Zitierweise: *Angew. Chem. Int. Ed.* **2024**, e202314869
doi.org/10.1002/anie.202314869

**Enzyme Catalysis** **Hot Paper**

# Engineering Biocatalysts for the C–H Activation of Fatty Acids by Ancestral Sequence Reconstruction**

*Bethan S. Jones, Connie M. Ross, Gabriel Foley, Nadiia Pozhydaieva, Joseph W. Sharratt, Nico Kress, Lisa S. Seibt, Raine E. S. Thomson, Yosephine Gumulya, Martin A. Hayes, Elizabeth M. J. Gillam,* and Sabine L. Flitsch**

**Abstract:** Selective, one-step C–H activation of fatty acids from biomass is an attractive concept in sustainable chemistry. Biocatalysis has shown promise for generating high-value hydroxy acids, but to date enzyme discovery has relied on laborious screening and produced limited hits, which predominantly oxidise the subterminal positions of fatty acids. Herein we show that ancestral sequence reconstruction (ASR) is an effective tool to explore the sequence-activity landscape of a family of multidomain, self-sufficient P450 monooxygenases. We resurrected 11 catalytically active CYP116B ancestors, each with a unique regioselectivity fingerprint that varied from subterminal in the older ancestors to mid-chain in the lineage leading to the extant, P450-TT. In lineages leading to extant enzymes in thermophiles, thermostability increased from ancestral to extant forms, as expected if thermophily had arisen de novo. Our studies show that ASR can be applied to multidomain enzymes to develop active, self-sufficient monooxygenases as regioselective biocatalysts for fatty acid hydroxylation.

[*] Dr. B. S. Jones, N. Pozhydaieva, J. W. Sharratt, Dr. N. Kress, L. S. Seibt, Prof. S. L. Flitsch
School of Chemistry, The University of Manchester, Manchester Institute of Biotechnology (MIB)
131 Princess Street, Manchester M1 7DN (UK)
E-mail: sabine.flitsch@manchester.ac.uk

Dr. C. M. Ross, Dr. G. Foley, Dr. R. E. S. Thomson, Dr. Y. Gumulya, Prof. E. M. J. Gillam
School of Chemistry and Molecular Biosciences, The University of Queensland
St. Lucia, Brisbane, 4072 (Australia)
E-mail: e.gillam@uq.edu.au

Prof. M. A. Hayes
Compound Synthesis and Management, Discovery Sciences, R&D, AstraZeneca
Gothenburg (SE)

[**] A previous version of this manuscript has been deposited on a preprint server (https://doi.org/10.26434/chemrxiv-2023-qn9gj).



## Introduction

Selective hydroxylation of fatty acids offers many opportunities for generating a wide range of pharmaceuticals, flavours, and fragrances from the limited starting materials that are available from biomass.[1] For instance, hydroxylation at each of the 19 C–H bonds in the highly abundant starting material, decanoic acid, would provide 17 unique products that can only be accessed by kinetic control over reaction conditions.[2] The challenge is to find catalysts that can distinguish chemically similar positions so as to selectively modify each of these C–H bonds. Biocatalytic methods offer the potential for exquisite control over the position of hydroxylation, but despite considerable efforts, only a limited number of effective biocatalysts for selective fatty acid hydroxylation have been identified to date. Most of these biocatalysts are from the superfamily of cytochrome P450 monooxygenases (P450s, CYP family) and have been largely limited to terminal or subterminal hydroxylases (Figure 1).[3] Many different P450 families hydroxylate fatty acids,[3] for example, CYP2,[4] CYP4,[2] CYP52A,[5] CYP102,[6–8] CYP116,[9] CYP147G,[10] CYP152A1,[11] and CYP153A.[12] Despite the potential of these enzymes for use in organic synthesis, difficulties with recombinant, heterologous expression, and low stability have limited their implementation. In addition, the identification and development of new biocatalysts is time-consuming; intensive screening of large numbers of natural and synthetic sequences is typically needed to identify a very small number of viable biocatalysts. Efforts to identify novel P450 biocatalysts from metagenomic databases have yielded very few P450s that are capable of selective mid-chain hydroxylation,[2,3,13–15] although much of the metagenomic sequence space remains unstudied, identifying forms that have the necessary specificity and kinetic selectivity is challenging. Directed evolution and rational protein design has improved the activity and selectivity of some terminal fatty acid hydroxylases, for example, CYP102A1 (P450-BM3)[8,16,17] and altered the regioselectivity of others, such as, CYP152A (P450-BSβ),[18,19] but engineering P450s, like metagenome mining, is time-consuming, and both labour- and resource-intensive, due to the need to screen large numbers of potential biocatalysts to identify one with the desired properties.

We recently identified CYP116B46 (P450-TT) from a panel of self-sufficient thermostable wild-type, extant

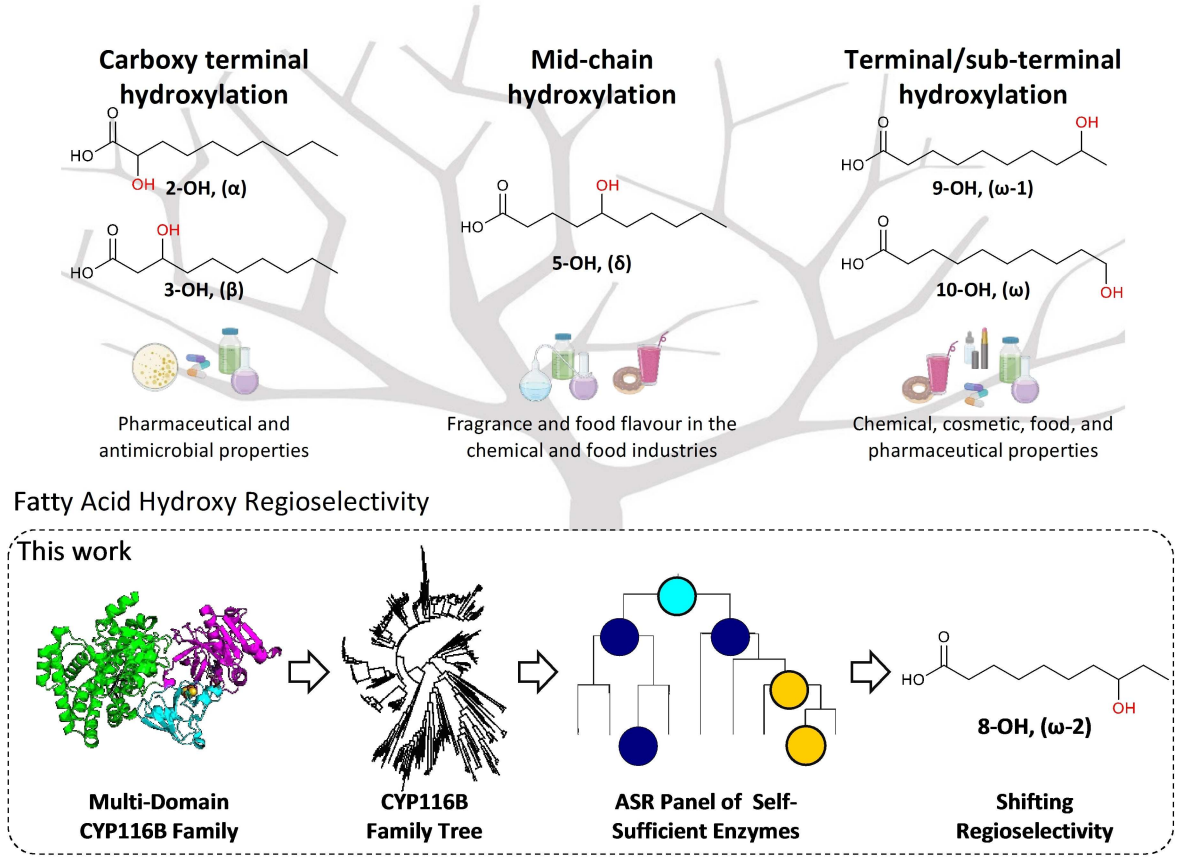



***Figure 1.*** Hydroxy acid regioselectivity classification: carboxy terminal (α, β), mid-chain (δ/ω-5), and terminal/subterminal (ω, ω-1) positions, and their properties. The previously identified self-sufficient extant CYP116B enzymes displayed activity towards fatty acids with a range of mid-chain and subterminal regioselectivities.[13] P450-TT uniquely favoured mid-chain hydroxylation of decanoic acid. Therefore, the CYP116B family was further explored through ancestral sequence reconstruction to create a panel of ancestral self-sufficient biocatalysts, where the nodes shown depict a simplified representation of the differing regioselective of decanoic acid hydroxylation from mid-chain (ω-5; orange) to different subterminal positions (ω-1, dark blue; ω-2, cyan).

CYP116B enzymes as being capable of selective mid-chain hydroxylation of decanoic acid to (*S*)-5-hydroxydecanoic acid,[13] while related enzymes from the same family (P450-AT, -TB, -AX, and -JT, Figure. 2) were either not regioselective or favoured ω-1 hydroxylation.[13] To further understand what made P450-TT selective for mid-chain hydroxylation, and to develop other P450s with similar capacities for mid-chain hydroxylation, we performed ancestral sequence reconstruction (ASR) on the CYP116B family.

ASR was employed as a tool to explore the sequence space of the CYP116B family, which enabled us to analyse a comparatively small number of potential biocatalysts with a greater chance of successfully identifying catalytically active enzymes with the desired properties of thermostability, self-sufficiency, and selective fatty-acid hydroxylation. ASR allows the generation of new proteins based upon a sequence alignment and inferred evolutionary tree that link a set of extant sequences; the most probable common ancestor and other evolutionary intermediates are inferred from the initial input.[20] The method has previously been used to identify and engineer enzymes with improved properties: greater thermal stability, enhanced expression, broader pH, and solvent tolerances.[21–34] This technique can identify novel sequences leading to functional diversity that otherwise may not have been investigated by conventional enzyme engineering and, unlike mutagenesis, most ancestors are functional.[35,36] ASR has also been used to improve the thermostability and solvent tolerance of animal P450s,[37–40] which can improve the heterologous expression and durability of biocatalysts;[41] typically, inferred ancestors have been found to be more thermostable than their extant counterparts. However, the evolution of thermostable enzymes within a largely mesophilic family had not been previously explored. As the extant CYP116B family has both mesophilic and thermophilic branches, and several previously characterised CYP116B sequences (e.g., P450-TT) originate from thermophilic (rather than mesophilic) bacteria we were interested in what evolutionary trends would emerge in the ASR of the CYP116B enzymes.

Although ASR has been successfully used for single-domain proteins, it was not clear if the approach could be used for multidomain P450s. In particular, in the CYP116B family, electron transfer from NAD(P)H to the P450 haem via a fused reductase domain and an iron-sulfur ferredoxin domain is essential for function.[42] These are self-sufficient enzymes, where the expressed proteins can be directly used

for catalysis without the need for additional redox partners.[41] Accordingly, natural and artificial self-sufficient enzymes are easier to apply in biotechnology. The proximity of the reductase and haem domains facilitates electron shuttling and enables a more effective transfer, enhancing their potential as biocatalysts.[43] It was not clear if ASR would produce functional multidomain enzymes with linker regions that would allow critical electron transfer between domains.[44–47] Since these regions are also likely to show greater variation across the extant three-domain proteins of the CYP116B family, than the haem and redox partner domains *per se*, more uncertainty was expected in the inference of the connecting sequences. Thus, we were interested if the interactions between individual domains would be retained when they were inferred together, in addition to the folding of the individual domains.

Using ASR to identify important mutational paths that lead to changes in enzymatic activity can provide insight into the mechanisms that control substrate specificity and activity,,[31–33,48–57] as demonstrated in the recent work on stereoselective C–H activation by flavin-dependent monooxygenases.[57] We were interested in how the regioselectivity towards fatty acid hydroxylation arose within the ancestral CYP116B family, especially in the pathway toward the unique mid-chain selectivity of P450-TT. In particular, we wondered whether some of these ancestors would allow us to access new regioselectivities, such as ω-2, which is not commonly observed in other biocatalysts.

We found ASR to be an effective method to generate eleven, self-sufficient, ancestral CYP116B P450 enzymes. A range of stabilities was observed, with a trend towards increasing thermostability along the branches leading to the more thermostable extant forms, suggesting that thermophilicity arose recently in the P450-TT and other extant thermophilic lineages, rather than being maintained from a highly stable ancestor. The inferred CYP116B ancestors were all catalytically active and showed unique regioselectivity profiles towards the hydroxylation of decanoic acid, including ancestors with novel selectivity for the ω-2 position. Finally, by examining the existing CYP116B structural information[42,58–60] and comparing the ancestral sequences, we were able to identify two active site residues that were partially responsible for changes to the regioselectivity of the CYP116B ancestors *vs.* the wild-type P450-TT from terminal to mid-chain hydroxylation.

## Results and Discussion

### Inference of the Phylogenetic and Ancestral Intermediates

The CYP116B family is highly diverse and CYP116B enzymes vary substantially in their thermostability, fatty acid hydroxylation regioselectivity,[13,43] and sequence identity (Tables S1 and S2). To generate an enzyme panel with diversified catalytic range, CYP116B sequences were collected by database searches and curated. We obtained 467 CYP116B sequences that included all three domains (P450/haem, iron-sulfur, and reductase). The sequences were mostly derived from two main phyla, Actinobacteria and Proteobacteria (Figure S1)[61,62] and covered twenty different bacterial orders, encompassing sequences that were $>45\,\%$ identical to P450-TT. In contrast to the previously characterised thermophilic panel of extant CYP116B sequences, most of the sequences were isolated from mesophiles (Figure S2),[43] although specific lineages appeared to be thermophilic, such as P450-TT and its closest homologues from *Caldimonas taiwanensis* and *Schlegelella thermodepolymerans*, which have been isolated from hot springs and activated sludge, respectively.[63,64] To explore the sampled sequence space, we inferred the ancestors from each major phylogenetic group (N374, N1, N213, N104, N118, N23, N70, and N3), as well as specific ancestors of the P450-TT lineage (N106 and N115) and the node (N0) connecting the major clade (Proteobacteria) with the outgroup (Actinobacteria). A total of 11 full-length multidomain CYP116B ancestors (Figure 2, the highlighted nodes within the tree) were inferred using the inference program Graphical Representation of Ancestral Sequence Predictions (GRASP), through a joint reconstruction maximum likelihood approach.[20] These self-sufficient ancestral enzymes were selected to explore the thermostability and fatty acid hydroxylation regioselectivity trends in the P450-TT lineage and to investigate nodes on divergent CYP116B branches corresponding to mesophilic wild-type sequences. All ancestors were expressed as holoenzymes with characteristic P450 spectra, with two forms reaching micromolar concentrations in culture (Table S3). N106 and N115 showed higher expression at 25 °C than all other CYP116B enzymes including their direct descendant, P450-TT, which was the most highly expressed extant form.

### Thermostability Profiles

Enhanced thermostability has been repeatedly observed in reconstructed ancestral proteins, including other P450s, compared to the extant forms from which they are inferred, where sequences have been derived from mesophilic organisms. However, to the best of our knowledge, no reconstructions have been performed to trace the evolution of stability in protein families from thermophiles. As the CYP116B family contains proteins from both mesophiles and thermophiles (Figure S2), we were interested in whether thermostability in CYP116B family enzymes is a recent development and a specific adaptation in CYP116B sequences from thermophilic organisms. We assessed the thermostability of the CYP116B ancestors by incubating the proteins at a range of temperatures (25–80 °C) and assessing the loss of the P450 holoenzyme by changes to the characteristic P450 spectrum. Two incubation times were used (15 and 60 minutes), which allowed us to determine both the $^{15}T_{50}$ and $^{60}T_{50}$ stabilities for each ancestor (Table 1; Figures S3 and S4). A wide range of thermostabilities ($^{15}T_{50}$ 38–54 °C and $^{60}T_{50}$ <30–55 °C) was observed among different ancestors across the CYP116B tree. As expected, the $^{60}T_{50}$ values were generally lower than the $^{15}T_{50}$ values, indicating a lower tolerance for prolonged heating, particularly in the ancestors

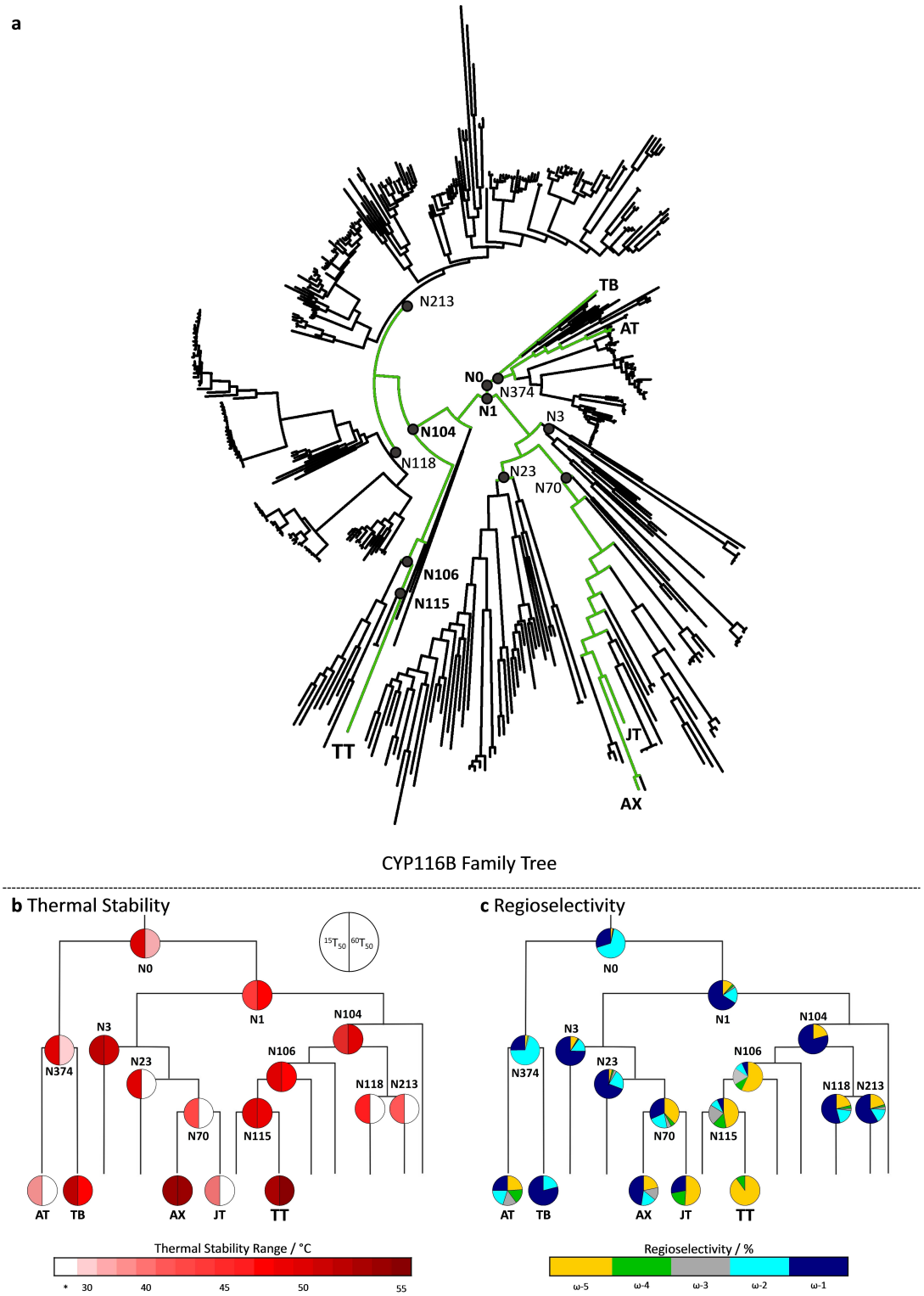



***Figure 2.*** Phylogeny of the 467 sequences from the CYP116B family (top) used for ASR depicting the "evolutionary pathway" (in green) towards the extant wild-type panel and the resurrected nodes (a). The simplified tree representations (bottom) highlight the different thermal stability (b) and hydroxylation regioselectivity distribution (c) of the ancestral and extant enzymes at each node. The thermal stability was determined from the residual P450 peak after incubating the enzyme at 25–80 °C for 15 min or 60 min. P450-TT and -AX were the most stable extant forms whereas N3 was the most thermostable ancestral enzyme. The regioselectivity of the ancestral enzymes in the pathway to P450-TT showed a change from the mid-chain selectivity in P450-TT towards subterminal selectivity in the older ancestors. In addition, there was a change in the ancestral nodes from ω-1 hydroxylation in the majority of the ancestral enzymes to ω-2 hydroxylation in N0 and N374. Data for the extant forms are from Manning et al.[13]

***Table 1:*** Thermostability of the CYP116B ancestral enzymes, where the $^{15}T_{50}$ and $^{60}T_{50}$ values are the temperature when 50% of the P450 has been lost after a 15- or 60-min incubation. Data are means ± standard error.

| Thermal stability | Ancestral CYP116B panel | | | | | | | | | | | Extant CYP116B panel | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | N0 | N1 | N3 | N23 | N70 | N104 | N106 | N115 | N118 | N213 | N374 | TT | JT | AX | AT | TB |
| $^{15}T_{50}$/ °C | 48 ± 1 | 44 ± 4 | 51 ± 1 | 40 ± 1 | 43 ± 2 | 45 ± 3 | 50 ± 1 | 50 ± 1 | 46 ± 4 | 42 ± 1 | 50 ± 1 | 53 ± 1 | 40 ± 1 | 54 ± 1 | 38 ± 3 | 52 ± 1 |
| $^{60}T_{50}$/ °C | 36 ± 8 | 48 ± 2 | 49 ± 2 | n.d.[a] | n.d.[a] | 50 ± 2 | 48 ± 1 | 49 ± 1 | n.d.[a] | n.d.[a] | < 30[a] | 55 ± 1 | n.d.[a] | 54 ± 3 | n.d.[a] | 48 ± 2 |

[a] An accurate estimate of $^{60}T_{50}$ could not be determined (n.d.) since a full sigmoidal denaturation curve was not obtained.

and wild-type enzymes with lower stability, for example, $^{15}T_{50}$ < 45 °C (Figure 2, Table 1). A full sigmoidal denaturation curve was not obtained for 6 of the enzymes at 60 min, so the $^{60}T_{50}$ values are not shown; however, $^{15}T_{50}$ correlated with $^{60}T_{50}$ where both values could be determined ($R^2$ = 0.874). Only five ancestral enzymes maintained their thermal stability after an hour incubation.

In contrast to most other reconstructions, particularly of P450 enzymes, we did not observe a general trend towards greater thermostability in the deeper nodes. Instead, specific lineages appeared to have lost and then sometimes developed thermostability de novo during CYP116B evolution (Figure 2, Figure S5). P450-AX, -TT, and -TB were the most thermostable extant forms, in agreement with previous studies[43] with $^{15}T_{50}$ values of 54, 53, and 52 °C, respectively, over 10 °C higher than those for P450-JT and -AT, forms that were markedly less thermostable and derived from organisms with generally lower optimal growth temperatures.[43] Amongst the ancestors, N3, which theoretically gave rise to mesophilic and uncharacterised extant CYP116B sequences, was the most stable with a $^{15}T_{50}$ of 51 °C (Table 1, Figure S5).

A closer look at the sequence comparisons in relation to three dimensional structures showed that variation was concentrated in the loops of the ancestors (Figure S6). Greater flexibility in the protein scaffold may explain the reduced thermostability of the ancestors relative to P450-TT.[65] Often proteins from thermophiles are shorter than those from mesophiles, by reduction in the length of loops and linkers; however, no such difference was apparent here between P450-TT and the ancestral CYP116B forms. A general comparison of amino acid composition (Table S4) showed a high degree of similarity between ancestors and extant forms, with a slight shift towards more hydrophobic amino acids in the ancestors compared to P450-TT, as noted previously for the CYP3 family reconstruction.[37]

We were interested in tracing the evolutionary paths leading to the development of thermostable extant enzymes, such as P450-TT; many ASR studies using sequences from mesophiles and one thermophilic family of proteins have reported increases in thermostability in deeper ancestors, but it was not clear if the same would be true for enzymes from thermophiles within a larger, predominantly mesophilic tree.[29,37,38,66] In our studies we found a range of stability profiles for the haem domain, however, no clear overall trends were observed across the tree, and the oldest ancestors were not the most thermostable. Overall, there appeared to be considerable variability in the $^{15}T_{50}$ values of ancestors of various lineages, with intermediate nodes varying in $^{15}T_{50}$ between approximately 42 and 51 °C. However, ignoring the most basal node of the tree, N0, which joins the proteobacterial clade with the actinobacterial outgroup so cannot be strictly considered as a potential historical ancestor, a general trend towards gain of thermostability over time can be seen in lineages terminating in a thermostable extant form (i.e., P450-AX, -TT, or -TB). For example, in the P450-TT lineage, for which more intermediate nodes were characterised, the ancestors closest to P450 TT were also thermostable, with both N115 and N106 having a $^{15}T_{50}$ of 50 °C, suggesting a trend towards gain of thermostability in this lineage. These differences in thermostability may reflect a transition from meso- to thermophilic bacteria in some lineages, for example, leading to *Tepidiphilus thermophilus*,[67] where there would have been selection pressure for enhanced thermostability. Characterisation of additional intermediate or descendant nodes in other lineages would allow better definition of trends in other branches of the tree, including the point at which thermostability arose in different CYP116B lineages. It would be of particular interest to assess descendants of N3 and N23, which showed relatively high $^{15}T_{50}$ values, and determine whether any of the uncharacterised source organisms of their descendants are thermophiles.

These findings contrast with those from the ASR of the thermophilic myo-inositol-3-phosphate synthase family proposed to have descended from a thermophilic archaeal ancestor, where the resurrected ancestral forms were more stable than the extant enzymes.[66] However, in both cases, the trends in thermostability in the inferred ancestors *vs.* their extant counterparts reflect the expected changes based on the hypothesised niche of the ancestral organism; here, the ancestor is presumed to have been a mesophile since the majority of the extant forms containing a CYP116B are mesophilic, whereas the ancestral myo-inositol-3-phosphate synthase is proposed to have existed in an organism that was more thermophilic than its descendants.

It has been proposed that ASR using maximum likelihood methods is biased towards predicting ancestors with enhanced thermostability.[68] However, the trends seen here in lineages leading to thermophiles suggest instead that any such bias is likely to be small relative to the trends related to evolution of biological niche (i.e., meso- to thermophile). This implies that it is not necessarily feasible to obtain a more thermostable enzyme simply by inferring and resurrecting ancestors of a group of extant sequences without considering the possible evolutionary pathway in the context

of the biological niche of the extant enzymes. Rather, as shown by comparison with other studies,[20,37–40] candidate families for which this strategy is most likely to succeed are those where there has been broad diversification of enzymes in a protein family from a common ancestor in mesophiles, leading to large extant families with diversified functions.

### *Ancestral Catalytic Activities*

In multidomain enzymes, such as the self-sufficient CYP116B forms, loop and linker variation across domains, such as seen here among in the ancestors (Figure S7), can influence the dynamics and interactions between the domains. In particular, efficient electron transfer across the domains to the haem domain is vital for hydroxylation activity. The ability of the ancestors to hydroxylate fatty acids was analysed using purified enzyme (10 and 50 μM; Figures S8–9), with decanoic acid (0.5 mM) as the substrate with a glucose dehydrogenase recycling system, in 0.1 M potassium phosphate buffer, pH 8 and at 30 °C, with agitation at 200 rpm for 20 h. Decanoic acid was chosen as a substrate because a highly selective and unusual mid-chain hydroxylation had been reported previously and the products have considerable commercial value as lactones in the food and flavour industries.[13] The biotransformations utilised glucose dehydrogenase for concomitant regeneration of NADPH, the preferred cofactor of the reductase domain of P450-TT.[43] At the lower enzyme concentration (10 μM), N70 displayed 69 % conversion of decanoic acid, whereas the conversion with the other ancestral enzymes was low or negligible (0–37 %, Table S5). Therefore, the ancestral regioselectivity fingerprints (Figure 2) were examined using a five-fold higher enzyme loading. Under these conditions, all self-sufficient ancestors were catalytically active (Tables S5 and S6), implying that electron transfer was maintained in all the ancestors, and attesting to the ability of ASR to facilitate the engineering of active electron transport chains. Interestingly, the highest turnover among ancestral forms at both P450 loadings was achieved with N70, N106, and N115, forms which also showed the best expression.

ASR has been widely used for single-domain proteins,[30,37,56,57] but enzymes composed of multiple domains linked by highly variable linkers, present a particular challenge since these regions influence conformational dynamics (e.g., needed for efficient electron transfer) but are also inferred with least confidence in the ASR. Here we show that ASR can also be applied successfully to the three-domain family of self-sufficient CYP116B monooxygenases, generating self-sufficient single polypeptide ancestral proteins that are all properly folded, with electron transfer across the domains and enzymatic activity retained. Conducting ASR on the full length CYP116B system has several advantages, including that the activity of the P450 domain can be directly evaluated because all ancestors are self-sufficient and do not need to be reconstituted. Furthermore, one single series of experiments can yield new synthetic reductases as well as P450s, the activity of which is verified by the ability to support P450 activity in each ancestor. These synthetic reductases should be a useful addition to natural sequences available for engineering self-sufficient fused chimeric enzyme systems,[46,69–74] an approach which has become a very popular for generating active P450s.

### *Evolutionary Shift in Regioselectivity*

It was previously shown that P450-TT was highly selective for mid-chain hydroxylation of decanoic acid.[13] Across the ancestral panel, hydroxylation of decanoic acid at each position between ω-5 and ω-1 was detected; yet there was no observation of any mid-chain ω-6 hydroxylation or hydroxylation closer to the carboxylate. In N0 and N374, there was an unexpected shift in selectivity to favour the ω-2 position over the ω-1 position. This novel subterminal preference for the ω-2 position was not seen in any of the wild-type product distributions. By contrast, the other deep ancestors, N1, N104, N3, and N23, showed a preference for ω-1 hydroxylation (Figure 2). The deeper ancestral enzymes tended to be more regioselective than the more recent ancestors or the extant enzymes that were examined, with the exception of P450-TT and -TB. Mid-chain hydroxylation arose independently in two different lineages (N70 and N106), from ancestors that were predominantly ω-1 hydroxylases (N1, N3 and N104). In the N70 lineage, a non-selective ancestor (N70: 38 % mid-chain and 32 % subterminal), gave rise to two non-selective wild-type enzymes P450-JT and -AX which displayed slight preferences for the ω-5 (51 %) and ω-1 (48 %) positions, respectively. In contrast, the P450-TT lineage developed a preference for mid-chain hydroxylation much earlier, in the N106 ancestor, and resulted in a switch to exclusively mid-chain hydroxylation in P450-TT. Indeed, the major product was hydroxylated at the ω-5 position, with 99 % conversion, for both N115 and N106, although these ancestors also showed a greater tendency towards hydroxylation at the ω-3 and ω-4 positions, compared to P450-TT. Therefore, this lineage shows a prominent shift in the regioselectivity, compared with the CYP116B bacterial ancestors, and suggests an evolutionary pathway towards the unusual regioselectivity of P450-TT. These results highlight the different ways in which selectivity might have evolved and provide biocatalysts for hydroxylation at alternative C–H positions in fatty acids.

### *Structural Insight into the Shifting Regioselectivity*

Having explored the range in regioselectivity in the ancestral CYP116B enzymes, we compared their sequences overall and within their active sites (Figure 3) to identify determinants of selectivity. Ancestral sequence identities (Table S2) ranged from 59 (earlier ancestors) to 77 % (more recent ancestors) compared to the P450-TT holoenzyme. Most of the non-conserved residues were found in the regions that are predicted to be surface exposed (Figure 3a), with more conservation in the core regions of each domain. The only changes in the active site were two consecutive amino acids

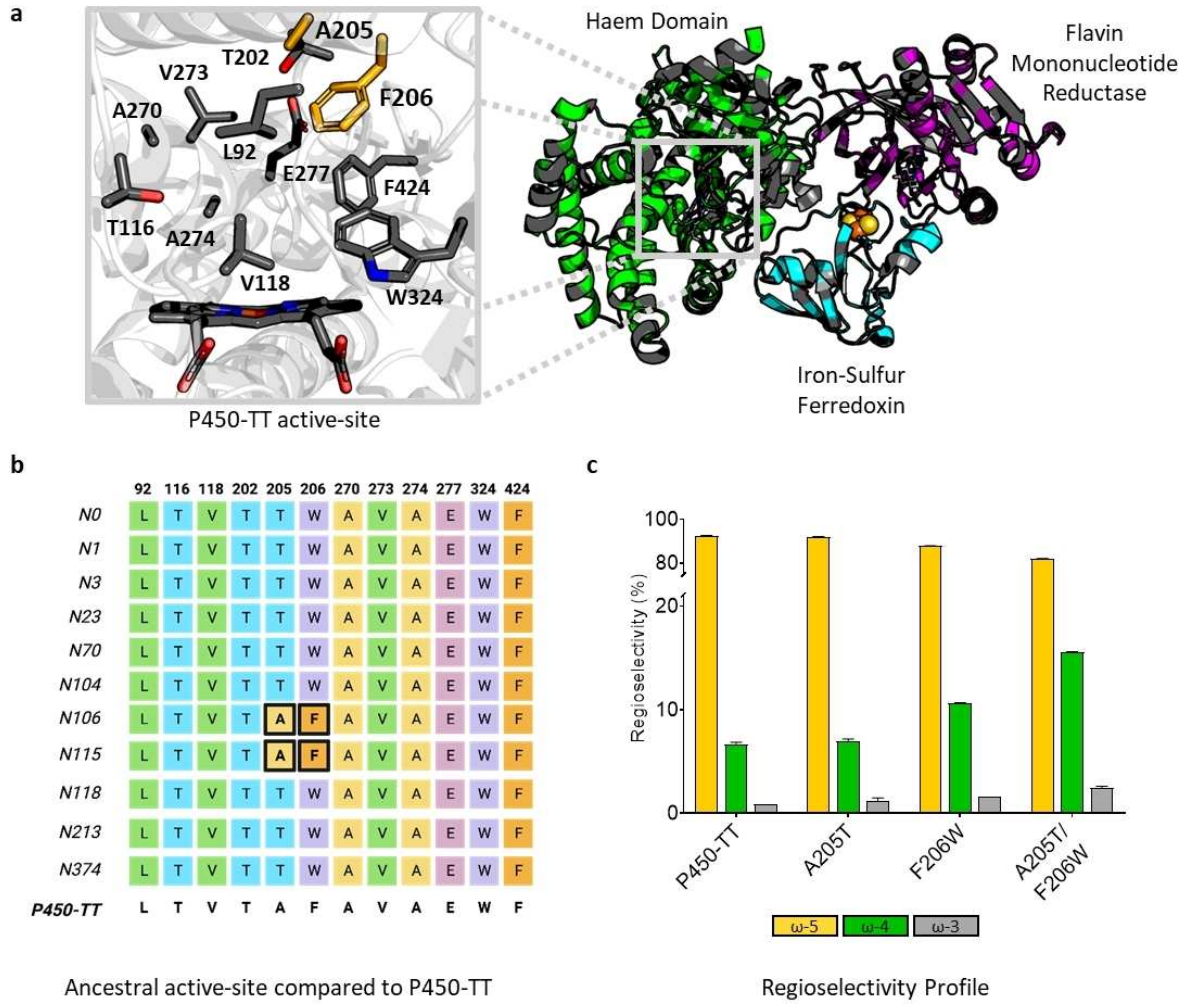



***Figure 3.*** Ancestral sequence comparison to P450-TT. a) Amino acids A205 and F206 (orange sticks; PDB ID: 6laa). All the residues that are not conserved between one or more of the ancestors and P450-TT are in grey. The conserved haem domain residues are in green, conserved reductase residues in pink, and conserved ferredoxin residues in cyan. b) The multiple sequence alignment of the active site residues in the ancestral sequences compared to P450-TT, where the order of residues corresponds to the order in which they appear in the P450-TT amino acid sequence. c) Regioselectivity observed after introducing the A205T and F206 W mutations separately and together into P450-TT.

(Ala205Thr and Phe206Trp), which increased the active site polarity and were present in all ancestors apart from the two closest to P450-TT (N115 and N106, Figure 3b). This suggested a more polar active site may favour subterminal hydroxylation in the CYP116B family. Only 17 of the 467 initial extant sequences in the alignment (Figure S10) lacked the Thr-Trp motif present in most of the characterised ancestral sequences, instead having either Ala-Phe, Ile-Trp or Gln-Phe residues, suggesting these could be key active site positions governing regioselectivity towards fatty acid hydroxylation. Additionally, molecular docking simulations (Figure S11) showed the substrate orientated towards the subterminal positions in the nine ancestral models containing the Thr-Trp active site motif. In contrast, to the two closest ancestors to P450-TT with the Ala-Phe motif (N115 and N106) the carboxyl group was anchored to an alternative Thr residue, favouring mid-chain selectivity.

To better understand the switch in selectivity from the ω-5 to ω-2 or ω-1 products associated with the Ala205Thr and Phe206Trp conserved motif, Ala205 and Phe206 in P450-TT were mutated to obtain the corresponding single and double mutants. The double mutant was more active than P450-TT (Figure S12), which may explain the prominence of these residues in other wild-type sequences. Although ω-5 hydroxylation was still the most favoured product, an increase in the proportion of ω-4 and ω-3 hydroxylated alkyl chains was observed (Figure 3c), suggesting that the increased polarity in the active site has an impact on regioselectivity.

The distinctive pathway to the unique mid-chain selectivity of wild-type, P450-TT, highlights the utility of ASR to provide insight into the determinants of fatty acid hydroxylation regioselectivity and other properties. The shifts in regioselectivity seen here appear to occur steadily along specific evolutionary lineages, with the most dramatic being from the poor overall conversion and subterminal hydroxylation in N104 to the higher conversion and mid-chain ω-5 selectivity of N106, ancestors that are separated by a single

node (N105) in the tree (which contains the AF motif at residues 205–206). Inspection of the sequence differences in this lineage compared to the rest of the ancestors may reveal sequence elements related to the enhanced overall turnover, thermostability or expression yield, not only regioselectivity, an investigation that could also be accelerated by a directed evolution approach.[75]

## Conclusion

In conclusion, facilitated by the steadily expanding pool of genome and metagenome sequences, ASR is a useful tool to both find biocatalysts with new reaction fingerprints, which can be accessed through considerably streamlined screening, as well as to understand trends in reaction selectivity and other properties. Fatty acid hydroxylation is an industrially important and chemically challenging target for regioselective C–H activation and ASR has been able to produce biocatalysts with new regioselectivities, compared to natural sequences explored previously. We have shown here that ASR can be used to obtain a panel of biocatalysts, each with a unique pattern of regioselectivity towards fatty acid hydroxylation. Notably, ancestor N374 would be a good starting point for engineering for efficient ω-2 hydroxylation, having shown approximately 70 % regioselectivity and 42 % product conversion here, as well as relatively high $^{15}T_{50}$, although its longer-term stability was not clear. Such predominant ω-2 hydroxylation had not been observed previously with other P450s and as such constitutes a new enzyme selectivity. Similarly, N70, N106, and N115 would be useful templates for engineering a higher efficiency biocatalyst for ω-5 hydroxylation, since they showed high regioselectivity, almost quantitative conversion of decanoic acid under the conditions used here, along with an overall expression yield that was comparable to or higher than the extant ω-5 hydroxylase, P450-TT.

Overall, ASR has enabled the identification of active site residues influencing the differing C–H regioselectivity in a panel of ancestral and wild-type CYP116B biocatalysts without the need for a large-scale mutagenic screen. These insights into the structure-function relationships leading to the unique regioselectivity of P450-TT, that is, a shift towards mid-chain hydroxylation. This approach has provided a library of diverse artificial templates both for the P450 and the reductase domains and some landmarks in the sequence space around these enzymes that can be used as starting points in future protein engineering campaigns to achieve efficient and regioselective C–H hydroxylation.

## Supporting Information

The authors have cited additional references within the Supporting Information.[76–86]

## Acknowledgements

This work was supported by the European Research Council (788231-ProgrES-ERC-2017-ADG). The Figures were partially created using Biorender.com.

## Conflict of Interest

The authors declare no conflict of interest.

## Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article.

**Keywords:** Ancestral Sequence Reconstruction · Biocatalysis · Cytochrome P450 Monooxygenase · Fatty Acid Hydroxylation · Regioselectivity

[1] K. R. Kim, D. K. Oh, *Biotechnol. Adv.* **2013**, *31*, 1473–1485.
[2] F. A. Thesseling, M. C. Hutter, C. Wiek, J. P. Kowalski, A. E. Rettie, M. Girhard, *Arch. Biochem. Biophys.* **2020**, *679*, 108216–108225.
[3] L. Hammerer, C. K. Winkler, W. Kroutil, *Catal. Lett.* **2018**, *148*, 787–812.
[4] Y.-H. Yang, J.-L. Wang, C. L. Miranda, D. R. Buhler, *Arch. Biochem. Biophys.* **1998**, *352*, 271–280.
[5] H. G. Park, V. Kim, H. Kim, R. Lee, M. A. Cho, S. W. Park, Y. J. Chun, D. Kim, *Arch. Biochem. Biophys.* **2019**, *671*, 27–34.
[6] J. L. Porter, J. Manning, S. Sabatini, M. Tavanti, N. J. Turner, S. L. Flitsch, *ChemBioChem* **2018**, *19*, 513–520.
[7] S. D. Munday, N. K. Maddigan, R. J. Young, S. G. Bell, *Biochim. Biophys. Acta* **2016**, *1860*, 1149–1162.
[8] C. J. C. Whitehouse, S. G. Bell, L.-L. Wong, *Chem. Soc. Rev.* **2012**, *41*, 1218–1260.
[9] J. L. Porter, S. Sabatini, J. Manning, M. Tavanti, J. L. Galman, N. J. Turner, S. L. Flitsch, *Enzyme Microb. Technol.* **2018**, *113*, 1–8.
[10] S. A. Child, V. P. Rossi, S. G. Bell, *Biochim. Biophys. Acta Gen. Subj.* **2019**, *1863*, 408–417.
[11] M. Girhard, S. Schuster, M. Dietrich, P. Dürre, V. B. Urlacher, *Biochem. Biophys. Res. Commun.* **2007**, *362*, 114–119.
[12] S. Honda Malca, D. Scheps, L. Kühnel, E. Venegas-Venegas, A. Seifert, B. M. Nestl, B. Hauer, *Chem. Commun.* **2012**, *48*, 5115–5117.
[13] J. Manning, M. Tavanti, J. L. Porter, N. Kress, S. P. De Visser, N. J. Turner, S. L. Flitsch, *Angew. Chem. Int. Ed.* **2019**, *58*, 5668–5671.
[14] P. C. Cirino, F. H. Arnold, *Adv. Synth. Catal.* **2002**, *344*, 932–937.
[15] M. J. Maseme, A. Pennec, J. van Marwijk, D. J. Opperman, M. S. Smit, *Angew. Chem. Int. Ed.* **2020**, *59*, 10359–10362.
[16] F. Brühlmann, L. Fourage, C. Ullmann, O. P. Haefliger, N. Jeckelmann, C. Dubois, D. Wahler, *J. Biotechnol.* **2014**, *184*, 17–26.
[17] T. W. B. Ost, C. S. Miles, J. Murdoch, Y. F. Cheung, G. A. Reid, S. K. Chapman, A. W. Munro, *FEBS Lett.* **2000**, *486*, 173–177.
[18] L. Hammerer, M. Friess, J. Cerne, M. Fuchs, G. Steinkellner, *ChemCatChem* **2019**, *11*, 5642–5649.

[19] K. Zhang, A. Yu, X. Chu, F. Li, J. Liu, L. Liu, W. J. Bai, C. He, X. Wang, *Angew. Chem. Int. Ed.* **2022**, *61*, e202204290.
[20] G. Foley, A. Mora, C. M. Ross, S. Bottoms, L. Sützl, M. L. Lamprecht, J. Zaugg, A. Essebier, B. Balderson, R. Newell, R. E. S. Thomson, B. Kobe, R. T. Barnard, L. Guddat, G. Schenk, J. Carsten, Y. Gumulya, B. Rost, D. Haltrich, V. Sieber, E. M. J. Gillam, M. Bodén, *PLoS Comput. Biol.* **2022**, *18*, e1010633.
[21] K. Watanabe, T. Ohkuri, S. I. Yokobori, A. Yamagishi, *J. Mol. Biol.* **2006**, *355*, 664–674.
[22] D. Gonzalez, J. Hiblot, N. Darbinian, J. C. Miller, G. Gotthard, S. Amini, E. Chabriere, M. Elias, *FEBS Open Bio* **2014**, *4*, 121–127.
[23] J. H. Whitfield, W. H. Zhang, M. K. Herde, B. E. Clifton, J. Radziejewski, H. Janovjak, C. Henneberger, C. J. Jackson, *Protein Sci.* **2015**, *24*, 1412–1422.
[24] C. Wilson, R. V. Agafonov, M. Hoemberger, S. Kutter, A. Zorba, J. Halpin, V. Buosi, R. Otten, D. Waterman, D. L. Theobald, D. Kern, *Science* **2015**, *347*, 882–886.
[25] V. A. Risso, F. Manssour-Triedo, A. Delgado-Delgado, R. Arco, A. Barroso-delJesus, A. Ingles-Prieto, R. Godoy-Ruiz, J. A. Gavira, E. A. Gaucher, B. Ibarra-Molero, J. M. Sanchez-Ruiz, *Mol. Biol. Evol.* **2015**, *32*, 440–455.
[26] A. Thomas, R. Cutlan, W. Finnigan, M. van der Giezen, N. Harmer, *Commun. Biol.* **2019**, *2*, 429–440.
[27] D. A. Hueting, S. R. Vanga, P.-O. Syrén, *J. Phys. Chem. B* **2022**, *126*, 3809–3821.
[28] L. Barandiaran, B. Alonso-Lerma, A. Reifs, I. Larraza, R. Olmos-Juste, A. Fernandez-Calvo, Y. Jabalera, A. Eceiza, R. Perez-Jimenez, *Commun. Mater.* **2022**, *3*, 55–64.
[29] R. E. S. Thomson, S. E. Carrera-Pacheco, E. M. J. Gillam, *J. Biol. Chem.* **2022**, *298*, 102435.
[30] J. Livada, A. Vargas, C. A. Martinez, R. Lewis, *ACS Catal.* **2023**, *13*, 2576–2585.
[31] P. Babkova, E. Sebestova, J. Brezovsky, R. Chaloupkova, J. Damborsky, *ChemBioChem* **2017**, *18*, 1448–1456.
[32] C. R. Nicoll, G. Bailleul, F. Fiorentini, M. L. Mascotti, M. W. Fraaije, A. Mattevi, *Nat. Struct. Mol. Biol.* **2020**, *27*, 14–24.
[33] Y. Sun, E. Calderini, R. Kourist, *ChemBioChem* **2021**, *22*, 1833–1840.
[34] Y. Joho, V. Vongsouthi, M. A. Spence, J. Ton, C. Gomez, L. L. Tan, J. A. Kaczmarski, A. T. Caputo, S. Royan, C. J. Jackson, A. Ardevol, *Biochemistry* **2023**, *62*, 437–450.
[35] Y. Gumulya, E. M. J. Gillam, *Biochemistry* **2017**, *474*, 1–19.
[36] M. A. Spence, J. A. Kaczmarski, J. W. Saunders, C. J. Jackson, *Curr. Opin. Struct. Biol.* **2021**, *69*, 131–141.
[37] Y. Gumulya, J. M. Baek, S. J. Wun, R. E. S. Thomson, K. L. Harris, D. J. B. Hunter, J. B. Y. H. Behrendorff, J. Kulig, S. Zheng, X. Wu, B. Wu, J. E. Stok, J. J. De Voss, G. Schenk, U. Jurva, S. Andersson, E. M. Isin, M. Bodén, L. Guddat, E. M. J. Gillam, *Nat. Catal.* **2018**, *1*, 878–888.
[38] Y. Gumulya, W. Huang, S. A. D'Cunha, K. E. Richards, R. E. S. Thomson, D. J. B. Hunter, J. M. Baek, K. L. Harris, M. Boden, J. J. De Voss, M. A. Hayes, E. M. Isin, S. Andersson, U. Jurva, E. M. J. Gillam, *ChemCatChem* **2019**, *11*, 841–850.
[39] P. Hartz, S. J. Strohmaier, B. M. EL-Gayar, A. Abdulmughni, M. C. Hutter, F. Hannemann, E. M. J. Gillam, R. Bernhardt, *FEBS J.* **2021**, *288*, 6510–6527.
[40] K. L. Harris, R. E. S. Thomson, Y. Gumulya, G. Foley, S. E. Carrera-Pacheco, P. Syed, T. Janosik, A. S. Sandinge, S. Andersson, U. Jurva, M. Boden, E. M. J. Gillam, *Mol. Biol. Evol.* **2022**, *39*, msac116.
[41] E. O'Reilly, V. Köhler, S. L. Flitsch, N. J. Turner, *Chem. Commun.* **2011**, *47*, 2490–2501.
[42] L. Zhang, Z. Xie, Z. Liu, S. Zhou, L. Ma, W. Liu, J. W. Huang, T. P. Ko, X. Li, Y. Hu, J. Min, X. Yu, R. T. Guo, C. C. Chen, *Nat. Commun.* **2020**, *11*, 2676–2681.
[43] M. Tavanti, J. L. Porter, S. Sabatini, N. J. Turner, S. L. Flitsch, *ChemCatChem* **2018**, *10*, 1042–1051.
[44] Z. Li, Y. Jiang, X. F. P. Guengerich, L. Ma, S. Li, W. Zhang, *J. Biol. Chem.* **2020**, *295*, 833–849.
[45] B. E. Eser, Y. Zhang, L. Zong, Z. Guo, *Chin. J. Chem. Eng.* **2021**, *30*, 121–135.
[46] H. Renata, *Biochemistry* **2023**, *62*, 253–261.
[47] C. C. Chen, J. Min, L. Zhang, Y. Yang, X. Yu, R. T. Guo, *ChemBioChem* **2021**, *22*, 1317–1328.
[48] T. Devamani, A. M. Rauwerdink, M. Lunzer, B. J. Jones, J. L. Mooney, M. A. O. Tan, Z. J. Zhang, J. H. Xu, A. M. Dean, R. J. Kazlauskas, *J. Am. Chem. Soc.* **2016**, *138*, 1046–1056.
[49] M. Wilding, T. S. Peat, S. Kalyaanamoorthy, J. Newman, C. Scott, L. S. Jermiin, *Green Chem.* **2017**, *19*, 5375–5380.
[50] U. Alcolombri, M. Elias, D. S. Tawfik, *J. Mol. Biol.* **2011**, *411*, 837–853.
[51] G. Conti, L. Pollegioni, G. Molla, E. Rosini, *FEBS J.* **2014**, *281*, 2443–2455.
[52] P. M. Shih, A. Occhialini, J. C. Cameron, P. J. Andralojc, M. A. J. Parry, C. A. Kerfeld, *Nat. Commun.* **2016**, *7*, 10382–10392.
[53] K. Schriever, P. Saenz-Mendez, R. S. Rudraraju, N. M. Hendrikse, E. P. Hudson, A. Biundo, R. Schnell, P. O. Syrén, *J. Am. Chem. Soc.* **2021**, *143*, 3794–3807.
[54] B. Zeng, S. Y. Zhao, R. Zhou, Y. H. Zhou, W. H. Jin, Z. W. Yi, G. Y. Zhang, *Comput. Struct. Biotechnol. J.* **2022**, *20*, 3313–3321.
[55] S. Chen, Z. Xu, B. Ding, Y. Zhang, S. Liu, C. Cai, M. Li, B. E. Dale, M. Jin, *Sci. Adv.* **2023**, *9*, eadd8835.
[56] G. Bailleul, G. Yang, C. R. Nicoll, A. Mattevi, M. W. Fraaije, M. L. Mascotti, *Nat. Commun.* **2023**, *14*, 1042–1031.
[57] C.-H. Chiang, T. Wymore, A. Rodríguez Benítez, A. Hussain, J. L. Smith, C. L. Brooks, A. R. H. Narayan, *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2218248120.
[58] M. Tavanti, J. L. Porter, C. W. Levy, J. R. Gómez Castellanos, S. L. Flitsch, N. J. Turner, *Biochem. Biophys. Res. Commun.* **2018**, *501*, 846–850.
[59] A. Ciaramella, G. Catucci, G. Gilardi, G. Di Nardo, *Int. J. Biol. Macromol.* **2019**, *140*, 577–587.
[60] R. Gong, L. J. Wu, Y. Zhang, Z. Lui, S. Dou, R. G. Zhang, J. H. Xu, C. Tang, J. H. Zhou, *Protein Data Bank* **2020**, https://doi.org/10.2210/pdb6KBH/pdb.
[61] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P. A. Chaumeil, P. Hugenholtz, *Nat. Biotechnol.* **2018**, *36*, 996–1004.
[62] G. A. Coleman, A. A. Davín, T. A. Mahendrarajah, L. L. Szánthó, A. Spang, P. Hugenholtz, G. J. Szöllsi, T. A. Williams, *Science* **2021**, *372*, eabe0511.
[63] W. M. Chen, J. S. Chang, C. H. Chiu, S. C. Chang, W. C. Chen, C. M. Jiang, *Syst. Appl. Microbiol.* **2005**, *28*, 415–420.
[64] K. Elbanna, T. Lütke-Eversloh, S. Van Trappen, J. Mergaert, J. Swings, A. Steinbüchel, *Int. J. Syst. Evol. Microbiol.* **2003**, *53*, 1165–1168.
[65] K. L. Harris, R. E. S. Thomson, S. J. Strohmaier, Y. Gumulya, E. M. J. Gillam, *Biochim. Biophys. Acta Proteins Proteomics* **2018**, *1866*, 97–115.
[66] N. C. Butzin, P. Lapierre, A. G. Green, K. S. Swithers, J. P. Gogarten, K. M. Noll, *PLoS One* **2013**, *8*, e84300.
[67] K. M. Hart, M. J. Harms, B. H. Schmidt, C. Elya, J. W. Thornton, S. Marqusee, *PLoS Biol.* **2014**, *12*, e1001994.
[68] P. D. Williams, D. D. Pollock, B. P. Blackburne, R. A. Goldstein, *PLoS Comput. Biol.* **2006**, *2*, 0598–0605.
[69] M. Nodate, M. Kubota, N. Misawa, *Appl. Microbiol. Biotechnol.* **2006**, *71*, 455–462.
[70] S. Li, L. M. Podust, D. H. Sherman, *J. Am. Chem. Soc.* **2007**, *129*, 12940–12941.

[71] A. Robin, G. A. Roberts, J. Kisch, F. Sabbadin, G. Grogan, N. Bruce, N. J. Turner, S. L. Flitsch, *Chem. Commun.* **2009**, *18*, 2478–2480.
[72] A. Robin, V. Köhler, A. Jones, A. Ali, P. P. Kelly, E. O'Reilly, N. J. Turner, S. L. Flitsch, *Beilstein J. Org. Chem.* **2011**, *7*, 1494–1498.
[73] C. C. Chen, M. Dai, L. Zhang, J. Zhao, W. Zeng, M. Shi, J. W. Huang, W. Liu, R. T. Guo, A. Li, *ACS Catal.* **2022**, *12*, 2831–2839.
[74] L. E. Zetzsche, J. A. Yazarians, S. Chakrabarty, M. E. Hinze, L. A. M. Murray, A. L. Lukowski, L. A. Joyce, A. R. H. Narayan, *Nature* **2022**, *603*, 79–85.
[75] J. B. Y. H. Behrendorff, C. D. Moore, K. H. Kim, D. H. Kim, C. A. Smith, W. A. Johnston, C. H. Yun, G. S. Yost, E. M. J. Gillam, *Chem. Res. Toxicol.* **2012**, *25*, 1964–1974.
[76] G. Foley, L. Sützl, S. A. D'Cunha, E. M. J. Gillam, M. Bodén, *BioTechniques* **2019**, *67*, 50–54.
[77] K. Katoh, H. Toh, *Bioinformatics* **2010**, *26*, 1899–1900.
[78] B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. Von Haeseler, R. Lanfear, E. Teeling, *Mol. Biol. Evol.* **2020**, *37*, 1530–1534.
[79] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, L. S. Jermiin, *Nat. Methods* **2017**, *14*, 587–589.
[80] D. Thi Hoang, O. Chernomor, A. von Haeseler, B. Quang Minh, L. Sy Vinh, M. S. Rosenberg, *Mol. Biol. Evol.* **2018**, *35*, 518–522.
[81] T. Pupko, I. Pe, R. Shamir, D. Graur, *Mol. Biol. Evol.* **2000**, *17*, 890–896.
[82] T. Omura, R. Sato, *J. Biol. Chem.* **1964**, *239*, 2370–2378.
[83] W. A. Johnston, W. Huang, J. J. De Voss, M. A. Hayes, E. M. J. Gillam, *J. Biomol. Screening* **2008**, *13*, 135–141.
[84] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. P. De Beer, C. Rempfer, L. Bordoli, R. Lepore, T. Schwede, *Nucleic Acids Res.* **2018**, *46*, W296–W303.
[85] E. Krieger, G. Vriend, *Bioinformatics* **2014**, *30*, 2981–2982.
[86] I. Letunic, P. Bork, *Nucleic Acids Res.* **2021**, *49*, W293–W296.

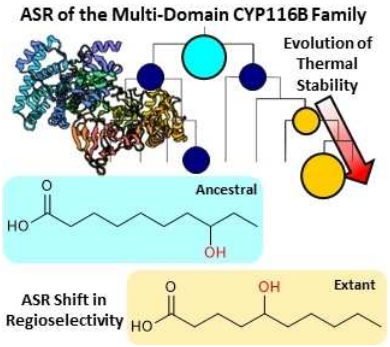
Manuscript received: October 4, 2023
Accepted manuscript online: January 1, 2024
Version of record online: ■■, ■■

# Forschungsartikel

*Enzyme Catalysis*

B. S. Jones, C. M. Ross, G. Foley, N. Pozhydaieva, J. W. Sharratt, N. Kress, L. S. Seibt, R. E. S. Thomson, Y. Gumulya, M. A. Hayes, E. M. J. Gillam,* S. L. Flitsch* _____ **e202314869**

Engineering Biocatalysts for the C–H Activation of Fatty Acids by Ancestral Sequence Reconstruction




A family of biocatalysts for the regioselective mid- and end-chain C–H activation of fatty acids to generate hydroxy acids is described. Ancestral sequence reconstruction was successfully applied to the multidomain enzyme family of CYP116B cytochrome P450 monooxygenases to generate ancestor variants with distinct product and thermostability profiles.