCCCGCCTATGGAACAGCTGCTCGCGGAGAAA
GTCCCGATCAATAAGCTGGACGGAAAGGTGCTGTGCCCATGGCCTGCACTCCCCCAGCCCCATG
GTTTAACTGGACCGGTTCCAGTTGTTGCATTGCAGGCAAACTTCGTGCGCGGCGGTTTAATCTT
GAATCTGTCTGCCCACCACACTATCATGGATGGAACCGCAAATTTTCAATTCCTTAAGCTTTTC
GCAACGGTTCTTAACGGTGGTGAAATCCCGGCGGCTGACCTGGAGCAAGCCAACCGCGATCGTA
ATCGCATTGTGCCTCTTATTCCTCACGGCGAGCCAGTAAAGGACTACAGCCACTTGCGTCCGCC
CCCGGGGTGGAAATTCGTAATGCCAACCTCTTGGCCAACATGGTGCTATTTCCTGATGCCCGTG
GACTCCCTGGCAGAGTTGGTTAAGACAGCTCGGGACGACGACGGTGAGTCCGACTCTAGTCAAC
GGATAAGCTCTGATGACATTCTCTGTGCCTTCTACTGGAAGCGCTTGTGTGCAATACGCTTGGC
CCGTGGAATGCCACCAGATACTGAAAGCAAGATCAGTCGAGCGATAAATGGCCGCACGCCTCTT
GGAATCCCGTCTTCTTACATGGGTGCTCAAGTCTGCCCCGCAATTACCCGTCTGCCCATGGGTC
GGGTCGATGAGTTAACCTTGCCGCAACTCGCACGCATTCTGCGGCGCGAATTAGTTGAAGCCGC
CACCCCCTGGGCTATACGCTCATTTGCGACGTTCATAGCGCGAGAACCGCCAGAGGATCGAGCG
CGACTGTTGTATACAGGTACACACAACAGTAACACTGATGTAGGGGCCACTAATGTAAGCCGGC
TTGCGACACCAAAGGGACCTTGGGGTCCACTGCTGGGTCCCTGCCGCTTCTATCGTCGTCCGAA
CGCCGCCCCGATCCCTGGAAGCTTTCGCATTCAAGAGCCAGAGAACGGAGCCCATCCTATAGCA
GTTTGTCTGCCTGAGGAAGACCTTGAGGCCCTTAAGAAGGATGAACTCTGGGGCCGTTATGCAA
CTTACATCGGCTGA

MGSSHHHHHHSSGLVPRGSHMMGSQDGEAPEYAEADIMGQFPVLNAYTQLAFGFELPPDVDRDA
VVSALQTSFDKLVEQIPWLGWQVATSESGVRTVVPWPQDVAKERVRVKICDDSIPPMEQLLAEK

VPINKLDGKVLCPWPALPQPHGLTGPVPVVALQANFVRGGLILNLSAHHTIMDGTANFQFLKLF
ATVLNGGEIPAADLEQANRDRNRIVPLIPHGEPVKDYSHLRPPPGWKFVMPTSWPTWCYFLMPV
DSLAELVKTARDDDGESDSSQRISSDDILCAFYWKRLCAIRLARGMPPDTESKISRAINGRTPL
GIPSSYMGAQVCPAITRLPMGRVDELTLPQLARILRRELVEAATPWAIRSFATFIAREPPEDRA
RLLYTGTHNSNTDVGATNVSRLATPKGPWGPLLGPCRFYRRPNAAPIPGSFRIQEPENGAHPIA
VCLPEEDLEALKKDELWGRYATYIG*

3.5.2 Protein production and purification

Proteins were heterologously expressed in *E. coli* BL21(DE3) cells. HisPur nickel-nitrilotriacetic acid resin (Ni-NTA resin) was purchased from Thermo Fisher Scientific. Proteins were concentrated at 3,500 x g at 4 °C using 30 kDa MWCO Amicon centrifuge filters purchased from EMD Millipore. Protein samples were analyzed by mini SDS-PAGE gels and visualized with Protein Ark Quick Coomassie Stain (Anatrace) and imaged with the Azure Gel Imaging System. The total protein concentration was quantified using a NanoDrop from Thermo Fisher Scientific, with absorbance measured at 280 nm (A280). All purification steps were performed at 4 °C.

3.5.3 96-well plate protein production for small-scale reactions using cell lysates

Chemically competent *E. coli* strain BL21(DE3) cells were transformed with a pET-28a(+) plasmid encoding acyltransferases using standard protocols and were grown on LB agar plates containing 50 µg/mL kanamycin overnight at 37 °C. Single colonies were used to inoculate 400 µL of LB media with 50 µg/mL kanamycin in each well in a 96-well plate and were grown overnight at 37 °C, 350 rpm. The overnight cultures (10 µL each) were used to inoculate 400 µL of TB media with 50 µg/mL kanamycin. The cultures in the 96-well plate were incubated at 37 °C, 350 rpm until their optical density at 600 nm (OD600) reached ~0.6 and were cooled to 16 °C and equilibrated for 1 h. Isopropyl-β-D-thiogalactopyranoside (IPTG) was added to 0.5 mM to induce overexpression. The cultures were incubated at 16 °C, 350 rpm for 18 h and harvested by centrifugation at 2,000 rpm for 10 min. Harvested cells were stored at −80 °C in the plate before lysis and use.

3.5.4 Large-scale protein production

Chemically competent *E. coli* strain BL21(DE3) cells were transformed with a pET-28a(+) plasmid encoding acyltransferases using standard protocols and grown on LB agar plates

containing 50 µg/mL kanamycin overnight at 37 °C. A single colony was used to inoculate 10 mL of LB media with 50 µg/mL kanamycin. The culture was incubated overnight at 37 °C, 200 rpm. The overnight culture was used to prepare a glycerol stock and inoculate 1 L of TB media containing 4% glycerol (v/v) and 50 µg/mL kanamycin in a 2.8-L flask. The 1-L culture was incubated at 37 °C, 200 rpm until reaching OD600 of ~0.6 and was cooled to 16 °C and equilibrated for 1 h. IPTG was added to 0.5 mM to induce overexpression. The culture was incubated at 16 °C, 200 rpm for 18 h and harvested by centrifugation at 13,881 x g for 30 min. Harvested cells were stored at −80 °C before lysis and purification.

3.5.5 General protein purification procedures

Harvested cell pellets from overexpression were resuspended in 30 mL of lysis buffer (50 mM Tricine pH 7.9, 500 mM NaCl, 5 mM imidazole) containing 1 mg/mL lysozyme, 100 units of DNase I, 2 mM dithiothreitol (DTT), and 1 mM phenylmethylsulfonyl fluoride (PMSF), incubated on a rocker at 4 °C for 45 min, and lysed by sonication for 100 s total in cycles of 5 seconds on and 50 seconds off. Insoluble material was removed by centrifugation (46,413 x g for 30 min). The clarified lysate was combined with equilibrated Ni-NTA resin (3 mL bed volume) and incubated on a rocker at 4 °C for 1 h. The resin was collected in a gravity-flow column and washed with 50 mL of lysis buffer containing 5 mM imidazole. Enriched His-tagged proteins were eluted from the resin with up to 10 mL lysis buffer containing 250 mM imidazole. The eluted proteins were concentrated and exchanged into storage buffer (50 mM Tricine pH 8.3, 100 mM NaCl, 10% (v/v) glycerol) using a Cytiva PD-10 desalting column, flash frozen with liquid nitrogen, and stored at −80 °C for future use. For the sample preparation for SDS-PAGE, the protein stock solutions were diluted 40 times by the storage buffer before they were mixed with 2x Laemmli sample buffer.

3.5.6 Protein thermal stability determined by differential scanning fluorimetry (DSF)

Protein samples were diluted to 4,000 µg/mL using KPi buffer (50 mM potassium phosphate pH 8.0, 100 mM NaCl, 10% (v/v) glycerol). SYPRO™ orange protein gel stain (5,000x concentrate in DMSO) was diluted to 500x in KPi buffer. 0.8 µL 500x dye was mixed with 79.2 µL KPi buffer as the control group, and 77.2 µL KPi buffer and 2 µL 4,000

μg/mL protein samples as the test group. The final protein concentration was targeted at 100 μg/mL. 20 μL of the mixtures was aliquoted to PCR tubes in a 96-well rack. Triplicate samples were analyzed by fluorescence in a StepOnePlus™ Real-Time PCR instrument operated with standard melt curve mode with ROX as a reporter and no quencher. The heating program followed a 2 min hold at 25 °C and a stepwise heating with a 1 °C/min ramp from 25–95 °C. The melting temperature (Tm) was calculated as the temperature corresponding to the maximum of the first derivative of the normalized fluorescence signal (Figure S3.2) using the DSFworld software available at https://gestwickilab.shinyapps.io/dsfworld/ (Table S1).

3.5.7 Biosynthetic gene cluster (BGC) visualization

The known biosynthetic gene clusters of azanigerone in *Aspergillus niger* (BGC0001143), rubropunctatin/monascorubrin in *Monascus spp.* (BGC0000027), chaetoviridin A in *Chaetomium globosum* (BGC0001405), and asperfuranone in *Aspergillus nidulans* (BGC0000022) were searched on the Minimum Information about a Biosynthetic Gene cluster (MIBiG). The GeneBank files of those clusters were downloaded and visualized by the clinker tool of the online CompArative GEne Cluster Analysis Toolbox (CAGECAT).[60]

3.5.8 Sequence similarity network (SSN) analysis

The sequence similarity network (SSN) for flavin-dependent monooxygenases (FDMOs) was created using AfoD (UniProt ID: Q5BEJ7) as a query to search the UniRef90 database via BLAST on the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST)[26-27] website. 500 homologous sequences were obtained using a default e-value of 5. The "Fragment Option" was activated to exclude UniProt-defined fragments from the results. The SSN was refined by setting the alignment score to 150 and filtering out sequences that fell outside the range of 300–600 aa, resulting in 463 sequences (Figure S3.3).

3.5.9 Genome neighborhood analysis and calculation of adjusted mutual information

The SSN in an xgmml file was used as input for genome neighborhood analysis on the Enzyme Function Initiative-Genome Neighborhood Tool (EFI-GNT)[27] website. For this calculation, a neighborhood size of 10 was utilized with a minimal co-occurrence percentage lower limit set to 20%. The analysis generated an output file, ALL_PFAM.txt, containing information about the neighbors for each query ID from the input SSN, the protein families (Pfams) they belonged to, and their distances from the query ID. Additionally, the output folder contained another file, pfam_neighbors_PF02458.txt, which specifically detailed the query ID that co-occurred with proteins from PF02458 (Transferase family). Both files were placed in the same folder titled “Neighbor Pfam domains at 0% minimal co-occurrence frequency,” which is located under the “Other Files” tab of the EFI-GNT output. The pfam_neighbors_PF02458.txt file was read using a Jupyter Notebook script (support.ipynb) and the data was exported to a csv file. 190 IDs of the PF01494 (FAD_binding_3) and PF02458 sequences that co-occurred with each other were saved separately in two txt files. These were then used as inputs for the second round of SSN analysis. Two SSNs were generated using the accession IDs acquired from the genome neighborhood analysis. Both were finalized using an alignment score of 150. Cluster analyses of these two SSNs were conducted on the EFI-EST website. The SSN of PF01494 contained was divided into 20 clusters, while that of PF02458 was divided into 21 clusters. All singletons were excluded from the subsequent calculation of the adjusted mutual information (AMI)[61] score. All co-occurred FDMO-AT pairs included in the following clustering comparison were assumed to be in the same BGCs for the synthesis of single azaphilone natural products.

The AMI score between the clusterings of PF01494 and PF02458 was calculated using the Jupyter Notebook script (*support.ipynb*). In the script, we utilized the adjusted_mutual_info_score function from the sklearn.metrics module (version 1.1.2). The AMI is an information-theoretic measure that quantifies the agreement between two clusterings, corrected for chance. Specifically, the AMI score returns a value of 1 when two clusterings are identical, and close to 0 when the mutual information between the two clusterings equals its expected value (random clustering). To be specific, the enzymes in two protein families were clustered by SSN separately and the sequences were labeled starting from the biggest cluster to the smallest cluster with at least two members in an

ascending order. The PF01494 sequences were sorted by the clustering label in an ascending order and the paired PF02458 sequences were found and ordered in the same way. For an example with five FDMO-AT pairs (A-A', B-B' C-C', D-D', and E-E'), the FDMOs and ATs are both separately clustered into two clusters. The FDMO sequences are ordered and labeled as (A:1, B:1, C:1, D:2, E:2) and the paired AT sequences are ordered accordingly as (A':1, B':1, C':2, D':2, E':2) with the cluster labels from the clustering of ATs. Then, the AMI score is computed for two lists from the clusterings of FDMO and AT, [1, 1, 1, 2, 2] and [1, 1, 2, 2, 2].

3.5.10 BLAST search of AfoD to find close homologs paired with an acyltransferase

BLASTP was conducted on the NCBI website using default parameters and AfoD (UniProt ID: Q5BEJ7) as the query. We downloaded the local genomic sequences (approximately 60k nt) centered around the AfoD homologs of the top 20 hits from the NCBI website. These sequences were then submitted to 2ndFind for genome neighborhood analysis at https://biosyn.nih.go.jp/2ndfind/. We searched for nearby acyltransferases in PF02458 to determine if they co-occurred with the AfoD homologs. Of the top 20 hits, only four AfoD homologs with sequence identities ranging from 55–60% were found to have an associated acyltransferase in PF02458 (Table S3.3). We then focused on the CazL-CazE (PF01494-PF02458) pair (Table S3.4) for subsequent studies, as this pair has been previously characterized.

3.5.11 Tree reconstruction

A set of 500 sequences was obtained by performing three iterations of PSI- BLAST[41] on the NCBI website by initially using CazE (UniProt ID: Q2GWL7) as the query. Sequences were clustered and filtered using CD-HIT[62] resulting in a set of 287 sequences where no sequence is related to another by over 95% sequence identity. The sequences were aligned using Muscle5[63] and the sequences with long indels were removed from the alignment. The process of multiple sequence alignment (MSA) and sequence filtering was iterated for six times until no sequence with long indels was seen. The alignment was visualized in AliView[64] to make minor edits to the MSA. The final alignment with 240 sequences in a fasta file was trimmed using the "Automated1" method in trimAl[65] and was converted to a phylip file for the subsequent selection of the best maximum likelihood

model. The evolutionary model selection was done by raxmlGUI 2.0[66] and the LG[67]+I+G4+F was chosen as the best model based on the Akaike Information Criterion (AIC)[68] and Bayesian Information Criterion (BIC).[69] The phylogenetic tree was reconstructed by PhyML 3.1[70] using ten random starting trees and was optimized using the subtree pruning and regrafting (SPR) algorithm. The best tree with the maximum likelihood was chosen and used for the subsequent ancestral sequence reconstruction. Branch supports were estimated using bootstrap with 100 replicates and the bootstrap values of the key branches were shown (Figure 3.6). The tree is unrooted without any outgroup sequences assigned.

3.5.12 Reconstruction of ancestral states

The ancestral acyltransferase sequences were reconstructed with the complete MSA and the best ML phylogenetic tree as input to PAML version 4.9f.[71] Within indels regions, the maximum parsimony Fitch's algorithm was used to determine if a gap should be placed at a specific position. The "AltAll" procedure[72] was used to infer an alternative ancestral sequence. This procedure replaces the most likely residue with the second most likely residue when the posterior probability difference between the two possibilities is 0.2 or less.

3.5.13 Active-site residue comparison

All descendants of AncAT4 were aligned, and the active site residues were identified and sorted by their distances from His166 in the AncAT4 AlphaFold2 model (Table S3.5). The analysis was done by the Jupyter Notebook script *support.ipynb*.

3.5.14 Reaction sequence using purified AncFDMO and clarified acyltransferase cell lysates

The frozen acyltransferase (AT) cell pellets, stored separately in a 96-well plate, were taken out from a −80 °C freezer, completely thawed in a room temperature water bath, and then frozen again by being placed into a liquid nitrogen bath. This freeze-thaw cycle was repeated five times. The cell pellets in each well were resuspended in 250 µL of lysis buffer (50 mM Tricine pH 7.9, 500 mM NaCl) containing 1 mg/mL lysozyme, 100 units of DNase I, and 1 mM PMSF. They were then incubated at 25 °C for 2 h with shaking at 350

rpm. The cell suspensions were centrifuged at 2,000 rpm and 4 °C for 20 min before use. During cell lysis, the oxidative dearomatization reaction of substrate **3.13** was performed at 30 °C for approximately 1 h until the full conversion of **3.13** to the product (*S*)-**3.12** using the purified ancestral FDMO, Anc321_L96F. The reaction contained 10 μM enzyme, 2.5 mM substrate **3.13**, 1.0 mM $NADP^+$, 2.0 U/mL glucose-6-phosphate dehydrogenase (G6PDH), 5.0 mM glucose-6-phosphate (G6P), and 50 mM potassium phosphate (KPi) buffer, pH 8.0 (Figure S3.4). After the first reaction, 50 μL of the reaction master mix was distributed to wells in a 96-well reaction plate and mixed with 50 μL of the clarified AT cell lysate and 1.1 equiv of thioester **3.14** (Figure S3.4). The mixture was then incubated at 30 °C for another 1 h (Figure S3.4). Given the red color of the final product (*S*)-**3.15**, the reactions were examined visually in this quick AT screening.

3.5.15 Reaction sequence using purified FDMO and AT

The enzymatic reaction sequence began with oxidative dearomatization catalyzed by purified FDMO (AncFDMO or AzaH). The first reaction was set up by including 10 μM FDMO, 2.5 mM substrate **3.13**, **S3.1**, **S3.2**, or **S3.3**, 1.0 mM $NADP^+$, 2.0 U/mL G6PDH, 5.0 mM G6P, and 50 mM KPi buffer, pH 8.0, in a 50 μL reaction mixture (Figure S3.5). The reaction proceeded at 30 °C for approximately 1 h or until complete conversion of the substrate. The second acylation reaction was initiated by adding 1.1 equiv of C5-thioester **3.14** and 10 μL of AT to achieve a final AT concentration of 2 mol % (Figure S3.5). The enzymatic reaction was then quenched after incubation at 30 °C for 1 h by adding 180 μL of methanol containing 3.5 mM pentamethylbenzene as an internal standard (ITS). Calibration curves for dearomatized substrates, (*S*)-**3.12**, (*R*)-**3.12**, (*S*)-**3.16**, (*S*)-**3.17**, and (*S*)-**3.18**, were prepared at 10, 20, 40, 60, 80, and 100% concentration by mixing the reaction mixture from the first step, the storage buffer, and dimethyl sulfoxide (DMSO) in the same ratio as the samples. Each sample, including those for the calibration curves, was prepared in duplicate.

Analysis was conducted on UPLC-DAD-CD (Method A for (*S*)-**3.12**) or UPLC-DAD-MS (Method B for (*R*)-**3.12** and C for (*S*)-**3.16**–**3.18**), and percentage conversions were quantified based on the consumption of the dearomatized substrates, as calculated by their calibration curves covering the range from 0–100% (Figure S3.6). The ratio of the

substrate peak area at 273 or 300 nm to the internal standard peak area at 273 nm was used for the y-axis, while the percentage substrate concentration served as the x-axis. The linear tricyclic products, (*S*)-**3.15**, (*R*)-**3.15**, (*S*)-**S3.5**, (*S*)-**S3.6**, and (*S*)-**S3.7** were detected at 440 nm.

Method A. Phenomenex Kinetex® 1.7 µm C18 100 Å, 100 x 2.1 mm column; phase A = deionized water with 0.1% formic acid, and phase B = acetonitrile with 0.1% formic acid; method = 20% B held for 1.1 min, to 100% B over 2.9 min, 100% B held for 0.5 min, back to 20% B over 0.1 min, 20% B held for 0.4 min for equilibration, 273, 300, and 440 nm UV detection, 230 nm CD detection, and 0.5 mL/min flow rate. Injection volume = 5 µL.

Method B. Phenomenex Kinetex® 1.7 µm C18 100 Å, 100 x 2.1 mm column; phase A = deionized water with 0.1% formic acid, and phase B = 95/5 acetonitrile/water with 0.1% formic acid; method = 20% B held for 1.1 min, to 100% B over 2.9 min, 100% B held for 1.0 min, back to 20% B over 0.1 min, 20% B held for 0.4 min for equilibration, 273, 300, and 440 nm UV detection, single quadrupole MS detection (positive and negative modes), and 0.5 mL/min flow rate. Injection volume = 2 µL.

Method C. Phenomenex Kinetex® 1.7 µm C18 100 Å, 100 x 2.1 mm column; phase A = deionized water with 0.1% formic acid, and phase B = 95/5 acetonitrile/water with 0.1% formic acid; method = 5% B held for 1.1 min, to 100% B over 2.9 min, 100% B held for 1.0 min, back to 5% B over 0.1 min, 5% B held for 0.4 min for equilibration, and another 0.5 min post-run equilibration, 273, 300, and 440 nm UV detection, single quadrupole MS detection (positive and negative modes), and 0.5 mL/min flow rate. Injection volume = 2 µL.

3.5.16 Screening of thioesters using purified AncFDMO and AncAT4

The first oxidative dearomatization reaction was performed at 30 °C following the abovementioned procedures. To the reaction mixture from the first step, AncAT4 was added to achieve a final AT concentration of 2 mol %. The thioesters **3.6** with different chain lengths (C1, C3, C5–7, C9) were tested and the reactivity was determined based

on the consumption of (*S*)-**3.12**, as calculated by its calibration curve (Figure S3.7). The analysis was conducted on UPLC-DAD-CD using Method A. The product can exist in close ((S)-**S3.16**) and open ((S)-**S3.17**) forms in solution (Figure S3.7).

3.5.17 General procedures for preparative scale in vitro reaction (8 mL scale)

The preparative-scale in-vitro reaction was done by Ye Wang. 2.5 mM orcinaldehyde substrate **3.12** (400 µL of a 50 mM stock solution in DMSO), 5 mM G6P (80 µL, 500 mM), 1 mM $NADP^+$ (80 µL, 100 mM), 1 U/mL G6PDH (80 µL, 100 U/mL), and 10 µM AncFDMO (103 µL of a 776 µM stock solution) was added to 7.26 mL of 50 mM potassium phosphate buffer, pH 8.0. The reaction was placed in an incubator at 30 °C with 100 rpm shaking. After 1 h, 2.75 mM thioester **3.14** (220 µL of a 100 mM stock solution in DMSO), and 33 µM AncAT4 (800 µL of a 373 µM stock solution) were then added, and the reaction was incubated at 30 °C for 25 min with 100 rpm shaking. The reaction was then diluted with brine (~3x reaction volume) and extracted 3 times with ~10 mL EtOAc (centrifuge if necessary). The combined organic layers were washed with ~30 mL of brine, then dried with sodium sulfate. The sodium sulfate and precipitated biomolecules were removed through gravity filtration and the flow-through was passed through a 0.22 µm filter before the solvent was removed using a rotary evaporator. The resulting crude oil was dissolved in 1 mL of HPLC grade acetonitrile and subjected to purification by preparative HPLC using a Phenomenex Kinetex® 5 µm C18 100 Å, 150 x 21.2 mm column under the following conditions: mobile phase A = deionized water and B = acetonitrile; method = 5% B hold 1 min, 5% to 20% B over 2 min, 20% to 50% B over 5 min, 50% to 90% B over 9 min, 90% to 100% B over 5 min, 100% B for 3 min; flow rate, 10 mL/min. Fractions containing the purified compound were pooled and the solvent was removed using a rotary evaporator.

(*S*)-3-hexanoyl-9a-methyl-6-propyl-2H-furo[3,2-g]isochromene-2,9(9aH)-dione (3.15). Following preparation by the general method described above, the title compound was isolated as an orange, amorphous solid. **^{1}H NMR** (600 MHz, $CDCl_3$) δ 7.85 (s, 1H), 6.82 (s, 1H), 6.15 (s, 1H), 2.99 – 2.89 (m, 2H), 2.41 (t, *J* = 7.5 Hz, 2H), 1.71 (s, 3H), 1.69

– 1.61 (m, 4H), 1.33 – 1.31 (m, 4H), 1.00 (t, *J* = 7.4 Hz, 3H), 0.88 (t, *J* = 6.8 Hz, 3H). $[\alpha]_D$ = +637 (*c* 0.001, MeOH).

3.5.18 Circular dichroism analysis of the model linear tricyclic azaphilone product

Both stereocomplementary purified azaphilone products, (*R*)-**3.15** and (*S*)-**3.15**, were dissolved in methanol and analyzed on the UPLC-DAD-CD using the conditions specified in Method D. The retention times of (*R*)-**3.15** and (*S*)-**3.15** were almost identical, and the normalized CD signals ($Height_{CD440}/Area_{UV440}$) for (*R*)-**3.15** and (*S*)-**3.15** were opposite in sign but similar in magnitude (−0.0756 vs. +0.0742, Figure S3.26), demonstrating the stereocomplementary feature of (*R*)-**3.15** and (*S*)-**3.15**.

Method D. Phenomenex Kinetex® 1.7 µm C18 100 Å, 100 x 2.1 mm column; phase A = deionized water with 0.1% formic acid, and phase B = acetonitrile with 0.1% formic acid; method = 60% B held for 5 min, 440 nm UV detection, 440 nm CD detection, and 0.5 mL/min flow rate. Injection volume = 5 µL.

3.5.19 Construction of models with phosphopantetheine 5C-thioester

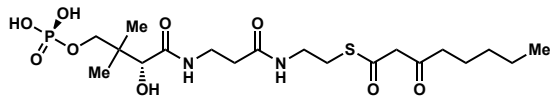
Models of AncAT4–6 and CazE were constructed using AlphaFold2 (version 2.3.0)[48] with the default pipeline and MSAs generated from the full database. The process included inference using the monomer model preset and relaxation of the best model. The top AlphaFold2 models were aligned to 2E1T,[49] a homologous malonyl transferase identified with BLAST,[41] using TM-align.[73] Model files were prepared for CHARMM[74] using convpdb.pl from the MMTSB[75] toolset. The aligned structures were then minimized in vacuum using the pyCHARMM[50] package, employing 1000 steps of steepest descents (SD) in the CHARMM36[76] force field. The histidine tautomer (HSD) was chosen to position the side chain nitrogen for interaction with the protein's open cavity.

The structure of malonyl-CoA in 2E1T guided the creation of a phosphopantetheine C5-thioester model in Avogadro[77]. Coordinates analogous to malonyl-CoA were transposed to the phosphopantetheine C5-thioester **3.19**, fixed, and integrated into the remaining structure. Newly added atoms were minimized using the GAFF[78] force field. Parameters for **3.19** and (*S*)-**3.12** were computed with the CHARMM General Force Field (CGenFF)[79]. **3.19** was incorporated into the vacuum-minimized

structures and minimized using SD and Adopted Basis Newton-Raphson (ABNR). During the first minimization round, only the backbone was restrained, followed by restraining the backbone and side chains outside a 5 Å radius from the thioester in the second round, and restraining all non-hydrogen atoms of the protein in the third round.

3.5.20 Rigid receptor and flexible receptor docking

Docking grid centers were set at the open cavity adjacent to histidine 166 and **3.19** in the aligned structures. Docking was performed using the cdocker module within pyCHARMM. For rigid CDOCKER,[80] 10 conformers of (*S*)-**3.12** were generated using obrotamer from OpenBabel.[81] Each conformer was placed 100 times within a 20 Å grid, using default softcore potentials for grid generation. Binding site residues were identified for each protein if a heavy atom was within 4.5 Å of any of the top 10 clustered poses. Flexible CDOCKER,[51-52] with a 40 Å grid, utilized 25 conformers and 20 placements each. Residues identified by Rigid CDOCKER informed the selection of flexible residues. The same grid center was used for both rigid and flexible docking. A modified Flexible CDOCKER script enabled flexibility in **3.19**, fixing all atoms from the phosphate to the first amine, similar to how backbones of flexible residues are fixed. Parameters from Flexible CDOCKER with only protein side chains were applied.



phosphopantetheine C5-thioester (**3.19**)

Figure 3.11. C5-Thioester **3.19** substrate for docking and molecular dynamics simulation.

3.5.21 Molecular dynamics

The initial structures for molecular dynamics (MD) simulations were based on the top clustered pose of (*S*)-**3.12** in AncAT4 identified by Flexible CDOCKER with a flexible **3.19**, which was in a putative reactive conformation. This pose was minimized in a similar manner to **3.19**, including three rounds of minimization, with the final round applying restraints to the thioester. The first 20 N-terminal residues were removed to decrease the size of the water box. The protein-substrate complexes were solvated in a cubic TIP3 water model with a 10 Å padding, neutralized with chloride ions, and the water molecules were minimized over 500 steps of SD. The MD simulations were carried out using the

OpenMM[82] CHARMM interface with a 2 fs timestep, particle mesh Ewald for long-range electrostatics, at 298 K. After a 0.5 ns equilibration period, a 30 ns production run was performed in an NPT ensemble. Trajectories were analyzed using the MDAnalysis[83] package.

3.6 Supporting Information

Table S3.1. Screened acyltransferases which co-localize with a flavin-dependent monooxygenase.

No.	Acyltransferase UniProt ID	Flavin-dependent monooxygenase UniProt ID	Percent identity to CazE
1	A0A094B1C4	A0A094B1B9	43.2%
2	A0A094J5P9	A0A094GTA2	32.6%
3	A0A084B9Z3	A0A084B9Z5	37.6%
4	A0A135LJ73	A0A135LJ57	51.1%
5	A0A194VYW6	A0A194W0Y1	44.7%

3.6.1 Protein purification and thermostability analysis

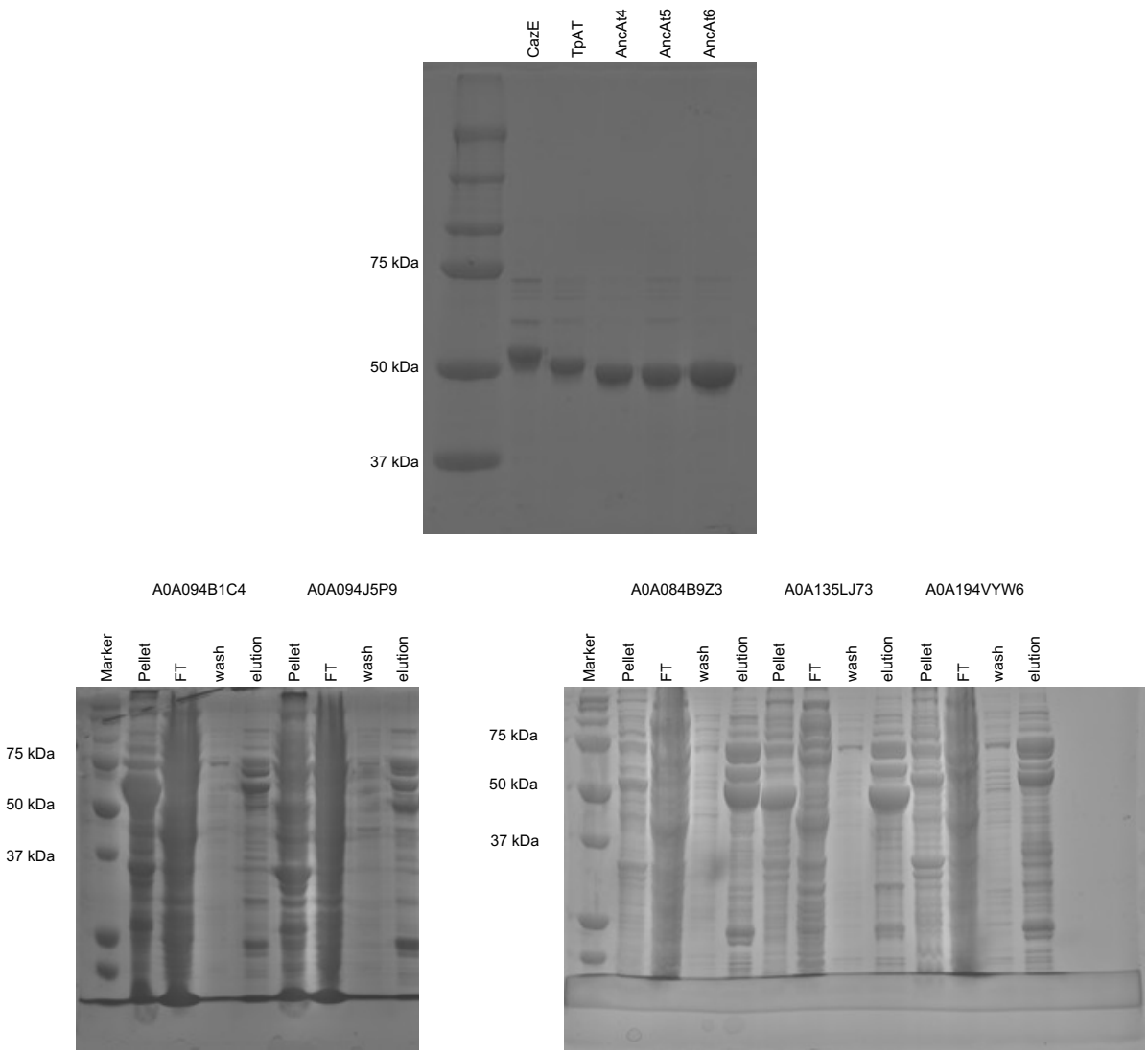


Figure S3.1. SDS-PAGE analysis of the different fractions of acyltransferases.
Proteins: CazE (55.51 kDa), TpAT (52.57 kDa), AncAT4 (52.14 kDa), AncAT5 (52.22 kDa), and AncAT6 (52.08 kDa), A0A094B1C4 (57.48 kDa), A0A094J5P9 (54.55 kDa), A0A084B9Z3 (55.43 kDa), A0A135LJ73 (55.49 kDa), and A0A194VYW6 (57.16 kDa). Fractions collected during the His-tag affinity purification: pellet, flow-through cell lysate (FT), wash, and elution.

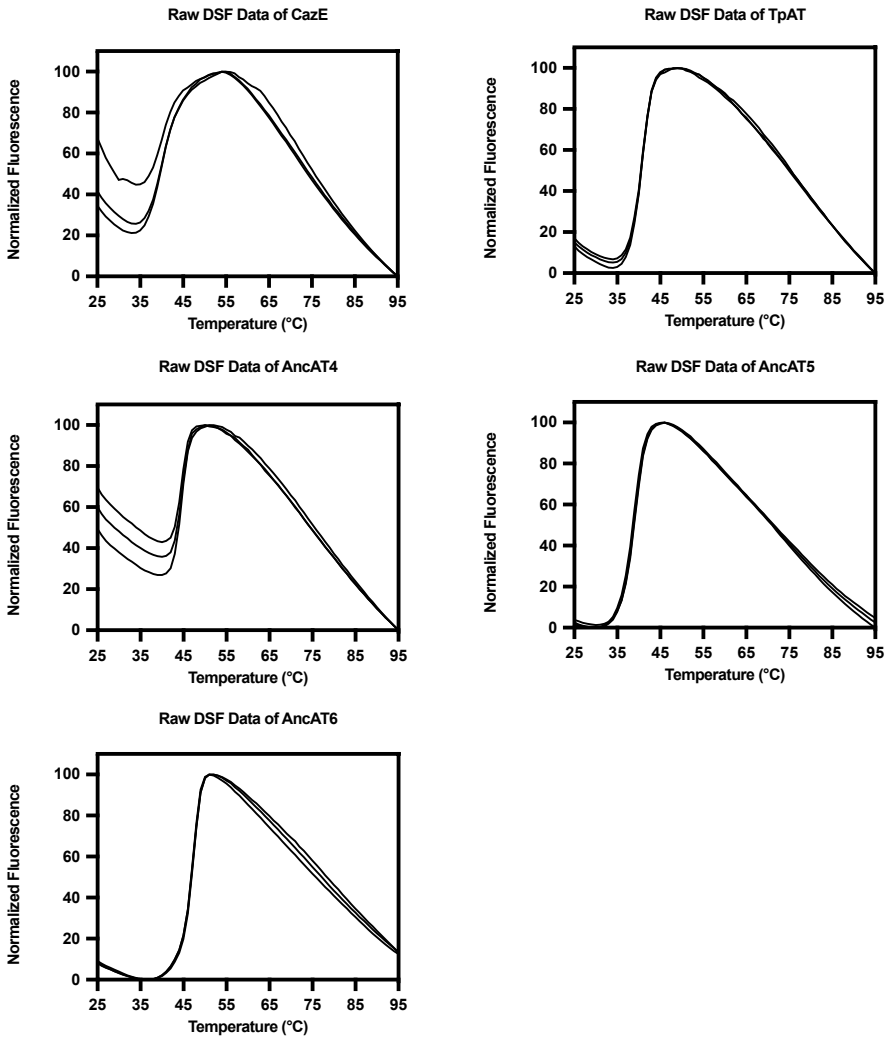


Figure S3.2. Normalized fluorescence data of acyltransferases from differential scanning fluorimetry. The concentration of the protein samples was 100 µg/mL and the SYPRO™ orange protein gel stain (5x) was used in the scanning from 25–95 °C (1 °C/min ramp).

Table S3.2. Protein thermal stability measured by differential scanning fluorimetry.

Enzymes	T_m (°C)	Stdev
CazE	40.1	0.1
TpAT	40.6	0.1
AncAT4	44.6	0.1
AncAT5	38.9	0.2
AncAT6	46.9	0.1

3.6.2 Bioinformatic analysis

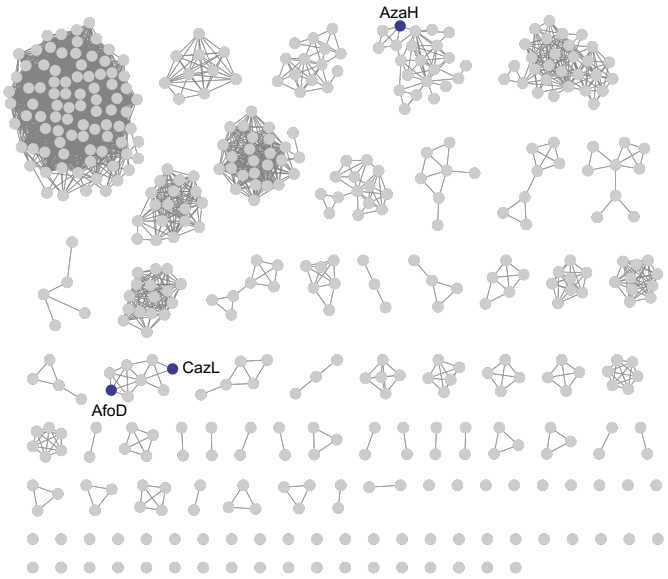


Figure S3.3. Sequence similarity network (SSN) for 463 flavin-dependent monooxygenases (FDMOs). The SSN was generated by the BLAST search using AfoD as the query. Characterized FDMOs including AzaH, AfoD, and CazL were highlighted in blue.

Table S3.3. Top 20 results of BLASTP using AfoD (UniProt ID: Q5BEJ7) as the query.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession	Co-occurred PF02458
protein afoD [Aspergillus nidulans FGSC A4]	Aspergillus nidulans FGSC A4	909	909	100%	0	100	440	XP_658637.1	FALSE
FAD-dependent monooxygenase afoD [Aspergillus mulundensis]	Aspergillus mulundensis	704	704	97%	0	81.78	442	XP_026602994.1	FALSE
hypothetical protein N7540_009298 [Penicillium herquei]	Penicillium herquei	625	625	97%	0	70.3	445	KAJ6014707.1	FALSE
uncharacterized protein N7483_005052 [Penicillium malachiteum]	Penicillium malachiteum	624	624	97%	0	67.04	466	XP_056948318.1	FALSE
hypothetical protein B0O99DRAFT_742098 [Bisporella sp. PMI_857]	Bisporella sp. PMI_857	624	624	97%	0	70.53	445	KAH8594184.1	FALSE
conserved hypothetical protein [Aspergillus terreus NIH2624]	Aspergillus terreus NIH2624	623	623	97%	0	70.14	435	XP_001216283.1	FALSE
hypothetical protein N7527_000383 [Penicillium freii]	Penicillium freii	623	623	97%	0	66.01	470	KAJ5534129.1	FALSE
uncharacterized protein N7508_004580 [Penicillium antarcticum]	Penicillium antarcticum	621	621	97%	0	65.13	468	XP_058321271.1	FALSE
hypothetical protein N7501_003441 [Penicillium viridicatum]	Penicillium viridicatum	620	620	97%	0	65.79	470	KAJ5967193.1	FALSE
hypothetical protein ACN42_g1501 [Penicillium freii]	Penicillium freii	619	619	96%	0	66.3	488	KUM65570.1	FALSE
hypothetical protein N7493_002478 [Penicillium malachiteum]	Penicillium malachiteum	611	611	96%	0	65.91	460	KAJ5733692.1	FALSE
hypothetical protein ATETN484_0011028500 [Aspergillus terreus]	Aspergillus terreus	607	607	97%	0	70.37	435	GES64796.1	FALSE
hypothetical protein HFD88_004906 [Aspergillus terreus]	Aspergillus terreus	605	605	97%	0	70.37	435	KAG2420107.1	FALSE
hypothetical protein N7528_006317 [Penicillium herquei]	Penicillium herquei	599	599	96%	0	65.91	460	KAJ5621534.1	FALSE
FAD-dependent monooxygenase [Penicillium meliponae]	Penicillium meliponae	519	519	97%	2e-179	61.24	444	WDW19278.1	FALSE
Salicylate hydroxylase [Tolypocladium ophioglossoides CBS 100239]	Tolypocladium ophioglossoides CBS 100239	516	516	96%	4e-178	57.98	451	KND87340.1	TRUE
hypothetical protein N7523_003416 [Penicillium sp. IBT 18751x]	Penicillium sp. IBT 18751x	511	511	98%	1e-176	60.92	443	KAJ6119136.1	FALSE
uncharacterized protein CHGG_07644 [Chaetomium globosum CBS 148.51]	Chaetomium globosum CBS 148.51	510	510	97%	8e-176	54.76	455	XP_001225300.1 **CazL**	TRUE **CazE**

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. ident	Acc. Len	Accession	Co-occurred PF02458
FAD/NAD(P)-binding domain-containing protein [Coniochaeta ligniaria NRRL 30616]	Coniochaeta ligniaria NRRL 30616	507	507	98%	9e-175	58.12	445	OIW25916.1	TRUE
hypothetical protein B0I37DRAFT_444651 [Chaetomium sp. MPI-CAGE-AT-0009]	Chaetomium sp. MPI-CAGE-AT-0009	506	506	97%	3e-174	54.67	447	KAH6850663.1	TRUE
Note: The AfoD homologs co-occurring with an acyltransferase (AT) in PF02458 were highlighted in gray.									

Table S3.4. Biosynthetic genes for chaetoviridin A in Chaetomium globosum CBS 148.51.

gene*	left	right	strand	nuc* len	aa* len	domain (Pfam)	homology (swissprot)
1	20896	22284	+	1389	462	Transferase	A0A084B9Z3.1 O-acetyltransferase SAT5; AltName: Ful... 273 1e-85
2	23390	31435	+	7311	2436	ADH_zinc_N, ADH_zinc_N_2, Acyl_transf_1, KR, Ketoacyl-synt_C, Methyltransf_12, Methyltransf_23, PS-DH, adh_short, ketoacyl-synt	Q0CF75.1 Highly reducing polyketide synthase ATEG_0... 1941 0.0
3	32513	34519	-	2007	668	PIF1	A6ZM04.1 ATP-dependent DNA helicase PIF1; AltName: ... 85.9 5e-16
4	34626	35497	+	795	264	no hits	Q19202.2 Apyrase apy-1; AltName: Full=Uridine-dipho... 31.6 6.4
5	36583	39672	+	1668	555	DDE_1, DDE_3, HTH_Tnp_Tc5	A0A345BJN8.1 MFS-type transporter clz9; AltName: Fu... 140 8e-34
6	40290	42203	-	1704	567	DAO, FAD_binding_3, NAD_binding_8, Pyr_redox_2, Trp_halogenase (x2)	Q0CCX4.2 Sulochrin halogenase gedL; AltName: Full=F... 605 0.0
7	43040	45497	-	1677	558	Fungal_trans, Zn_clus	Q0CF68.2 Transcription factor ATEG_07666; AltName: ... 479 1e-162
8	47209	48855	-	1326	441	MFS_1, TRI12	Q5BEJ9.1 Efflux pump afoB; AltName: Full=Asperfuran... 486 3e-168
9	48908	49303	-	396	131	no hits	Q5BEJ9.1 Efflux pump afoB; AltName: Full=Asperfuran... 131 7e-36
10	50733	52251	+	1392	463	FAD_binding_3	Q0CF72.1 FAD-dependent monooxygenase ATEG_07662; Al... 537 0.0
11	53085	56700	+	3534	1177	Acyl_transf_1, Ketoacyl-synt_C, ketoacyl-synt	A0A0K0MCJ4.1 Iterative polyketide synthase CazM [Ch... 2388 0.0
12	56738	61328	+	4494	1497	Epimerase, Methyltransf_12, Methyltransf_23, NAD_binding_4, PP-binding, adh_short	A0A0K0MCJ4.1 Iterative polyketide synthase CazM [Ch... 3075 0.0
13	62301	64973	+	2073	690	CBM-like, fn3_3	Q2U5P7.1 Probable rhamnogalacturonate lyase C; Flag... 602 0.0
14	65857	67328	-	1413	470	BBE, FAD_binding_4	B6HNK6.1 FAD-linked oxidoreductase sorD; AltName: F... 464 6e-160
15	67895	69009	-	1032	343	no hits	A0A097ZPD4.1 Oxidoreductase andH; AltName: Full=And... 202 3e-61
16	70160	71209	-	1050	349	no hits	A6NDX4.1 PUTATIVE PSEUDOGENE: Putative transmembran... 32.0 2.5
17	72787	74691	-	1548	515	no hits	A0A1P8AUY4.1 Midasin; Short=AtMDN1; AltName: Full=D... 36.6 0.62
18	75740	77653	-	1632	543	DUF1688	A5H0J3.1 Uracil catabolism protein 4; AltName: Full... 348 3e-114
19	81448	83090	-	1643	546	PHD, PLU-1	Q9HDV4.1 Lid2 complex component lid2; Short=Lid2C c... 259 8e-74

Note: *cazL, a* flavin-dependent monooxygenase is the query gene 10 and its paired acyltransferase gene in PF02458, *cazE*, is labeled as gene 1 in the table.

Table S3.5. Alignment and distances of 96 residues centered around the active-site His166.

		Common	AncAT5 clade				AncAT6 clade			
Position	Distance	AncAT4	AncAT5	KAH6850668.1	KAH6641043.1	CazE	AncAT6	RYP17742.1	TpAT	KAF7554211.1
166	3.1	H	H	H	H	H	H	H	H	H
167	3.7	T	T	**A**	T	T	T	T	T	T
38	5.2	T	T	T	T	T	T	T	T	T
169	5.4	M	M	M	M	M	M	M	M	M
170	5.5	D	D	D	D	D	D	D	D	D
168	5.6	V	V	V	V	V	**M**	**M**	**M**	**M**
36	6.0	A	A	A	A	A	A	A	A	A
165	6.1	H	H	H	H	H	H	H	H	H
34	6.2	L	L	L	L	L	L	L	L	L
37	6.3	Y	Y	Y	Y	Y	Y	Y	Y	Y
35	6.3	N	N	N	N	N	**D**	**D**	**D**	**D**
171	6.7	G	G	G	G	G	G	G	G	G
164	7.7	S	S	S	S	S	S	S	S	**A**
29	7.8	G	G	G	G	G	G	G	G	G
312	7.9	A	A	A	A	A	A	A	A	**C**
134	8.0	L	L	L	L	L	L	L	L	L
39	8.0	H	H	H	H	H	H	H	**Q**	H
30	8.6	Q	Q	Q	Q	Q	Q	Q	Q	Q
31	8.9	F	F	F	F	F	F	F	F	F
33	8.9	V	V	V	V	V	**I**	**I**	**I**	**I**
313	9.0	Q	Q	**L**	Q	Q	Q	Q	Q	**H**
314	9.1	V	V	V	**I**	V	V	V	V	**I**
174	9.3	N	N	N	N	N	N	N	N	N
172	9.5	T	T	T	T	T	T	T	T	T
173	9.8	A	A	A	A	A	A	A	A	A
135	9.8	P	P	P	P	P	P	P	P	P
32	10.0	P	P	P	P	P	P	P	P	P
28	10.4	M	M	M	M	M	M	M	M	M
40	10.7	L	L	L	**I**	**I**	L	L	L	L
311	10.7	G	G	G	G	G	G	G	G	G
136	10.8	Q	Q	**H**	Q	Q	Q	Q	Q	**L**
163	11.0	S	S	S	S	S	S	S	S	S
147	11.0	I	I	I	I	I	I	I	**L**	I
315	11.0	Y	**H**	**H**	**H**	**H**	Y	Y	Y	Y
133	11.0	A	A	**G**	**S**	A	A	**G**	A	A
138	11.2	H	H	H	H	H	H	H	H	H
146	11.5	V	V	V	V	V	V	V	V	V
27	11.5	L	L	L	L	L	L	L	L	L
175	11.6	F	F	F	**L**	**L**	F	F	F	F
137	11.7	P	P	P	P	P	P	P	P	P
85	11.7	T	T	T	T	T	T	T	T	T
369	11.7	Y	Y	Y	Y	Y	Y	Y	Y	Y
86	11.9	A	**T**	**T**	**T**	**T**	A	A	A	A
26	12.0	D	D	D	D	D	D	D	D	D
148	12.3	T	T	T	T	T	T	T	**A**	T
84	12.5	R	**L**	**L**	**L**	R	R	R	R	R
370	12.5	N	N	N	**T**	N	N	N	N	N
77	12.6	M	M	M	M	**I**	M	M	**I**	**I**
297	12.7	I	I	I	I	I	I	I	I	I
296	12.8	A	A	A	A	A	A	A	A	A
310	13.0	M	**L**	**L**	**L**	**L**	M	M	M	M
316	13.0	T	**M**	**M**	**M**	**M**	T	T	T	T
145	13.1	P	P	P	P	P	P	P	P	P
382	13.1	T	T	T	T	T	T	T	T	T
298	13.2	D	D	D	D	D	D	D	D	D
132	13.3	P	P	P	P	P	P	P	P	P
162	13.3	L	**F**	**F**	**F**	**F**	L	L	L	L
176	13.5	Q	Q	Q	Q	Q	Q	Q	Q	Q
177	13.5	V	V	V	V	V	V	V	V	**L**
383	13.5	S	S	S	S	S	S	S	S	S
78	13.5	G	G	**D**	G	G	G	G	G	G
299	13.7	G	**A**	**A**	**A**	**A**	G	G	G	G
139	13.7	G	G	G	G	G	G	G	G	G
41	14.0	A	A	A	**T**	**T**	**C**	**C**	**C**	A
381	14.0	A	A	A	A	A	A	A	A	A
368	14.2	L	L	L	L	L	**M**	**M**	**M**	L
131	14.3	W	W	W	W	W	W	W	W	W
411	14.4	A	A	**P**	**T**	A	A	A	A	**T**
413	14.4	I	I	I	I	I	I	I	I	I
407	14.4	R	R	R	R	R	R	R	R	R
408	14.6	S	S	S	S	S	S	S	S	S
178	14.6	L	**P**	**P**	**P**	**P**	L	L	L	L
384	14.6	S	S	S	S	S	S	S	S	S
144	14.8	V	V	V	V	V	V	V	V	V
73	14.8	L	L	L	L	L	L	L	L	L
76	14.9	Q	Q	Q	Q	Q	Q	Q	Q	Q
300	15.1	R	R	R	R	R	R	R	R	R
140	15.1	L	L	**I**	L	L	L	L	**I**	L
418	15.2	T	T	**S**	T	T	**V**	**V**	**V**	**I**
380	15.2	G	G	G	G	G	G	G	G	G
130	15.3	P	P	P	P	P	P	P	P	P
25	15.3	P	P	P	P	P	P	P	P	P
367	15.3	L	L	L	L	L	L	L	L	L
295	15.4	R	R	R	R	R	R	R	R	R
410	15.4	G	G	G	G	G	G	G	**S**	G
416	15.5	S	S	S	S	S	**A**	**A**	**A**	**A**
309	15.6	Y	Y	**C**	Y	Y	Y	Y	Y	Y
83	15.9	V	V	V	V	V	V	V	V	V
129	16.5	T	T	T	**A**	**A**	T	T	T	T
161	16.5	N	N	**T**	N	N	N	N	N	N
358	16.6	I	I	I	I	I	I	I	I	I
364	16.7	R	R	R	R	R	R	R	R	R
202	16.9	R	R	R	R	R	R	R	R	R
72	16.9	W	W	W	W	W	W	W	W	W
387	17.8	T	T	T	**M**	**M**	T	T	T	T
206	18.0	I	I	I	I	I	I	I	I	I
Residue differences		-	8	18	17	13	7	8	13	13
Active-site identity w/ AT4		-	91.7%	81.3%	82.3%	86.5%	92.7%	91.7%	86.5%	86.5%

3.6.3 Reaction sequence using purified AncFDMO and clarified AT cell lysates

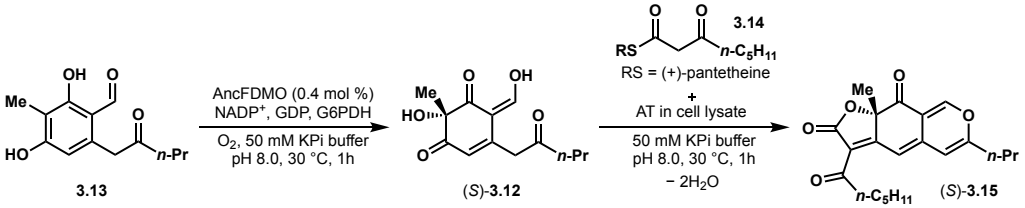


Figure S3.4. Enzymatic reaction sequence using purified AncFDMO and clarified AT cell lysates.

3.6.4 Reaction sequence using purified FDMO and AT

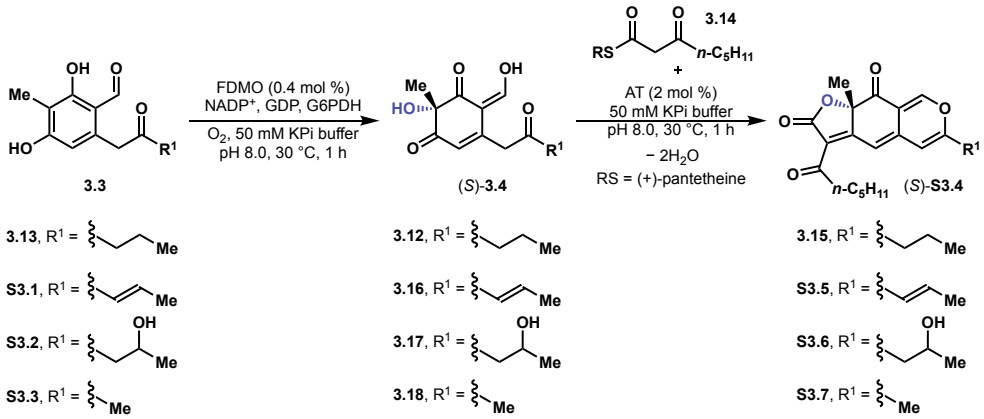


Figure S3.5. Enzymatic reaction sequence using purified AncFDMO and AT.

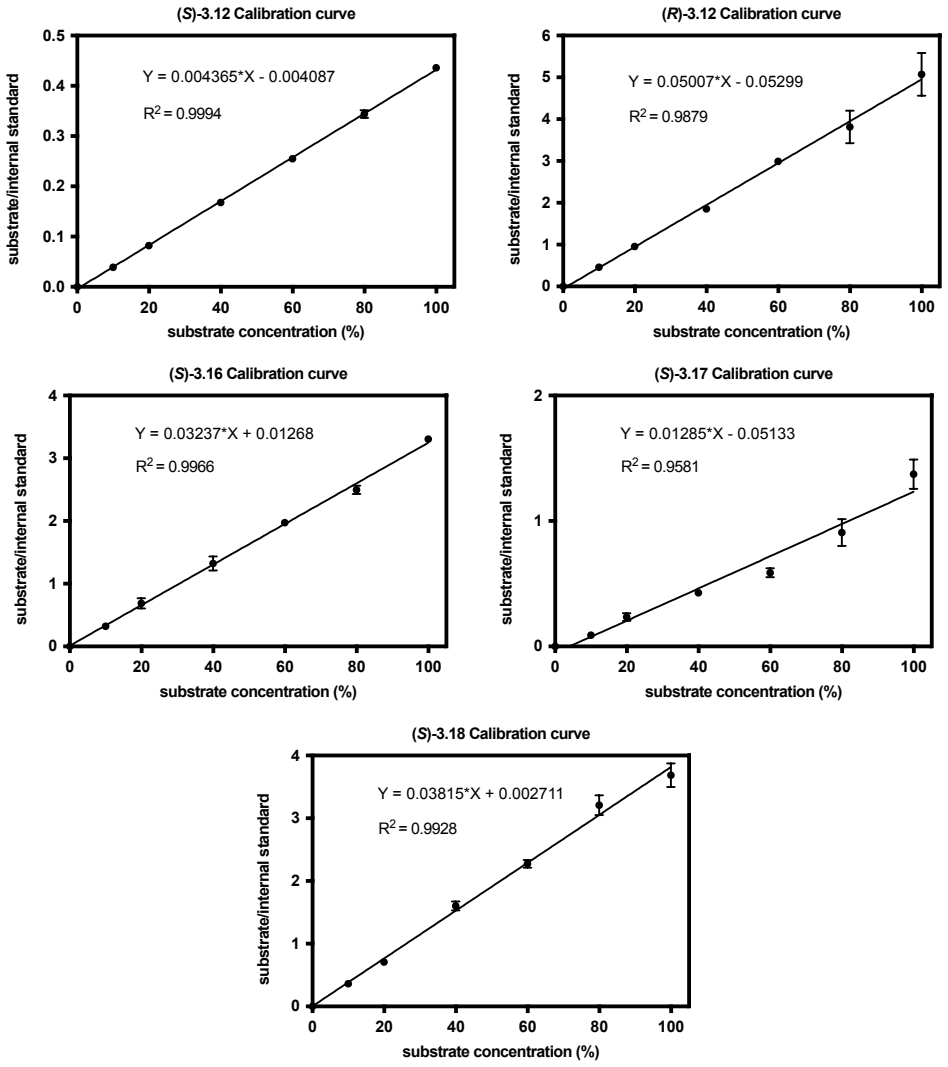


Figure S3.6. Standard calibration curves for (*S*)-**3.12** and (*R*)-**3.12**, (*S*)-**3.16**–**18**.
The substrate peak area at 273 or 300 nm was normalized by the peak area of the internal standard at 273 nm. The calibration curves cover the range from 0–100% and each sample was prepared in duplicate.

Enzymatic reactions of the substrate (*S*)-**3.12**

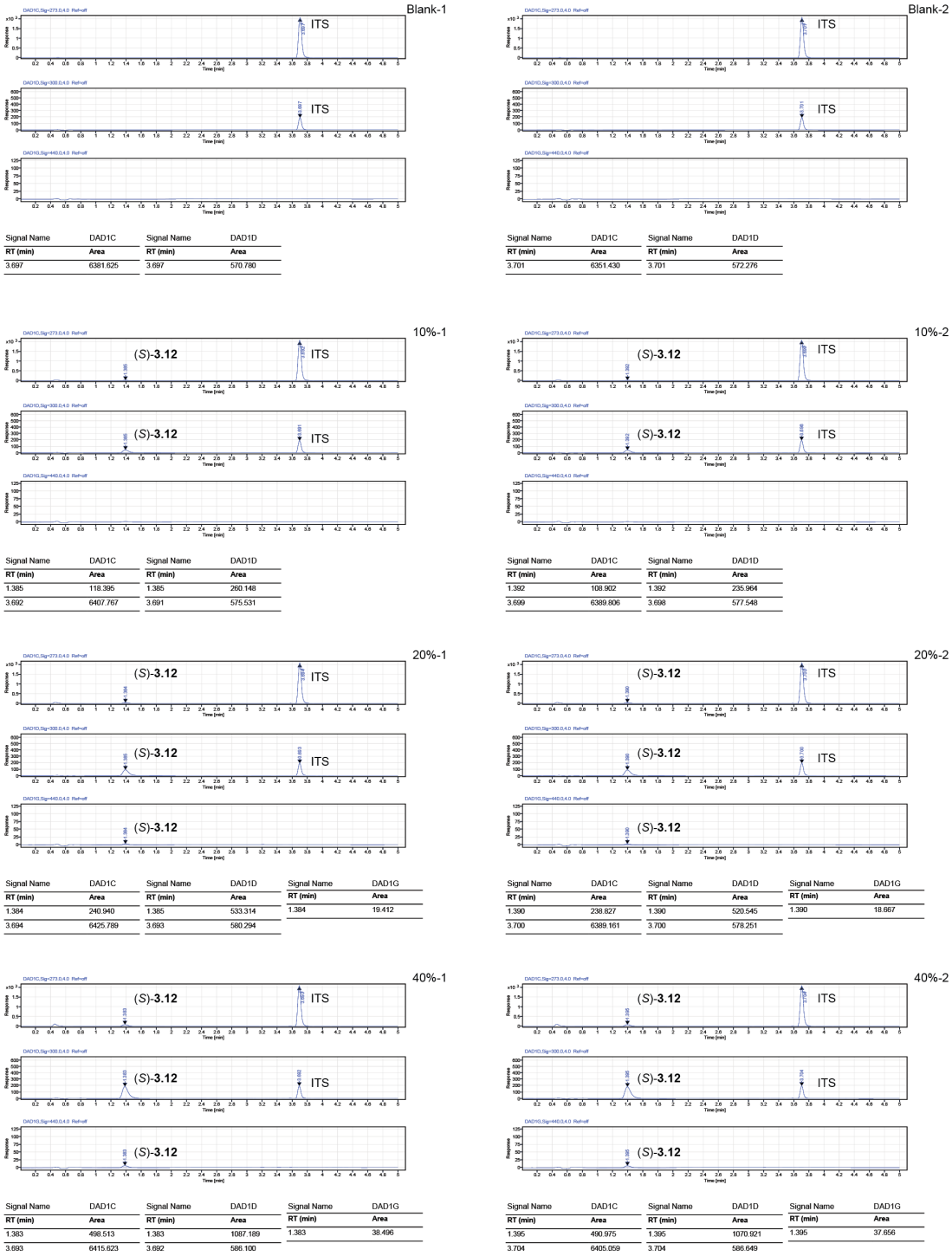


Figure S3.7. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

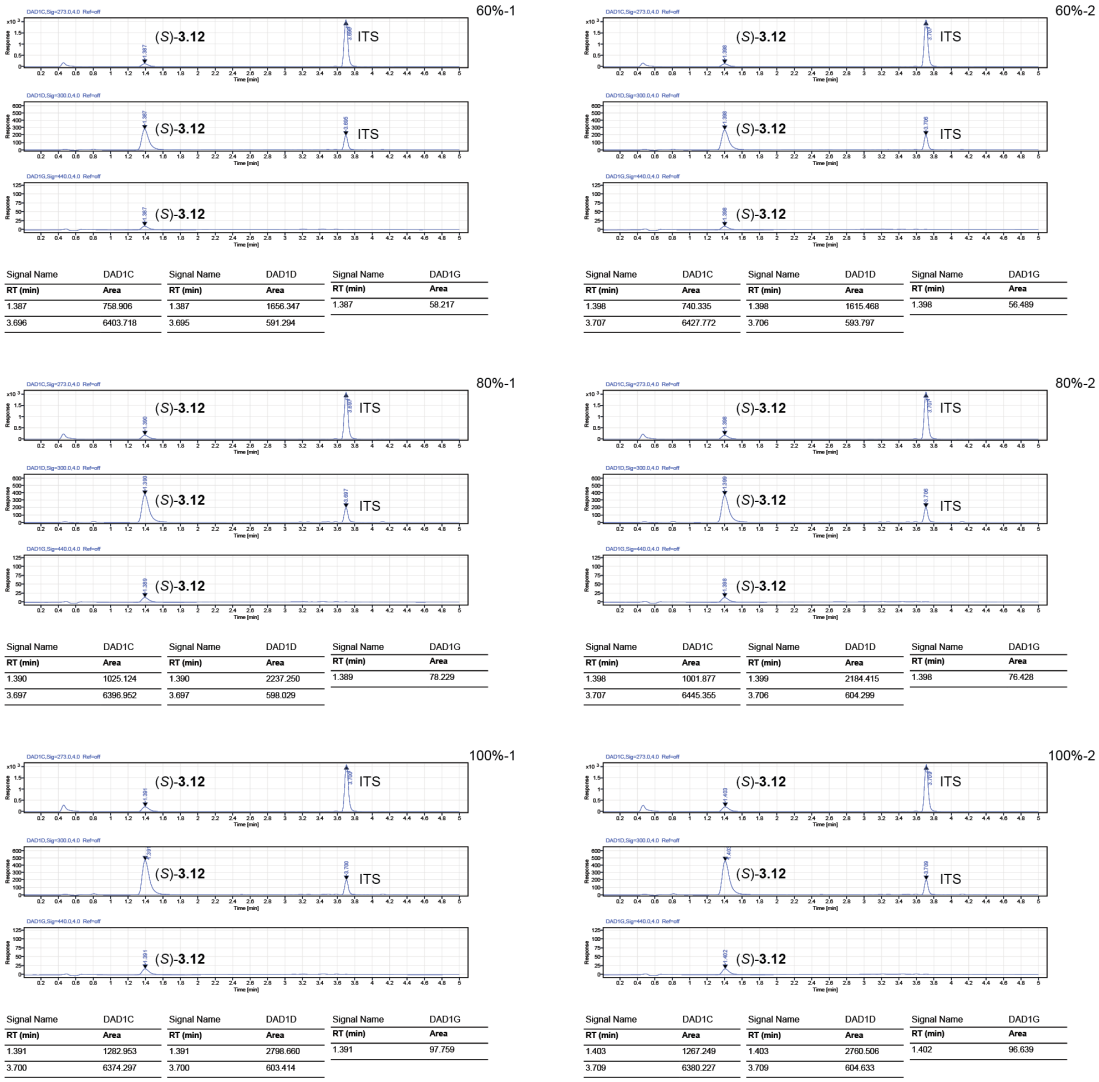


Figure S3.7 Continued. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

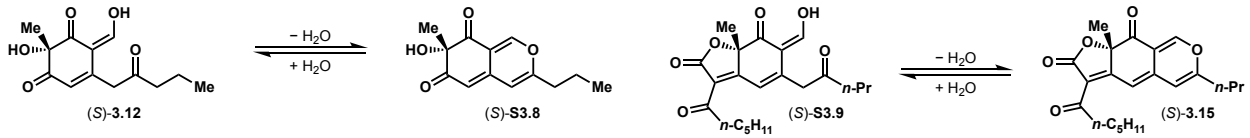
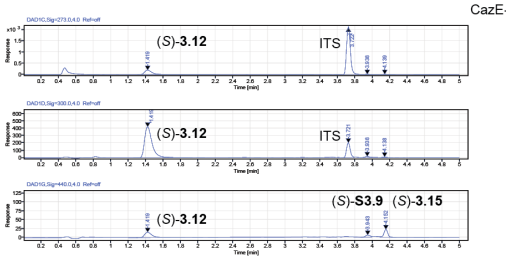


Figure S3.8. Substrate and product exist in closed ((*S*)-**S3.8** and (*S*)-**3.15**) and open ((*S*)-**3.12** and (*S*)-**S3.9**) forms.

CazE-1

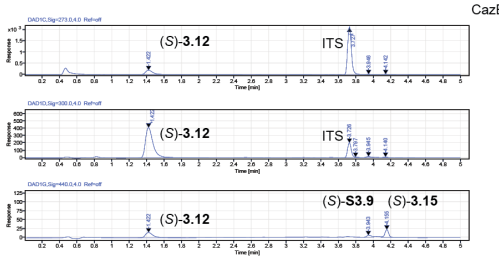


Signal Name	DAD1C
RT (min)	Area
1.419	1162.402
3.722	6437.011

Signal Name	DAD1D
RT (min)	Area
1.419	2534.705
3.721	593.230

Signal Name	DAD1G
RT (min)	Area
1.419	88.720
3.943	21.597
4.152	70.351

CazE-2

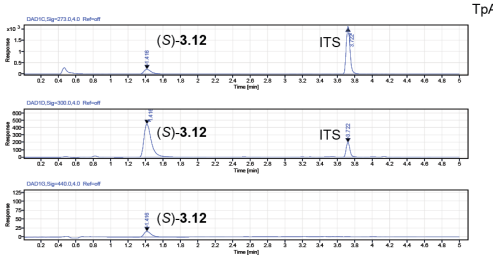


Signal Name	DAD1C
RT (min)	Area
1.422	1143.740
3.727	6407.581

Signal Name	DAD1D
RT (min)	Area
1.422	2485.047
3.726	594.623

Signal Name	DAD1G
RT (min)	Area
1.422	87.040
3.943	22.023
4.155	68.658

TpAT-1

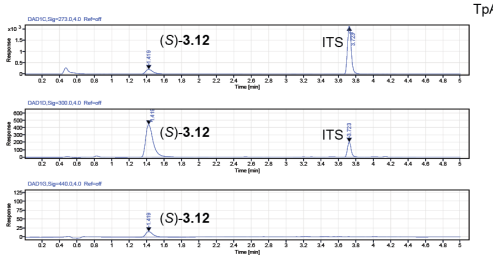


Signal Name	DAD1C
RT (min)	Area
1.416	1241.713
3.722	6462.737

Signal Name	DAD1D
RT (min)	Area
1.416	2705.206
3.722	598.682

Signal Name	DAD1G
RT (min)	Area
1.416	94.661

TpAT-2

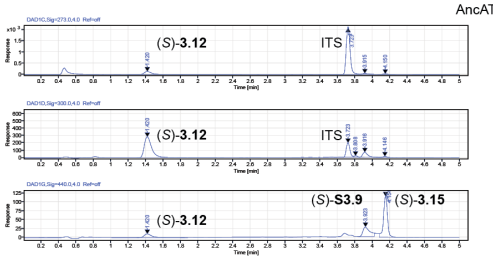


Signal Name	DAD1C
RT (min)	Area
1.419	1234.198
3.723	6459.971

Signal Name	DAD1D
RT (min)	Area
1.419	2685.459
3.723	600.632

Signal Name	DAD1G
RT (min)	Area
1.419	94.118

AncAT4-1

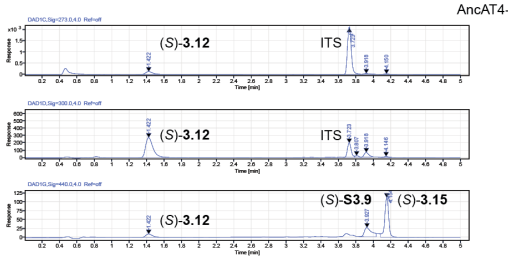


Signal Name	DAD1C
RT (min)	Area
1.420	779.553
3.723	6521.862
3.915	129.982

Signal Name	DAD1D
RT (min)	Area
1.420	1701.089
3.723	585.200
3.916	279.837

Signal Name	DAD1G
RT (min)	Area
1.420	59.472
3.923	149.635
4.154	361.129

AncAT4-2

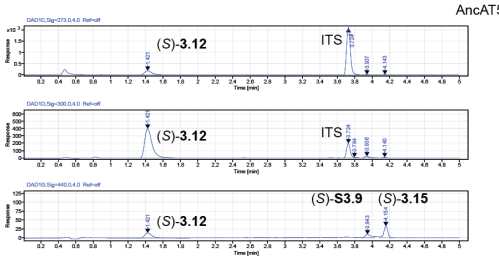


Signal Name	DAD1C
RT (min)	Area
1.422	761.354
3.723	6470.586
3.918	120.784

Signal Name	DAD1D
RT (min)	Area
1.422	1649.558
3.723	586.111

Signal Name	DAD1G
RT (min)	Area
1.422	57.823
3.927	146.290
4.154	347.879

AncAT5-1

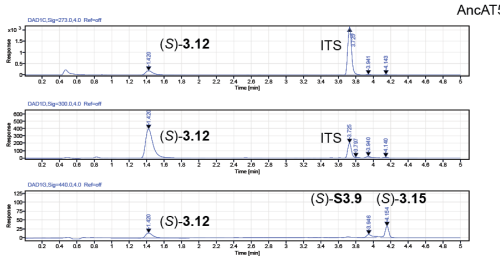


Signal Name	DAD1C
RT (min)	Area
1.421	1106.513
3.724	6461.140

Signal Name	DAD1D
RT (min)	Area
1.421	2410.336
3.724	589.959

Signal Name	DAD1G
RT (min)	Area
1.421	84.241
3.943	39.404
4.154	104.402

AncAT5-2



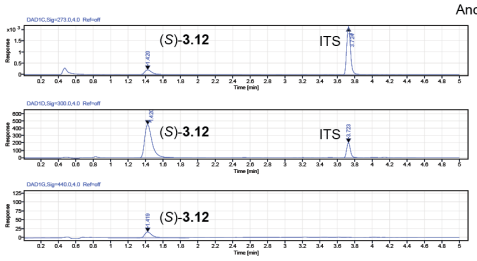
Signal Name	DAD1C
RT (min)	Area
1.420	1105.754
3.725	6489.247

Signal Name	DAD1D
RT (min)	Area
1.420	2382.408
3.725	601.464

Signal Name	DAD1G
RT (min)	Area
1.420	83.274
3.946	39.470
4.154	103.009

Figure S3.9. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

AncAT6-1

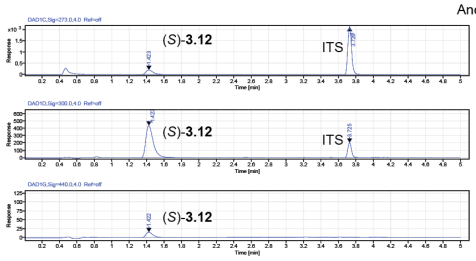


Signal Name	DAD1C
RT (min)	Area
1.420	1242.862
3.724	6516.362

Signal Name	DAD1D
RT (min)	Area
1.420	2707.497
3.723	606.726

Signal Name	DAD1G
RT (min)	Area
1.419	94.727

AncAT6-2



Signal Name	DAD1C
RT (min)	Area
1.423	1223.077
3.726	6507.237

Signal Name	DAD1D
RT (min)	Area
1.423	2654.060
3.725	608.239

Signal Name	DAD1G
RT (min)	Area
1.422	92.878

Figure 3.9 Continued. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

Enzymatic reactions of the substrate (*R*)-**3.12**

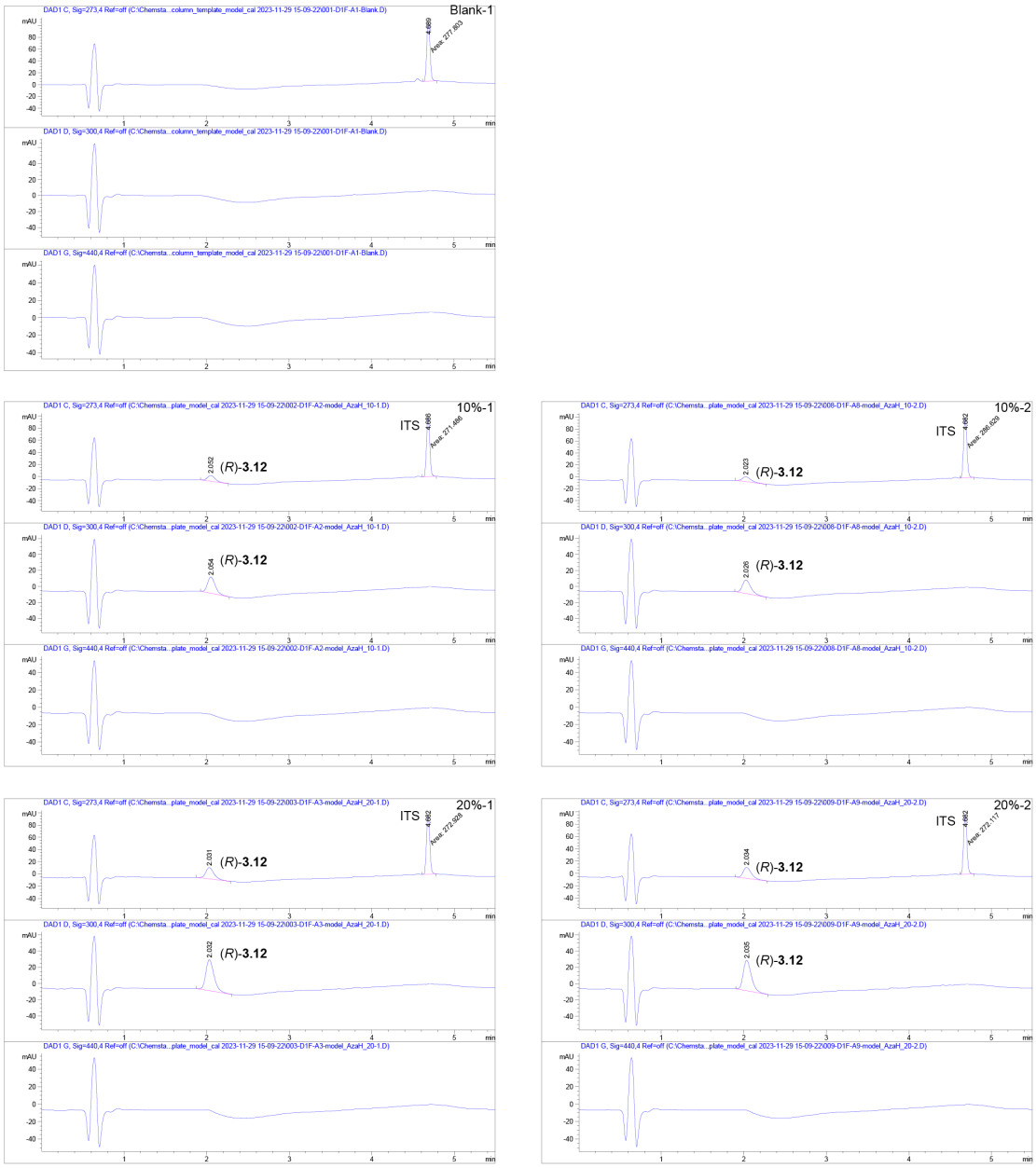


Figure S3.10. UPLC chromatograms of the samples for the standard calibration of (*R*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

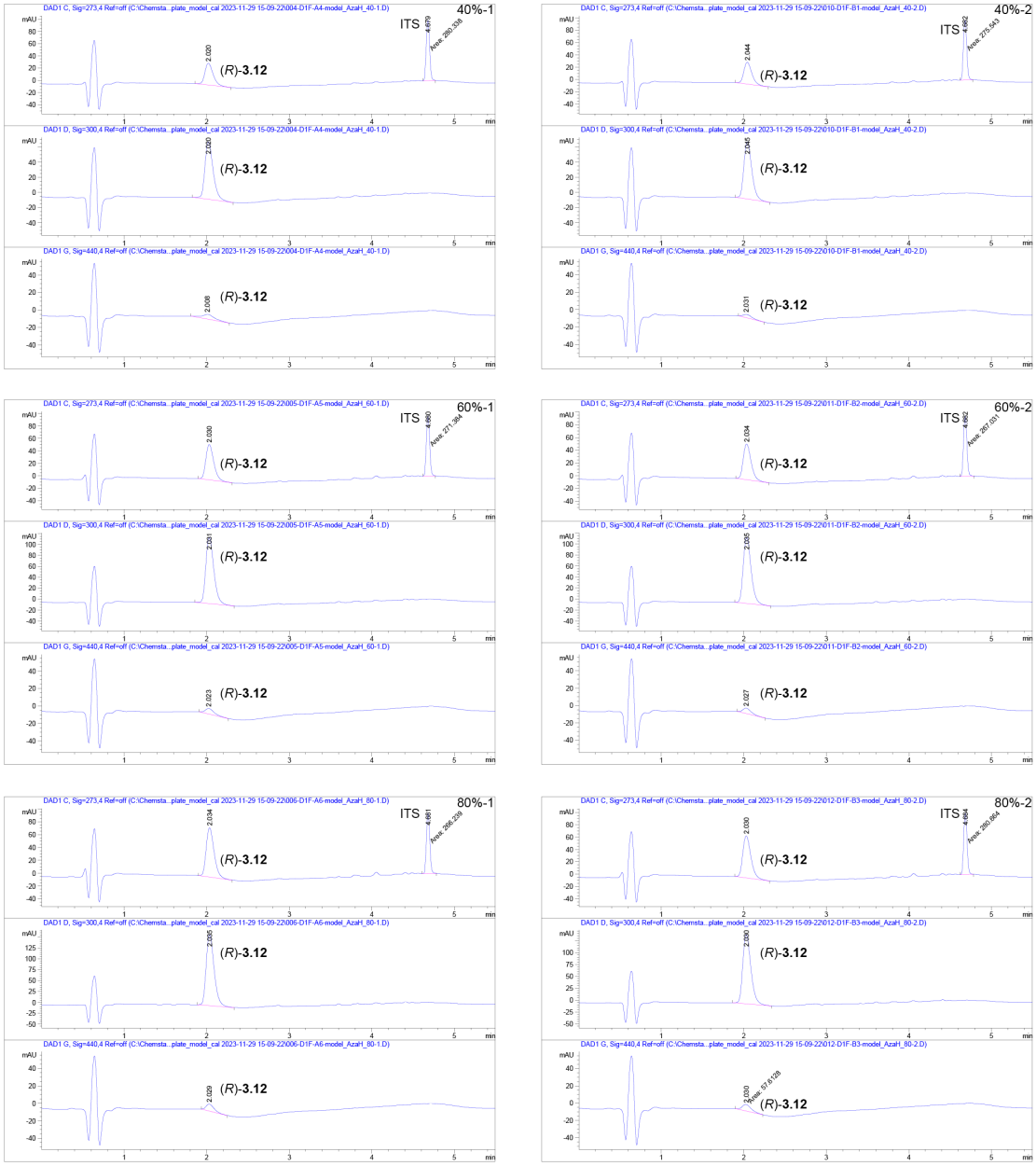


Figure 3.10 Continued. UPLC chromatograms of the samples for the standard calibration of (*R*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

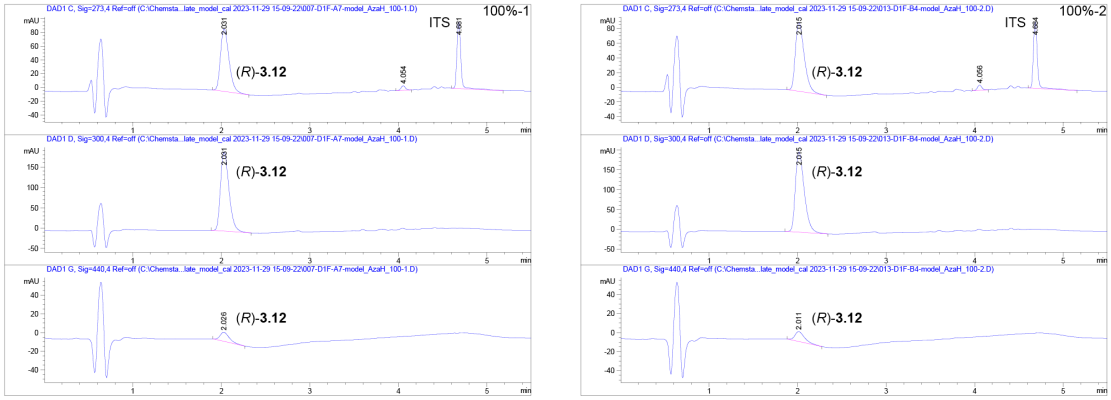


Figure 3.10 Continued UPLC chromatograms of the samples for the standard calibration of (*R*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

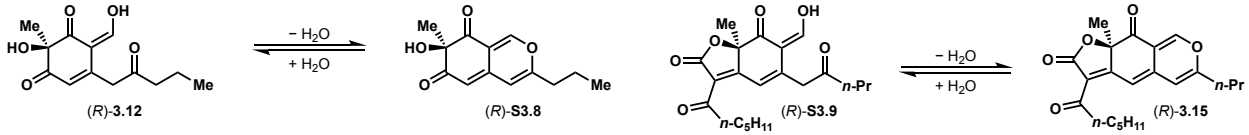


Figure S3.11. Substrate and product exist in closed ((*R*)-**S3.8** and (*R*)-**3.15**) and open ((*R*)-**3.12** and (*R*)-**S3.9**) forms.

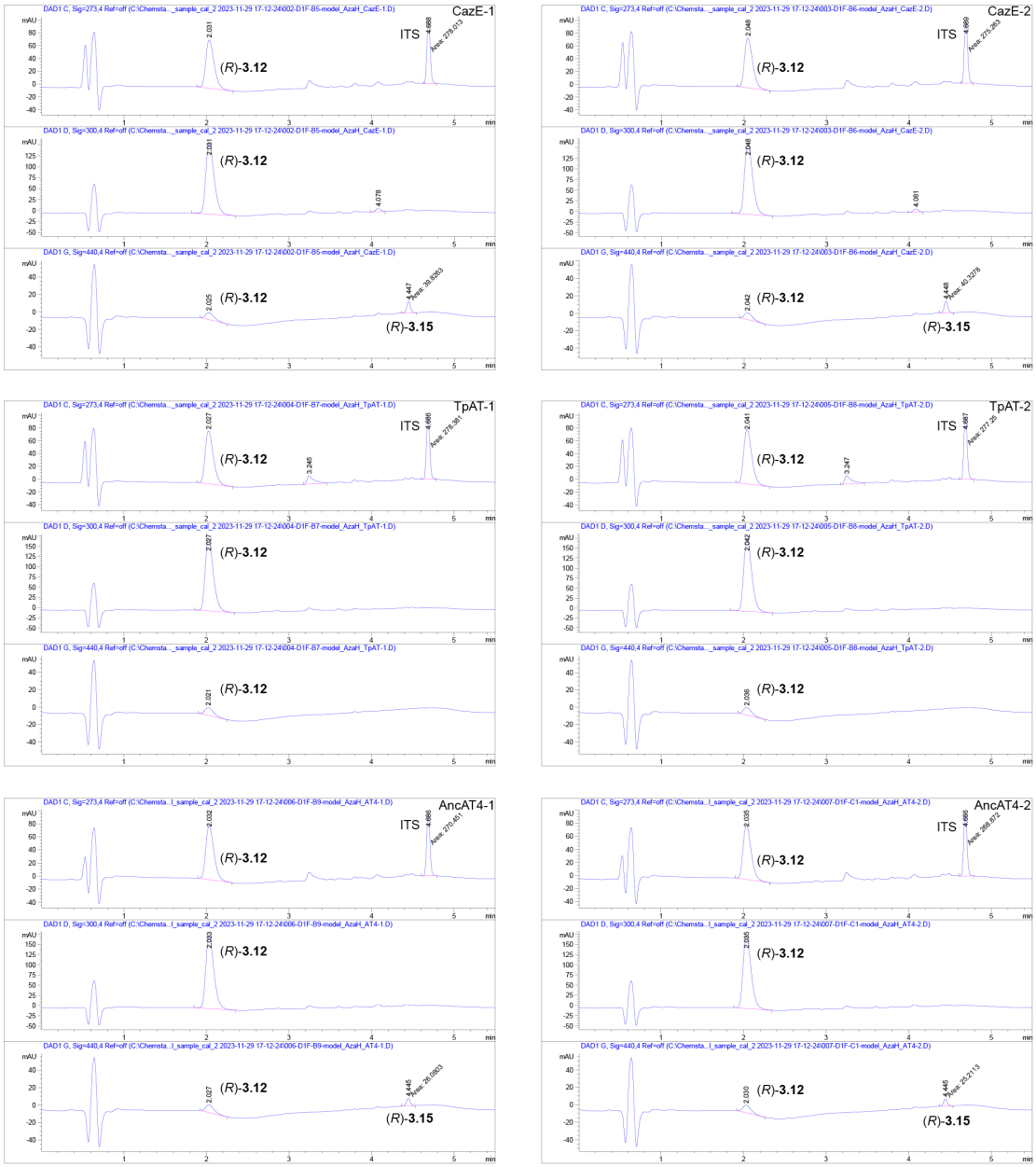


Figure S3.12. UPLC chromatograms of the samples from the enzymatic reactions of (*R*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

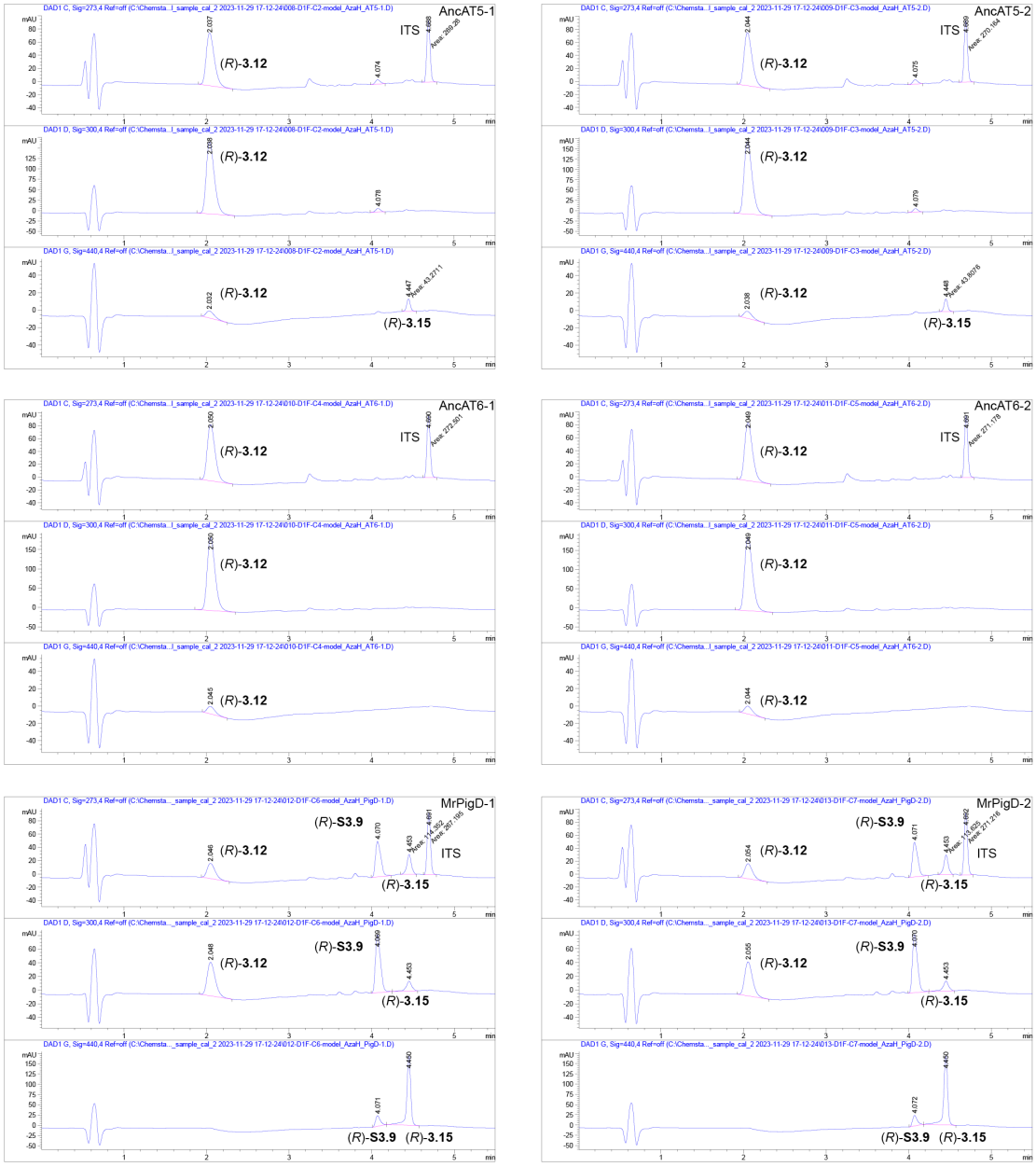


Figure 3.12 Continued. UPLC chromatograms of the samples from the enzymatic reactions of (*R*)-**3.12** at 273, 300, and 440 nm (ITS: internal standard).

Enzymatic reactions of the substrate (*S*)-**3.16**

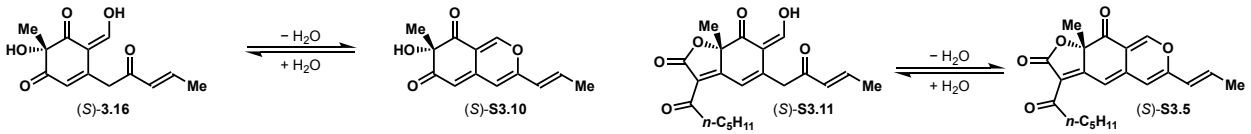


Figure S3.13. Substrate and product exist in closed ((*S*)-**S3.10** and (*S*)-**S3.5**) and open ((*S*)-**3.16** and (*S*)-**S3.11**) forms.

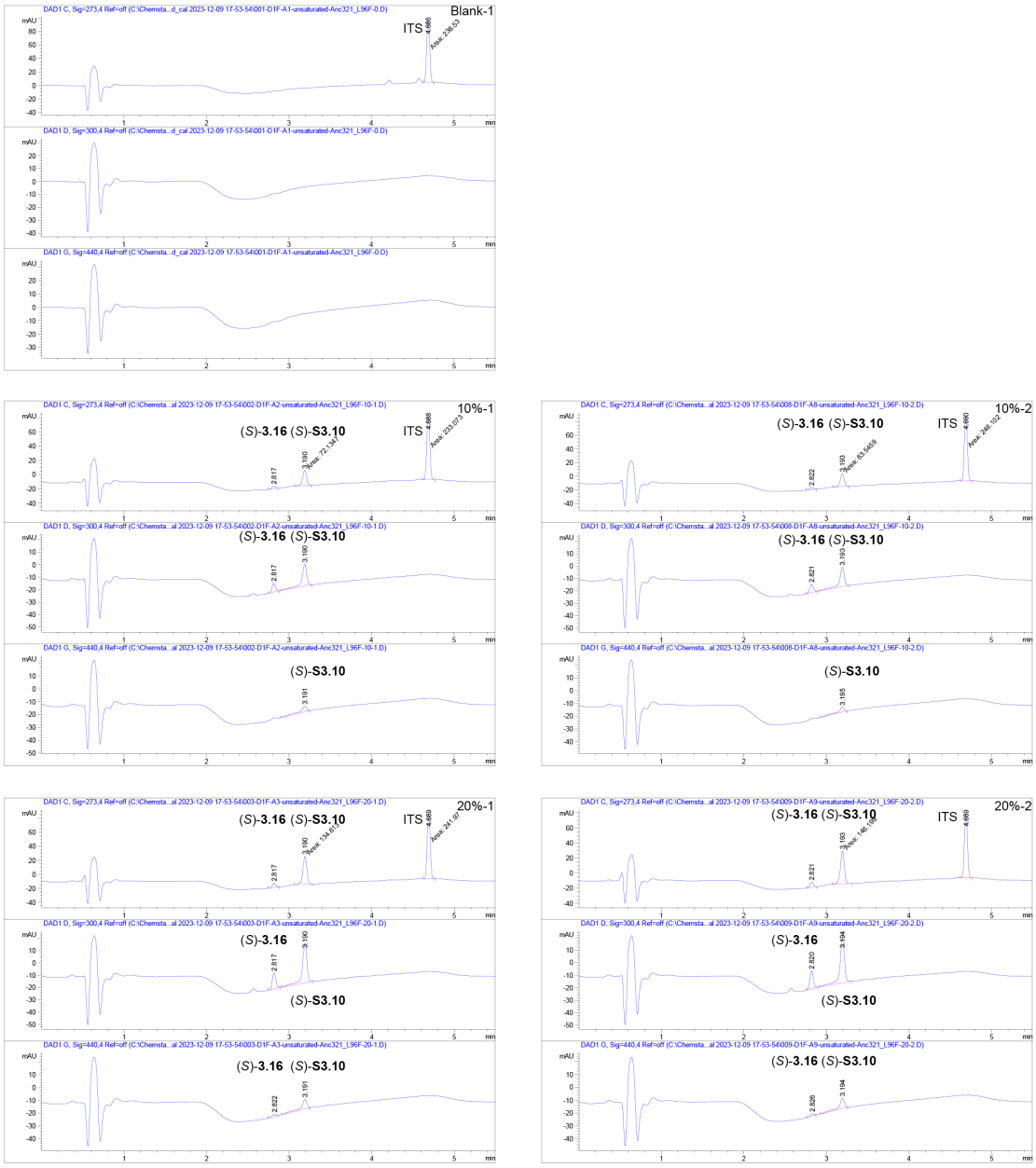


Figure S3.14. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.16** at 273, 300, and 440 nm (ITS: internal standard).

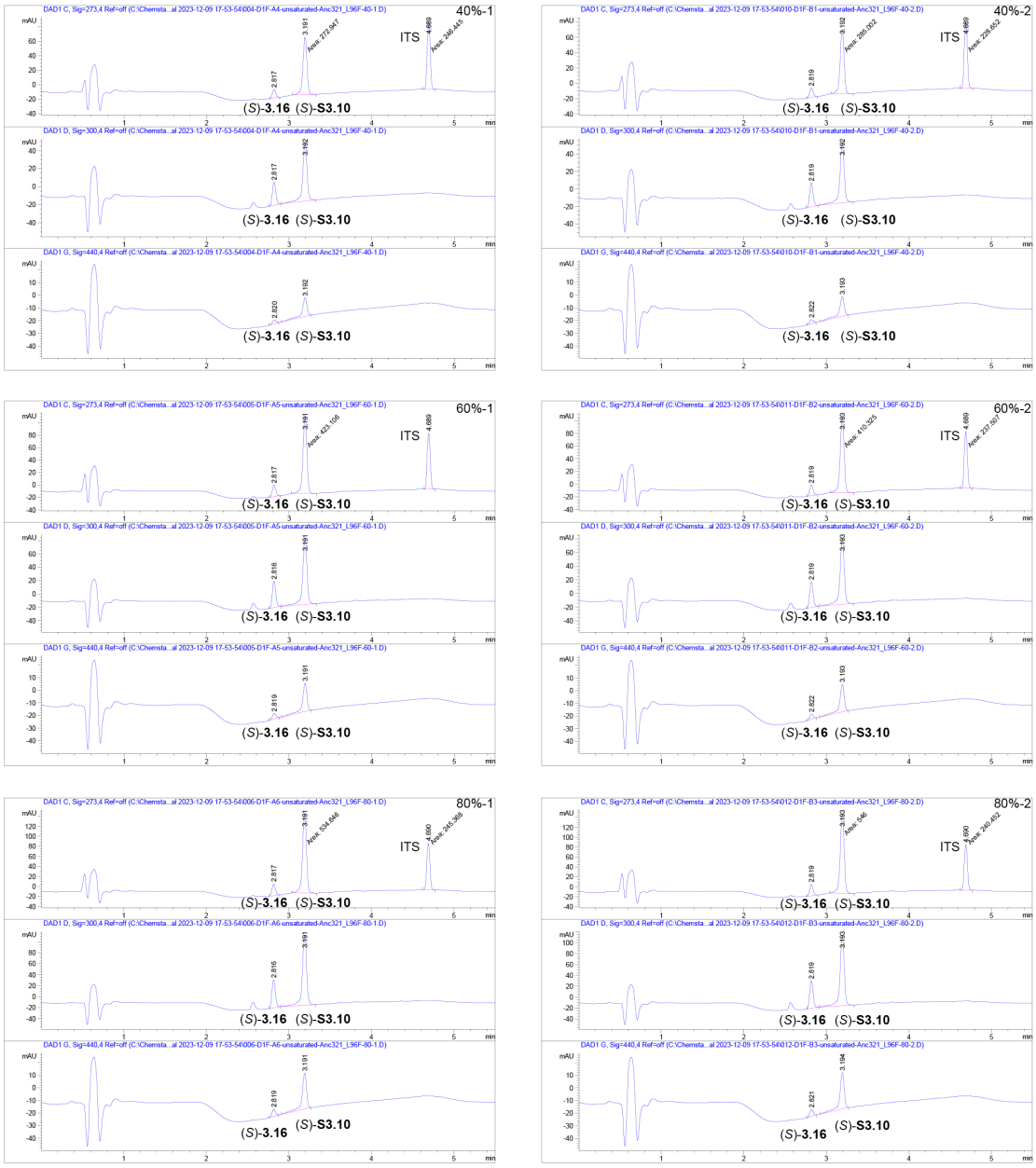


Figure S3.14 Continued. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.16** at 273, 300, and 440 nm (ITS: internal standard).

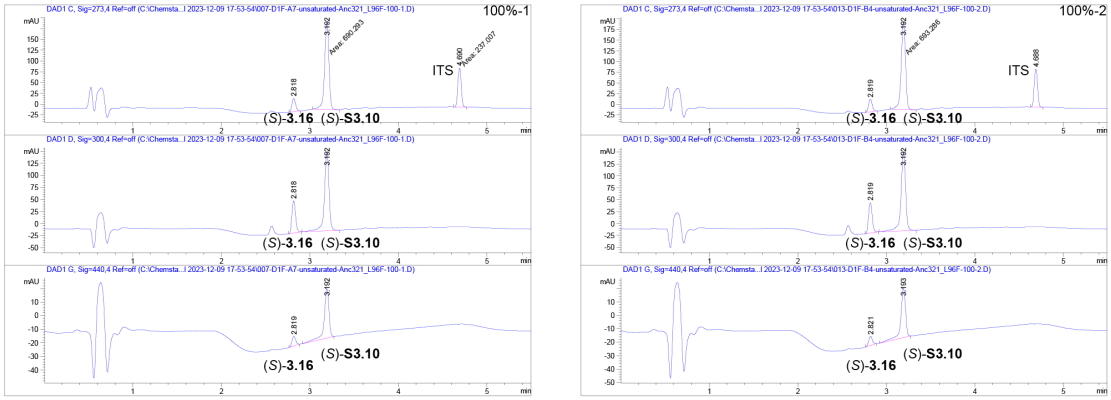


Figure S3.14 Continued. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.16** at 273, 300, and 440 nm (ITS: internal standard).

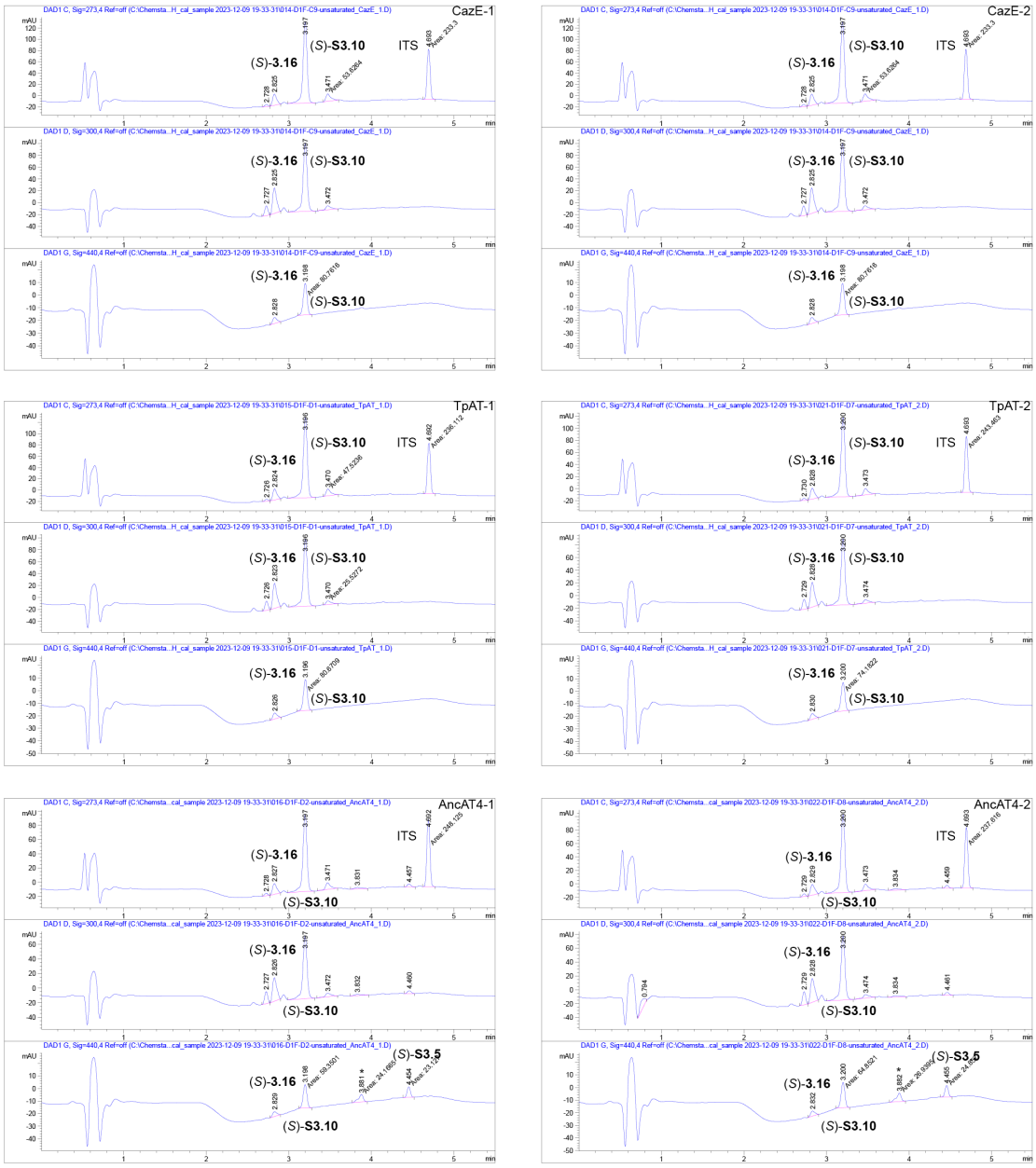


Figure S3.15. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.16** at 273, 300, and 440 nm (ITS: internal standard).
The peak labeled with a star is a side product coming from thiol-Michael addition.

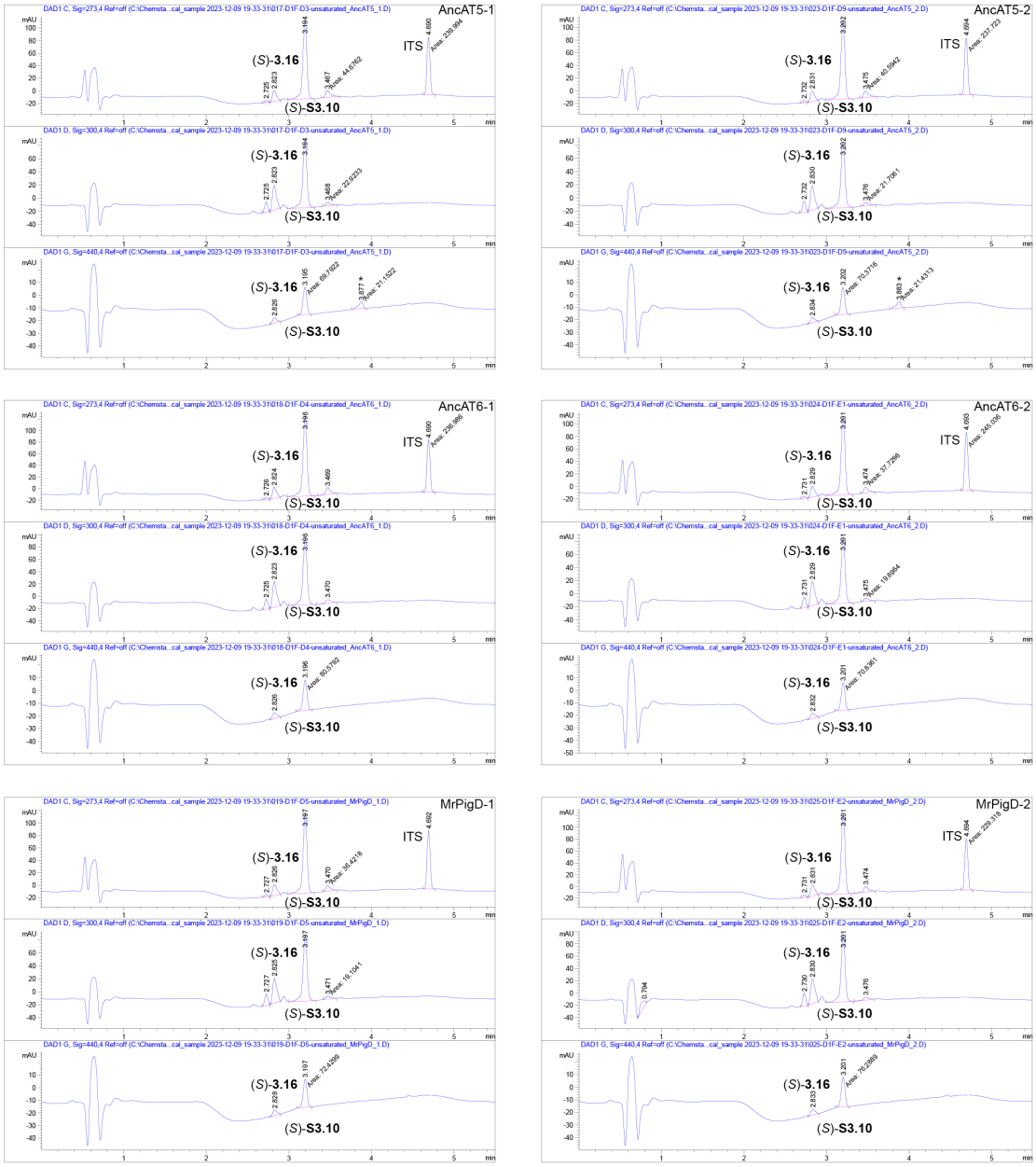


Figure S3.15 Continued. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.16** at 273, 300, and 440 nm (ITS: internal standard).
The peak labeled with a star is a side product coming from thiol-Michael addition.

Enzymatic reactions of the substrate (*S*)-**3.17**

Both the substrate and product can exist in closed ((*S*)-**S3.12** and (*S*)-**S3.6**) and open ((*S*)-**3.17** and (*S*)-**S3.13**) forms in solution. The equilibration of two forms in solution is very common and has been reported in our previous studies,[32,36] leading to the two peaks (retention time = 2.43 and 2.57 min) in the UPLC chromatograms. Each peak includes a mixture of diastereomers of (*S*)-**3.17** or (*S*)-**S3.12** because of the chiral alcohol, which results in a broader peak shape.

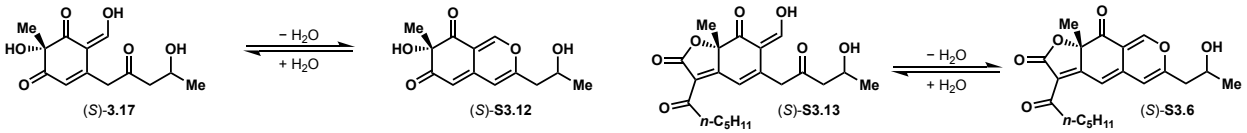


Figure S3.16. Substrate and product exist in closed ((*S*)-**S3.12** and (*S*)-**S3.6**) and open ((*S*)-**3.17** and (*S*)-**S3.13**) forms.

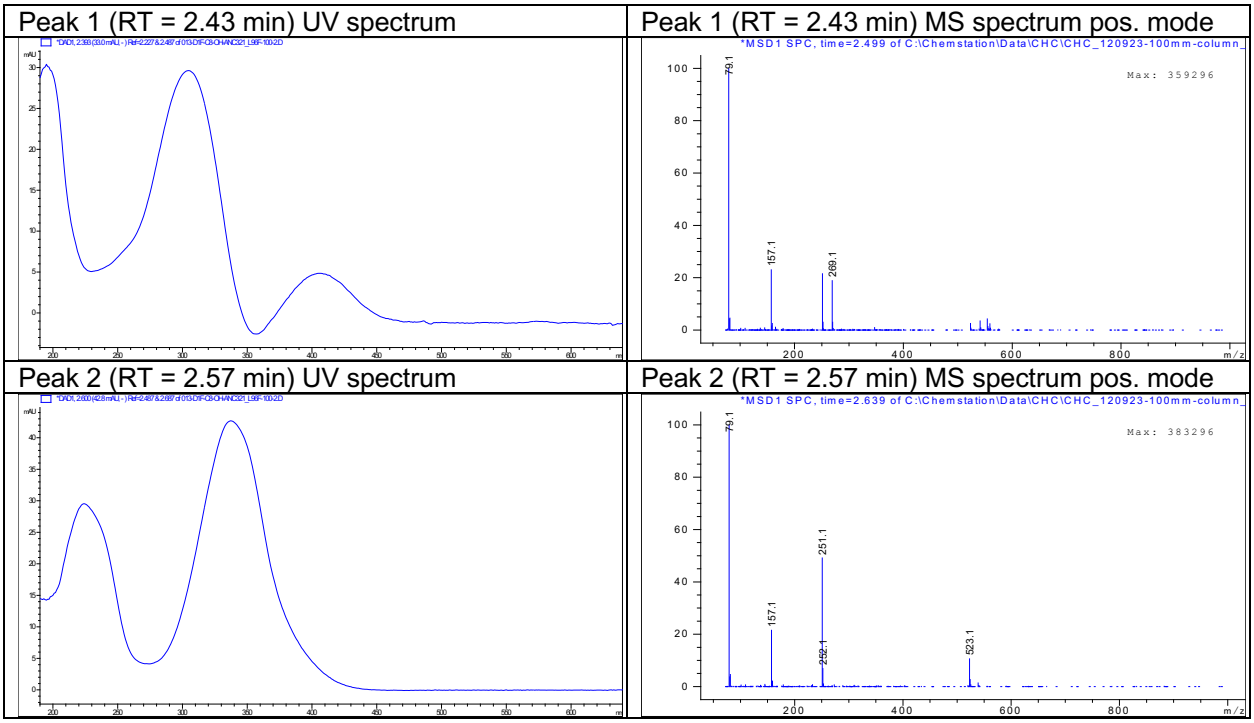


Figure S3.17. UV and MS spectra of the open (*S*)-**3.17** and closed (*S*)-**S3.12** forms of the acylation substrate eluted respectively from UPLC at retention time (RT) 2.43 and 2.57 min.

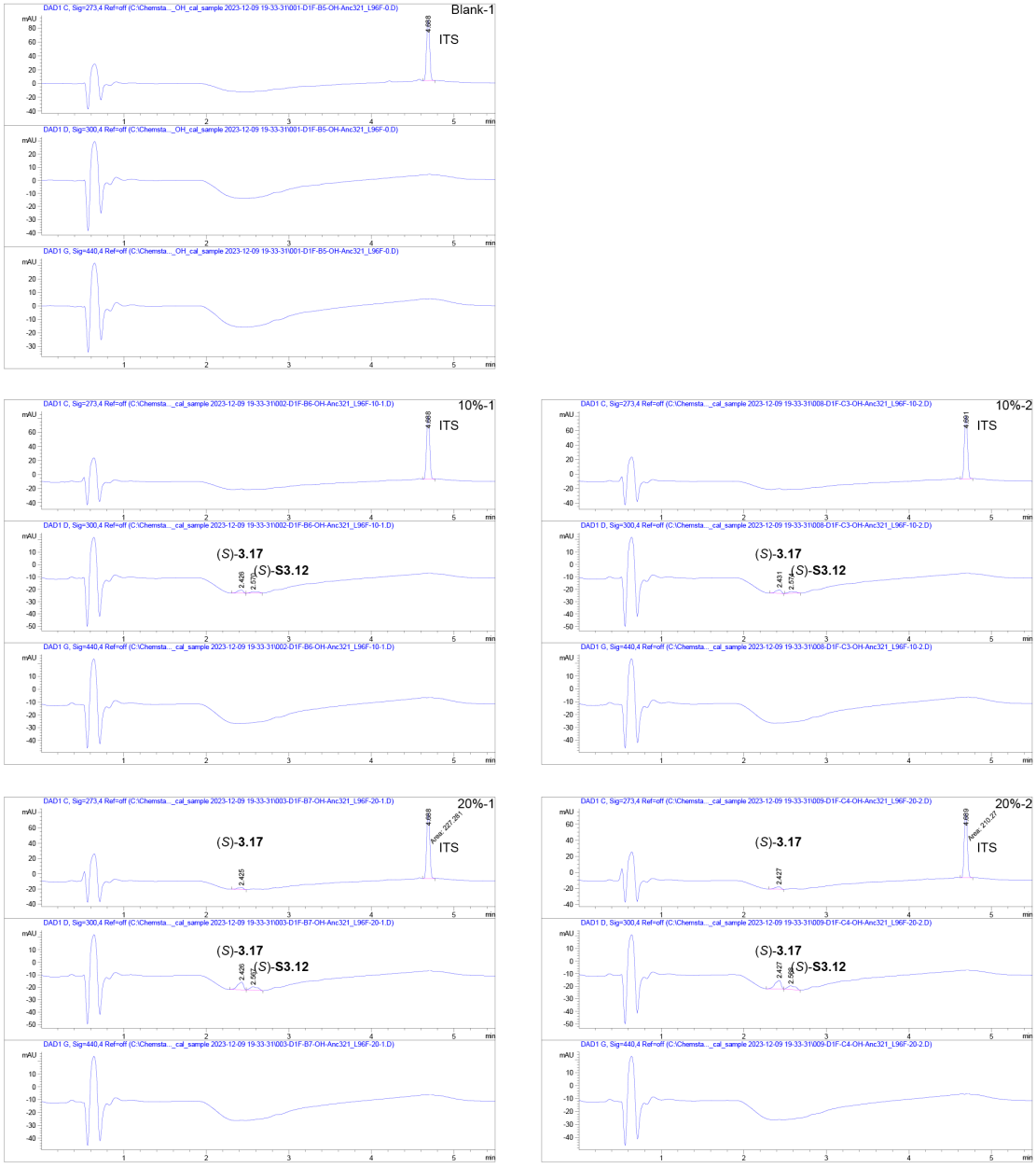


Figure S3.18. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.17** at 273, 300, and 440 nm (ITS: internal standard).

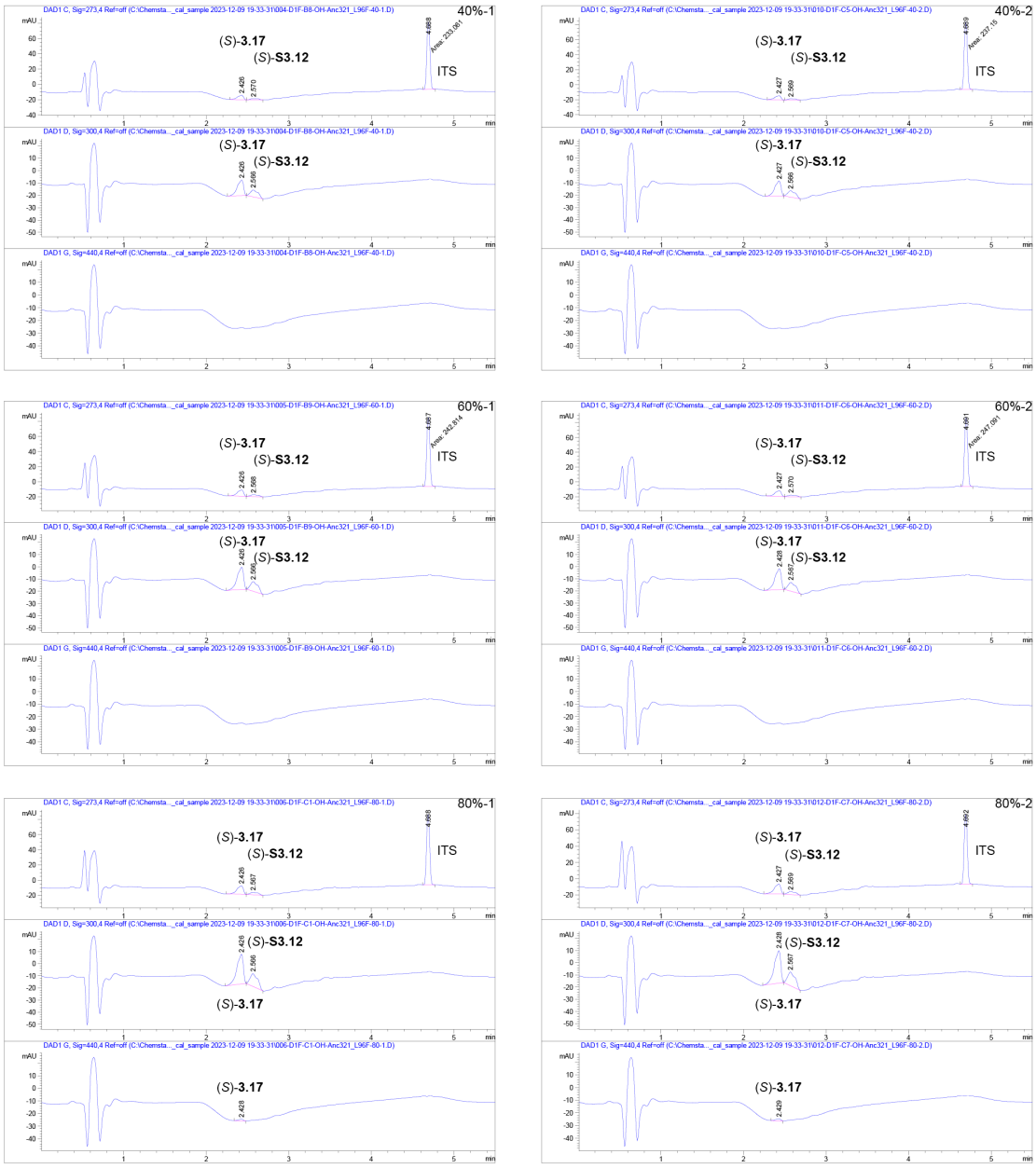


Figure S3.18 Continued. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.17** at 273, 300, and 440 nm (ITS: internal standard).

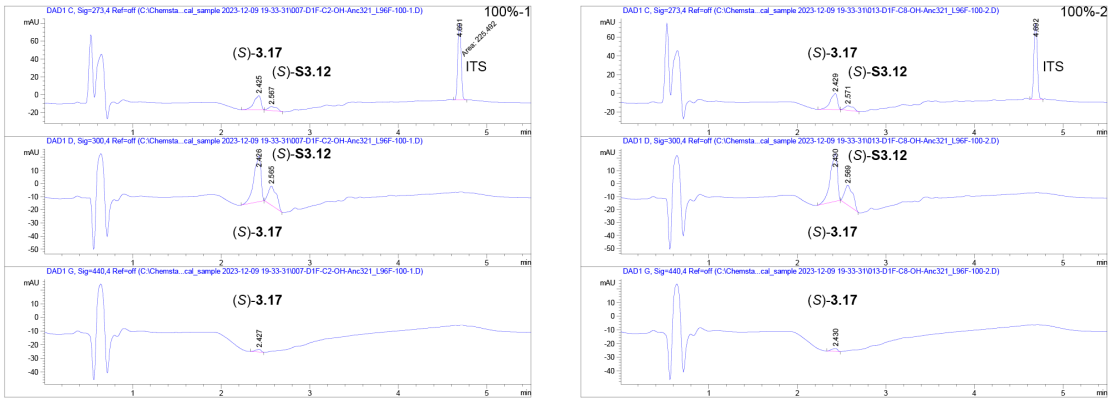


Figure S3.18 Continued. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.17** at 273, 300, and 440 nm (ITS: internal standard).

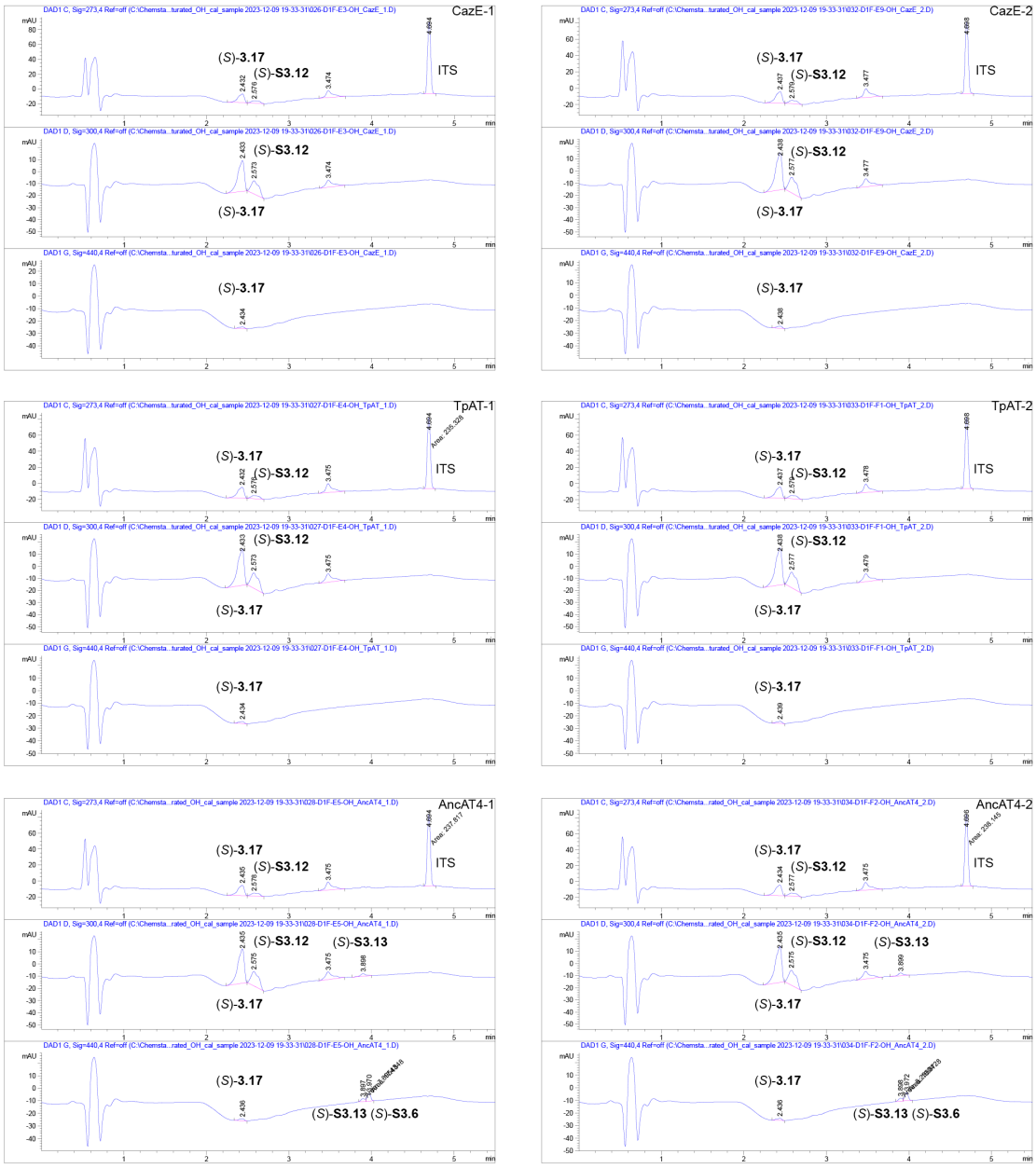


Figure S3.19. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.17** at 273, 300, and 440 nm (ITS: internal standard).

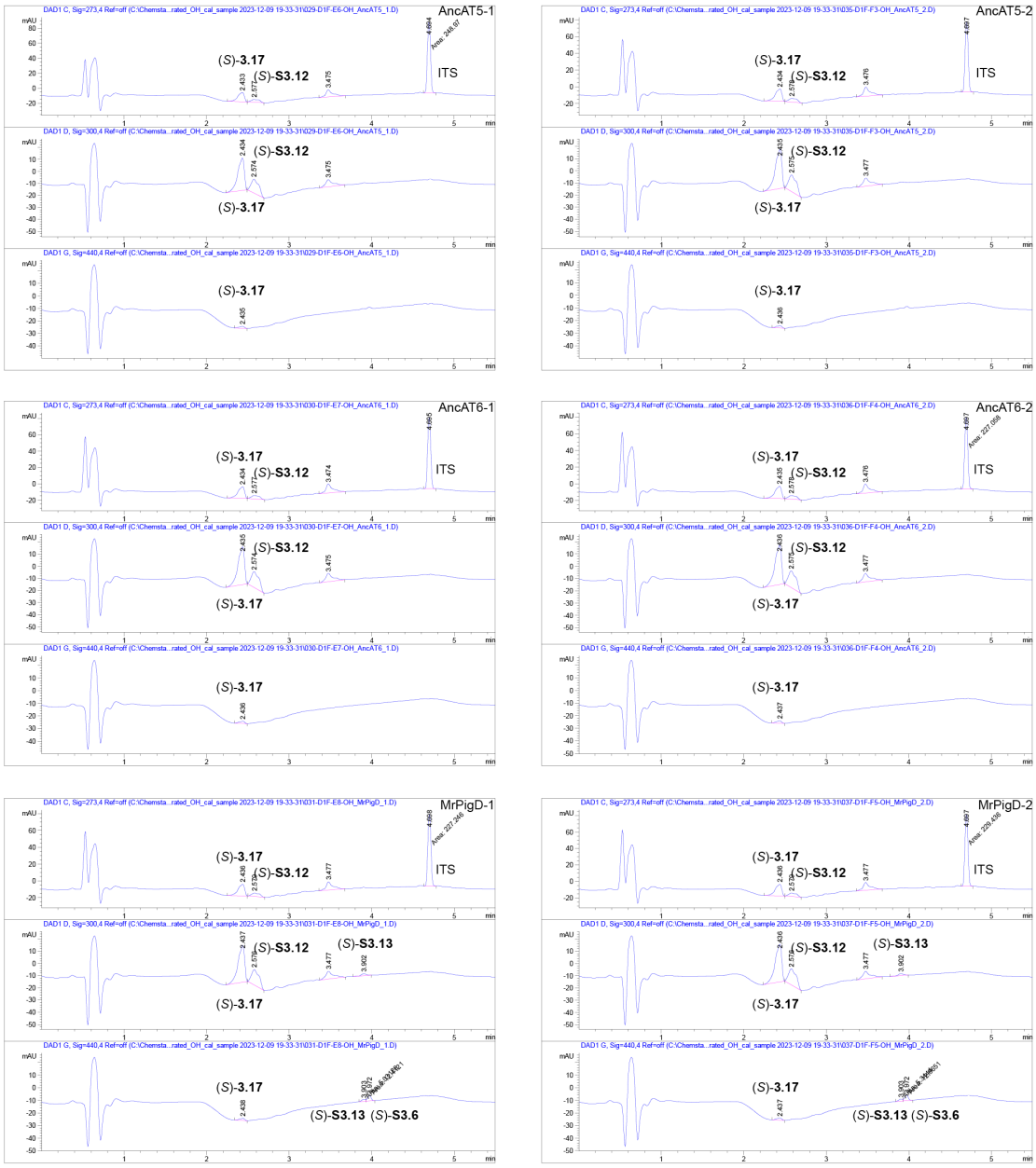


Figure 3.19 Continued. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.17** at 273, 300, and 440 nm (ITS: internal standard).

Enzymatic reactions of the substrate (*S*)-**3.18**

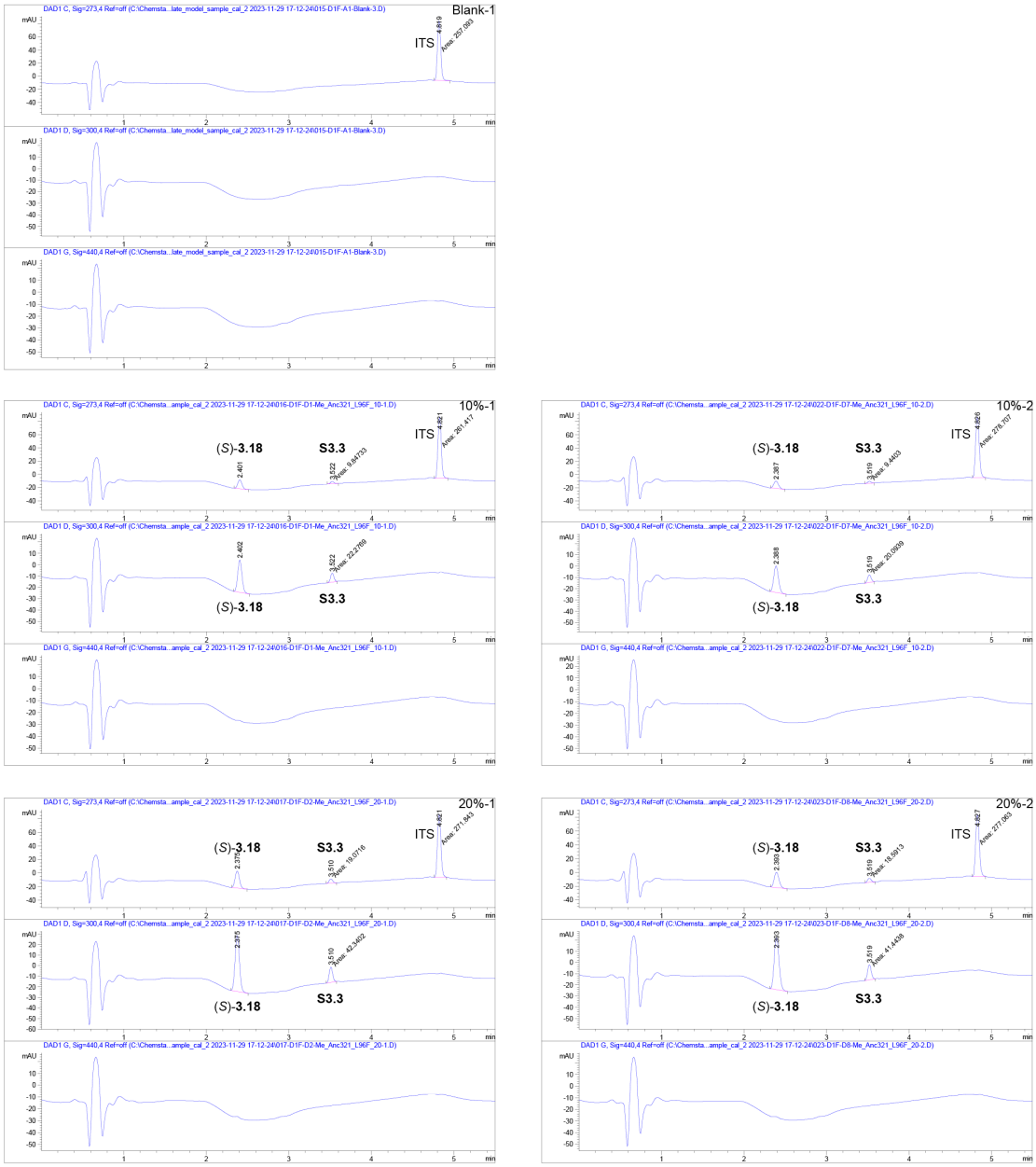


Figure S3.20. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.18** at 273, 300, and 440 nm (ITS: internal standard).

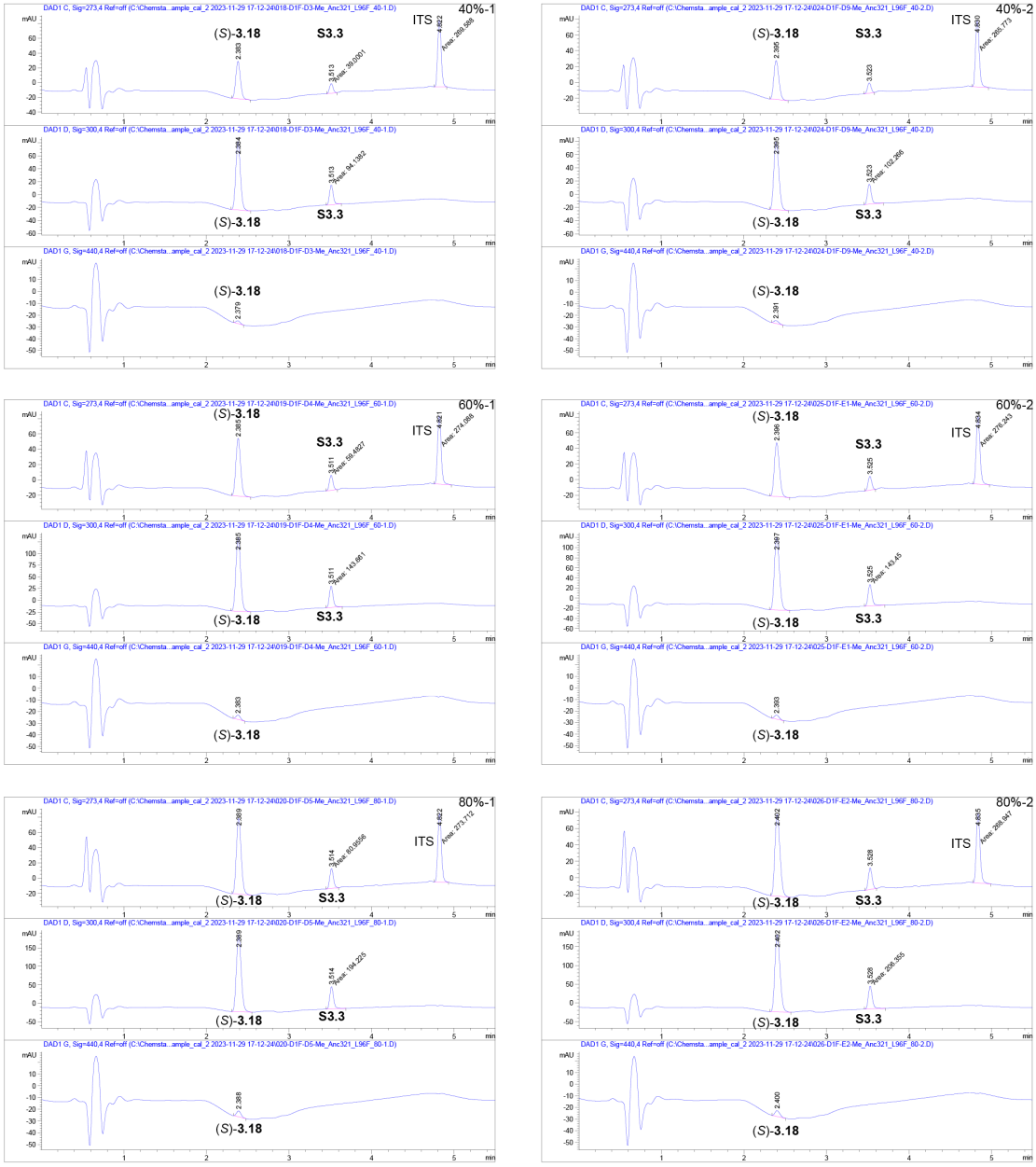


Figure 3.20 Continued. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.18** at 273, 300, and 440 nm (ITS: internal standard).

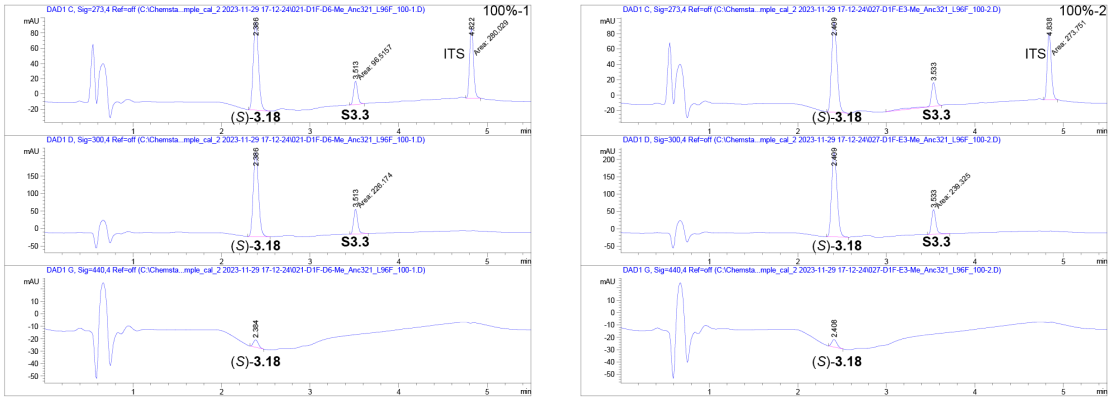


Figure 3.20 Continued. UPLC chromatograms of the samples for the standard calibration of (*S*)-**3.18** at 273, 300, and 440 nm (ITS: internal standard).

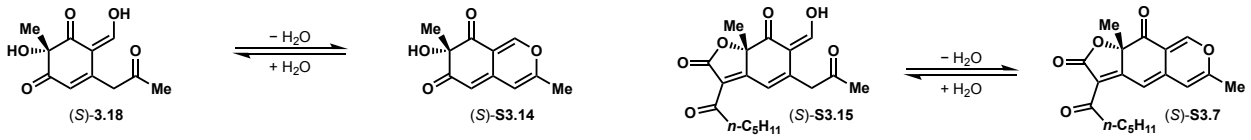


Figure S3.21. Substrate and product exist in closed ((*S*)-**S3.14** and (*S*)-**S3.7**) and open ((*S*)-**3.18** and (*S*)-**S3.15**) forms.

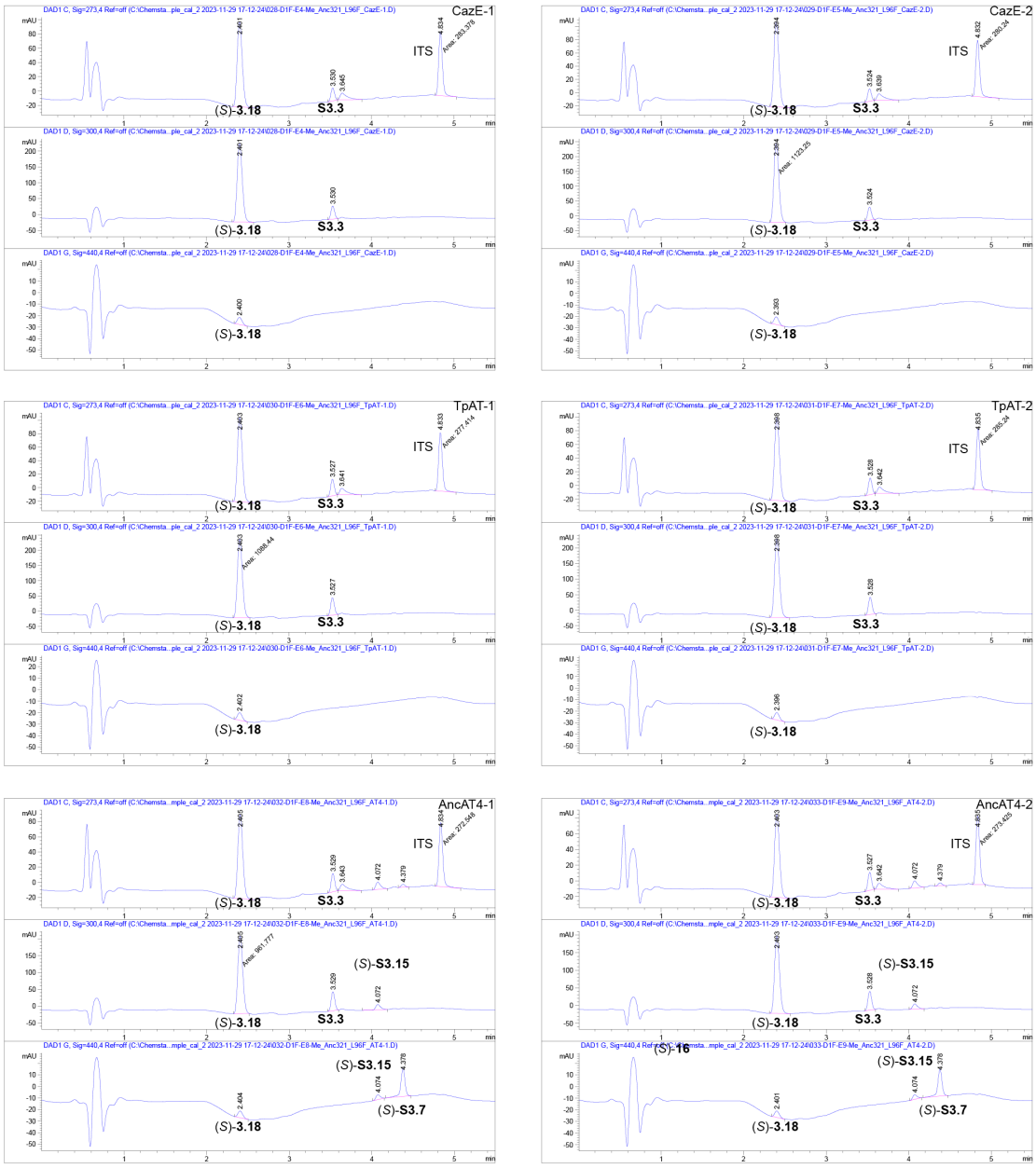


Figure S3.22. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.18** at 273, 300, and 440 nm (ITS: internal standard).

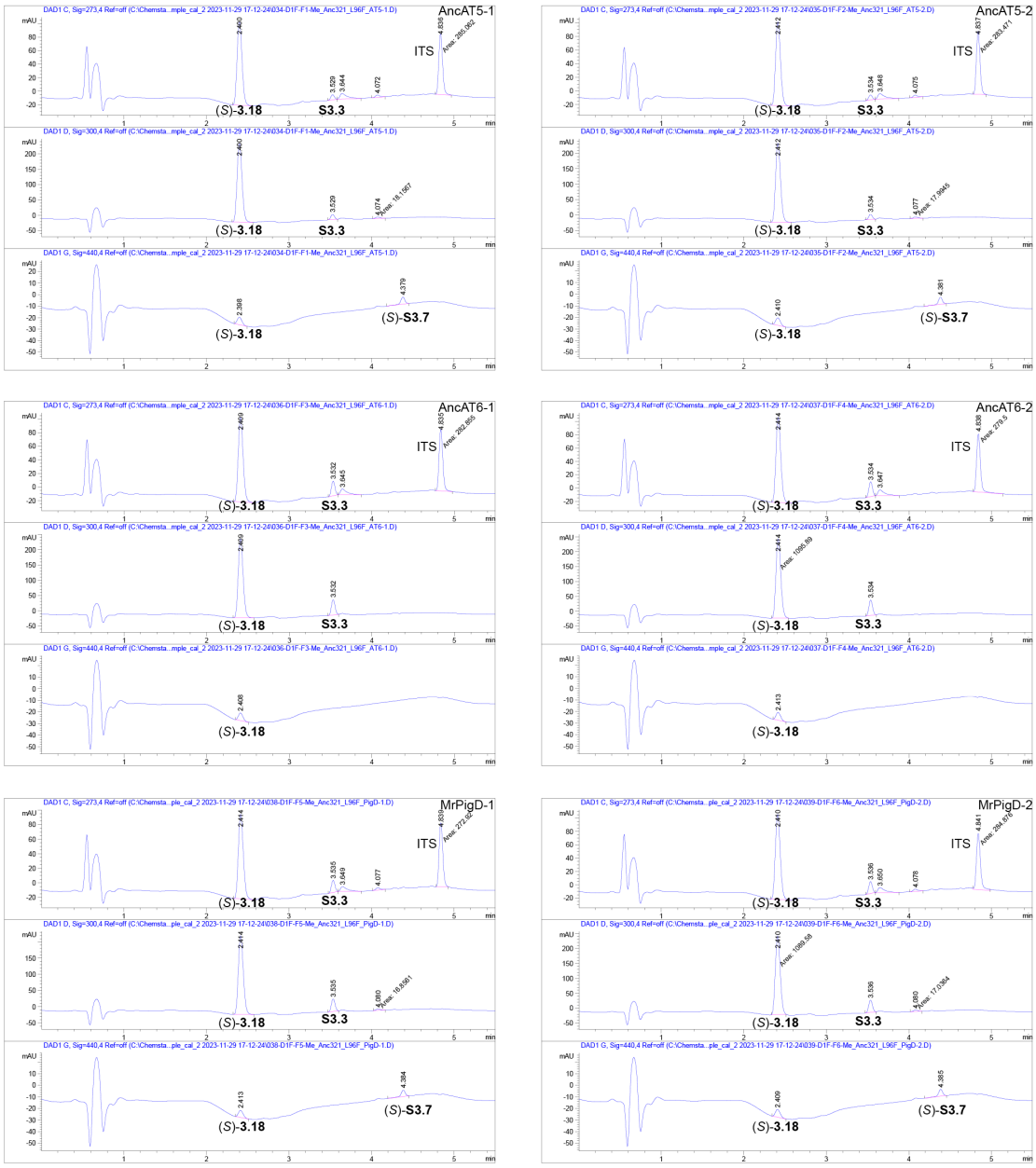


Figure 3.22 Continued. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.18** at 273, 300, and 440 nm (ITS: internal standard).

Screening of thioesters using purified AncFDMO and AncAT4

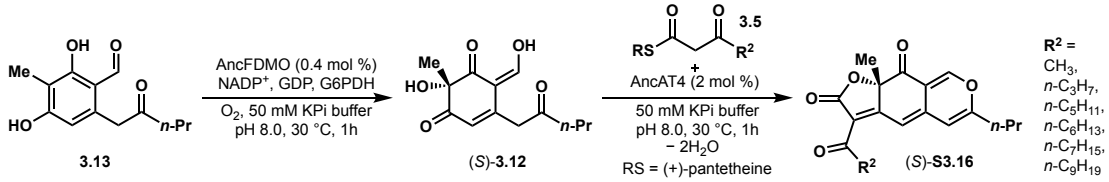


Figure S3.23. Screening of thioesters from C1 to C9 using purified AncFDMO and AncAT4.

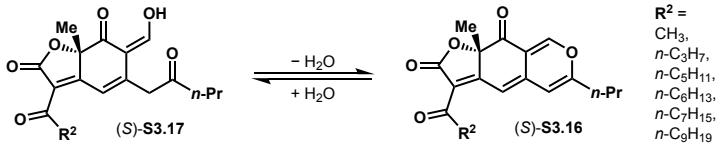
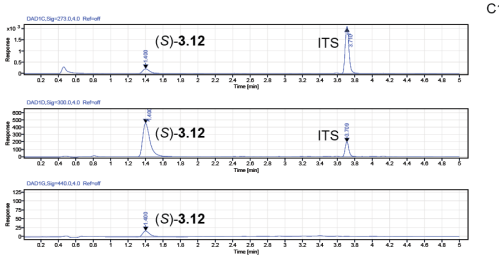


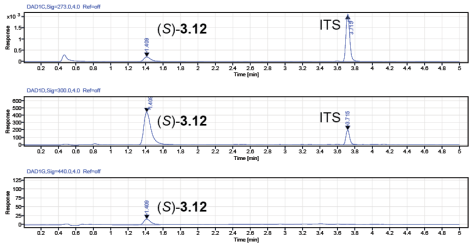
Figure S3.24. Product can exist in closed (*S*)-**S3.16** and open (*S*)-**S3.17** forms.

C1-1



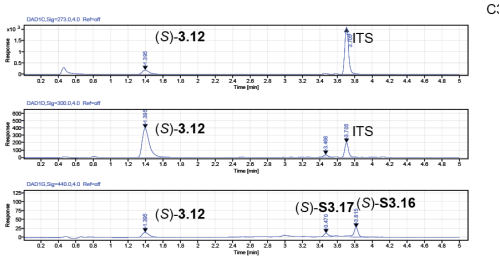
Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.400	1232.877	1.400	2683.486	1.400	94.622
3.710	6427.256	3.709	597.632		

C1-2



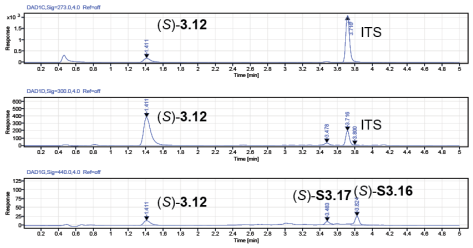
Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.409	1194.585	1.409	2607.527	1.409	91.849
3.715	6442.322	3.715	596.593		

C3-1



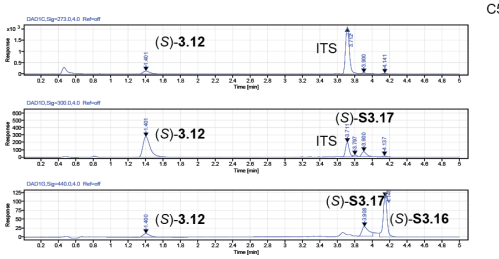
Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.395	1061.532	1.395	2309.320	1.395	81.243
3.705	6415.414	3.705	597.677	3.470	36.777
				3.815	90.169

C3-2



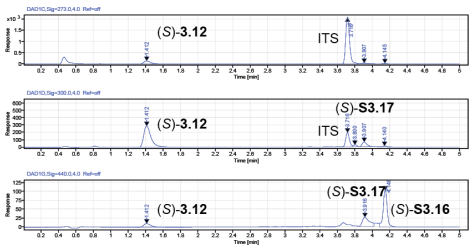
Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.411	1036.254	1.411	2251.389	1.411	79.339
3.716	6429.449	3.716	588.953	3.483	34.101
				3.824	86.344

C5-1



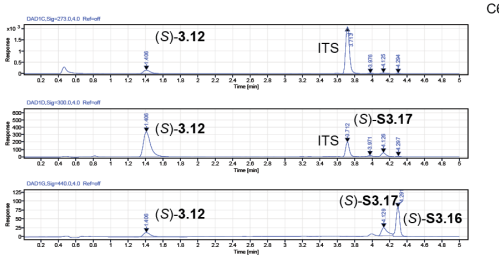
Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.401	753.112	1.401	1648.482	1.400	57.994
3.712	6407.798	3.711	570.277	3.908	147.593
3.900	127.637	3.900	265.279	4.145	362.730

C5-2



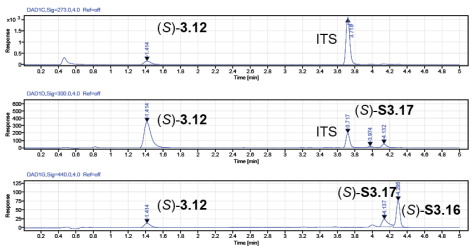
Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.412	775.724	1.412	1695.862	1.412	59.461
3.716	6455.732	3.716	576.675	3.916	143.533
3.907	105.599			4.148	333.272

C6-1



Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.406	915.571	1.406	2004.032	1.406	70.492
3.713	6414.237	3.712	597.302	4.129	109.278
				4.291	245.012

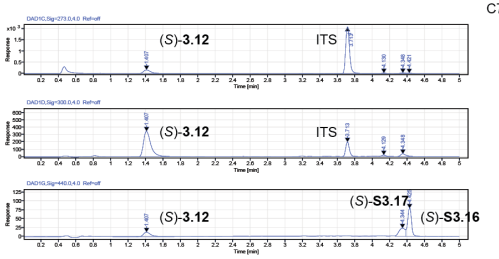
C6-2



Signal Name	DAD1C	Signal Name	DAD1D	Signal Name	DAD1G
RT (min)	Area	RT (min)	Area	RT (min)	Area
1.414	936.854	1.414	2045.644	1.414	71.916
3.718	6462.906	3.717	601.107	4.137	98.600
				4.295	226.145

Figure S3.25. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.12** with the thioesters of different chain lengths (C1, C3, C5–7, and C9) at 273, 300, and 440 nm (ITS: internal standard).

C7-1

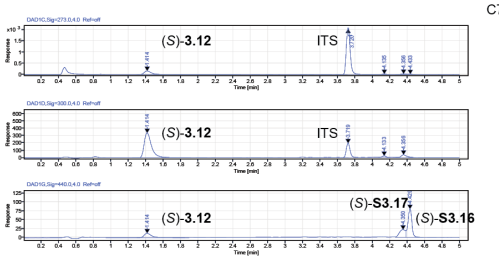


Signal Name	DAD1C
RT (min)	Area
1.407	938.889
3.713	6403.698

Signal Name	DAD1D
RT (min)	Area
1.407	2055.632
3.713	589.540

Signal Name	DAD1G
RT (min)	Area
1.407	72.163
4.344	112.962
4.425	242.070

C7-2

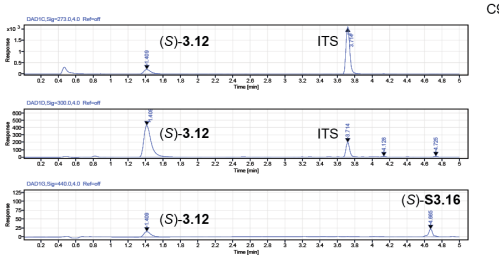


Signal Name	DAD1C
RT (min)	Area
1.414	914.413
3.720	6427.761

Signal Name	DAD1D
RT (min)	Area
1.414	1996.789
3.719	591.064

Signal Name	DAD1G
RT (min)	Area
1.414	70.173
4.350	103.312
4.429	246.827

C9-1

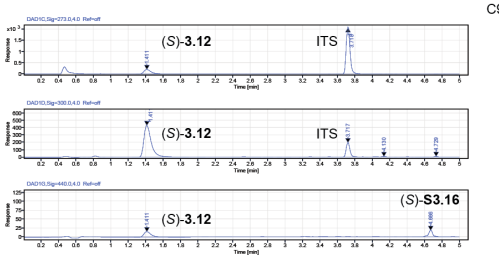


Signal Name	DAD1C
RT (min)	Area
1.409	1170.646
3.714	6423.730

Signal Name	DAD1D
RT (min)	Area
1.409	2553.676
3.714	593.772

Signal Name	DAD1G
RT (min)	Area
1.409	89.916
4.665	78.156

C9-2

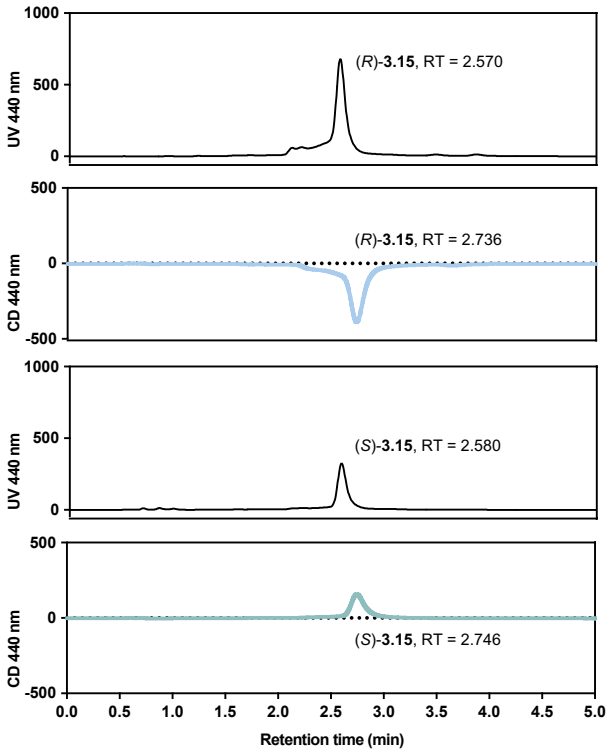


Signal Name	DAD1C
RT (min)	Area
1.411	1167.952
3.718	6424.406

Signal Name	DAD1D
RT (min)	Area
1.411	2548.745
3.717	595.008

Signal Name	DAD1G
RT (min)	Area
1.411	89.522
4.666	68.138

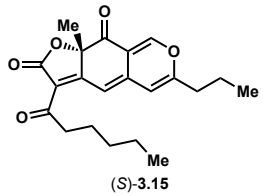
Figure 3.25 Continued. UPLC chromatograms of the samples from the enzymatic reactions of (*S*)-**3.12** with the thioesters of different chain lengths (C1, C3, C5–7, and C9) at 273, 300, and 440 nm (ITS: internal standard).



	RT (min)	$Area_{UV440}$	$Height_{CD440}$	Height/Area
(*R*)-**3.15**	2.570	5039.598	-381.231	-0.0756
(*S*)-**3.15**	2.580	2101.840	+155.922	+0.0742

Figure S3.26. UPLC chromatograms of (*R*)-**3.15** and (*S*)-**3.15** under the 440 nm wavelength of the UV and CD detectors.

Compound (*S*)-**3.15** 1H NMR (600 MHz, CDCl3)



(*S*)-**3.15**

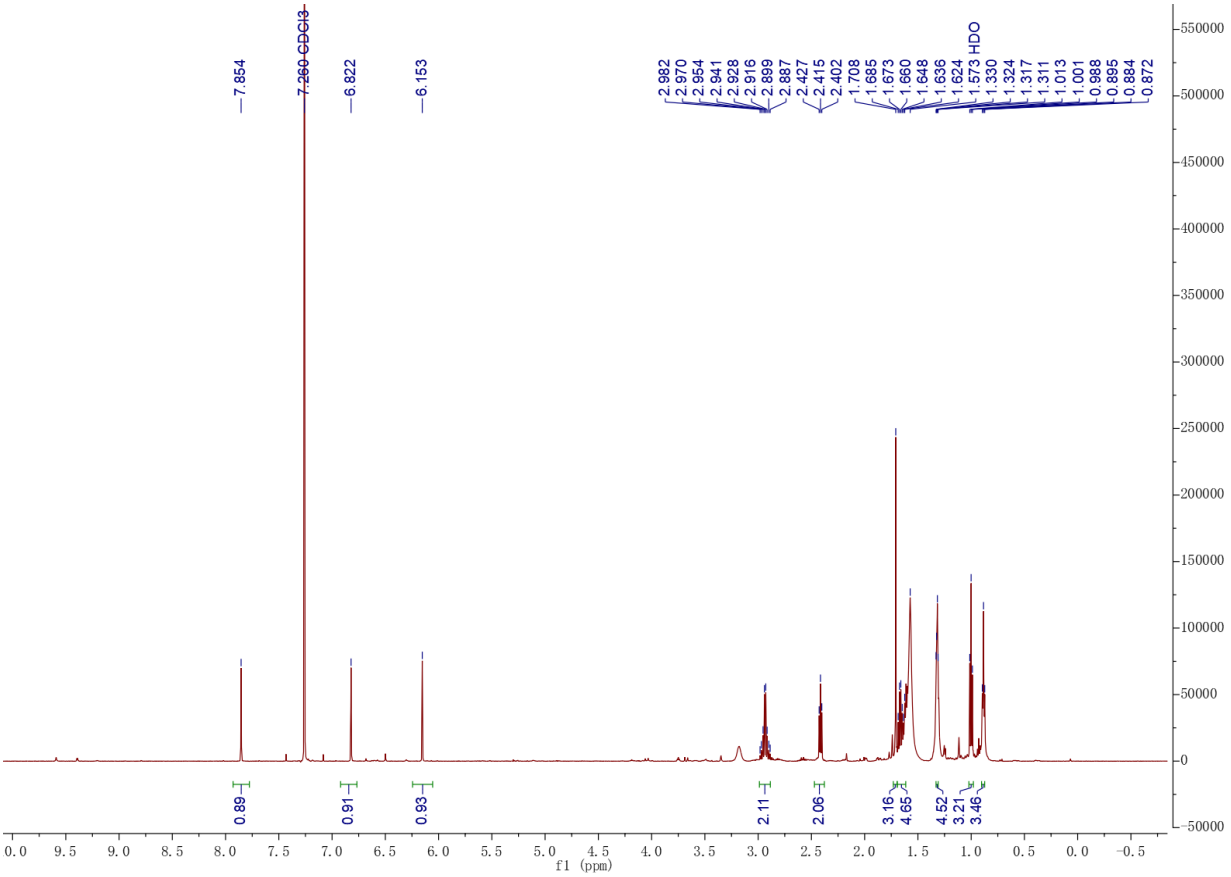


Figure S3.27. 1H NMR (600 MHz, CDCl3) spectrum of (*S*)-**3.15**.

Molecular modeling of (*R*)-**3.12**

The molecular modeling was performed by Azam Hussain. The results shown in Figure S3.28 indicate that with the current modeling approach a much smaller fraction of poses in AT4 are reactive compared to the (*S*)-enantiomer, and AT4 still appears best at capturing the substrate in a reactive geometry. Another distinction that correlates with enzyme activity is the oxygen-histidine distance where AT6 exhibits the largest average with both the (*S*)- and (*R*)-enantiomer. This potentially could help explain why the studied acyltransferases have distinct activity differences with the (*R*)-enantiomer compared to

AT6 which has no activity with (*S*)-**3.12** or (*R*)-**3.12**. Future work outside the scope of this study could potentially look at additional docking, enhanced sampling, and additional histidine tautomers to identify reactive geometries with (*R*)-**3.12**.

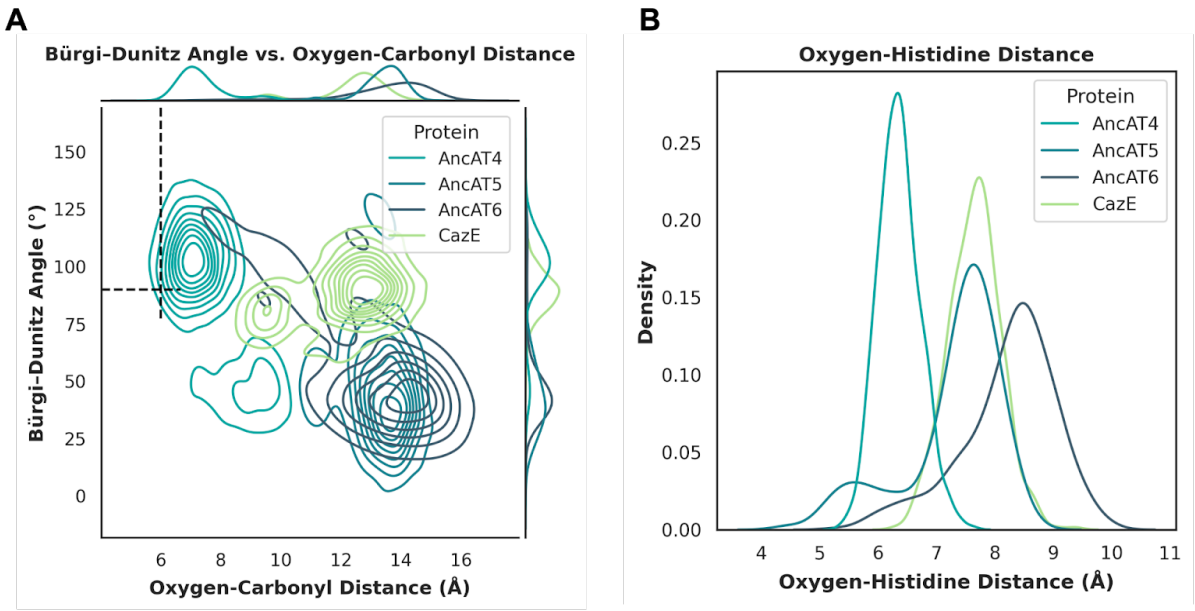


Figure S3.28. Molecular dynamics analysis of (*R*)-**3.12**.
A. Joint plot of the OH-carbonyl carbon distance of (*R*)-**3.12** with the Bürgi-Dunitz angle obtained from 40 ns simulation. The black dashed lines indicate an OH-carbonyl carbon distance of less than 6 Å and a Bürgi-Dunitz angle of greater than 90º. **B.** Plot of the histidine-OH distance from 40 ns simulation.

3.7 References

1. Huffman, M. A.; Fryszkowska, A.; Alvizo, O.; Borra-Garske, M.; Campos, K. R.; Canada, K. A.; Devine, P. N.; Duan, D.; Forstater, J. H.; Grosser, S. T., et al., Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science* **2019,** *366* (6470), 1255–1259.
2. Fryszkowska, A.; An, C.; Alvizo, O.; Banerjee, G.; Canada, K. A.; Cao, Y.; DeMong, D.; Devine, P. N.; Duan, D.; Elgart, D. M., et al., A chemoenzymatic strategy for site-selective functionalization of native peptides and proteins. *Science* **2022,** *376* (6599), 1321–1327.
3. Buller, R.; Lutz, S.; Kazlauskas, R. J.; Snajdrova, R.; Moore, J. C.; Bornscheuer, U. T., From nature to industry: Harnessing enzymes for biocatalysis. *Science* **2023,** *382* (6673), eadh8615.
4. Chen, K.; Arnold, F. H., Engineering new catalytic activities in enzymes. *Nat. Catal.* **2020,** *3* (3), 203–213.
5. Wang, Y.; Xue, P.; Cao, M.; Yu, T.; Lane, S. T.; Zhao, H., Directed Evolution: Methodologies and Applications. *Chem. Rev.* **2021,** *121* (20), 12384–12444.
6. Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H., Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (15), 5869–5874.

7. Romero, P. A.; Arnold, F. H., Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **2009,** *10* (12), 866–876.
8. Tokuriki, N.; Stricher, F.; Serrano, L.; Tawfik, D. S., How Protein Stability and New Functions Trade Off. *PLoS Comput. Biol.* **2008,** *4* (2), e1000002.
9. Sternke, M.; Tripp, K. W.; Barrick, D., Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2019,** *116* (23), 11275–11284.
10. Sumida, K. H.; Nunez-Franco, R.; Kalvet, I.; Pellock, S. J.; Wicky, B. I. M.; Milles, L. F.; Dauparas, J.; Wang, J.; Kipnis, Y.; Jameson, N., et al., Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **2024,** *146* (3), 2054–2061.
11. Bridgham, J. T.; Carroll, S. M.; Thornton, J. W., Evolution of Hormone-Receptor Complexity by Molecular Exploitation. *Science* **2006,** *312* (5770), 97–101.
12. Harms, M. J.; Thornton, J. W., Analyzing protein structure and function using ancestral gene reconstruction. *Curr. Opin. Struct. Biol.* **2010,** *20* (3), 360–366.
13. Hochberg, G. K. A.; Thornton, J. W., Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu. Rev. Biophys.* **2017,** *46*, 247–269.
14. Nicoll, C. R.; Bailleul, G.; Fiorentini, F.; Mascotti, M. L.; Fraaije, M. W.; Mattevi, A., Ancestral-sequence reconstruction unveils the structural basis of function in mammalian FMOs. *Nat. Struct. Mol. Biol.* **2020,** *27* (1), 14–24.
15. Schriever, K.; Saenz-Mendez, P.; Rudraraju, R. S.; Hendrikse, N. M.; Hudson, E. P.; Biundo, A.; Schnell, R.; Syren, P. O., Engineering of Ancestors as a Tool to Elucidate Structure, Mechanism, and Specificity of Extant Terpene Cyclase. *J. Am. Chem. Soc.* **2021,** *143* (10), 3794–3807.
16. Chiang, C. H.; Wymore, T.; Rodríguez Benítez, A.; Hussain, A.; Smith, J. L.; Brooks III, C. L.; Narayan, A. R. H., Deciphering the evolution of flavin-dependent monooxygenase stereoselectivity using ancestral sequence reconstruction. *Proc. Natl. Acad. Sci. U. S. A.* **2023,** *120* (15), e2218248120.
17. DeMars II, M. D.; O'Connor, S. E., Evolution and diversification of carboxylesterase-like [4+2] cyclases in aspidosperma and iboga alkaloid biosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **2024,** *121* (7), e2318586121.
18. Devamani, T.; Rauwerdink, A. M.; Lunzer, M.; Jones, B. J.; Mooney, J. L.; Tan, M. A.; Zhang, Z. J.; Xu, J. H.; Dean, A. M.; Kazlauskas, R. J., Catalytic Promiscuity of Ancestral Esterases and Hydroxynitrile Lyases. *J. Am. Chem. Soc.* **2016,** *138* (3), 1046–1056.
19. Zou, T.; Risso, V. A.; Gavira, J. A.; Sanchez-Ruiz, J. M.; Ozkan, S. B., Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Mol. Biol. Evol.* **2015,** *32* (1), 132–143.
20. Wheeler, L. C.; Lim, S. A.; Marqusee, S.; Harms, M. J., The thermostability and specificity of ancient proteins. *Curr. Opin. Struct. Biol.* **2016,** *38*, 37–43.
21. Risso, V. A.; Sanchez-Ruiz, J. M.; Ozkan, S. B., Biotechnological and protein-engineering implications of ancestral protein resurrection. *Curr. Opin. Struct. Biol.* **2018,** *51*, 106–115.
22. Gumulya, Y.; Baek, J. M.; Wun, S. J.; Thomson, R. E. S.; Harris, K. L.; Hunter, D. J. B.; Behrendorff, J. B. Y. H.; Kulig, J.; Zheng, S.; Wu, X. M., et al., Engineering highly

functional thermostable proteins using ancestral sequence reconstruction. *Nat. Catal.* **2018,** *1* (11), 878–888.
23. Livada, J.; Vargas, A. M.; Martinez, C. A.; Lewis, R. D., Ancestral Sequence Reconstruction Enhances Gene Mining Efforts for Industrial Ene Reductases by Expanding Enzyme Panels with Thermostable Catalysts. *ACS Catal.* **2023,** *13* (4), 2576–2585.
24. Jones, B. S.; Ross, C. M.; Foley, G.; Pozhydaieva, N.; Sharratt, J. W.; Kress, N.; Seibt, L. S.; Thomson, R. E. S.; Gumulya, Y.; Hayes, M. A., et al., Engineering Biocatalysts for the C–H Activation of Fatty Acids by Ancestral Sequence Reconstruction. *Angew. Chem. Int. Ed. Engl.* **2024,** *63* (18), e202314869.
25. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L., NCBI BLAST: a better web interface. *Nucleic Acids Res.* **2008,** *36* (Web Server issue), W5–9.
26. Zallot, R.; Oberg, N.; Gerlt, J. A., The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019,** *58* (41), 4169–4182.
27. Oberg, N.; Zallot, R.; Gerlt, J. A., EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J. Mol. Biol.* **2023,** *435* (14), 168018.
28. Gao, J. M.; Yang, S. X.; Qin, J. C., Azaphilones: Chemistry and Biology. *Chem. Rev.* **2013,** *113* (7), 4755–4811.
29. Chen, C.; Tao, H.; Chen, W.; Yang, B.; Zhou, X.; Luo, X.; Liu, Y., Recent advances in the chemistry and biology of azaphilones. *RSC Adv.* **2020,** *10* (17), 10197–10220.
30. Williams, K.; Greco, C.; Bailey, A. M.; Willis, C. L., Core Steps to the Azaphilone Family of Fungal Natural Products. *ChemBioChem* **2021,** *22* (21), 3027–3036.
31. Chen, W.; Chen, R.; Liu, Q.; He, Y.; He, K.; Ding, X.; Kang, L.; Guo, X.; Xie, N.; Zhou, Y., et al., Orange, red, yellow: biosynthesis of azaphilone pigments in Monascus fungi. *Chem. Sci.* **2017,** *8* (7), 4917–4925.
32. Wang, Y.; Torma, K. J.; Pyser, J. B.; Zimmerman, P. M.; Narayan, A. R. H., Substrate-Selective Catalysis Enabled Synthesis of Azaphilone Natural Products. *ACS Cent Sci* **2024,** *10* (3), 708–716.
33. Fielding, B. C.; Holker, J. S. E.; Jones, D. F.; Powell, A. D. G.; Richmond, K. W.; Robertson, A.; Whalley, W. B., 898. The chemistry of fungi. Part XXXIX. The structure of monascin. *J. Chem. Soc.* **1961**, 4579–4589.
34. Zabala, A. O.; Xu, W.; Chooi, Y. H.; Tang, Y., Characterization of a silent azaphilone gene cluster from Aspergillus niger ATCC 1015 reveals a hydroxylation-mediated pyran-ring formation. *Chem. Biol.* **2012,** *19* (8), 1049–1059.
35. Chiang, Y. M.; Szewczyk, E.; Davidson, A. D.; Keller, N.; Oakley, B. R.; Wang, C. C., A Gene Cluster Containing Two Fungal Polyketide Synthases Encodes the Biosynthetic Pathway for a Polyketide, Asperfuranone, in *Aspergillus nidulans*. *J. Am. Chem. Soc.* **2009,** *131* (8), 2965–2970.
36. Pyser, J. B.; Dockrey, S. A. B.; Benítez, A. R.; Joyce, L. A.; Wiscons, R. A.; Smith, J. L.; Narayan, A. R. H., Stereodivergent, Chemoenzymatic Synthesis of Azaphilone Natural Products. *J. Am. Chem. Soc.* **2019,** *141* (46), 18551–18559.
37. Chiang, C. H.; Wymore, T.; Rodriguez Benitez, A.; Hussain, A.; Smith, J. L.; Brooks III, C. L.; Narayan, A. R. H., Deciphering the evolution of flavin-dependent

monooxygenase stereoselectivity using ancestral sequence reconstruction. *Proc. Natl. Acad. Sci. U. S. A.* **2023,** *120* (15), e2218248120.
38. Fischbach, M. A.; Walsh, C. T.; Clardy, J., The evolution of gene collectives: How natural selection drives chemical innovation. *Proc. Natl. Acad. Sci. U. S. A.* **2008,** *105* (12), 4601–4608.
39. Winter, J. M.; Sato, M.; Sugimoto, S.; Chiou, G.; Garg, N. K.; Tang, Y.; Watanabe, K., Identification and Characterization of the Chaetoviridin and Chaetomugilin Gene Cluster in Chaetomium globosum Reveal Dual Functions of an Iterative Highly-Reducing Polyketide Synthase. *J. Am. Chem. Soc.* **2012,** *134* (43), 17900–17903.
40. Vinh, N. X.; Epps, J.; Bailey, J., Information Theoretic Measures for Clusterings Comparison: is a Correction for Chance Necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, Association for Computing Machinery: Montreal, Quebec, Canada, 2009; pp 1073–1080.
41. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997,** *25* (17), 3389–3402.
42. Chen, K.; Arnold, F. H., Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc. Natl. Acad. Sci. U. S. A.* **1993,** *90* (12), 5618–5622.
43. Arnold, F. H., Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem. Int. Ed. Engl.* **2018,** *57* (16), 4143–4148.
44. Wang, T.; Badran, A. H.; Huang, T. P.; Liu, D. R., Continuous directed evolution of proteins with improved soluble expression. *Nat. Chem. Biol.* **2018,** *14* (10), 972–980.
45. Risso, V. A.; Gavira, J. A.; Mejia-Carmona, D. F.; Gaucher, E. A.; Sanchez-Ruiz, J. M., Hyperstability and Substrate Promiscuity in Laboratory Resurrections of Precambrian Beta-Lactamases. *J. Am. Chem. Soc.* **2013,** *135* (8), 2899–2902.
46. Nicoll, C. R.; Massari, M.; Fraaije, M. W.; Mascotti, M. L.; Mattevi, A., Impact of ancestral sequence reconstruction on mechanistic and structural enzymology. *Curr. Opin. Struct. Biol.* **2023,** *82*, 102669.
47. Chaloupkova, R.; Liskova, V.; Toul, M.; Markova, K.; Sebestova, E.; Hernychova, L.; Marek, M.; Pinto, G. P.; Pluskal, D.; Waterman, J., et al., Light-Emitting Dehalogenases: Reconstruction of Multifunctional Biocatalysts. *ACS Catal.* **2019,** *9* (6), 4810–4823.
48. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Zidek, A.; Potapenko, A., et al., Highly accurate protein structure prediction with AlphaFold. *Nature* **2021,** *596* (7873), 583–589.
49. Unno, H.; Ichimaida, F.; Suzuki, H.; Takahashi, S.; Tanaka, Y.; Saito, A.; Nishino, T.; Kusunoki, M.; Nakayama, T., Structural and Mutational Studies of Anthocyanin Malonyltransferases Establish the Features of BAHD Enzyme Catalysis. *J. Biol. Chem.* **2007,** *282* (21), 15812–15822.
50. Buckner, J.; Liu, X.; Chakravorty, A.; Wu, Y.; Cervantes, L. F.; Lai, T. T.; Brooks III, C. L., pyCHARMM: Embedding CHARMM Functionality in a Python Framework. *J. Chem. Theory Comput.* **2023,** *19* (12), 3752–3762.
51. Gagnon, J. K.; Law, S. M.; Brooks III, C. L., Flexible CDOCKER: Development and application of a pseudo-explicit structure-based docking method within CHARMM. *J. Comput. Chem.* **2016,** *37* (8), 753–762.

52. Wu, Y.; Brooks III, C. L., Flexible CDOCKER: Hybrid Searching Algorithm and Scoring Function with Side Chain Conformational Entropy. *J. Chem. Inf. Model.* **2021,** *61* (11), 5535–5549.
53. Bürgi, H. B.; Dunitz, J. D.; Lehn, J. M.; Wipff, G., Stereochemistry of reaction paths at carbonyl centres. *Tetrahedron* **1974,** *30* (12), 1563–1572.
54. Panina, I. S.; Krylov, N. A.; Chugunov, A. O.; Efremov, R. G.; Kordyukova, L. V., The Mechanism of Selective Recognition of Lipid Substrate by hDHHC20 Enzyme. *Int. J. Mol. Sci.* **2022,** *23* (23), 14791.
55. Arora, B.; Mukherjee, J.; Gupta, M. N., Enzyme promiscuity: using the dark side of enzyme specificity in white biotechnology. *Sustain. Chem. Process.* **2014,** *2* (1), 25.
56. Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Mc, Q. G. S.; Roodveldt, C.; Tawfik, D. S., The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **2005,** *37* (1), 73–76.
57. Leveson-Gower, R. B.; Mayer, C.; Roelfes, G., The importance of catalytic promiscuity for enzyme design and evolution. *Nat. Rev. Chem.* **2019,** *3* (12), 687–705.
58. Champagne, S. E.; Chiang, C.-H.; Gemmel, P. M.; Brooks III, C. L.; Narayan, A. R. H., Biocatalytic Stereoselective Oxidation of 2-Arylindoles. *J. Am. Chem. Soc.* **2024,** *146* (4), 2728–2735.
59. Gumulya, Y.; Gillam, E. M., Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. *Biochem. J.* **2017,** *474* (1), 1–19.
60. Gilchrist, C. L. M.; Chooi, Y. H., clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics* **2021,** *37* (16), 2473–2475.
61. Vinh, N. X. E., J.; Bailey, J., Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *J. Mach. Learn. Res.* **2010,** *11*, 2837–2854.
62. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W., CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **2010,** *26* (5), 680–682.
63. Edgar, R. C., Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **2022,** *13* (1), 6968.
64. Larsson, A., AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **2014,** *30* (22), 3276–3278.
65. Capella-Gutierrez, S.; Silla-Martinez, J. M.; Gabaldon, T., trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009,** *25* (15), 1972–1973.
66. Edler, D.; Klein, J.; Antonelli, A.; Silvestro, D.; Matschiner, M., raxmlGUI 2.0: A graphical interface and toolkit for phylogenetic analyses using RAxML. London: Wiley Subscription Services, Inc: 2021; Vol. 12.
67. Le, S. Q.; Gascuel, O., An Improved General Amino Acid Replacement Matrix. *Mol. Biol. Evol.* **2008,** *25* (7), 1307–1320.
68. Hirotogu, A., Information Theory and an Extension of the Maximum Likelihood Principle. New York, NY: Springer New York: 1973.
69. Schwarz, G., Estimating the Dimension of a Model. *Ann. Stat.* **1978,** *6* (2), 461–464.

70. Guindon, S.; Dufayard, J. F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O., New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst. Biol.* **2010,** *59* (3), 307–321.
71. Yang, Z., PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **2007,** *24* (8), 1586–1591.
72. Hanson-Smith, V.; Kolaczkowski, B.; Thornton, J. W., Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol. Biol. Evol.* **2010,** *27* (9), 1988–1999.
73. Zhang, Y.; Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **2005,** *33* (7), 2302–2309.
74. Brooks, B. R.; Brooks III, C. L.; Mackerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S., et al., CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009,** *30* (10), 1545–1614.
75. Feig, M.; Karanicolas, J.; Brooks III, C. L., MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. *J. Mol. Graph. Model.* **2004,** *22* (5), 377–395.
76. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., Jr., CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017,** *14* (1), 71–73.
77. Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R., Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **2012,** *4* (1), 17.
78. Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A., Development and testing of a general amber force field. *J. Comput. Chem.* **2004,** *25* (9), 1157–1174.
79. Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I., et al., CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010,** *31* (4), 671–690.
80. Ding, X.; Wu, Y.; Wang, Y.; Vilseck, J. Z.; Brooks III, C. L., Accelerated CDOCKER with GPUs, Parallel Simulated Annealing, and Fast Fourier Transforms. *J. Chem. Theory Comput.* **2020,** *16* (6), 3910–3919.
81. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *J. Cheminform.* **2011,** *3*, 33.
82. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D., et al., OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017,** *13* (7), e1005659.
83. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O., MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **2011,** *32* (10), 2319–2327.

Chapter 4 Leveraging Latent Space Models to Navigate Enzyme Sequence and Function Space

4.1 Summary

The exponential growth in available protein sequence data has broadened enzyme discovery opportunities but simultaneously highlighted a significant gap between sequence and function information. Traditional tools like phylogenetic trees and sequence similarity networks (SSNs) are widely adopted for sampling enzymes for novel transformations. However, their utility suffers from inherent limitations, which are exacerbated for large enzyme families. Phylogenetic trees, while useful for studying evolutionary relationships, become computationally intensive and difficult to visualize for larger protein datasets. SSNs, on the other hand, are sensitive to user-defined thresholds for sequence clustering and easily fail to capture more distant relationships between clusters. Additionally, both tools are alignment based and cannot capture higher-order interactions between residues. In this study, we address these limitations by optimizing a variational autoencoder (VAE)-based latent space model to visualize and explore enzyme sequence-function landscapes. By training our models on simulated datasets and real enzyme families, such as cyclases and flavin-dependent monooxygenases (FDMOs), we demonstrated that the optimized latent space effectively preserves phylogenetic relationships and enables high-resolution clustering for functionally distinct enzymes. The models further outperform traditional SSNs in capturing local and global relationships in a continuous two-dimensional space, enabling the discovery of multiple uncharacterized FDMOs for oxidative dearomatization and decarboxylative hydroxylation that illustrates their application. Our findings show that low-dimensional latent spaces can serve as valuable tools for enzyme discovery, allowing for interpolation and extrapolation to guide novel enzyme sampling for biocatalytic reactions.

4.2 Introduction

The increasing number of protein sequences in databases such as UniProt,[1] Pfam,[2] InterPro,[3] and the National Center for Biotechnology Information (NCBI)[4] provides scientists with access to an astonishing number of enzyme sequences for novel functions in the lab. For example, there are approximately 250 million sequence entries in the UniProt Knowledgebase (UniProtKB) as of February 2024.[5] However, only 0.22% of protein sequences have been manually reviewed and annotated by experts in UniProtKB,[5] and this percentage will become lower as the number of sequences grows exponentially. This highlights a widening gap between sequence and function information, underscoring the need for tools to explore the protein sequence space and facilitate the understanding of sequence-function relationships based on available data.

Phylogenetic analysis is an essential tool for investigating protein sequence-function relationships, commonly used by bioinformaticians and evolutionary biologists. Phylogenetic trees are constructed from a set of aligned homologous sequences obtained through multiple sequence alignment (MSA), often with a chosen substitution model (Figure 4.1).[6] Phylogeny can then be used to infer sequence-function relationships. However, visualizing trees can be challenging,[7-8] and high-quality phylogenetic analysis requires extensive curation and is computationally demanding especially for larger protein dataset.[9-10]

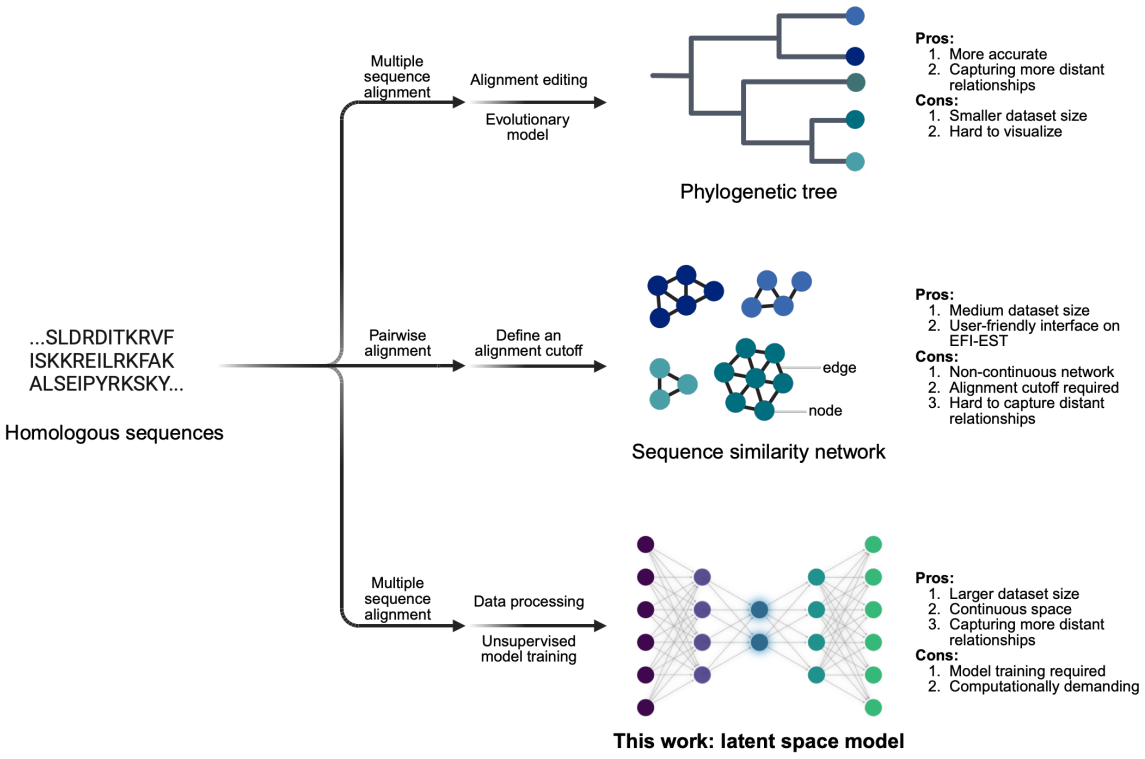


Figure 4.1. Comparison of three tools for enzyme discovery and sampling, including phylogenetic trees, sequence similarity networks (SSNs), and latent space models.

Sequence similarity networks (SSNs), as an alternative, have been developed to display pairwise sequence similarity relationships among homologous proteins.[11] SSN generation is typically performed on the EFI-EST website,[12] beginning with pairwise sequence alignment, followed by the selection of a user-defined alignment score (AS), and visualization in Cytoscape (Figure 4.1).[12-13] Each SSN comprises multiple clusters, with nodes representing enzymes and edges indicating relatedness between the nodes within each cluster. The number of clusters is determined by the user determined AS for "isofunctional" pairs of sequences. A higher threshold results in more putatively isofunctional clusters, while a lower threshold causes all enzyme nodes to merge into one giant cluster. SSNs are renowned for their user-friendliness and easy integration with downstream tools like genome neighborhood analysis on EFI-GNT.[14] However, SSNs have several intrinsic limitations. The primary limitation is that the SSN is constructed based on pairwise identity alignment and no information regarding higher-order residue interactions are considered. Another key limitation is the selection of an appropriate

alignment threshold to define "isofunctional," which directly impacts subsequent hypothesis formation and the sampling of uncharacterized enzymes. Although an AS corresponding to 35–40% sequence identity is generally recommended for the initial SSN, the selection of the score is very case-dependent and sometimes depends on whether there are enough annotations to guide the selection. Also limiting is that relationships between clusters only exist when the clusters are connected by edges; once separated, these relationships are fully lost, limiting the SSNs' ability to capture distant relationships between sequences. Collectively, the limitations associated with phylogenetic trees or SSNs led us to consider if a superior tool can be developed to visualize/study sequence-function relationships of larger enzyme families.

Representation (feature) learning is a burgeoning field in machine learning.[15] The purpose of representation learning is to extract meaningful data representations as numerical vectors, which can then be used in downstream machine learning tasks such as regression, classification, and clustering. Learning the protein sequence-function relationship is one of many applications of representation learning. Classic examples include variational autoencoders for predicting mutational effects[16] and deciphering protein evolution,[17-19] a recurrent neural network (RNN), UniRep, for improving efficiency in protein engineering,[20] a Transformer model, ESM-1b, for various biological applications,[21] and the state-of-the-art model like ProtTrans.[22] Most contemporary representation learning models were trained on extremely large datasets with millions of cross-family sequences and the learned representations are usually presented as vectors in high dimensions (5700 for UniRep, 1280 for ESM-1b, and 1024 for ProtT5-XL).[23] Those large protein language models (PLMs) generally show good performance on downstream tasks; however, they require further dimensional reduction for visualization. In addition, there is little to no pattern after the high-dimensional representations is projected into lower and visualizable dimensions, making them often difficult to interpret for humans and less suitable for visualizing sequence-function relationships.[18] In contrast, the two-dimensional (2D) representations acquired through training VAE models for single protein families have been demonstrated to capture phylogeny.[17-19] As funcational similarity between enzymes can usually be inferred according to phylogenetic similarity, the

capability of capturing phylogeny within large enzyme families shows great potential in visualizing sequence-function relationships.

VAE model architectures comprise an encoder and a decoder (Figure 4.2A). The encoder maps input data x_i to a Gaussian distribution over latent variable z_i, from which z_i is sampled and passed through the decoder to output $\hat{x}_i$ to reconstruct the input. The encoder and decoder are coupled and are parameterized together during model training. In our case, the training process forces the model to learn meaningful representations of input protein sequences in a lower-dimensional and continuous latent space for the decoder to better reconstruct the input sequence data. Several features of VAEs make them suitable for learning protein representations for visualization. First, VAEs capture high-order interactions among residues within protein sequences, which play an important role in epistasis, yet are hard to capture/not considered within traditional alignment-based methods. Secondly, the latent space is continuous, enabling users to easily relate two proteins according to the Euclidean distance of their latent variables. Such a concept can be expanded and applied in novel enzyme discovery to infer/predict the function or substrate of uncharacterized enzymes by operations such as interpolation and extrapolation in the latent space.[17] Taken together, we suggest that the low-dimensional representations from VAEs can be used as an alternative to either phylogenetic trees or SSNs for the visualization of sequence-function relationships of large enzyme familes.

In this study, we explore the potential of developing a VAE-trained latent space model as a tool for enzyme discovery and sampling. Specifically, we hypothesized that the specific phylogeny-correlated distribution of 2D protein representations and the continuity of the model could facilitate the understanding of local and global sequence-function relationships, which are beneficial in novel enzyme discovery for biocatalytic purposes. To explore this hypothesis, we built a simulated protein dataset for model optimization, performing a side-by-side comparison between the latent space model and a constructed SSN derived from these sequences (Figure 4.2B). We demonstrate the performance of the optimized latent space model using the dataset of cyclases in PF05147 and compare it with state-of-the-art protein language models on a downstream classification task. Finally, leveraging the continuity of the latent space, we apply our optizmized model design to demonstrate reactivity trends of several uncharacterized

flavin-dependent monooxygenases (FDMOs) for decarboxylative hydroxylation and oxidative dearomatization in PF01494. In summary, we illustrate that an optimized latent space model shows significant promise in applications that include visualizing seqeunce-function relationships and guiding the discovery of novel enzymes.

4.3 Results and Discussion

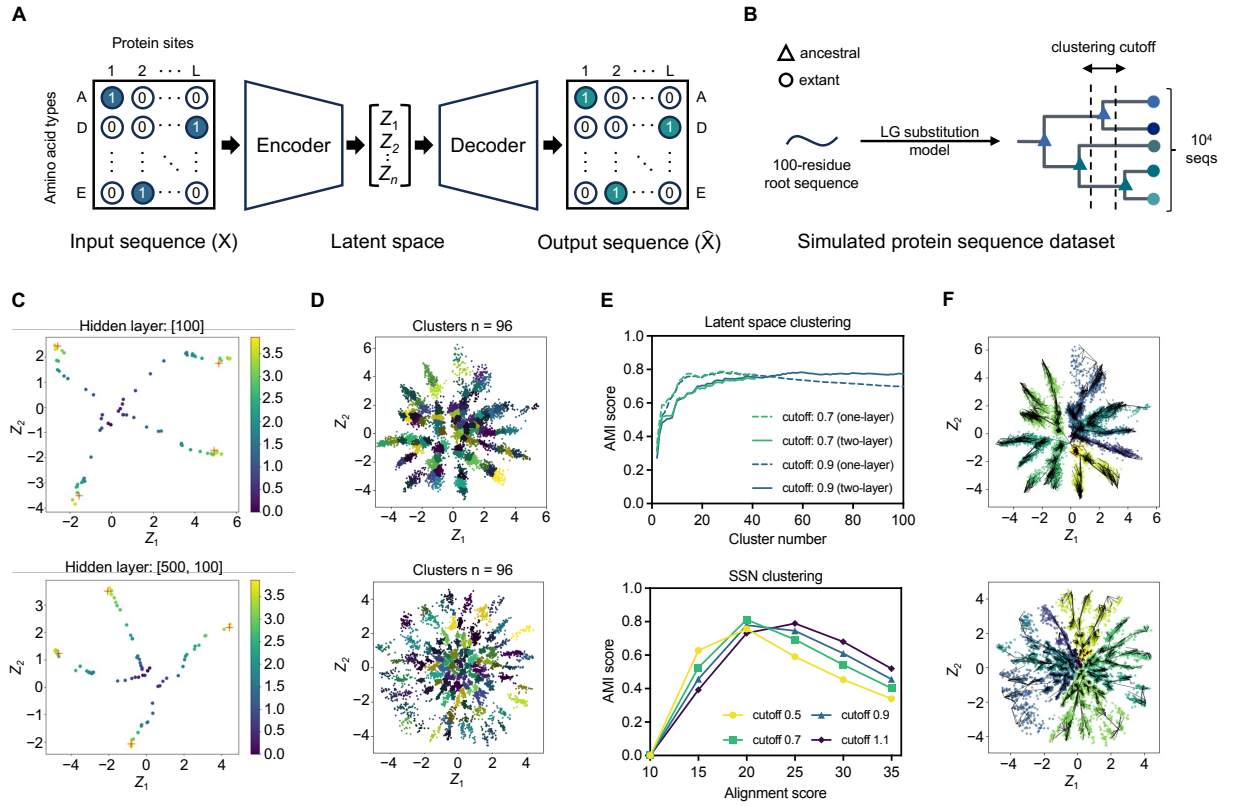


Figure 4.2. Latent space models of sequences from a simulated protein sequence dataset.
A. Latent space models trained by a variational autoencoder (VAE). **B.** Generation of a simulated protein sequence dataset by feeding a 100-residue root sequence to a simulated tree with a substitution model. **C.** Latent space representations of four leaf sequences and their ancestors. **D.** Hierarchical clustering (n = 96) of latent space representations. **E.** Adjusted mutual information (AMI) scores between the latent space-based clustering or sequence similarity network (SSN)-based clustering with the tree-based clustering as ground truth. **F.** Linkages of latent space representations of leaf sequences by their representations of major ancestors.

Optimizing latent space models using a simulated protein sequence dataset. Latent space models are known to capture phylogenetic relationships among proteins, a feature that suggests potential for further optimization as tools in enzyme discovery and sampling. Initial model design mostly with a single hidden layer in both the encoder and decoder has been shown to encode proteins into a 2D latent space with a star-like distribution

(Figures 4.2C and 4.2D).[17] In this distribution, major protein clusters within a family were separated by spikes radiating from a central point; however, proteins within the same clusters were not well-differentiated. This original model structure is effective for distinguishing major groups but lacks the granularity needed for applications such as enzyme discovery and sampling that require differentiating closely related proteins within clusters (see Figure 4.2D).

To address this limitation, we increased the number of hidden layers to enhance the model's capacity for nuanced separation within clusters (Table S4.4). We created a simulated dataset containing 10,000 ancestral and extant sequences, each 100 residues long, by feeding a root sequence through a LG evolutionary model (Figure 4.2B).[24] Accordingly, a simulated rooted phylogenetic tree was generated in the process, with clustering labels derived from the various branches of the tree providing a ground truth for evaluating model performance (Figure 4.2B). Two versions of the latent space model—one with a single hidden layer and another with two hidden layers—were trained on 8,000 extant sequences, with the remaining 2,000 sequences reserved for validation. Both models successfully preserved phylogenetic relationships, as demonstrated by embedding ancestral proteins and color-coding each by its evolutionary distance from the root. This visual inspection confirmed that both models effectively captured evolutionary relationships, with ancestral proteins positioned according to their distance from the root (Figure 4.2C).

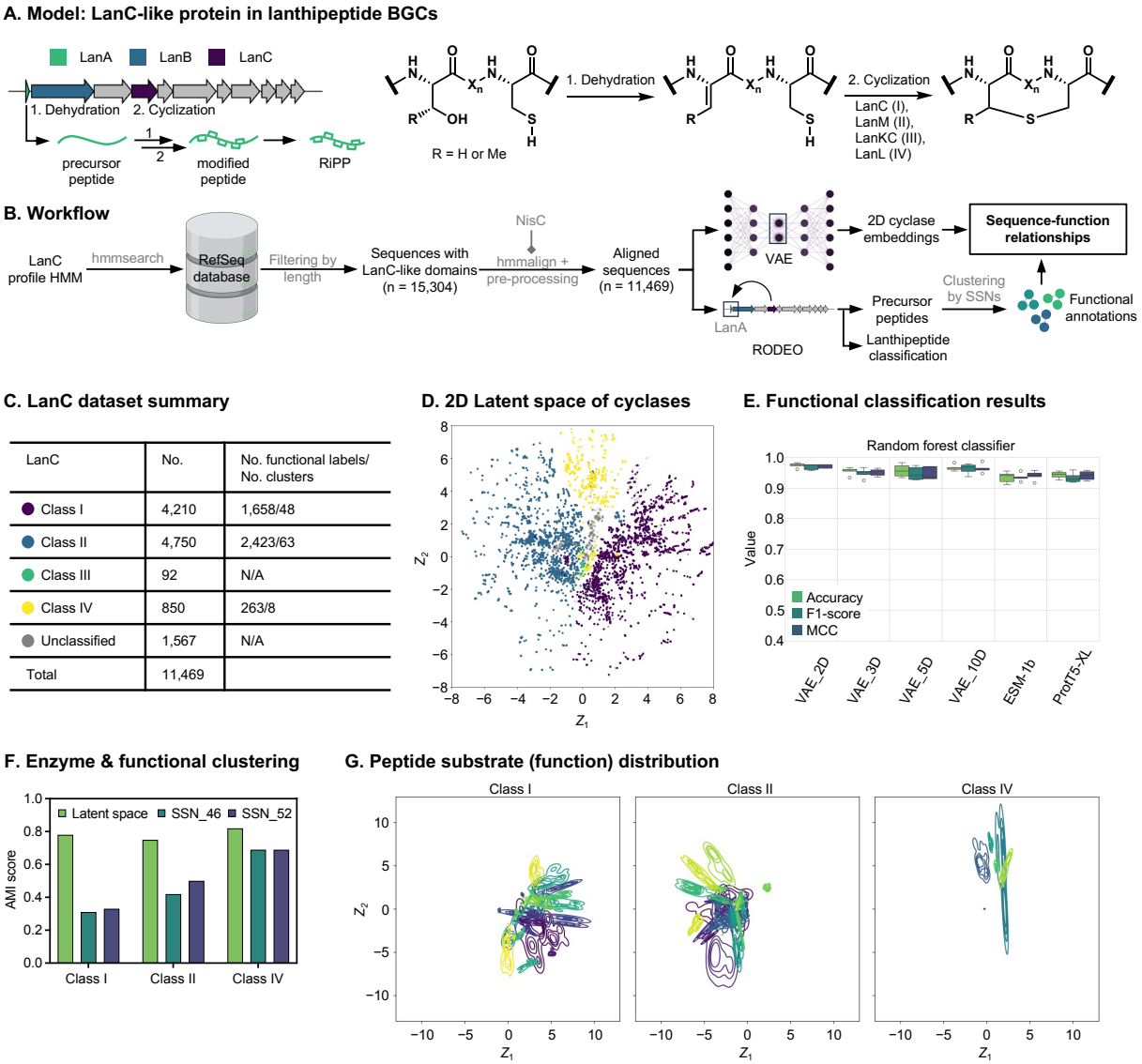
We compared the ability of the model to accurately group the protein sequences into clusters. Using the phylogenetic tree clustered at an evolutionary distance cutoff of 0.9 (where 90% of residues differ), we identified 100 clusters in the tree to serve as true labels. Proteins in each latent space model were also clustered hierarchically into 100 groups (Figure 4.2D), and their clustering labels were compared against the true tree-based labels using adjusted mutual information (AMI) scores, which measure the alignment between two sets of clusters.[25] The one-layer model achieved an AMI score of 0.70, while the two-layer model scored 0.77, suggesting that the two-layer model produced clusters more consistent with the true phylogenetic structure. To further assess the models, we repeated the tree-based clustering at a lower resolution with a 0.7 cutoff, yielding 41 tree clusters. We found that the one-layer model AMI score peaked when

assigned with 28 clusters, after which performance declined, whereas the two-layer model maximized performance at 59 clusters when using the 0.9 cutoff (Figure 4.2E). This suggests that the two-layer model supports higher-resolution clustering, effectively differentiating more groups within the protein family.

For benchmarking purposes, we compared the latent space models to SSNs, clustering sequences across alignment scores of 10–35 and comparing these clusters to the tree-based clusters at various cutoffs (0.5–1.1). The highest AMI score obtained by the SSNs was around 0.8 at an AS of 20, which corresponds to 35–40% sequence identity (Figure 4.2E). However, unlike the latent space models, the SSN clusters were entirely separate without connecting edges (Figure S4.1), losing distant phylogenetic relationships in the process. In contrast, the latent space models kept continuity between related clusters through phylogenetic linkages back to the root sequence, underscoring their advantage in preserving both local and global relationships within the protein family (Figure 4.2F).

Testing latent space models on a unique cyclase-lanthipeptide dataset. Latent space models trained with multiple hidden layers and optimized hyperparameters showed significant potential to be used as a novel enzyme sampling tool based on our hypothesis that functional relationships can be inferred by phylogenetic relationships. To further explore this idea, we trained a model on a real enzyme family. Functional labels are sparse among most enzyme families. However, we took inspiration from previous studies of ribosomally synthesized and post-translationally modified peptides (RiPPs), and curated our own dataset consisting of cyclases from the protein family of Lanthionine synthetase C-like (LANC_like, PF05147), which are responsible for side-chain cyclization of peptides through the formation of thioester bridges (Figure 4.3A).[26] The uniqueness of this dataset is that the peptide substrate of each cyclase can be predicted by bioinformatic tools such as Rapid ORF Description and Evaluation Online (RODEO).[27] Unlike non-peptide and small-molecule substrates, peptide substrates can be easily compared and clustered by alignment-based methods like SSNs, providing valuable functional annotations of each corresponding cyclase in the same gene cluster. Lanthipeptides can be divided into four major classes according to their biosynthetic enzymes.[26,28-30] In class

I, two separate enzymes, Lan B and LanC,[31-33] are responsible for dehydration and cyclization of precursor peptides respectively, while in class II–IV, a single enzyme performs both steps by different domains and the C-terminal domains (LanM, LanKC, LanL) are for cyclization.[34-38]



LanC	No.	No. functional labels/ No. clusters
Class I	4,210	1,658/48
Class II	4,750	2,423/63
Class III	92	N/A
Class IV	850	263/8
Unclassified	1,567	N/A
Total	11,469	

Figure 4.3. Latent space models for a family of cyclases responsible for cyclization of lanthipeptides.
A. Biosynthetic gene cluster (BGC) of a LanC cyclase gene (LanC) and its corresponding LanA substrate gene. The modification of lanthipeptides undergoes two major steps: dehydration and cyclization. **B.** Preparation of a cyclase dataset with putative peptide substrates predicted using RODEO analysis. **C.** Latent space representations of cyclases. **D.** LanC dataset with four major classes. **E.** Evaluation of using different embeddings as inputs for a downstream classification task by a random forest classifier. **F.** Comparison between the latent space-based clustering and SSN-based clustering with two alignment scores (46 and 52) for the cyclases in Class I, II, and IV. **G.** Contours of the latent space representations of cyclases clustered and color coded by their clustering of precursor peptide substrates (functional labels).

To prepare a cyclase dataset for training a latent space model, potential sequences were identified by searching the non-redundant RefSeq database from NCBI (October 2023) with the profile hidden Markov model (HMM) of PF05147 using HMMER (Figure 4.3B).[39] This search resulted in 15,304 sequences within the range of 200–500 residues with LanC-like domains. The sequences within the preliminary dataset were then aligned and pre-processed using a well-studied cyclase, NisC (WP_043991124.1/31-402), as query to filter out sequences with >20% gaps, reducing the dataset to contain 11,469 sequences for the downstream model training and analysis (Figure 4.3B). The sequences in our dataset were classified by first comparing their IDs with the dataset reported by Mitchell and van der Donk.[40] The remaining sequences not included in the previous dataset were submitted to RODEO for analysis and further classified by the same criteria applied by Mitchell and van der Donk.[40] The classification resulted in 4,210 putative class I cyclases, 4,750 class II cyclases, 92 class III cyclases, 850 class IV cyclases, and 1,567 unclassified sequences (Figure 4.3C). Comparing with the previous dataset reported, we found significantly fewer class III cyclases in our dataset: most likely because the LanKC cyclization domain of class III cyclases lacks the zinc-binding residues, conserved in other three classes, and were excluded in the pre-processing step using the class-I NisC as the query. In summary, our cyclase dataset predominantly consists of sequences involved in the production of class I and II lanthipeptides, some sequences associated with class IV, very few sequences for class III lanthipeptides, and a number of uncharacterized sequences.

From this dataset we built a 2D latent space model for cyclases by training a VAE with two hidden layers in its encoder and decoder. The VAE was trained using five-fold cross-validation and the best model with the lowest score for the validation dataset was selected for downstream analysis (Table S4.4). We first tested our latent space model to see if different classes of cyclases can be differentiated in the space. In Figure 4.3D, enzymes labeled as Class I–IV are shown in different colors illustrating visually that sequences from different classes are distributed in separate regions in the space. This observation aligns with conclusions from a previous study that LanM (class II) and LanL (class IV) evolved independently from LanC (class I), and that the divergence likely only happened once over the evolution.[41]

To further test whether the latent space model could generate meaningful representations for downstream classification tasks, the precursor peptides of each putative cyclase from class I, II, and IV were predicted by RODEO and were clustered separately based on classes using SSNs (Figures 4.3B and 4.3C). By having the clustering results as the functional annotations of cyclases, six classic classifier methodologies including logistic regression, support vector classifier (SVC), naïve Bayes, k-nearest neighbors (kNN), decision tree, and random forest, were trained using the cyclase embeddings as inputs.[42] A five-fold cross validation was performed, and the classifiers were evaluated by accuracy, F1-score, and Matthew's correlation coefficient (MCC) scores between −1 and +1 (+1: perfect prediction, 0: random prediction, and −1: inverse prediction). Out of six the classifiers tested, random forest, decision tree, and kNN produced the best models for class-I cyclase classification with MCC scores of 0.97, 0.95, and 0.95, respectively (Figure S4.7). Similar trends were seen for class II and IV (Figures S4.8 and S4.9). We were curious if increasing the dimensionality of the latent space could further improve the performance of the classifiers. Under the same architecture, three-, five-, and ten-dimensional models were trained (Tables S4.1–S4.4). We found that the classifiers based on logistic regression and SVC showed significant improvement in MCC while the others had equal or slightly smaller MCC. At the same time, we embedded the cyclases into representations of 1280 and 1024 dimensions, by using the PLMs, ESM-1b[21] and ProtT5-XL,[22] respectively, and used their representations to train the same set of classifiers. Surprisingly, except for the logistic regression classifier, our 2D latent space model is in general on par with or better than ESM-1b and ProtT5-XL (Figure 4.3E). This is very likely a result of the more focused training on specific aligned sequence families, compared with two transformer models using unaligned sequences.[43]

To explore the potential of using the latent space model for enzyme discovery, we investigated the agreement between the clustering of cyclase embeddings and that of putative peptide precursors (functional annotations) by calculating the AMI score. Our anticipation is that strong overlap between the clustering were indicative of the model ability to assign function. We found that the highest AMI scores were 0.78, 0.75, and 0.82 for class I, II, and IV cyclases, respectively (Figures 4.3F and S4.10–S4.12). For comparison, we also clustered the cyclases by generating two SSNs with alignment

scores 46 and 52, which corresponded to recommended 35% and 40% sequence identity. From this analysis, we found that most cyclases, either in class I, II, or IV, were clustered into individual giant clusters and could not be separated (Figures S4.3 and S4.4). The corresponding AMI scores were calculated to be 0.31, 0.42, and 0.69 for class I, II, and IV using the alignment score of 46 and 0.33, 0.50, and 0.69 for the alignment score of 52 (Figure 4.3F). The results provide further evidence that the clustering from SSNs is less aligned with functional annotations compared with the clustering from our 2D latent space models. Increasing the alignment scores to 66 and 80 (45% and 50% sequence identity) when generating SSNs helped separate giant clusters into smaller clusters in SSNs and further improved the AMI scores. However, without any experimental support, increasing the alignment score may also risk losing information about more distant relationships between enzymes, which is one intrinsic limitation of using SSNs for enzyme discovery and sampling.

After demonstrating the application of the 2D latent space model for downstream classification tasks and its superior agreement with functional annotations compared with SSNs, we asked whether functional relationships exist in the latent space, and if so, in what pattern? The ability of the VAE-based latent space models to capture phylogenetic relationships was previously reported by using a simulated protein sequence dataset with known ancestral sequences and evolutionary relationships.[17] Previous studies showed that more ancient protein sequences tend to sit closer to the origin of a 2D latent space and branch out following a radial path in the space as the evolution progresses.[17] The same pattern emerges in our optimized model with more hidden layers (Figure 4.2C). Based on our hypothesis that functional relationships can be inferred by phylogenetic relationships, we anticipated that functional relationships would present in a star shape as well. To test this hypothesis, we labeled cyclases with different colors corresponding to their functional clusters and visualized the latent space using contour maps (Figure 4.3G). As anticipated, the functional relationships of the cyclases exist in a spiked shape with some overlapping in the central region of the latent space. These findings suggest that in the sampling of functionally related enzymes, a radial pattern should be followed. Ultimately, our demonstration that the optimized latent space model constructed in this cyclase-lanthipeptide example outperformed or matched other methods highlights its

potential as an alternative tool to SSNs for visualizing sequence-function relationships and guiding novel enzyme sampling and discovery.

Novel flavin-dependent monooxygenase discovery guided by a latent space model. The VAE-based latent space model of protein sequence families has shown significant potential as a tool for enzyme discovery and sampling. To further explore this potential, we turned to the family of flavin-dependent monooxygenases (FDMOs) to ascertain whether we could leverage our 2D latent space models to navigate the sequence space for more diverse chemical transformations. We trained a 2D latent space model on this family and used the 2D relationship of sequence and function to explore the chemistry accessible to the flavin adenine dinucleotide (FAD) cofactor, including hydroxylation, epoxidation, oxidative dearomatization, and even halogenation via the formation of HOX species.[44] Many of the FDMOs are involved in natural product biosynthesis and show great potential for biocatalytic purposes. For example, TropB performs site- and stereoselective oxidative dearomatization on a resorcinol substrate in the biosynthetic pathway for stipitatic acid.[45-46] Following the same fundamental mechanism of FDMOs, enzymes such as OpS4,[47] AtA,[48] and NahG[49-50] can catalyze decarboxylative hydroxylation on orsellinic, 6-methylsalicylic, and salicylic acid, which are important steps in the biosynthetic or degradation pathways. Although significant efforts have been dedicated to the investigation of additional FDMOs and their functions, only 0.16% of approximately 250,000 protein sequences within the UniProtKB database have been characterized, highlighting the huge gap between sequence and function and the need to explore more broadly in FDMO sequence space.

A 2D latent model with four hidden layers was developed by training a VAE on truncated aligned sequences of about 34,000 FDMOs from UniRef90 and was optimized and validated using a small dataset of computationally reconstructed ancestral FDMOs that have not been seen during the training (Figures 4.4A and 4.4B and Table S4.4).[51] Guided by the phylogenetic tree, the ancestral FDMOs were clustered into 11 groups (Figure 4.4C). The tree-based clustering was then compared with the latent space-based clustering of embedded ancestral sequences according to the AMI score. Different numbers of clusters (n = 1–11) were assigned and the AMI scores reached 0.72–0.79 when n is more than seven, indicating some agreement between two clusterings (Figures

4.4D and 4.4E). The Euclidean distances between 11 common ancestors were calculated and correlated with their corresponding evolutionary distances from the tree. The Pearson correlation coefficient is 0.78 (Figure 4.4D), illustrating that the local phylogenetic relationship was also captured in the latent space. Based on all the features observed from this testing, we envisioned that sampling uncharacterized enzymes around previously reported enzymes across the entire latent space could be a useful tool to push the boundary of unexplored sequence space.

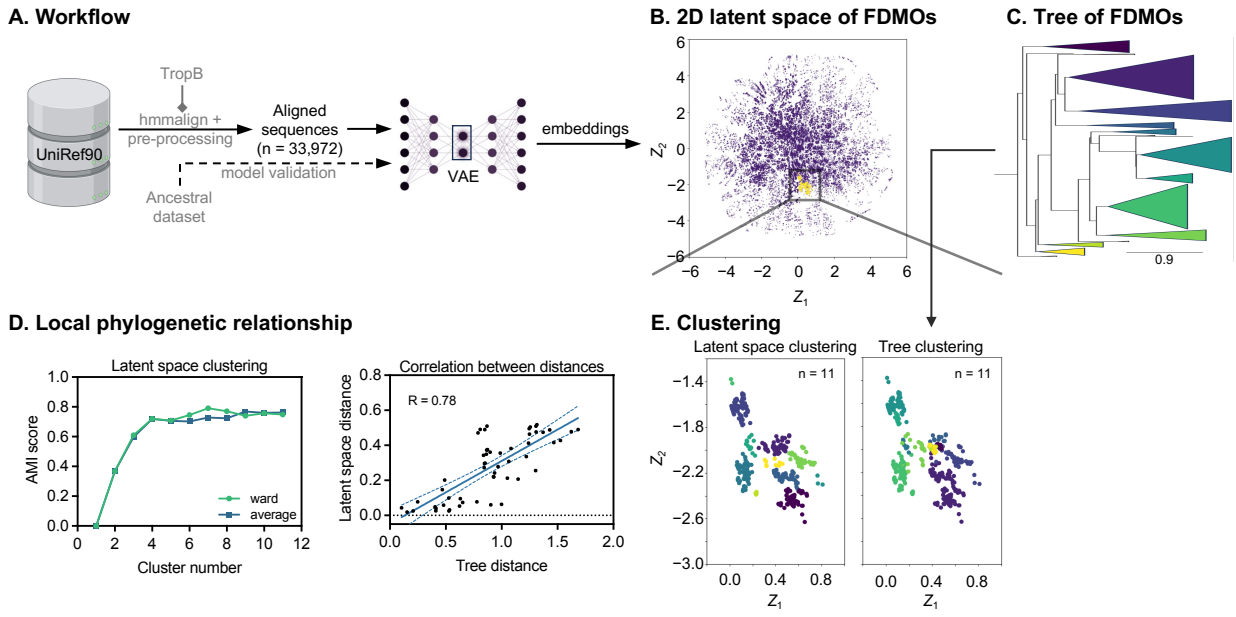


Figure 4.4. Latent space model of novel flavin-dependent monooxygenases (FDMOs).
A. Workflow showing the collection of FDMO homologs from UniRef90 and subsequent model training and validation. **B.** Embedded two-dimensional latent space representations of flavin-dependent monooxygenases (FDMOs). Purple dots represent sequences in the training dataset and yellow dots represent sequences in the ancestral dataset for validation. **C.** Phylogenetic relationship of the extant and ancestral FDMOs used for model validation. **D.** Agreement between the tree-based and latent space-based clustering and the correlation between evolutionary distances of 11 common ancestors in the tree and their Euclidean distances in the latent space. **E.** The ancestral FDMOs colored by the latent space-based and tree-based and clustering.

Latent space model-guided discovery of novel FDMOs for decarboxylative hydroxylation. Our group has focused on exploring the local sequence space of FDMOs for oxidative dearomatization while seeking opportunities to characterize homologous enzymes with alternative transformations. However, traditional tools for enzyme sampling often provide insights limited to local sequence and function space. To overcome this limitation, we curated an FDMO library by sampling select FDMOs and latent space neighboring uncharacterized enzymes. This strategy allowed us to perform screening based on prior knowledge, such as substrate scaffolds and reaction types, derived from

characterized enzymes, while simultaneously exploring more potential transformations. For example, using this approach, we successfully discovered a novel enzyme, Champase (UniProt ID: J3KA16), which catalyzes the stereoselective oxidation of 2-arylindoles.[52] Additionally, by embedding characterized FDMOs in the latent space and leveraging the continuity of the two-dimensional latent space, we found that FDMOs associated with oxidative dearomatization are positioned close to two characterized enzymes, AtA and NahG (UniProt IDs P23262 and Q53552), reported for decarboxylative hydroxylation (Figure 4.5A). The latent space model's "spiked shape" organization of phylogenetic relationships led us to hypothesize that oxidative dearomatization FDMOs are phylogenetically and functionally related to these decarboxylative hydroxylation enzymes.

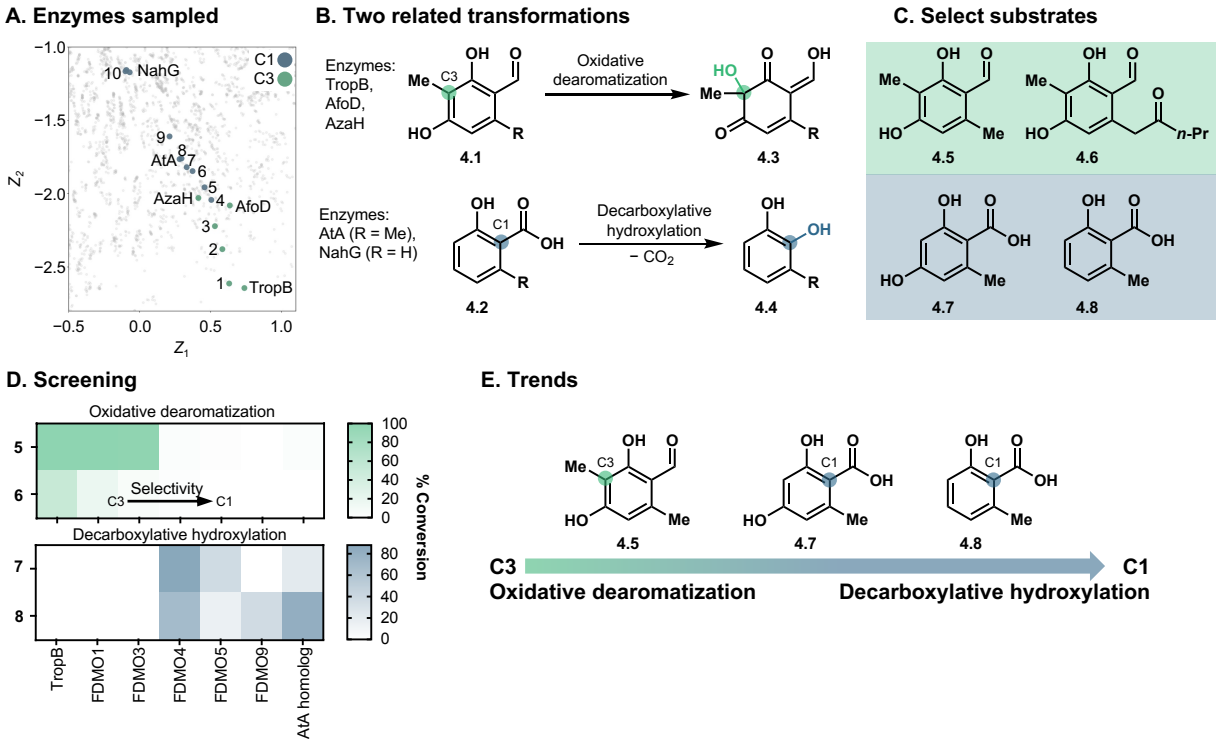


Figure 4.5. Latent space model-guided discovery of novel flavin-dependent monooxygenases (FDMOs) for decarboxylative hydroxylation.
A. Local latent space where ten uncharacterized enzymes were sampled: 1. A0A101MPG3, 2. A0A4Q7JZ50, 3. A0A0L0NHZ5, 4. A7UX11, 5. A0A1W2TF28, 6. A0A2J6S3A7, 7. A0A2J6T8P1, 8. E3RKN0, 9. A0A3D8QGL3, 10. A0A239M104. **B.** TropB-, AfoD-, or AzaH-catalyzed oxidative dearomatization and AtA-, or NahG-catalyzed decarboxylative hydroxylation observed in the local latent space. **C.** Select substrates for screening. **D.** Screening of five FDMOs on the radial path from TropB to NahG in the space. **E.** Trends observed in the FDMOs sampled from the latent space transitioning from C3- to C1-selectivity and from oxidative dearomatization to decarboxylative hydroxylation.

To investigate this hypothesis, we sampled ten FDMOs from the latent space along a path connecting NahG, AtA, and TropB, radiating from approximately (0, −1.2) in the latent space (Figure 4.5A). Of these enzymes, five (FDMO1, 3, 4, 5, and 9) were sufficiently soluble for downstream profiling (Figure S4.13). Based on the enzymatic mechanism, we postulated that oxidative dearomatization and decarboxylative hydroxylation could be distinguished by whether the enzyme hydroxylates the C1- or C3-position (Figure 4.5B). Using this insight, we proposed four substrates (**4.5**-**4.8**) as ideal candidates for testing (Figure 4.5C).

Our results revealed that FDMO1 and FDMO3, which are closer to TropB in the latent space, catalyzed oxidative dearomatization on either **4.5** or **4.6**, while FDMO4 was the first enzyme along the proposed trajectory identified to perform decarboxylative hydroxylation instead of oxidative dearomatization. The transformation is expected to begin with C1-selective hydroxylation, followed by decarboxylation to afford a rearomatized catechol product. For the substrate **4.7** (orsellinic acid) selected for decarboxylation, we clearly saw that the reaction mixtures turned red after adding FDMO4 for 1 h. This is due to that the enzymatic hydroquinone product spontaneously oxidized and partially further dimerized to form quinone dimers. Both quinone monomer and dimer were detected by mass spectrometry, which is consistent with the fungal biosynthesis of the bibenzoquinone oosporein, as reported for OpS4.[47] Beyond FDMO4, all tested enzymes, including FDMO5 and FDMO9, functioned as decarboxylases on either **4.7** or **4.8**. Further traversing the proposed trajectory in the 2D space leads to NahG, which was reported to transform salicylic acid into catechol.[49-50]

Collectivly, moving from TropB to NahG, we observed several critical trends: substrate preference shifted from resorcinol to phenol and from aldehyde to carboxylic acid, while site selectivity transitioned from hydroxylation at the C3 position to the C1 position (Figures 4.5D and 4.5E). In addition, we observed that the substrate preference seemed to transition from relatively complex aldehyde substrates to carboxylic acid substrates with less methyl groups from orsellinic acid **4.7** to 6-methyl salicylic acid **4.8**. While previous reports identified enzymes with C3-selectivity (e.g., TropB, AfoD, and AzaH) and C5-selectivity (e.g., SorbC), our latent space approach uncovered a group of C1-selective FDMOs catalyzing decarboxylative hydroxylation. These C1- and C3-

selective FDMOs are phylogenetically related, illustrating nature's versatility in leveraging selectivity to diversify downstream products. Notably, no enzyme, including FDMO4 appearing nearby the boundary of two transformations in the latent space, demonstrated proficiency in both transformations, indicating that the transition between the two functions is likely more abrupt than gradual. Although the biosynthetic pathways and natural products associated with C1-selective FDMOs remain unreported, our findings underscore the utility of latent space models in capturing enzyme relationships across local and global scales. This approach enables substrate and function prediction, further demonstrating its potential for enzyme discovery and functional analysis.

4.4 Conclusions

In conclusion, the latent space model developed in this study represents an alternative approach to traditional phylogenetic analysis and SSNs in enzyme discovery, providing a continuous, interpretable map for exploring the vast sequence-function space. By capturing both local and global sequence relationships, our model enables high-resolution clustering of closely related enzymes, facilitating the discovery of novel functionalities within enzyme families. Our application of this model to cyclases and FDMOs highlights its major advantages over SSNs, as it reliably groups enzymes by both phylogenetic and functional similarity. The continuity of the latent space model also allows for intuitive sampling strategies that align with evolutionary patterns, presenting a powerful tool for navigating the uncharacterized regions of enzyme sequence space. We envision that the model can be applied across diverse enzyme classes, reinforcing its potential to accelerate biocatalyst discovery and development.

4.5 Experimental Procedures

4.5.1 Data availability

Python scripts, Jupyter notebooks, and computational results to reproduce the model training and analysis: https://github.com/BrooksResearchGroup-UM/latent_space_paper

4.5.2 Generation of a simulated protein sequence dataset

A simulated protein sequence dataset was generated following the procedures reported by our group.[17] A random phylogenetic tree with 10,000 leaves was generated first using the Environment for Tree Exploration (ETE) toolkit in Python.[53] In the simulation of protein

evolution, the LG[24] evolutionary (substitution) model was used to specify transition probabilities between amino acid states by incorporating both the relative substitution rates and the equilibrium frequencies. The protein evolution was modeled as a continuous-time Markov chain commencing with a random 100-amino-acid root sequence. In total, 10,000 ancestral and extant simulated sequences were generated through the process and were used as the training and validation datasets.

4.5.3 Training of variational autoencoder models on the simulated protein dataset

The simulated extant sequences were used for training. No alignment was required since all simulated sequences were in an equal length and each sequence was assigned an equal weight during training.

The VAE training was implemented using PyTorch. The encoder was parameterized using a fully connected artificial neural network with different depths. The first design includes one hidden layer with 100 units and the second design includes two hidden layers with 500 and 100 units in the first and second hidden layer, respectively. We used Tanh as the activation function for the hidden layers. The encoder outputs two vectors: a mean (μ) and a log variance (log σ^2), both of dimension two. The decoder comprises one or two layers, mirroring the encoder's architecture in reverse. The final layer uses a log softmax activation to produce log probability distributions over the 20 amino acids for each position in the sequence. A Gaussian prior ($\mathcal{N}(0, I)$) was assumed for the latent space, and the Kullback-Leibler (KL) divergence between the encoder's output distribution and this prior was used for regularization.

The VAE models were trained using the Adam optimizer with a learning rate of 1e−3 and a weight decay of 1e−2 by optimizing the ELBO function.[54] Training was performed for 1000 epochs with a batch size of 64. Five-fold cross validation was performed and the model with the lowest loss was chosen as the final model.

4.5.4 Visualization, clustering, and analysis of 2D latent space representations of simulated protein sequences

The simulated ancestral sequences were embedded into a two-dimensional latent space for visualization and downstream analysis using the Jupyter Notebook script *analyze_ancestors.ipynb.* The clustering and subsequent analysis were performed using the Jupyter Notebook script *latent_space_clustering.ipynb*. The visualization of latent

space representations of extant sequences and their ancestral sequences at different tree cutoffs was performed using the Jupyter Notebook script *connect_ancestors.ipynb*.

4.5.5 Sequence similarity networks of simulated extant sequences

Sequence similarity networks (SSNs) of the simulated protein sequence dataset were created using the default setting by submitting the fasta file of the dataset to the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) website (Option C).[12,55] The initial SSN was further clustered based on different alignment scores 10, 15, 20, 25, 30, and 35, resulting in six separate SSNs for downstream analysis using the Jupyter Notebook script *ssn_clustering.ipynb*.

4.5.6 Curation of a cyclase dataset

To prepare a lanthipeptide cyclase dataset, non-redundant protein sequences of bacteria and archaea were first downloaded from RefSeq separately, and then combined to prepare a database to be searched against. The LanC-Like Pfam HMM (PF05147.12) was downloaded to run the *hmmsearch* using HMMER3[39]. The hits acquired from *hmmsearch* were filtered first to only include sequences within 200–500 amino acids and the sequences were aligned using *hmmalign*. The following commands were used to perform these tasks.

```
# Download (bacteria)
wget -r -nH --cut-dirs=6 -A '*.protein.faa.gz'
ftp://ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/

# Download (archaea)
wget -r -nH --cut-dirs=6 -A '*.protein.faa.gz'
ftp://ftp.ncbi.nlm.nih.gov/refseq/release/archaea/

# Combine fasta files (bacteria)
cat bacteria_nonredundant_protein/*.protein.faa.gz >
bacteria_nonredundant_protein.faa.gz

# Combine fasta files (archaea)
cat archaea_nonredundant_protein/*.protein.faa.gz >
archaea_nonredundant_protein.faa.gz

# Combine bacteria and archaea
cat *_protein.faa.gz > refseq_protein_bacteria_archaea.faa.gz

# Decompression
gunzip refseq_protein_bacteria_archaea.faa.gz

# The hmmpress step is required for hmmscan to work.
```

```
hmmpress PF05147.hmm

# Run the hmmsearch
hmmsearch -A PF05147_hits.sto PF05147.hmm
refseq_protein_bacteria_archaea.fasta > PF05147.out

# Output the hits from hmmsearch to a fasta file
esl-reformat fasta PF05147_hits.sto > PF05147_hits.fasta

# Multiple sequence alignment
hmmalign --outformat afa PF05147.hmm PF05147_hits_200_500aa.fasta >
PF05147_hits_200_500aa_MSA.fasta
```

4.5.7 Training of a variational autoencoder model on the cyclase dataset

The aligned sequences were processed by first excluding positions at which the query sequence, NisC (WP_043991124.1/31-402), has gaps, followed by removing sequences with amino acid codes like X, Z, and O and removing sequences with gaps more than 20% of the total length of NisC, and finally excluding positions with more than 20% of sequences having gaps. The truncated aligned sequences were one-hot encoded to represent the standard 20 amino acids and gaps.

To further reduce redundancy and ensure diversity during training, each sequence used for the training was given different weights using the position-based sequence weighting method.[56] For example, the sequence weight of the n^{th} sequence is determined by the sum of weights of positions (w_j^n), which are calculated as

$$w_j^n = \frac{1}{C_j} \times \frac{1}{C_j^n}$$

where C_j represents the number of unique amino acid types at the j^{th} position of the MSA, and C_j^n denotes the number of sequences in the MSA that have the same amino acid type at the j^{th} position as the n^{th} sequence. Intuitively, positions with fewer amino acid types across the MSA (i.e., more conserved positions) are assigned higher weights. Additionally, positions where the amino acid type is less frequently identical to other sequences are also given higher weights.

The VAE training was implemented using PyTorch. The encoder was parameterized using a fully connected artificial neural network with two layers with 500 and 100 units in the first and second hidden layer. We used Tanh as the activation function for the hidden layers. The encoder outputs two vectors: a mean (μ) and a log variance (log σ^2), both of dimension two. The decoder comprises four layers, mirroring

the encoder's architecture in reverse. The final layer uses a log softmax activation to produce log probability distributions over the 20 amino acids and gap for each position in the sequence. A Gaussian prior ($\mathcal{N}(0, I)$) was assumed for the latent space, and the Kullback-Leibler (KL) divergence between the encoder's output distribution and this prior was used for regularization.

Hyperparameters such as learning rate, weight decay, and batch size were optimized using Optuna.[57] Given the weighted protein sequences, the VAE was trained using the Adam optimizer with a learning rate of 0.00154 and a weight decay of 0.000113 by optimizing the ELBO function.[54] Training was performed for 1000 epochs with a batch size of 128.

4.5.8 Classification of the cyclases and predicted RiPP precursor peptides

The cyclase sequences in the dataset were first compared with the dataset reported by Walker et al.[40] The remaining unmatched sequences were then submitted to Rapid ORF Description and Evaluation Online (RODEO)[27] to identify biosynthetic gene clusters and predict RiPP precursor peptides. The classification of gene clusters was performed based on their encoding proteins and their matches with specific Pfam HMM profiles. Gene clusters encoding proteins that matched the Lant_dehydr_N Pfam HMM (PF04738.13) and the Lant_dehydr_C Pfam HMM (PF14028.6) were designated as class I clusters. Clusters containing proteins that matched the DUF4135 Pfam HMM (PF13575.6) were identified as class II clusters. For class III and IV clusters, gene clusters encoding proteins that matched both the LANC_Like Pfam HMM and the Pkinase Pfam HMM (PF00069.25) were categorized accordingly. Gene clusters lacking matches to these HMMs were excluded as unclassified. Proteins encoded in class III and IV gene clusters were subsequently analyzed against the HMMs reported by Mitchell and van der Donk.[40] Proteins were classified as class III if their E-value for the class III HMM was lower than that for the class IV HMM, and as class IV if the opposite was true. After the cyclases were classified, their corresponding potential RiPP precursor peptides from the same class that were not shorter than 30 amino acids or longer than 120 amino acids were saved into fasta files and combined with previously reported RiPP precursor peptides for building SSNs. The classification was done using the Jupyter Notebook *rodeo_classification.ipynb*.

4.5.9 Sequence similarity networks of RiPP precursor peptides

Sequence similarity networks (SSNs) of the RiPP precursor peptides from the class I, II, or IV were created using the default setting by submitting the fasta files to the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) website (Option C).[12,55] The initial SSNs were clustered using an alignment score of 14. From the RODEO prediction, one cyclase may have more than one predicted RiPP precursor peptides. To avoid this situation that one cyclase was assigned more than one substrates (functions), only the RiPP precursor peptide from the largest cluster was included in the final dataset and assigned as the corresponding peptide substrate of that cyclase. Finally, clusters with less than ten members were excluded from the dataset and peptide clustering labels were used as functional annotations for downstream applications. The processing of the SSN data was done using the Jupyter Notebook script *ssn_analysis.ipynb*.

4.5.10 Sequence similarity networks of the cyclase dataset

Sequence similarity networks (SSNs) of the cyclase dataset were created using the default setting by submitting the fasta file of the dataset to the Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST) website (Option C).[12,55] The initial SSN was further clustered based on two alignment scores 46 and 52, which correspond to approximately 35% and 40% sequence identity, resulting in eight and fourteen clusters, respectively. The processing of the SSN data was done using the Jupyter Notebook script *ssn_analysis.ipynb*.

4.5.11 Visualization of 2D latent space representations of cyclases

The tasks were done using the Jupyter Notebook script *figure3_analyze.ipynb*.

4.5.12 Embeddings of the cyclases from ESM-1b and ProtT5-XL

The embeddings of the cyclases (whole sequences) from ESM-1b were acquired by running the Jupyter Notebook script *esm1b_embeddings.ipynb* with required modules installed. The embeddings from ProtT5-XL were acquired by running the command below with ProtTrans installed and models downloaded.

```
# Embeddings from ProtT5-XL
python3 protT5_embedder.py --input PF05147_hits_200_500aa.fasta --output
protein_embeddings.h5 --per_protein 1
```

4.5.13 Training of downstream classification models

The tasks were done using the Jupyter Notebook script *figure3_analyze.ipynb*.

4.5.14 Calculation of adjusted mutual information score of latent space and SSN clusterings

The tasks were done using the Jupyter Notebook script *figure3_analyze.ipynb*.

4.5.15 Visualization of peptide substrate distribution in the 2D latent space

The tasks were done using the Jupyter Notebook script *figure3_analyze.ipynb*.

4.5.16 Training of a variational autoencoder model on the flavin-dependent monooxygenase dataset

Protein sequences in the protein family PF01494 were retrieved from UniRef90 in December, 2020. Multiple sequence alignment (MSA) was performed using HMMER by aligning those sequences to the PF01494 profile created from a stable curated seed alignment of this family. The aligned sequences were processed by first excluding positions at which the query sequence, TropB (UniProt ID: B8M9J8), has gaps, followed by removing sequences with amino acid codes like X, Z, and O and removing sequences with gaps more than 20% of the total length of TropB, and finally excluding positions with more than 20% of sequences having gaps. The truncated aligned sequences were one-hot encoded to represent the standard 20 amino acids and gaps.

To further reduce redundancy and ensure diversity during training in addition to the strategy of retrieving the sequences from UniRef90, each sequence used for the training was given different weights using the position-based sequence weighting method.[56] For example, the sequence weight of the n^{th} sequence is determined by the sum of weights of positions (w_j^n), which are calculated as

$$w_j^n = \frac{1}{C_j} \times \frac{1}{C_j^n}$$

where C_j represents the number of unique amino acid types at the j^{th} position of the MSA, and C_j^n denotes the number of sequences in the MSA that have the same amino acid type at the j^{th} position as the n^{th} sequence. Intuitively, positions with fewer amino acid types across the MSA (i.e., more conserved positions) are assigned higher weights. Additionally,

positions where the amino acid type is less frequently identical to other sequences are also given higher weights.

The VAE training was implemented using PyTorch. The encoder was parameterized using a fully connected artificial neural network with four layers with 2048 units in the first hidden layer, 1024 units in the second,128 units in the third, and 32 in the fourth. We used Tanh as the activation function for the hidden layers. The encoder outputs two vectors: a mean (μ) and a log variance (log σ^2), both of dimension two. The decoder comprises four layers, mirroring the encoder's architecture in reverse. The final layer uses a log softmax activation to produce log probability distributions over the 20 amino acids and gap for each position in the sequence. A Gaussian prior ($\mathcal{N}(0, I)$) was assumed for the latent space, and the Kullback-Leibler (KL) divergence between the encoder's output distribution and this prior was used for regularization.

Given the weighted protein sequences, the VAE was trained using the Adam optimizer with a learning rate of 0.0005 and a weight decay of 0.0005 by optimizing the ELBO function.[54] Training was performed for 1000 epochs with a batch size of 256.

4.5.17 Visualization, clustering, analysis of the ancestral dataset in the latent space

The tasks were done using the Jupyter Notebook script *figure4_analyze.ipynb*.

4.5.18 Plasmids, DNA, and protein sequences

General information

The gene encoding each FDMO was codon-optimized for overexpression in *E. coli* synthesized by the Joint Genome Institute (JGI). The JGI cloned the synthesized sequence into pMCSG7 vector conferring the T7 expression system, ampicillin resistance, and an N-terminal 6x His tag encoded upstream from the inserted gene. No further modification to the construct was necessary.

Sequence information

FDMO1 (A0A101MPG3) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGCAGACGCGATCGATGCGCAGGATGGTAGCTTTACCGTGGGGATCGTCGGGGC
CGGCTTAATCGGTGTGATGCTCGCGGCGGGGTTAGTTCGCCGCGGTATCTCCGTGAAAGTGTAC
GAACAGGCGCGTGGCTTTCGCGAAATCGGTGCTGGTATGGCTTTCACTGCGACAAGTGTGCGGT
GTATGGAACTCCTCGACCCGGCAATTGTTTACGCTCTGCGTAGCTCCGGCGCAGTTCCCATCTC
CATTGGTGACCATCAGGGCGAAGCTCGCGATTACTTGCGTTGGGTGGATGGCTACAACCAGGTA
CCGGGCAAAGACACCACCGATGAACGCAAACTGTACGAGATCGACGCTGGCATTCGCGGCTTTG
AGGGCTGCCGCCGCGACCAGTTTCTCGAAACGCTGGTCAAAGTGCTGCCCGAAGGTGTCATCGA
ATGCCAAAAGCGCTTACAACAGGTTTACGAAAAAGAACTGGGTGACAAAATCACCTTGGAATTT
GCGGATGGCGTGAGTGTTGAGGTGGACTGCGTGATTGGCACCGATGGCATCCGTAGCCGCGTGC
GTCAGCATCTGTTTGGGAAGGATAGTCCACATTCTTTTCCGCACTACTCGCACAAATTTGCCTA
CCGTGGTCTGATCAGTATGGAGAACGCCATCTCGGCCCTGGGAGAATCAAAGGCCCATACCCTC
AACATGCACGTCGGCCCCAACGCGCACCTGATTCATTACCCGGTGGCGAACCAGACCATGGTCA
ATGTGGCTGCGTTTGTCTCAGATCCAAACGACTGGACCGATTCCACCTCCCTGGTGCGCCCAGC
GACGCGCCTTGACGCTATGAATGACTTCGCGAACTGGAATACATGCTTACGCGCAGTTGTTGGG
TTTTTGCCGGAAAAAATCGATCGCTGGGCGATGTTTGACACCTGGGACTACCCGGCGCCGTACT
TCTCGCGCGGTAAAATCTGCCTGGCGGGTGACGCGGCGCATGCAGCCGTGCCGCATCATGGCGC
CGGCGCGTGCATTGGTGTAGAAGATGCCCTGTGCCTGGCAATCCTGCTGGGTGAAGTGAGCGTG
TCGATGCGTGGCCAGAGCGCAATGGATTACGGTCGCGCGGTAGTCGCAGCTTTCAATACGTACG
ACGCGGTCCGTCGTGTACGTGCGCAATGGTTCGTGGATAGCTCCCGTCGTGTATGCGACCTGTA
CCAGCAACCGGAGTGGGCTGATCCAGAACGTTGGGCGAAAGCTGAGACCTGTTTCGAAGAAATC
AAGGACCGCAGCCACAAAATTTGGCATTTTGACTACAATTCAATGCTGCAGGAGGCGATCGAGG
GTTACCGTCAGCGTATGTCAGAGTAA

MHHHHHHSSGVDLGTENLYFQSNMADAIDAQDGSFTVGIVGAGLIGVMLAAGLVRRGISVKVYE
QARGFREIGAGMAFTATSVRCMELLDPAIVYALRSSGAVPISIGDHQGEARDYLRWVDGYNQVP
GKDTTDERKLYEIDAGIRGFEGCRRDQFLETLVKVLPEGVIECQKRLQQVYEKELGDKITLEFA
DGVSVEVDCVIGTDGIRSRVRQHLFGKDSPHSFPHYSHKFAYRGLISMENAISALGESKAHTLN
MHVGPNAHLIHYPVANQTMVNVAAFVSDPNDWTDSTSLVRPATRLDAMNDFANWNTCLRAVVGF
LPEKIDRWAMFDTWDYPAPYFSRGKICLAGDAAHAAVPHHGAGACIGVEDALCLAILLGEVSVS
MRGQSAMDYGRAVVAAFNTYDAVRRVRAQWFVDSSRRVCDLYQQPEWADPERWAKAETCFEEIK
DRSHKIWHFDYNSMLQEAIEGYRQRMSE*

FDMO2 (A0A4Q7JZ50) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGCCGTCCTACTCTAAAGATACCGAGTCTGTGGAAGTGGCGGTCATCGGCGGCGG
CATCGTGGGGCTGGTACTCGCCGCCGGTTTAACCCGTCGTCAAATTAAAGTTAAGGTTTACGAA
CAGAGTCAGGGTTTCCGTGACATCGGTGCAGGCATTGGTTTTAACGGCGCCGCCAAAGCGTGTA
TGCAGATGATTGACCCAGGCGTCATCACGGCGTTGCATCGGGGAGGCGGTGTAGCGGTCTCCGC
CGCAGATGAGGACGATCTTCACGATTACTTGCGTTGGATTGATGGATTTGACCGCGGTAAAATT
CAGCACTTACACGATCAAAAACTTTACTGCAAAGTGGACGCAGGATACAAAAGTATTGAAGGCA
CCCGTCGTGATCGTTTTCTGGAAGAACTTTCGAAAGATCTGCCGGAGGGTATGGTAGAATTCAA
GAAACGTTTGCGTACGGTGGAGGAAGGTGGAGATGATTGCAAACTGCAATTACACTTTGAGGAT
GGTACCATTGCCGAGGCAGATGCGGGATGCGATGGTATCAAATCTCGCATCCGTGAGATCGTGC
TGAGTGAGGCGAGTGTCGCTTCCAAACCCAGCTACACCCATGTGAACTTTTACAGCAGTCTGAT
CCCAATGAACAAGGCAGTTGATATTCTGGGCAAATTCAAAGCATCCGTGTTCCATAACCATATT
GGACCTGGGGCTAACGTGTTGCATTACCCGGTAGCGAATGGTACCCTGTGCAACGTGAGTGCCT
TCGTGCATGACGCTAATGAATGGCCGGCGGACAAATCGCCGACGAGCATCGGTTTCCGCAAACA
TATCCAGGAAAAACTGGTCGGCTGGAGCCCTGTAGTGCGCGGCTTGATTGATTTGTTCCCGGAC
ACTCTGCCGGTCTGGGCCGTTTTTGATTTATGGGAACACCCCATGCCGTACTACAATCGTGGTC
GTATCTGTGTTGCAGGTGATGCTGCACATGCTTCATCACCGCATCATGGCGCTGGGGCCGGCAT
GGGCATCGAGGATGCGTTATGCTTAAGCGTGCTTCTTGATGAAGTCTCTTCTTCCATTCGGCTG
GAAGCGGCAAGTCGCCGCGATGCTATTAGCGTGGCCTTCCAGGTTTACGATTCGATCCGCCGTC
GTCGCTCGCAATGGCTCGTGAATAGCTCTCGTCGCTTGTGTGATCTGCAACAGCACCATGACTG
GGCCGATCCGGCCAAATTGGTCAAAGCGGAGACGTGTTTCGAGGAAATTACCGACCGTACGTAC
AAGATCTGGAACTTTGATTCGAACGGTATGATCAAAGAAAGTATCGAGAAGTACGGCCGTGCGA
TCAACAGCTTGCGGCGTAATGGTCTCGCCACGAATACTGATTGTAAAGGTAACGGCCACGTAAA
TGGGGTCCGTGCCTAA

MHHHHHHSSGVDLGTENLYFQSNMPSYSKDTESVEVAVIGGGIVGLVLAAGLTRRQIKVKVYEQ
SQGFRDIGAGIGFNGAAKACMQMIDPGVITALHRGGGVAVSAADEDDLHDYLRWIDGFDRGKIQ
HLHDQKLYCKVDAGYKSIEGTRRDRFLEELSKDLPEGMVEFKKRLRTVEEGGDDCKLQLHFEDG
TIAEADAGCDGIKSRIREIVLSEASVASKPSYTHVNFYSSLIPMNKAVDILGKFKASVFHNHIG
PGANVLHYPVANGTLCNVSAFVHDANEWPADKSPTSIGFRKHIQEKLVGWSPVVRGLIDLFPDT
LPVWAVFDLWEHPMPYYNRGRICVAGDAAHASSPHHGAGAGMGIEDALCLSVLLDEVSSSIRLE
AASRRDAISVAFQVYDSIRRRRSQWLVNSSRRLCDLQQHHDWADPAKLVKAETCFEEITDRTYK
IWNFDSNGMIKESIEKYGRAINSLRRNGLATNTDCKGNGHVNGVRA*

FDMO3 (A0A0L0NHZ5) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGCAGGGGATGAAACTGCACATGCCACGGAACCCCGTGTAGCGATCGTTGGTGG
CGGGATCGTAGGCGCAATTCTCACGCTGGGCCTGTTGCGCCAAAATATCTCTGTTCGTGTATAC
GAACGCGCAGGTGGGTTCCGTGAACTGGGTGCCGGTATGGCCTTCACCGCCGCCGCCCGCCGCT
GCATGACCCTGATGGATCCGGCCATTGCCGAGGCACTCCGCCGTTGCGGTGCAGTGTCCGTAAG
CGATGACGACTCGGATGACGACTACCTTCGTTGGATTGATGGCTACAATCAACATCGCGCTGAT
GATCCGAGCTACCAGAAACCGCTGGCGCAGATCGGCGGTAACGGGTTTCAGGGCTGCCGCCGTG
ATCAGTTCCTGGAAGAACTGAGCAAGAATATCCCGGATGGGGTAATTGAATTCCGTAAACATCT

GGAGTCAATTGAGGAACGGGAAGATGGGCGTCTGGTCCTGCGTTTTGTTGACGGAACGACGGCA
GAATCTGACGCAGTGATTGGTTGCGATGGTATTAAATCGCGTGTGCGCCAACATCTGATGGGAC
GCGACCACGCGGCGAGCCATCCACAGTACACTCATAAGGTCGCCTACCGCGGCCTGGTTCCGAT
GGATCGTGCAGTGGCGGTGCTGGGGGAATGGAAAGCGCATAACTTTCATCACCACGTGGGCCCC
GGAGCGCACCTGACCCACTACCCAGTTGCAAATGGCGCGGCATTAAATGTCACCGTCTTTCTGT
CAGACCCGGGCCGCTGGCCAGATCATACCGCTATGGTAGCTGCAGGCAGTCGTGATGAGGTCGT
TGCTGCGTTACGCGGCTGGCACCCAGTCGTGCTCGGATTGGTGCGCTTACTCCCGCATGAATTA
GTCAAATGGGCGCTGTTCGATCTGGCGGAATTTCCAGTCCCGCGCTACAATGCGGGCCGCGTCT
GCATCGCGGGGGATGCAGCGCATGCAAGCTCGCCACATCATGGCGCCGGCGCCTGCCTTGGCGT
GGAAGATGCCTTGTGCTTGAGCACTTTACTGGGACAAGTGCGCGAGACGGTCCAGGTAGCGGCG
GCAGGTACCACGACCACGGGCCGTGCACTCGAAGCAGCATTTGAAGCGTTCGACGCTGTACGTC
GTACACGTACCCAGTGGTTGGTCAATAGTTCGCGTCGCGCGTGTGATCTTTACCATCAGCCAGA
ATGGGCCGATCCGACTCGGTGGATTAAAGCCGAGACCTGCTTCGAAGAGCTGCGCGACCGTTCT
TACAAAATCTGGCATTTCGATGCGGGCGACATGGTTCGCCAGACTGTCGCGGAGTACAGCCGTC
GTCTCGAAGCGTCGTCATGCGAAGGTCAGGGCCATGTGGATGGTAAAGGCAGCCATTCTAGCAT
CGTGGCCGACCTGGATGGTAAAGCCGGCATGCCACTTCGCGGGAAACATTAA

MHHHHHHSSGVDLGTENLYFQSNMAGDETAHATEPRVAIVGGGIVGAILTLGLLRQNISVRVYE
RAGGFRELGAGMAFTAAARRCMTLMDPAIAEALRRCGAVSVSDDDSDDDYLRWIDGYNQHRADD
PSYQKPLAQIGGNGFQGCRRDQFLEELSKNIPDGVIEFRKHLESIEEREDGRLVLRFVDGTTAE
SDAVIGCDGIKSRVRQHLMGRDHAASHPQYTHKVAYRGLVPMDRAVAVLGEWKAHNFHHHVGPG
AHLTHYPVANGAALNVTVFLSDPGRWPDHTAMVAAGSRDEVVAALRGWHPVVLGLVRLLPHELV
KWALFDLAEFPVPRYNAGRVCIAGDAAHASSPHHGAGACLGVEDALCLSTLLGQVRETVQVAAA
GTTTTGRALEAAFEAFDAVRRTRTQWLVNSSRRACDLYHQPEWADPTRWIKAETCFEELRDRSY
KIWHFDAGDMVRQTVAEYSRRLEASSCEGQGHVDGKGSHSSIVADLDGKAGMPLRGKH*

FDMO4 (A7UX11) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGCCCAGTATCCCGCAGAGCCGCTTTCCGTCTGGTCCGAGCCCGATCCCTTTTCA
TGTCGCTATTATTGGCGGAGGCATTACAGGCATTAATTTGGCTCTCGGACTCGAACATCGCGGG
ATTTCTTACACCGTCTACGAGCGCAGCCCGGGCTTTCGTGAAATTGGTGCAGGCATTGGCTTTT
CTCCGAATGCGGAGGAAGCCATGAAACTGTTGAACCCGGACATCTTCCACGCATACAAACGCAC
AGCTAACCCGAACGGCGAAGATAAGTTCCAGTGGATCAACGGACTGACAGATGAACGGCTGTAC
AGCCTGCCGGTGCCGAAGGACGGCTTCCTCGGCTGCAAACGCAGTGAAATCTTAGAAGAATGGG
GGCGCTTGGTACCGAGTCAAAACATTCAGTTCAATAAAACTATTGAAGTGGTCCAAGAGCCTGG
TGATCAGGGTTACAATCCCGACGAAGCGTCGGGTAAGATTCTCCTGCGTTTCACGGATGGGACT
CAGGCTTCGGTGGATTTGGTGATTGGGTGCGATGGGATCCGTTCGCGCCTTCGCCAGCACGTGC
AGTCGCACCTGGACCGTAGCTCTATCCAGCACGTGGCCAAAGAGTCCTACACCTACAAATACTG
TTACCGGGCTCTGGTCCCAATGTCACGCGCCATTGAAGCGGTGGGACAGCATCGGGCCTCCACC
CGCTTCATGTACAATGGCCCGGGTGCGCATATTATTACCTACCCGATTGGGAATAACTCAGTTT
TAAACATGCTGGCAGTCATCAGTGACCCAAATGAATGGCCGGATGAACAGCGTCATGTCTTGCC
GGGCAATAAAGATGATATCATTCAGGCCTTTAATGGCTGGCATCCTACGGTGCAGAAACTGGCA
GATTTATTTCCTAAAGAAAATGAACGCGGCGACGGCAACATCGTAAAATGGGGTATTTTCGATA
CACTGGATTTTCCTCTGCCTAAATACTACTCGGGCCGTGTAGCGGTTGCCGGTGATGCGGCCCA
CGCTACAGGTCCGCATCTGGGCGCGGGTGGGGGACTGGGTATTGAGGATGCCCTTATCTTAGCG
GAGCTGCTCGAAAAAGTAACCGAGATTCAGGCAAGCAACGGCAGCGAGAGCGAACTCGGTGCGG

ATATTATGCTGGAACGGGCGTTATACCAATACAACGATTTTCGTTACGCACGCACCCAAGATGT
AGTGAGCTGGACTCGTAAAGCGGTAGACTTATTCCAGTGGAAAGATGAGAACGTTGCAAAAGAC
GGCAAAAAGTTCGGCCGTCTGGTTAAGAAACTGTTCCATCAGGTTTGGTACGACCTGGATCTGC
ATCAGATGGTGGATTGGAGCCGTGGCCGGTTGGAATTTGATGTGTTGGAACGTGCACTTGAACC
GATTGAACGCGAAGCCCGTGAGACGCGGGGGGGCGACAATCTGTCCAGCTTATCCGATATTGAT
TCGGAAGAAGGTGATTACCGTGTTTGA

MHHHHHHSSGVDLGTENLYFQSNMPSIPQSRFPSGPSPIPFHVAIIGGGITGINLALGLEHRGI
SYTVYERSPGFREIGAGIGFSPNAEEAMKLLNPDIFHAYKRTANPNGEDKFQWINGLTDERLYS
LPVPKDGFLGCKRSEILEEWGRLVPSQNIQFNKTIEVVQEPGDQGYNPDEASGKILLRFTDGTQ
ASVDLVIGCDGIRSRLRQHVQSHLDRSSIQHVAKESYTYKYCYRALVPMSRAIEAVGQHRASTR
FMYNGPGAHIITYPIGNNSVLNMLAVISDPNEWPDEQRHVLPGNKDDIIQAFNGWHPTVQKLAD
LFPKENERGDGNIVKWGIFDTLDFPLPKYYSGRVAVAGDAAHATGPHLGAGGGLGIEDALILAE
LLEKVTEIQASNGSESELGADIMLERALYQYNDFRYARTQDVVSWTRKAVDLFQWKDENVAKDG
KKFGRLVKKLFHQVWYDLDLHQMVDWSRGRLEFDVLERALEPIEREARETRGGDNLSSLSDIDS
EEGDYRV*

FDMO5 (A0A1W2TF28) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGCCGCAGCGGCGACTTCCGATGGCGCATCCGCGACCCCAGGTCTGCATGTTGC
GATCGTGGGAGGTGGTATCACGGGTGTGATGTTAGGTCTGGGATTAGAACGTCGGGGTGTCTCC
TACACAATTTACGAACGGGCAGCGGCCAGTGCGGAGATCGGCGCCGGGATCGGCTTCTCGCCGA
ACGCCGAACGTGCCATGGCATTGCTTTGCCCTTCCGCCCATGCTACGTACAAAGCCGTCGCTAG
TTCGACTGGCGGGGAAGAGGAGTACTTCACATGGGTGGATGGCCATGGGACAAACGAAGTGGTG
GCCCGCTTGCTGATTGGCGTCGATGCGTTTCAAGGCGGGCGTCGTTGTGATTTTGTGCGCGCGT
GGGGAGCCCTGCTTCCGGAAGGTCGTCAGCGGTTTGGGAAAGAAGTGGATGCCCTCATTCAGCG
GCCCGACGGCAAAGTGACTCTGACCTTTACCGATCGTTCTGAGGCCGAAGCTGATATTGTGATT
GGCTCGGATGGGGTGCGGAGTCGTGTTCGTCAGCTGATGCTTGGTGCGGACGATCCTGCCAGCC
GGCCGGGTTACACGCATAAATACTGTTACCGTGCGCTGTTACCAATGGCAGCAGCTGTACGCGC
AGTAGGTGCTGCACGCACTCGTACACGGTTCATGTACAACGGGCCAGGTGCCCATGCGATTACC
TACGCCGTTGCGGGTGGCACCCTCCTGAACGCCTTGTTTGTGGTCTCCGATCCTGATCCGTGGG
ACGGTGGGGAACGTCACACCGCCACTGGCACGCGTCGTGACGTGGCCGATGCCTTTGCAGGTTG
GCACCCAGCGGTCCGCGCCTTAGTTGCTCTGTTGCCGGATGAATTGGATCGCTGGGCCATCTTC
GATACCCATGAACGCCCAGCGCCGCGTTACCACGCTGGCTGCGTTGCGCTGGCTGGCGACGCCG
CCCGTACCAGTGGCCCACATTTGGGCTCTGGCGCGGGCTTTGGCATTGAGGACGCCCTGGCGCT
GTGCGGGGCGCTGGGCGTGGCAGATGAAGAAGTCCGTTCCGGACGTCTGGGCGGTGGTGCGGGC
GGAAGCCGTGCGGAAATGTGTCGGGCTGCCCTGGCAGCCTACGATGCAGTGCGCTACGAACGCG
GTCAATGGCTTCCCGGGGCGACTCGTAAGGCTGGCGAGTTGTTTCAATGGCGCCACCCAGATGT
AGGCAGCGATCCGGATAAGTTTTTACCTCGCATCACGCGTTTATTCCATACAATCTGGGATAAC
GATGTGGCAGCGATGGAACGCGCCGCGGCGGCGGAAGCAAAAAAACTGGCCTACGGCCAAGCGG
ATGAAAGTCCTGCAGTAGCAGCGGCAACTGCAACCGCAGCTGCTGCAGCGGTAGATAGTGTGGC
ATGA

MHHHHHHSSGVDLGTENLYFQSNMAAAATSDGASATPGLHVAIVGGGITGVMLGLGLERRGVSY
TIYERAAASAEIGAGIGFSPNAERAMALLCPSAHATYKAVASSTGGEEEYFTWVDGHGTNEVVA
RLLIGVDAFQGGRRCDFVRAWGALLPEGRQRFGKEVDALIQRPDGKVTLTFTDRSEAEADIVIG

SDGVRSRVRQLMLGADDPASRPGYTHKYCYRALLPMAAAVRAVGAARTRTRFMYNGPGAHAITY
AVAGGTLLNALFVVSDPDPWDGGERHTATGTRRDVADAFAGWHPAVRALVALLPDELDRWAIFD
THERPAPRYHAGCVALAGDAARTSGPHLGSGAGFGIEDALALCGALGVADEEVRSGRLGGGAGG
SRAEMCRAALAAYDAVRYERGQWLPGATRKAGELFQWRHPDVGSDPDKFLPRITRLFHTIWDND
VAAMERAAAAEAKKLAYGQADESPAVAAATATAAAAAVDSVA*

FDMO6 (A0A2J6S3A7) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGCATCCTCATCACTGCCTATTCATATTGCGATCATCGGCGGTGGAATCGGTGG
TCTGTCCCCGGCGATCGGCCTGCAGAAACACCCACAGATCTCTTTTACAATTTACGAAAGCCAC
TCATCATTCAGCGAAATCGGCGCAGGTCTGGGGTTCACTACAAACGCCCATCGCGCCATGTCCC
CGATTAGCCCCGCACTGTACGAAAAATTTAAGTCGGTGGCACTTTTTAACGGTACCGAGGAAAA
ACGTGGGTTTGCGTTTCGTCATCAGGTTGGTGAAGAAGGCCCGGACGAAGGTAAAGAGATCATC
GAAGTGCAGATTCCGCCAGGTCTTGAACAGAGCTCGGTTCACCGTAACGATTTTTTGGATGTTC
TGGTGGGTTGTTTGCCAGATGGGGGTCGCTCTTGTGTCGAGTTTGGTAAGCGTTTGGTTGAAAT
CCGTGATGGGGAAGACGAGAAACTGATTTGCATGTTTGCAGACGGTACCTCCGTGGATGCTGAT
GTGGTGGTAGGTTGCGACGGCTTAAAATCTGTGTCCCGTCTGTACGTTTTTGGGAGCGAATCCG
AACTGTCGCGGCCACGGTTCACAGGCAAGATCGCATACCGCGCGATCATTCCGATGGAACGCGC
CATTAAGGCAATTGGGGAAAAGCAGGCACGCAACCGTCAGATGTACCTGGGTCACCACGGTTTT
ATGATCACATACGCAATGGCGAAAGGCACTCTGCTGAACGTCATCGGTTTTCGCACAACTGAGT
CGAACACCTGGGACGGGGAATGGTTTCGCCCGGCCAAACAAGAAGAGTACGAACGCGACTTCAA
AGGTTGGGGCAAAACAGTCAAATCTCTGATCTCCCTGGTAGAACAGCCGGATCTCTGGGCTATT
TTTGATCACCCTCCTGTTTCTACCTACCATAAAGGTCGTCTGTGTTTACTGGGCGATGCCGCGC
ATACGACCGCCCCGCATTTCGGCCAGGGTGCGGGTATGGCATTGGAGGACGCTCATGTGCTGAG
CTCGCTGTTAGGCCGCTGTACGAGTACCAGTGAATTGACTGCCGCAAGCCGGGCTTTTGATTAC
GCGCGTGTCCCCCGTACCACTCGGGTGATGGCAGCGAGTCGCGGTCAGGGCAATCTGGTCTGCT
TCGAAGACCCGGATGCCGGCAGCGACTTGGAAAAAGTTGCGGAACTGATCAACTGGGATCGTCG
TATGTGGCTGTGGGACTACGATGTGGACGGCAACTGCAAGCAGGCCCTCGAAAAATTTGAAGAA
GAACGCGGAAGCTTTCGTTAA

MHHHHHHSSGVDLGTENLYFQSNMASSSLPIHIAIIGGGIGGLSPAIGLQKHPQISFTIYESHS
SFSEIGAGLGFTTNAHRAMSPISPALYEKFKSVALFNGTEEKRGFAFRHQVGEEGPDEGKEIIE
VQIPPGLEQSSVHRNDFLDVLVGCLPDGGRSCVEFGKRLVEIRDGEDEKLICMFADGTSVDADV
VVGCDGLKSVSRLYVFGSESELSRPRFTGKIAYRAIIPMERAIKAIGEKQARNRQMYLGHHGFM
ITYAMAKGTLLNVIGFRTTESNTWDGEWFRPAKQEEYERDFKGWGKTVKSLISLVEQPDLWAIF
DHPPVSTYHKGRLCLLGDAAHTTAPHFGQGAGMALEDAHVLSSLLGRCTSTSELTAASRAFDYA
RVPRTTRVMAASRGQGNLVCFEDPDAGSDLEKVAELINWDRRMWLWDYDVDGNCKQALEKFEEE
RGSFR*

FDMO7 (A0A2J6T8P1) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGATTCTCCGCTGTCCCCGCCGCCGTTAAGTATCGCGATTATTGGAGCAGGTAT
CGCCGGCGTTACATTGGCTATTGCGCTGAGTGAACATAATCCAAACCTTTCACTCACAATCTAC
GAATCTCGCCGCGAGTTCTCGGCTGTTAGTGCGGGCGTTGGGTTCGGCCCGAATGCGTTAAAAG
CAATGGCTTTAATTTCCCCAAAACTCATTGCAGCGTACGATCAGGTGAAGACTCCCAACTTATG

GCCTGAGAAAGACCACCTGTGGTACGAATTCCGTTGGGCCGATACCGGTAAACTCGTACTGCAG
GTGGAAAGCGAGAAAGGCTTTGCCCACTGCAATGCCTCGCGTGTACATCTGCTCGCGTCACTGG
CTGCCATCGTTTCCGGGACCGTGAAAGTGGAGTTTGGGAAACGTCTGATCGATGTCCAGGATTC
TGAAGATGCAGGTTTTGGCAACGGTGGTACCGGCAAGACGCGCATCCGTTTCGAGGACGGTACA
GATGCGTTTGCCGACGCGGTTGTGGGCTGCGATGGCATCCGGAGCGCGTGCCGCCAAATCTTAC
TGGGCGAAGATGAGAAAAGCGCGCATGCAGTATACTCTGGTAAGTACGCCTACCGTAAGATGGT
CCAAATGAAAAAGGCGATCGAGGCCGTTGGTCCGGATGTTGAGAACCGCCAAATCTACCTGGGG
AAGGGTCGCCACGTATTAACTTTTGCGATCCGCGGCGGTCAGATGCTGAATATCGTCGCTTTTA
AAGACTCAGGTGGTGCCCCCTGGACGCAGCGTCAGTGGGTTGTCACCTCTAGCCGTGAGGCGCT
GCTGAAGGATTTCGCTGGTTGCGGCGACCGCGCTTCGAAAATTTTAGAGTTGATCGACGAACCA
GAAAAATGGGCGCTGTTCAACCATCTGCCGGCCCCGACCTACGCCAAAGGCAACTTGTGCATCC
TGGGCGATGCCGCGCATGCGACCACCCCCCATTTAGGTTCTGGAGCGGGGTGCGCGATTGAAGA
CGTGCATCTGCTGAGCGGTCTGCTGACGCCGTCGCTGATCAAGTCGGCATCTGACATCAAGTGC
GCCTTTCTCGCGTACGACGCATTACGTCGTCCGCGCAGTCAAGATCTGGTGCGCAAAAGCCGCG
CTCAAGGACATATCTTGCAGCTGGAAAGCGAGCTTCCGGATCTGGCCAACGGTAACCGGGAAGC
GTGGCAGGAATACCTGGAACGCGAAATGGATGTTAATCCGCGCTGGGTGTGGAACGTGGATCTG
GAAGGTATGCTTGTCCAGGCGCGCGAACTTTTTGAGTCGATCCGTCAGGGGGAAAAGTGCGCTG
TGGACGCTTGA

MHHHHHHSSGVDLGTENLYFQSNMDSPLSPPPLSIAIIGAGIAGVTLAIALSEHNPNLSLTIYE
SRREFSAVSAGVGFGPNALKAMALISPKLIAAYDQVKTPNLWPEKDHLWYEFRWADTGKLVLQV
ESEKGFAHCNASRVHLLASLAAIVSGTVKVEFGKRLIDVQDSEDAGFGNGGTGKTRIRFEDGTD
AFADAVVGCDGIRSACRQILLGEDEKSAHAVYSGKYAYRKMVQMKKAIEAVGPDVENRQIYLGK
GRHVLTFAIRGGQMLNIVAFKDSGGAPWTQRQWVVTSSREALLKDFAGCGDRASKILELIDEPE
KWALFNHLPAPTYAKGNLCILGDAAHATTPHLGSGAGCAIEDVHLLSGLLTPSLIKSASDIKCA
FLAYDALRRPRSQDLVRKSRAQGHILQLESELPDLANGNREAWQEYLEREMDVNPRWVWNVDLE
GMLVQARELFESIRQGEKCAVDA*

AtA (Q0CJ62) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGACCTTTTCGAACCGGGTGTCCGTTGCGATTATCGGCGGCGGGATCGGTGGTCT
GTCATTAGCCATTGGCCTGCTGCAAAATAAAAACTTGGACGTTGCGATTTACGAGACTGCACCG
AAATTTGCGGAAATCGGTGCAGGTGTGGCGCTTGGACCTAACGCACAGCACGCCCTTGCTCTGA
TCAGTCCGGCCGCCGAGCACGCCTTCCGTATCCACGCTACTACTTCTCTTAGCCCGGAGTTTGA
ACATGTTTGGTTTGACTTTCGTAACGGGAATGCGGGGGAAAAAGACGGCGAGGTACTGTCTAAA
GTTGAAAATGAGACCGGCCAGCAGACTGTCCATCGCGCGAAGTTCCTGGATGAACTGGTCAAGC
TGATCCCGCGTGAGATTGCGCACTTTGGTAAACGTCTGATCCATATCCAGAAGGATCCCGTGTC
GGGCGGCGCTCAATACAAACTGTTCTTTGAAGATGGAACTACCGCCTCAGCCGACTGTGTGATC
GGCGCCGACGGTATCCATAGCTCAGTGCGCAAGCACCTGCTGGGCGAGTCCCATCCTGCTGCTA
CGCCTGTGTTCACGGGAACGGTGGTCTACCGCGGTCTGATCCCGATGGACGTTGCCCGTGATGC
TATTGGCGAATTTGCTGACAACAGCTACATGTGGTGTGGTGATGGTGGCATGGTGATGACGTAC
CCGATCGACCACGGCGAGACGCTTAATGTGGTTGGCACTCGCAATGATAAAGGTCGTTGGGATG
GCCCACCTTACACACGCCCCGTCGATGAAGAAACCGTTCGCAATGACTTTATGGGCTGGGGGGA
GATCCCGAGCAAAGTAATCCAGCCGACTATGTGGGCGATCTTAGATCACTACCCGGCTCCATAC
TACTACAGCGGTAACGTGGCCATCATGGGCGATGCAGCGCACGCAACTACGCCATTCCAGGGAG
CCGGTGCGGGTCAGGCAATTGAAGATGCCCTGGTACTGTCCACCTTGTTTCAACGCGTAACGCA

TAGCGGACTTGTGCGGCCAGCACTGGCAGCATACAACAACGTCCGCCTGCGTCGTACCCAAAAA
GTGGTGGCTACGAGCCGCGATGCCTTGCGCTTATTCTGCTTCAATGACCGCTACGTTGATGGCG
ACGCCCAGCGGTGGCGTGAAGTCTGGAACGGCCGCATGGATTGGCTGTGGGGAATGGACTTGGA
AAAACAGAATCGTGACGCGGTGAATCTGTTCGCTGATATCGTCGAAAAACAGAGCTCAGCACGC
GAAACCATGCCGAAATTTGGGCCCGTTCCGTTTCCGGTGGCGACCAGCACTGTCTAA

MHHHHHHSSGVDLGTENLYFQSNMTFSNRVSVAIIGGGIGGLSLAIGLLQNKNLDVAIYETAPK
FAEIGAGVALGPNAQHALALISPAAEHAFRIHATTSLSPEFEHVWFDFRNGNAGEKDGEVLSKV
ENETGQQTVHRAKFLDELVKLIPREIAHFGKRLIHIQKDPVSGGAQYKLFFEDGTTASADCVIG
ADGIHSSVRKHLLGESHPAATPVFTGTVVYRGLIPMDVARDAIGEFADNSYMWCGDGGMVMTYP
IDHGETLNVVGTRNDKGRWDGPPYTRPVDEETVRNDFMGWGEIPSKVIQPTMWAILDHYPAPYY
YSGNVAIMGDAAHATTPFQGAGAGQAIEDALVLSTLFQRVTHSGLVRPALAAYNNVRLRRTQKV
VATSRDALRLFCFNDRYVDGDAQRWREVWNGRMDWLWGMDLEKQNRDAVNLFADIVEKQSSARE
TMPKFGPVPFPVATSTV*

FDMO8 (E3RKN0) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAAT
ATGTCGGAGATGAAAGAAAAGCCGTTTGAGGTGGCGATCTGCGGCGGTGGCATCGGTGGTTTAT
GTCTGGCTATTGGGCTGCATAACCATGGTATTCCGTTCACGGTGTACGAAGCAGCAAGTGAATT
CGCAGAAGTTGGCGCGGGTATGTCGTTTGGCTGCAACTCTCTGCGTGCCATGAGCCTGATCGAT
CCGGCAATTCTGACAGCTTACCGTGCCGAGGAAACGGTTAACGCATGGGCGAATAAGCGCCAGA
ATTGGTTTGAGTTTCGTCACGGTTACCGCAGCTCTTCAGATTACGAAGCATTTAGCGACCCGGG
CTCCACGAGCACCACGGCCGACGGACTCATCTACGAATTTCCGGAAGGCTACATGGGTCAGAGT
AGCATTCATCGCGCTCGGTTCATCAACGCTCTGCTGACTCTGGTGCCACCCGAGCGCATCAAAC
TGGGCAAACGCGTGCAACATGTCTCCGCTGATTCTGAAGATGATGCAAAAGATGAAATTGTGGA
TGCCGAGGAAAGCAATGGTCCTAAACTGCGCAATTTTGCCTTCGTAAACGAGAATCCGGCCAGT
CGCAGTCAAAAGCCGCGCATTCAGTTTACCGATGGCACGTCAACCACCGCGGATGTGGTGGTAG
GCTGTGATGGCGTGAAAAGTCGCGTGCGTCGCTTTGTCGTCGGCTCAGACCATCCTGAGGTGGA
ACCGCGTTTCACCGGTAAATACTGCTACCGTGCTATCGTGGATATGGATCGTTTACGCGCGTTA
GTGGGCTCCGAACGTGCTACTAATAGCGTGATGTACCTGGGCCATCATGGCCACGTCGTGACGT
ACCCTATTGAACACGGAACCAAAGCCAATATTGTTGCGTTTCATAGTGAAGAAAATGGTAAATG
GACCCCGGAGGACTGGGTGGTACCATGCACATTTGATCGTATCGTCGCCGATTTCCAGGGTTGG
GATCGTACTGTTCTGTCTATCATCTCCCTCCTCGACGAAAACCCATCCGTGTGGGCTCTGTTTA
GTCACCCAAACACCCCGAATTACTCTCGCAACCACATCTGTTTGTTGGGTGACGCAGCACATGC
GGGAACGCCACATCAAGGCGCAGCTGGCGGCCAGGCCATTGAAGATGCCTTGGTGCTGAGTGAA
CTGCTTGGACGTATTACGGATAATAATAGTGATGCAATTAAAAACGCCTTACAGGCATACGATG
CAGTTCGTCGTCCACGTGCCCAGCGGGTAGTACGCACTTCGGATGCGGCGGGAGCCATTTACGA
CTTTGAGGATAAAGAACTGGGTAGCAACTGTATGCGTATTCGCGAAGATCTGCGTTGCCGTTTC
GACTGGATTTGGGAACATGACCCACAGGAGGACGTCAATAACGCGCTGCGTCTTAGCGGTATCG
CAGTTTAA

MHHHHHHSSGVDLGTENLYFQSNMSEMKEKPFEVAICGGGIGGLCLAIGLHNHGIPFTVYEAAS
EFAEVGAGMSFGCNSLRAMSLIDPAILTAYRAEETVNAWANKRQNWFEFRHGYRSSSDYEAFSD
PGSTSTTADGLIYEFPEGYMGQSSIHRARFINALLTLVPPERIKLGKRVQHVSADSEDDAKDEI
VDAEESNGPKLRNFAFVNENPASRSQKPRIQFTDGTSTTADVVVGCDGVKSRVRRFVVGSDHPE

VEPRFTGKYCYRAIVDMDRLRALVGSERATNSVMYLGHHGHVVTYPIEHGTKANIVAFHSEENG
KWTPEDWVVPCTFDRIVADFQGWDRTVLSIISLLDENPSVWALFSHPNTPNYSRNHICLLGDAA
HAGTPHQGAAGGQAIEDALVLSELLGRITDNNSDAIKNALQAYDAVRRPRAQRVVRTSDAAGAI
YDFEDKELGSNCMRIREDLRCRFDWIWEHDPQEDVNNALRLSGIAV*

FDMO9 (A0A3D8QGL3) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGCACAAAAACCGGATAACCGTTTGTCCGTGGCTATTGTCGGTGGAGGCCTGGG
AGGGGTAGCTTTGGCGATTGGCTTACTCCGCCATGGTATCAGCACACACATCTACGAAGCCGCT
GCTGGGTTTGGTGAAATCGGGGCTGGTATCATGCACGGAACGAACTCTGTCCGTGCGATCGGCT
TAATTTCCCCAGACATGCTGAGTGCATTTAAAAAATGTGCCACCTACAACACTGACCCAGCAAA
GGAGAACACTTTTATCGAGCATCGCTACGGCACCGACTCCAACAATGGGAACGGGAAGAAGAGC
GGTGATAGCATTTGCGACGTTCAAACCAACCTGCCGGAACATCGTTGGACCGGTACCCGTGCGA
TGACGACTGTTCATCGTGCGCGCTTTCTGGACGAAATCGTGAAAATTATCCCGGAGGGTACGGC
CAGCTTCTCCAAATCCTTAGTGGACATCGTCGAGTCCGATGATGGCGTGCGTCTGTACTTTGCG
GATGGATCGCAGGCAGAAGCTTCAACTGTCATTGGCTGTGATGGGATCAAATCTAAAACCCGTG
AAGTTTTGTTTCGTAACACTACCCCTGTCGTTCCGCAGTACTCAGGGGTGTACGCGTACCGTAG
CCTGGTGCCGCGTGATATCGCCATCGATGTTCTGGGTGAAGACCTCACGACCAACGGATCTTTG
TGGATGGGTTACGGTGGCTACTGGATCACTTACCCAGTAGATCATGGCGAATACCTGAATGTGG
TGGCGATGCCGCGTAAAGAAGACTGGCAGGTTGCCGTTGAAAACCTGGTGCTGCCAGGCACTCG
CGAAGATTTACTGAACGACTGGCCTGGTTGGGATCCGCGGCTGTTGAAACTGTCAGAACATTTT
CAGAGTTACAACCGTTGGGCCATCTTTGACTTGATCCATGATGAACGTTACTTCAAAGGCCGGA
TCTGCCTGTTGGGCGATGCTGCGCACGCCTCGGGACCGCACATGGGCGCTGGTGCAGGGATGGC
TATGGAAGATGCCTACATCTTGAGCAACCTGCTGGCGGGTGCGAAAGATTCGTCCGATTTGGAA
CGTGTGTTTGAAGCGTACGACTCAGTGCGCCGCACCCGTACCCAAGATTTAGTCCGCCTGTCCC
GCTACGCAGCCGAAGTCAACGAACTGGCTGCCCCAGGCATTATGGATAACTTCGAGGCGCTGAA
GCATCACTCAGACCAGCGCTACCGCTGGATTTGGAACGTCGACTTGGAGGCGCAGCTTGAAGAA
GCAAAACAAATCATGAATAAACATATTGACCACAGCACGTCAATGCAGTAA

MHHHHHHSSGVDLGTENLYFQSNMAQKPDNRLSVAIVGGGLGGVALAIGLLRHGISTHIYEAAA
GFGEIGAGIMHGTNSVRAIGLISPDMLSAFKKCATYNTDPAKENTFIEHRYGTDSNNGNGKKSG
DSICDVQTNLPEHRWTGTRAMTTVHRARFLDEIVKIIPEGTASFSKSLVDIVESDDGVRLYFAD
GSQAEASTVIGCDGIKSKTREVLFRNTTPVVPQYSGVYAYRSLVPRDIAIDVLGEDLTTNGSLW
MGYGGYWITYPVDHGEYLNVVAMPRKEDWQVAVENLVLPGTREDLLNDWPGWDPRLLKLSEHFQ
SYNRWAIFDLIHDERYFKGRICLLGDAAHASGPHMGAGAGMAMEDAYILSNLLAGAKDSSDLER
VFEAYDSVRRTRTQDLVRLSRYAAEVNELAAPGIMDNFEALKHHSDQRYRWIWNVDLEAQLEEA
KQIMNKHIDHSTSMQ*

FDMO10 (A0A239M104) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGGCACCCACATGACAACTAAACCAGCCTTACGGATTGGCATTATCGGCGGTGG
CATCTCCGGCGTTGCACTTGCCCTGGACTTGTGCCGCCATTCCCATTTGGATGTGCAGCTGTTT
GAAAGCGCGCCGGCGTTTGGGGAGGTAGGCGCGGGCGTTAGCTTCGGGGCTAACGCGGTACGTG
CGATCGCTGGGCTGGGCCTGGCGGAAGCATACGCCCAGGTTGCAGACCGTACCCCGGAGCCTTG
GCAAGACGTTTGGTTTGAGTGGCGTCGTGGCTCTGATGCAGGCTACCTGGGCGCATCGGTCGCC

CCGGGAGTTGGCCAGAGTAGTGTGCACCGCGCTGACTTCCTGGACGTCTTAGCATCACGGCTGC
CGGCCGGCGTAGCTCGCTTTGGTAAGCGCGCTACGGGTGTGGCGCAGCAGGGCGAGGAAGCGTG
TGTTCGTTTTACAGATGGAAGTGATTACCGTGGTGACCTGCTGATTGCTGCCGACGGGATTAAA
AGTGCGTTACGTGGGCATGTGCTGGAGGGTCTGGGAGCCGCACCCGCAGCCCCACGCTACTCTG
GCACGCGTGCTTACCGTGGGCTGGTCGATAGTCAGCGTTTACGCGAAGCATACCGTGCGGCGGG
TGTCGATGAGCATCTGGTTGATGTGCCTCAGATGTACCTGGGTCCGGACGCCCACATTCTGACC
TTTCCTGTCAAACAGGGGCGTATTGTGAATGTGGTTGCGTTTGTCTCCGACCGCTCAGAACCGA
ACCCGCAGTGGCCGCAGGATGCTCCGTGGGTGCGCGAAGCATCACAACGGGAAATGTTAGATGT
ATTCGCAGGCTGGGGCGACGCAGCGCGTATTCTGCTGCAGTGTATCCCAACACCGAGCCATTGG
GCCCTGCACGATCTTGCAGAGCTGCCGGGTTACGTTCATGGTCGCGTGGCCCTGATTGGCGATG
CCGCCCATGCGATGTTACCGCATCAGGGGGCGGGCGCTGGCCAAGGTCTGGAAGATGCCTGGTT
TCTTGCGCGTCTTCTCGGTGATCCCCGCGCGTGTCATGGGAATCTTGGCGAACTGCTGGCTGCA
TACGATGCTCTGCGTCGCCCGCGGGCGTGCCGCGTACAACGTACCAGCTGGGAAGCGGGGGAGT
TGTACGAATTGCGTGACGCTGCTGTGGGCGACGACGAACGCGCACTGGGCGCCACCTTAGCCAG
CCGCTTTGATTGGCTCTGGAATCATGATCTGGAGGCTGATTTAGCCGCGGCACGTGCGCGCTTA
GGTTGGCAGGCGCTGTCACATTGCGCTTAA

MHHHHHHSSGVDLGTENLYFQSNMGTHMTTKPALRIGIIGGGISGVALALDLCRHSHLDVQLFE
SAPAFGEVGAGVSFGANAVRAIAGLGLAEAYAQVADRTPEPWQDVWFEWRRGSDAGYLGASVAP
GVGQSSVHRADFLDVLASRLPAGVARFGKRATGVAQQGEEACVRFTDGSDYRGDLLIAADGIKS
ALRGHVLEGLGAAPAAPRYSGTRAYRGLVDSQRLREAYRAAGVDEHLVDVPQMYLGPDAHILTF
PVKQGRIVNVVAFVSDRSEPNPQWPQDAPWVREASQREMLDVFAGWGDAARILLQCIPTPSHWA
LHDLAELPGYVHGRVALIGDAAHAMLPHQGAGAGQGLEDAWFLARLLGDPRACHGNLGELLAAY
DALRRPRACRVQRTSWEAGELYELRDAAVGDDERALGATLASRFDWLWNHDLEADLAAARARLG
WQALSHCA*

NahG (P23262) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGAAAAATAACAAACTGGGTCTCCGTATTGGGATCGTTGGTGGTGGGATCTCTGG
CGTTGCCCTGGCCCTGGAGTTGTGCCGCTACAGCCACATTCAAGTGCAACTGTTTGAGGCCGCC
CCGGCATTCGGTGAGGTCGGCGCCGGGGTATCGTTCGGCCCAAACGCAGTTCGCGCAATTGTTG
GCCTGGGCCTCGGAGAAGCGTACCTGCAGGTGGCCGATCGTACCTCAGAACCTTGGGAAGATGT
GTGGTTTGAATGGCGCCGTGGCTCTGACGCATCATACCTTGGTGCAACAATCGCACCTGGTGTC
GGTCAGTCCAGCGTTCATCGCGCGGATTTTATCGATGCCCTGGTTACGCACCTGCCAGAAGGTA
TCGCGCAGTTCGGGAAGCGCGCCACCCAGGTCGAACAGCAGGGCGGGGAAGTCCAAGTATTATT
TACGGATGGCACCGAATACCGCTGTGACCTTCTCATTGGTGCGGACGGCATCAAAAGCGCGCTT
CGTTCACATGTGCTTGAAGGTCAGGGCCTGGCACCGCAGGTACCTCGTTTTTCGGGCACCTGCG
CATACCGTGGTATGGTTGACAGCCTGCATCTCCGTGAAGCGTACCGTGCACACGGCATCGATGA
ACATCTGGTGGACGTGCCTCAGATGTACTTGGGCCTCGACGGTCATATTCTGACGTTCCCTGTG
CGTAACGGCGGCATTATTAACGTGGTGGCGTTTATTTCAGATCGTTCAGAGCCGAAACCAACCT
GGCCGGCTGATGCGCCGTGGGTTCGTGAGGCGAGTCAGCGCGAAATGCTGGATGCTTTCGCCGG
TTGGGGAGACGCTGCGCGCGCCCTGCTGGAATGTATTCCGGCCCCGACATTGTGGGCACTGCAT
GATCTGGCTGAGTTGCCTGGTTACGTGCACGGCCGCGTCGTGTTAATTGGCGATGCGGCCCACG
CTATGTTGCCACATCAAGGTGCGGGTGCAGGGCAGGGGCTGGAAGACGCGTACTTTCTGGCGCG
TCTGTTAGGTGATACGCAGGCGGATGCGGGCAACCTGGCGGAACTGCTGGAAGCGTACGATGAT
CTGCGCCGCCCTCGCGCATGCCGCGTTCAACAAACATCATGGGAGACCGGTGAACTGTACGAGC

TCCGTGATCCGGTCGTTGGCGCAAACGAGCAGTTGTTGGGGGAAAATTTGGCAACCCGCTTTGA
TTGGCTGTGGAATCATGATCTTGATACGGACCTGGCAGAAGCTCGTGCGCGTTTGGGTTGGGAG
CACGGTGGCGGTGGCGCCCTGCGTCAAGGTTAA

MHHHHHHSSGVDLGTENLYFQSNMKNNKLGLRIGIVGGGISGVALALELCRYSHIQVQLFEAAP
AFGEVGAGVSFGPNAVRAIVGLGLGEAYLQVADRTSEPWEDVWFEWRRGSDASYLGATIAPGVG
QSSVHRADFIDALVTHLPEGIAQFGKRATQVEQQGGEVQVLFTDGTEYRCDLLIGADGIKSALR
SHVLEGQGLAPQVPRFSGTCAYRGMVDSLHLREAYRAHGIDEHLVDVPQMYLGLDGHILTFPVR
NGGIINVVAFISDRSEPKPTWPADAPWVREASQREMLDAFAGWGDAARALLECIPAPTLWALHD
LAELPGYVHGRVVLIGDAAHAMLPHQGAGAGQGLEDAYFLARLLGDTQADAGNLAELLEAYDDL
RRPRACRVQQTSWETGELYELRDPVVGANEQLLGENLATRFDWLWNHDLDTDLAEARARLGWEH
GGGGALRQG*

NahG (Q53552) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGAATGACATGAACGCGAAGAAGCCCGCTCTGCGCGTAGCAATTGTGGGGGGTGG
TATTTCCGGCTTAGCTCTGGCTCTGTCGCTGTGCAAGCACAGCCATTTAAACGTCCAACTGTTT
GAAGCAGCGCCTGCGTTTGGGGAGGTAGGTGCCGGCGTTAGCTTCGGTCCGAACGCGGTCCGTG
CCATTGTGGGTCTTGGCCTCGGGCAAGCGTACTTCCAGGTGGCAGATCGTACGCCACAGCCTTG
GGAAGACATCTGGTTCGAATGGCGTCGTGGCTCGGACGCTTCATACCTGGGCGCAACCATCGCA
GGCGTGGGTCAGAGCTCTGTCCATCGCGCCGACTTCTTGGACGCCCTCGTCAAGCATCTGCCCG
AAGGCATTGCACAGTTCCGTAAACGCGCCACGCAGATTGAACAGCAGGGTGATGAGCTGCAAGT
CCTGTTTCGTGATGGTACAGAATACCGCTGCGACTTACTGATCGGTCGCGATGGTATCAAATCC
GCGCTGCGTTCTTACGTTCTGGAGGGGCAGGGTCAAGACCACCTGGAACCCCGCTTTAGTGGTA
CCTGTGCTTACCGCGGCATGGTTGACTCCTTACAATTACGCGAAGCATACCGTATTAATGGCAT
TGACGAGCATCTGGTGGATGTACCGCAAATGTACCTGGGTTTATACGGACATATTTTGACATTC
CCGGTGCGGAAAGGCCGTATTGTTAATGTGGTCGCCTTCACCAGTGATCGTTCCCAGCCGGAAC
CGACCTGGCCGGCTGATGCGCCGTGGGTGCGCGAAGCCTCTCAGCGTGAAATGCTTGATGCCTT
CGCAGGCTGGGGCGATGCGCGCGCTCTGCTTGAATGCATCCCAGCGCCGACCTTGTGGGCATTG
CATGATCTGCCTGAATTACCAGGCTACGTTCATGGCCGCGTGGCCCTGATCGGGGATGCAGCTC
ATGCAATGTTGCCGCACCAAGGCGCCGGTGCGGGTCAAGGGTTAGAGGACGCGTACTTTCTCGC
CCGTCTGCTGGGAGATAGTCGTACAGAGACTGGTAACCTGCCAGAATTACTTGGGGCCTACGAC
GATCTGCGTCGCCCACACGCGTGCCGTGTGCAGCGTACGACTGTCGAGACCGGCGAGTTGTACG
AATTGCGTGATCCAATTGTTGGAGCCGATGAGCAGCTTGTTGGAGAAATTCTGGCAACCCGGTT
TGATTGGCTGTGGAACCATGATCTGGACGCAGATGTCGCCGAAGCTCGCTTGCGCATGGGTTGG
GAAGCACATGAACAGATCGCGCTTCGTCAGGGCTAA

MHHHHHHSSGVDLGTENLYFQSNMNDMNAKKPALRVAIVGGGISGLALALSLCKHSHLNVQLFE
AAPAFGEVGAGVSFGPNAVRAIVGLGLGQAYFQVADRTPQPWEDIWFEWRRGSDASYLGATIAG
VGQSSVHRADFLDALVKHLPEGIAQFRKRATQIEQQGDELQVLFRDGTEYRCDLLIGRDGIKSA
LRSYVLEGQGQDHLEPRFSGTCAYRGMVDSLQLREAYRINGIDEHLVDVPQMYLGLYGHILTFP
VRKGRIVNVVAFTSDRSQPEPTWPADAPWVREASQREMLDAFAGWGDARALLECIPAPTLWALH
DLPELPGYVHGRVALIGDAAHAMLPHQGAGAGQGLEDAYFLARLLGDSRTETGNLPELLGAYDD
LRRPHACRVQRTTVETGELYELRDPIVGADEQLVGEILATRFDWLWNHDLDADVAEARLRMGWE
AHEQIALRQG*

AtA homolog (I1S0U1) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGTTACGAGTGAACCGCTTTCTGTGGCCATCGTCGGCGGAGGTATCAGCGGTCT
GATTCTGGCCATCGGTCTCCTGCATAACCCGCAATTAGAAGTTACGATTTTTGAGGCGAGCTCG
GCGTTTACTGAGATTGGTGCAGGGTTAGCCCTTGGTCCTAATGCTCAACGTGCACTTCGCCTGA
TTAGCCAAGATGTGGCACAGGCGTTCGAAGACCTCGCCACTGGCAACTTATCTTCTGAATTTTC
CAAAGTATGGTTTAACTTCCGTCGCGATGTGCAAGGTCAGGAATGCGATCTGGACCTGGGTGCT
ATCGAAAATGATACGGGTCAGCAGACAGTCCACCGTGCCAAATTTCTGGATGCATTGACCAGTC
TGGTCCCGGAAGATAAAGTGCAGTTTGGTAAACGCCTGCTTAATGTTGTTCAGCATGATCACCA
TGTAGTGCTGCATTTCGAAGACGGCACTACAGCGAAAGCCAGCTGCGCTATTGGTGCGGACGGA
GTTCACTCCAATATGCGCAAATACCTGCTCGGCACCACAGCGCCCGAAGCAGTCCCGGTTTTTT
CGGGCTTTGTTGCCTACCGTGGCCTCATTCCTATGGAAATTGCTGATAAGACACTGGGCAAGTT
TGCCCGCGACGCGTACCTTTGGTGCGGCAACGAGTGCATGGTAATGATTTACCCGATCGATGAC
GGCCAGGTCGTTAACGTCGTGGCGGCCTGTCATCGTAAAACCTGGGGTAATAAAGCCTTCGTAG
TAAACAGCACCCAAGAACAGTTCCGTAAAGATTTTGATAGCCTGGATCGTATGCCGCGTAAGAT
TATCGAACTGCTGGATAGCCCGATGCTGTGGGGTTTATTAGATCATACCTCTGCCTCGTGCTAC
TACCTCAAAAATACCGCTGTAATTGGAGACGCAGCGCATGCTACCACGCCATTTCAGGGTGCAG
GTGCCGGCCAAGCGATCGAGGATGCGTTAGTTATGTCTAAATTGTTTGGCCAGGTCCAATCATT
GCATGATATTGAACCAGCGTTTGCAGCGTACGATGCAATTCGCCGTCCACGTACGGAACGCGTC
GTCCAGACTTCGCGTCAGGCAATGAAACTTTTTACATTTACCGACGCGGCGGTGGGGGATGACG
TAGAAAAATGGAAGCAGGCATGGGAAGGCCGCATGAATTGGATTTGGGATATTGACTTAGAGGC
GCATGTTGCGGATGCGCTGGGCATGTTTGAGCTGAATCGCTGCGCATGTGAGGAATAA

MHHHHHHSSGVDLGTENLYFQSNMVTSEPLSVAIVGGGISGLILAIGLLHNPQLEVTIFEASSA
FTEIGAGLALGPNAQRALRLISQDVAQAFEDLATGNLSSEFSKVWFNFRRDVQGQECDLDLGAI
ENDTGQQTVHRAKFLDALTSLVPEDKVQFGKRLLNVVQHDHHVVLHFEDGTTAKASCAIGADGV
HSNMRKYLLGTTAPEAVPVFSGFVAYRGLIPMEIADKTLGKFARDAYLWCGNECMVMIYPIDDG
QVVNVVAACHRKTWGNKAFVVNSTQEQFRKDFDSLDRMPRKIIELLDSPMLWGLLDHTSASCYY
LKNTAVIGDAAHATTPFQGAGAGQAIEDALVMSKLFGQVQSLHDIEPAFAAYDAIRRPRTERVV
QTSRQAMKLFTFTDAAVGDDVEKWKQAWEGRMNWIWDIDLEAHVADALGMFELNRCACEE*

OpS4 (J4VWM7) sequences

ATGCACCATCATCATCATCATCATTCTTCTGGTGTAGATCTGGGTACCGAGAACCTGTACTTCC
AATCCAATATGGGCAGCATTCGTGAACCACTCCACCTGGTTGTGATCGGGGGGGGTCTGGCCGG
TCTGAGCGCCGCTATCGCGACTCGTTTGGAAGGCCATCGCTGCACCCTGCTGGAAAAAGCGCCT
GAATTCAATGAAGTAGGCGCTGGACTGCAATTGACCCCGAATTCGACACGTTTGTTACGCCGCT
GGGGTGTGTTAGATAAACTGCGTAGTAAAGCCGGCATTCCTACACAGCTTACTGTGCGTCGCTA
CGATGGCTCAAAGGTTCTGTCCCGCGCGGACGGCTGGGATGAAACGATGCAGTCTCAGTACGAT
GCTCCGTTCTGGGATATGCATCGTGCAGATTTGCAAGCGGCAATGGTAGCTCGTGCTCGTCATC
TTGGTGTAGATGTACGTACTGGTGCTGAGGTCGAAAGCATCGATACGGATGGTGTTGCCGTAAT
TCTGGCTGGTACTCGCGAACGTCTGCAGGGTGATGTAGTGCTGGCGGCGGATGGCCTGTGGTCG
CGTACTCGCGCGGCGCTGTTCCCGGATCTCGGCACCGCACCACAGCCTACGGGTGATCTCGCGT
ACCGCATCATCTTACGCTTGGAGAACCTGCAACACGATCCTGAGTTGGCCGCCTGGGTTGCAAA
ACCGACTGTGAACTTTTGGGTTGGCGCGGATGCCCACGCAGTCGCATACTCTGTTCGTGGTGGT
AGTGAATTAAACCTGGTGCTGCTGTGCCCTGATGATTTGCCCGAAGGCTGCGCACGCGCGCAGG
CAGATTTAGAGGAAATGCGCGCTCGGTTTCAGGGTTGGGATCCTCTTCTGTGTCGCTTCCTGGA

CAATGTTAAAACCGTTGAGAAATGGCGTCTGATGCATATGCCATCCCTGCCGAAATGGAACCAC
GAATCCGGATACTTCACGATGGCAGGTGACAGCTGCCACCCCATGTTACCGTACCTGGCACAAG
GCGCGAATTCGGCAATGGAGGATGGCGCAGTTCTGGGACGCCTCTTAGGCAGCATTCATGAAGC
TAGTCGGATCCCAGATGTGCTTGCAGTCTACCAAGAGATCCGCAAAGTACGTGTTGAAAAGATC
GCGAAACAGGCACTTAAACAACGTTACAATTTTCATATGCCGGATGGTCCGCTTCAAGAAGCCC
GTGACGAAGCTATGACGACGCACCAGCAACGCGAAGAAGAATACGCTTCCCAGTGGACCTGCCC
GATCATGCAGCCCTGGCTTTACGGGTACGATGCGATCGCCGTAGCGGATAGTGCAGTCGCCAAT
ACGAAGCTGTAA

MHHHHHHSSGVDLGTENLYFQSNMGSIREPLHLVVIGGGLAGLSAAIATRLEGHRCTLLEKAPE
FNEVGAGLQLTPNSTRLLRRWGVLDKLRSKAGIPTQLTVRRYDGSKVLSRADGWDETMQSQYDA
PFWDMHRADLQAAMVARARHLGVDVRTGAEVESIDTDGVAVILAGTRERLQGDVVLAADGLWSR
TRAALFPDLGTAPQPTGDLAYRIILRLENLQHDPELAAWVAKPTVNFWVGADAHAVAYSVRGGS
ELNLVLLCPDDLPEGCARAQADLEEMRARFQGWDPLLCRFLDNVKTVEKWRLMHMPSLPKWNHE
SGYFTMAGDSCHPMLPYLAQGANSAMEDGAVLGRLLGSIHEASRIPDVLAVYQEIRKVRVEKIA
KQALKQRYNFHMPDGPLQEARDEAMTTHQQREEEYASQWTCPIMQPWLYGYDAIAVADSAVANT
KL*

4.5.19 Large-scale protein production and purification

Chemically competent *E. coli* strain BL21(DE3) cells were transformed with a pMCSG7 plasmid encoding any FDMO using standard protocols and grown on LB agar plates containing 100 µg/mL ampicillin at 37 °C. A single colony was used to inoculate 10 mL of LB media containing 100 µg/mL ampicillin. The culture was incubated overnight at 37 °C, 200 rpm. The overnight culture was used to prepare a glycerol stock and inoculate a 2.8 L flask with 1 L of TB media containing 4% glycerol (v/v) and 100 µg/mL ampicillin. The 1 L culture was incubated at 37 °C, 200 rpm until reaching an optical density at 600 nm (OD600) of ~0.8 and cooled to 16 °C, 200 rpm for 1 h. IPTG was added to 0.1 mM to induce overexpression. The culture was incubated at 16 °C, 200 rpm for 18 h and harvested by centrifugation at 13,881 x g for 30 min. Harvested cells were stored at −80 °C before lysis and purification.

Harvested cell pellets from overexpression were resuspended in 40 mL of lysis buffer (50 mM Tris-HCl pH 7.8, 300 mM NaCl, 10 mM imidazole, and 10% (v/v) glycerol) containing 1 mg/mL lysozyme, 100 units of DNase I, 0.1 mM flavin adenine dinucleotide (FAD), and 1 mM phenylmethylsulfonyl fluoride (PMSF), incubated on a rocker at 4 °C for 45 min, and lysed by sonication for 100 s total in cycles of 20 seconds on and 50 seconds off. Insoluble material was removed by centrifugation (46,413 x g for 30 min). The clarified lysate was combined with equilibrated HisPur nickel-nitrilotriacetic acid (Ni-NTA) resin (3 mL bed volume) from Thermo Fisher Scientific and incubated on a rocker at 4 °C for 1 h.

The resin was collected in a gravity-flow column and washed with 50 mL of lysis buffer containing 10 mM imidazole. Enriched His-tagged proteins were eluted from the resin with up to 10 mL lysis buffer containing 400 mM imidazole. The eluted proteins were concentrated at 3,500 x g at 4 °C using 30 kDa MWCO Amicon centrifuge filters purchased from EMD Millipore and exchanged into storage buffer (50 mM Tris-HCl pH 7.8, 100 mM NaCl, 10% (v/v) glycerol) using a Cytiva PD-10 desalting column, flash frozen with liquid nitrogen, and stored at −80 °C for future use. Protein samples were analyzed by mini SDS-PAGE gels and visualized with Protein Ark Quick Coomassie Stain (Anatrace) and imaged with the Azure Gel Imaging System. Total protein concentration was quantified by Pierce™ 660nm protein assay using Pierce™ 660nm Protein Assay Reagent purchased from Thermo Fisher Scientific. All purification steps were performed at 4 °C.

4.5.20 Analytical-scale enzymatic reactions

Analytical-scale reactions were performed using purified enzymes. Each reaction contained 10 μM enzyme, 50 mM potassium phosphate (KPi) buffer pH 8.0, 2.5 mM substrate (**4.5**, **4.6**, **4.7**, or **4.8**), 1.0 mM $NADP^+$, 2.0 U/mL glucose-6-phosphate dehydrogenase (G6PDH), 5.0 mM glucose-6-phosphate (G6P), and deionized water to a final volume of 100 μL. Reactions were carried out at 30 °C for 1 h and quenched with the addition of 300 μL methanol containing 2.5 mM benzophenone as an internal standard (IS). Quenched reaction mixtures were filtered through Pall AcroPrep Advance 350 μL 0.2 um PTFE 96-well filter plates by centrifugation at 2,000 rpm for 15 min, and the filtrates were analyzed by methods on UPLC-SQ. Percentage conversions were quantified by comparison of the ratio of $Area_{substrate}$ to $Area_{IS}$ of each sample to that of the no enzyme control.

UPLC-SQ method. Phenomenex Kinetex® 1.7 μm C18 100 Å, 100 x 2.1 mm column; phase A = deionized water with 0.1% formic acid, and phase B = 95/5 acetonitrile/water with 0.1% formic acid; method = 5% B held for 1.0 min, to 100% B over 2.5 min, 100% B held for 1.0 min, back to 5% B over 0.1 min, 5% B held for 0.4 min for equilibration, 235, 254, 273, 300, 350, 400, 450, and 500 nm UV detection, single quadrupole MS

detection (positive and negative modes), and 0.5 mL/min flow rate. Injection volume = 1 µL.

4.6 Supporting Information

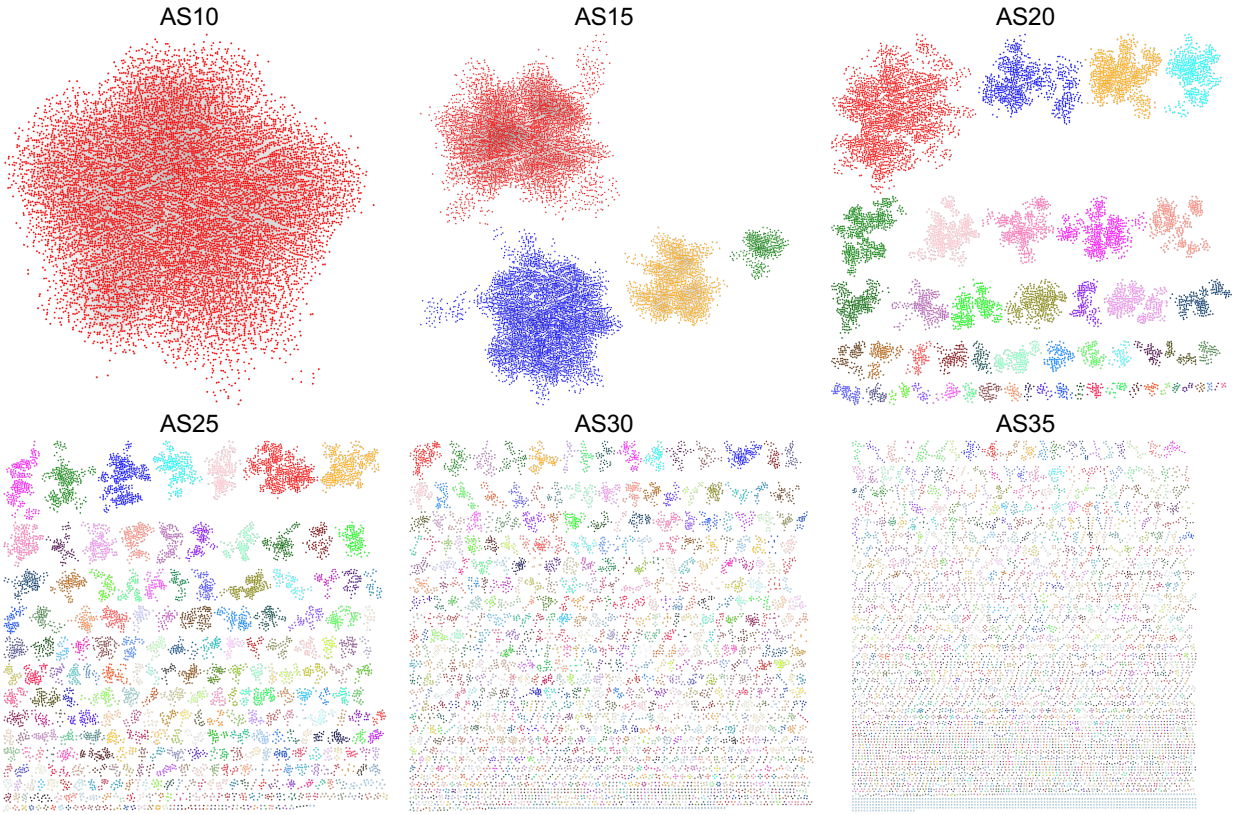
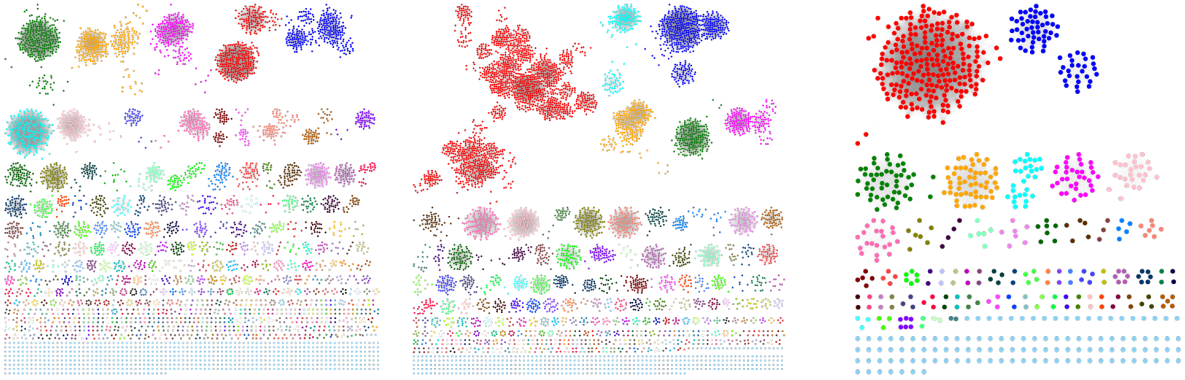


Figure S4.1. Sequence similarity networks of 10,000 simulated extant sequences built using different alignment scores (ASs) and subsequent cluster analysis.



Class I precursor peptides | Class II precursor peptides | Class IV precursor peptides

Figure S4.2. Clustering of precursor peptides from Class I, II, or IV by sequence similarity networks.

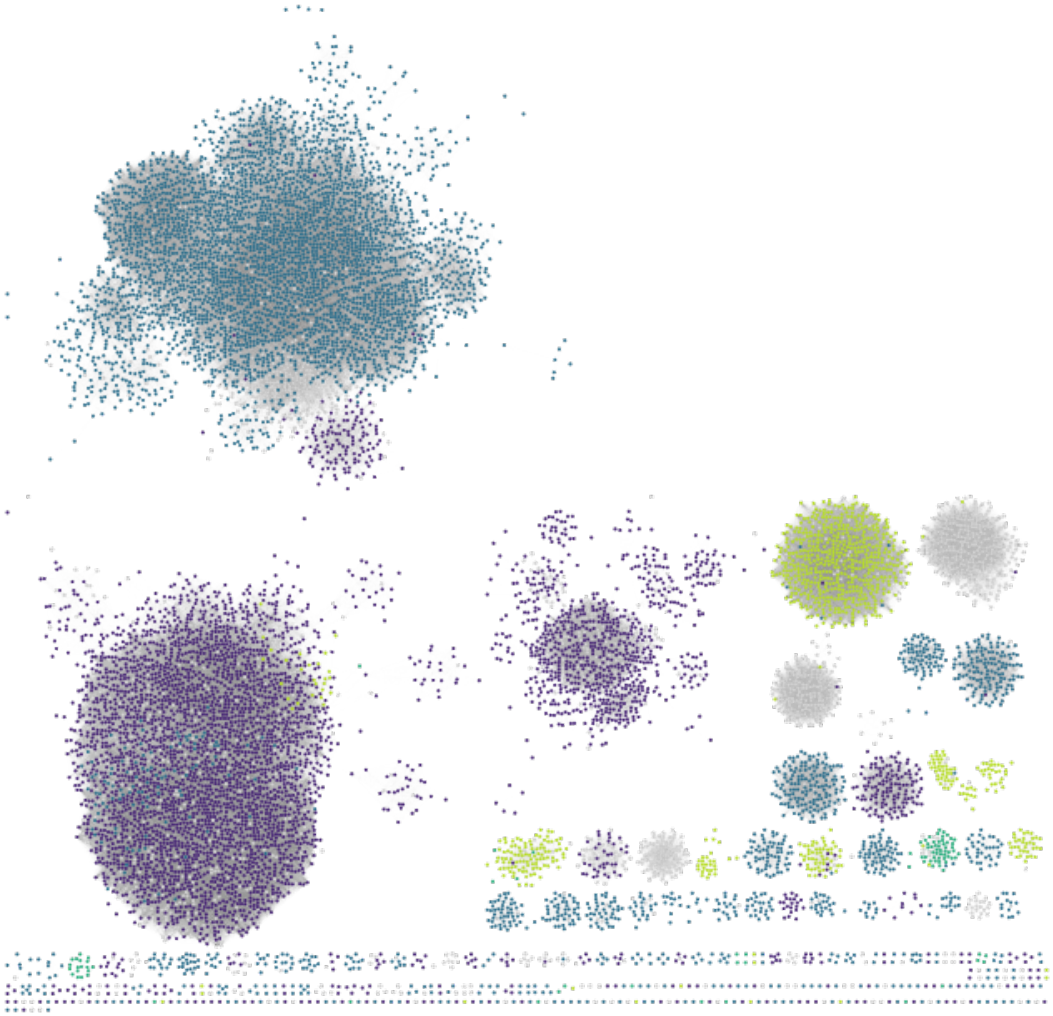


Figure S4.3. Sequence similarity network (alignment score of 46) of the cyclase dataset with nodes colored by classes. Purple: class I cyclases, Teal: class II cyclases, Green: class III cyclases, and Yellow: class IV cyclases.

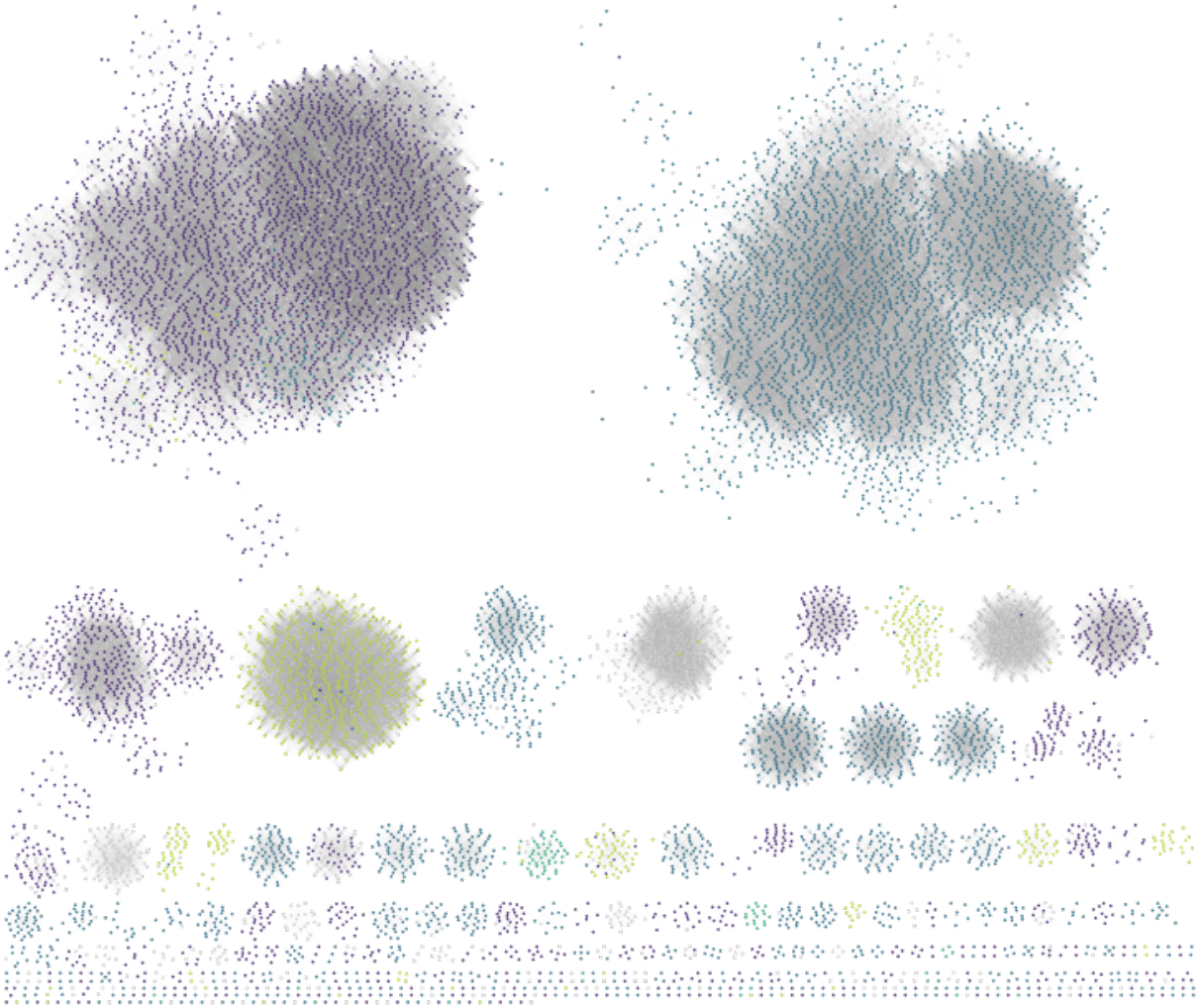


Figure S4.4. Sequence similarity network (alignment score of 52) of the cyclase dataset with nodes colored by classes. Purple: class I cyclases, Teal: class II cyclases, Green: class III cyclases, and Yellow: class IV cyclases.

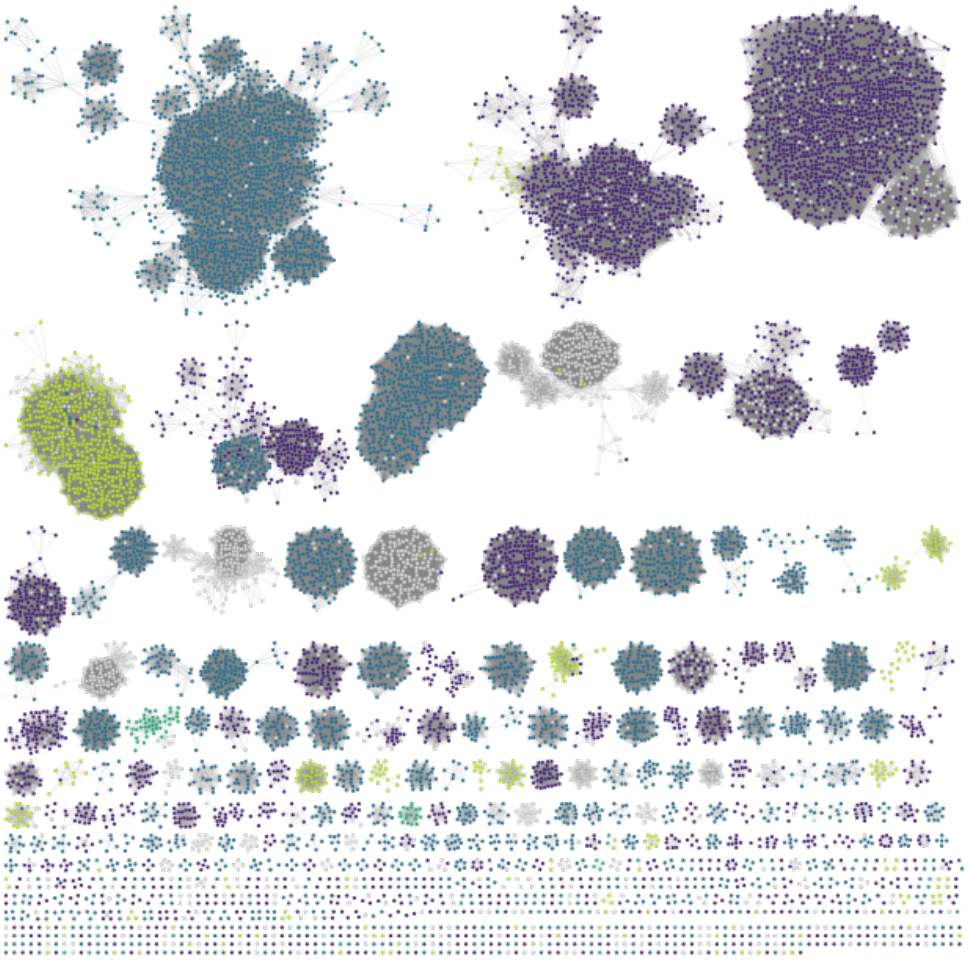


Figure S4.5. Sequence similarity network (alignment score of 66) of the cyclase dataset with nodes colored by classes. Purple: class I cyclases, Teal: class II cyclases, Green: class III cyclases, and Yellow: class IV cyclases.

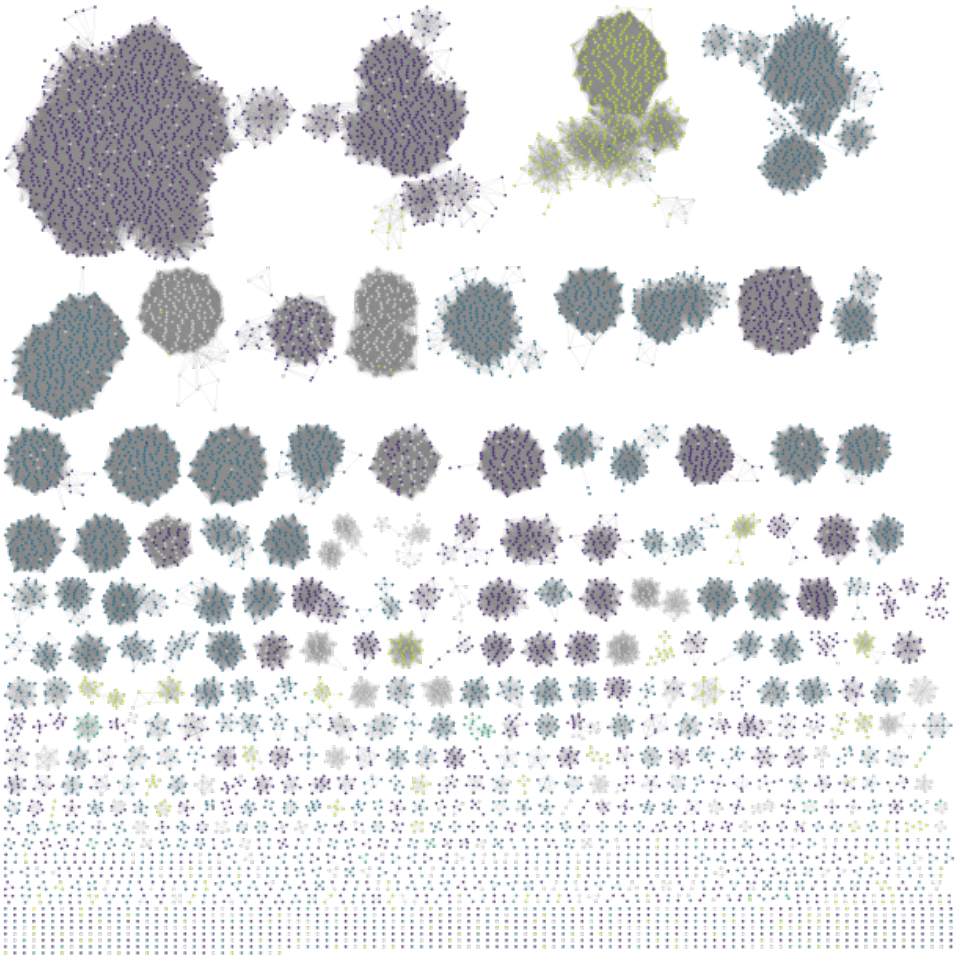


Figure S4.6. Sequence similarity network (alignment score of 80) of the cyclase dataset with nodes colored by classes. Purple: class I cyclases, Teal: class II cyclases, Green: class III cyclases, and Yellow: class IV cyclases.

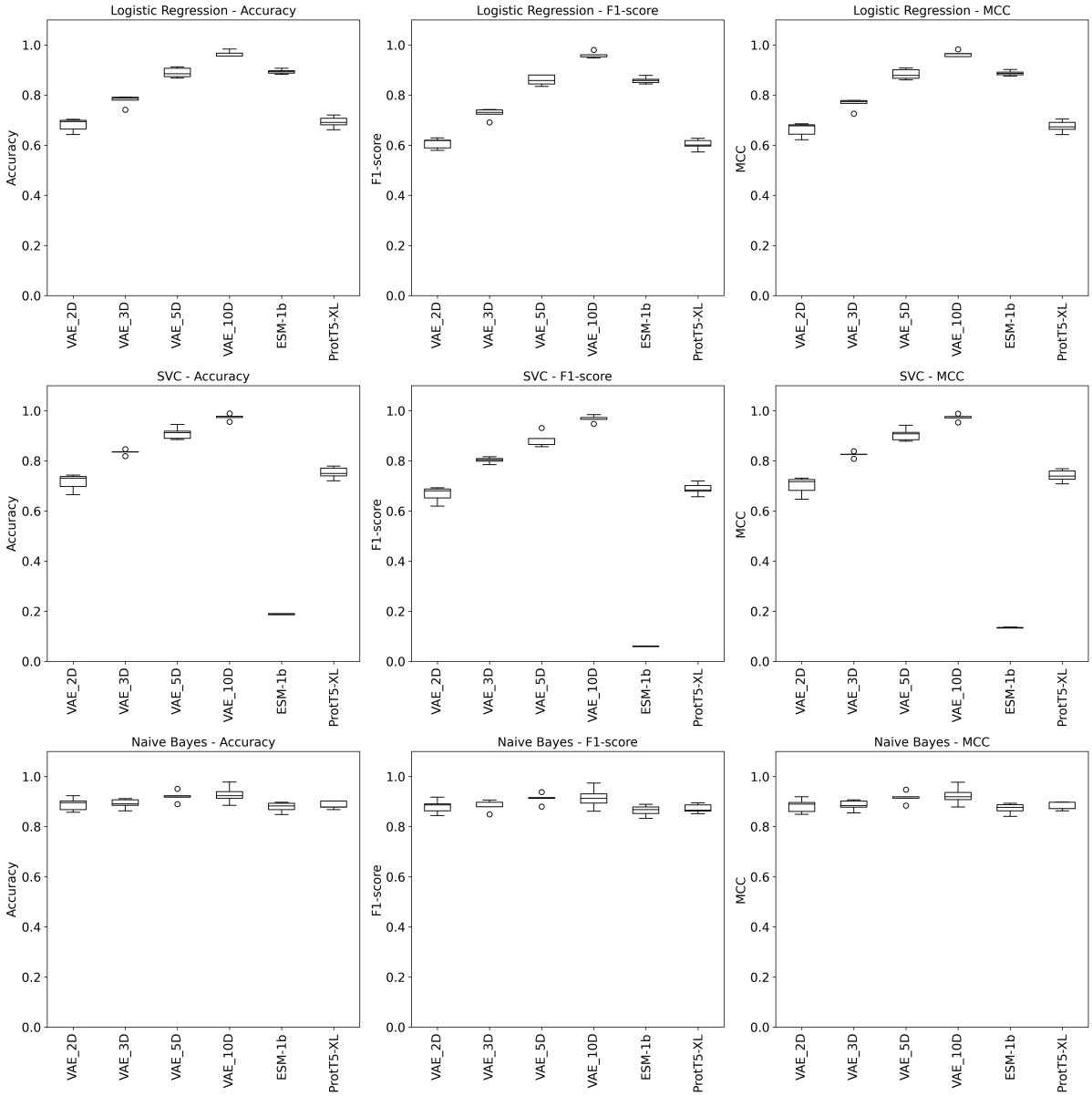


Figure S4.7. Performance comparison of various classification models trained on the class I cyclase embeddings. Classification models (Logistic Regression, SVC, and Naïve Bayes) were trained using six different embeddings. Three metrics, Accuracy, F1-score, and MCC, are included.

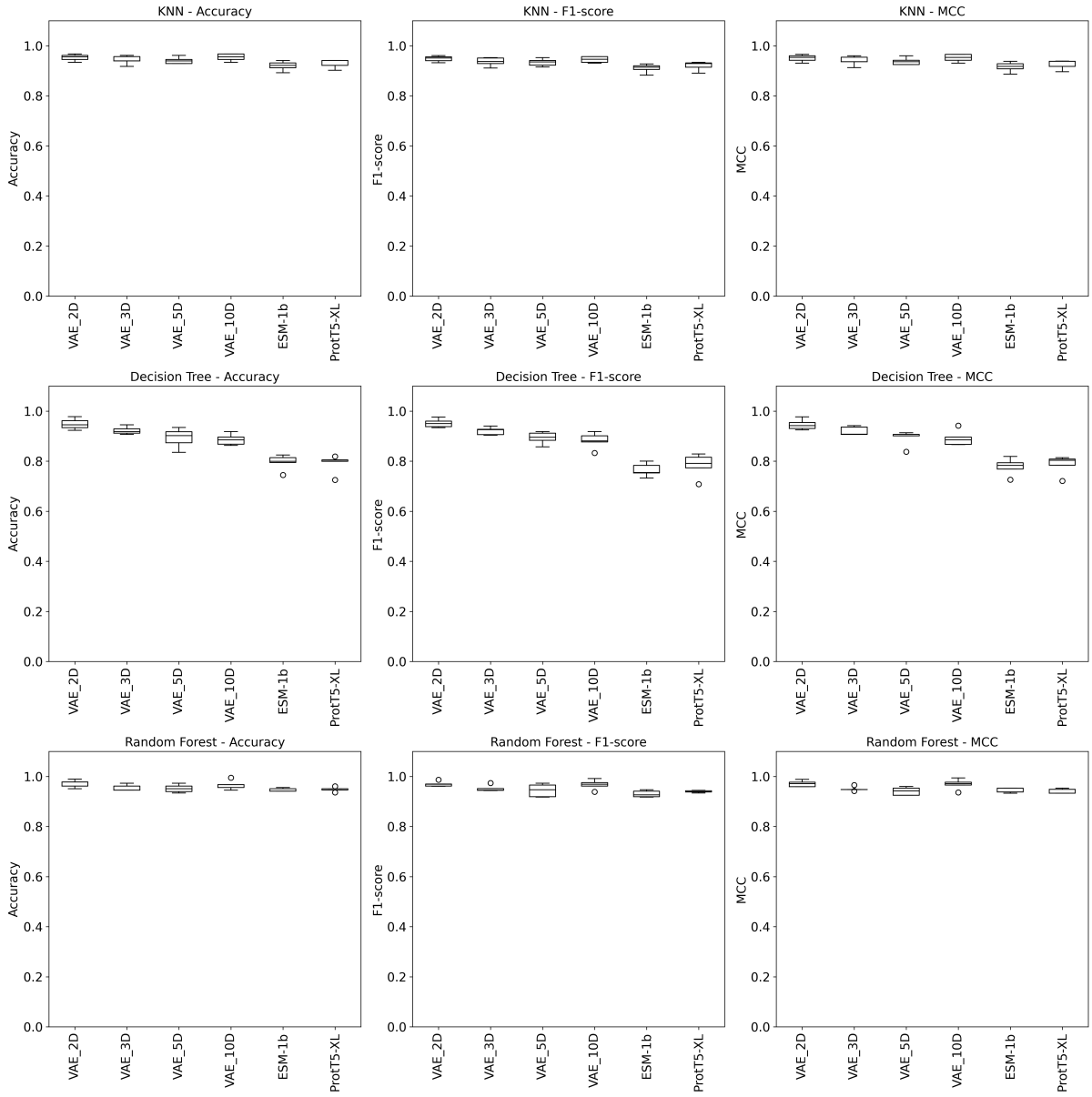


Figure S4.7 Continued. Performance comparison of various classification models trained on the class I cyclase embeddings.
Classification models (KNN, Decision Tree, and Random Forest) were trained using six different embeddings. Three metrics, Accuracy, F1-score, and MCC, are included.

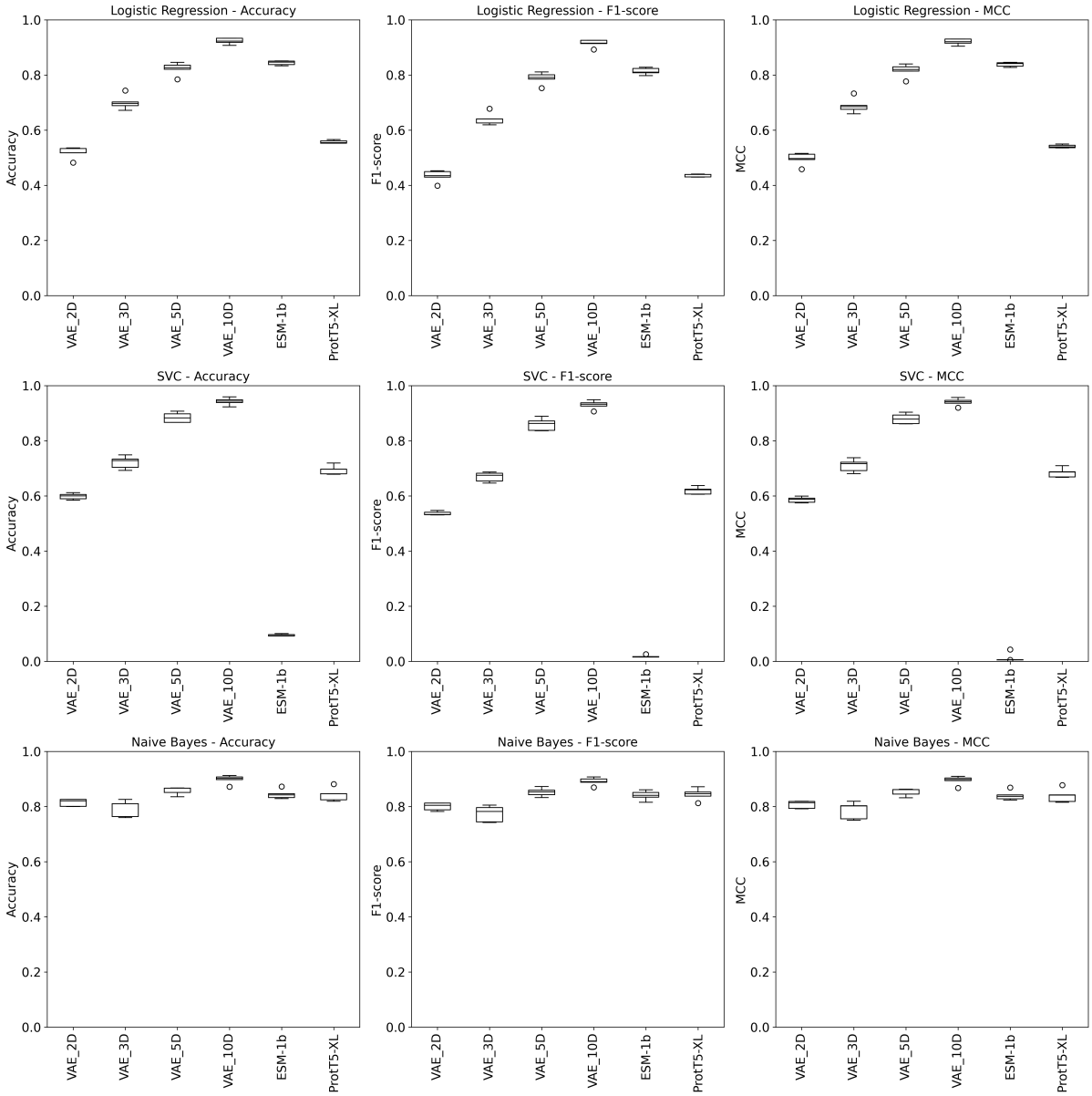


Figure S4.8. Performance comparison of various classification models trained on the class II cyclase embeddings. Classification models (Logistic Regression, KNN, and Naïve Bayes) were trained using six different embeddings. Three metrics, Accuracy, F1-score, and MCC, are included.

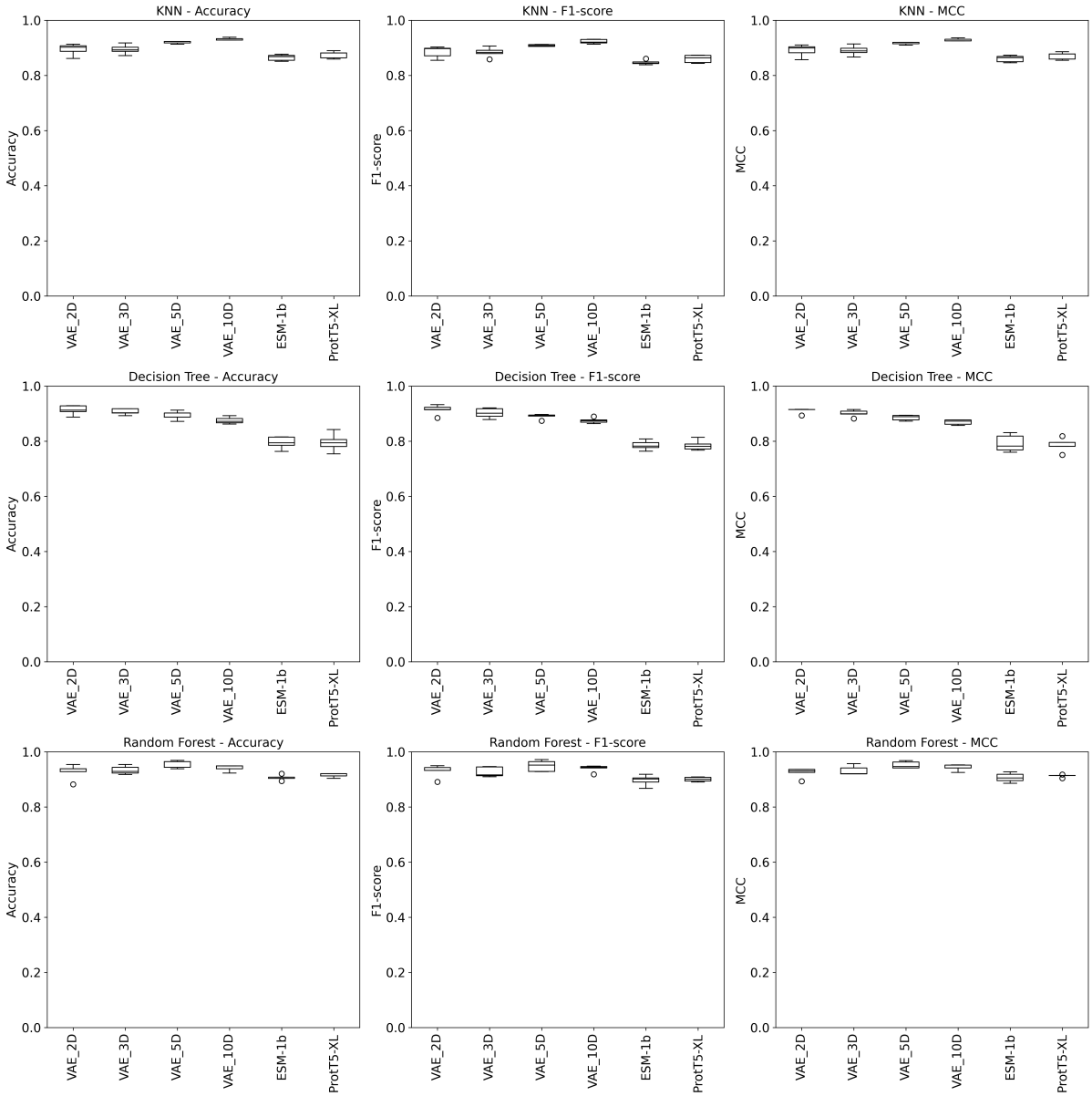


Figure S4.8 Continued. Performance comparison of various classification models trained on the class II cyclase embeddings.
Classification models (Decision Tree, Random Forest, and SVC) were trained using six different embeddings. Three metrics, Accuracy, F1-score, and MCC, are included.

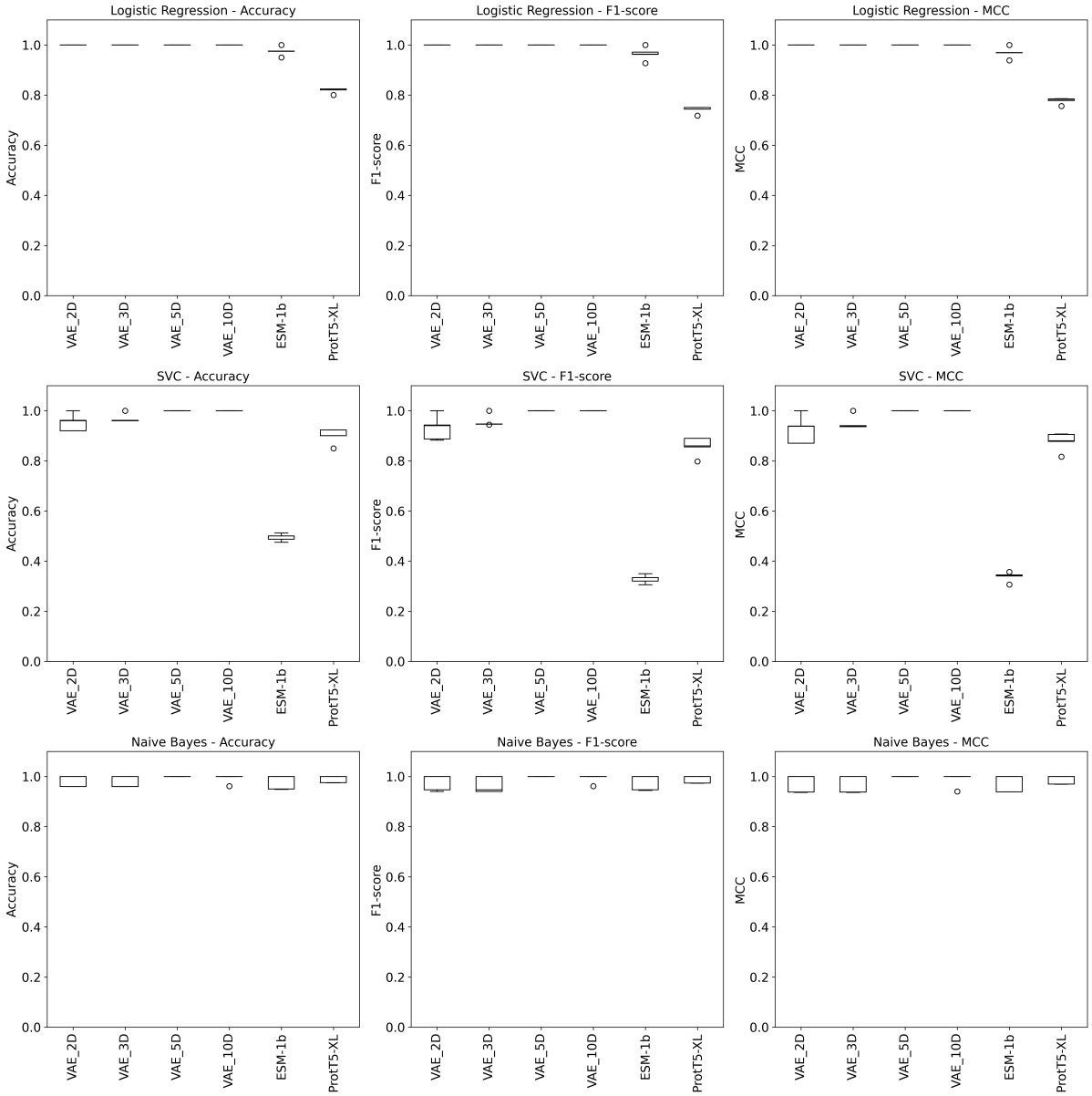


Figure S4.9. Performance comparison of various classification models trained on the class IV cyclase embeddings. Classification models (Logistic Regression, KNN, and Naïve Bayes) were trained using six different embeddings. Three metrics, Accuracy, F1-score, and MCC, are included.

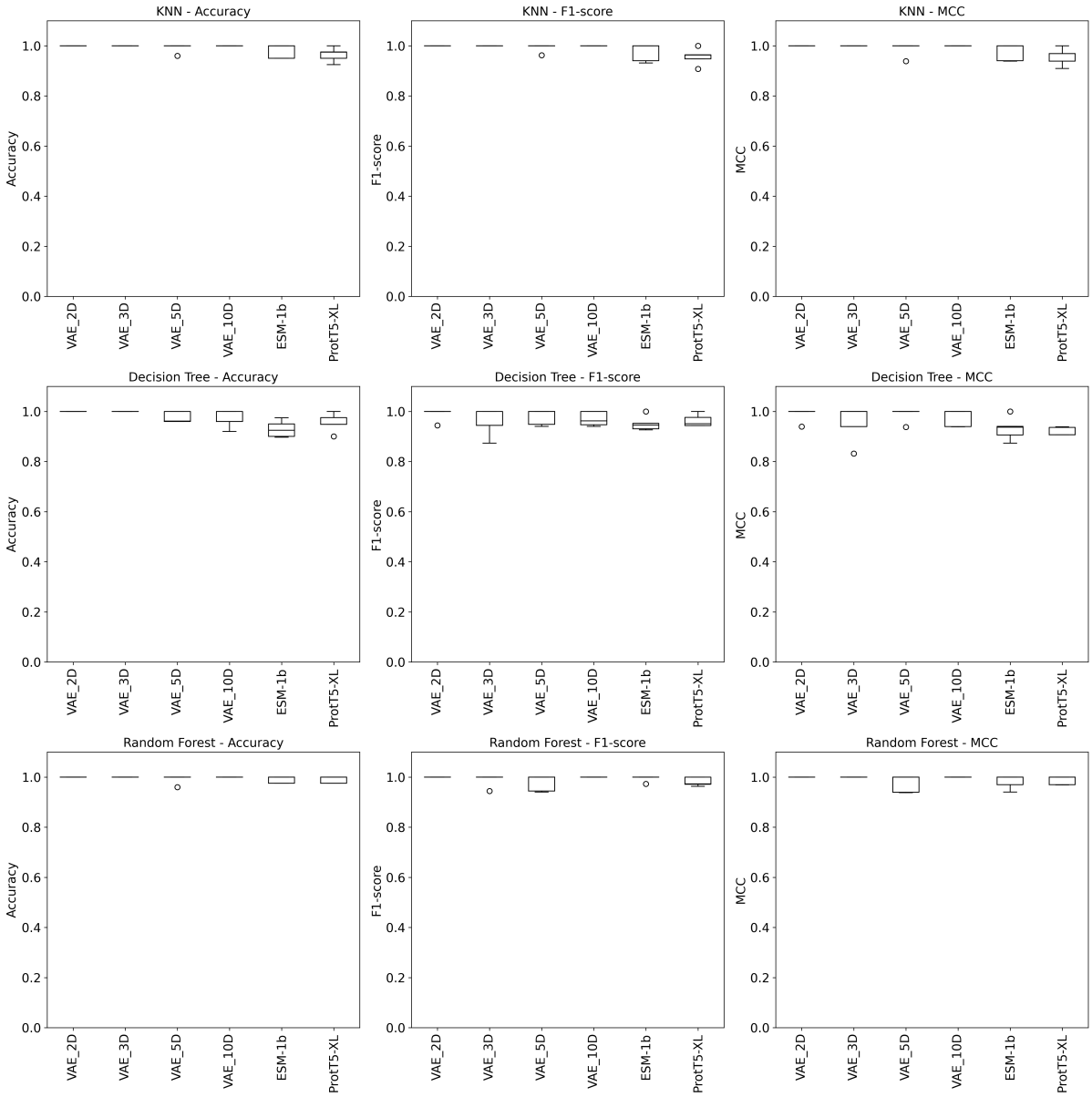


Figure S4.9 Continued. Performance comparison of various classification models trained on the class IV cyclase embeddings.
Classification models (Decision Tree, Random Forest, and SVC) were trained using six different embeddings. Three metrics, Accuracy, F1-score, and MCC, are included.

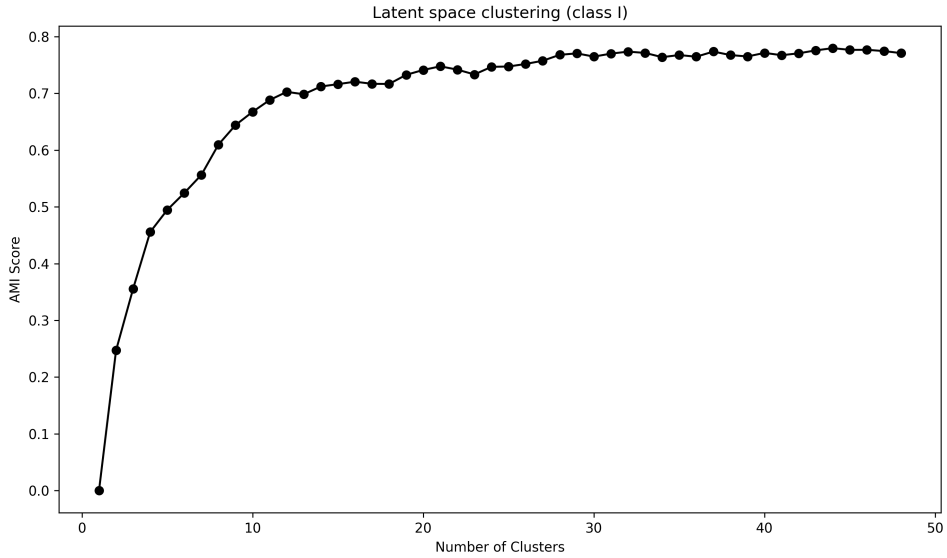


Figure S4.10. Agreement between the latent space clustering for class I cyclases and their peptide clustering. The precursor peptide clustering was treated as functional labeling. The Adjusted Mutual Information (AMI) score cahnges as the number of clusters increase.

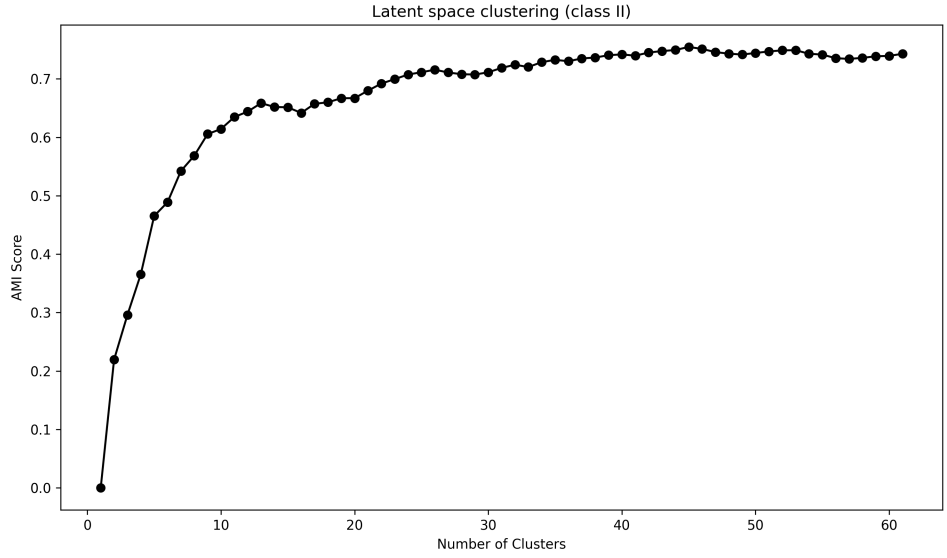


Figure S4.11. Agreement between the latent space clustering for class II cyclases and their peptide clustering. The precursor peptide clustering was treated as functional labeling. The Adjusted Mutual Information (AMI) score cahnges as the number of clusters increase.

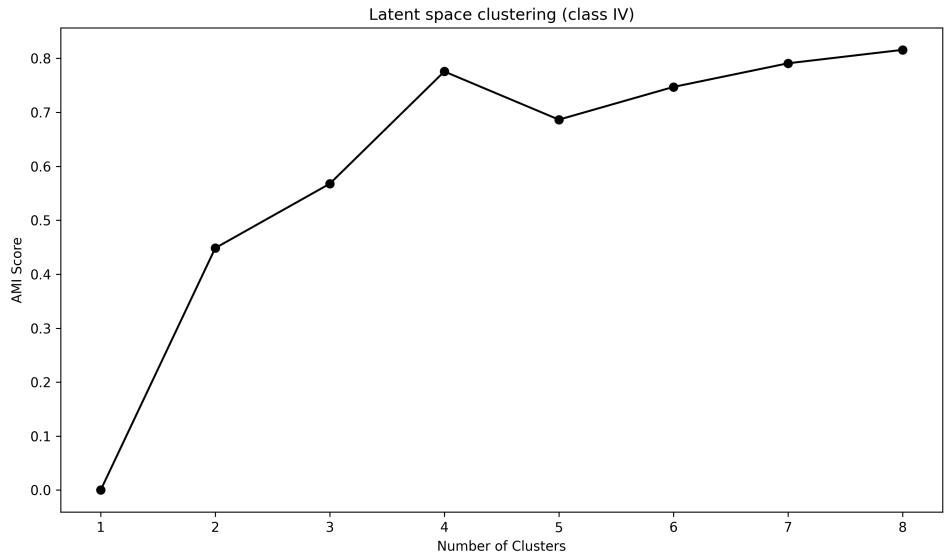


Figure S4.12. Agreement between the latent space clustering for class IV cyclases and their peptide clustering. The precursor peptide clustering was treated as functional labeling. The Adjusted Mutual Information (AMI) score cahnges as the number of clusters increase.

4.6.1 Summary of latent space model training

Table S4.1. Input datasets for all latent space models.

No.	Dataset	Dataset size	Input length	Residue type
1	Simulated protein	10,000	100	21
2	Cyclase	11,469	308	21
3	FDMO	33,972	352	21

Table S4.2. Architectures of all latent space models.

No.	Dataset	Description	Hidden layers (encoder, decoder)	Hidden layer nodes (encoder, decoder)	Latent space dimensions
1	Simulated protein	One-layer model	(1, 1)	[100], [100]	2
2	Simulated protein	Two-layer model	(2, 2)	[500, 100], [100, 500]	2
3	Cyclase	VAE_2D	(2, 2)	[500, 100], [100, 500]	2
4	Cyclase	VAE_3D	(2, 2)	[500, 100], [100, 500]	3
5	Cyclase	VAE_5D	(2, 2)	[500, 100], [100, 500]	5
6	Cyclase	VAE_10D	(2, 2)	[500, 100], [100, 500]	10
7	FDMO	N/A	(4, 4)	[2048, 1024, 128, 32], [32, 128, 1024, 2048]	2

Table S4.3. Hyperparameters for all latent space models.

No.	Dataset	Description	Learning rate	Weight decay	Batch size
1	Simulated protein	One-layer model	0.001	0.01	64
2	Simulated protein	Two-layer model	0.001	0.01	64
3	Cyclase	VAE_2D	0.00154	0.000113	128
4	Cyclase	VAE_3D	0.00184	0.000788	128
5	Cyclase	VAE_5D	0.00109	0.0000310	128
6	Cyclase	VAE_10D	0.00136	0.0000193	128
7	FDMO	-	0.0005	0.0005	256

Table S4.4. Training results of all latent space models.

No.	Dataset	Description	Epochs	Output files	Training/ validation loss	Training/testing reconstruction accuracy
1	Simulated protein	One-layer model	1000	slurm-8522.out	229.37/229.40	-
2	Simulated protein	Two-layer model	1000	slurm-8382.out	196.16/196.61	-
3	Cyclase	VAE_2D	1000	slurm-237303.out	245.73/218.38	-
4	Cyclase	VAE_3D	1000	slurm-442349.out	170.87/159.64	-
5	Cyclase	VAE_5D	1000	slurm-468708.out	100.08/99.51	-
6	Cyclase	VAE_10D	1000	slurm-442360.out	97.17/92.44	-
7	FDMO	N/A	1000	slurm-468708.out	381.6/N/A	63.9%/64.2%

SDS-PAGE analysis of enzyme samples

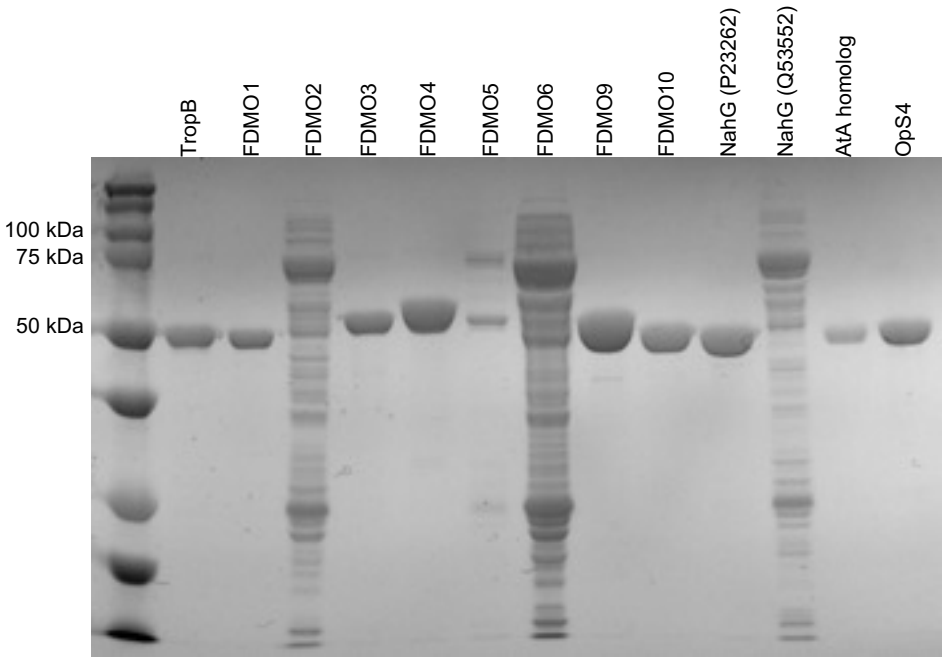
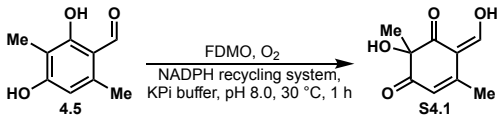


Figure S4.13. SDS-PAGE analysis of enzyme stock solutions.
Enzymes include TropB (B8M9J8), FDMO1 (A0A101MPG7, 53.07 kDa), FDMO2 (A0A407JZ50, 54.63 kDa), FDMO3 (A0A0L0NHZ5, 55.48 kDa), FDMO4 (A7UX11, 58.18 kDa), FDMO5 (A0A1W2TF28, 51.54 kDa), FDMO6 (A0A2J6S3A7, 50.13 kDa), FDMO9 (A0A3D8QGL3, 51.41 kDa), FDMO10 (A0A239M104, 49.11 kDa), NahG (P23262, 49.63 kDa), NahG (Q53552, 50.58 kDa), AtA homolog (I1S0U1, 48.95 kDa), and OpS4 (J4VWM7, 49.96 kDa).

Enzymatic reactions of the substrate **4.5**



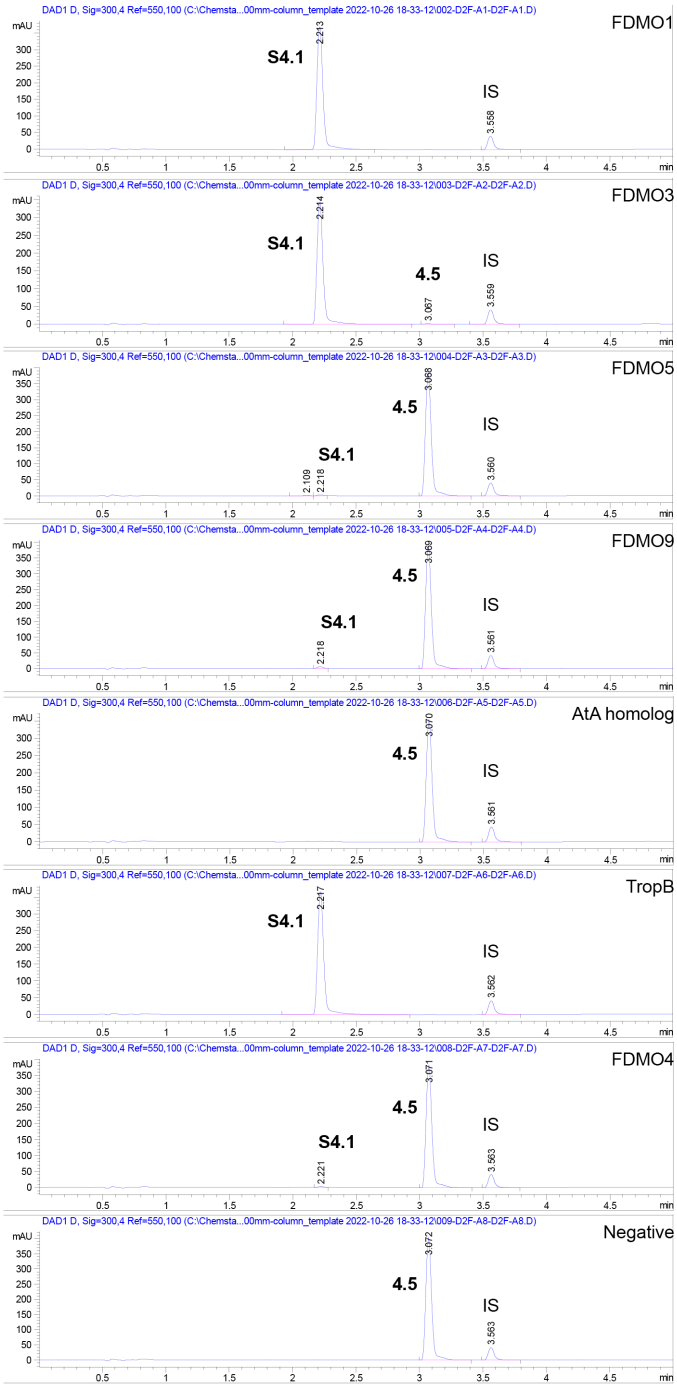
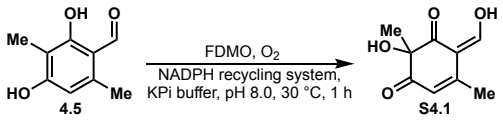


Figure S4.14. UPLC chromatograms for oxidative dearomatization to form **S4.1**.
Percentage conversions of different enzymes were calculated based on the ratio of the peak area (300 nm) of **4.5** to the peak area of IS (benzophenone) to that of the negative (no enzyme) control.

Enzymatic reactions of the substrate **4.5**



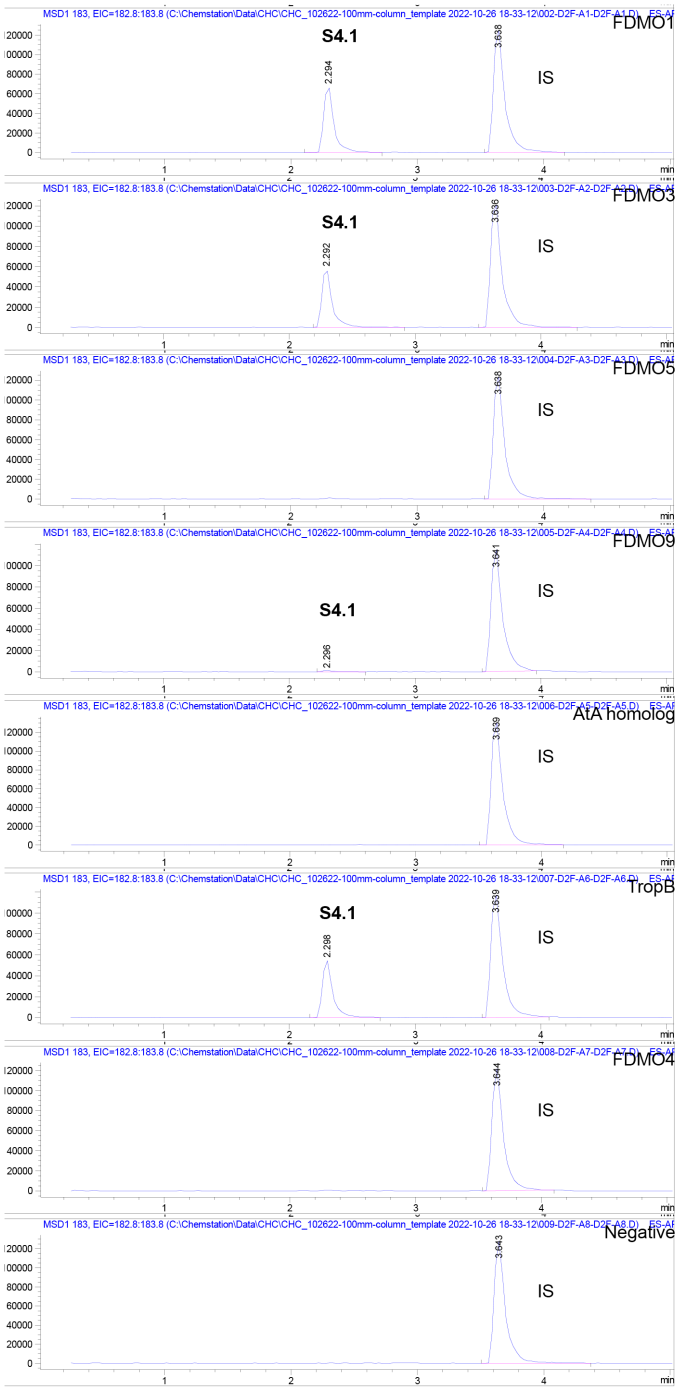
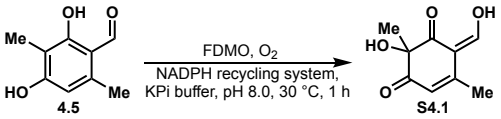


Figure S4.15. Extracted ion chromatograms of the product **S4.1** (with $[M+H]^+$ of 183.0662) of the oxidative dearomatization reaction and the internal standard benzophenone (IS, with $[M+H]^+$ of 183.0804).

Enzymatic reactions of the substrate **4.5**



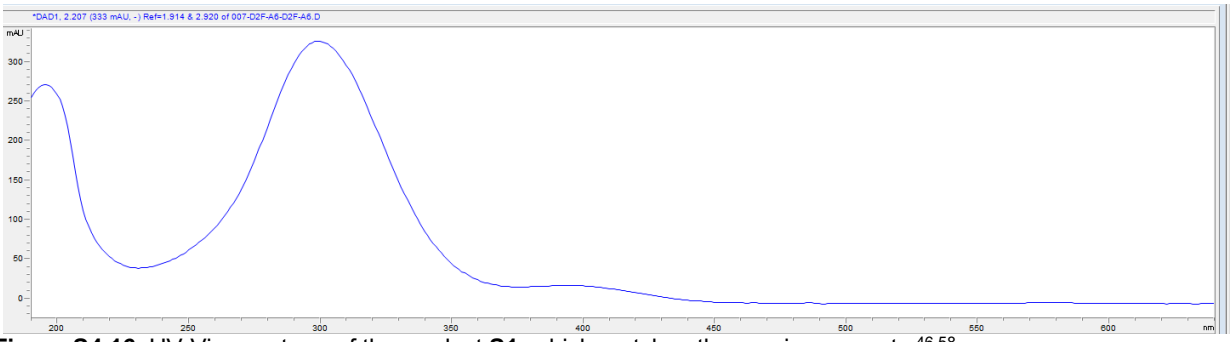
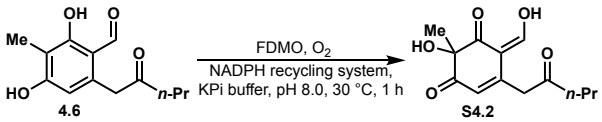


Figure S4.16. UV-Vis spectrum of the product **S1**, which matches the previous reports.[46,58]

Enzymatic reactions of the substrate **4.6**



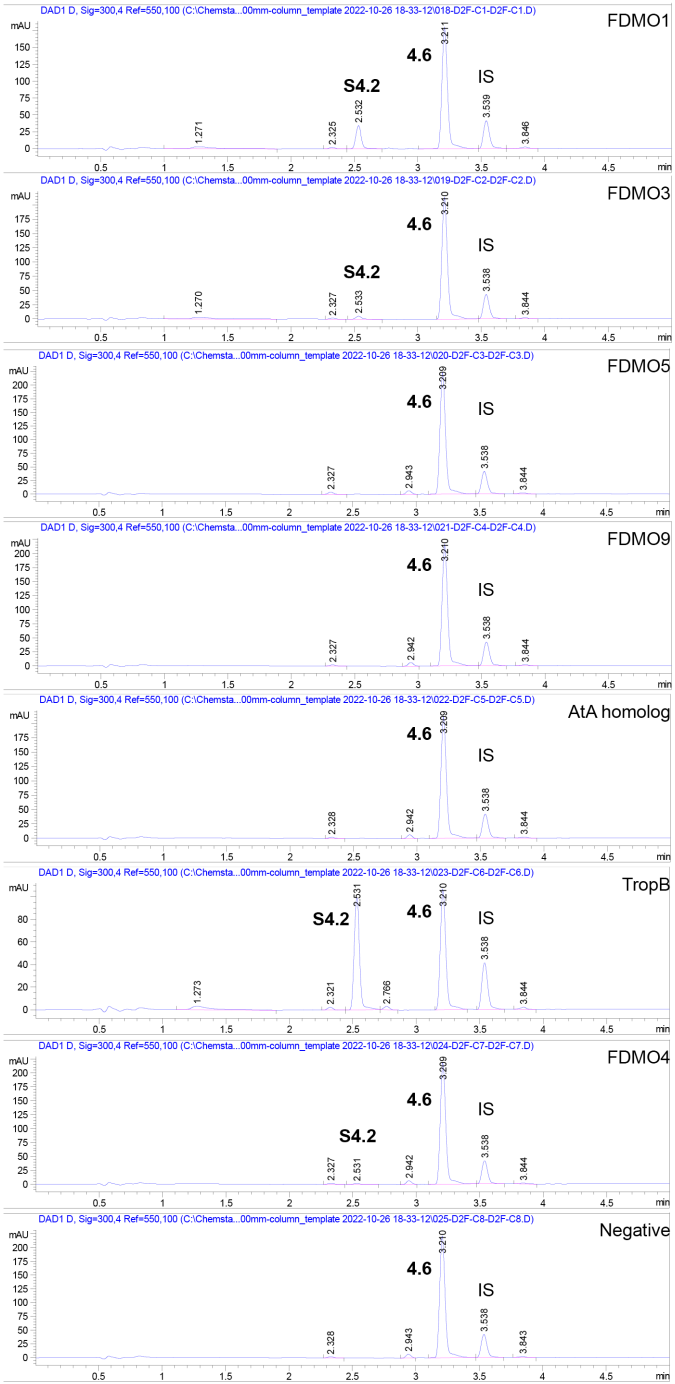
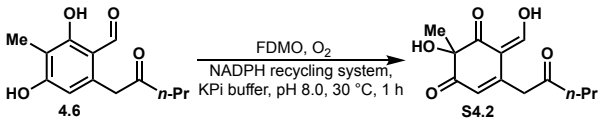


Figure S4.17. UPLC chromatograms for oxidative dearomatization to form **S4.2**.
Percentage conversions of different enzymes were calculated based on the ratio of the peak area (300 nm) of **4.6** to the peak area of IS (benzophenone) to that of the negative (no enzyme) control.

Enzymatic reactions of the substrate **4.6**



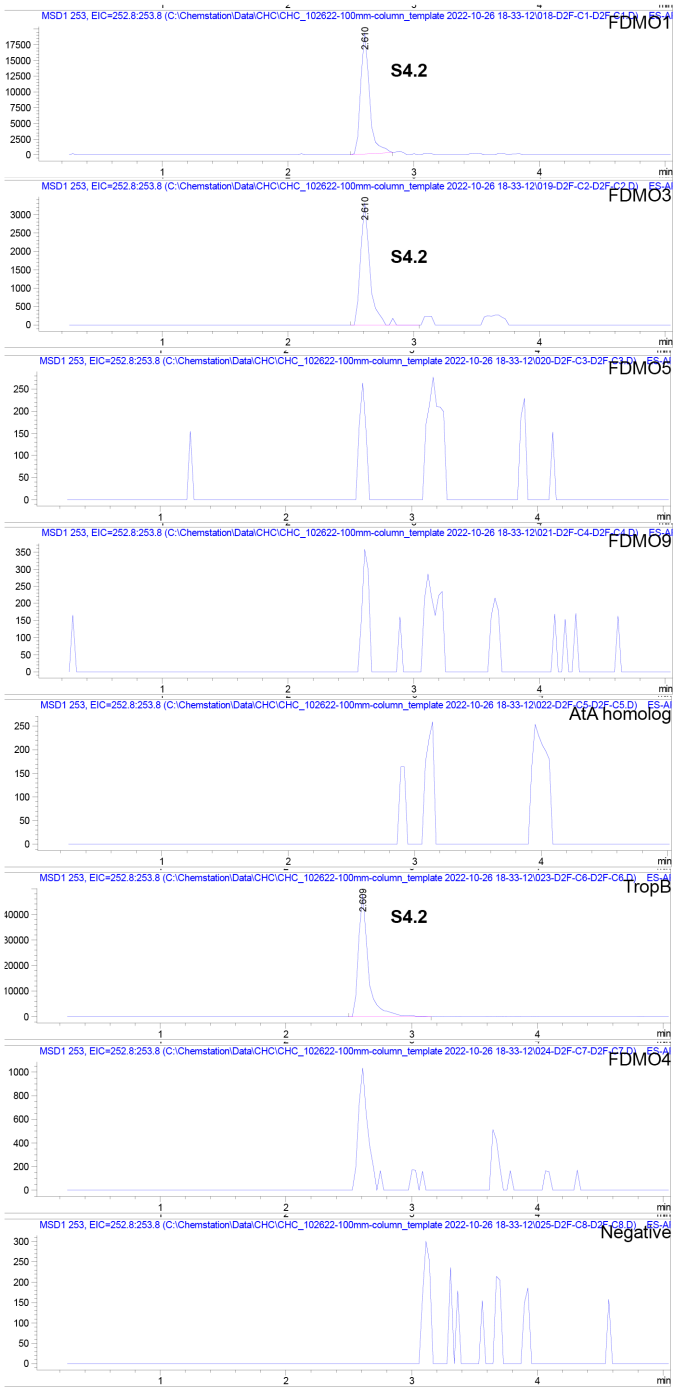
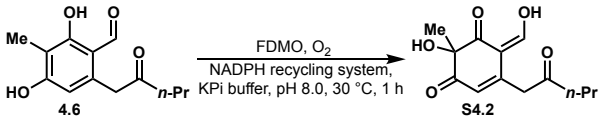


Figure S4.18. Extracted ion chromatograms of the product **S2** (with $[M+H]^+$ of 253.1071) of the oxidative dearomatization reaction.

Enzymatic reactions of the substrate **4.6**



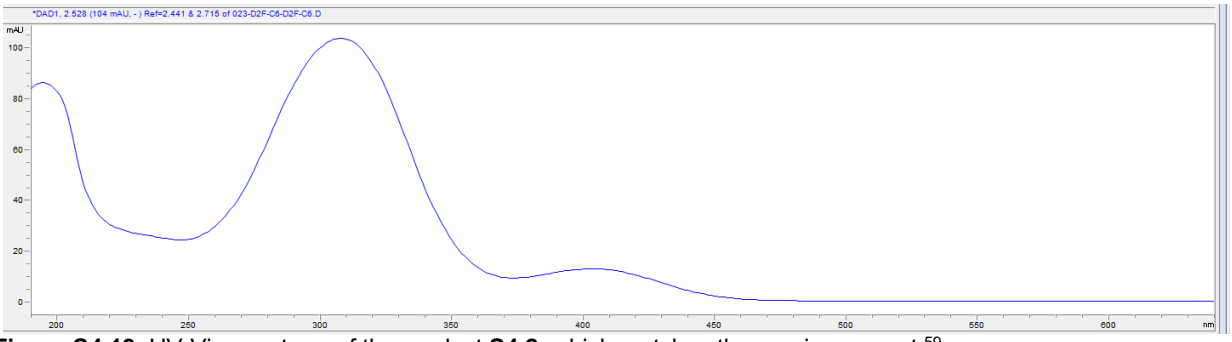
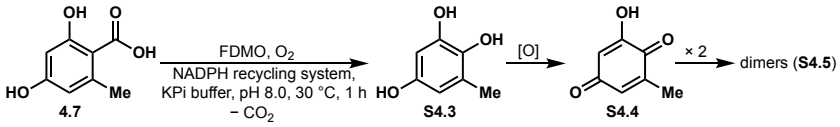


Figure S4.19. UV-Vis spectrum of the product **S4.2**, which matches the previous report.[59]

Enzymatic reactions of the substrate **4.7**



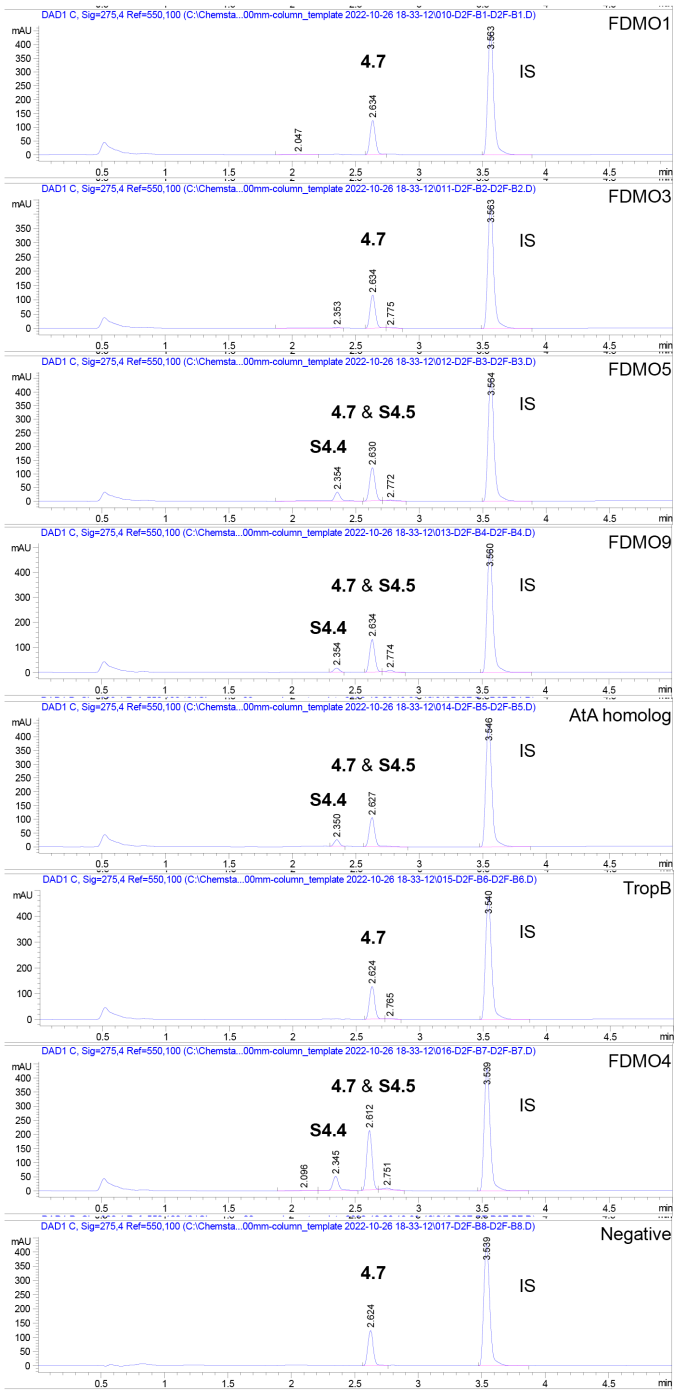
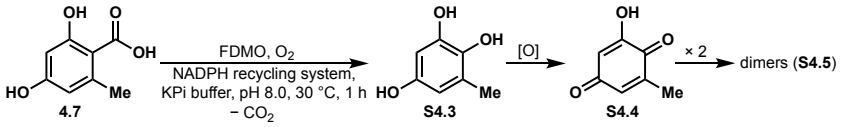


Figure S4.20. UPLC chromatograms for decarboxylative hydroxylation to form **S4.4** and **S4.5**.
The substrate **4.7** and the major product **S4.5** co-elute. Percentage conversions were calculated based on the ratio of peak area (EIC) of **4.7** to the peak area of IS (benzophenone) to that of the negative (no enzyme) control.

Enzymatic reactions of the substrate **4.7**



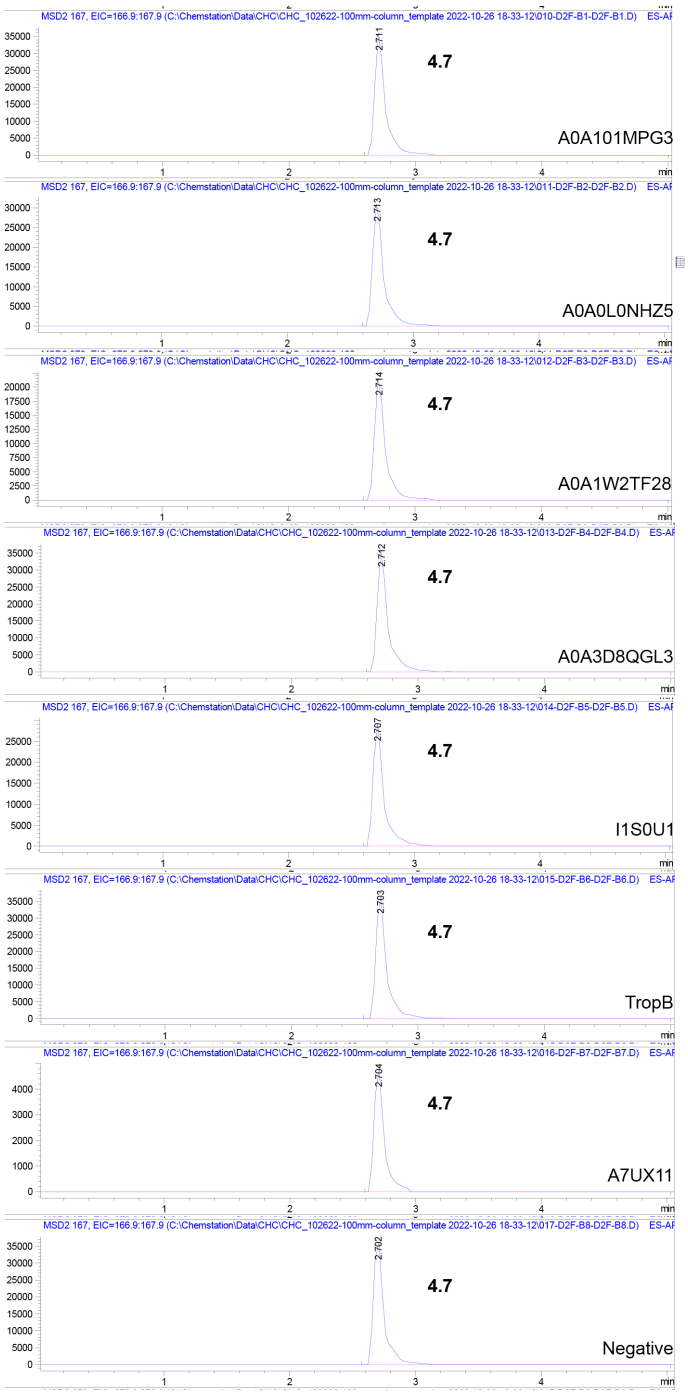
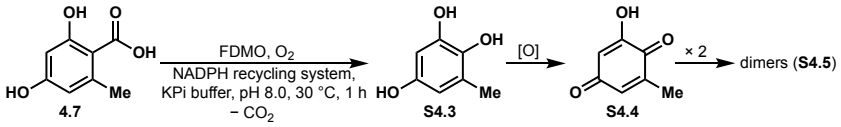


Figure S4.21. Extracted ion chromatograms of the substrate **4.7** (with [M-H]- 167.0350) used for the calculation of percentage conversions.

Enzymatic reactions of the substrate **4.7**



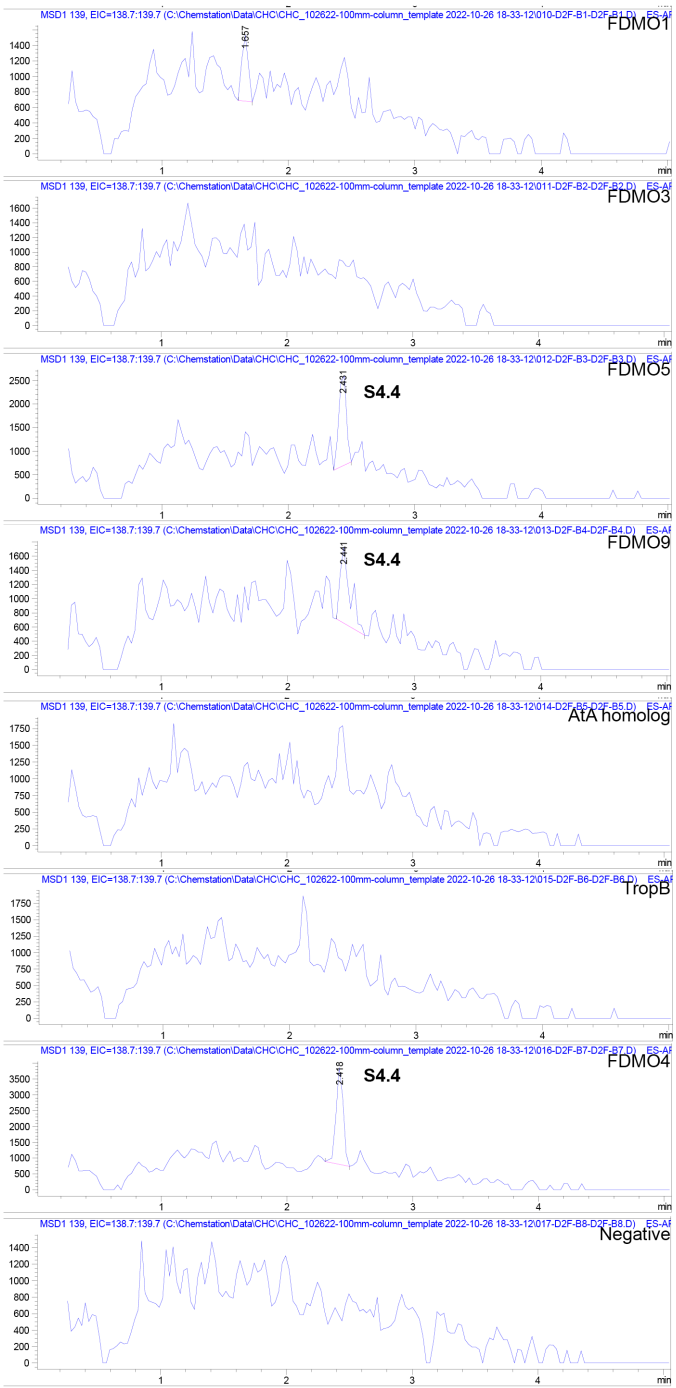
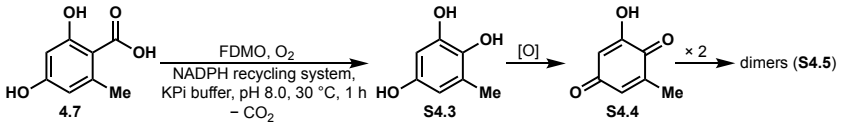


Figure S4.22. Extracted ion chromatograms of the oxidized product **S4.4** (with $[M+H]^+$ of 139.0390) of the decarboxylative hydroxylation reaction.

Enzymatic reactions of the substrate **4.7**



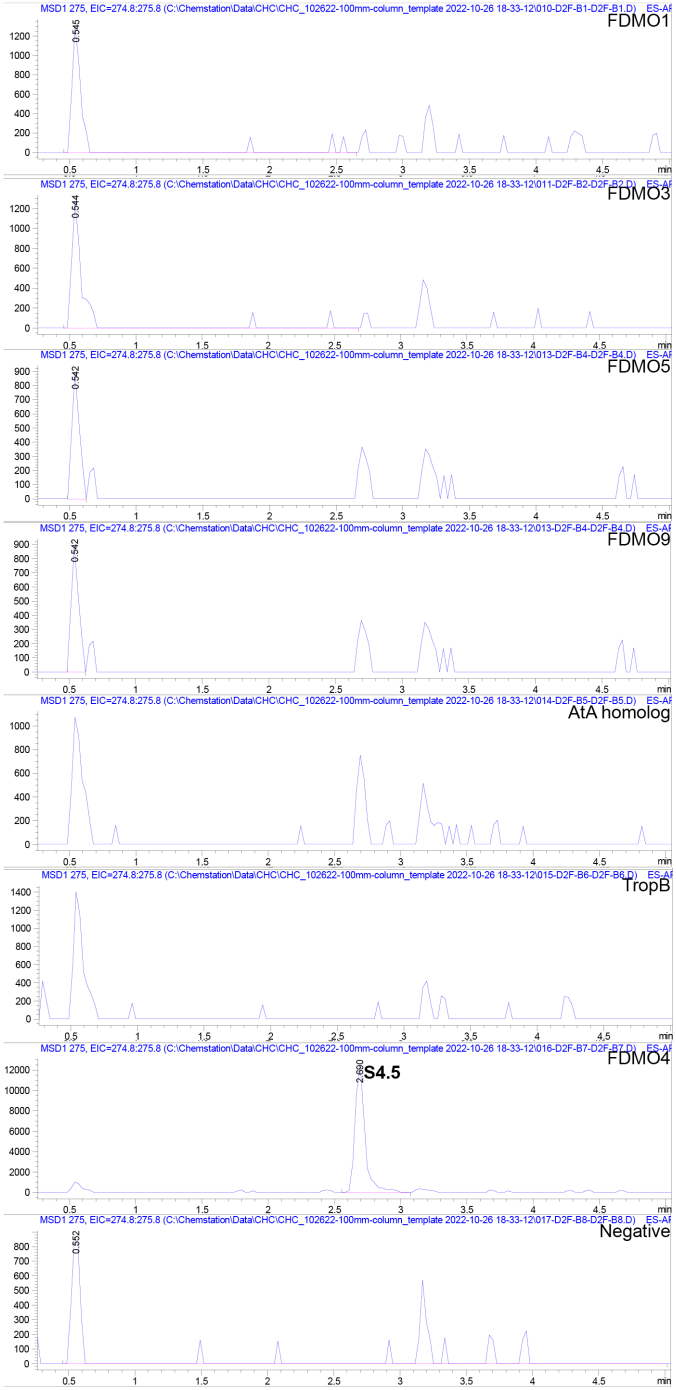
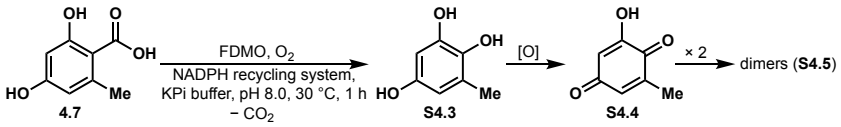


Figure S4.23. Extracted ion chromatograms of the dimerized product **S4.5** (with $[M+H]^+$ of 275.0550) of **S4.4**.

Enzymatic reactions of the substrate **4.7**



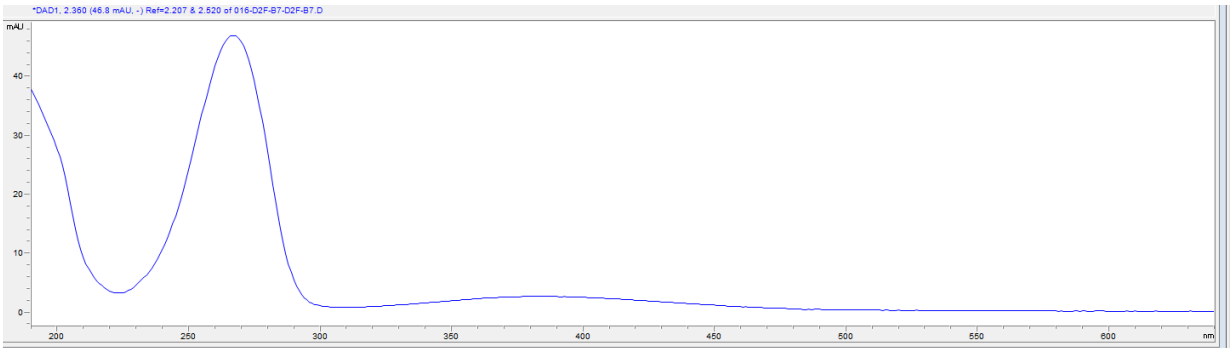


Figure S4.24. UV-Vis spectrum of the oxidized product **S4.4**.

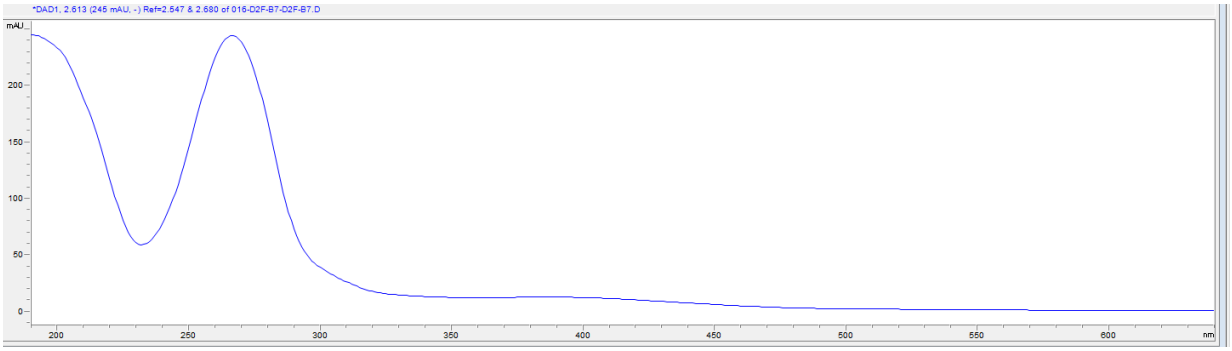
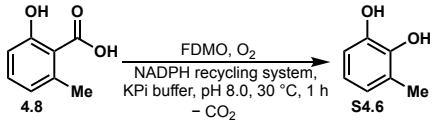


Figure S4.25. UV-Vis spectrum of the dimerized product **S4.5**.

Enzymatic reactions of the substrate **4.8**



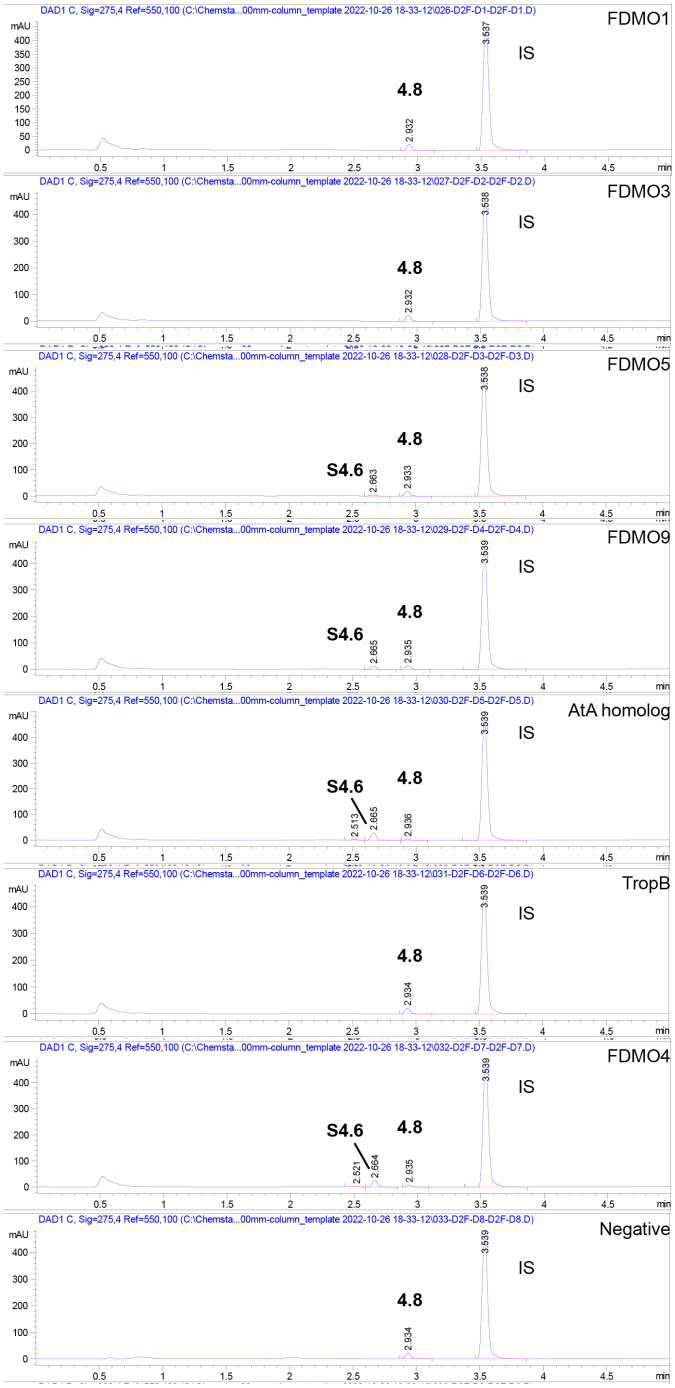
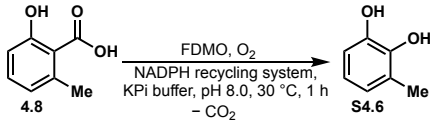


Figure S4.26. UPLC chromatograms for decarboxylative hydroxylation to form **S4.6**.
Percentage conversions of different enzymes were calculated based on the ratio of the peak area (275 nm) of **4.8** to the peak area of IS (benzophenone) to that of the negative (no enzyme) control.

Enzymatic reactions of the substrate **4.8**



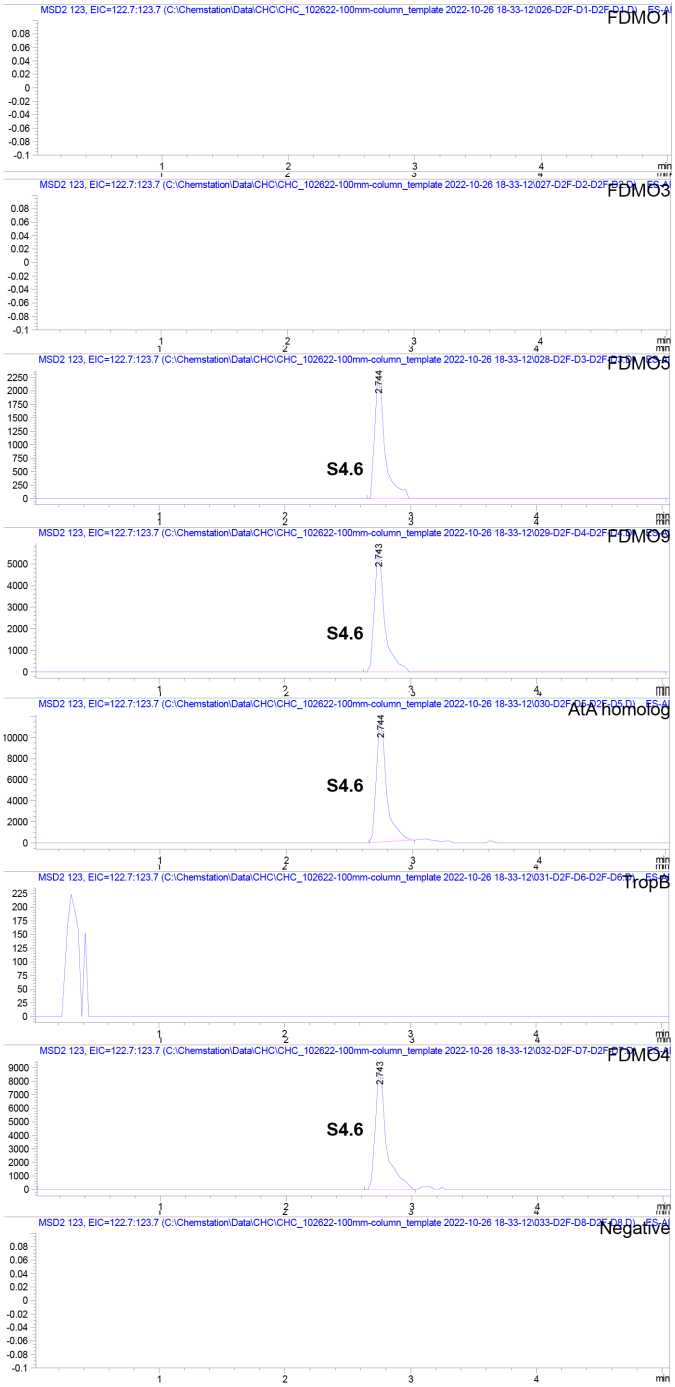
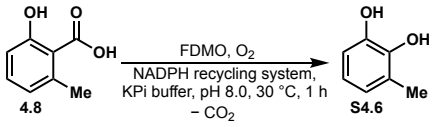


Figure S4.27. Extracted ion chromatograms of the product **S4.6** (with $[M-H]^-$ of 123.0452) of the decarboxylative hydroxylation reaction.

Enzymatic reactions of the substrate **4.8**



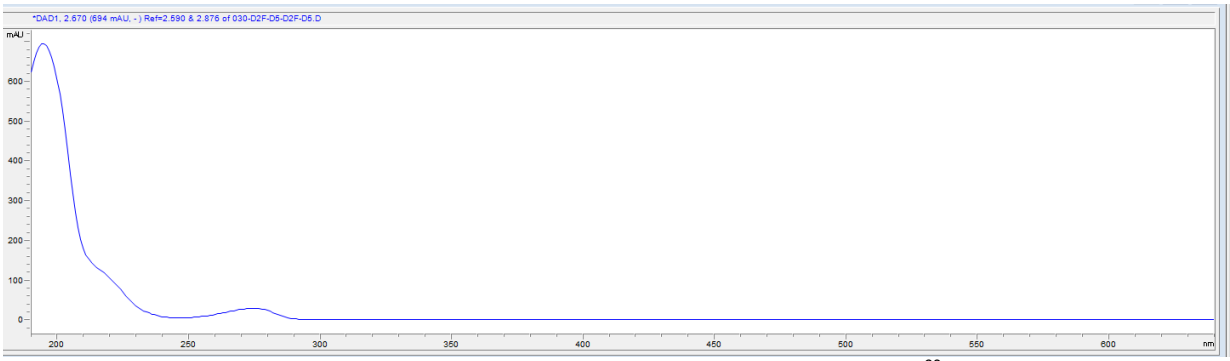


Figure S4.28. UV-Vis spectrum of the product **S4.6**, which matches the previous report.[60]

4.7 References

1. UniProt, C., UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021,** *49* (D1), D480–D489.
2. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J., et al., Pfam: The protein families database in 2021. *Nucleic Acids Res.* **2021,** *49* (D1), D412–D419.
3. Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; Pinto, B. L.; Salazar, G. A.; Bileschi, M. L.; Bork, P.; Bridge, A.; Colwell, L., et al., InterPro in 2022. *Nucleic Acids Res.* **2023,** *51* (D1), D418–D427.
4. Sayers, E. W.; Beck, J.; Bolton, E. E.; Brister, J. R.; Chan, J.; Comeau, D. C.; Connor, R.; DiCuccio, M.; Farrell, C. M.; Feldgarden, M., et al., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2024,** *52* (D1), D33–D43.
5. UniProtKB. https://www.uniprot.org/uniprotkb (accessed May 31, 2024).
6. Liu, K.; Raghavan, S.; Nelesen, S.; Linder, C. R.; Warnow, T., Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science* **2009,** *324* (5934), 1561–1564.
7. Letunic, I.; Bork, P., Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **2007,** *23* (1), 127–128.
8. Subramanian, B.; Gao, S.; Lercher, M. J.; Hu, S.; Chen, W. H., Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res.* **2019,** *47* (W1), W270–W275.
9. Price, M. N.; Dehal, P. S.; Arkin, A. P., FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **2010,** *5* (3), e9490.
10. Sanderson, M. J.; Driskell, A. C., The challenge of constructing large phylogenetic trees. *Trends Plant Sci.* **2003,** *8* (8), 374–379.
11. Atkinson, H. J.; Morris, J. H.; Ferrin, T. E.; Babbitt, P. C., Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **2009,** *4* (2), e4345.

12. Zallot, R.; Oberg, N.; Gerlt, J. A., The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019,** *58* (41), 4169–4182.
13. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003,** *13* (11), 2498–2504.
14. Oberg, N.; Zallot, R.; Gerlt, J. A., EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J. Mol. Biol.* **2023,** *435* (14), 168018.
15. Bengio, Y.; Courville, A.; Vincent, P., Representation Learning: A Review and New Perspectives. *arXiv e-prints* **2012**, arXiv:1206.5538.
16. Riesselman, A. J.; Ingraham, J. B.; Marks, D. S., Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods* **2018,** *15* (10), 816–822.
17. Ding, X.; Zou, Z.; Brooks III, C. L., Deciphering protein evolution and fitness landscapes with latent space models. *Nat. Commun.* **2019,** *10* (1), 5644.
18. Detlefsen, N. S.; Hauberg, S.; Boomsma, W., Learning meaningful representations of protein sequences. *Nat. Commun.* **2022,** *13* (1), 1914.
19. Ziegler, C.; Martin, J.; Sinner, C.; Morcos, F., Latent generative landscapes as maps of functional diversity in protein sequence space. *Nat. Commun.* **2023,** *14* (1), 2222.
20. Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M., Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods* **2019,** *16* (12), 1315–1322.
21. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J., et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021,** *118* (15).
22. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M., et al., ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022,** *44* (10), 7112–7127.
23. Unsal, S.; Atas, H.; Albayrak, M.; Turhan, K.; Acar, A. C.; Doğan, T., Learning functional properties of proteins with language models. *Nat. Mach. Intell.* **2022,** *4* (3), 227–245.
24. Le, S. Q.; Gascuel, O., An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **2008,** *25* (7), 1307–1320.
25. Vinh, N. X.; Epps, J.; Bailey, J., Information Theoretic Measures for Clusterings Comparison: is a Correction for Chance Necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, Association for Computing Machinery: Montreal, Quebec, Canada, 2009; pp 1073–1080.
26. Repka, L. M.; Chekan, J. R.; Nair, S. K.; van der Donk, W. A., Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chem. Rev.* **2017,** *117* (8), 5457–5520.
27. Tietz, J. I.; Schwalen, C. J.; Patel, P. S.; Maxson, T.; Blair, P. M.; Tai, H. C.; Zakai, U. I.; Mitchell, D. A., A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.* **2017,** *13* (5), 470–478.

28. Pag, U.; Sahl, H. G., Multiple activities in lantibiotics–models for the design of novel antibiotics? *Curr. Pharm. Des.* **2002,** *8* (9), 815–833.
29. Goto, Y.; Li, B.; Claesen, J.; Shi, Y.; Bibb, M. J.; van der Donk, W. A., Discovery of unique lanthionine synthetases reveals new mechanistic and evolutionary insights. *PLoS Biol.* **2010,** *8* (3), e1000339.
30. Knerr, P. J.; van der Donk, W. A., Discovery, biosynthesis, and engineering of lantipeptides. *Annu. Rev. Biochem.* **2012,** *81*, 479–505.
31. Ortega, M. A.; Hao, Y.; Zhang, Q.; Walker, M. C.; van der Donk, W. A.; Nair, S. K., Structure and mechanism of the tRNA-dependent lantibiotic dehydratase NisB. *Nature* **2015,** *517* (7535), 509–512.
32. van der Donk, W. A.; Nair, S. K., Structure and mechanism of lanthipeptide biosynthetic enzymes. *Curr. Opin. Struct. Biol.* **2014,** *29*, 58–66.
33. Li, B.; Yu, J. P.; Brunzelle, J. S.; Moll, G. N.; van der Donk, W. A.; Nair, S. K., Structure and mechanism of the lantibiotic cyclase involved in nisin biosynthesis. *Science* **2006,** *311* (5766), 1464–1467.
34. Rahman, I. R.; Acedo, J. Z.; Liu, X. R.; Zhu, L.; Arrington, J.; Gross, M. L.; van der Donk, W. A., Substrate Recognition by the Class II Lanthipeptide Synthetase HalM2. *ACS Chem. Biol.* **2020,** *15* (6), 1473–1486.
35. Muller, W. M.; Schmiederer, T.; Ensle, P.; Sussmuth, R. D., In vitro biosynthesis of the prepeptide of type-III lantibiotic labyrinthopeptin A2 including formation of a C-C bond as a post-translational modification. *Angew. Chem. Int. Ed. Engl.* **2010,** *49* (13), 2436–2440.
36. Kodani, S.; Hudson, M. E.; Durrant, M. C.; Buttner, M. J.; Nodwell, J. R.; Willey, J. M., The SapB morphogen is a lantibiotic-like peptide derived from the product of the developmental gene ramS in Streptomyces coelicolor. *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101* (31), 11448–11453.
37. Hernandez Garcia, A.; Nair, S. K., Structure and Function of a Class III Metal-Independent Lanthipeptide Synthetase. *ACS Cent Sci* **2023,** *9* (10), 1944–1956.
38. Goto, Y.; Okesli, A.; van der Donk, W. A., Mechanistic studies of Ser/Thr dehydration catalyzed by a member of the LanL lanthionine synthetase family. *Biochemistry* **2011,** *50* (5), 891–898.
39. Eddy, S. R., Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011,** *7* (10), e1002195.
40. Walker, M. C.; Eslami, S. M.; Hetrick, K. J.; Ackenhusen, S. E.; Mitchell, D. A.; van der Donk, W. A., Precursor peptide-targeted mining of more than one hundred thousand genomes expands the lanthipeptide natural product family. *BMC Genomics* **2020,** *21* (1), 387.
41. Zhang, Q.; Yu, Y.; Velasquez, J. E.; van der Donk, W. A., Evolution of lanthipeptide synthetases. *Proc. Natl. Acad. Sci. U. S. A.* **2012,** *109* (45), 18361–18366.
42. James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J., *An Introduction to Statistical Learning: with Applications in Python*. 1st ed. 2023. ed.; Springer International Publishing : Imprint: Springer: 2023.
43. Hsu, C.; Nisonoff, H.; Fannjiang, C.; Listgarten, J., Learning protein fitness models from evolutionary and assay-labeled data. *Nat. Biotechnol.* **2022,** *40* (7), 1114–1122.
44. Huijbers, M. M.; Montersino, S.; Westphal, A. H.; Tischler, D.; van Berkel, W. J., Flavin dependent monooxygenases. *Arch. Biochem. Biophys.* **2014,** *544*, 2–17.

45. Abood, A.; Al-Fahad, A.; Scott, A.; Hosny, A. E.-D. M. S.; Hashem, A. M.; Fattah, A. M. A.; Race, P. R.; Simpson, T. J.; Cox, R. J., Kinetic characterisation of the FAD dependent monooxygenase TropB and investigation of its biotransformation potential. *RSC Adv.* **2015,** *5* (62), 49987–49995.
46. Benítez, A. R.; Tweedy, S. E.; Dockrey, S. A. B.; Lukowski, A. L.; Wymore, T.; Khare, D.; Brooks III, C. L.; Palfey, B. A.; Smith, J. L.; Narayan, A. R. H., Structural basis for selectivity in flavin-dependent monooxygenase-catalyzed oxidative dearomatization. *ACS Catal.* **2019,** *9* (4), 3633–3640.
47. Feng, P.; Shang, Y.; Cen, K.; Wang, C., Fungal biosynthesis of the bibenzoquinone oosporein to evade insect immunity. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (36), 11365–11370.
48. Guo, C. J.; Sun, W. W.; Bruno, K. S.; Wang, C. C., Molecular genetic characterization of terreic acid pathway in Aspergillus terreus. *Org. Lett.* **2014,** *16* (20), 5250–5253.
49. You, I. S.; Murray, R. I.; Jollie, D.; Gunsalus, I. C., Purification and characterization of salicylate hydroxylase from Pseudomonas putida PpG7. *Biochem. Biophys. Res. Commun.* **1990,** *169* (3), 1049–1054.
50. Costa, D. M. A.; Gomez, S. V.; de Araujo, S. S.; Pereira, M. S.; Alves, R. B.; Favaro, D. C.; Hengge, A. C.; Nagem, R. A. P.; Brandao, T. A. S., Catalytic mechanism for the conversion of salicylate into catechol by the flavin-dependent monooxygenase salicylate hydroxylase. *Int. J. Biol. Macromol.* **2019,** *129*, 588–600.
51. Chiang, C. H.; Wymore, T.; Rodríguez Benítez, A.; Hussain, A.; Smith, J. L.; Brooks III, C. L.; Narayan, A. R. H., Deciphering the evolution of flavin-dependent monooxygenase stereoselectivity using ancestral sequence reconstruction. *Proc. Natl. Acad. Sci. U. S. A.* **2023,** *120* (15), e2218248120.
52. Champagne, S. E.; Chiang, C. H.; Gemmel, P. M.; Brooks III, C. L.; Narayan, A. R. H., Biocatalytic Stereoselective Oxidation of 2-Arylindoles. *J. Am. Chem. Soc.* **2024,** *146* (4), 2728–2735.
53. Huerta-Cepas, J.; Serra, F.; Bork, P., ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016,** *33* (6), 1635–1638.
54. Kingma, D. P.; Welling, M., Auto-Encoding Variational Bayes. *arXiv e-prints* **2013**, arXiv:1312.6114.
55. Gerlt, J. A.; Bouvier, J. T.; Davidson, D. B.; Imker, H. J.; Sadkhin, B.; Slater, D. R.; Whalen, K. L., Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **2015,** *1854* (8), 1019–1037.
56. Henikoff, S.; Henikoff, J. G., Position-based sequence weights. *J. Mol. Biol.* **1994,** *243* (4), 574–578.
57. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., Optuna: A Next-generation Hyperparameter Optimization Framework. In *KDD '19: The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery: Anchorage AK USA, 2019; pp 2623–2631.
58. Dockrey, S. A. B.; Lukowski, A. L.; Becker, M. R.; Narayan, A. R. H., Biocatalytic site- and enantioselective oxidative dearomatization of phenols. *Nat. Chem.* **2018,** *10* (2), 119–125.

59. Pyser, J. B.; Dockrey, S. A. B.; Benítez, A. R.; Joyce, L. A.; Wiscons, R. A.; Smith, J. L.; Narayan, A. R. H., Stereodivergent, Chemoenzymatic Synthesis of Azaphilone Natural Products. *J. Am. Chem. Soc.* **2019,** *141* (46), 18551–18559.
60. Prakash, D.; Pandey, J.; Tiwary, B. N.; Jain, R. K., A process optimization for bio-catalytic production of substituted catechols (3-nitrocatechol and 3-methylcatechol. *BMC Biotechnol.* **2010,** *10*, 49.

Chapter 5 Workflow Optimization of Directed Evolution: Integrating Automated Primer Design, Protein Language Models, Design of Experiments, and High-Throughput Screening

5.1 Summary

Enzyme engineering through directed evolution (DE) is a powerful but resource-intensive process due to the vast protein sequence space and the sparsity of functional proteins. To address these challenges, this work integrates computational tools and experimental strategies to streamline and optimize common DE workflows. While the approaches developed here have broader applicability, their utility is demonstrated through the engineering of a bacterial P450 enzyme for the selective dimerization of apigenin. This P450-catalyzed reaction yields biflavonoids—compounds with diverse biological activities that are difficult to synthesize using traditional chemical methods. By employing this challenging target, we illustrate how our methodologies can accelerate DE.

Firstly, an automated primer generation pipeline for site-saturation mutagenesis (SSM) was developed and hosted on Google Colab. This pipeline streamlines selection of mutation sites and primer design by automating steps such as adjusting primer lengths and calculating melting temperatures, significantly reducing manual labor. Testing the pipeline demonstrated its efficiency and accuracy, successfully generating primers that amplified 95 out of 96 target gene variants.

Using these primers, we conducted the first round of DE on P450_D12 to enhance its apigenin dimerization activity. Screening identified multiple beneficial mutations, primarily in regions associated with conformational flexibility upon substrate binding. In the second round, variants from the regions were sequenced and analyzed using protein fitness models like EVcouplings and ESM-1v to predict mutational effects. While the

In this work, enzyme engineering experiments, including mutational site identification, DNA library preparation, protein library preparation, screening, hit validation, and sequencing, were done together with Di Yang.

models showed only low to moderate correlation with experimental results for individual mutations, they highlighted potential mutational hotspots for further exploration.

Further optimization involved combining beneficial mutations through multiple site-directed mutagenesis and conducting additional rounds of SSM. This led to a P450 variant exhibiting about 38% conversion in apigenin dimerization, producing a major dimeric product that constitutes about 50% of all products. To refine the enzymatic reaction conditions, Design of Experiments (DoE) was employed, identifying key factors such as temperature and pH that significantly impact enzymatic activity.

In conclusion, the integration of automated primer design, in-silico predictions, and DoE optimization demonstrated a more efficient workflow for DE. This approach not only streamlined the process but also provided insights into mutational effects, even when targeting non-native functions. Current efforts are focused on identifying the primary dimeric product of apigenin produced by the engineered P450 enzyme.

5.2 Introduction

Enzyme engineering by directed evolution (DE) is powerful but time-consuming and resource-intensive. The core of DE involves randomly introducing mutations and screening variants to identify beneficial mutations that optimize the fitness of desired proteins. The randomness and selection of mutants compensate for our limited understanding of the effects of substitutions. However, the astronomical size of protein sequence space and the sparsity of functional proteins within the space constrain our ability to traverse the fitness landscape,[1] making DE very challenging and inefficient. For example, twelve rounds of evolution were required to reach ideal stereoselectivity in a key galactose oxidase (GOase) used in a biocatalytic cascade for the manufacture of islatravir.[2] Similarly, evolving a cyclic guanosine-adenosine synthase (cGAS) in an enzymatic cascade toward MK-1454 required ten rounds to enhance reactivity.[3] These examples underscore the labor-intensive nature of DE and highlight the need for tools that streamline the process.

Our strategy does not aim to reduce the number of engineering rounds required but instead focuses on making each round less labor-intensive by optimizing key steps in the workflow. To illustrate the utility of our approach, we selected the engineering of a

bacterial P450 enzyme for the selective dimerization of apigenin as our target. The dimerization can be mediated by the P450 enzyme through oxidative C–C or C–O coupling, which offers enzyme-controlled selectivity as previously demonstrated by our group.[4] Apigenin dimers (biflavonoids), which are formed by coupling different positions of apigenin monomers, exhibit diverse biological activities (Figure 5.1).[5-10] However, their selective chemical synthesis is challenging,[11] emphasizing the value of an enzymatic approach to achieve selective and high-yielding dimerization of apigenin.

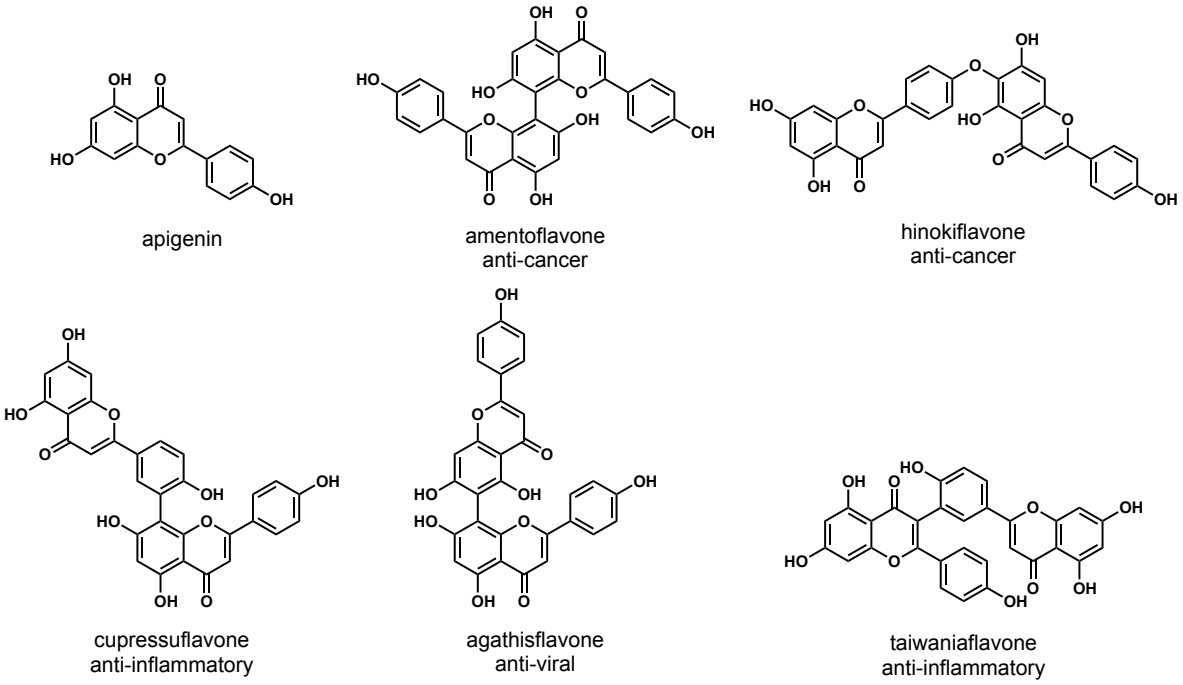


Figure 5.1. Apigenin and its dimers (biflavanoids) coupled between different positions.

Based on the DE workflow in our lab,[4] we first recognized the need to develop an automated primer generation pipeline for site-saturation mutagenesis (SSM). SSM is one of the most common methods to introduce mutations, providing a way to target specific areas for mutations while maintaining randomness by substituting an amino acid with all 19 non-wild-type amino acids.[12] Technically, SSM of a specific position can be achieved by polymerase chain reaction (PCR) using a forward and reverse primer pair, with the forward primer containing a 5' degenerate codon NNK.[13] Through PCR, the whole plasmid harboring the gene variant is amplified and can be efficiently transformed into *Escherichia coli* after nicks are ligated by T4 polynucleotide kinase and ligase. Several primer design tools such as Primer3[14] have been developed. However, they mostly are

less user-friendly, which requires users to have some programming background, and are difficult to be adapted to the SSM-based DE workflow.

Generation of variant libraries by SSM requires designing tens to hundreds of primer pairs, depending on the number of target positions. This entails a huge amount of labor if done traditionally, which in our experience involves copying and pasting an initial primer design for each position from DNA visualization software into a New England Biolabs (NEB) T_m calculator,[15] adjusting the primer length to match the target annealing temperature, and ensuring the primer ends in either G or C. The time and effort required are proportional to the number of target positions. For example, designing a primer library covering 96 positions requires repeating the aforementioned steps 192 times, which may take one to two days to accomplish. To streamline this time-consuming process, we envisioned that building an automated primer generation pipeline could provide an efficient and error-free solution to SSM primer design, significantly saving time.

Screening is a critical component of enzyme evolution. Establishing robust biocatalytic reactions can significantly enhance the effectiveness of screening and influence the success of subsequent scale-up reactions. Consequently, biocatalytic reaction conditions should be carefully optimized during an evolution campaign. Traditionally, this optimization has been achieved using the one-factor-at-a-time (OFAT) method, in which one variable is adjusted while others are kept constant (Figure 5.2). However, due to the complexity of biocatalytic reactions, the OFAT method may be neither efficient nor effective. Enzymatic reactions can be influenced by multiple factors, including buffer composition, pH, temperature, shaking speed, and cell lysate loading, among others. For oxidation-reduction reactions, such as those catalyzed by flavin-dependent monooxygenases or cytochrome P450 monooxygenases, an additional NAD(P)H recycling system with dehydrogenases is usually necessary to complete the catalytic cycle, adding even more complexity to the system. Here, we embraced design of experiments (DoE)[16] to enable a more comprehensive study of the factor space of our target biocatalytic reaction (Figure 5.2). DoE offers several advantages over the OFAT approach, including higher efficiency, the ability to study interactions between factors, and the estimation of experimental error.[17] We anticipated that using DoE could speed up the identification of optimal biocatalytic conditions for screening.

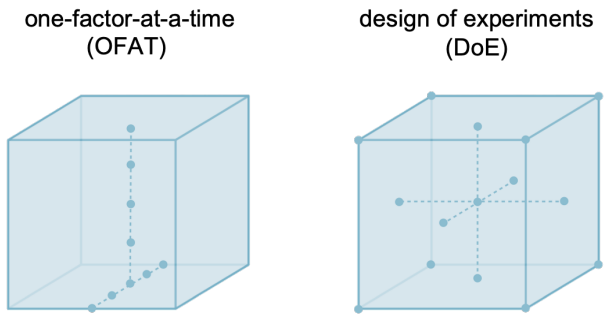


Figure 5.2. One-factor-at-a-time (OFAT) and design of experiments (DoE) for optimizing biocatalytic reactions.

Lastly, we questioned whether the engineering of a promiscuous function for apigenin dimerization by the bacterial P450 could benefit from the integration of machine learning. Specifically, we aimed to sequence every variant in the screening and feed the variants into protein fitness models such as a Potts model[18] or the Evolutionary Scale Modeling (ESM)-1v transformer[19], to evaluate whether the protein fitness predictions correlate with experimental reactivity data. This study was expected to determine whether machine learning, especially protein language models, is useful for predicting the mutational effects in DE targeting non-native (promiscuous) functions, and to provide guidance for future machine learning-guided DE.

5.3 Results and Discussion

Design and development of an automated primer generation pipeline. The primary purpose of this pipeline is to simplify the generation of site-saturation mutagenesis (SSM) primer pairs required for constructing enzyme variant libraries. Users can easily generate these primers by inputting their enzyme's DNA and protein sequences, along with the corresponding protein model or structure (Figure 5.3). The pipeline is hosted on Google Colab, making it accessible to anyone with the link, without requiring the setup of Python environments on local computers. This pipeline is currently optimized for DE efforts targeting enzyme reactivity or selectivity by identifying residues based on their proximity to the active site. However, the pipeline can be adapted for other engineering goals, such as enhancing enzyme solubility.

The process begins with a BLASTp-based homology search,[20] using the enzyme of interest as a query. This search can be performed within Colab, as long as no more than 3,000 homologous enzymes are targeted (Figure 5.4). For larger datasets or users

preferring alternative homology search tools or databases, results from external searches can be uploaded in FASTA format. After the homology search, the pipeline aligns the sequences using MAFFT[21] by default, and calculates residue conservation relative to the query enzyme (Figure 5.4). To select mutation sites, the query enzyme's structure or model is required, and users can define both low and high conservation thresholds to control whether they include or exclude lowly or highly conserved positions. This flexibility allows users to fix residues that are evolutionarily important or less significant during the SSM process.

Two options are provided for defining the active site center. Users can either select the geometric center of a ligand in the enzyme's structure or model, or define it by selecting the geometric center of several residues known to be near the active site (Figure 5.4). The final residues for SSM are chosen based on the number of mutation sites and their proximity to the active site. For further customization, users can define a region that only includes residues within a specified distance cutoff from the active site. Finally, the pipeline requires the DNA sequence of the query enzyme and outputs all primer pairs ready to be ordered (Figure 5.4).

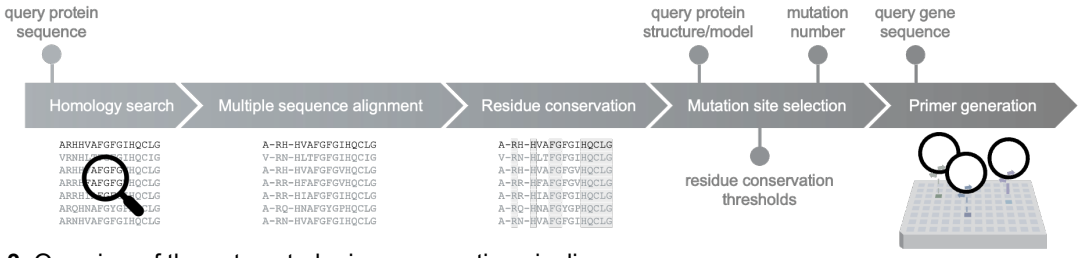


Figure 5.3. Overview of the automated primer generation pipeline.
The pipeline includes five modules including homology search, multiple sequence alignment, residue conservation calculation, mutation site selection, and primer generation.

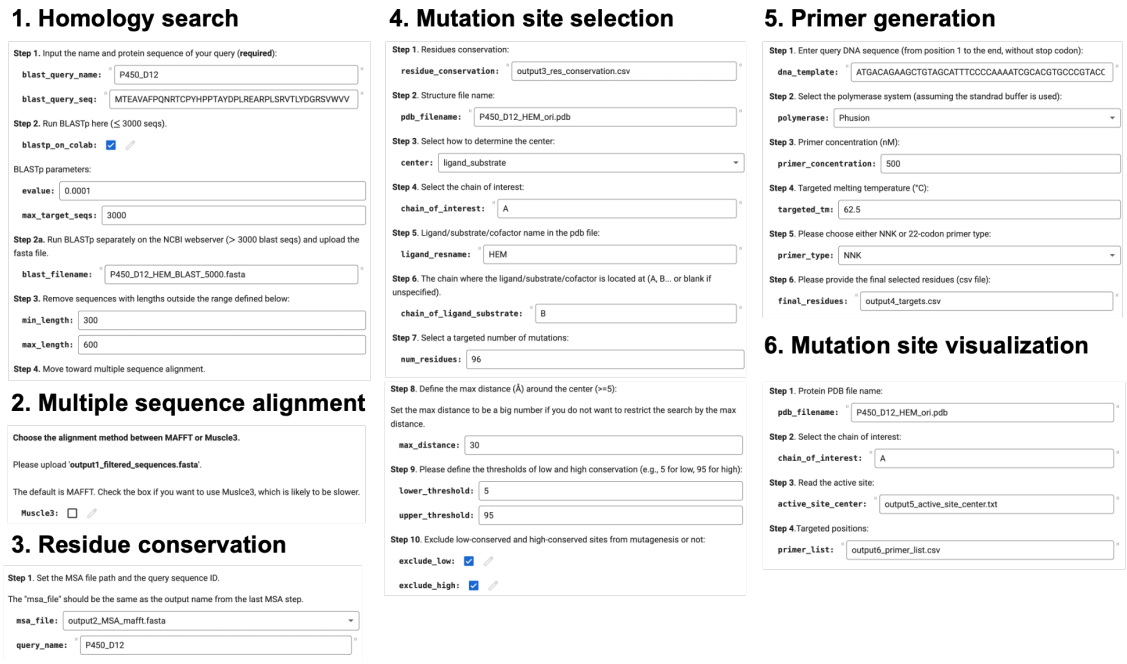


Figure 5.4. Interfaces of the automated primer generation pipeline on Google Colab.

The T_m calculation in this pipeline follows the method used by the NEB T_m calculator, ensuring compatibility with NEB polymerases, which are the primary source of widely-used polymerases. The T_m values are calculated based on the nearest-neighbor thermodynamics method[22] and are further corrected for salt concentration according to the specific polymerase being used.[23-24] Since detailed salt concentrations for all polymerase systems were not provided, they were back-calculated using multiple primer inputs. Within the pipeline, users can set their target annealing temperature (T_m), select between different polymerase systems, and choose between NNK or 22-codon primer design options (Figure 5.4). The final module of the pipeline can also function as a standalone primer generator for cases where individual primer pairs are needed for subsequent rounds of evolution.

Testing the automated primer generation pipeline. To validate the automated primer generation pipeline, we sought to determine whether the pipeline could streamline primer design and produce primers capable of successfully amplifying gene variants. A bacterial P450 from *Streptomyces ficellus* (UniProt ID: A0A6I6FIL6) was selected from our P450 library as the evolution target for cross-coupling reactions. This P450 enzyme was chosen due to its broader substrate scope compared to other P450s, as demonstrated by library screening performed by Lauren Murray. Initial reactivity in the dimerization of apigenin

using clarified P450 cell lysate was observed, making this promiscuous enzyme a promising starting point for evolution aimed at enhancing selective apigenin dimerization. This reaction could ultimately provide access to a variety of biflavonoid natural products, each of which is challenging to synthesize using traditional chemical methods.

Starting with an AlphaFold model of P450_D12, with a heme group filled in by AlphaFill,[25] residues around the geometric center of the heme were selected. Residues exhibiting greater than 95% conservation (based on a MAFFT alignment of up to 5,000 homologous sequences from the homology search) were excluded. The pipeline identified 96 target residues and automatically generated 96 primer sets for site-saturation mutagenesis (SSM). These target residues were randomly mutated to the other 19 amino acids using SSM. To assess the primers' performance, PCR samples for each residue were analyzed via gel electrophoresis (Figure S5.1). The results indicated that 95 out of 96 plasmids harboring gene variants were successfully amplified. Of these, 84 showed a single major DNA product (Table 5.1), whereas 11 contained additional undesired DNA products, which could be removed through gel extraction. In summary, the pipeline provides a much more efficient way to generate all SSM primers in minutes than our traditional approach by manual design.

Table 5.1. DNA amplification by polymerase chain reaction with SSM primers.

	Number
Successful amplification ▪ one major product ▪ multiple products	95 (84) (11)
Failed amplification	1
Total	96

Evolution of P450_D12 for apigenin dimerization. With the SSM primers in hand, Di Yang and I embarked on the first round of evolution of P450_D12 using SSM first aiming for higher reactivity of apigenin dimerization. For the P450 to function, the enzyme is tethered to a redox partner, RhFRed,[26-27] originally from *Rhodococcus aetherivorans* by a two-amino-acid linker EF (Figure 5.5). The system enables the P450 to acquire electrons transported from the fused reductase and ferrodoxin required for catalytic activity. Variants generated through SSM were screened using *E.coli* cell lysate in 96-well plates and each reaction was analyzed for the formation of apigenin dimers by RapidFire-mass spectrometry (MS, Figure 5.5). The screen resulted in the identification of multiple beneficial substitutions including S55K, T86L, H97R, I104M, etc (Figure 5.6). Based on the assignment of the secondary structures of the P450 by SecStrAnnotator,[28] these substitutions are mostly at the B-helix, B′-C loop, and C-helix of P450s, where are reported to have large conformational flexibility upon substrate binding.[29-33]

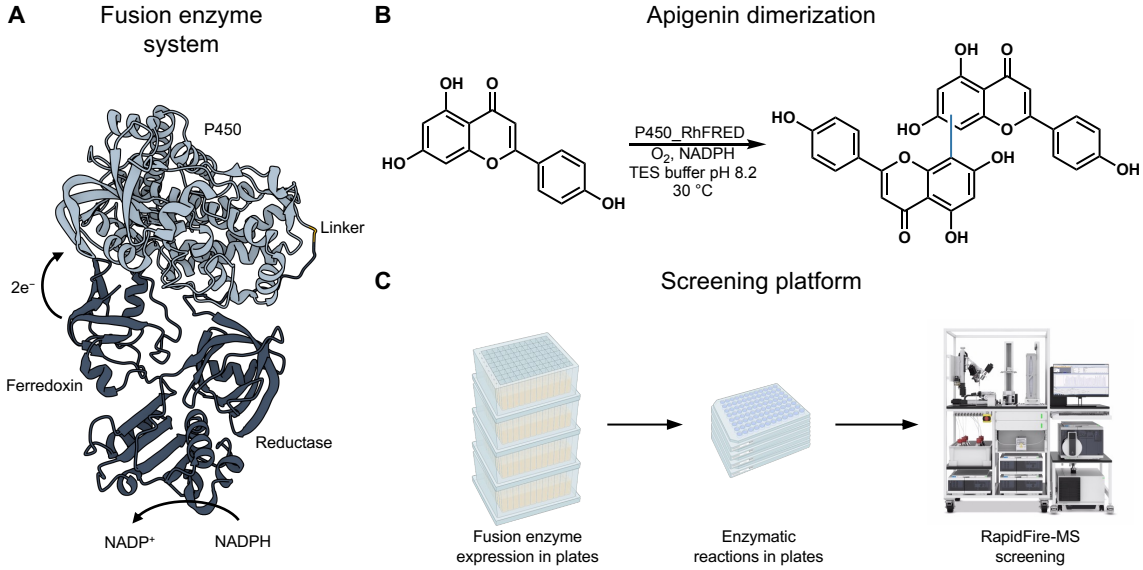


Figure 5.5. Evolution of a P450 enzyme fused with a redox partner, RhFRed, for apigenin dimerization using RapidFire-mass spectrometry (MS)-based platform.

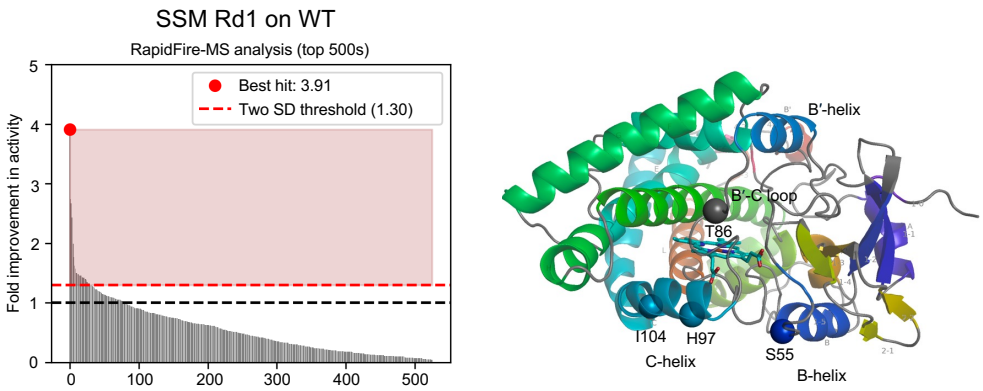


Figure 5.6. First round of site-saturation mutagenesis and major hits highlighted in spheres on the P450_D12 model.

Evaluation of protein fitness models by the data of P450 evolution. Using the top hit T86L as our template, we proceeded to the second round of site-saturation mutagenesis (SSM). Unlike the first round, mutations in the second round were categorized and pooled into four groups based on their regions: G1 (positions 50–119), G2 (positions 144–208), G3 (positions 228–288), and G4 (positions 309–389). This design enabled the subsequent sequencing of every variant using evSeq.[34] In this round, two separate experiments were conducted: one targeting mutations in G1 and G2, and the other targeting mutations in G2 and G3. Given that most beneficial mutations in the first round originated from G1, we decided to evaluate the potential of using in-silico protein fitness prediction in DE by sequencing every variant from G1. G1 variants from nine 96-well plates were barcoded, normalized, pooled into one sample, and submitted to a NGS-service provider. Sequencing results were deconvoluted using scripts from evSeq, revealing 483 qualified data points, among which 253 were unique mutations (Table 5.2).

Table 5.2. evSeq results of every variant in the group 1 (positions 55–119).

Total samples ▪ unknown ▪ positive control ▪ negative control	768 (756) (9) (3)
Template number	56 (47 + 9)
Qualified data points	483
Qualified unique mutations	253
Total mutation number	608 (32 * 19)
Library coverage	0.42

The sequence data of the variants were input into two protein fitness models, EVcouplings (Potts model) and Evolutionary Scale Modeling (ESM)-1v, to predict the mutational effects of each variant. These effects were calculated based on the log odds

ratio of sequence probabilities between the template (T86L) and the mutant sequences (Table S5.1). We envisioned that this approach could help explore the potential of using in-silico prediction to guide the evolution of P450. To begin, we investigated whether EVcouplings and ESM-1v exhibited similar trends in predicting mutational effects. The prediction scores from both models were plotted against each other, yielding a correlation coefficient R of 0.87, indicating that both models provide similar predictions of mutational effects (Figure 5.7). We then correlated the RapidFire-MS screening data of the qualified variants with the predictions from each model (EVcouplings_epistatic, EVcouplings_independent, and ESM-1v). The results showed that the prediction scores from all models had only moderate correlation with the experimental data and perfomred comparably (Figure 5.7 and S5.1), highlighting a potential limitatation of using the evolutionary models trained on sequences with native functions to predict mutational effects on non-native functions. However, what has been learned from these evolutionary models can still improve performance and generalizability of machine learning tasks for non-native functions through transfer learning in the future.

The EVcouplings model (epistatic) demonstrated the highest correlation with the experimental data among all model tested. Although the model could not accurately predict individual mutations, we investigated whether it could provide position-level predictions to guide us in identifying positions that are likely to accept substitutions. Mean prediction scores to other 19 amino acids for positions 15–397 of P450_D12 (T86L) were calculated (Table S5.2), and its AlphaFold protein model was color-coded based on these scores with blue, white, and red representing low, medium, and high substitution likelihoods, respectively (Figure 5.8). The top and bottom twenty positions and their corresponding secondary structures are listed in Table 5.3. Generally, residues at the B-B′, B′-C, and H-I loops were more favorable to substitutions, while residues at the C-, I-, K-, L-helices, and the K′-L loop were less tolerant of changes. The top hit, F71L, acquired in the second SSM was from the second most favorable position for changes. In addition to our own data, the analysis aligns with numerous examples of P450 engineering. For instance, a major substitution in P450 BM3 (F87A), which activated the oxidative hydroxylation activity of testosterone, is located at the B′-C loop.[35] In CYP105A1, two key residues, R73 and R84, responsible for the hydroxylation of vitamin D_3 of, are located at

the B-B′ loop and the B′-helix, respectively.[36] In summary, not limited to the enzyme, we envision that this position-level prediction of mutational effects could be potentially useful for the hotspot identification in future engineering.

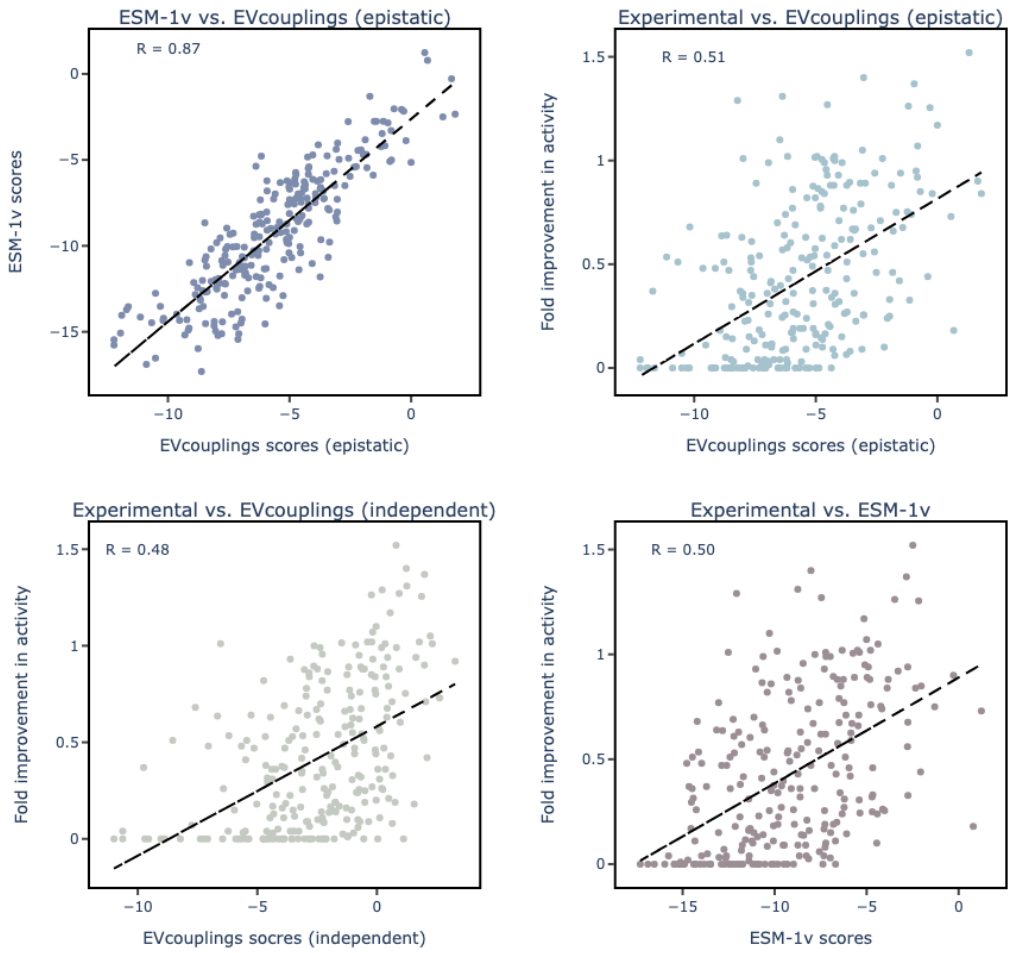


Figure 5.7. Predicted mutational effects by EVcouplings and ESM-1v.
Correlation of predicted mutational effects by EVcouplings and ESM-1v models with each other (top left) and with experimental fold improvement in activity (top right, bottom left, bottom right).

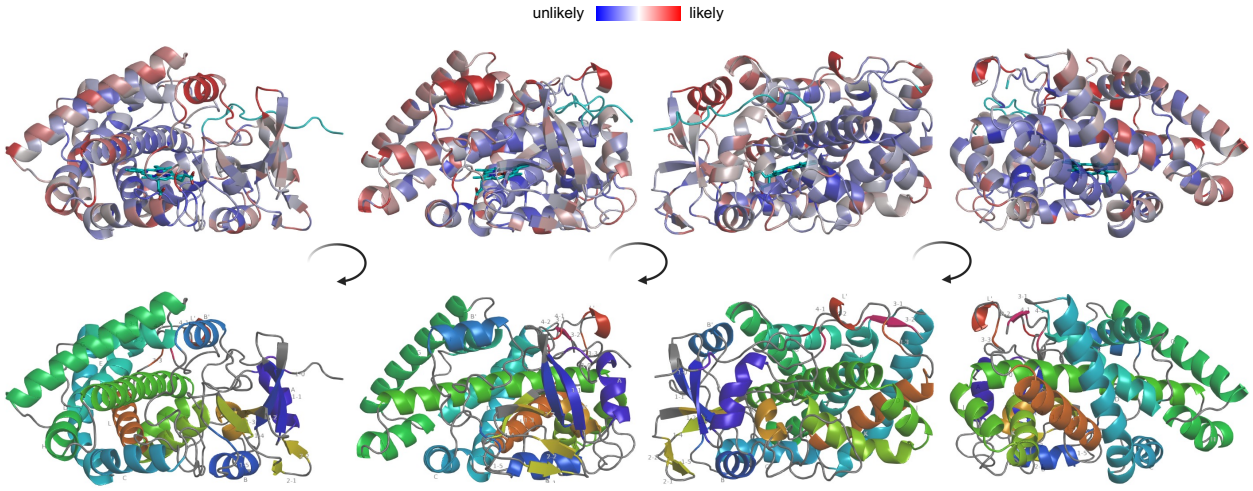


Figure 5.8. Mean prediction scores by EVcouplings for positions 15–397 of P450_D12 (T86L). Substitution likelihood from low, medium to high is shown in blue, white, and red (top row). Key secondary structures of P450 enzymes (bottom row).

Table 5.3. Top and bottom 20 positions in P450_D12 (T86L) ranked by their mean prediction scores.

Top 20	Positions	Prediction scores	Secondary structures	Bottom 20	Positions	Prediction scores	Secondary structures
1	221	-0.6646	H-I loop	20	235	-9.3151	I-helix
2	71	-0.6647	B-B′ loop	19	158	-9.3362	E-F loop
3	91	-0.8597	B′-C loop	18	347	-9.4320	K′-L loop
4	83	-0.9785	B′-C loop	17	276	-9.4346	K-helix
5	224	-1.1172	H-I loop	16	345	-9.4615	K′-L loop
6	86	-1.1931	B′-C loop	15	343	-9.4709	K′-L loop
7	219	-1.2349	H-I loop	14	156	-9.5033	E-F loop
8	202	-1.3612	G-H loop	13	242	-9.5422	I-helix
9	122	-1.6358	D-helix	12	348	-9.6195	K′-L loop
10	52	-1.7388	B-helix	11	371	-9.6361	L-L’ loop
11	81	-1.7984	B′-C loop	10	92	-9.8301	B′-C loop
12	177	-1.8016	F-G loop	9	100	-9.8824	C-helix
13	131	-1.8033	D-helix	8	353	-9.9467	L-helix
14	72	-1.8566	B-B′ loop	7	139	-10.0667	E-helix
15	226	-1.9260	I-helix	6	279	-10.0673	K-helix
16	70	-2.0090	B-B′ loop	5	364	-10.1143	L-helix
17	191	-2.0189	G-helix	4	341	-10.2098	K′-L loop
18	266	-2.0784	J-K loop	3	107	-10.2491	C-helix
19	97	-2.1248	C-helix	2	277	-10.3447	K-helix
20	203	-2.1809	G-H loop	1	253	-10.5499	I-helix

Multiple site-directed mutagensis and subsequent evolution. We did not find significantly beneficial mutations in the second SSM. We proceeded with combining all beneficial mutations in the first SSM using multiple site-directed mutagenesis (MSDM), which resulted in 2.2-fold improvement compared with the template T86L (Figure 5.9). The triply-mutated variant with S55R, T86L, and I104M was selected as the new template for two subsequent rounds of SSM, in which more sites were targeted. To enable a better coverage of variants in screening, newly-ordered primers were designed using the 22-

codon trick instead of the NNK design,[37] which decreases the number of variants statistically required for 95% library coverage from 96 using NNK to 66 using the 22-codon trick.[37] Although using the 22-codon trick costs twice as much as the NNK design on primers, the higher coverage of this trick reduces the resources and screening effort required in each step after the library generation. Following the third and fourth SSM, more slightly beneficial mutations were identified and were combined using MSDM (Figure 5.9). The final top hit is a variant with H16N, H47L, S55R, V75Q, T86L, H97R, I104M, and Q169L. At this stage, we had engineered a variant achieving an overall apigenin dimerization activity with ~38% conversion based on peak areas, producing a major dimeric product that constitutes approximately 50% of the total.

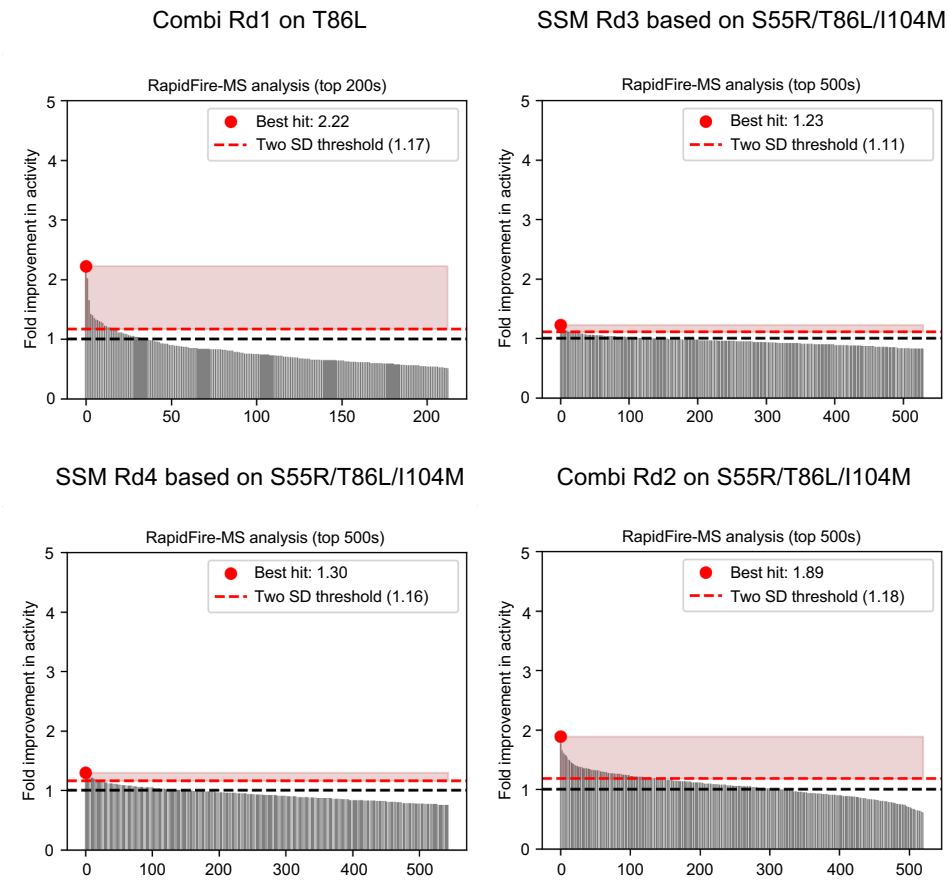


Figure 5.9. Multiple site-directed mutagenesis and site-saturation mutagenesis identified more beneficial mutations.

Optimization of the P450-catalyzed dimerization reaction by DoE. The chemical synthesis of any dimeric product standards of apigenin can be challenging. To overcome this challenge, our strategy was to scale up the enzymatic reaction using the top variant

and to extract and identify the major dimeric product. Before the scale-up reaction, we sought to find the optimal enzymatic dimerization reaction using design of experiments (DoE). Multiple critical factors, such as buffer, pH, sodium phosphite concentration, phosphite dehydrogenase (PTDH) concentration, were investigated in three runs of DoEs (Table 5.4). Each condition was run in triplicate and 31 conditions were included in one DoE. Through three runs of DoEs, two major factors, temperature and pH, were shown to be significant at $\alpha = 0.05$. Two other factors such as the TES buffer and shaking speed at 350 rpm, though not shown as significant, tended to favor the yield and robustness of the reaction. The amount of sodium phosphite and PTDH does not have any significant effect on the yield. In summary, the optimal reaction condition was found to be in the TES buffer, pH 8.2 with 20 μM PTDH, 2 mM $NADP^+$, and 10–100 mM soudium phosphite at 30 °C, which is the same as the default screening condition.

Table 5.4. Three runs of design of experiments.

First run				
Factors	**Levels**			
Buffer	Tris-HCl	HEPES	**TES**	Tricine
pH	7.5		**8.2**	
Na_2HPO_3 concetration (mM)	20		**100**	
PTDH concentration (μM)	**20**		40	
$NADP^+$ concentration (mM)	0.2		0.4	
Second run				
Factors	**Levels**			
Temperature (°C)	**30**		32	
pH	7.2	7.5	7.8	**8.2**
Na_2HPO_3 concetration (mM)	20		**100**	
PTDH concentration (μM)	10	**20**	40	
$NADP^+$ concentration (mM)	0.2		0.4	
Third run				
Factors	**Levels**			
Shaking speed (rpm)	200		**350**	
Na_2HPO_3 concetration (mM)	10	20	40	60
PTDH concentration (μM)	10	**20**	30	
$NADP^+$ concentration (mM)	1		**2**	

* The concentrations of Na_2HPO_3, PTDH, and $NADP^+$ shown are those in the reaction mixture portion, which was diluted to half after mixing with the cell lysate portion.
* The bolded conditions are those used in regular high-throughput reactions with cell lysate.
* PTDH: phosphite dehydrogenase

5.4 Conclusions

In conclusion, the work in this chapter began as an initiative to test an automated primer generation pipeline, ultimately evolving into a directed evolution campaign to engineer P450 for selective biocatalytic dimerization of apigenin. Although directed evolution is a powerful tool for enzyme engineering, it often faces challenges such as limited screening throughput and high labor demands. To address this, we developed an automated

pipeline that significantly streamlined the primer design process. Building on this improvement, we explored strategies to enhance the efficiency and effectiveness of directed evolution further. Specifically, we investigated whether evolutionary information from protein fitness models trained on wild-type enzymes could predict mutational effects for non-native reactions. Our findings indicated that predictions from both EVcouplings and ESM-1v exhibited weak to moderate correlations with experimental data for individual mutations. However, evidence from our evolution experiments and previous studies suggests that these models may be valuable for identifying mutational hotspots. We envision that predictions from protein models could be integrated into future directed evolution campaigns, using site-saturation mutagenesis (SSM) to target key positions. Current efforts are focused on characterizing the primary dimeric product of apigenin.

5.5 Experimental Procedures

5.5.1 Plasmids, DNA and protein sequences

General information

The gene encoding the P450RhF reductase domain (RhFRed) in a pET28b vector was a generous gift from Professor David Sherman's lab at the University of Michigan. The codon-optimized P450_D12 gene in a pET28b vector harboring the RhFRed domain was purchased from Twist Bioscience. The DNA and amino acid characters highlighted in red are not from the genes but from the templates instead.

Sequence information

P450_D12 (A0A6I6FIL6) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGACAGAAGCTGTAGCATTTCCCCAAAATCGCACGTGCCCGTACCATCCGCCAACGGCCTACGA
TCCGCTGCGTGAAGCGAGGCCACTTAGCCGTGTGACCCTGTATGATGGGCGTTCCGTTTGGGTT
GTCACGGGTCACGCAGCTGCACGTCAGTTGCTGTCGGATCAACGTCTGTCGACCGATAGCACCA
ATGAAGCCTTCCCGATTTTCGCGGAGCGCGTGGCTGCGGTTCGTTCTCGTCGTCCGAGACGTAC
GGCATTGCTGGGCGAAGACGACCCGGTGCATCATGCTCAAAGACGGATGCTGATCCCGTCTTTT
ACCTTAAAGCGCACCGCCGACCTTCGTCCGCGGATTCAGCGTGCGGTCGACGAACTCATCGACG
CCATGACCGAACAAGGTCCGCCTGCTGAATTGGTTAGCGCGTTCGCCCTGCCGGTCCCGAGCAT
CGTGATATGCGATCTGTTGGGCGTACCGTACGCGGACCACGATTTCTTCGAGGGCCAAAGCAGA
CGTCTGTTACGCGGTAGCACTGCGGAGGAGATCGAAGACGCACGCGACCAGCTGGAAGGTTATC
TGGGCGACTTGGTCGAGCGCAAACGTGACAAGCCGGGTGATGGTGTGCTGGACGATCTGATCGC
GCAGCAATTGGCGCATGGTCACCTGGACAAGCAGCAGTTGGTGGAGCTCGCTACGATTCTGTTG

GTGGCCGGTCACGAAACCACAGCGAACATGATTAGCCTGGGCACCTTCACCCTGCTACAGCATC
CGGAGCAGTTGGCGGAGTTACGTTCTGCGCCAGAACTGATGCCGACTGCGGTGGAGGAGTTGCT
GCGCTTTCTCTCCATCGCAGACGGCATGCTGCGCGTTGCCAAAGAAGATATTGATGTTGACGGC
GTGACCATTCGTGCGGATGAGGGTGTTGTGTTTAGCACCTCCGTTATTAACCGTGACACCACCG
CGTACCCGGAGCCGGATACCCTGGATTGGCATCGTAGTGCTCGTCACCACGTGGCGTTCGGCTT
TGGCATTCACCAGTGTCTGGGCCAAAACCTGGCGCGCGCAGAGATGGAAATCGCGCTGCGCACC
CTGTTTGATCGTCTGCCTGGTCTGCGTCTCGCTGCGCCGGCAGAAGAGATCCCGTTTAAACCGG
GTGATACCCTGCAAGGTATGATCGAGCTGCCGGTTACTTGGGAATTC

MGSSHHHHHHSSGLVPRGSHMMTEAVAFPQNRTCPYHPPTAYDPLREARPLSRVTLYDGRSVWV
VTGHAAARQLLSDQRLSTDSTNEAFPIFAERVAAVRSRRPRRTALLGEDDPVHHAQRRMLIPSF
TLKRTADLRPRIQRAVDELIDAMTEQGPPAELVSAFALPVPSIVICDLLGVPYADHDFFEGQSR
RLLRGSTAEEIEDARDQLEGYLGDLVERKRDKPGDGVLDDLIAQQLAHGHLDKQQLVELATILL
VAGHETTANMISLGTFTLLQHPEQLAELRSAPELMPTAVEELLRFLSIADGMLRVAKEDIDVDG
VTIRADEGVVFSTSVINRDTTAYPEPDTLDWHRSARHHVAFGFGIHQCLGQNLARAEMEIALRT
LFDRLPGLRLAAPAEEIPFKPGDTLQGMIELPVTWEF

RhFRed (P450RhF reductase domain, WP_029546987.1) sequences

GTGCTGCACCGCCATCAACCGGTCACCATCGGAGAACCCGCCGCCCGGGCGGTGTCCCGCACCG
TCACCGTCGAGCGCCTGGACCGGATCGCCGACGACGTGCTGCGCCTCGTCCTGCGCGACGCCGG
CGGAAAGACATTGCCCACGTGGACTCCCGGCGCCCATATCGACCTCGACCTCGGCGCGCTGTCG
CGCCAGTACTCCCTGTGCGGCGCGCCCGATGCGCCGAGCTACGAGATTGCCGTGCACCTGGATC
CCGAGAGCCGCGGCGGTTCGCGCTACATCCACGAACAGCTCGAGGTGGGAAGCCCGCTCCGGAT
GCGCGGCCCTCGGAACCATTTCGCGCTCGACCCCGGCGCCGAGCACTACGTGTTCGTCGCCGGC
GGCATCGGCATCACCCCAGTCCTGGCCATGGCCGACCACGCCCGCGCCCGGGGGTGGAGCTACG
AACTGCACTACTGCGGCCGAAACCGTTCCGGCATGGCCTATCTCGAGCGTGTCGCCGGGCACGG
TGACCGGGCCGCCCTGCACGTGTCCGAGGAAGGCACCCGGATCGACCTCGCCGCCCTCCTCGCC
GAGCCCGCCCCCGGCGTCCAGATCTACGCGTGCGGGCCCGGGCGGCTGCTCGCCGGACTCGAGG
ACGCGAGCCGGAACTGGCCCGACGGGGCGCTGCACGTCGAGCACTTCACCTCGTCCCTCGCGGC
GCTCGATCCGGACGTCGAGCACGCCTTCGACCTCGAACTGCGTGACTCGGGGCTGACCGTGCGG
GTCGAACCCACCCAGACCGTCCTCGACGCGTTGCGCGCCAACAACATCGACGTGCCCAGCGACT
GCGAGGAAGGCCTCTGCGGCTCGTGCGAGGTCGCCGTCCTCGACGGCGAGGTCGACCATCGCGA
CACGGTGCTGACCAAGGCCGAGCGGGCGGCGAACCGGCAGATGATGACCTGCTGCTCGCGTGCC
TGTGGCGACCGGCTGGCCCTGCGCCTCTGA

VLHRHQPVTIGEPAARAVSRTVTVERLDRIADDVLRLVLRDAGGKTLPTWTPGAHIDLDLGALS
RQYSLCGAPDAPSYEIAVHLDPESRGGSRYIHEQLEVGSPLRMRGPRNHFALDPGAEHYVFVAG
GIGITPVLAMADHARARGWSYELHYCGRNRSGMAYLERVAGHGDRAALHVSEEGTRIDLAALLA
EPAPGVQIYACGPGRLLAGLEDASRNWPDGALHVEHFTSSLAALDPDVEHAFDLELRDSGLTVR
VEPTQTVLDALRANNIDVPSDCEEGLCGSCEVAVLDGEVDHRDTVLTKAERAANRQMMTCCSRA
CGDRLALRL*

D12-RhFRed (fusion protein) sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGACAGAAGCTGTAGCATTTCCCCAAAATCGCACGTGCCCGTACCATCCGCCAACGGCCTACGA
TCCGCTGCGTGAAGCGAGGCCACTTAGCCGTGTGACCCTGTATGATGGGCGTTCCGTTTGGGTT
GTCACGGGTCACGCAGCTGCACGTCAGTTGCTGTCGGATCAACGTCTGTCGACCGATAGCACCA
ATGAAGCCTTCCCGATTTTCGCGGAGCGCGTGGCTGCGGTTCGTTCTCGTCGTCCGAGACGTAC
GGCATTGCTGGGCGAAGACGACCCGGTGCATCATGCTCAAAGACGGATGCTGATCCCGTCTTTT
ACCTTAAAGCGCACCGCCGACCTTCGTCCGCGGATTCAGCGTGCGGTCGACGAACTCATCGACG
CCATGACCGAACAAGGTCCGCCTGCTGAATTGGTTAGCGCGTTCGCCCTGCCGGTCCCGAGCAT
CGTGATATGCGATCTGTTGGGCGTACCGTACGCGGACCACGATTTCTTCGAGGGCCAAAGCAGA
CGTCTGTTACGCGGTAGCACTGCGGAGGAGATCGAAGACGCACGCGACCAGCTGGAAGGTTATC
TGGGCGACTTGGTCGAGCGCAAACGTGACAAGCCGGGTGATGGTGTGCTGGACGATCTGATCGC
GCAGCAATTGGCGCATGGTCACCTGGACAAGCAGCAGTTGGTGGAGCTCGCTACGATTCTGTTG
GTGGCCGGTCACGAAACCACAGCGAACATGATTAGCCTGGGCACCTTCACCCTGCTACAGCATC
CGGAGCAGTTGGCGGAGTTACGTTCTGCGCCAGAACTGATGCCGACTGCGGTGGAGGAGTTGCT
GCGCTTTCTCTCCATCGCAGACGGCATGCTGCGCGTTGCCAAAGAAGATATTGATGTTGACGGC
GTGACCATTCGTGCGGATGAGGGTGTTGTGTTTAGCACCTCCGTTATTAACCGTGACACCACCG
CGTACCCGGAGCCGGATACCCTGGATTGGCATCGTAGTGCTCGTCACCACGTGGCGTTCGGCTT
TGGCATTCACCAGTGTCTGGGCCAAAACCTGGCGCGCGCAGAGATGGAAATCGCGCTGCGCACC
CTGTTTGATCGTCTGCCTGGTCTGCGTCTCGCTGCGCCGGCAGAAGAGATCCCGTTTAAACCGG
GTGATACCCTGCAAGGTATGATCGAGCTGCCGGTTACTTGGGAATTCGTGCTGCACCGCCATCA
ACCGGTCACCATCGGAGAACCCGCCGCCCGGGCGGTGTCCCGCACCGTCACCGTCGAGCGCCTG
GACCGGATCGCCGACGACGTGCTGCGCCTCGTCCTGCGCGACGCCGGCGGAAAGACATTGCCCA
CGTGGACTCCCGGCGCCCATATCGACCTCGACCTCGGCGCGCTGTCGCGCCAGTACTCCCTGTG
CGGCGCGCCCGATGCGCCGAGCTACGAGATTGCCGTGCACCTGGATCCCGAGAGCCGCGGCGGT
TCGCGCTACATCCACGAACAGCTCGAGGTGGGAAGCCCGCTCCGGATGCGCGGCCCTCGGAACC
ATTTCGCGCTCGACCCCGGCGCCGAGCACTACGTGTTCGTCGCCGGCGGCATCGGCATCACCCC
AGTCCTGGCCATGGCCGACCACGCCCGCGCCCGGGGGTGGAGCTACGAACTGCACTACTGCGGC
CGAAACCGTTCCGGCATGGCCTATCTCGAGCGTGTCGCCGGGCACGGTGACCGGGCCGCCCTGC
ACGTGTCCGAGGAAGGCACCCGGATCGACCTCGCCGCCCTCCTCGCCGAGCCCGCCCCCGGCGT
CCAGATCTACGCGTGCGGGCCCGGGCGGCTGCTCGCCGGACTCGAGGACGCGAGCCGGAACTGG
CCCGACGGGGCGCTGCACGTCGAGCACTTCACCTCGTCCCTCGCGGCGCTCGATCCGGACGTCG
AGCACGCCTTCGACCTCGAACTGCGTGACTCGGGGCTGACCGTGCGGGTCGAACCCACCCAGAC
CGTCCTCGACGCGTTGCGCGCCAACAACATCGACGTGCCCAGCGACTGCGAGGAAGGCCTCTGC
GGCTCGTGCGAGGTCGCCGTCCTCGACGGCGAGGTCGACCATCGCGACACGGTGCTGACCAAGG
CCGAGCGGGCGGCGAACCGGCAGATGATGACCTGCTGCTCGCGTGCCTGTGGCGACCGGCTGGC
CCTGCGCCTCTGA

MGSSHHHHHHSSGLVPRGSHMMTEAVAFPQNRTCPYHPPTAYDPLREARPLSRVTLYDGRSVWV
VTGHAAARQLLSDQRLSTDSTNEAFPIFAERVAAVRSRRPRRTALLGEDDPVHHAQRRMLIPSF
TLKRTADLRPRIQRAVDELIDAMTEQGPPAELVSAFALPVPSIVICDLLGVPYADHDFFEGQSR
RLLRGSTAEEIEDARDQLEGYLGDLVERKRDKPGDGVLDDLIAQQLAHGHLDKQQLVELATILL
VAGHETTANMISLGTFTLLQHPEQLAELRSAPELMPTAVEELLRFLSIADGMLRVAKEDIDVDG
VTIRADEGVVFSTSVINRDTTAYPEPDTLDWHRSARHHVAFGFGIHQCLGQNLARAEMEIALRT
LFDRLPGLRLAAPAEEIPFKPGDTLQGMIELPVTWEFVLHRHQPVTIGEPAARAVSRTVTVERL
DRIADDVLRLVLRDAGGKTLPTWTPGAHIDLDLGALSRQYSLCGAPDAPSYEIAVHLDPESRGG
SRYIHEQLEVGSPLRMRGPRNHFALDPGAEHYVFVAGGIGITPVLAMADHARARGWSYELHYCG
RNRSGMAYLERVAGHGDRAALHVSEEGTRIDLAALLAEPAPGVQIYACGPGRLLAGLEDASRNW

PDGALHVEHFTSSLAALDPDVEHAFDLELRDSGLTVRVEPTQTVLDALRANNIDVPSDCEEGLC
GSCEVAVLDGEVDHRDTVLTKAERAANRQMMTCCSRACGDRLALRL*

P450_D12_T86L sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGACAGAAGCTGTAGCATTTCCCCAAAATCGCACGTGCCCGTACCATCCGCCAACGGCCTACGA
TCCGCTGCGTGAAGCGAGGCCACTTAGCCGTGTGACCCTGTATGATGGGCGTTCCGTTTGGGTT
GTCACGGGTCACGCAGCTGCACGTCAGTTGCTGTCGGATCAACGTCTGTCGACCGATAGCACCA
ATGAAGCCTTCCCGATTTTCGCGGAGCGCGTGGCTGCGGTTCGTTCTCGTCGTCCGAGACGTTT
GGCATTGCTGGGCGAAGACGACCCGGTGCATCATGCTCAAAGACGGATGCTGATCCCGTCTTTT
ACCTTAAAGCGCACCGCCGACCTTCGTCCGCGGATTCAGCGTGCGGTCGACGAACTCATCGACG
CCATGACCGAACAAGGTCCGCCTGCTGAATTGGTTAGCGCGTTCGCCCTGCCGGTCCCGAGCAT
CGTGATATGCGATCTGTTGGGCGTACCGTACGCGGACCACGATTTCTTCGAGGGCCAAAGCAGA
CGTCTGTTACGCGGTAGCACTGCGGAGGAGATCGAAGACGCACGCGACCAGCTGGAAGGTTATC
TGGGCGACTTGGTCGAGCGCAAACGTGACAAGCCGGGTGATGGTGTGCTGGACGATCTGATCGC
GCAGCAATTGGCGCATGGTCACCTGGACAAGCAGCAGTTGGTGGAGCTCGCTACGATTCTGTTG
GTGGCCGGTCACGAAACCACAGCGAACATGATTAGCCTGGGCACCTTCACCCTGCTACAGCATC
CGGAGCAGTTGGCGGAGTTACGTTCTGCGCCAGAACTGATGCCGACTGCGGTGGAGGAGTTGCT
GCGCTTTCTCTCCATCGCAGACGGCATGCTGCGCGTTGCCAAAGAAGATATTGATGTTGACGGC
GTGACCATTCGTGCGGATGAGGGTGTTGTGTTTAGCACCTCCGTTATTAACCGTGACACCACCG
CGTACCCGGAGCCGGATACCCTGGATTGGCATCGTAGTGCTCGTCACCACGTGGCGTTCGGCTT
TGGCATTCACCAGTGTCTGGGCCAAAACCTGGCGCGCGCAGAGATGGAAATCGCGCTGCGCACC
CTGTTTGATCGTCTGCCTGGTCTGCGTCTCGCTGCGCCGGCAGAAGAGATCCCGTTTAAACCGG
GTGATACCCTGCAAGGTATGATCGAGCTGCCGGTTACTTGGGAATTCGAATTC

MGSSHHHHHHSSGLVPRGSHMMTEAVAFPQNRTCPYHPPTAYDPLREARPLSRVTLYDGRSVWV
VTGHAAARQLLSDQRLSTDSTNEAFPIFAERVAAVRSRRPRRLALLGEDDPVHHAQRRMLIPSF
TLKRTADLRPRIQRAVDELIDAMTEQGPPAELVSAFALPVPSIVICDLLGVPYADHDFFEGQSR
RLLRGSTAEEIEDARDQLEGYLGDLVERKRDKPGDGVLDDLIAQQLAHGHLDKQQLVELATILL
VAGHETTANMISLGTFTLLQHPEQLAELRSAPELMPTAVEELLRFLSIADGMLRVAKEDIDVDG
VTIRADEGVVFSTSVINRDTTAYPEPDTLDWHRSARHHVAFGFGIHQCLGQNLARAEMEIALRT
LFDRLPGLRLAAPAEEIPFKPGDTLQGMIELPVTWEF

P450_D12_S55R/T86L/I104M sequences

ATGGGCAGCAGCCATCATCATCATCATCACAGCAGCGGCCTGGTGCCGCGCGGCAGCCATATGA
TGACAGAAGCTGTAGCATTTCCCCAAAATCGCACGTGCCCGTACCATCCGCCAACGGCCTACGA
TCCGCTGCGTGAAGCGAGGCCACTTAGCCGTGTGACCCTGTATGATGGGCGTTCCGTTTGGGTT
GTCACGGGTCACGCAGCTGCACGTCAGTTGCTGCGTGATCAACGTCTGTCGACCGATAGCACCA
ATGAAGCCTTCCCGATTTTCGCGGAGCGCGTGGCTGCGGTTCGTTCTCGTCGTCCGAGACGTCT
GGCATTGCTGGGCGAAGACGACCCGGTGCATCATGCTCAAAGACGGATGCTGATGCCGTCTTTT
ACCTTAAAGCGCACCGCCGACCTTCGTCCGCGGATTCAGCGTGCGGTCGACGAACTCATCGACG
CCATGACCGAACAAGGTCCGCCTGCTGAATTGGTTAGCGCGTTCGCCCTGCCGGTCCCGAGCAT
CGTGATATGCGATCTGTTGGGCGTACCGTACGCGGACCACGATTTCTTCGAGGGCCAAAGCAGA
CGTCTGTTACGCGGTAGCACTGCGGAGGAGATCGAAGACGCACGCGACCAGCTGGAAGGTTATC
TGGGCGACTTGGTCGAGCGCAAACGTGACAAGCCGGGTGATGGTGTGCTGGACGATCTGATCGC

GCAGCAATTGGCGCATGGTCACCTGGACAAGCAGCAGTTGGTGGAGCTCGCTACGATTCTGTTG
GTGGCCGGTCACGAAACCACAGCGAACATGATTAGCCTGGGCACCTTCACCCTGCTACAGCATC
CGGAGCAGTTGGCGGAGTTACGTTCTGCGCCAGAACTGATGCCGACTGCGGTGGAGGAGTTGCT
GCGCTTTCTCTCCATCGCAGACGGCATGCTGCGCGTTGCCAAAGAAGATATTGATGTTGACGGC
GTGACCATTCGTGCGGATGAGGGTGTTGTGTTTAGCACCTCCGTTATTAACCGTGACACCACCG
CGTACCCGGAGCCGGATACCCTGGATTGGCATCGTAGTGCTCGTCACCACGTGGCGTTCGGCTT
TGGCATTCACCAGTGTCTGGGCCAAAACCTGGCGCGCGCAGAGATGGAAATCGCGCTGCGCACC
CTGTTTGATCGTCTGCCTGGTCTGCGTCTCGCTGCGCCGGCAGAAGAGATCCCGTTTAAACCGG
GTGATACCCTGCAAGGTATGATCGAGCTGCCGGTTACTTGGGAATTCGAATTC

MGSSHHHHHHSSGLVPRGSHMMTEAVAFPQNRTCPYHPPTAYDPLREARPLSRVTLYDGRSVWV
VTGHAAARQLLRDQRLSTDSTNEAFPIFAERVAAVRSRRPRRLALLGEDDPVHHAQRRMLMPSF
TLKRTADLRPRIQRAVDELIDAMTEQGPPAELVSAFALPVPSIVICDLLGVPYADHDFFEGQSR
RLLRGSTAEEIEDARDQLEGYLGDLVERKRDKPGDGVLDDLIAQQLAHGHLDKQQLVELATILL
VAGHETTANMISLGTFTLLQHPEQLAELRSAPELMPTAVEELLRFLSIADGMLRVAKEDIDVDG
VTIRADEGVVFSTSVINRDTTAYPEPDTLDWHRSARHHVAFGFGIHQCLGQNLARAEMEIALRT
LFDRLPGLRLAAPAEEIPFKPGDTLQGMIELPVTWEF

5.5.2 Generation of libraries for site-saturation mutagenesis

Degenerate forward and reverse primers were output by the automated primer generation pipeline for each of the 96 active site residues being targeted and purchased from IDT (Table 5.5). Blunt-end whole-plasmid PCR reactions were carried out using a 96-well PCR plate. Each 25 µL reaction mixture included 1× HF Phusion buffer, 4% dimethyl sulfoxide (DMSO), 400 µM deoxynucleotide triphosphates (dNTPs), 0.1 ng/µL of template plasmid DNA, 0.4 units of Phusion DNA polymerase, and 0.8 µM each of a degenerate forward primer and a reverse primer. The PCR program was set to an initial denaturation at 98 °C for 2 minutes, followed by 25 cycles of 98 °C for 20 seconds, 58 °C for 20 seconds, and 72 °C for 4.5 minutes, with a final extension at 72 °C for 10 minutes. Following amplification, 5 µL from each PCR product was combined into a single tube. The pooled PCR products underwent gel purification using a 0.8% agarose gel and Qiagen spin columns. To eliminate any remaining template DNA, the mixture was treated overnight at 37 °C with 5 µL of DpnI (NEB) in 1× rCutsmart buffer, then purified again using Qiagen PCR clean-up spin columns. The linear DNA fragments were phosphorylated by incubating them at 37 °C for 30 minutes in 1× Quick Ligase buffer (NEB) with 2 µL of polynucleotide kinase (PNK) enzyme (NEB). Immediately after phosphorylation, 1 µL of Quick Ligase enzyme (NEB) was added directly to the reaction, and the mixture was

incubated at room temperature for 15 minutes to facilitate ligation. Salts were removed from the ligation reaction using Qiagen PCR clean-up spin columns. For transformation, 2 µL of the purified DNA was introduced into high-efficiency 10-beta electrocompetent *E. coli* cells (NEB) using standard electroporation protocols. The resulting bacterial colonies were harvested, and plasmid DNA was extracted via miniprep to create a plasmid DNA library. This library was then used to transform chemically competent Lemo21(DE3) cells, and the colonies obtained were directly utilized for library screening.

5.5.3 Generation of combinatorial libraries using multiple site-directed mutagenesis

Forward primers flanking the desired mutation site—with more than 15 complementary nucleotides on each side—were designed and procured from IDT (refer to Table 5.6). Equal molar amounts of each forward primer were mixed together and diluted to a final concentration of 10 µM. Polymerase chain reactions (PCRs) were conducted using the QuikChange Lightning Multi Site-Directed Mutagenesis Kit from Agilent, following the manufacturer's recommended protocols. After amplification, 2 µL of the supplied DpnI enzyme was added to each reaction, and the mixtures were incubated at 37 °C for 1 hour to digest the template DNA. High-efficiency 10-beta electrocompetent *E. coli* cells (NEB) were then transformed with 2 µL of the purified DNA using standard electroporation methods. The resulting colonies were harvested, and plasmid DNA was extracted via miniprep to generate a plasmid DNA library.This library was used to transform chemically competent Lemo21(DE3) cells. The colonies obtained from this transformation were directly employed in the library screening process.

5.5.4 Sample preparation for next-generation sequencing

For the preparation of next-generation sequencing (NGS) samples, the protocol reported by Wittmann et al. was used. This workflow has been optimized for compatibility with Laragen’s pipeline, which allows pooling of library samples from multiple customers, thereby significantly reducing per-sample sequencing costs. Initial steps involve designing inner primers with a melting temperature of 58°C, ending in either guanine (G) or cytosine (C), and verifying the absence of secondary structures using the IDT tool.

Amplification fragments should be limited to under 600 base pairs (bp) to ensure robust results.

The first PCR step requires a mastermix consisting of ThermoPol buffer, 10 mM dNTPs, Taq polymerase, DMSO, and an inner primer mix, into which 1 µL of overnight LB culture (stored at -80°C) is added to each PCR well. To increase the sequencing efficiency, nine overnight culture plates were merged into eight sequencing plates to avoid redundant sequencing of positive and negative controls. The thermocycling protocol for the first PCR includes an initial 95°C denaturation for five minutes, followed by 20-second denaturation at 95°C, a touchdown phase from 63°C to 54°C (decreasing by 1°C per cycle) for annealing, and a 30-second extension at 68°C.

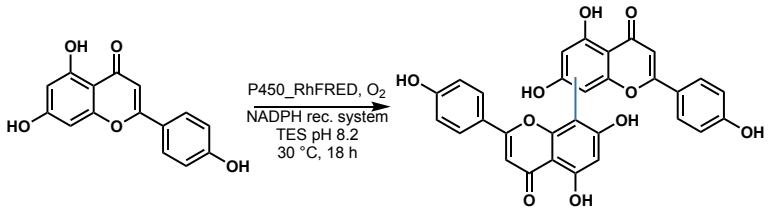
During the first PCR, eight 1 µM dual-indexed barcode plates were prepared from the 100 µM forward and reverse primer stocks in plates by serial dilution. Each well in eight plates was uniquely barcoded according to the different combinations of forward and reverse primers. For the second PCR, 2 µL of the barcode solution was added to each well, followed by 24 amplification cycles with a denaturation step at 95°C and annealing/extension at 68°C. Once completion, 4 parts of the reaction mixtures were quenced using 1 part of 100 mM EDTA to stabilize the samples. Samples from the same plate were then pooled, supplemented with loading dye (without SDS), and analyzed on a 2% TAE agarose gel with SYBR gold. After the gel has sufficiently separated, the DNA bands were extracted and eluted in 10 µL ddH2O. For sequencing, equimolar concentrations of each gel-extracted pool were combined, creating a final prepared library ready for submission.

5.5.5 Expression of P450-RhFRed libraries in 96-well plates.

Chemically competent Lemo21(DE3) cells were transformed with a pET28b plasmid encoding P450-RhFRed using standard heat shock protocols. The transformed cells were directly inoculated into 400 µL of LB medium containing 50 µg/mL kanamycin in a 96-well plate and incubated overnight at 37 °C with shaking at 350 rpm. From each overnight culture, 10 µL was used to inoculate 400 µL of TB supplemented with 4% (v/v) glycerol and 50 µg/mL kanamycin. These cultures were grown at 37 °C and 350 rpm until they reached an optical density at 600 nm (OD600) of approximately 0.8 to 1.0. To induce protein expression, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final

concentration of 0.5 mM, and 5-aminolevulinic acid (5-ALA) was supplemented at 1 mM to facilitate heme synthesis. The cultures were then incubated at 20 °C with shaking at 350 rpm for an additional 20 hours. Cells were harvested by centrifugation at 1,000 × g for 15 minutes at 4 °C. The harvested cell pellets were stored at –80 °C in their plates until lysate reactions were performed.

5.5.6 High-throughput reactions with P450-RhFRed lysate



A AA

Harvested cells in 96-well plates underwent five freeze-thaw cycles and were resuspended in 250 µL of a lysis buffer containing 50 mM TES at pH 8.2–8.3, 150 mM NaCl, 10% glycerol, 2.5 U/mL DNase I (Thermo Scientific #90083), 2 mg/mL lysozyme, and 100 µM PMSF. The cells were enzymatically lysed by incubating them at 25 °C with shaking at 350 rpm for 2 hours. Following lysis, the plates were centrifuged at 1,000 × g for 20 minutes at 4 °C to clarify the lysate. A 50 µL portion of the clarified lysate was transferred to a new plate, where each well contained 50 µL of a reaction mixture composed of 100 mM Na_2HPO_3, 2 mM $NADP^+$, 20 µM phosphite dehydrogenase, 2 mM apigenin in 50 mM TES buffer with 150 mM NaCl and 5% glycerol at pH 8.2–8.3. The reactions were incubated at 30 °C for approximately 18 hours and then quenched by adding 300 µL of methanol containing 1 µM 2,6-dihydroxyacetophenone as an internal standard. The quenced samples were then centrifuged for the screening on RapidFire-MS.

5.5.7 High-throughput RapidFire-MS screen for activity

To evaluate the total dimerization activity in the library reactions, a high-throughput screening was conducted using an Agilent RapidFire 365 Mass Spectrometry System connected to an Agilent 6545 LC–Q-TOF. The library samples were prepared by transfering 10 µL of the quenched samples to 250 µL of the RapidFire solution (2:1 ratio of the quenching solution to water containing 666 nM 2,6'-dihydroxyacetophenone as an

internal standard). These samples were placed into shallow 96-well plates from Axygen and sealed with pierceable aluminum covers supplied by Agilent.

Each injection on the RapidFire system was programmed with specific parameters: a 0.6-second aspiration of the sample, followed by a 3-second loading onto a C4 cartridge using water at a flow rate of 0.5 mL/min. The sample was then eluted from the cartridge over 8 seconds using a solvent mixture of 90% acetonitrile and 10% water (v/v), buffered with 10 mM ammonium formate at pH 8, delivered at a flow rate of 0.8 mL/min. A 0.5-second re-equilibration step was included at the end of each cycle. To minimize carry-over between samples, a blank injection was performed after each sample injection.

For each sample, peak areas corresponding to the internal standard, substrate, and dimerization product were extracted from the ion chromatograms using MassHunter Qualitative Analysis 10.0 software. The relative percentage conversion to the dimerization product was then calculated as [2 × **AA**/(2 × **AA** + 1 × **A**)] × 100%. To normalize the data, the average percent conversion from the template reactions on each plate (with $n = 9$) was set as a reference value of 1.0. The conversion rates of the variant reactions were normalized against this template average, resulting in fold improvement scores for each variant.

5.5.8 LC-MS analysis of biocatalytic reactions for validation

The RapidFire samples for validation were then analyzed using ultra-performance liquid chromatography (UPLC) coupled with photodiode array (PDA) spectrometry and mass spectrometry. This was performed on an Agilent 6230 time-of-flight mass spectrometer equipped with a Dual AJS ESI source, along with an Agilent 1290 Infinity Series II diode array detector, autosampler, and binary pump.

MassHunter software was used to extract ion chromatograms at the target masses, allowing for the characterization of percent yields and reaction selectivity. To determine site-selectivity, samples were analyzed in negative ion mode using a Waters Acquity Premier HSS T3 C18 column (1.8 µm, 2.1 × 50 mm). The mobile phases consisted of Phase A: 95% water and 5% acetonitrile (v/v), and Phase B: 95% acetonitrile and 5% water (v/v). The gradient elution program was as follows: maintain 30% Phase B for 0.5 minutes; ramp to 80% B over 1.5 minutes; increase to 100% B over the next 0.5 minutes; holding at 100% B for 1.0 minute. The flow rate was set at 0.7 mL/min, and the column

temperature was maintained at 45 °C. After each injection, the system was equilibrated at 30% Phase B for 0.5 minutes.

5.5.9 Data availability

Python scripts, Jupyter notebooks, and computational results to reproduce the analysis: https://github.com/Chang-Hwa/SSM_workflow

5.5.10 Computational predictions of mutational effects

Mutational effects on the template, P450_D12 (T86L), were predicted using two protein fitness models, EVcouplings and ESM-1v. To build an evolutionary coupling model, the template protein sequence was submitted to the EVcouplings server (https://v2.evcouplings.org/) for homology search against UniRef90 using a bitscore of 0.7, followed by a sequence alignment. The computational predictions of mutational effects, including both epistatic and independent modes, were output to a csv file named as "P450_D12_T86L_b0.7_single_mutant_matrix.csv".

The computational predictions from ESM-1v were acquired by running the command below with ESM installed. The input csv file for ESM-1v was prepared using the Jupyter Notebook script *P450D12_evolution_NGS.ipynb*.

```
# Predict variant effects by ESM-1v
python predict.py \
    --model-location esm1v_t33_650M_UR90S_1 esm1v_t33_650M_UR90S_2
esm1v_t33_650M_UR90S_3 esm1v_t33_650M_UR90S_4 esm1v_t33_650M_UR90S_5 \
    --sequence MTEAV… \
    --dms-input ./data/P450/P450_D12_T86L_wholelength.csv \
    --mutation-col mutant \
    --dms-output ./data/P450/P450_D12_T86L_wholelength_labeled.csv \
    --offset-idx 0 \
    --scoring-strategy wt-marginals
```

5.5.11 Annotation of secondary structure elements of P450_D12

The annotation of secondary structure elements of P450_D12 was done by submitting its AlphaFold2 model to the SecStrAnnotator[28] server (https://sestra.ncbr.muni.cz/submit) using the template from Cytochromes P450 2nnjA.

5.5.12 Mean computational prediction scores of positions from 15–397 of P450_D12_T86L

The task was done by using the Jupyter Notebook script *P450D12_evolution_NGS.ipynb*.

5.5.13 Color coding P450_D12 (T86L) by the mean computational prediction scores

The structural model was opened in PyMOL. In PyMOL command line, run the Python script *EV_pymol_coloring.py*.

```
# Load pdb structure
cmd.load(model.pdb)

# Run the Python script
run EV_pymol_coloring.py
```

Table 5.5. Primers used for site-saturation mutagenesis of P450_D12 active site residues.

Mutation	Forward primer (5' to 3')		Reverse primer (5' to 3')	
F7X	Rd4-for	NDT/VHG/TGG-CCCCAAAATCGCACGTGC	Rd4-rev	TGCTACAGCTTCTGTCATCATATGG
P8X	Rd4-for	NDT/VHG/TGG-CACGTGCCCG	Rd4-rev	AAATGCTACAGCTTCTGTCATCATATGG
Q9X	Rd4-for	NDT/VHG/TGG-AATCGCACGTGCCCG	Rd4-rev	GGGAAATGCTACAGCTTCTGTC
N10X	Rd4-for	NDT/VHG/TGG-CGCACGTGCCCGTAC	Rd4-rev	TTGGGGAAATGCTACAGCTTC
Y15X	Rd4-for	NDT/VHG/TGG-CATCCGCCAACGGCC	Rd4-rev	CGGGCACGTGCGATTTTG
H16X	Rd4-for	NDT/VHG/TGG-CCGCCAACGGCCTAC	Rd4-rev	GTACGGGCACGTGCG
P17X	Rd4-for	NDT/VHG/TGG-CCAACGGCCTACGATCCG	Rd4-rev	ATGGTACGGGCACGTGC
P18X	Rd4-for	NDT/VHG/TGG-ACGGCCTACGATCCGC	Rd4-rev	CGGATGGTACGGGCACG
Y21X	Rd4-for	NDT/VHG/TGG-GATCCGCTGCGTGAAGC	Rd4-rev	GGCCGTTGGCGGATG
L24X	Rd4-for	NDT/VHG/TGG-CGTGAAGCGAGGCCAC	Rd4-rev	CGGATCGTAGGCCGTTGG
V33X	Rd4-for	NDT/VHG/TGG-ACCCTGTATGATGGGCGTTC	Rd4-rev	ACGGCTAAGTGGCCTCG
L35X	Rd4-for	NDT/VHG/TGG-TATGATGGGCGTTCCGTTTG	Rd4-rev	GGTCACACGGCTAAGTGG
Y36X	Rd4-for	NDT/VHG/TGG-GATGGGCGTTCCGTTTGG	Rd4-rev	CAGGGTCACACGGCTAAGTG
D37X	Rd4-for	NDT/VHG/TGG-GGGCGTTCCGTTTGGG	Rd4-rev	ATACAGGGTCACACGGCTAAG
V41X	Rd4-for	NDT/VHG/TGG-TGGGTTGTCACGGGTCAC	Rd4-rev	GGAACGCCCATCATACAGG
V43X	Rd4-for	NDT/VHG/TGG-GTCACGGGTCACGCAG	Rd4-rev	CCAAACGGAACGCCCATC
V44X	Rd4-for	NDT/VHG/TGG-ACGGGTCACGCAGCTG	Rd4-rev	AACCCAAACGGAACGCC
T45X	Rd4-for	NDT/VHG/TGG-GGTCACGCAGCTGCAC	Rd4-rev	GACAACCCAAACGGAACGC
G46X	Rd4-for	NDT/VHG/TGG-CACGCAGCTGCACGTC	Rd4-rev	CGTGACAACCCAAACGGAAC
H47X	Rd4-for	NDT/VHG/TGG-GCAGCTGCACGTCAGTTG	Rd4-rev	ACCCGTGACAACCCAAAC
A49X	Rd4-for	NDT/VHG/TGG-GCACGTCAGTTGCTGCG	Rd4-rev	TGCGTGACCCGTGACAAC
A50X	Rd1-for Rd3-for	NNKCGTCAGTTGCTGTCGGATC NDT/VHG/TGG-CGTCAGTTGCTGCGTGATC	Rd1-rev Rd3-rev	AGCTGCGTGACCCGTG AGCTGCGTGACCCGTG
R51X	Rd1-for Rd3-for	NNKCAGTTGCTGTCGGATCAACG NDT/VHG/TGG-CAGTTGCTGCGTGATCAACG	Rd1-rev Rd3-rev	TGCAGCTGCGTGACC TGCAGCTGCGTGACCC
L53X	Rd4-for	NDT/VHG/TGG-CTGCGTGATCAACGTCTGTC	Rd4-rev	CTGACGTGCAGCTGCG
L54X	Rd1-for Rd3-for	NNKTCGGATCAACGTCTGTCGAC NDT/VHG/TGG-CGTGATCAACGTCTGTCGAC	Rd1-rev Rd3-rev	CAACTGACGTGCAGCTGC CAACTGACGTGCAGCTGC
S55X R55X	Rd1-for Rd3-for	NNKGATCAACGTCTGTCGACCG NDT/VHG/TGG-GATCAACGTCTGTCGACCG	Rd1-rev Rd3-rev	CAGCAACTGACGTGCAGC CAGCAACTGACGTGCAGC
L59X	Rd1-for Rd3-for	NNKTCGACCGATAGCACCAATGAAG NDT/VHG/TGG-TCGACCGATAGCACCAATGAAG	Rd1-rev Rd3-rev	ACGTTGATCCGACAGCAAC ACGTTGATCACGCAGCAAC
T61X	Rd1-for Rd3-for	NNKGATAGCACCAATGAAGCCTTCC NDT/VHG/TGG-GATAGCACCAATGAAGCCTTCC	Rd1-rev Rd3-rev	CGACAGACGTTGATCCGAC CGACAGACGTTGATCACGC
D62X	Rd1-for Rd3-for	NNKAGCACCAATGAAGCCTTCC NDT/VHG/TGG-AGCACCAATGAAGCCTTCC	Rd1-rev Rd3-rev	GGTCGACAGACGTTGATCC GGTCGACAGACGTTGATCAC
S63X	Rd1-for	NNKACCAATGAAGCCTTCCCG	Rd1-rev	ATCGGTCGACAGACGTTGATC
T64X	Rd4-for	TGGAATGAAGCCTTCCCGATTTTCG	Rd4-rev	GCTATCGGTCGACAGACGTTG
N65X	Rd4-for	TGGGAAGCCTTCCCGATTTTCGC	Rd4-rev	GGTGCTATCGGTCGACAGAC
F68X	Rd4-for	TGGCCGATTTTCGCGGAGCG	Rd4-rev	GGCTTCATTGGTGCTATCGG
I70X	Rd4-for	TGGTTCGCGGAGCGCGTG	Rd4-rev	CGGGAAGGCTTCATTGGTG
F71X	Rd1-for	NNKGCGGAGCGCGTGGCTG	Rd1-rev	AATCGGGAAGGCTTCATTGG
A72X	Rd4-for	NDT/VHG/TGG-GAGCGCGTGGCTGCG	Rd4-rev	GAAAATCGGGAAGGCTTCATTGG
E73X	Rd4-for	NDT/VHG/TGG-CGCGTGGCTGCGGTTC	Rd4-rev	CGCGAAAATCGGGAAGGC
R74X	Rd4-for	NDT/VHG/TGG-GTGGCTGCGGTTCGTTC	Rd4-rev	CTCCGCGAAAATCGGGAAG
V75X	Rd4-for	NDT/VHG/TGG-GCTGCGGTTCGTTCTCG	Rd4-rev	GCGCTCCGCGAAAATCG
A76X	Rd4-for	NDT/VHG/TGG-GCGGTTCGTTCTCGTCG	Rd4-rev	CACGCGCTCCGCGAAAATC
A77X	Rd4-for	NDT/VHG/TGG-GTTCGTTCTCGTCGTCCGAG	Rd4-rev	AGCCACGCGCTCCGC
V78X	Rd4-for	NDT/VHG/TGG-CGTTCTCGTCGTCCGAGAC	Rd4-rev	CGCAGCCACGCGCTC
R79X	Rd4-for	NDT/VHG/TGG-TCTCGTCGTCCGAGACG	Rd4-rev	AACCGCAGCCACGCG
R85X	Rd1-for Rd2-for	NNKACGGCATTGCTGGGC NNKTTGGCATTGCTGGGCG	Rd1-rev Rd2-rev	TCTCGGACGACGAGAACG TCTCGGACGACGAGAACG
T86X	Rd1-for	NNKGCATTGCTGGGCGAAGAC	Rd1-rev	ACGTCTCGGACGACGAG
A87X	Rd1-for Rd2-for	NNKTTGCTGGGCGAAGACG NNKTTGCTGGGCGAAGACGAC	Rd1-rev Rd2-rev	CGTACGTCTCGGACGACG CAAACGTCTCGGACGACG
L88X	Rd1-for	NNKCTGGGCGAAGACGACCC	Rd1-rev	TGCCGTACGTCTCGGAC

Mutation	Forward primer (5' to 3')		Reverse primer (5' to 3')	
	Rd2-for	NNKCTGGGCGAAGACGACCC	Rd2-rev	TGCCAAACGTCTCGGACG
L89X	Rd1-for	NNKGGCGAAGACGACCCGG	Rd1-rev	CAATGCCGTACGTCTCGG
	Rd2-for	NNKGGCGAAGACGACCCGG	Rd2-rev	CAATGCCAAACGTCTCGGAC
G90X	Rd1-for	NNKGAAGACGACCCGGTGC	Rd1-rev	CAGCAATGCCGTACGTCTC
	Rd2-for	NNKGAAGACGACCCGGTGCATC	Rd2-rev	CAGCAATGCCAAACGTCTCG
E91X	Rd1-for	NNKGACGACCCGGTGCATC	Rd1-rev	GCCCAGCAATGCCGTAC
	Rd2-for	NNKGACGACCCGGTGCATCATG	Rd2-rev	GCCCAGCAATGCCAAACG
D93X	Rd1-for	NNKCCGGTGCATCATGCTCAAAG	Rd1-rev	GTCTTCGCCCAGCAATGC
V95X	Rd1-for	NNKCATCATGCTCAAAGACGGATGC	Rd1-rev	CGGGTCGTCTTCGCCC
H97X	Rd1-for	NNKGCTCAAAGACGGATGCTGATC	Rd1-rev	ATGCACCGGGTCGTCTTC
	Rd3-for	NDT/VHG/TGG-GCTCAAAGACGGATGCTGATG	Rd3-rev	ATGCACCGGGTCGTCTTC
Q99X	Rd1-for	NNKAGACGGATGCTGATCCCG	Rd1-rev	AGCATGATGCACCGGG
	Rd3-for	NDT/VHG/TGG-AGACGGATGCTGATGCCG	Rd3-rev	AGCATGATGCACCGGGTC
R101X	Rd1-for	NNKATGCTGATCCCGTCTTTTACC	Rd1-rev	TCTTTGAGCATGATGCACCG
	Rd3-for	NDT/VHG/TGG-ATGCTGATGCCGTCTTTTACC	Rd3-rev	TCTTTGAGCATGATGCACCG
M102X	Rd1-for	NNKCTGATCCCGTCTTTTACCTTAAAGC	Rd1-rev	CCGTCTTTGAGCATGATGCAC
	Rd3-for	NDT/VHG/TGG-CTGATGCCGTCTTTTACCTTAAAGC	Rd3-rev	CCGTCTTTGAGCATGATGCAC
L103X	Rd1-for	NNKATCCCGTCTTTTACCTTAAAGCG	Rd1-rev	CATCCGTCTTTGAGCATGATGC
	Rd3-for	NDT/VHG/TGG-ATGCCGTCTTTTACCTTAAAGCG	Rd3-rev	CATCCGTCTTTGAGCATGATGC
I104X	Rd1-for	NNKCCGTCTTTTACCTTAAAGCGCAC	Rd1-rev	CAGCATCCGTCTTTGAGCATG
M104X	Rd3-for	NDT/VHG/TGG-CCGTCTTTTACCTTAAAGCGCAC	Rd3-rev	CAGCATCCGTCTTTGAGCATG
P105X	Rd1-for	NNKTCTTTTACCTTAAAGCGCACCG	Rd1-rev	GATCAGCATCCGTCTTTGAGC
	Rd3-for	NDT/VHG/TGG-TCTTTTACCTTAAAGCGCACCG	Rd3-rev	CATCAGCATCCGTCTTTGAGC
S106X	Rd1-for	NNKTTTACCTTAAAGCGCACCGC	Rd1-rev	CGGGATCAGCATCCGTC
	Rd3-for	NDT/VHG/TGG-TTTACCTTAAAGCGCACCGC	Rd3-rev	CGGCATCAGCATCCGTC
T108X	Rd1-for	NNKTTAAAGCGCACCGCCG	Rd1-rev	AAAAGACGGGATCAGCATCC
	Rd3-for	NDT/VHG/TGG-TTAAAGCGCACCGCCG	Rd3-rev	AAAAGACGGCATCAGCATCC
L109X	Rd1-for	NNKAAGCGCACCGCCGAC	Rd1-rev	GGTAAAAGACGGGATCAGCATC
	Rd3-for	NDT/VHG/TGG-AAGCGCACCGCCGAC	Rd3-rev	GGTAAAAGACGGCATCAGCATC
R111X	Rd1-for	NNKACCGCCGACCTTCGTC	Rd1-rev	CTTTAAGGTAAAAGACGGGATCAGC
	Rd3-for	NDT/VHG/TGG-ACCGCCGACCTTCGTC	Rd3-rev	CTTTAAGGTAAAAGACGGCATCAGC
T112X	Rd1-for	NNKGCCGACCTTCGTCCGC	Rd1-rev	GCGCTTTAAGGTAAAAGACGGG
	Rd3-for	NDT/VHG/TGG-GCCGACCTTCGTCCGC	Rd3-rev	GCGCTTTAAGGTAAAAGACGGC
L115X	Rd1-for	NNKCGTCCGCGGATTCAGC	Rd1-rev	GTCGGCGGTGCGCTTTAAG
I119X	Rd1-for	NNKCAGCGTGCGGTCGAC	Rd1-rev	CCGCGGACGAAGGTCG
A144X	Rd1-for	NNKCTGCCGGTCCCGAGC	Rd1-rev	GAACGCGCTAACCAATTCAGC
L145X	Rd1-for	NNKCCGGTCCCGAGCATCG	Rd1-rev	GGCGAACGCGCTAACC
V147X	Rd1-for	NNKCCGAGCATCGTGATATGCG	Rd1-rev	CGGCAGGGCGAACGC
S149X	Rd1-for	NNKATCGTGATATGCGATCTGTTGG	Rd1-rev	CGGGACCGGCAGGGC
V151X	Rd1-for	NNKATATGCGATCTGTTGGGCG	Rd1-rev	GATGCTCGGGACCGGC
C153X	Rd1-for	NNKGATCTGTTGGGCGTACCG	Rd1-rev	TATCACGATGCTCGGGACC
L155X	Rd1-for	NNKTTGGGCGTACCGTACGC	Rd1-rev	ATCGCATATCACGATGCTCG
Q169X	Rd4-for	NDT/VHG/TGG-AGCAGACGTCTGTTACGCG	Rd4-rev	GCCCTCGAAGAAATCGTGG
S170X	Rd1-for	NNKAGACGTCTGTTACGCGGTAG	Rd1-rev	TTGGCCCTCGAAGAAATCG
R171X	Rd4-for	NDT/VHG/TGG-CGTCTGTTACGCGGTAGC	Rd4-rev	GCTTTGGCCCTCGAAGAAATC
R172X	Rd4-for	NDT/VHG/TGG-CTGTTACGCGGTAGCACTG	Rd4-rev	TCTGCTTTGGCCCTCGAAG
L173X	Rd1-for	NNKTTACGCGGTAGCACTGC	Rd1-rev	ACGTCTGCTTTGGCCC
L174X	Rd1-for	NNKCGCGGTAGCACTGCG	Rd1-rev	CAGACGTCTGCTTTGGCC
R175X	Rd1-for	NNKGGTAGCACTGCGGAGGAG	Rd1-rev	TAACAGACGTCTGCTTTGGC
G176X	Rd4-for	NDT/VHG/TGG-AGCACTGCGGAGGAGATC	Rd4-rev	GCGTAACAGACGTCTGCTTTG
S177X	Rd3-for	NDT/VHG/TGG-ACTGCGGAGGAGATCGAAG	Rd3-rev	ACCGCGTAACAGACGTCTG
T178X	Rd4-for	NDT/VHG/TGG-GCGGAGGAGATCGAAGACG	Rd4-rev	GCTACCGCGTAACAGACG
R199X	Rd3-for	NDT/VHG/TGG-AAACGTGACAAGCCGGG	Rd3-rev	CTCGACCAAGTCGCCCAG
D202X	Rd3-for	NDT/VHG/TGG-AAGCCGGGTGATGGTGTG	Rd3-rev	ACGTTTGCGCTCGACC
V208X	Rd1-for	NNKCTGGACGATCTGATCGCG	Rd1-rev	ACCATCACCCGGCTTGTC
V228X	Rd1-for	NNKGAGCTCGCTACGATTCTGTTG	Rd1-rev	CAACTGCTGCTTGTCCAGG
E229X	Rd1-for	NNKCTCGCTACGATTCTGTTGGTG	Rd1-rev	CACCAACTGCTGCTTGTCC
L230X	Rd1-for	NNKGCTACGATTCTGTTGGTGGC	Rd1-rev	CTCCACCAACTGCTGCTTG
A231X	Rd1-for	NNKACGATTCTGTTGGTGGCC	Rd1-rev	GAGCTCCACCAACTGCTG
T232X	Rd1-for	NNKATTCTGTTGGTGGCCGG	Rd1-rev	AGCGAGCTCCACCAACTG
I233X	Rd1-for	NNKCTGTTGGTGGCCGGTC	Rd1-rev	CGTAGCGAGCTCCACCAAC
L234X	Rd1-for	NNKTTGGTGGCCGGTCACG	Rd1-rev	AATCGTAGCGAGCTCCACC
V236X	Rd1-for	NNKGCCGGTCACGAAACCAC	Rd1-rev	CAACAGAATCGTAGCGAGCTC
A237X	Rd1-for	NNKGGTCACGAAACCACAGCG	Rd1-rev	CACCAACAGAATCGTAGCGAG
E240X	Rd1-for	NNKACCACAGCGAACATGATTAGC	Rd1-rev	GTGACCGGCCACCAAC
T242X	Rd1-for	NNKGCGAACATGATTAGCCTGGG	Rd1-rev	GGTTTCGTGACCGGCC
A243X	Rd1-for	NNKAACATGATTAGCCTGGGCAC	Rd1-rev	TGTGGTTTCGTGACCGG
N244X	Rd1-for	NNKATGATTAGCCTGGGCACCTTC	Rd1-rev	CGCTGTGGTTTCGTGACC
M245X	Rd1-for	NNKATTAGCCTGGGCACCTTCAC	Rd1-rev	GTTCGCTGTGGTTTCGTGAC
I246X	Rd1-for	NNKAGCCTGGGCACCTTCAC	Rd1-rev	CATGTTCGCTGTGGTTTCGTG
S247X	Rd4-for	NDT/VHG/TGG-CTGGGCACCTTCACCCTG	Rd4-rev	AATCATGTTCGCTGTGGTTTCG
V274X	Rd1-for	NNKGAGGAGTTGCTGCGCTTTC	Rd1-rev	CGCAGTCGGCATCAGTTC
E275X	Rd1-for	NNKGAGTTGCTGCGCTTTCTCTC	Rd1-rev	CACCGCAGTCGGCATC
L277X	Rd1-for	NNKCTGCGCTTTCTCTCCATCG	Rd1-rev	CTCCTCCACCGCAGTCG
L278X	Rd1-for	NNKCGCTTTCTCTCCATCGCAG	Rd1-rev	CAACTCCTCCACCGCAG
F280X	Rd1-for	NNKCTCTCCATCGCAGACGGC	Rd1-rev	GCGCAGCAACTCCTCC
L281X	Rd1-for	NNKTCCATCGCAGACGGCATG	Rd1-rev	AAAGCGCAGCAACTCCTC

Mutation	Forward primer (5' to 3')		Reverse primer (5' to 3')	
S282X	Rd1-for	NNKATCGCAGACGGCATGC	Rd1-rev	GAGAAAGCGCAGCAACTCC
I283X	Rd1-for	NNKGCAGACGGCATGCTGC	Rd1-rev	GGAGAGAAAGCGCAGCAAC
A284X	Rd1-for	NNKGACGGCATGCTGCGC	Rd1-rev	GATGGAGAGAAAGCGCAGC
D285X	Rd1-for	NNKGGCATGCTGCGCGTTG	Rd1-rev	TGCGATGGAGAGAAAGCG
G286X	Rd1-for	NNKATGCTGCGCGTTGCC	Rd1-rev	GTCTGCGATGGAGAGAAAGC
M287X	Rd1-for	NNKCTGCGCGTTGCCAAAGAAG	Rd1-rev	GCCGTCTGCGATGGAG
L288X	Rd1-for	NNKCGCGTTGCCAAAGAAGATATTG	Rd1-rev	CATGCCGTCTGCGATGG
V290X	Rd4-for	NDT/VHG/TGG-GCCAAAGAAGATATTGATGTTGACG	Rd4-rev	GCGCAGCATGCCGTC
G307X	Rd4-for	NDT/VHG/TGG-GTTGTGTTTAGCACCTCCGTTATTAAC	Rd4-rev	CTCATCCGCACGAATGGTC
V308X	Rd4-for	NDT/VHG/TGG-GTGTTTAGCACCTCCGTTATTAACC	Rd4-rev	ACCCTCATCCGCACGAATG
V309X	Rd1-for	NNKTTTAGCACCTCCGTTATTAACCG	Rd1-rev	AACACCCTCATCCGCAC
F310X	Rd1-for	NNKAGCACCTCCGTTATTAACCG	Rd1-rev	CACAACACCCTCATCCGC
S311X	Rd1-for	NNKACCTCCGTTATTAACCGTGACAC	Rd1-rev	AAACACAACACCCTCATCCG
T312X	Rd1-for	NNKTCCGTTATTAACCGTGACACC	Rd1-rev	GCTAAACACAACACCCTCATCC
S313X	Rd1-for	NNKGTTATTAACCGTGACACCACCG	Rd1-rev	GGTGCTAAACACAACACCCTC
V314X	Rd4-for	NDT/VHG/TGG-ATTAACCGTGACACCACCGC	Rd4-rev	GGAGGTGCTAAACACAACACC
I315X	Rd1-for	NNKAACCGTGACACCACCGC	Rd1-rev	AACGGAGGTGCTAAACACAAC
R317X	Rd4-for	NDT/VHG/TGG-GACACCACCGCGTACCC	Rd4-rev	GTTAATAACGGAGGTGCTAAACACAAC
T319X	Rd3-for	NDT/VHG/TGG-ACCGCGTACCCGGAG	Rd3-rev	GTCACGGTTAATAACGGAGGTG
Y322X	Rd3-for	NDT/VHG/TGG-CCGGAGCCGGATACCC	Rd3-rev	CGCGGTGGTGTCACG
H336X	Rd1-for	NNKCACGTGGCGTTCGGC	Rd1-rev	ACGAGCACTACGATGCCAATC
H337X	Rd1-for	NNKGTGGCGTTCGGCTTTGG	Rd1-rev	GTGACGAGCACTACGATGC
V338X	Rd1-for	NNKGCGTTCGGCTTTGGCATTC	Rd1-rev	GTGGTGACGAGCACTACG
A339X	Rd1-for	NNKTTCGGCTTTGGCATTCACC	Rd1-rev	CACGTGGTGACGAGCAC
F342X	Rd1-for	NNKGGCATTCACCAGTGTCTGG	Rd1-rev	GCCGAACGCCACGTG
I344X	Rd1-for	NNKCACCAGTGTCTGGGCC	Rd1-rev	GCCAAAGCCGAACGCC
Q346X	Rd1-for	NNKTGTCTGGGCCAAAACCTG	Rd1-rev	GTGAATGCCAAAGCCGAAC
L348X	Rd1-for	NNKGGCCAAAACCTGGCGC	Rd1-rev	ACACTGGTGAATGCCAAAGC
Q350X	Rd1-for	NNKAACCTGGCGCGCGCAG	Rd1-rev	GCCCAGACACTGGTGAATG
N351X	Rd1-for	NNKCTGGCGCGCGCAGAG	Rd1-rev	TTGGCCCAGACACTGGTG
A355X	Rd1-for	NNKGAGATGGAAATCGCGCTGC	Rd1-rev	GCGCGCCAGGTTTTGG
M357X	Rd1-for	NNKGAAATCGCGCTGCGCAC	Rd1-rev	CTCTGCGCGCGCCAG
E358X	Rd1-for	NNKATCGCGCTGCGCACC	Rd1-rev	CATCTCTGCGCGCGC
K383X	Rd1-for	NNKCCGGGTGATACCCTGCAAG	Rd1-rev	AAACGGGATCTCTTCTGCCG
P384X	Rd4-for	TGGGGTGATACCCTGCAAGGTATG	Rd4-rev	TTTAAACGGGATCTCTTCTGCC
G385X	Rd4-for	TGGGATACCCTGCAAGGTATGATCG	Rd4-rev	CGGTTTAAACGGGATCTCTTCTG
D386X	Rd1-for	NNKACCCTGCAAGGTATGATCGAG	Rd1-rev	ACCCGGTTTAAACGGGATC
T387X	Rd1-for	NNKCTGCAAGGTATGATCGAGCTG	Rd1-rev	ATCACCCGGTTTAAACGGG
L388X	Rd1-for	NNKCAAGGTATGATCGAGCTGCC	Rd1-rev	GGTATCACCCGGTTTAAACGG
Q389X	Rd1-for	NNKGGTATGATCGAGCTGCCG	Rd1-rev	CAGGGTATCACCCGGTTTAAAC
M391X	Rd4-for	NDT/VHG/TGG-ATCGAGCTGCCGGTTACTTG	Rd4-rev	ACCTTGCAGGGTATCACCC

Table 5.6. Primers for multiple site-directed mutagenesis combinatorial libraries.

Mutation(s)	Forward primer (5' to 3')	
S55K	D12Rd1-combi-1	TGCACGTCAGTTGCTGAAAGATCAACGTCTGTCGACCGATAGC
S55R	D12Rd1-combi-2	AGCTGCACGTCAGTTGCTGCGTGATCAACGTCTGTCGAC
F71L	D12Rd1-combi-3	AATGAAGCCTTCCCGATTCTGGCGGAGCGCGTG
T86L	D12Rd1-combi-4	TCGTCCGAGACGTCTGGCATTGCTGGGCGAAG
T86V	D12Rd1-combi-5	TCGTCCGAGACGTGTGGCATTGCTGGGCGAAG
H97R	D12Rd1-combi-6	CGACCCGGTGCATCGTGCTCAAAGACGG
I104M	D12Rd1-combi-7	GACGGATGCTGATGCCGTCTTTTACCTTAAAGCGC
I104Q	D12Rd1-combi-8	CATCATGCTCAAAGACGGATGCTGCAGCCGTCTTTTACCTTAAAGC
I104A	D12Rd1-combi-9	CATGCTCAAAGACGGATGCTGGCGCCGTCTTTTACCTTAAAGC
S106F	D12Rd1-combi-10	CGGATGCTGATCCCGTTTTTTACCTTAAAGCGCACC
I104M, S106F	D12Rd1-combi-11	AGACGGATGCTGATGCCGTTTTTTACCTTAAAGCGCACCG
I104Q, S106F	D12Rd1-combi-12	GACGGATGCTGCAGCCGTTTTTTACCTTAAAGCGCACCGCCGAC
I104A, S106F	D12Rd1-combi-13	CTCAAAGACGGATGCTGGCCCCGTTTTTTACCTTAAAGCGCAC
R111A	D12Rd1-combi-14	TCCCGTCTTTTACCTTAAAGGCGACCGCCGACCTTCGTC
S170T	D12Rd1-combi-15	TCTTCGAGGGCCAAACCAGACGTCTGTTACGC
H337V	D12Rd1-combi-16	ATCGTAGTGCTCGTCACGTGGTGGCGTTCGGCTTTGG
N351W	D12Rd1-combi-17	CCAGTGTCTGGGCCAATGGCTGGCGCGCGCAG
Y15M	D12Rd2-combi-1	ATTTCCCCAAAATCGCACGTGCCCGATGCATCCGCCAAC
H16N	D12Rd2-combi-2	ACGTGCCCGTACAACCCGCCAACGGCCTAC
Y15M, H16N	D12Rd2-combi-3	CCCAAAATCGCACGTGCCCGATGAACCCGCCAACGGCCTACG
H47L	D12Rd2-combi-4	TTTGGGTTGTCACGGGTCTGGCAGCTGCACGTC
F71L	D12Rd2-combi-5	AATGAAGCCTTCCCGATTCTGGCGGAGCGCGTG
A72Q	D12Rd2-combi-6	AATGAAGCCTTCCCGATTTTCCAGGAGCGCGTGGC
E73S	D12Rd2-combi-7	GCCTTCCCGATTTTCGCGAGCCGCGTGGCTGCG
V75D	D12Rd2-combi-8	ATTTTCGCGGAGCGCGATGCTGCGGTTCGTTCTC
V75Q	D12Rd2-combi-9	ATTTTCGCGGAGCGCCAGGCTGCGGTTCGTTC
H97R	D12Rd2-combi-10	CGACCCGGTGCATCGTGCTCAAAGACGG
Q169L	D12Rd2-combi-11	ACGATTTCTTCGAGGGCCTGAGCAGACGTCTGTTACG
R171V	D12Rd2-combi-12	ATTTCTTCGAGGGCCAAAGCGTGCGTCTGTTACGCGGTAG
Q169L, R171V	D12Rd2-combi-13	GACCACGATTTCTTCGAGGGCCTGAGCGTGCGTCTGTTACGCGG
P384E	D12Rd2-combi-14	AAGAGATCCCGTTTAAAGAAGGTGATACCCTGCAAGGTATGATCG

Table 5.7. Inner primer pairs for the amplification of the target regions in P450_D12.

Target region	Forward primer (5' to 3')	Reverse primer (5' to 3')
Positions 50–119	CACCCAAGACCACTCTCCGGTTGTCACGGGTCACGCAGC	CGGTGTGCGAAGTAGGTGCCGATGAGTTCGTCGACCGCAC
Positions 144–208	CACCCAAGACCACTCTCCGGCCTGCTGAATTGGTTAGCGCGTTC	CGGTGTGCGAAGTAGGTGCTTGCTGCGCGATCAGATCGTC
Positions 228–288	CACCCAAGACCACTCTCCGGGCATGGTCACCTGGACAAGCAG	CGGTGTGCGAAGTAGGTGCGTCAACATCAATATCTTCTTTGGCAACGCG
Positions 309–389	CACCCAAGACCACTCTCCGGCCATTCGTGCGGATGAGGGTGTT	CGGTGTGCGAAGTAGGTGCGAATTCCCAAGTAACCGGCAGCTCGATC

Table 5.8. Forward primers for next-generation sequencing barcoding.

Well Position	Forward primer for barcoding (5' to 3')
A01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATCATGCACCCAAGACCACTCTCCGG
A02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTACATGGCACCCAAGACCACTCTCCGG
A03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAAGCACCCACCCAAGACCACTCTCCGG
A04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGGCTCACACCCAAGACCACTCTCCGG
A05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTGCTCCACCCAAGACCACTCTCCGG
A06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGCGTCACCCAAGACCACTCTCCGG
A07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCTCCATCACCCAAGACCACTCTCCGG
A08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGAAGGCACCCAAGACCACTCTCCGG
A09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAATGTCCACCCAAGACCACTCTCCGG
A10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCTCCACACCCAAGACCACTCTCCGG
A11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCGTTATCACCCAAGACCACTCTCCGG
A12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCACCACACCCAAGACCACTCTCCGG
B01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCCTATCACCCAAGACCACTCTCCGG
B02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGGAATCCACCCAAGACCACTCTCCGG
B03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCACTGCACCCAAGACCACTCTCCGG
B04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGTACCCACCCAAGACCACTCTCCGG
B05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCGAGTCACCCAAGACCACTCTCCGG
B06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCAGCCACCCAAGACCACTCTCCGG
B07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTGCACACCCAAGACCACTCTCCGG
B08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCTGTCCACCCAAGACCACTCTCCGG
B09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGCCATTCACCCAAGACCACTCTCCGG
B10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCCATGACACCCAAGACCACTCTCCGG
B11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCACAACGCACCCAAGACCACTCTCCGG
B12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTCGCTCACCCAAGACCACTCTCCGG
C01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGTGAACACCCAAGACCACTCTCCGG
C02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTATCGGCACCCAAGACCACTCTCCGG
C03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGACCATCACCCAAGACCACTCTCCGG
C04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACATGAGCACCCAAGACCACTCTCCGG
C05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACGTACTCACCCAAGACCACTCTCCGG
C06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCACCTCACACCCAAGACCACTCTCCGG
C07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTGGAGCACCCAAGACCACTCTCCGG
C08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTTCTGCACCCAAGACCACTCTCCGG
C09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTTACGTCACCCAAGACCACTCTCCGG
C10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGGTTGCACCCAAGACCACTCTCCGG
C11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGGACACACCCAAGACCACTCTCCGG
C12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACACTGACACCCAAGACCACTCTCCGG
D01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATCTGTGCACCCAAGACCACTCTCCGG
D02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATGTGCCACCCAAGACCACTCTCCGG
D03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGTTGACACCCAAGACCACTCTCCGG
D04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCACCACCCAAGACCACTCTCCGG
D05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGAAGCGCACCCAAGACCACTCTCCGG
D06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCTACAACACCCAAGACCACTCTCCGG
D07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAGAACCACCCAAGACCACTCTCCGG
D08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGAGTGCACCCAAGACCACTCTCCGG
D09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCCGAACACCCAAGACCACTCTCCGG
D10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTACGACCACCCAAGACCACTCTCCGG
D11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACTCTTGCACCCAAGACCACTCTCCGG
D12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCAACCACACCCAAGACCACTCTCCGG
E01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCTAGACACCCAAGACCACTCTCCGG
E02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATCGGACACCCAAGACCACTCTCCGG
E03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGTCCTACACCCAAGACCACTCTCCGG
E04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGAATGTCACCCAAGACCACTCTCCGG
E05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCAAGCCACCCAAGACCACTCTCCGG
E06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCACCTACACCCAAGACCACTCTCCGG
E07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTGCGTTCACCCAAGACCACTCTCCGG
E08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGGATTCACCCAAGACCACTCTCCGG
E09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGCATACACCCAAGACCACTCTCCGG
E10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGTTGAGCACCCAAGACCACTCTCCGG
E11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGCTACTCACCCAAGACCACTCTCCGG
E12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGATCCCACCCAAGACCACTCTCCGG

Well Position	Forward primer for barcoding (5' to 3')
F01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCATGGTCACCCAAGACCACTCTCCGG
F02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTCGTTACACCCAAGACCACTCTCCGG
F03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTGACACACCCAAGACCACTCTCCGG
F04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTGTAGCACCCAAGACCACTCTCCGG
F05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGAGCAACACCCAAGACCACTCTCCGG
F06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTACTCCCACCCAAGACCACTCTCCGG
F07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATGCCACACCCAAGACCACTCTCCGG
F08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACCGATCACCCAAGACCACTCTCCGG
F09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACGTTGGCACCCAAGACCACTCTCCGG
F10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGAGCGCACCCAAGACCACTCTCCGG
F11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTACTCCGCACCCAAGACCACTCTCCGG
F12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATTCACCACCCAAGACCACTCTCCGG
G01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGACGCCACCCAAGACCACTCTCCGG
G02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTTGTTCACCCAAGACCACTCTCCGG
G03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTACTTGCACCCAAGACCACTCTCCGG
G04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAGCAAGCACCCAAGACCACTCTCCGG
G05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGCCATCACCCAAGACCACTCTCCGG
G06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGAACACACCCAAGACCACTCTCCGG
G07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTATAGCCACCCAAGACCACTCTCCGG
G08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGATGGACACCCAAGACCACTCTCCGG
G09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGCAGTACACCCAAGACCACTCTCCGG
G10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAAGAAGCACCCAAGACCACTCTCCGG
G11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGCGGTTCACCCAAGACCACTCTCCGG
G12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTAAGGCCCACCCAAGACCACTCTCCGG
H01	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAACCTGTCACCCAAGACCACTCTCCGG
H02	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTACACCACCCAAGACCACTCTCCGG
H03	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCGTAGCACCCAAGACCACTCTCCGG
H04	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTAGGTGCACCCAAGACCACTCTCCGG
H05	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGATACCCACCCAAGACCACTCTCCGG
H06	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCGGCTACACCCAAGACCACTCTCCGG
H07	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGGTTGTCACCCAAGACCACTCTCCGG
H08	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTGCCTCACCCAAGACCACTCTCCGG
H09	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCATTCGACACCCAAGACCACTCTCCGG
H10	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCACAATCACCCAAGACCACTCTCCGG
H11	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCAGTAACACCCAAGACCACTCTCCGG
H12	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTAATCCACCCAAGACCACTCTCCGG

Table 5.9. Reverse primers for next-generation sequencing barcoding.

Well Position	Sequence
A01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAACTGCCGGTGTGCGAAGTAGGTGC
A02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCAGGTCGGTGTGCGAAGTAGGTGC
A03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTAGAGCGGTGTGCGAAGTAGGTGC
A04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCACACAACGGTGTGCGAAGTAGGTGC
A05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGGAACCGGTGTGCGAAGTAGGTGC
A06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATATGCCCGGTGTGCGAAGTAGGTGC
A07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTCTGACGGTGTGCGAAGTAGGTGC
A08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGAGATCGGTGTGCGAAGTAGGTGC
A09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTGGCAGCGGTGTGCGAAGTAGGTGC
A10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGCCTGCGGTGTGCGAAGTAGGTGC
A11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCGAAGCGGTGTGCGAAGTAGGTGC
A12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCTTACCGGTGTGCGAAGTAGGTGC
B01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGTTGGCCGGTGTGCGAAGTAGGTGC
B02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACGATGCGGTGTGCGAAGTAGGTGC
B03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTACCGCGGTGTGCGAAGTAGGTGC
B04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTGTCTCGGTGTGCGAAGTAGGTGC
B05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCAGCTTCGGTGTGCGAAGTAGGTGC
B06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTAGACGCGGTGTGCGAAGTAGGTGC
B07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCATACCGGTGTGCGAAGTAGGTGC
B08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACGACTCGGTGTGCGAAGTAGGTGC
B09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTCACCTCGGTGTGCGAAGTAGGTGC
B10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTGATGCGGTGTGCGAAGTAGGTGC
B11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTTCCTCGGTGTGCGAAGTAGGTGC
B12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACGTCGGTGTGCGAAGTAGGTGC
C01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGGACTTCGGTGTGCGAAGTAGGTGC
C02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACCGGAACGGTGTGCGAAGTAGGTGC
C03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGAAGTCGGTGTGCGAAGTAGGTGC
C04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCACGCACGGTGTGCGAAGTAGGTGC
C05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATCCTCGCGGTGTGCGAAGTAGGTGC
C06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGAATAGCGGTGTGCGAAGTAGGTGC
C07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTATCCGGCGGTGTGCGAAGTAGGTGC
C08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCAAGACGGTGTGCGAAGTAGGTGC
C09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTCGACCGGTGTGCGAAGTAGGTGC
C10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCCATGCGGTGTGCGAAGTAGGTGC
C11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCAATCGCGGTGTGCGAAGTAGGTGC
C12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGAGTGGCGGTGTGCGAAGTAGGTGC

Well Position	Sequence
D01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGGAGACGGTGTGCGAAGTAGGTGC
D02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGTCAGTCGGTGTGCGAAGTAGGTGC
D03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGCTGTCGGTGTGCGAAGTAGGTGC
D04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAACAACCGGTGTGCGAAGTAGGTGC
D05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAATAGCCCGGTGTGCGAAGTAGGTGC
D06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTGTGACGGTGTGCGAAGTAGGTGC
D07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTGGTACGGTGTGCGAAGTAGGTGC
D08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCGTATGCGGTGTGCGAAGTAGGTGC
D09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGTTACGCGGTGTGCGAAGTAGGTGC
D10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCCTGCCGGTGTGCGAAGTAGGTGC
D11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTATGTCGCGGTGTGCGAAGTAGGTGC
D12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGAGAGACGGTGTGCGAAGTAGGTGC
E01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTTAGGCGGTGTGCGAAGTAGGTGC
E02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTATCCCGGTGTGCGAAGTAGGTGC
E03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAACCTGCGGTGTGCGAAGTAGGTGC
E04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTGATGACGGTGTGCGAAGTAGGTGC
E05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAGACAGCGGTGTGCGAAGTAGGTGC
E06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCTCGTCGGTGTGCGAAGTAGGTGC
E07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATTGCGCGGTGTGCGAAGTAGGTGC
E08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCCTTCACGGTGTGCGAAGTAGGTGC
E09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCACAGGCGGTGTGCGAAGTAGGTGC
E10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGAGCTGCGGTGTGCGAAGTAGGTGC
E11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCTGTCGGTGTGCGAAGTAGGTGC
E12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTCGAACGGTGTGCGAAGTAGGTGC
F01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGTCTCCGGTGTGCGAAGTAGGTGC
F02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTGAGGCGGTGTGCGAAGTAGGTGC
F03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACAATCCGGTGTGCGAAGTAGGTGC
F04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCACTTGCCGGTGTGCGAAGTAGGTGC
F05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGAACGCCGGTGTGCGAAGTAGGTGC
F06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTAGCACGGTGTGCGAAGTAGGTGC
F07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGGTTCCCGGTGTGCGAAGTAGGTGC
F08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTACACACGGTGTGCGAAGTAGGTGC
F09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATAGGTCGGTGTGCGAAGTAGGTGC
F10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGCCTCCGGTGTGCGAAGTAGGTGC
F11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCAGCCCGGTGTGCGAAGTAGGTGC
F12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGATTCACGGTGTGCGAAGTAGGTGC
G01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGAGCCTCGGTGTGCGAAGTAGGTGC
G02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACGCGACGGTGTGCGAAGTAGGTGC
G03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCATTGCCGGTGTGCGAAGTAGGTGC
G04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGCATCTCGGTGTGCGAAGTAGGTGC
G05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTGGTCTCGGTGTGCGAAGTAGGTGC
G06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAAGGATCGGTGTGCGAAGTAGGTGC
G07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGACGTCCGGTGTGCGAAGTAGGTGC
G08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAGGTCAACGGTGTGCGAAGTAGGTGC
G09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATGCTACCGGTGTGCGAAGTAGGTGC
G10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCTGATCGGTGTGCGAAGTAGGTGC
G11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCAAGTCCGGTGTGCGAAGTAGGTGC
G12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCGAGCTCGGTGTGCGAAGTAGGTGC
H01	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAGTCTCGGTGTGCGAAGTAGGTGC
H02	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGATACCGGTGTGCGAAGTAGGTGC
H03	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTACGTTCCGGTGTGCGAAGTAGGTGC
H04	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACGGTTCCGGTGTGCGAAGTAGGTGC
H05	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATCGTCCGGTGTGCGAAGTAGGTGC
H06	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTACGCATCGGTGTGCGAAGTAGGTGC
H07	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTAGACCGGTGTGCGAAGTAGGTGC
H08	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAACTGACCGGTGTGCGAAGTAGGTGC
H09	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTTGCACGGTGTGCGAAGTAGGTGC
H10	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACGCGATCGGTGTGCGAAGTAGGTGC
H11	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCGACACCGGTGTGCGAAGTAGGTGC
H12	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACTCAACCGGTGTGCGAAGTAGGTGC

5.6 Supporting Information

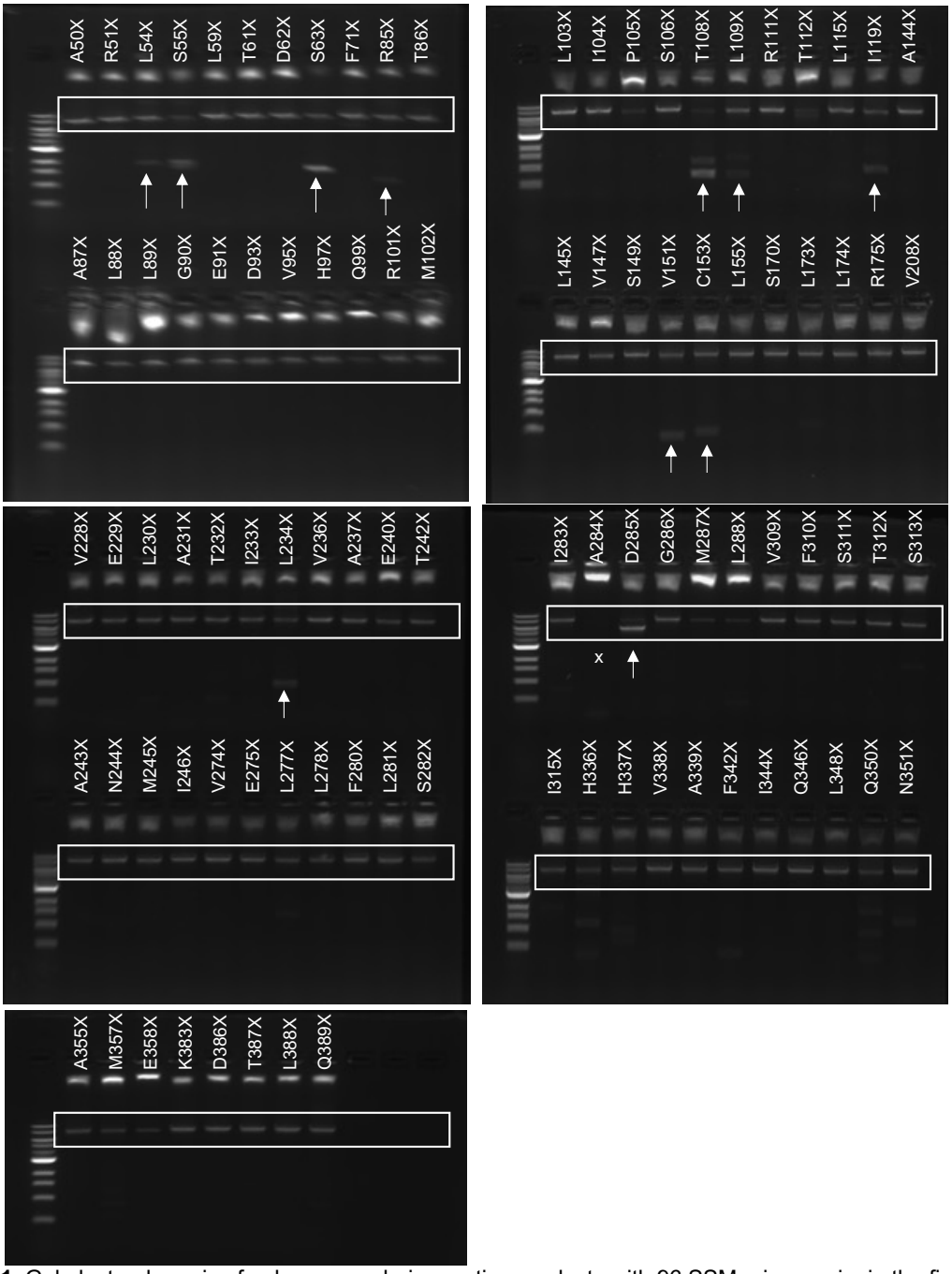


Figure S5.1. Gel electrophoresis of polymerase chain reaction products with 96 SSM primer pairs in the first round. Note: the arrows indicate products with additional bands and the cross means no amplification.

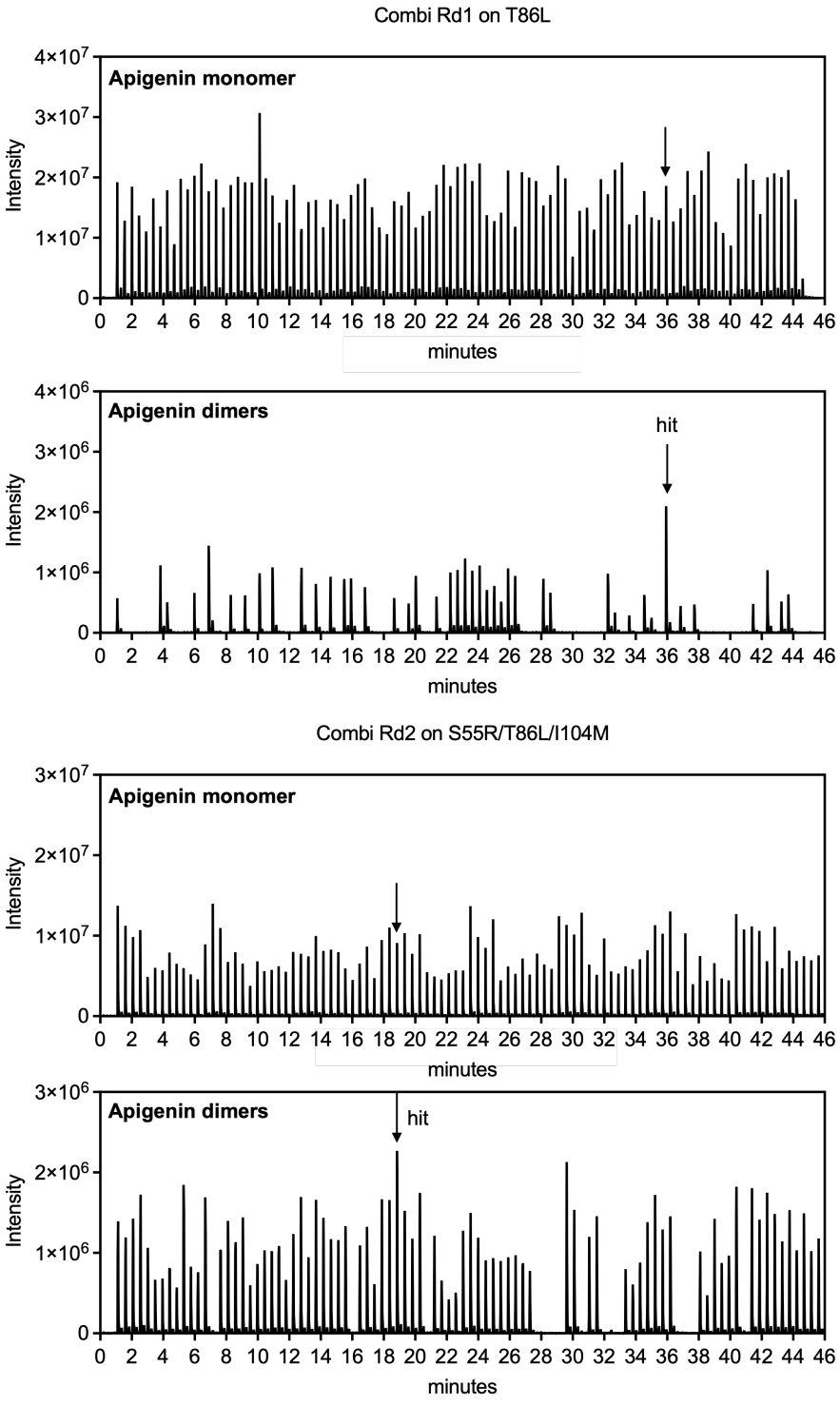


Figure S5.2. Apigenin dimerization of major P450_D12 variants analyzed by RapidFire-MS. Extracted ion chromatograms include dimerized apigenin (with $[M\text{-}H]^-$ 537.0827) and apigenin monomer (with $[M-H]^-$ of 269.0455) The major variants include S55R/T86L/I104M and the final variant with eight substitutions (H16N/H47L/S55R/V75Q/T86L/H97R/I104M/Q169L).

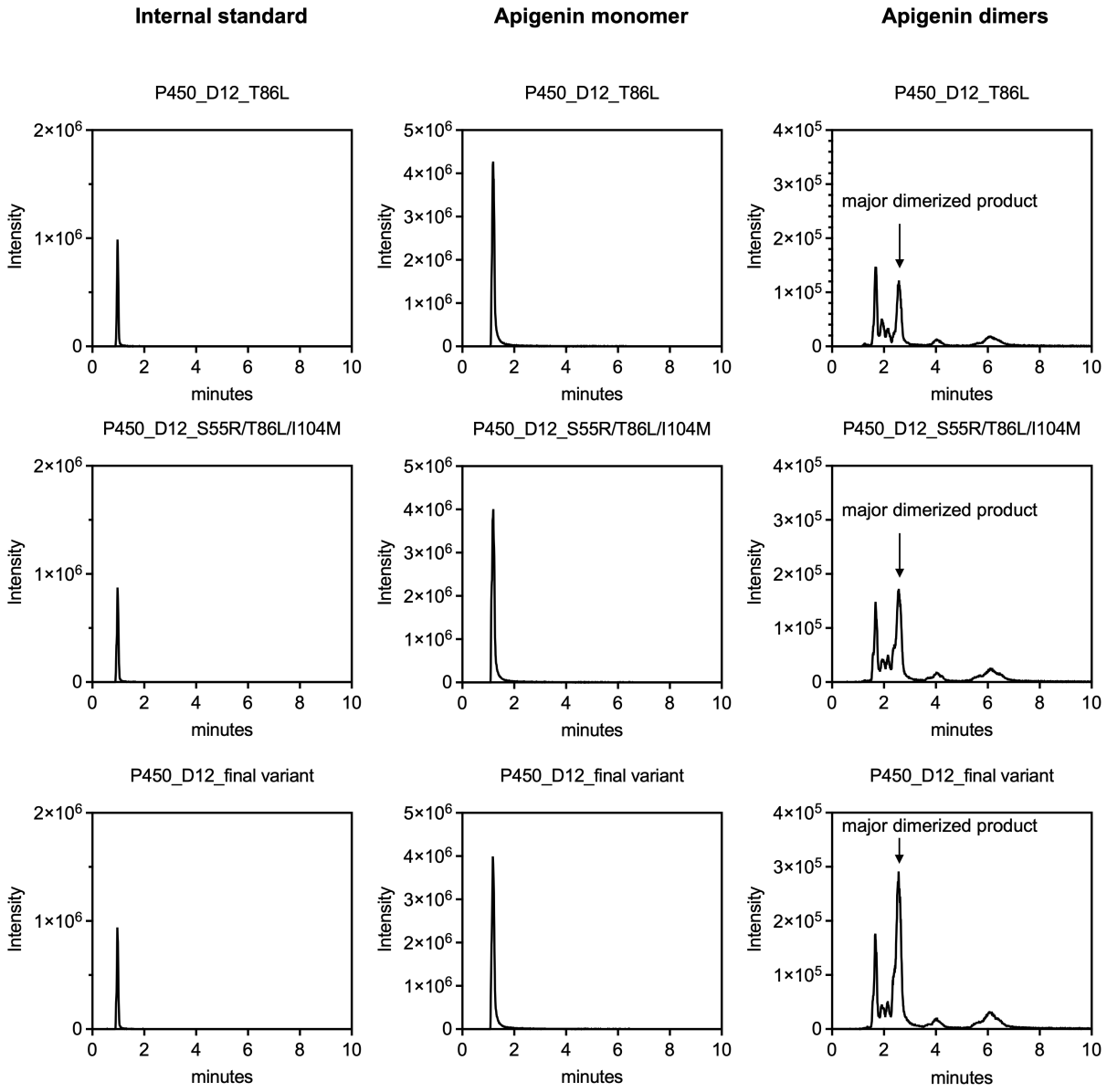


Figure S5.3. Apigenin dimerization of major P450_D12 variants analyzed by LC-MS.
Extracted ion chromatograms include dimerized apigenin (with $[M-H]^-$ 537.0827), apigenin monomer (with $[M-H]^-$ of 269.0455), and internal standard 2,6'-dihydroxyacetophenone (with $[M-H]^-$ of 151.0401). Three major variants include T86L, S55R/T86L/I104M, and the final variant (H16N/H47L/S55R/V75Q/T86L/H97R/I104M/Q169L).

Table S5.1. Experimental results and computational predictions of mutational effects.

Mutant	Experimental	EV (epistatic)	EV (independent)	ESM-1v	Mutant	Experimental	EV (epistatic)	EV (independent)	ESM-1v
A50I	0.00	-8.418816538	0.033685055	-11.19180512	L89Q	0.17	-4.615064148	-3.189090348	-11.38935304
A50R	0.00	-11.95149685	-9.703119648	-15.069947	L89D	0.00	-5.266933941	-5.482246877	-12.89826512
A50D	0.00	-10.31208859	-7.385050809	-13.51552343	L89Y	0.00	-6.22640805	-4.74079742	-12.35598564
A50G	0.10	-4.729203384	-2.473606072	-7.220741749	L89W	0.04	-6.723438575	-6.222211466	-12.62222838
A50P	0.00	-11.6310108	-9.033277045	-13.53115463	L89M	0.90	-3.608272007	-2.641257396	-6.810650945
A50V	0.00	-4.871133047	1.112018022	-7.378046036	L89G	0.07	-3.742704739	-2.81774433	-11.8056376
A50Q	0.00	-7.061033436	-3.025259015	-10.97384048	L89E	0.00	-5.415497935	-4.334534507	-13.4797287
A50E	0.07	-10.51145498	-7.743972672	-12.75476956	L89I	0.95	-0.875607329	-0.743460966	-5.09515667
A50L	0.02	-6.926883835	-1.209544496	-11.39860654	G90P	0.00	-7.16435295	-2.857522722	-10.04434896
A50S	0.74	-6.246428425	-2.962711693	-8.196434617	G90T	0.21	-4.485094144	0.494468863	-8.244710803
A50C	0.44	-7.555172944	-0.933732541	-10.04020596	G90S	0.99	-3.920801919	0.19634	-7.208656192
A50K	0.04	-12.21999916	-10.63192851	-15.76692748	G90V	0.00	-5.00234618	-0.552142671	-9.50105834
A50W	0.00	-12.21999916	-10.6319286	-15.46056318	G90L	0.09	-7.8540445	-0.713838008	-9.539644241
A50H	0.00	-8.551750116	-5.90666048	-13.51787996	E91F	0.85	-0.702619037	-0.228909791	-2.032340169
A50T	0.36	-5.92401201	-2.28313033	-9.09818697	E91W	0.46	-1.841362076	-3.001529261	-4.670358121
R51D	0.00	-9.16646651	-2.247054169	-14.90376496	E91A	0.18	0.672754629	1.559889154	0.798163116
R51L	0.17	-5.963230315	-1.575761207	-11.32860613	E91K	0.26	-3.024468112	0.24329658	-4.02459234
R51K	0.15	-3.9331767	-1.062067755	-6.958171248	E91R	0.33	-1.14398156	0.656296683	-2.75324297
R51G	0.04	-7.108864389	-3.616629395	-11.3805933	E91M	0.73	0.558749249	2.620682728	1.237301201
R51V	0.10	-6.821365433	-1.260435915	-11.18433213	D93Y	0.51	-8.778405458	-5.444255294	-12.4240725

Mutant	Experimental	EV (epistatic)	EV (independent)	ESM-1v	Mutant	Experimental	EV (epistatic)	EV (independent)	ESM-1v
L54T	0.00	-6.184818043	-3.536378246	-11.86366081	D93V	0.18	-7.327387892	-3.812753539	-10.06160021
L54S	0.00	-7.889881847	-2.149452613	-12.19430637	D93I	0.18	-8.93922465	-5.024504058	-12.14111376
L54H	0.00	-6.258565108	-4.443580684	-12.46162391	D93E	0.63	-4.73610508	-1.009441209	-5.852906704
L54F	0.34	-4.348768668	-2.250942107	-6.443132401	D93A	0.27	-5.930943842	-0.201826369	-6.225539446
L54P	0.00	-10.2010626	-8.87013221	-14.32940388	D93S	0.32	-6.233456883	-2.510880414	-8.61661458
L54G	0.02	-7.208105844	-4.452473221	-14.71864057	D93K	0.35	-7.627376423	-3.130262922	-10.19286394
S55W	1.01	-7.98936824	-6.537853913	-12.52162313	D93R	0.40	-6.453176429	-1.741578582	-9.531466842
S55R	1.40	-3.034831984	1.228275926	-8.027295589	D93C	0.37	-11.70467412	-9.758601529	-13.70480394
L59M	0.62	-3.945959894	-2.420757835	-7.489236355	D93G	1.27	-4.524533701	0.932145979	-7.457519054
L59H	0.00	-7.086750224	-2.428119936	-15.115345	D93P	0.71	-4.045899029	2.037329715	-6.1231637
L59N	0.00	-7.123653276	-2.699485905	-15.44476914	D93H	0.34	-7.997714811	-3.145676486	-9.822127342
L59R	0.01	-7.790964362	-1.22650495	-15.11342573	D93L	0.26	-6.265477125	-3.110898034	-10.28757691
L59P	0.00	-9.223622215	-4.607071225	-14.29702139	D93F	0.47	-8.575136105	-5.139741361	-12.60427356
L59K	0.00	-8.623663076	-3.871095999	-17.30720782	D93W	0.48	-9.134016028	-5.665623754	-14.78057408
L59Q	0.00	-7.945525299	-3.565092492	-15.22349405	D93Q	0.43	-6.126569839	-1.922355033	-8.730979562
L59C	0.31	-6.877947098	-4.370835979	-11.4073751	V95S	0.59	-4.225038077	-1.087942132	-5.841755152
L59V	0.32	-4.045644064	-0.667870883	-4.706929088	V95L	1.07	-0.811183605	-0.176757477	-5.002213001
T61P	0.17	-4.999643083	-1.939255987	-7.314813733	V95C	0.41	-6.857150913	-5.489886994	-10.70435572
T61I	0.89	-3.166973645	-1.181126478	-5.665272117	V95G	0.24	-2.060035844	-0.799544397	-4.416896045
T61N	0.67	-3.490597781	-1.119673519	-5.799340844	V95Y	0.54	-5.068483983	-1.621212122	-7.603082061
T61V	1.26	-1.193368206	-0.229759902	-3.473904133	V95Q	0.56	-2.590102593	0.118435735	-2.777763128
T61E	0.00	-6.553897421	-2.768161435	-10.53136706	V95T	0.76	-3.455004374	-0.716518764	-4.769456744
T61K	0.47	-5.939338779	0.290074356	-8.031914592	V95A	0.68	-1.424766604	0.430455599	-2.757731318
T61A	0.75	-1.697007647	-0.04807507	-1.300765619	V95R	0.75	-1.237105846	1.18913285	-4.229492188
T61Q	0.14	-5.919743047	-1.716483692	-8.544961572	V95E	0.90	1.66008657	1.986918827	-0.27707234
T61D	0.00	-5.842852327	-3.293255975	-9.824991703	H97R	1.37	-0.954882045	1.989701254	-2.836956173
T61R	0.08	-3.593058189	0.643026892	-7.259787261	H97W	0.78	-4.808020433	-2.946687966	-9.099781871
T61G	0.15	-4.780478343	-1.783146006	-6.088223577	H97T	0.92	-0.821807416	3.271096344	-3.291104734
D62M	0.11	-7.062604425	-3.899570792	-11.23477244	H97K	1.17	-0.000813738	0.555347489	-5.141774893
D62V	0.14	-4.721795933	-1.997511167	-8.42181778	H97L	0.94	-2.573929505	0.449585909	-4.909465313
D62L	0.09	-6.242694548	-2.946519347	-9.477085352	H97P	0.26	-3.040548705	-0.746688487	-7.287032843
D62P	0.00	-5.67591891	-2.126325737	-12.42859411	H97A	1.26	-0.305078554	1.86872712	-2.172561705
D62R	0.16	-5.004620222	-1.283952852	-8.121744275	Q99H	0.77	-4.807908793	-0.080818857	-7.572267413
D62C	0.23	-7.785922049	-4.203979237	-10.85891104	Q99S	0.14	-8.433209736	-2.321069928	-10.62596345
D62Q	0.26	-5.770428132	-2.20720226	-8.979334712	Q99R	0.52	-5.356074582	-0.559529282	-8.791040182
D62G	0.19	-5.514986731	-1.745893729	-8.928813219	R101L	0.51	-7.65273548	-3.683257929	-10.56804013
D62F	0.13	-7.534151289	-4.317832886	-10.93690896	R101Y	0.48	-9.642285856	-7.050729292	-13.96754622
D62S	0.19	-5.923342152	0.761239825	-6.808398128	R101G	0.49	-5.139968709	-1.39883431	-9.181841135
D62Y	0.11	-6.861455928	-4.471964667	-10.06916726	R101I	0.64	-8.66643708	-5.263263772	-13.15362692
F71L	1.52	1.305040703	0.802167306	-2.490788043	R101V	0.63	-8.041951994	-4.451413312	-11.95820618
F71V	0.84	1.808191484	0.383733539	-2.342654884	R101H	0.53	-5.578608174	-3.533371662	-9.307699203
F71Y	0.10	-2.18590311	-0.796729025	-4.450547695	R101P	0.04	-7.382438767	-5.505863351	-13.96902061
F71K	0.25	-1.965247179	-0.501756275	-4.692065179	R101C	0.68	-10.18250262	-7.594679508	-14.21082354
F71Q	0.84	-0.210363023	-0.07466607	-3.883829296	R101K	0.89	-2.964142032	-0.673379402	-4.97244215
F71H	0.74	-1.079516876	-1.069819224	-4.172094405	R101A	0.58	-6.176954057	-1.329766873	-8.212349415
F71N	0.33	-1.963527485	-0.7739977	-4.665137112	R101T	0.54	-7.317638199	-2.52826322	-9.267600536
F71S	0.44	-0.395811802	-0.045621074	-2.06793654	R101E	0.15	-7.739856104	-3.508719833	-11.13258958
R85I	0.19	-5.654223259	-1.73423314	-7.268754244	R101W	0.51	-10.67103855	-8.539214666	-14.46979761
R85S	1.02	-4.229540566	-0.396474541	-5.494306087	M102R	0.27	-3.713572676	-1.857886761	-9.63903451
R85D	0.06	-4.222949521	-0.790333844	-7.566621423	M102L	0.42	-1.562267708	2.096405563	-5.891945243
R85C	0.36	-8.009313772	-4.395744967	-12.41166019	M102E	0.09	-5.399188807	-2.897743613	-11.75632453
R85Y	0.00	-6.448654411	-2.3222925	-9.31218791	M102G	0.12	-6.806362815	-3.444250093	-11.58784437
R85E	0.03	-5.591453359	-0.972093311	-6.97396934	M102T	0.55	-3.05499277	-1.020875273	-8.553766489
R85V	0.51	-3.718106302	-0.536998644	-6.156351805	M102S	0.37	-5.132335595	-1.641632324	-9.624593854
R85L	0.24	-5.384086314	-0.852704688	-6.329074264	M102V	0.36	-4.214082989	0.35242587	-8.615214348
R85G	0.31	-5.105027229	0.110972989	-6.210016966	M102P	0.19	-5.10508826	-0.506781346	-9.82388854
R85N	0.85	-4.753708262	-1.248932658	-6.403926134	M102N	0.16	-7.20807268	-3.955682322	-11.40733194
A87R	0.11	-4.995002309	-2.277664708	-6.694710612	M102C	0.36	-6.997898569	-4.237554299	-14.48531175
A87E	0.09	-3.102338894	-2.256507031	-8.312366843	M102A	0.36	-3.423871257	0.039754224	-7.833918095
A87S	0.94	-1.515128131	1.63102793	-2.763719499	M102W	0.38	-6.689453819	-4.644972078	-11.86579895
A87L	0.20	-4.413376509	-0.031550471	-6.621531487	M102Y	0.57	-6.370331962	-3.380403752	-11.79328084
A87F	0.62	-3.58100616	-0.738903225	-7.232941747	M102D	0.03	-6.955618413	-5.593632656	-13.01967311
A87K	0.04	-6.478579486	-3.46428687	-8.271636963	L103Q	0.00	-8.452293647	-5.669932062	-12.69732356
A87M	0.43	-3.734506752	0.577777867	-6.66191864	L103G	0.00	-8.041779116	-4.13839532	-14.9623003
A87G	0.26	-3.818556025	0.204749781	-4.12790364	L103S	0.03	-8.383284959	-4.34859333	-12.65967202
A87W	0.09	-4.293537853	-0.346260032	-8.733571649	L103C	0.32	-7.7617547	-4.582074657	-14.42252254
A87V	0.46	-2.513318759	-0.068572008	-5.953579783	L103D	0.00	-10.5136994	-8.220437544	-16.53930664
A87T	0.88	-1.88946916	0.930999389	-3.827005446	L103H	0.05	-8.716832624	-4.138000735	-13.6551075
L88W	0.00	-9.116945315	-7.080584713	-10.98198271	L103I	0.47	-4.504441896	-0.934569497	-6.92773211
L88D	0.00	-8.479662776	-4.022602951	-10.81731331	L103W	0.00	-8.097480494	-4.742166315	-14.39834666
L88A	0.04	-6.202317121	-3.265060321	-7.323117733	L103T	0.09	-6.357462934	-2.214814155	-10.40109563
L88P	0.03	-6.163090282	-1.425291559	-8.151365042	L103R	0.00	-11.90150275	-10.99643372	-14.04777288
L88R	0.00	-8.231042384	-5.785032857	-10.1860013	L103M	0.00	-4.362365654	-1.786414565	-6.690453529
L88K	0.00	-8.580704309	-6.452027553	-12.02193093	L103V	0.71	-2.34643704	0.632331382	-5.42570734
L88V	0.24	-6.382809986	-2.04057022	-5.363765359	I104M	0.99	-4.647241843	0.815918541	-6.978939176
L88G	0.04	-8.483020435	-4.608193536	-8.657185674	I104G	0.60	-6.036597613	0.976732499	-11.33454823
L88N	0.00	-7.582998097	-4.764404417	-8.974219561	L103S	0.03	-8.383284959	-4.34859333	-12.65967202
L88I	0.31	-4.204673045	-1.03641556	-5.403187752	L103C	0.32	-7.7617547	-4.582074657	-14.42252254
L89A	0.23	-3.354386512	-2.102688502	-9.785561681	L103D	0.00	-10.5136994	-8.220437544	-16.53930664
I104R	1.29	-8.219874869	0.220263765	-12.07073379					
I104V	1.01	-2.258800672	0.34314892	-5.390209675					
I104A	1.01	-4.279666067	2.315655549	-7.226296425					
P105M	0.53	-7.668359828	-4.667306722	-12.1889987					
P105R	0.58	-5.026288639	0.058913812	-6.57362628					
P105K	0.93	-4.53180552	-0.169287878	-7.299167514					
P105H	0.82	-5.366596263	-2.292147957	-9.227360129					
P105G	1.01	-4.74876485	-0.891916201	-5.740999579					
P105E	0.70	-6.58003603	-2.352250079	-11.12453437					
P105V	0.00	-7.33767141	-3.92637179	-11.73560238					
P105L	0.23	-7.911818614	-3.63257279	-11.88670421					
P105A	0.77	-3.858031449	-1.140246788	-6.23426795					
P105S	0.66	-1.816964304	-1.345909461	-5.460596323					
S106M	1.02	-5.03471489	-1.475642329	-9.863280296					
S106T	0.88	-3.723989073	-0.604309955	-6.591669083					
S106A	1.05	-3.082600669	2.229949165	-4.382025361					
S106C	0.64	-6.292880388	-2.912505783	-10.55482364					
S106G	0.79	-3.148001455	0.484667031	-6.500198245					
S106V	0.63	-5.415347589	0.415797494	-8.601485014					
S106R	0.68	-2.725257734	-0.653714653	-6.494644642					
L109M	1.02	-6.161030846	-0.214675998	-4.790029883					
L109K	1.10	-6.480345529	-0.030907406	-10.28333569					
L109G	0.89	-7.450724867	0.273279435	-9.329262733					

Mutant	Experimental	EV (epistatic)	EV (independent)	ESM-1v	Mutant	Experimental	EV (epistatic)	EV (independent)	ESM-1v
L109R	1.31	-6.373593594	1.252640927	-8.743046284					
L109C	0.54	-11.13438561	-6.190781748	-14.13772845					
R111V	0.77	-5.959206299	-2.076707455	-13.04668617					
R111K	0.88	-4.395814589	-2.491050492	-6.265537977					
R111A	0.54	-4.546084709	-0.637472202	-10.41024661					
R111G	0.54	-5.227773132	-3.099970357	-10.79772711					
R111P	0.17	-6.005207114	-4.356104821	-14.54848576					
R111H	0.84	-4.647782697	-2.828003941	-8.581133127					
R111N	0.99	-6.94692217	-2.378976307	-10.62758756					
R111L	0.42	-6.842104564	-3.147050937	-11.70210242					
R111D	0.30	-8.454628815	-4.576313284	-14.52420449					
R111Q	0.96	-4.576431672	-2.285762871	-8.453365684					
R111S	0.87	-4.913024202	-2.059132749	-10.68276882					
R111C	0.64	-8.872004568	-6.666783721	-12.83119106					
R111Y	0.47	-7.677813974	-3.851572426	-12.8791461					
R111M	0.61	-7.637703672	-2.362396812	-12.21473408					
R111F	0.37	-8.104620923	-3.868862513	-14.65201283					
R111W	0.15	-8.684386361	-5.961398163	-13.05101132					
L115H	0.93	-4.070949841	-3.622202681	-11.02972722					
L115N	0.69	-5.845664103	-3.874395874	-12.25533509					
L115A	0.38	-4.569881786	-3.972203894	-10.04753375					
L115R	0.66	-3.411126894	-3.268605707	-10.65844345					
L115I	1.00	-4.219033128	-2.627521823	-7.810911894					
L115Y	0.86	-5.954426964	-2.840825026	-10.19352984					
L115D	0.26	-6.956126737	-6.418161001	-14.26371241					
L115T	0.82	-5.021297733	-4.738330347	-10.40242743					
L115V	0.88	-5.114349386	-3.032330183	-8.721170425					
I119G	0.00	-8.760367962	-7.237613316	-15.9735961					
I119W	0.00	-9.305788129	-4.593389147	-15.18306255					
I119S	0.00	-7.133560348	-5.355679618	-13.82632041					
I119R	0.00	-10.88772353	-9.549833291	-16.88993549					
I119V	0.88	-4.256298057	-0.524717203	-4.827870727					
I119M	0.67	-3.368838386	-2.526704241	-7.90785265					
L115R	0.66	-3.411126894	-3.268605707	-10.65844345					
L115I	1.00	-4.219033128	-2.627521823	-7.810911894					
L115Y	0.86	-5.954426964	-2.840825026	-10.19352984					
L115D	0.26	-6.956126737	-6.418161001	-14.26371241					
L115T	0.82	-5.021297733	-4.738330347	-10.40242743					
L115V	0.88	-5.114349386	-3.032330183	-8.721170425					
I119G	0.00	-8.760367962	-7.237613316	-15.9735961					
I119W	0.00	-9.305788129	-4.593389147	-15.18306255					
I119S	0.00	-7.133560348	-5.355679618	-13.82632041					
I119R	0.00	-10.88772353	-9.549833291	-16.88993549					
I119V	0.88	-4.256298057	-0.524717203	-4.827870727					
I119M	0.67	-3.368838386	-2.526704241	-7.90785265					

Table S5.2. Average predicted scores of mutations at positions 15–397 from EVcouplings.

Positions	Average prediction scores (epistatic)	Positions	Average prediction scores (epistatic)	Positions	Average prediction scores (epistatic)	Positions	Average prediction scores (epistatic)
15	-8.368182869	125	-4.988576275	226	-1.926007746	326	-5.838540686
16	-6.098008	126	-5.912279304	227	-6.438815188	327	-3.141797184
17	-7.48029976	127	-7.336461609	228	-5.373377093	328	-8.475302817
18	-8.514412366	128	-7.156428709	229	-2.190433605	329	-7.562812094
19	-5.110640567	129	-4.550047215	230	-4.605491722	330	-7.357626405
20	-4.993021985	130	-6.495809797	231	-7.453108405	331	-4.828977182
21	-6.020790064	131	-1.803336172	232	-2.562384097	332	-6.637177537
22	-3.304229816	132	-4.084781503	233	-6.094569004	333	-5.237796813
23	-7.025581917	133	-5.012818946	234	-7.697641102	334	-5.419591512
24	-4.406506718	134	-6.134271214	235	-9.315127158	335	-7.36314154
25	-5.839262453	136	-4.849547476	236	-6.390061127	336	-6.165712516
26	-5.167552481	137	-8.020066021	237	-8.127109949	337	-8.521357735
27	-3.022016785	138	-7.114401794	238	-9.298544228	338	-5.884144009
28	-6.095707587	139	-10.06669679	239	-8.623332028	339	-7.800527426
29	-6.530544129	140	-7.215093903	240	-7.988720468	340	-8.569392268
30	-5.14318852	141	-5.529507389	241	-9.176741806	341	-10.20980188
31	-4.049052781	142	-5.497832292	242	-9.542189236	342	-7.092049047
32	-6.04484571	143	-9.111373745	243	-6.755972995	343	-9.470867691
33	-5.729811693	144	-7.983460215	244	-6.772832949	344	-6.408533144
34	-5.144241297	145	-6.106818151	245	-8.042506913	345	-9.461482811
38	-7.394914761	146	-7.668595215	246	-8.550617561	346	-8.619102885
39	-6.803736781	147	-7.82759925	247	-7.761354691	347	-9.432004235
40	-3.670024899	148	-8.962536385	248	-8.74310367	348	-9.619534963
41	-6.08009091	149	-7.579528959	249	-7.539686509	349	-9.271180432
42	-7.76263439	150	-2.206426469	250	-7.75256185	350	-7.410777053
43	-3.342898627	151	-7.68424035	251	-5.88368955	351	-8.459072085
44	-9.16785354	152	-7.551559928	252	-7.824101965	352	-8.555434127
45	-6.331786021	153	-7.7577854	253	-10.54986428	353	-9.946725899
46	-6.837352873	154	-2.721343578	254	-8.032616633	354	-8.949325387
47	-6.281958005	155	-7.375238947	255	-4.107201973	355	-7.817713368
48	-4.936652722	156	-9.503290893	256	-7.054650223	356	-8.25840394
49	-4.376571133	157	-7.960342723	257	-8.254609347	357	-7.34801961
50	-8.333145833	158	-9.336218533	258	-6.303229409	358	-7.354573006
51	-7.242959445	159	-8.862783681	259	-4.13530843	359	-8.043555896
52	-1.738813313	160	-7.872746111	260	-6.620987688	360	-8.42067631
53	-9.274539823	161	-4.30982221	261	-7.196157732	361	-7.81776537
54	-7.031373146	162	-7.719624298	262	-6.394821526	362	-4.472267107
55	-4.060591747	163	-7.963737548	263	-7.371889482	363	-6.479098326
56	-8.353352192	164	-4.63358023	264	-6.78915127	364	-10.11431077
57	-3.879776275	165	-6.55523921	265	-2.589117333	365	-7.254390317
58	-8.626924386	166	-9.039203388	266	-2.078410253	366	-3.666781597
59	-6.81838918	167	-8.433598069	267	-5.249400532	367	-7.321244147

Positions	Average prediction scores (epistatic)	Positions	Average prediction scores (epistatic)	Positions	Average prediction scores (epistatic)	Positions	Average prediction scores (epistatic)
60	-7.648534612	168	-2.406490846	268	-4.996912675	368	-7.273426794
61	-4.579645891	169	-5.391664749	269	-5.458097223	369	-8.167054899
62	-5.945772755	170	-6.32628314	270	-4.785142697	370	-5.90303636
63	-3.132990721	171	-7.018396734	271	-4.897393483	371	-9.636086576
64	-3.298221174	172	-5.002460016	272	-2.859029141	372	-6.154639634
65	-4.18660468	173	-6.201940169	273	-8.477250362	373	-6.924241008
66	-2.84109894	174	-6.200927627	274	-8.366317976	374	-7.213040292
67	-3.727376794	175	-6.168891602	275	-8.554054522	375	-4.280868045
68	-6.189743199	176	-5.303300745	276	-9.434634308	376	-7.458358168
69	-7.332502801	177	-1.801571154	277	-10.34473877	377	-4.682967048
70	-2.009041612	178	-4.889425826	278	-6.133581842	381	-6.0301633
71	-0.664699372	179	-4.492666067	279	-10.06726119	382	-6.36671842
72	-1.856596481	180	-3.307068049	280	-6.234069876	383	-7.325291161
73	-4.047993616	181	-4.770344617	281	-7.807397647	384	-5.975831197
81	-1.798402594	182	-2.493341044	282	-8.10102257	385	-6.614364958
82	-3.871701581	183	-2.439220507	283	-8.266604722	386	-6.31711608
83	-0.97853084	184	-3.946633076	284	-6.15379397	387	-5.394611459
84	-4.929851349	185	-6.960971319	285	-8.712001493	388	-5.072014292
85	-5.228055646	186	-5.152028924	286	-5.5002605	389	-8.287682581
86	-1.193053666	187	-3.89222428	287	-4.837166996	390	-7.089830255
87	-3.947369065	188	-3.388111054	288	-4.38705395	391	-7.178305075
88	-7.360801686	189	-8.308501011	289	-7.596360251	392	-2.712639688
89	-4.533699555	190	-4.185014827	290	-6.674213981	393	-5.621736189
90	-5.865241033	191	-2.018922193	291	-8.660110773	394	-8.289089198
91	-0.859655387	192	-8.361762103	292	-2.97624857	395	-7.168159703
92	-9.83014845	193	-7.463209754	293	-6.862344135	396	-8.980749083
93	-7.258486889	194	-4.608069554	294	-7.992735892	397	-6.001109158
94	-9.228586141	195	-3.649726685	295	-8.044525867		
95	-3.125007254	196	-7.01190726	296	-4.320147743		
96	-8.191554234	197	-6.145420709	297	-6.399881334		
97	-2.124785742	198	-3.095794969	298	-2.586051323		
98	-3.438165856	199	-3.340415161	299	-7.727636265		
99	-7.567314568	200	-6.391512393	300	-3.454918998		
100	-9.882356145	201	-5.517934563	301	-6.090780236		
101	-7.172680466	202	-1.361205383	302	-7.938245625		
102	-5.267860116	203	-2.180866066	303	-5.980307048		
103	-7.762741837	204	-4.617109062	304	-6.442342547		
104	-6.783906429	205	-5.942503424	305	-6.019913083		
105	-6.140276528	206	-4.505411461	306	-7.254936097		
106	-4.387172131	207	-6.557009764	307	-7.368308965		
107	-10.24911143	208	-4.567070395	308	-8.411790122		
108	-6.520258882	209	-7.239276307	309	-5.789706228		
109	-6.559165205	210	-8.84866301	310	-7.784319631		
110	-7.637124984	211	-4.783884101	311	-5.107006461		
111	-6.545836721	212	-7.87337003	312	-5.715480529		
112	-5.208636824	213	-5.406063033	313	-6.793286665		
113	-4.711699427	214	-4.137682331	314	-5.231177199		
114	-2.264072475	215	-2.913784472	315	-7.430673112		
115	-5.111893402	216	-4.213214982	316	-8.631841911		
116	-6.013384293	217	-4.209819086	317	-8.162782359		
117	-7.962821546	218	-2.526283638	318	-9.106749502		
118	-4.038326028	219	-1.234935617	319	-2.937376073		
119	-8.206716593	220	-7.425400168	320	-2.258858638		
120	-6.726134726	221	-0.664603939	321	-4.301646818		
121	-4.380862815	222	-3.457002457	322	-6.333775394		
122	-1.635849053	223	-5.685867232	323	-4.584177984		
123	-8.435755284	224	-1.117196872	324	-4.065580691		
124	-6.629788596	225	-2.91102705	325	-8.175305293		

5.7 References

1. Romero, P. A.; Arnold, F. H., Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **2009,** *10* (12), 866–876.
2. Huffman, M. A.; Fryszkowska, A.; Alvizo, O.; Borra-Garske, M.; Campos, K. R.; Canada, K. A.; Devine, P. N.; Duan, D.; Forstater, J. H.; Grosser, S. T., et al., Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science* **2019,** *366* (6470), 1255–1259.
3. McIntosh, J. A.; Liu, Z.; Andresen, B. M.; Marzijarani, N. S.; Moore, J. C.; Marshall, N. M.; Borra-Garske, M.; Obligacion, J. V.; Fier, P. S.; Peng, F., et al., A kinase-cGAS cascade to synthesize a therapeutic STING activator. *Nature* **2022,** *603* (7901), 439–444.

4. Zetzsche, L. E.; Yazarians, J. A.; Chakrabarty, S.; Hinze, M. E.; Murray, L. A. M.; Lukowski, A. L.; Joyce, L. A.; Narayan, A. R. H., Biocatalytic oxidative cross-coupling reactions for biaryl bond formation. *Nature* **2022,** *603* (7899), 79–85.
5. Salehi, B.; Venditti, A.; Sharifi-Rad, M.; Kregiel, D.; Sharifi-Rad, J.; Durazzo, A.; Lucarini, M.; Santini, A.; Souto, E. B.; Novellino, E., et al., The Therapeutic Potential of Apigenin. *Int. J. Mol. Sci.* **2019,** *20* (6).
6. Lee, J. S.; Sul, J. Y.; Park, J. B.; Lee, M. S.; Cha, E. Y.; Song, I. S.; Kim, J. R.; Chang, E. S., Fatty acid synthase inhibition by amentoflavone suppresses HER2/neu (erbB2) oncogene in SKBR3 human breast cancer cells. *Phytother. Res.* **2013,** *27* (5), 713–720.
7. Pawellek, A.; Ryder, U.; Tammsalu, T.; King, L. J.; Kreinin, H.; Ly, T.; Hay, R. T.; Hartley, R. C.; Lamond, A. I., Characterisation of the biflavonoid hinokiflavone as a pre-mRNA splicing modulator that inhibits SENP. *eLife* **2017,** *6*, e27402.
8. Chaves, O. A.; Lima, C. R.; Fintelman-Rodrigues, N.; Sacramento, C. Q.; de Freitas, C. S.; Vazquez, L.; Temerozo, J. R.; Rocha, M. E. N.; Dias, S. S. G.; Carels, N., et al., Agathisflavone, a natural biflavonoid that inhibits SARS-CoV-2 replication by targeting its proteases. *Int. J. Biol. Macromol.* **2022,** *222*, 1015–1026.
9. Al-Sayed, E.; Gad, H. A.; El-Shazly, M.; Abdel-Daim, M. M.; Nasser Singab, A., Anti-inflammatory and analgesic activities of cupressuflavone from Cupressus macrocarpa: Impact on pro-inflammatory mediators. *Drug Dev. Res.* **2018,** *79* (1), 22–28.
10. Pokharel, Y. R.; Yang, J. W.; Kim, J. Y.; Oh, H. W.; Jeong, H. G.; Woo, E. R.; Kang, K. W., Potent inhibition of the inductions of inducible nitric oxide synthase and cyclooxygenase-2 by taiwaniaflavone. *Nitric Oxide* **2006,** *15* (3), 217–225.
11. Rahman, M.; Riaz, M.; Desai, U. R., Synthesis of biologically relevant biflavanoids--a review. *Chem. Biodivers.* **2007,** *4* (11), 2495–2527.
12. Arnold, F. H.; Georgiou, G., *Directed evolution library creation: methods and protocols*. Humana Press: Totowa, N.J., 2003; Vol. Totowa, N.J.
13. Byrappa, S.; Gavin, D. K.; Gupta, K. C., A highly efficient procedure for site-specific mutagenesis of full-length plasmids using Vent DNA polymerase. *Genome Res.* **1995,** *5* (4), 404–407.
14. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B. C.; Remm, M.; Rozen, S. G., Primer3–new capabilities and interfaces. *Nucleic Acids Res.* **2012,** *40* (15), e115.
15. NEB Tm Calculator. https://tmcalculator.neb.com/#!/main (accessed 2024-09-29).
16. Anderson, M. J., *DOE Simplified: Practical Tools for Effective Experimentation*. 3rd ed.; CRC Press: 2017.
17. Lendrem, D. W.; Lendrem, B. C.; Woods, D.; Rowland-Jones, R.; Burke, M.; Chatfield, M.; Isaacs, J. D.; Owen, M. R., Lost in space: design of experiments and scientific exploration in a Hogarth Universe. *Drug Discov. Today* **2015,** *20* (11), 1365–1371.
18. Hopf, T. A.; Ingraham, J. B.; Poelwijk, F. J.; Scharfe, C. P.; Springer, M.; Sander, C.; Marks, D. S., Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **2017,** *35* (2), 128–135.
19. Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A., Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* **2021**, 2021.2007.2009.450648.

20. Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L., NCBI BLAST: a better web interface. *Nucleic Acids Res.* **2008,** *36* (Web Server issue), W5–9.
21. Katoh, K.; Standley, D. M., MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013,** *30* (4), 772–780.
22. SantaLucia, J., Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. U. S. A.* **1998,** *95* (4), 1460–1465.
23. Schildkraut, C., Dependence of the melting temperature of DNA on salt concentration. *Biopolymers* **1965,** *3* (2), 195–208.
24. Owczarzy, R.; You, Y.; Moreira, B. G.; Manthey, J. A.; Huang, L.; Behlke, M. A.; Walder, J. A., Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures. *Biochemistry* **2004,** *43* (12), 3537–3554.
25. Hekkelman, M. L.; de Vries, I.; Joosten, R. P.; Perrakis, A., AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **2023,** *20* (2), 205–213.
26. Roberts, G. A.; Grogan, G.; Greter, A.; Flitsch, S. L.; Turner, N. J., Identification of a new class of cytochrome P450 from a Rhodococcus sp. *J. Bacteriol.* **2002,** *184* (14), 3898–3908.
27. Li, S.; Podust, L. M.; Sherman, D. H., Engineering and analysis of a self-sufficient biosynthetic cytochrome P450 PikC fused to the RhFRED reductase domain. *J. Am. Chem. Soc.* **2007,** *129* (43), 12940–12941.
28. Midlik, A.; Navratilova, V.; Moturu, T. R.; Koca, J.; Svobodova, R.; Berka, K., Uncovering of cytochrome P450 anatomy by SecStrAnnotator. *Sci. Rep.* **2021,** *11* (1), 12345.
29. Poulos, T. L., Cytochrome P450 flexibility. *Proc. Natl. Acad. Sci. U. S. A.* **2003,** *100* (23), 13121–13122.
30. Stout, C. D., Cytochrome p450 conformational diversity. *Structure* **2004,** *12* (11), 1921–1922.
31. Williams, P. A.; Cosme, J.; Vinkovic, D. M.; Ward, A.; Angove, H. C.; Day, P. J.; Vonrhein, C.; Tickle, I. J.; Jhoti, H., Crystal structures of human cytochrome P450 3A4 bound to metyrapone and progesterone. *Science* **2004,** *305* (5684), 683–686.
32. Xu, L. H.; Fushinobu, S.; Ikeda, H.; Wakagi, T.; Shoun, H., Crystal structures of cytochrome P450 105P1 from Streptomyces avermitilis: conformational flexibility and histidine ligation state. *J. Bacteriol.* **2009,** *191* (4), 1211–1219.
33. McLean, K. J.; Hans, M.; Meijrink, B.; van Scheppingen, W. B.; Vollebregt, A.; Tee, K. L.; van der Laan, J. M.; Leys, D.; Munro, A. W.; van den Berg, M. A., Single-step fermentative production of the cholesterol-lowering drug pravastatin via reprogramming of Penicillium chrysogenum. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (9), 2847–2852.
34. Wittmann, B. J.; Johnston, K. E.; Almhjell, P. J.; Arnold, F. H., evSeq: Cost-Effective Amplicon Sequencing of Every Variant in a Protein Library. *ACS Synth. Biol.* **2022,** *11* (3), 1313–1324.
35. Kille, S.; Zilly, F. E.; Acevedo, J. P.; Reetz, M. T., Regio- and stereoselectivity of P450-catalysed hydroxylation of steroids controlled by laboratory evolution. *Nat. Chem.* **2011,** *3* (9), 738–743.

36. Hayashi, K.; Sugimoto, H.; Shinkyo, R.; Yamada, M.; Ikeda, S.; Ikushiro, S.; Kamakura, M.; Shiro, Y.; Sakaki, T., Structure-Based Design of a Highly Active Vitamin D Hydroxylase from Streptomyces griseolus CYP105A1. *Biochemistry* **2008,** *47* (46), 11964–11972.
37. Kille, S.; Acevedo-Rocha, C. G.; Parra, L. P.; Zhang, Z.-G.; Opperman, D. J.; Reetz, M. T.; Acevedo, J. P., Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis. *ACS Synth. Biol.* **2013,** *2* (2), 83–92.

Chapter 6 Conclusions and Future Directions

6.1 Conclusions

This thesis presents strategies to address key challenges in biocatalyst development through the integration of bioinformatics, machine learning, and ancestral sequence reconstruction (ASR). By combining these approaches, we have demonstrated their value in enzyme identification, mechanistic study, and engineering.

In Chapter 2, we explored ASR as a robust tool for enzyme mechanistic studies. The findings underscore the limitations of traditional mutagenesis on extant enzymes, where stability issues and epistasis can obscure critical substitutions involved in selective control. Evolutionarily reconstructed enzymes, however, provide a unique view of nature's strategies for achieving chemical selectivity across different circumstances and selective pressure. By analyzing ancestral enzymes that represent key evolutionary stages, we demonstrated how ASR can overcome the constraints of stability and epistasis, revealing insights into stereo-control mechanisms that would otherwise be challenging to decipher. ASR has broad applicability beyond stereoselectivity, as it can also be applied to investigate other selectivities, such as chemo-, site-, or atroposelectivity, as well as different transformations. By profiling and resurrecting ancestral enzymes, researchers can better understand how enzymes have evolved for specific tasks, gaining insights into nature’s design principles. This chapter suggests that ASR, when combined with high-throughput screening and phylogenetic analysis, is a powerful approach for decoding enzyme function across diverse biochemical contexts.

The high stability and solubility of ancestral enzymes highlighted in Chapter 2 inspired their use in enzyme engineering, as demonstrated in Chapter 3. Here, we applied ASR to develop ancestral enzymes for a stereocomplementary biocatalytic synthesis of azaphilones, a transformation not previously reported. Ancestral enzymes offer advantages in stability, solubility, reactivity, and evolvability, making them ideal starting points for future enzyme engineering. To focus our efforts on enzymes likely to perform

the desired transformation, we used the evolutionary relationship between co-located genes in biosynthetic gene clusters (BGCs). Recognizing that genes within the same BGC often co-evolve to produce complex natural products, we leveraged this co-evolution to predict substrate specificity and selectivity. This allowed us to narrow the ancestral sequence space and resurrect only seven ancestral acyltransferases (ATs) that showed high potential for our target acylation reaction. This strategy enabled us to bypass extensive enzyme engineering by directed evolution, as we quickly identified an ideal AT with four-fold reactivity for practical applications. Our approach not only accelerates enzyme discovery but also demonstrates a generalizable method for developing enzyme cascades for biocatalytic synthesis. Given the prevalence of BGCs across a wide range of natural products with potential medical applications, this strategy could be widely adopted in future studies. Furthermore, the general stability and solubility of ancestral enzymes make them particularly suitable for further engineering, reinforcing ASR as a foundational tool in biocatalyst development.

In Chapter 3, we developed variational autoencoder (VAE)-based latent space models as an alternative to traditional tools like sequence similarity networks (SSNs) and phylogenetic analysis to guide enzyme sampling and discovery. The primary advantages of the two-dimensional latent space models are their ability to handle large enzyme datasets, capture both local and global relationships within enzyme families, and provide a continuous space where functional relationships can be inferred based on the Euclidean distances between sequence representations. We trained, optimized, and validated the models using three datasets: a simulated protein sequence dataset, a lanthipeptide cyclase dataset, and a flavin-dependent monooxygenase (FDMO) dataset. With the simulated dataset, we observed that the latent space models trained with deeper encoder and decoder layers could better distinguish protein clusters, allowing for more precise enzyme sampling and navigation within the sequence space. The cyclase dataset, which included peptide substrate information, enabled a comprehensive evaluation of the latent space models using real data. We tested the VAE-based latent space representations in two, three, five, and ten dimensions as input features to train various classifiers, assessing whether the models generated meaningful embeddings. We compared these VAE-derived embeddings to those from popular protein language models, ESM-1b and

ProtT5-XL, finding that our models performed equivalently or better, likely due to weak positive labeling achieved through multiple sequence alignment. Finally, we applied the latent space model trained on the FDMO dataset to explore the continuity of the space, leading to the discovery of multiple novel FDMOs involved in decarboxylative hydroxylation. Altogether, these examples demonstrate the utility of VAE-based latent space models in enzyme sampling and discovery.

In Chapter 4, we focused on streamlining enzyme engineering through directed evolution, a critical stage in enzyme development. Recognizing the need for an efficient approach to design large-scale primer pairs required for site-saturation mutagenesis, we developed an automated primer generation pipeline on Google Colab. This tool significantly reduces the time and effort required for manually identifying mutation sites and designing primer pairs. To validate the pipeline, we applied it to evolve a cytochrome P450 for a selective apigenin dimerization reaction. Through several rounds of directed evolution, we improved the conversion rate of the P450-catalyzed dimerization from under 5% to approximately 40%. Additionally, we incorporated ESM-1b and EVcouplings into our engineering workflow to assess the feasibility of computationally predicting mutational effects. By sequencing every variant in a specific hot region of the P450, we compared our screening data with the computational predictions. While the correlation was not strong, we found that averaging the computational predictions of mutational effects at each position could help identify hot mutational sites. This approach allowed us to pinpoint the top 20 mutational positions, many of which aligned with beneficial mutations identified in our directed evolution experiments and are located at the regions reported in many previous P450 engineering studies. In the future, we anticipate that this approach can be incorporated into our automated primer generation pipeline, enabling scientists to expand mutational studies beyond residues close to active sites and explore distant positions that may have indirect or epistatic effects on enzyme function.

6.2 Future Directions

This thesis has demonstrated that ancestral enzymes hold significant promise for biocatalysis, largely due to their general stability, solubility, and promiscuity, which make them more amenable to engineering.[1-2] These properties have proven advantageous in previous studies.[3-5] Recognizing the importance of robustness of enzymes, researchers

have already begun incorporating similar strategies into their workflows, specifically by using enzymes from thermophilic organisms that exhibit higher stability for transformations.[6-9] Looking forward, there is great potential to expand this approach through the development of more accessible and user-friendly tools for ASR.[5] Such tools would enable large-scale ancestral enzyme sampling, allowing researchers to systematically screen more robust enzymes for desirable properties. This would open up new possibilities for identifying and harnessing stable, soluble, and promiscuous enzymes across a broader range of reactions, further accelerating biocatalyst discovery and optimization.

A second promising avenue for future research lies in enhancing the sequence-function relationship models we have developed by integrating structural information, which is becoming increasingly accessible owing to breakthroughs in predictive modeling tools such as AlphaFold,[10] RoseTTAFold,[11] and ESMFold.[12] Our current strategies have primarily relied on sequence-based approaches to infer functional relationships. However, as these structure prediction tools continue to improve in accuracy and availability, it will become feasible to incorporate high-throughput generation of enzyme structural models into biocatalyst development workflows.[13] Access to reliable structural information at scale could significantly enhance our ability to design and predict enzyme functions, identify active sites, and locate potential allosteric sites, ultimately leading to a more detailed understanding of enzyme mechanisms.[14]

By integrating structural data with sequence-based methods, future studies could develop more comprehensive models that capture both sequence and structural determinants of enzyme function. This combined approach would enable researchers to perform high-throughput computational screening, identifying initial enzyme hits with enhanced precision. Molecular modeling techniques, powered by structural predictions, could simulate enzyme-substrate interactions, predict binding affinities, and assess the effects of mutations on enzyme stability and activity, all of which would streamline the discovery of promising biocatalysts before proceeding to experimental validation.[14]

In summary, the future of biocatalyst development will likely benefit from the dual advancements of large-scale ancestral enzyme sampling and the integration of structural data into sequence-function modeling. Together, these developments have the potential

to transform the field by making enzyme engineering more efficient, predictable, and accessible, ultimately leading to the creation of more robust and specialized biocatalysts for industrial, pharmaceutical, and environmental applications.

6.3 References

1. Aharoni, A.; Gaidukov, L.; Khersonsky, O.; Mc, Q. G. S.; Roodveldt, C.; Tawfik, D. S., The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **2005,** *37* (1), 73–76.
2. Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H., Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **2006,** *103* (15), 5869–5874.
3. Chiang, C. H.; Wymore, T.; Rodríguez Benítez, A.; Hussain, A.; Smith, J. L.; Brooks III, C. L.; Narayan, A. R. H., Deciphering the evolution of flavin-dependent monooxygenase stereoselectivity using ancestral sequence reconstruction. *Proc. Natl. Acad. Sci. U. S. A.* **2023,** *120* (15), e2218248120.
4. Chiang, C. H.; Wang, Y.; Hussain, A.; Brooks III, C. L.; Narayan, A. R. H., Ancestral Sequence Reconstruction to Enable Biocatalytic Synthesis of Azaphilones. *J. Am. Chem. Soc.* **2024,** *146* (44), 30194–30203.
5. Prakinee, K.; Phaisan, S.; Kongjaroon, S.; Chaiyen, P., Ancestral Sequence Reconstruction for Designing Biocatalysts and Investigating their Functional Mechanisms. *JACS Au* **2024**.
6. Buller, A. R.; Brinkmann-Chen, S.; Romney, D. K.; Herger, M.; Murciano-Calles, J.; Arnold, F. H., Directed evolution of the tryptophan synthase beta-subunit for stand-alone function recapitulates allosteric activation. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (47), 14599–14604.
7. Sosa, Y.; Kapur, B.; Hurtak, J.; Kingsley, L. J.; Wu, H.; Gruber, S.; Nar, H.; Khattabi, S.; Moral, J. S.; Lucas, M. F., et al., In silico enzyme screening identifies an SDR ketoreductase from Thermus caliditerrae as an attractive biocatalyst and promising candidate for protein engineering. *Front. Chem. Biol.* **2024,** *3*.
8. Wang, T. C.; Mai, B. K.; Zhang, Z.; Bo, Z.; Li, J.; Liu, P.; Yang, Y., Stereoselective amino acid synthesis by photobiocatalytic oxidative coupling. *Nature* **2024,** *629* (8010), 98–104.
9. Bruffy, S. K.; Meza, A.; Soler, J.; Doyon, T. J.; Young, S. H.; Lim, J.; Huseth, K. G.; Willoughby, P. H.; Garcia-Borràs, M.; Buller, A. R., Biocatalytic asymmetric aldol addition into unactivated ketones. *Nat. Chem.* **2024**.
10. Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J., et al., Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024,** *630* (8016), 493–500.
11. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D., et al., Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021,** *373* (6557), 871–876.
12. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y., et al., Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023,** *379* (6637), 1123–1130.
13. Hussain, A.; Brooks Iii, C. L., Guiding discovery of protein sequence-structure-function modeling. *Bioinformatics* **2024,** *40* (1).

14. Kouba, P.; Kohout, P.; Haddadi, F.; Bushuiev, A.; Samusevich, R.; Sedlar, J.; Damborsky, J.; Pluskal, T.; Sivic, J.; Mazurenko, S., Machine Learning-Guided Protein Engineering. *ACS Catal* **2023,** *13* (21), 13863–13895.

Appendix

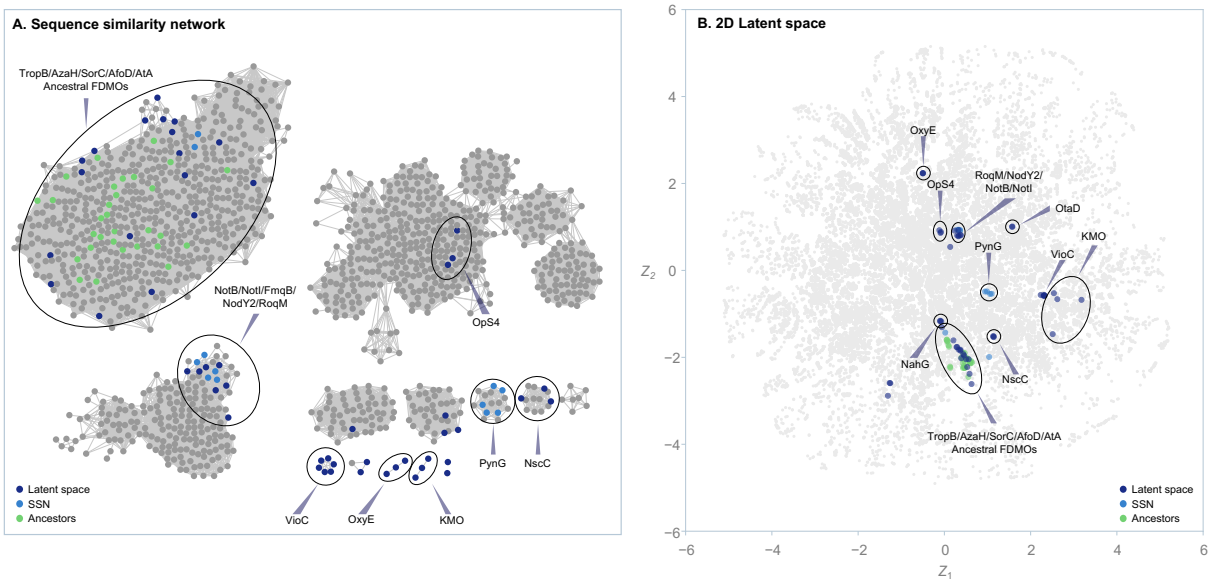
Appendix A: Curation of a Flavin-Dependent Monooxygenase Family

A.1 Results and Discussion

The goal of curating a flavin-dependent monooxygenase (FDMO) library here was to identify enzymes with more diverse transformations in addition to oxidative dearomatization well studied in our group.

Enzymes in the library primarily came from three sources, including ancestral sequences, sequence similarity network (SSN) sampling, and latent space sampling (Figure A.1). The first portion of the library includes 30 ancestral FDMOs reconstructed from a phylogenetic tree of TropB, AfoD, and AzaH and other 274 extant fungal FDMOs. 12 enzymes were directly sampled from a SSN constructed from approximately 3500 fungal sequences from the Ascomycota phylum. 45 sequences in the library were sampled from the variational autoencoder (VAE)-based two-dimensional (2D) latent space model of PF01494 in Chapter 4. Lastly, five biocatalysts in the library are artificial enzymes created using the generative aspects of VAE to take a point in latent space and decode it to a sequence, and three enzymes are from other protein families. We also included a negative control, which is an empty vector control (pMCSG7 vector without an FDMO gene). All enzymes and the control were housed separately in a 96-well plate. For this library, our strategy for enzyme sampling was to include some characterized FDMOs with their connected uncharacterized sequence nodes from the same clusters in the SSN or their close uncharacterized neighbors in the 2D latent space. Following this strategy, we were able to explore the enzyme sequence space with some guidance from previous studies.

Portions of this appendix were adapted from Champagne, S. E.; Chiang, C.-H.; Gemmel, P. M.; Brooks III, C. L.; Narayan, A. R. H. Biocatalytic Stereoselective Oxidation of 2-Arylindoles. *J. Am. Chem. Soc.* **2024,** *146* (4)*,* 2728–2735. Copyright 2024 American Chemical Society.



Appendix Figure A.1. Sequences selected from the SSN (blue), latent space (purple), and computationally reconstructed ancestral sequences (green).
A. SSN of FDMOs in PF01494. **B.** Latent space of FDMOs in PF01494.

In the 96-well plate shown in Table A.1, ancestral sequences are housed in the row 1–3, and row 4 (A–F), sequences related to TropB[1-2] or AtA[3] are in the row 4 (G–H), row 5, and row 6 (A–B), artificial sequences generated from decoding are in the row 6 (C–G), sequences related to OtaD[4] are in the row 6 (H) and row 7 (A), sequences related to OpS4[5] or NahG[6-7] are from the row 7 (B–G), sequences potentially related to kynurenine 3-monooxygenases[8] are in the row 7 (H) and row 8 (A–C), sequences related to VioC[9] are in the row 8 (E–H) and row 9 (A–B), sequences related to FmqB[10] are in the row 9 (C–E), sequences related to RoqM[11-12] are in the row 9 (F–H) and row 10 (A–C), sequences related to PynG[13] are in the row 10 (D–H), sequences related to ubiquinone biosynthesis monooxygenase Coq6 are in the row 11 (A–C), sequences related to OxyE[14] are in the row 11 (D–F), NotB[15] and NotI[16] are in G11 and H11, sequences related to NodY2[17] are in A12 and B12, sequences related to NscC[18] are in the row 12 (C–E), sequences in D8, F12–G12 are from other protein families, and H12 is the negative control.

The enzymes in the library were provide by the Joint Genome Institute (JGI) as glycerol stocks of *E. coli* TOP10 in plates. The plasmid DNAs were isolated and used to transform chemically competent *E. coli* BL21(DE3). The enzymes were overproduced by *E. coli* BL21(DE3) in plates and were used for downstream screening or analysis.

Appendix Table A.1. Curated libarary of flavin-dependent monooxygenases.

	1	2	3	4	5	6
A	Anc321_v1	Anc322_v1	Anc373_v1	Anc416_v1	A0A0L0NHZ5	I1S0U1
B	Anc333_v1	Anc322a_v1	Anc373a_v1	Anc416a_v1	A7UX11	A0A0F7ZML0
C	Anc369a_v1	Anc366_v1	Anc374_v1	Anc457_v1	A0A1W2TF28	cP23262
D	Anc382_v1	Anc366a_v1	Anc374a_v1	Anc457a_v1	A0A2J6S3A7	cQ5BEJ7
E	Anc331_v1	Anc367_v1	Anc280_v1	Anc488_v1	A0A2J6T8P1	cG3XMC2
F	Anc331a_v1	Anc367a_v1	Anc280a_v1	Anc488a_v1	A0A3D8QGL3	cB6HN76
G	Anc332_v1	Anc368_v1	Anc412_v1	A0A101MPG3	E3RKN0	cQ0CJ62
H	Anc332a_v1	Anc368a_v1	Anc412a_v1	A0A4Q7JZ50	Q0CJ62	A0A1R3RGJ2
	7	**8**	**9**	**10**	**11**	**12**
A	A0A317WVG8	Q1DDU6	A0A430HSE9	A0A0F8U263	B6K5M4	A0A2I6PIZ8
B	J4VWM7	A8LVF4	A0A0S4NSV0	A0A0C9V1F0	S9VNB0	A0A1D8G3K4
C	A0A559M954	Q2GQG8	Q4WLW7	J3KA16	Q9Y7Z9	C5FM59
D	A0A177BZG4	G3XQJ1	A0A0K8LS04	A5ABG5	Q3S8R0	E4V2N4
E	P23262	Q9S3U9	K9G8U3	V5FKW7	A0A370B0Q9	F2S701
F	Q53552	A0A2R2IYY2	B6HJU4	R7YWS5	A0A178XEJ6	KPJ13581.1
G	A0A239M104	A0A4R7BCP1	S8AW72	S3CM43	E1ACP7	Q5ATK1
H	Q84HF5	C7EDM6	A0A135LJG1	A0A0B4HMN8	E1ACQ4	Empty_pMCSG7

A.2 Experimental Procedures

Generation and preparation of stock library. Glycerol stock library preparation All genes in the library were provided by the Joint Genome Institute (JGI), cloned into a pMCSG7 vector, and delivered as glycerol stocks of *E. coli* strain TOP10 in 96-well plates. In a sterile 96-well plate, each well received 400 µL of LB media with 100 µg/mL ampicillin, and was inoculated with a glycerol stock. The plates were then incubated overnight at 37 °C, 85% humidity, and 350 rpm in a Kuhner plate incubator. The overnight *E. coli* cultures were then pelleted, and the supernatant was removed. Plasmids were isolated using a QIAprep Spin Miniprep Kit from Qiagen and an EconoSpin™ 96-well filter plate (Epoch Life Science) was used for batch elution of plasmid DNA. The eluted pMCSG7 plasmids were then transformed into chemically competent *E. coli* strain BL21(DE3) cells, prepared in a 96-well plate format. After a 1-hour incubation of the transformed *E. coli* cells in LB media without ampicillin, 50 µL of the cell/media mixture was transferred into LB media containing 100 µg/mL ampicillin, and incubated overnight at 37 °C, 85% humidity, and 350 rpm in the plate incubator. The overnight cultures were used to prepare glycerol stocks of *E. coli* strain BL21(DE3) by mixing 50 µL of the culture with 50 µL of a 50% (v/v) glycerol solution. The glycerol stocks were then stored at −80 °C for future use.

Generation of library plates. Enzyme library plates were generated from overnight cultures. To generate the overnight culture, 400 µL of sterilized LB media with 100 µg/mL

ampicillin antibiotic was added to sterilized 96-well plates. Then, scrapings of glycerol stocks were used to inoculate each well. The plate was covered with a breathable film and incubated at 37 °C with 350 rpm overnight. The next day 400 µL of TB media containing 100 µg/mL ampicillin antibiotic were added to 96-well plates. Each well was inoculated with 50 µL of overnight culture. The cultures were then incubated at 37 °C with 350 rpm for 4 h. The plates and incubator were cooled to 16 °C. Expression was induced by the addition of 10 µL of 4.6 mM isopropyl-b-D-1-thiogalactopyranoside solution (IPTG) (final concentration 0.1 mM). Cultures were incubated overnight (~18 h) at 16 °C and 350 rpm. The next day, the pellets were collected by centrifuging at 2000 rpm for 10 min. Then, the supernatant was discarded. Finally, the plates were covered with a sticky foil covers and placed in the −80 °C freezer.

Lysis of library plates. The library plates were lysed by first removing the plates from the −80 °C freezer. The plates were thawed for 20 min in room temperature water. The bottom of the plates were dried with paper towels between thawing and freezing to prevent ice buildup. Then, the plates were frozen in liquid nitrogen for 5 min. This thaw/freeze process was repeated two more times. After a last thawing step, the pellets were resuspended with 400 µL of 50 mM KPi buffer pH 8.0 using a multichannel pipette to pipette up and down until each well contained homogenous crude cell lysate.

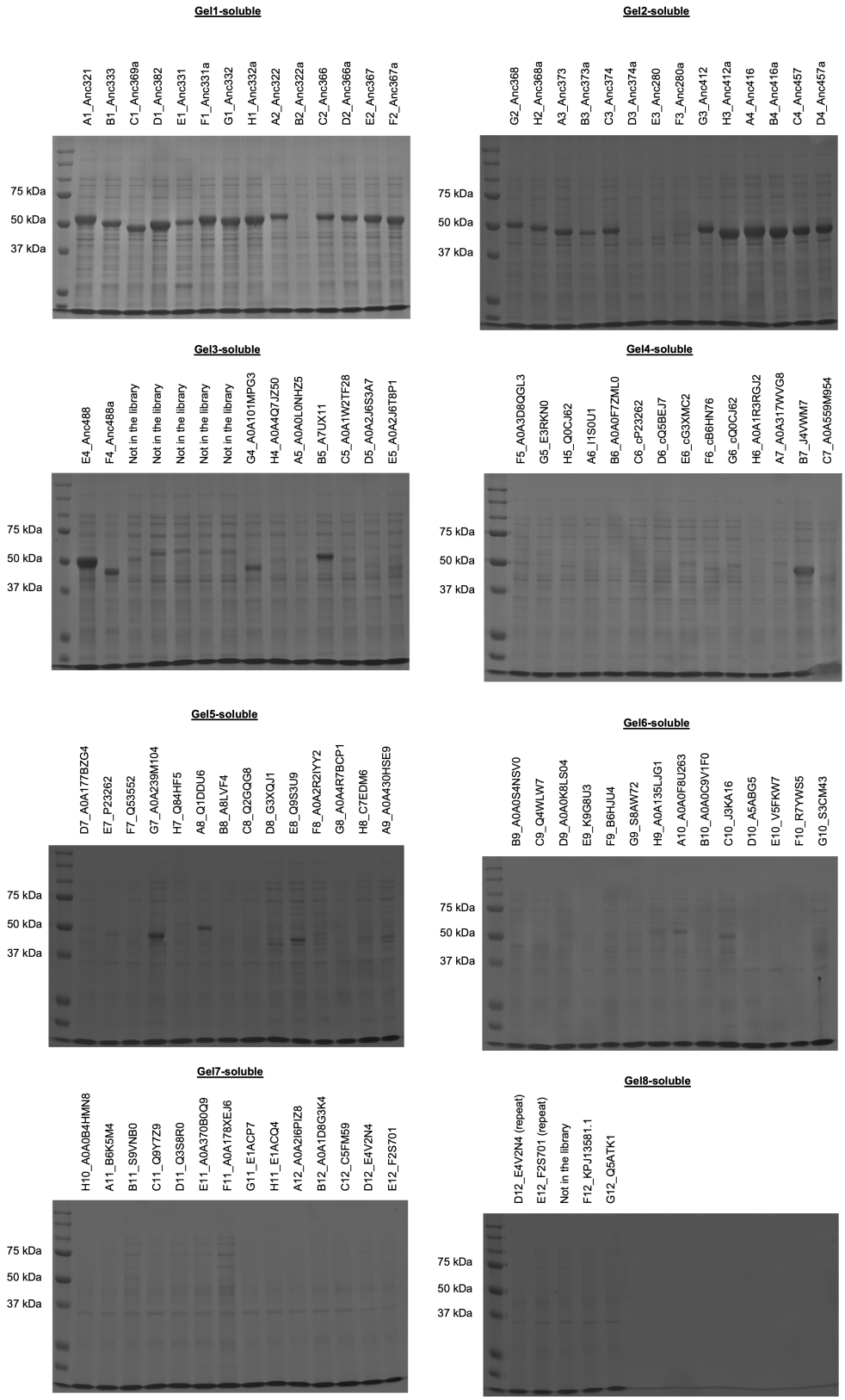
A.3 Supporting Information

Appendix Table A.2. Summary of the curated FDMO library.

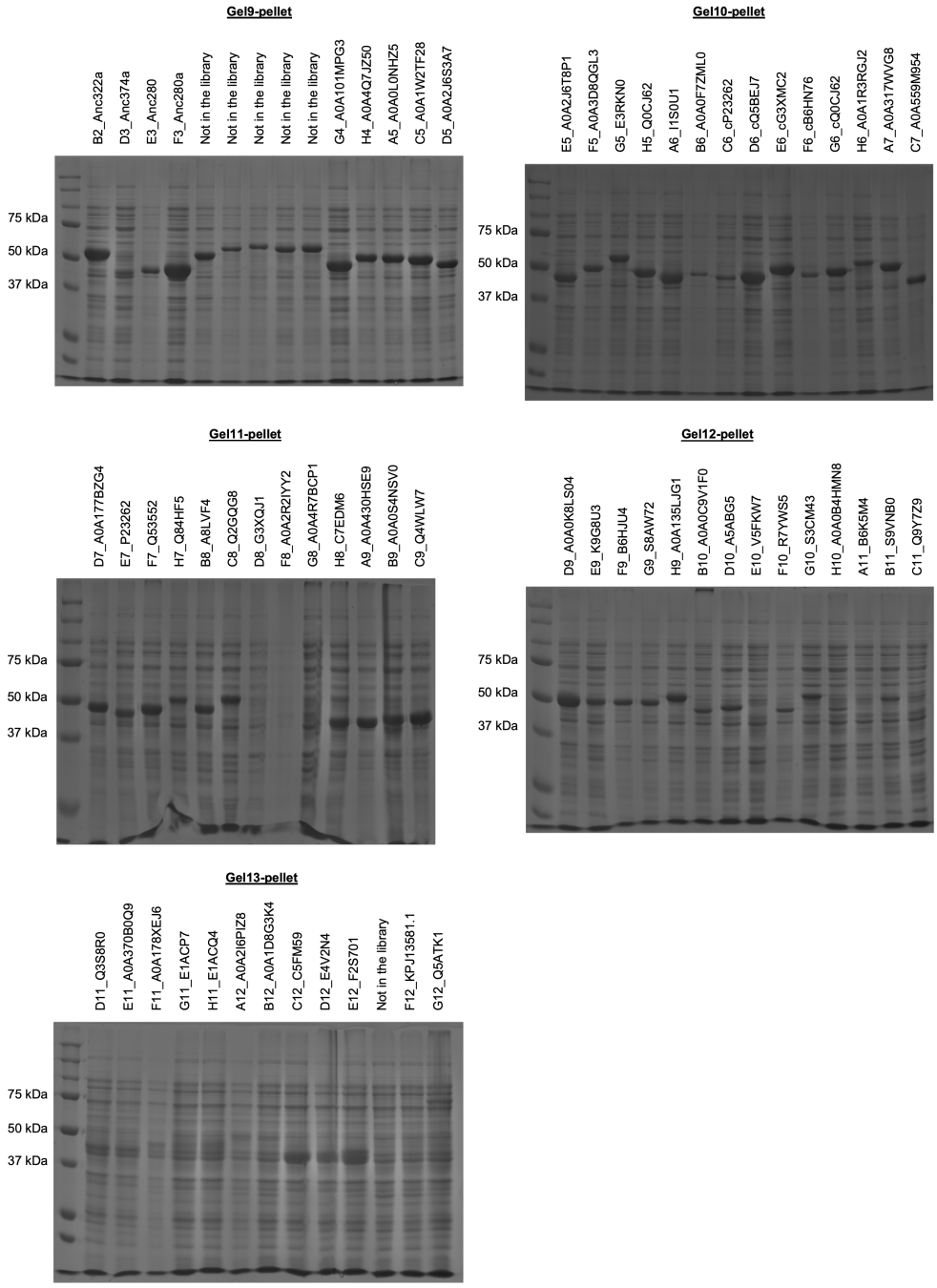
Well	Protein ID	Source	Within the latent space or not	Z_1, Z_2
A1	Anc321_v1	ancestors	TRUE	0.42785758, -2.1298318
B1	Anc333_v1	ancestors	TRUE	0.55580246, -2.4514384
C1	Anc369a_v1	ancestors	TRUE	0.62330353, -2.0881517
D1	Anc382_v1	ancestors	TRUE	0.63912016, -2.1231296
E1	Anc331_v1	ancestors	TRUE	0.47844255, -2.2377806
F1	Anc331a_v1	ancestors	TRUE	0.43238145, -2.2448068
G1	Anc332_v1	ancestors	TRUE	0.47282797, -2.2418659
H1	Anc332a_v1	ancestors	TRUE	0.44973582, -2.2017217
A2	Anc322_v1	ancestors	TRUE	0.42835933, -2.0317755
B2	Anc322a_v1	ancestors	TRUE	0.46082568, -2.2428555
C2	Anc366_v1	ancestors	TRUE	0.3934651, -2.007549
D2	Anc366a_v1	ancestors	TRUE	0.48227692, -2.117263
E2	Anc367_v1	ancestors	TRUE	0.50392044, -2.0163705
F2	Anc367a_v1	ancestors	TRUE	0.51695365, -2.0154443
G2	Anc368_v1	ancestors	TRUE	0.54089904, -2.0383174
H2	Anc368a_v1	ancestors	TRUE	0.5683586, -2.052679
A3	Anc373_v1	ancestors	TRUE	0.59208727, -2.095837

Well	Protein ID	Source	Within the latent space or not	Z_1, Z_2
B3	Anc373a_v1	ancestors	TRUE	0.5944696, -2.1101387
C3	Anc374_v1	ancestors	TRUE	0.6008625, -2.1331348
D3	Anc374a_v1	ancestors	TRUE	0.6004483, -2.1360612
E3	Anc280_v1	ancestors	TRUE	0.433953, -1.8963025
F3	Anc280a_v1	ancestors	TRUE	0.47192103, -1.9017115
G3	Anc412_v1	ancestors	TRUE	0.06863125, -1.6325436
H3	Anc412a_v1	ancestors	TRUE	0.056804046, -1.5961931
A4	Anc416_v1	ancestors	TRUE	0.0570053, -1.5946522
B4	Anc416a_v1	ancestors	TRUE	0.07253605, -1.6301149
C4	Anc457_v1	ancestors	TRUE	0.10615361, -1.755038
D4	Anc457a_v1	ancestors	TRUE	0.093696, -1.7214332
E4	Anc488_v1	ancestors	TRUE	0.1329694, -2.2139335
F4	Anc488a_v1	ancestors	TRUE	0.13725735, -2.2406435
G4	A0A101MPG3	latent space	TRUE	0.6302328, -2.613682
H4	A0A4Q7JZ50	latent space	TRUE	0.5830799, -2.3796587
A5	A0A0L0NHZ5	latent space	TRUE	0.5288805, -2.2225268
B5	A7UX11	latent space	TRUE	0.5054158, -2.0425935
C5	A0A1W2TF28	latent space	TRUE	0.4567477, -1.9578559
D5	A0A2J6S3A7	latent space	TRUE	0.3722806, -1.8468466
E5	A0A2J6T8P1	latent space	TRUE	0.33171004, -1.8206255
F5	A0A3D8QGL3	latent space	TRUE	0.21060883, -1.6096447
G5	E3RKN0	latent space	TRUE	0.29419452, -1.7611613
H5	Q0CJ62	latent space	TRUE	0.2814979, -1.7650052
A6	I1S0U1	ssn	TRUE	0.020846575, -1.4316106
B6	A0A0F7ZML0	ssn	FALSE	nan, nan
C6	cP23262	latent space	TRUE	-0.05757896, -1.3068638
D6	cQ5BEJ7	latent space	TRUE	0.56383264, -2.0426793
E6	cG3XMC2	latent space	TRUE	0.39018196, -2.0204973
F6	cB6HN76	latent space	TRUE	0.43819672, -1.9255714
G6	cQ0CJ62	latent space	TRUE	0.37225515, -1.8242621
H6	A0A1R3RGJ2	latent space	TRUE	1.5692405, 1.0027628
A7	A0A317WVG8	latent space	TRUE	1.578102, 1.0101352
B7	J4VWM7	latent space	TRUE	-0.09063016, 0.86221987
C7	A0A559M954	latent space	TRUE	-0.07892035, 0.8847529
D7	A0A177BZG4	latent space	TRUE	-0.11671233, 0.9246275
E7	P23262	latent space	TRUE	-0.09360162, -1.1585988
F7	Q53552	latent space	TRUE	-0.07139816, -1.1740268
G7	A0A239M104	latent space	TRUE	-0.10140179, -1.1674757
H7	Q84HF5	latent space	TRUE	2.6188369, -0.6619251
A8	Q1DDU6	latent space	TRUE	3.1795316, -0.67714715
B8	A8LVF4	latent space	TRUE	2.5396464, -0.523248
C8	Q2GQG8	latent space	TRUE	2.5123463, -1.465749
D8	G3XQJ1	others	FALSE	nan, nan
E8	Q9S3U9	latent space	TRUE	2.3174944, -0.58064985
F8	A0A2R2IYY2	latent space	TRUE	2.297859, -0.5838471
G8	A0A4R7BCP1	latent space	TRUE	2.3206859, -0.5796577
H8	C7EDM6	latent space	TRUE	2.3131979, -0.56473655
A9	A0A430HSE9	latent space	TRUE	2.3149319, -0.57617575
B9	A0A0S4NSV0	latent space	TRUE	2.2330315, -0.5645902
C9	Q4WLW7	ssn	FALSE	nan, nan
D9	A0A0K8LS04	ssn	TRUE	0.33116597, 0.93554175
E9	K9G8U3	ssn	TRUE	0.29451096, 0.93318164
F9	B6HJU4	latent space	TRUE	0.32562095, 0.78432983
G9	S8AW72	latent space	TRUE	0.33167952, 0.7978469
H9	A0A135LJG1	ssn	TRUE	0.36294097, 0.9174993
A10	A0A0F8U263	ssn	TRUE	0.3644328, 0.92820364
B10	A0A0C9V1F0	latent space	TRUE	0.23494612, 0.9167208
C10	J3KA16	latent space	TRUE	0.32778883, 0.92722696

Well	Protein ID	Source	Within the latent space or not	Z_1, Z_2
D10	A5ABG5	ssn	TRUE	0.9430018, -0.48505482
E10	V5FKW7	ssn	TRUE	0.98867875, -0.48476204
F10	R7YWS5	ssn	TRUE	1.0380437, -1.9911587
G10	S3CM43	ssn	TRUE	1.0967946, -0.53608805
H10	A0A0B4HMN8	ssn	TRUE	1.0596166, -0.5414662
A11	B6K5M4	latent space	TRUE	-1.3092986, -2.8845394
B11	S9VNB0	latent space	TRUE	-1.268124, -2.5899985
C11	Q9Y7Z9	latent space	TRUE	-1.254292, -2.5952544
D11	Q3S8R0	latent space	TRUE	-0.4963346, 2.237866
E11	A0A370B0Q9	latent space	TRUE	-0.49609274, 2.2340467
F11	A0A178XEJ6	latent space	TRUE	-0.50635767, 2.2349386
G11	E1ACP7	latent space	TRUE	0.37850755, 0.8064661
H11	E1ACQ4	latent space	TRUE	0.37851292, 0.81871647
A12	A0A2I6PIZ8	latent space	TRUE	0.29558492, 0.8059675
B12	A0A1D8G3K4	latent space	TRUE	0.13491593, 0.54040366
C12	C5FM59	latent space	TRUE	1.1292031, -1.5131166
D12	E4V2N4	latent space	TRUE	1.1414809, -1.5358471
E12	F2S701	latent space	TRUE	1.1455326, -1.5185566
F12	KPJ13581.1	others	FALSE	nan, nan
G12	Q5ATK1	others	FALSE	nan, nan
H12	Empty pMCSG7	others	FALSE	nan, nan



Appendix Figure A.2. Soluble fractions of protein overproduction after cell lysis.



Appendix Figure A.3. Insoluble fractions of protein overproduction after cell lysis.

A.4 References

1. Abood, A.; Al-Fahad, A.; Scott, A.; Hosny, A. E.-D. M. S.; Hashem, A. M.; Fattah, A. M. A.; Race, P. R.; Simpson, T. J.; Cox, R. J., Kinetic characterisation of the FAD dependent monooxygenase TropB and investigation of its biotransformation potential. *RSC Adv.* **2015,** *5* (62), 49987–49995.
2. Benítez, A. R.; Tweedy, S. E.; Dockrey, S. A. B.; Lukowski, A. L.; Wymore, T.; Khare, D.; Brooks III, C. L.; Palfey, B. A.; Smith, J. L.; Narayan, A. R. H., Structural basis for

selectivity in flavin-dependent monooxygenase-catalyzed oxidative dearomatization. *ACS Catal.* **2019,** *9* (4), 3633–3640.
3. Guo, C. J.; Sun, W. W.; Bruno, K. S.; Wang, C. C., Molecular genetic characterization of terreic acid pathway in Aspergillus terreus. *Org. Lett.* **2014,** *16* (20), 5250–5253.
4. Wang, Y.; Wang, L.; Wu, F.; Liu, F.; Wang, Q.; Zhang, X.; Selvaraj, J. N.; Zhao, Y.; Xing, F.; Yin, W. B., et al., A Consensus Ochratoxin A Biosynthetic Pathway: Insights from the Genome Sequence of Aspergillus ochraceus and a Comparative Genomic Analysis. *Appl. Environ. Microbiol.* **2018,** *84* (19), e01009–01018.
5. Feng, P.; Shang, Y.; Cen, K.; Wang, C., Fungal biosynthesis of the bibenzoquinone oosporein to evade insect immunity. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (36), 11365–11370.
6. You, I. S.; Murray, R. I.; Jollie, D.; Gunsalus, I. C., Purification and characterization of salicylate hydroxylase from Pseudomonas putida PpG7. *Biochem. Biophys. Res. Commun.* **1990,** *169* (3), 1049–1054.
7. Costa, D. M. A.; Gomez, S. V.; de Araujo, S. S.; Pereira, M. S.; Alves, R. B.; Favaro, D. C.; Hengge, A. C.; Nagem, R. A. P.; Brandao, T. A. S., Catalytic mechanism for the conversion of salicylate into catechol by the flavin-dependent monooxygenase salicylate hydroxylase. *Int. J. Biol. Macromol.* **2019,** *129*, 588–600.
8. Matthijs, S.; Baysse, C.; Koedam, N.; Tehrani, K. A.; Verheyden, L.; Budzikiewicz, H.; Schafer, M.; Hoorelbeke, B.; Meyer, J. M.; De Greve, H., et al., The Pseudomonas siderophore quinolobactin is synthesized from xanthurenic acid, an intermediate of the kynurenine pathway. *Mol. Microbiol.* **2004,** *52* (2), 371–384.
9. Shinoda, K.; Hasegawa, T.; Sato, H.; Shinozaki, M.; Kuramoto, H.; Takamiya, Y.; Sato, T.; Nikaidou, N.; Watanabe, T.; Hoshino, T., Biosynthesis of violacein: a genuine intermediate, protoviolaceinic acid, produced by VioABDE, and insight into VioC function. *Chem. Commun.* **2007**, 4140–4142.
10. Lim, F. Y.; Ames, B.; Walsh, C. T.; Keller, N. P., Co-ordination between BrlA regulation and secretion of the oxidoreductase FmqD directs selective accumulation of fumiquinazoline C to conidial tissues in Aspergillus fumigatus. *Cell. Microbiol.* **2014,** *16* (8), 1267–1283.
11. Ali, H.; Ries, M. I.; Nijland, J. G.; Lankhorst, P. P.; Hankemeier, T.; Bovenberg, R. A.; Vreeken, R. J.; Driessen, A. J., A branched biosynthetic pathway is involved in production of roquefortine and related compounds in Penicillium chrysogenum. *PLoS One* **2013,** *8* (6), e65328.
12. Ries, M. I.; Ali, H.; Lankhorst, P. P.; Hankemeier, T.; Bovenberg, R. A.; Driessen, A. J.; Vreeken, R. J., Novel key metabolites reveal further branching of the roquefortine/meleagrin biosynthetic pathway. *J. Biol. Chem.* **2013,** *288* (52), 37289–37295.
13. Yamamoto, T.; Tsunematsu, Y.; Noguchi, H.; Hotta, K.; Watanabe, K., Elucidation of Pyranonigrin Biosynthetic Pathway Reveals a Mode of Tetramic Acid, Fused gamma-Pyrone, and exo-Methylene Formation. *Org. Lett.* **2015,** *17* (20), 4992–4995.
14. Zhang, W.; Ames, B. D.; Tsai, S. C.; Tang, Y., Engineered biosynthesis of a novel amidated polyketide, using the malonamyl-specific initiation module from the oxytetracycline polyketide synthase. *Appl. Environ. Microbiol.* **2006,** *72* (4), 2573–2580.

15. Li, S.; Finefield, J. M.; Sunderhaus, J. D.; McAfoos, T. J.; Williams, R. M.; Sherman, D. H., Biochemical characterization of NotB as an FAD-dependent oxidase in the biosynthesis of notoamide indole alkaloids. *J. Am. Chem. Soc.* **2012,** *134* (2), 788–791.
16. Fraley, A. E.; Tran, H. T.; Kelly, S. P.; Newmister, S. A.; Tripathi, A.; Kato, H.; Tsukamoto, S.; Du, L.; Li, S.; Williams, R. M., et al., Flavin-Dependent Monooxygenases NotI and NotI' Mediate Spiro-Oxindole Formation in Biosynthesis of the Notoamides. *ChemBioChem* **2020,** *21* (17), 2449–2454.
17. Van de Bittner, K. C.; Nicholson, M. J.; Bustamante, L. Y.; Kessans, S. A.; Ram, A.; van Dolleweerd, C. J.; Scott, B.; Parker, E. J., Heterologous Biosynthesis of Nodulisporic Acid F. *J. Am. Chem. Soc.* **2018,** *140* (2), 582–585.
18. Chooi, Y. H.; Fang, J.; Liu, H.; Filler, S. G.; Wang, P.; Tang, Y., Genome mining of a prenylated and immunosuppressive polyketide from pathogenic fungi. *Org. Lett.* **2013,** *15* (4), 780–783.