nature catalysis

ARTICLES

https://doi.org/10.1038/s41929-022-00743-0

Biocatalytic synthesis of non-standard amino acids by a decarboxylative aldol reaction

Jonathan M. Ellis[1,3], Meghan E. Campbell[1,3], Prasanth Kumar[1], Eric P. Geunes[1], Craig A. Bingman[2] and Andrew R. Buller[1,2] ✉

Enzymes are renowned for their catalytic efficiency and selectivity. Despite the wealth of carbon–carbon bond-forming transformations in traditional organic chemistry and nature, relatively few C–C bond-forming enzymes have found their way into the biocatalysis toolbox. Here we show that the enzyme UstD performs a highly selective decarboxylative aldol addition with diverse aldehyde substrates to make non-standard γ-hydroxy amino acids. We increased the activity of UstD through three rounds of classic directed evolution and an additional round of computationally guided engineering. The enzyme that emerged, UstD$^{v2.0}$, is efficient in a whole-cell biocatalysis format. The products are highly desirable, functionally rich bioactive γ-hydroxy amino acids that we demonstrate can be prepared stereoselectively on the gram scale. The X-ray crystal structure of UstD$^{v2.0}$ at 2.25 Å reveals the active site and provides a foundation for probing the UstD mechanism.

Major advances have been made in the practical use of enzymes for enantioselective functional group manipulations[1]. For example, the asymmetric reduction of ketones and enantiospecific hydrolysis of racemic esters are now routine in process chemistry. Also, impressive strides have been made in enzymatic C–H activation[2]. However, the development of enzymes to form C–C bonds on a preparative scale lags far behind that of traditional synthetic organic methodology[3]. Although nature is rife with C–C bond-forming enzymes[4,5], these catalysts often have substantial limitations, such as a limited substrate scope or poor heterologous expression[6]. Engineering can overcome these challenges, but a more severe limitation is thermodynamic in nature: reactions that form carbon nucleophiles via C–H deprotonation, such as classic aldol transformations, are typically reversible[7]. In nature, metabolic flux drives reactions and preserves the stereochemical purity of the products. Laboratory approaches mimic nature by coupling reversible biocatalytic C–C bond-forming reactions to a thermodynamic sink, such as a subsequent transformation or selective crystallization[8–11]. Although these advances are substantial, the potential of biocatalytic enzymes in assembling carbon chains is still hindered by the simple lack of high-quality exergonic transformations[12]. Hence, development of scalable and thermodynamically favourable C–C bond-forming reactions may open diverse avenues of biocatalytic synthesis.

To fill this gap, we were drawn to a recently described pyridoxal 5′-phosphate (PLP)-dependent enzyme involved in the biosynthesis of Ustiloxin B, an inhibitor of microtubilin polymerization (Fig. 1a)[13]. This enzyme, UstD, decarboxylates the side chain of L-aspartate (**1**) to form a putative nucleophilic enamine intermediate (Fig. 1b). This enamine then attacks an aliphatic aldehyde appended to a cyclic tetrapeptide, which results in the formation of a γ-hydroxy amino acid side chain. The loss of CO_2 renders this enantioselective C–C bond-forming reaction effectively irreversible. This decarboxylative aldol addition mechanism is distinct from the classic aldolases, transketolases and PLP-dependent Thr aldolases, which catalyse the tautomerization of an imine to form an enamine nucleophile[14,15]. It has been shown that the transketolase catalytic cycle can be non-natively entered through decarboxylation, and that the reactions initially proceed to a high conversion. However, the native proton transfer machinery eventually breaks down the product into an equilibrium mixture with starting materials[16]. Although the detailed mechanism of this UstD addition has not yet been explored, Ye et al. reported that the UstD reaction cannot be initiated from L-Ala, which indicates that enamine formation through tautomerization is not viable. Therefore, UstD is mechanistically distinct from classic aldolases and may have unique properties as a biocatalyst.

The native substrate for UstD is a complex, cyclic peptide, and it was not known if this enzyme would react promiscuously with alternative substrates. If so, the enzyme would directly produce γ-hydroxy amino acids (Fig. 1b). Such non-standard amino acids (nsAAs) are found in bioactive natural products, such as caspofungin and clavalanine (Fig. 1a)[17]. Although nature employs side-chain hydroxylation to tune bioactivity, these nsAAs are virtually absent from medicinal chemistry[18] because they require multistep synthesis[17]. The need for multistep synthesis to prepare these nsAAs has begun to be addressed by biocatalysis, in which an elegant multienzyme cascade was recently developed by Clapés and co-workers to access γ-hydroxy nsAAs[19,20]. However, the ability to use a single enzyme to produce the same motif offers a greater practical utility and versatility. Beyond their use in pharmaceuticals, nsAAs can be enabling for a host of synthetic and chemical biology applications[21,22]. Therefore, the development of UstD for organic synthesis would introduce a valuable and much-needed enantioselective C–C bond-forming enzyme into the biocatalytic toolbox and provide direct access to a structurally complex synthon.

Here we show that the enzyme UstD performs a highly selective decarboxylative aldol addition with diverse aldehyde substrates to make non-standard γ-hydroxy amino acids. We increased the activity of UstD through three rounds of classic directed evolution and an additional round of computationally guided engineering. The enzyme that emerged, UstD$^{v2.0}$, is efficient in a whole-cell biocatalysis format, which circumvents the need for enzyme purification, and thereby facilitates its use in traditional organic settings on a gram scale. The X-ray crystal structure of UstD$^{v2.0}$ at 2.25 Å reveals the active site and the molecular basis for the promiscuity of this catalyst.

[1]Department of Chemistry, University of Wisconsin–Madison, Madison, WI, USA. [2]Department of Biochemistry, University of Wisconsin–Madison, Madison, WI, USA. [3]These authors contributed equally: Jonathan M. Ellis, Meghan E. Campbell. ✉e-mail: arbuller@wisc.edu

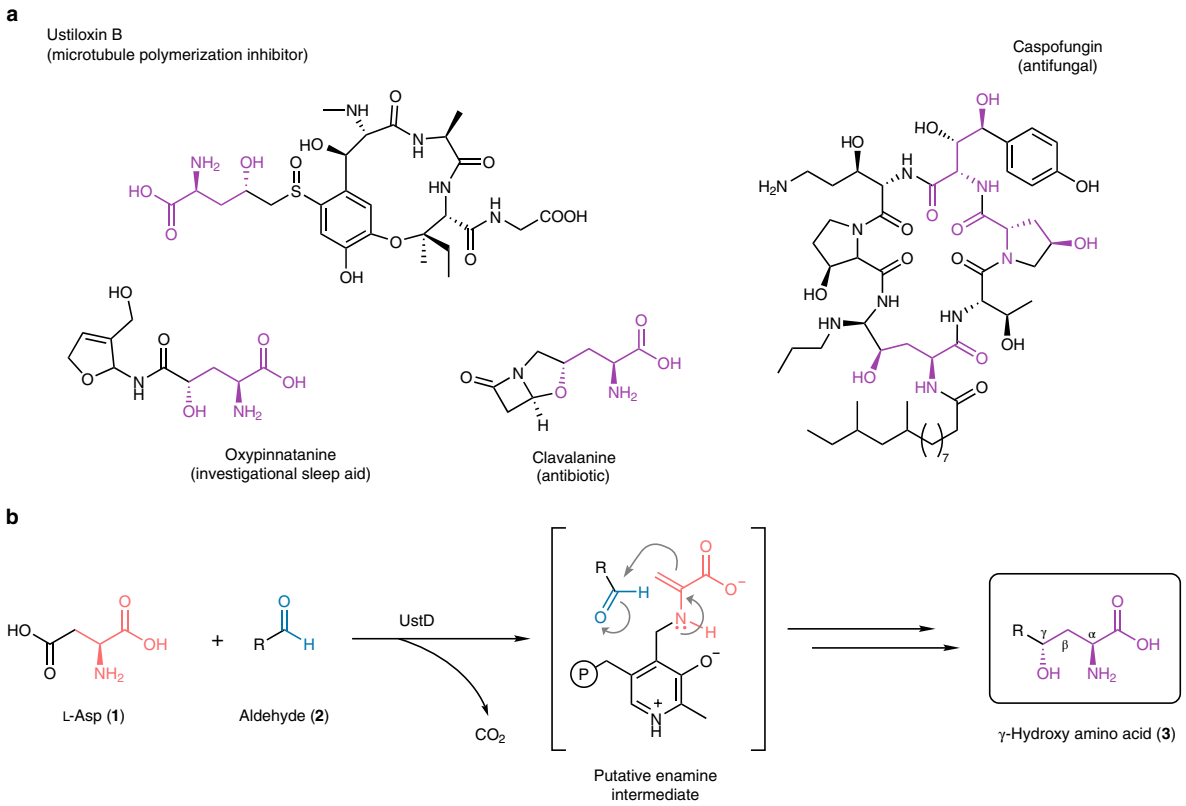

Fig. 1 | Relevance and mechanism of enzymatic C–C bond formation. a, Bioactive molecules with a γ-hydroxy amino acid motif, shown in purple. The native product of UstD is Ustiloxin B. **b**, The generalized decarboxylative aldol reaction of UstD showing the putative enamine nucleophilic intermediate.

Results

Initial characterization of UstD. We expressed C-His-UstD (wild-type (wt)-UstD) in *Escherichia coli* (Supplementary Fig. 1), but were uncertain whether molecular recognition for the structurally complex native substrate would be required for catalytic activity. We therefore assessed the reactivity of wt-UstD with benzaldehyde **2a** and were pleased to observe a successful decarboxylative aldol addition to afford the γ-hydroxy nsAA **3a** by ultra-high pressure liquid chromatography–mass spectrometry (UPLC–MS) (Supplementary Fig. 2). A preparative scale reaction with 0.125 mol% catalyst gave the product in 43% yield, and analysis by NMR spectroscopy indicated a single diastereomer predominated (d.r. >98:2). To determine the absolute stereochemical preference for the enzyme, we analysed the product from a reaction with 4-bromobenzaldehyde (**2b**). The crystal structure of the product (**3b**) revealed that the aldol addition occurred with the same stereochemical outcome as that of the native reaction (Supplementary Fig. 2). These transformations indicated that wt-UstD has the potential for organic synthesis, but the comparatively modest activity (<1,000 turnovers with the initial reaction conditions) and low catalyst expression would hinder routine use of the natural enzyme. Given the inherent structural differences between the native tetrapeptide substrate and simpler commercially available aldehydes (such as **2a**), we hypothesized that directed evolution and reaction-condition optimization could be used to increase the catalytic efficiency of UstD towards non-native substrates.

Directed evolution of UstD for improved catalytic activity. To inform our engineering process, we used a homology model of wt-UstD derived from a distantly related cysteine desulfurase (27% identity)[23,24]. Six residues in the predicted active site were chosen for saturation mutagenesis, and we used benzaldehyde (**2a**) as a model substrate for the directed evolution (Fig. 2a). Mutation at positions predicted to form direct contacts with the cofactor resulted in inactivation of the catalyst, a common trend among PLP-dependent enzymes[25]. Nevertheless, these libraries yielded a single variant in a putative loop region that flanked the substrate binding site, C392L, with a 2.3-fold boost in activity (Fig. 2b). Concurrently, we employed global random mutagenesis on wt-UstD to search throughout the protein sequence for activating mutations. A second activating mutation was discovered, L393M, immediately adjacent to Cys392. We combined these mutations to yield the double variant UstD^{C392L,L393M}, which had a further increase in activity to 4.9-fold above the wild type (Supplementary Fig. 3). It is common for the mutation of neighbouring residues to display cooperativity[26,27], and we chose to test additional mutations in this region of the sequence (Fig. 2b). We used a degenerate codon mutagenesis strategy on four contiguous residues from Ile391 to Ala394. We restricted the sequence space to residues commonly found among UstD homologues, which provided a good structural diversity in a focused set of mutations (see Supplementary Information for the details). Screening this library revealed that mutation of Ala394 was generally deleterious. However, multiple highly active variants retained Ala394 and contained mutations at Ile391, Cys392 and Leu393. To best capture the relative rate effects of mutations, catalysts were compared under dilute conditions. Variants UstD^{TLM} and UstD^{FVF} (the superscript refers to the identity of the residues at positions 391–393) had a 5.1-fold and 4.1-fold increase in activity relative to wt-UstD, respectively.

We next optimized the reaction conditions for the most active variant, UstD^{TLM}. Reaction mixtures were initially coloured yellow (Supplementary Fig. 1) by the presence of PLP that co-purified with the enzyme, but became colourless over time, which suggests the cofactor degraded during the reaction. Gratifyingly,

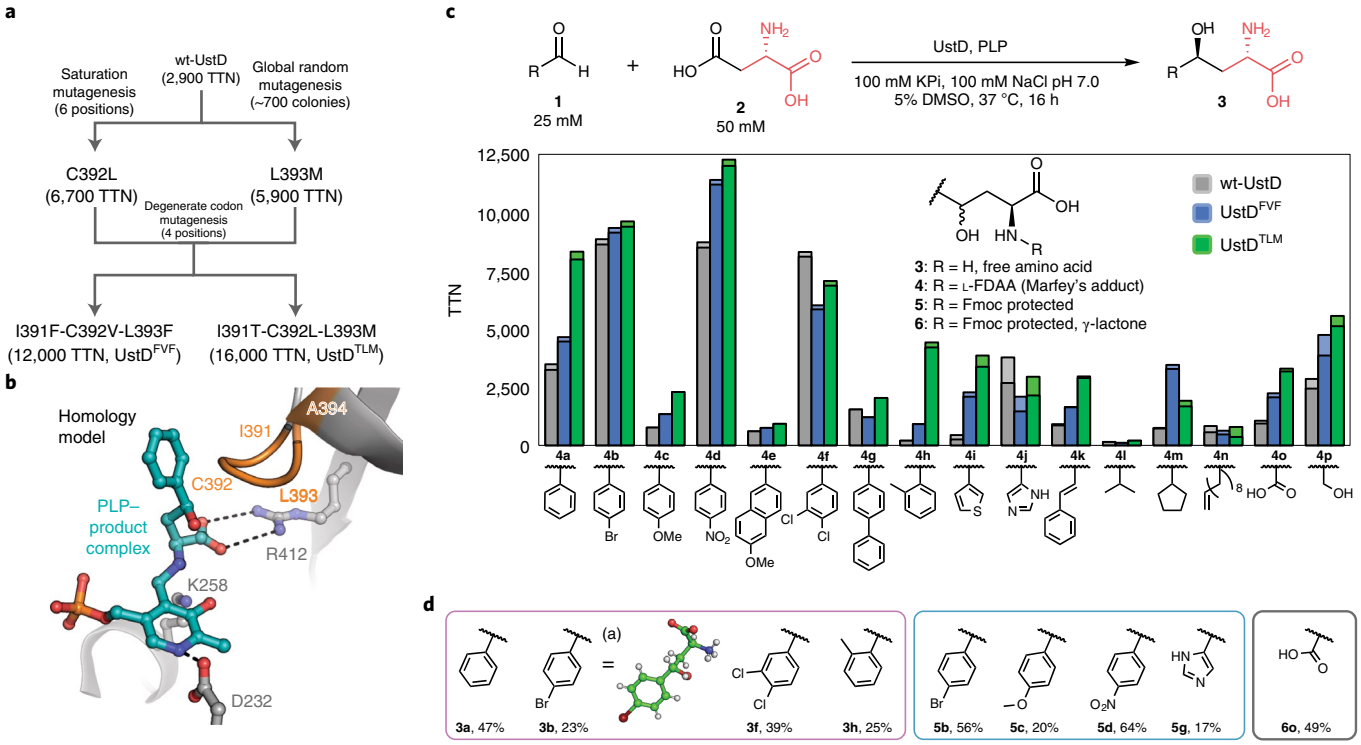

Fig. 2 | Directed evolution of UstD and the evaluation of variants. a, Lineage of activated UstD variants. Standard screening conditions: 25 mM **2a**, 50 mM **1**, buffer (100 mM KPi, pH 7.0, 100 mM NaCl), 5% dimethylsulfoxide (DMSO), 37 °C, 16 h. Catalyst activity was measured by the total turnover number (TTN). **b**, Computational model of UstD bound to **3a**, derived through homology modelling. Active site residues are shown as sticks and the loop residues targeted for mutagenesis are coloured in orange. Potential hydrogen bonds are shown as black dashes. **c**, Performance evaluation of UstD and activated variants measured by Marfey's analysis of the enzymatic products. Exact values and standard deviations are available in Supplementary Table 1 ($n = 3$ individual experiments per substrate and variant), and the error was generally below 10%. The bar sections in a lighter colour represent the amount of the other Cγ epimer from which the d.r. values are calculated. Absolute configuration was assigned by analogy with the product **3b** and the native Ustiloxin D stereochemistry[13] (see Supplementary Information for the details). **d**, Synthesis of select products at a 0.2 mmol scale with isolated yields. The different purification strategies are denoted by the different colours, free amino acid (purple), Fmoc-protected amino acid (blue) and lactonization with Fmoc protection (grey). The letter (a) denotes that the reactions from which **3b** was purified used wt-UstD.

supplementation of PLP led to a large increase in product formation (Supplementary Fig. 4). We did not observe a notable change when the concentration of **1** was increased (Supplementary Fig. 4). However, we observed the formation of L-Ala in the reactions, which indicates some **1** was lost to a non-productive protonation of the nucleophilic enamine intermediate[13]. We therefore used aldehyde as the limiting reagent and 2 equiv. **1** for subsequent experiments, which identified an optimal initial pH of 7.0 (Supplementary Fig. 4). Last, we varied the catalyst loading and found that UstD^{TLM} was capable of a high conversion (~70%) with just a 0.01 mol% catalyst loading (Supplementary Fig. 4). With these optimized conditions, we evaluated the performance of wt-UstD and both activated variants, UstD^{TLM} and UstD^{FVF}, with a more diverse set of aldehyde substrates. We anticipated that the striking sequence divergence in the putative loop would lead to distinct trends in substrate selectivity.

Performance analysis of UstD and its variants. Engineering enzymes for activity on a model substrate often leads to specialist catalysts with a diminished activity on substrate analogues[28,29]. The initial comparisons among wt-UstD, UstD^{FVF} and UstD^{TLM} with a small panel of aldehydes suggested that both variants had evolved towards an improved overall activity (Supplementary Fig. 5). We therefore expanded the substrate scope. Marfey's reagent cleanly derivatized the diverse enzymatic products to provide a uniform chromophore for the quantitative measurement of turnover and selectivity via UPLC–MS[30]. Product formation was observed with virtually every substrate tested, from the large and hydrophobic biphenyl aldehyde (**2g**) to the small and hydrophilic glycolaldehyde (**2p**) (Fig. 2c). Generally, the variant UstD^{TLM} performed the most turnovers and displayed an excellent diastereoselectivity, typically forming a d.r. of 95:5. Although UstD^{FVF} usually performed fewer turnovers than UstD^{TLM} with most substrates, UstD^{FVF} generally had a higher selectivity than wt-UstD or UstD^{TLM} (Supplementary Table 1). Reactions with *p*-substituted aromatic aldehydes exhibited a Hammett-like reactivity trend: more product was formed as the aldehyde electrophilicity increased. Activity was lowest with the electron rich *p*-anisaldehyde (**2c**), but a high activity was observed for the electron deficient *p*-NO_2-benzaldehyde (**2d**) with both engineered enzymes. To better capture the maximum turnover number (TON) with **2d**, we repeated the reactions at lower catalyst loadings, which revealed that the engineered variants can perform ~34,000 turnovers (Supplementary Fig. 6). Active-site mutagenesis had little apparent impact on reactions with some highly hydrophobic substrates, such as the methoxynaphthyl (**2e**), 3,4-dichlorobenzyl (**2f**) and biphenyl (**2g**) aldehydes; reactivity in these cases may be limited by poor aqueous solubility (Fig. 2c). In contrast, the reactivity of *o*-tolualdehyde (**2h**) and thiophene-3-carboxaldehyde (**2i**) increased dramatically during evolution. UstD^{TLM} displayed a ninefold increase in activity on **2i** and a remarkable 23-fold increase in turnovers with **2h** compared with those of wt-UstD. Activity with the imidazole substrate

2j was demonstrated and was one of the few substrates for which wt-UstD had the higher activity. To the best of our knowledge, the product is a previously unreported analogue of histidine. Reactivity with the cinnamaldehyde (**2k**) improved with both variants relative to that with wt-UstD. The reactions proceeded smoothly with several aliphatic substrates, which included isobutyraldehyde (**2l**), cyclopentylaldehyde (**2m**) and even 10-undecenal (**2n**); in the third case, the reactivity appeared to be limited by solubility. Pivaldehyde, however, was unreactive with all three enzymes, an observation we attribute to steric bulk near the carbonyl. The engineered UstD enzymes were active with glyoxylic acid (**2o**), which resulted in the formation of γ-hydroxyglutamate, an intermediate in hydroxyproline metabolism[31]. Last, we observed good reactivity with glycolaldehyde to yield the dihydroxylated amino acid **3p**. Previously, a protected form of **3p** was identified as a key intermediate in the synthesis of clavalanine (Fig. 1b)[17], an antibiotic that inhibits the biosynthesis of methionine[32]. Activity on **2p** increased twofold, with an improved diastereoselectivity and pristine enantioselectivity, for UstD^{TLM} relative to the wt enzyme. These substrates collectively demonstrate that the active site of UstD is remarkably permissive of diverse functional groups and that catalytic activity and selectivity can be rapidly optimized by mutation at residues 391–393.

These engineered enzymes enable a stereoselective synthesis of γ-hydroxy nsAAs in a single step from cheap, commercially available starting materials. The production of unprotected amino acids affords complete flexibility with regards to subsequent manipulation, but isolation of the free amino acids themselves is challenging due to their hydrophilic, zwitterionic nature. Therefore, we selected a representative set of products to demonstrate isolation strategies (Fig. 2d). Sufficiently hydrophobic products were isolated as the free amino acid, whereas for others we utilized protection with fluorenylmethoxycarbonyl (Fmoc) to increase the hydrophobicity, and simultaneously added a handle commonly used in solid-phase peptide synthesis. Diverse manipulations, such as lactonization with the γ-hydroxy group, can also be employed to facilitate isolation and downstream manipulation[19]. Throughout these reactions, a second, minor diastereomer was observed. The mixture of configurations at the γ-C arises through imperfect selectivity with the aldol addition and could be aggravated by reversible retro-aldol cleavage of the major diastereomer. We tested the latter possibility by resubjecting products **3a** and **3d** to the reaction conditions and observed no change in the d.r. by Marfey's analysis (Supplementary Fig. 7). However, in the case of **3a**, the formation of Ala was observed concomitant with a decrease in product peak area. This observation is consistent with slow product re-entry into the catalytic cycle via retro-aldol cleavage of **3a** to reform **2a** and Ala.

Linear regression guided protein engineering. The above studies relied on purified protein for preparative-scale reactions. However, access to enzymes in sufficient quantity is a common and often underappreciated limitation of biocatalysis. As is observed for many proteins, UstD had a relatively low expression titres in *E. coli* (8 mg l^{-1} culture) due to poor solubility (Supplementary Fig. 1). Although enzyme immobilization can be used to increase the utility of purified protein catalysts[33], a complementary synthetic methodology would use whole-cell preparations of UstD; this latter approach is attractive to process chemists[34]. Whole-cell catalysts are operationally simple to generate, stable over long periods and obviate the need for expensive protein purification.

We sought to further engineer UstD^{TLM} to increase the soluble heterologous expression in *E. coli* for whole-cell biocatalysis. This enzyme contains nine Cys residues, and our homology model suggested five are surface exposed (Supplementary Fig. 8). It is well known among protein crystallographers that removing surface Cys residues can increase the soluble expression and increase the probability of crystallization[35]. However, we found that the mutation of all five putative surface Cys residues to Ala eliminated catalytic activity. To identify mutations that would retain the activity while increasing the soluble expression, we performed sequence-similarity network analysis to identify non-Cys residues at these positions common among UstD homologues. Based on this analysis, we constructed a five-site degenerate codon library (Fig. 3a and Supplementary Fig. 8).

To efficiently navigate this sequence space, we employed linear regression modelling to predict sequence–activity relationships[36]. We hypothesized that this simple computational approach would be effective because the target residues are dispersed throughout the protein, which should make non-linear, pairwise mutational effects unlikely. We screened and sequenced 176 random clones from this library for increased activity in lysate, which is sensitive to changes in both soluble enzyme expression and enzymatic efficiency. Although most variants in this library were inactive, we were heartened to observe several apparently improved variants (Fig. 3a). Linear regression model testing using leave-one-out cross-validation of the full dataset indicated a poor predictive behaviour of the model for high-activity variants (Supplementary Fig. 9). We suspected that the model quality was diminished by the abundance of inactive variants, for which activity measurements are indistinguishable from experimental noise. We therefore restricted our analysis to variants for which bona fide activity could be measured, which left just 26 sequence–activity relationships. Despite the sparsity of these data (~5% of the sequence space), leave-one-out cross-validation showed the model was dramatically improved (see Supplementary Information for details).

We evaluated the three most active variants predicted by the model, UstD$^{TLM\text{-}ACASC}$, UstD$^{TLM\text{-}ASCSC}$ and UstD$^{TLM\text{-}ASASC}$. Comparisons of expression and whole-cell activity were made between these variants, the parent enzyme and the most active variant identified from screening, UstD$^{TLM\text{-}SCASC}$. We were delighted to find the expression titre increased relative to that for UstD^{TLM} for all the variants, up to 48 mg protein l^{-1} culture (Supplementary Fig. 10). Although purified enzyme activity is slightly decreased for the new variants, their overall activity in whole cells is substantially improved (Fig. 3a and Supplementary Fig. 10). Tests at the analytical scale showed, at a 0.25% w/v cell loading, that UstD^{TLM} formed **3a** with just a 13% yield, which highlights the challenges associated with translating in vitro activity into large-scale reaction formats. In contrast, the variant with the highest whole-cell activity, the computationally predicted UstD$^{TLM\text{-}ACASC}$ (designated UstD$^{v2.0}$), produced **3a** in a 31% yield, a 2.4-fold boost over that of UstD^{TLM} and a cumulative 15-fold boost over the wild type. Higher conversions were achieved by increasing the cell loading of UstD$^{v2.0}$ to 1% w/v, which afforded **3a** in 78% yield on an analytical scale (Fig. 3a). To demonstrate the utility of UstD$^{v2.0}$, large-scale reactions were carried out with **2a** and **2d**. The reaction with **2a** at a 0.5% w/v catalyst loading afforded 0.80 g of **3a** in a 77% isolated yield with pristine stereoselectivity after purification by reverse-phase chromatography. The reaction with **2d** at just a 0.1% w/v catalyst loading provided 1.4 g of **3d** in a 98% isolated yield with a high stereoselectivity (see Supplementary Information for details). Notably, these cell loadings are sufficient for process-scale biocatalytic reactions[37], which illustrates that UstD$^{v2.0}$ can operate on the scale needed to meet the demands of practical organic synthesis.

Crystallography of UstD$^{v2.0}$. Although the engineering we report here produced a generalist variant of UstD, structural information could guide more targeted engineering for the production of specific γ-hydroxy nsAAs. Despite extensive efforts, we were unable to produce crystals of wt-UstD. In contrast, UstD$^{v2.0}$ readily crystallized, which we attribute to the decrease in surface Cys residues. The 2.25 Å crystal structure of UstD$^{v2.0}$ was determined using experimental phases from a Au(III) derivative (Fig. 3b, Protein Data Bank ID

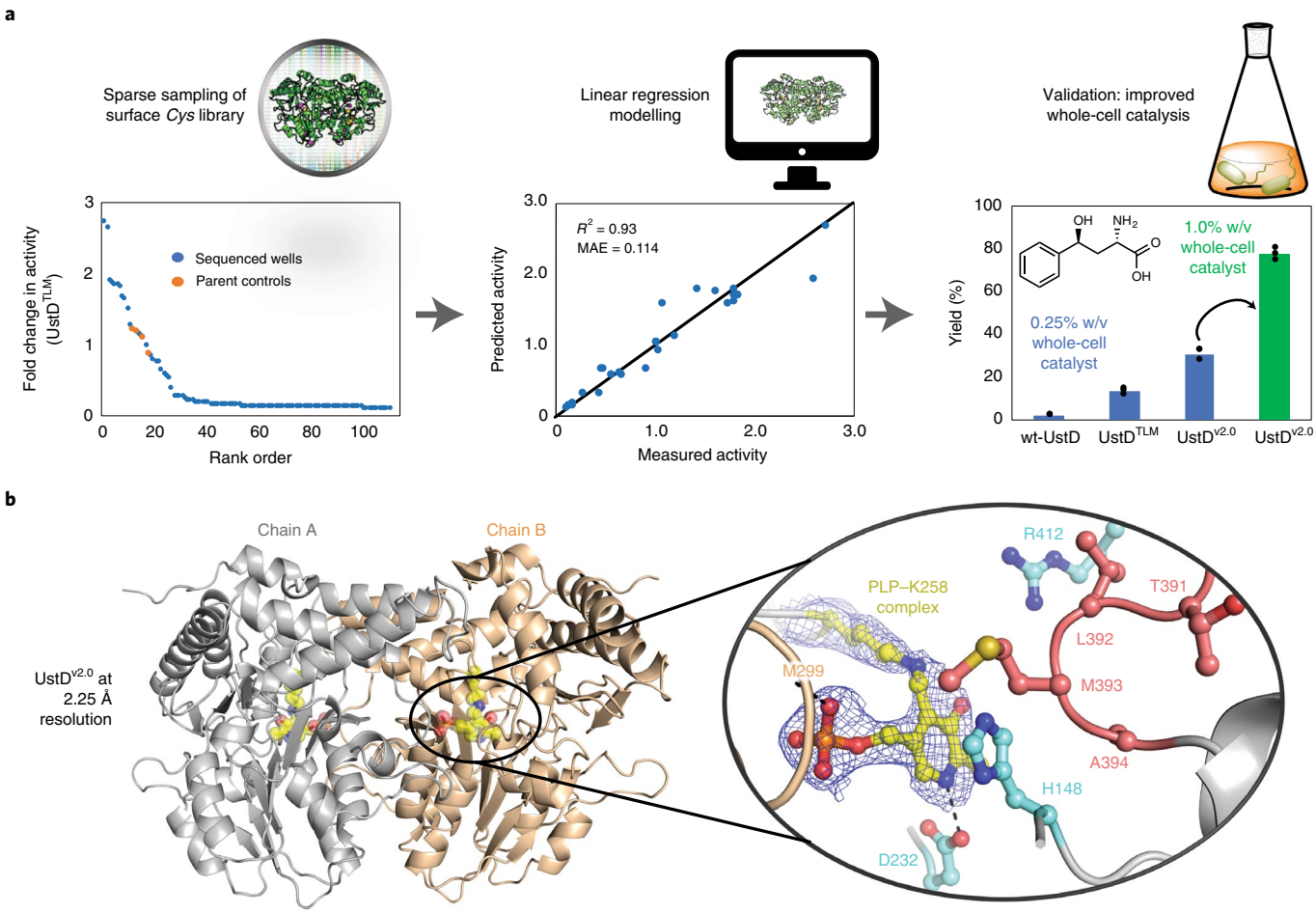

Fig. 3 | Engineering UstD for an increased crystallizability and activity in whole-cell catalysis. a, Experimental process for bioinformatic and regression-guided mutagenesis of UstD. In the first stage, a small mutagenesis library is sampled to collect sequence and/or activity data. The second stage builds a linear regression model to correlate sequences to activity. This regression model is then used to predict the activated sequences, which are validated in the last stage using whole-cell catalysis. The dots in the bar graph represent the individual measurements of triplicate technical replicates. **b**, Representation of the overall structure of UstD$^{2.0}$. Individual monomers are coloured grey (chain A) and brown (chain B). The PLP-K258 complex is shown as yellow spheres and sticks. Inset: active-site residues superimposed on the 2mFo–DFc electron density map (blue mesh, $\sigma = 1.2$) are shown as sticks. The TLMA loop residues are coloured pink. Hydrogen bonds are shown as black dashes. MAE, mean absolute error.

7MKV). This structure revealed an active site at the dimer interface, which is common among fold-type I PLP-dependent enzymes[38]. The internal aldimine that involves a Schiff base linkage to Lys258 and a salt bridge between the pyridinium N1 and Asp232 is clearly resolved in the active site. The 391–393 loop, which harbours the activating TLM mutations, projects over the top of the active site that forms part of the substrate binding pocket. The remainder of the pocket appears to be solvent exposed, which explains the tolerance of UstD for diverse aldehyde substrates (Supplementary Fig. 11).

In the future, we envision engineering UstD for increased activity with non-aldehyde substrates. As an initial demonstration, we showed that purified UstD$^{v2.0}$ performs ~50 turnovers with the ketone substrate trifluoroacetone to produce a nsAA that bears a tertiary alcohol side chain (Supplementary Fig. 12). The comparatively low turnover highlights the challenges associated with aldol addition into ketones. When nucleophilic attack is sufficiently slow, irreversible protonation of the enamine can quench the reactive intermediate and, indeed, we observed substantial accumulation of L-Ala in this reaction. A similar scenario was observed with the hydrolysis of an electrophilic PLP intermediate formed by TrpB and the reactions with attenuated substrates were enabled by directed evolution that increased the lifetime of the reactive intermediate[39,40]. Hence, future engineering to decrease the rate of enamine protonation in UstD$^{v2.0}$ may further expand the substrate scope.

Discussion

Here we improved a C–C bond-forming enzyme, UstD, that catalyses a decarboxylative aldol addition using the loss of CO_2 from L-Asp as a thermodynamic driving force to produce γ-hydroxy amino acids. This mechanism of action and the innate tolerance of diverse aldehydes marked UstD as a candidate for directed evolution into a versatile catalyst for organic synthesis. To screen for improved catalysts, we used a combination of globally random, site-saturation and degenerate codon mutagenesis libraries. We illustrated the engineering potential of the active site with two variants, UstD^{FVF} and UstD^{TLM}, that share no mutations in common and display a commensurate or superior activity to wt-UstD with the vast majority of aldehydes tested. We demonstrated how a simple regression-modelling approach to protein engineering can increase protein-soluble expression and crystallizability. The evolved variant, UstD$^{v2.0}$, is poised to deliver desirable nsAA precursors for medicinal chemistry, and the crystal structure will facilitate future work to explore the mechanism and reactivity of this intriguing enzyme.

Methods

All chemicals and reagents were purchased from commercial suppliers (Sigma-Aldrich, VWR, Chem-Impex International, Alfa Aesar, Combi-blocks and Oakwood Products) at the highest quality available and used without further purification unless stated otherwise. Genes were purchased as gBlocks from Integrated DNA Technologies. *E. coli* cells were electroporated with an Eppendorf E-porator at 2,500 V. New Brunswick I26R shaker incubators (Eppendorf) were used for cell growth. Cell disruption via sonication was performed with a Sonic Dismembrator 550 (Fisher Scientific) sonicator. Ultraviolet–visible spectroscopic measurements were collected on a UV-2600 Shimadzu spectrophotometer. Optical density measurements were collected using an optical density reader (Amersham Biosciences). UPLC–MS data were collected on an Acquity UPLC (Waters) equipped with an Acquity PDA and QDA MS detector using either a BEH C18 column (Waters) for the substituted benzaldehyde reactions, or an Intrada Amino Acid column (Imtakt) for the aliphatic aldehyde reactions. All UPLC–MS data were processed using Empower 3 (Waters). Preparative column separations were performed on an Isolera One Flash Purification system (Biotage). NMR data were collected on Bruker 400 or 500 MHz spectrometers equipped with BBFO and DCH cryoprobes, respectively. All NMR chemical shifts were referenced either to a residual solvent peak or tetramethylsilane internal standard. Spectra recorded using DMSO-d_6 were referenced to the residual DMSO signal at 2.5 ppm for ^{1}H and 39.52 ppm for ^{13}C NMR analysis. Spectra recorded using $CDCl_3$ were referenced to the residual $CHCl_3$ peak at 7.26 ppm for ^{1}H and 77.16 ppm for ^{13}C NMR spectroscopy. Spectra recorded using CD_3OD were referenced to the CH_3OD residual solvent peak at 3.31 ppm for ^{1}H and 49.00 ppm for ^{13}C NMR analysis. Spectra recorded using D_2O:acetonitrile-d_3 as the solvent were referenced to the residual H_2O signal at 4.79 ppm for ^{1}H and absolute referenced to the ^{1}H spectrum for ^{13}C NMR analysis. Signal positions were recorded in ppm with the abbreviations s, d, t, q, dd and m denoting singlet, doublet, triplet, quartet, doublet of doublets and multiplet, respectively. All the coupling constants *J* were measured in Hertz. High-resolution mass spectrometry data were collected with a Q Extractive Plus Orbitrap (NIH 1S10OD020022-1) instrument with samples ionized by electrospray ionization.

Cloning of wt-UstD. A codon-optimized copy of the *Aspergillus flavus* UstD gene was purchased as a gBlock from Integrated DNA Technologies. This DNA fragment was inserted into a pET-22b(+) vector by the Gibson Assembly method[41] and transformed into electrocompetent BL21(DE3) *E. coli* cells via electroporation. After a 30 min recovery period in Luria–Burtani (LB) media, cells were plated onto LB plates that contained 100 μg ml^{-1} Amp (LB_{Amp} plates) and incubated overnight. A single colony was then used to inoculate 50 ml of Terrific Broth II media that contained 100 μg ml^{-1} Amp (TB_{Amp}), which was then incubated overnight at 37 °C with shaking at 200 r.p.m. A 500 μl aliquot of the saturated cell culture was then mixed with 500 μl of sterile 80% glycerol and snap frozen in liquid nitrogen to generate a glycerol stock.

Plasmid preparations. A 5 ml overnight culture of *E. coli* that harboured the plasmid of interest was grown overnight at 37 °C with shaking at 200 r.p.m. The plasmid was isolated and purified using Zymo Plasmid Miniprep kits and sequenced through Functional Biosciences.

Protein expression. An overnight culture of *E. coli* BL21(DE3) that harboured a pET-22b(+) plasmid encoding a given UstD variant was created by inoculating 50 ml of TB_{Amp} media with a single colony. This culture was shaken at 37 °C and 200 r.p.m. for ~16 h. A 10 ml aliquot of the overnight culture was then used to inoculate 1 l of TB_{Amp}, which was shaken at 37 °C and 200 r.p.m. for approximately 1.5 h or until an optical density of 0.4–0.6 was reached. Cultures were removed from the incubator and cooled on ice for 30 min, followed by induction with 100 μM isopropyl β-D-1-thiogalactopyranoside. The cultures were allowed to continue to grow for about an additional 16 h at 20 °C and shaken at 200 r.p.m. Cells were then harvested by centrifugation (4 °C, 30 min, 4,000*g*), and the cell pellets were stored at −20 °C overnight.

Whole cell preparation of *E. coli* that harboured UstD and variants. After protein expression, cells were harvested by centrifugation (4 °C, 30 min, 4,000*g*). The cell pellets were then resuspended in water and centrifuged twice to remove all media. The cell pellets were transferred to 50 ml conical tubes and freeze dried by lyophilization. The dried cells were stored at −80 °C until further use.

Protein purification of UstD and variants. To purify UstD, cell pellets were thawed on ice and then resuspended in lysis buffer, which comprised enzyme storage buffer (100 mM potassium phosphate buffer, pH 7.0, 100 mM sodium chloride) that contained 20 mM imidazole, 1 mg ml^{-1} Hen Egg White Lysozyme (GoldBio), 0.2 mg ml^{-1} DNase (GoldBio), 1 mM $MgCl_2$ and 150 μM PLP. A ratio of 4 ml of lysis buffer per gram of wet cell pellet was used. Cells lysis began by shaking for 1 h at 37 °C. The resuspended cells were subsequently sonicated (20 min, 0.8 s on, 0.2 s off, power setting 5). The resulting lysate was then spun down at 75,600*g* to pellet the cellular debris. Ni/NTA beads were pre-equilibrated in storage buffer that contained 20 mM imidazole. A 1 ml aliquot of resin per 25 g of cells was added to the cleared lysis supernatant and incubated with nutation on ice for 1 h. The beads were then collected in a gravity column with a plastic frit, and the flow through was repassed once to collect any remaining beads from the original vessel. The collected beads were washed with 10–20 column volumes of storage buffer that contained 60 mM imidazole. Protein was eluted with 5 ml of storage buffer that contained 250 mM imidazole and the flow through collected until the eluent was no longer yellow (the colour is due to the enzymatically bound PLP cofactor). The eluent was then transferred to a centrifugal filter tube (Amicon Ultra-15, 30k MWCO) and concentrated by centrifugation (4000*g*, 15 min). Imidazole was then removed either through dialysis or through repeated dilution (with enzyme storage buffer) and concentration steps until <1 μM imidazole.

Generation of random mutagenesis libraries. Random mutagenesis was carried out via error-prone PCR. The reaction conditions were optimized to generate 1–2 codon mutations per plasmid. Reactions were set up by adding the following to a PCR tube: 5 μl 10X Taq buffer (New England Biolabs), 1 μl of a 10 mM dNTP mix, 1 μl of 10 μM 22b-intF, 1 μl of 10 μM 22b-intR, 1 μl of the ~100 ng μl^{-1} parent plasmid, 5.5 μl of 50 mM $MgCl_2$, 2.5 or 5 μl of 1 mM $MnCl_2$, 1 μl of DMSO, 0.5 μl of Taq polymerase (New England Biolabs) and the total volume was made up to 50 μl with H_2O.

The PCR product was purified using a preparative agarose gel. Purified DNA fragment was inserted into a pET-22b(+) vector by the Gibson Assembly method[41]. BL21 (DE3) *E. coli* cells were subsequently transformed with the resulting cyclized DNA product via electroporation. After 45 min of recovery in LB media that contained 0.4% glucose at 37 °C, cells were plated onto LB_{Amp} plates and incubated overnight. Single colonies were used to inoculate 5 ml LB_{Amp} plates, which were grown overnight at 37 °C with shaking at 200 r.p.m. Colonies were sequenced and there were 1–2 coding mutations for both the concentrations of $MnCl_2$.

Protein engineering (library expression, screening and validation). Electrocompetent BL21(DE3) cells were transformed with mutagenized plasmid DNA and allowed to recover for 45 min in 800 μl of TB. After recovery, the cells were plated onto LB_{Amp} plates and incubated overnight. A 96-well plate that contained 500 μl of TB_{Amp} per well was inoculated with single colonies. Each plate included parent positive controls (from a fresh transformation), negative controls and a sterile control that was not inoculated. The plates were grown overnight at 37 °C with shaking at 200 r.p.m. Expression plates were prepared with 630 μl of TB_{Amp} per well and inoculated with 20 μl of overnight culture. Glycerol stocks of each starter plate well were made from the remaining culture to ensure the sequence of any mutants of interest could be determined. The expression cultures were grown at 37 °C with shaking at 200 r.p.m. for 2.5 h. Expression plates were then placed on ice for 30 min and induced with a final concentration of 0.1 mM isopropyl β-D-1-thiogalactopyranoside in 50 μl of fresh TB_{Amp}. The expression culture was grown overnight at 20 °C with shaking at 200 r.p.m. After overnight growth, the plate was centrifuged (4,000*g*, 30 min, 4 °C) and all the media was removed by striking plates against a paper towel on a table. Expression plates were stored at −20 °C until further use.

A lysis buffer that contained a 100 mM potassium phosphate buffer (pH 7.0), 100 mM sodium chloride, 1 mg ml^{-1} Hen Egg White Lysozyme (GoldBio), 0.2 mg ml^{-1} DNase (GoldBio), 1 mM $MgCl_2$ and 150 μM PLP was added to each well and the plate was subsequently lysed for 1 h at 37 °C. The lysate was pelleted at 4,000*g* for 30 min. Clarified lysate was added to a 96-well reaction plate in which each well contained a master mix solution, such that the end reaction concentrations were 25 mM aldehyde, 25 mM L-Asp, PLP and buffer (100 mM KPi + NaCl, pH 7.0). The ratio of clarified lysate to reaction master mix was varied over the course of the engineering to maintain a reasonable product measurement dynamic range. The reactions were allowed to incubate overnight at 37 °C, and were subsequently quenched with 100 μl of acetonitrile (ACN) and pelleted at 4,000*g* for 30 min. The cleared reaction mixture was transferred to a 0.2 μm centrifuge filter plate (PALL) and filtered at 1,500*g* for 10 min into a clean 96-well plate before being sealed prior to analysis by UPLC–MS.

The relative amount of product formed in the reactions compared with that in the positive control reaction was measured by absorbance at 210 nm via UPLC–MS. Given the relatively high variability in the parent signal in this assay, wells typically required an apparent 1.5-fold increase in product compared with that of the parent to be carried forward for the validation of hits. Using the glycerol stocks from the starter culture plate (described above), the wells of interest could be streaked on to a fresh LB_{Amp} plate for subsequent sequencing and validation.

Every mutant of interest was validated by heterologous expression and Ni-NTA purification, which accounted for changes in the soluble enzyme concentration as well as changes in the activity. To study how the activity profile of UstD changed over the course of engineering, each key variant in the evolutionary lineage was expressed and purified in tandem, as described above (Supplementary Fig. 3). Parallel triplicate 200 μl reactions that contained 25 mM benzaldehyde, 50 mM L-Asp sodium salt monohydrate, 2.5 μM PLP and 0.25 μM UstD variant (0.001 mol% catalyst, 100,000 maximum TON) were allowed to react at 37 °C for 16 h. Afterwards, each reaction was quenched with 200 μl of ACN that contained 1 mM tryptamine as an internal standard, and the reaction mixtures were analysed by UPLC–MS. A standard curve was made using previously purified **3a** to

facilitate the TTN calculations. The variants were also trialled against several other aldehydes, which included biphenyl-4-carboxaldehyde (20,000 maximum TON), *p*-anisaldehyde (20,000 maximum TON) and glycolaldehyde (100,000 maximum TON). Reactions were run using the same reaction conditions and procedure, with catalyst loading changed to match the indicated maximum TON. Simple fold-response measurements were used to quantify the activity differences between variants (Supplementary Fig. 5).

UstD^{TLM} reaction condition optimization. All the optimization reactions were conducted in triplicate on an analytical scale (100 µl). PLP and L-Asp stock solutions were made with a 100 mM potassium phosphate buffer that contained 100 mM sodium chloride (reaction buffer) at the indicated pH. Postreaction quenching was done by adding 100 µl of 99:1 ACN:ethanol with 1 mM tryptamine as an internal standard. Quenched reactions were then centrifuged at 15,000*g* to remove aggregated protein, and diluted with 200 µl of 1:1 water:ACN. Quantification was performed by UPLC–MS analysis. Measurements of the internal standard, benzaldehyde and product concentrations was done by separation on a BEH C18 column (Waters) and measurement of the corresponding 210 nm ultraviolet peak areas. Measurements of the internal standard, product, L-Asp and L-Ala concentrations were done by separation on an Intrada amino acids column (Imtakt) using a positive-mode single-ion readout for the M + H mass peak. Variability in the injection volumes was corrected by dividing peak areas by the observed internal standard peak area for each injection. Optimization for each reaction condition component is listed below.

PLP concentration. A reaction master mix that contained 27.5 mM L-Asp monosodium monohydrate, 27.5 mM benzaldehyde and 5.5% DMSO was made in a 100 mM potassium phosphate buffer (pH 8.0) with a 100 mM sodium chloride reaction buffer. Stocks of PLP (1.00, 0.20, 0.08 and 0.02 mM) were made by diluting a 20 mM PLP stock solution in the reaction buffer. Glass vials (0.5 dram) were charged with 90.9 µl of the reaction master mix and 5 µl of the appropriate PLP stock (or buffer, in the case of no added PLP), and catalysis was initiated by the addition of 4.1 µl of 25 µM UstD^{TLM} (final concentrations: 25 mM L-Asp, 2.5 µmol; 25 mM benzaldehyde, 2.5 µmol; 1 µM UstD^{TLM}, 0.004 mol% catalyst, 25,000 maximum TON; 50 µM, 10 µM, 4 µM, 1 µM or 0 µM PLP; 5% DMSO). Reactions were allowed to proceed in a 37 °C incubator for 16 h prior to quenching with 100 µl of ACN and quantification (Supplementary Fig. 4a).

L-Asp concentration. A reaction master mix that contained 55.6 mM benzaldehyde, 111.1 µM PLP and 11.1% DMSO was made in a 100 mM potassium phosphate buffer (pH 8.0) with a 100 mM sodium chloride reaction buffer. Stocks of L-Asp monosodium monohydrate (500, 250, 100 and 50 mM) were made in the reaction buffer. Glass vials (0.5 dram) were charged with 45 µl of the reaction master mix and 50 µl of the appropriate L-Asp stock, and catalysis was initiated by the addition of 5 µl of 5 µM UstD^{TLM} (final concentrations: 25 mM benzaldehyde, 2.5 µmol; 25, 50, 125 and 250 mM L-Asp, 2.5, 5.0, 12.5 and 25. µmol, respectively; 0.25 µM UstD^{TLM}, 0.001 mol% catalyst, 100,000 maximum TON; 2.5 µM PLP, 10 equiv. relative to UstD^{TLM}; 5% DMSO). Reactions were allowed to proceed in a 37 °C incubator for 16 h prior to quenching with 100 µl of ACN and quantification (Supplementary Fig. 4b).

pH. Five separate master mix solutions that contained 25 mM benzaldehyde, 130 mM L-Asp monosodium monohydrate, 1.3 mM PLP and 5.2% DMSO were prepared in 100 mM potassium phosphate with a 100 mM NaCl reaction buffer at pH 6.0, 6.5, 7.0, 7.5 and 8.0 (pH of the buffer was not altered after the addition of the reaction components). Glass vials (0.5 dram) were charged with 96.1 µl of the appropriate reaction master mix, and catalysis was initiated by the addition of 3.9 µl of 6 µM UstD^{TLM} (final concentrations: 25 mM benzaldehyde, 2.5 µmol; 125 mM L-Asp, 12.5 µmol; 0.25 µM UstD^{TLM}, 0.001 mol% catalyst, 100,000 maximum TON; 2.5 µM PLP, 10 equiv. relative to UstD^{TLM}; 5% DMSO). Reactions were allowed to proceed in a 37 °C incubator for 16 h prior to quenching with 100 µl of ACN and quantification (Supplementary Fig. 4c).

Catalyst loading. A reaction master mix that contained 50 mM benzaldehyde, 100 mM L-Asp monosodium monohydrate and 10% DMSO was made in 100 mM potassium phosphate, pH 7.0, with 100 mM sodium chloride. UstD^{TLM} stock solutions that contained 50 µM, 5.0 µM, 1.7 µM and 1 µM were made, each of which contained 10 equiv. PLP. Glass vials (0.5 dram) were charged with 50 µl of reaction master mix, and catalysis was initiated by the addition of 50 µl of the appropriate UstD stocks (final concentrations: 25 mM benzaldehyde, 2.5 µmol; 50 mM L-Asp, 5.0 µmol; 25 µM (0.1 mol% catalyst, 1,000 maximum TON), 2.5 µM (0.01 mol% catalyst, 10,000 maximum TON), 0.83 µM (0.003 mol% catalysis, 30,000 maximum TON), 0.25 µM (0.001 mol% catalyst, 100,000 maximum TON) UstD^{TLM}; 10 equiv. PLP relative to UstD^{TLM}; 5% DMSO). Reactions were allowed to proceed in a 37 °C incubator for 16 h prior to quenching with 100 µl of ACN and quantification (Supplementary Fig. 4D).

UstD performance evaluation using Marfey's derivatization. A 0.5 dram glass vial was charged with a master mix of L-Asp sodium salt monohydrate (0.005 mmol, 2 equiv., 50 mM final concentration), PLP (10 equiv. relative to the final UstD concentration) and buffer. The master mix composition was varied to ensure a uniform concentration of each UstD variant at the completion of the reaction set-up. To this solution the aldehydes that correspond to compounds **2a**–**2p** (0.0025 mmol, 1 equiv., 25 mM final concentration) were added to the reaction mixtures. The reactions were initiated by the addition of UstD (0.007 mol% catalyst, 15,000 maximum TON). The reaction vessels were placed in a dark 37 °C incubator for 18 h and subsequently quenched with 200 µl of ACN. A Marfey's derivatization reaction was then performed to determine the e.e. and d.r. of each enzymatic reaction. In a new flat-bottom glass LC vial, 6 µl of a quenched reaction mix (1 equiv., 0.5 mM final total amines from unreacted L-Asp and formed L-Ala and the γ-hydroxy amino acid product) was added to a solution of 144 µl of 10.41 mM $NaHCO_3$ (10 equiv., 5 mM final concentration) and 0.21 mM of either L-Arg (0.1 mM final concentration, aldehydes **2a**–**2k**) or tryptamine (0.1 mM final concentration, aldehydes **2l**–**2p**), followed by the addition of 150 µl of 5 mM L-FDAA (Marfey's reagent) dissolved in ACN (5 equiv., 2.5 mM final concentration) to bring the total reaction volume to 300 µl. Each reaction vial was sealed with a pierceable LC vial cap, placed in a dark 37 °C incubator for 18 h and then quenched with 300 µl of 1:1 ACN:60 mM HCl (15 mM postquench). Quenched reaction mixtures were analysed by UPLC–MS no later than 24 h after quenching; the results are shown in Supplementary Table 1 and Supplementary Figs. 13–28.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The structure of UstD$^{v2.0}$ is available through the Protein Data Bank ID 7MKV. The sequence-activity data used for linear regression modelling is available through GitHub[42]. All the other data are available from the authors upon reasonable request.

Code availability

The linear regression modelling code used during the final round of protein engineering is available through GitHub[42] under the MIT License.

Received: 16 August 2021; Accepted: 6 January 2022;
Published online: 21 February 2022

References

1. Nestl, B. M., Hammer, S. C., Nebel, B. A. & Hauer, B. New generation of biocatalysts for organic synthesis. *Angew. Chem. Int. Ed.* **53**, 3070–3095 (2014).
2. Zhang, X. et al. Divergent synthesis of complex diterpenes through a hybrid oxidative approach. *Science* **369**, 799–806 (2020).
3. Brown, D. G. & Boström, J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? *J. Med. Chem.* **59**, 4443–4458 (2016).
4. Fesko, K. & Gruber-Khadjawi, M. Biocatalytic methods for C–C bond formation. *ChemCatChem* **5**, 1248–1272 (2013).
5. Schmidt, N. G., Eger, E. & Kroutil, W. Building bridges: biocatalytic C–C-bond formation toward multifunctional products. *ACS Catal.* **6**, 4286–4311 (2016).
6. Fujii, I. Heterologous expression systems for polyketide synthases. *Nat. Prod. Rep.* **26**, 155–169 (2009).
7. Heine, A. et al. Observation of covalent intermediates in an enzyme mechanism at atomic resolution. *Science* **294**, 369–374 (2001).
8. Wang, Z. J. et al. Improved cyclopropanation activity of histidine-ligated cytochrome P450 enables the enantioselective formal synthesis of levomilnacipran. *Angew. Chem. Int. Ed.* **53**, 6810–6813 (2014).
9. Berkeš, D., Kolarovič, A., Manduch, R., Baran, P. & Považanec, F. Crystallization-induced asymmetric transformations (CIAT): stereoconvergent acid-catalyzed lactonization of substituted 2-amino-4-aryl-4-hydroxybutanoic acids. *Tetrahedron Asymm.* **16**, 1927–1934 (2005).
10. Goldberg, S. L. et al. Preparation of β-hydroxy-α-amino acid using recombinant D-threonine aldolase. *Org. Process Res. Dev.* **19**, 1308–1316 (2015).
11. Steinreiber, J. et al. Overcoming thermodynamic and kinetic limitations of aldolase-catalyzed reactions by applying multienzymatic dynamic kinetic asymmetric transformations. *Angew. Chem. Int. Ed.* **46**, 1624–1626 (2007).
12. Zetzsche, L. E. & Narayan, A. R. H. Broadening the scope of biocatalytic C–C bond formation. *Nat. Rev. Chem.* **4**, 334–346 (2020).
13. Ye, Y. et al. Unveiling the biosynthetic pathway of the ribosomally synthesized and post-translationally modified peptide ustiloxin B in filamentous fungi. *Angew. Chem. Int. Ed.* **55**, 8072–8075 (2016).
14. Prier, C. K. & Arnold, F. H. Chemomimetic biocatalysis: exploiting the synthetic potential of cofactor-dependent enzymes to create new catalysts. *J. Am. Chem. Soc.* **137**, 13992–14006 (2015).
15. Di Salvo, M. L. et al. On the catalytic mechanism and stereospecificity of *Escherichia coli* L-threonine aldolase. *FEBS J.* **281**, 129–145 (2014).

16. Marsden, S. R., Gjonaj, L., Eustace, S. J. & Hanefeld, U. Separating thermodynamics from kinetics—a new understanding of the transketolase reaction. *ChemCatChem* **9**, 1808–1814 (2017).
17. Ariza, J., Font, J. & Ortuño, R. M. An efficient and concise entry to (–)-4,5-dihydroxy-D-threo-L-norvaline. Formal synthesis of clavalanine. *Tetrahedron Lett.* **32**, 1979–1982 (1991).
18. Blaskovich, M. A. T. Unusual amino acids in medicinal chemistry. *J. Med. Chem.* **59**, 10807–10836 (2016).
19. Moreno, C. J. et al. Synthesis of γ-hydroxy-α-amino acid derivatives by enzymatic tandem aldol addition–transamination reactions. *ACS Catal.* **11**, 4660–4669 (2021).
20. Hernandez, K. et al. Combining aldolases and transaminases for the synthesis of 2-amino-4-hydroxybutanoic acid. *ACS Catal.* **7**, 1707–1711 (2017).
21. Vargas-Rodriguez, O., Sevostyanova, A., Söll, D. & Crnković, A. Upgrading aminoacyl-tRNA synthetases for genetic code expansion. *Curr. Opin. Chem. Biol.* **46**, 115–122 (2018).
22. Marchand, J. A. et al. Discovery of a pathway for terminal-alkyne amino acid biosynthesis. *Nature* **567**, 420–424 (2019).
23. Yang, J. et al. The I-TASSER suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2014).
24. Ho, T. H. et al. Catalytic intermediate crystal structures of cysteine desulfurase from the Archaeon *Thermococcus onnurineus* NA1. *Archaea* **2017**, 1–11 (2017).
25. Kumar, P. et al. L-Threonine transaldolase activity is enabled by a persistent catalytic intermediate. *ACS Chem. Biol.* **16**, 95 (2021).
26. Reetz, M. T., Prasad, S., Carballeira, J. D., Gumulya, Y. & Bocola, M. Iterative saturation mutagenesis accelerates laboratory evolution of enzyme stereoselectivity: rigorous comparison with traditional methods. *J. Am. Chem. Soc.* **132**, 9144–9152 (2010).
27. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
28. Reetz, M. T., Bocola, M., Carballeira, J. D., Zha, D. & Vogel, A. Expanding the range of substrate acceptance of enzymes: combinatorial active-site saturation test. *Angew. Chem. Int. Ed.* **44**, 4192–4196 (2005).
29. Romney, D. K., Sarai, N. S. & Arnold, F. H. Nitroalkanes as versatile nucleophiles for enzymatic synthesis of noncanonical amino acids. *ACS Catal.* **9**, 8726–8730 (2019).
30. Marfey, P. Determination of D-amino acids. II. Use of a bifunctional reagent, 1,5-difluoro-2,4-dinitrobenzene. *Carlsberg Res. Commun.* **49**, 591–596 (1984).
31. Wu, G. et al. Proline and hydroxyproline metabolism: implications for animal and human nutrition. *Amino Acids* **40**, 1053–1063 (2011).
32. Müller, J.-C., Toome, V., Pruess, D. L., Blount, J. F. & Weigele, M. Ro 22-5417, a new clavam antibiotic from *Streptomyces clavuligerus*. III Absolute stereochemistry. *J. Antibiot.* **36**, 217–225 (1983).
33. Wahab, R. A., Elias, N., Abdullah, F. & Ghoshal, S. K. On the taught new tricks of enzymes immobilization: an all-inclusive overview. *React. Func. Pol.* **152**, 104613 (2020).
34. Wachtmeister, J. & Rother, D. Recent advances in whole cell biocatalysis techniques bridging from investigative to industrial scale. *Curr. Opin. Biotechnol.* **42**, 169–177 (2016).
35. Al-Ayyoubi, M., Gettins, P. G. W. & Volz, K. Crystal structure of human maspin, a serpin with antitumor properties: reactive center loop of maspin is exposed but constrained. *J. Biol. Chem.* **279**, 55540–55544 (2004).
36. Fox, R. Directed molecular evolution by machine learning and the influence of nonlinear interactions. *J. Theor. Biol.* **234**, 187–199 (2005).
37. Huffman, M. A. et al. Design of an in vitro biocatalytic cascade for the manufacture of islatravir. *Science* **366**, 1255–1259 (2019).
38. Eliot, A. C. & Kirsch, J. F. Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. *Annu. Rev. Biochem.* **73**, 383–415 (2004).
39. Romney, D. K., Murciano-Calles, J., Wehrmüller, J. E. & Arnold, F. H. Unlocking reactivity of TrpB: a general biocatalytic platform for synthesis of tryptophan analogues. *J. Am. Chem. Soc.* **139**, 10769–10776 (2017).
40. Boville, C. E. et al. Engineered biosynthesis of β-alkyl tryptophan analogues. *Angew. Chem. Int. Ed.* **57**, 14764–14768 (2018).
41. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
42. Ellis, J. M. Linear regression analysis of UstD-TLM. *Zenodo* https://doi.org/10.5281/zenodo.5719389 (2021).

Acknowledgements

We thank I. Guzei for small-molecule X-ray structure determinations and S.H. Gellman and members of the Buller group for critical reading of the manuscript. The crystal mounting and data collection were mediated by the Collaborative Crystallography Core, Department of Biochemistry, UW–Madison, and data were collected at the Life Sciences Collaborative Access Team beamline 21ID-D at the Advanced Photon Source, Argonne National Laboratory, and we thank Z. Wawrzak for technical assistance during data collection. Use of LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (grant 085P1000817). This work was supported by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison, Wisconsin Alumni Research Foundation, National Institute of Health (grant DP2-GM137417, A.R.B.), Morgridge Institute for Research—Metabolism Theme Fellowship (P.K.) and the NIH Biotechnology Training Grant (T32-GM008349, J.M.E.). The Bruker AVANCE III-500 NMR spectrometers were supported by the Bender Fund. The Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract no. W-31-109-Eng-38. The Bruker D8 VENTURE Photon III X-ray diffractometer was partially funded by a NSF Award (no. CHE-1919350) to the UW–Madison Department of Chemistry.

Author contributions

A.R.B. and J.M.E. conceptualized the goals and aims of the project. J.M.E., M.E.C., P.K., E.P.G., C.A.B. and A.R.B. carried out the development of the chemistry and enzymes. J.M.E. developed the code for data analysis and developed the linear regression model. J.M.E. and M.E.C. verified the results. J.M.E., M.E.C., P.K. and A.R.B. prepared the figures and data visualizations. A.R.B. secured funding for the project that led to this publication. A.R.B. coordinated team members for the development of the chemistry and enzyme evolution. C.A.B. supervised the data acquisition of protein crystals that led to the resolved crystal structure. A.R.B. supervised the research activity planning and execution. J.M.E., M.E.C. and A.R.B. prepared the initial manuscript. J.M.E., M.E.C., P.K. and A.R.B. reviewed and edited the initial manuscript and provided critical commentary and revisions.

Competing interests

A.R.B., J.M.E. and P.K. have a patent pending on the use of engineered UstD for the synthesis of nsAAs, US Patent application no. 20210115480A1. All other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41929-022-00743-0.

Correspondence and requests for materials should be addressed to Andrew R. Buller.

Peer review information *Nature Catalysis* thanks the anonymous reviewers for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Andrew R. Buller

Last updated by author(s): Nov 29, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Empower 3, CCP4 7.1, XDS Build 20210205, Coot 0.9.6
Data analysis	SnapGene 5.3.1, Python 3.8.6, scikit-learn 0.23.2, https://doi.org/10.5281/zenodo.5719389

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The dataset generated during and/or analysed during the current study are available in the Buller Lab GitHub repository, https://github.com/bullerlaboratory/NatCatalUstDPaper. The UstD2.0 crystal structure is available via the Protein Data Bank, PDB-ID: 7MKV

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Triplicate measurements were made when quantitative analysis of individual activities was determined for standard deviation calculations. Sampling of protein sequence space was done with single measurements, as screening of protein engineering libraries is not conducive to replicate measurements, nor do replicate measurements typically impact discovery of activated variants
Data exclusions	Measurements indistinguishable from noise were excluded from the final model.
Replication	Claims were verified by experimental replicates. No claims were made which were not validated by multiple replicate experiments (triplicate or higher, in most cases)
Randomization	N/A, there are no claims in which randomization would be considered standard practice or beneficial
Blinding	N/A, there are no claims where blinding would have been applicable as no test subjects were used.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging