(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
06 November 2025 (06.11.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/230736 A1

(51) International Patent Classification:
C12N 9/10 (2006.01) *C12P 13/12* (2006.01)

(21) International Application Number:
PCT/US2025/025073

(22) International Filing Date:
17 April 2025 (17.04.2025)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/640,709 30 April 2024 (30.04.2024) US

(71) Applicant: **WISCONSIN ALUMNI RESEARCH FOUNDATION** [US/US]; 614 Walnut Street, 13th Floor, Madison, Wisconsin 53726 (US).

(72) Inventors: **BULLER, Andrew**; 210 Forest Street, Madison, Wisconsin 53726 (US). **CAMPBELL, Meghan**; 611 San Luisito Way Unit 10, Sunnyvale, California 94085 (US). **OHLER, Amanda**; 2872 Pleasant View Rd. Unit 201, Middleton, Wisconsin 53562 (US). **MCGILL, Matthew**; 1736 Old Hwy C, Saint Germain, Wisconsin 54558 (US).

(74) Agent: **LEONE, Joseph** et al.; 25 W. Main Street, Suite 800, Madison, Wisconsin 53703 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UY, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

(54) Title: ENGINEERED BIOCATALYSTS FOR THE SYNTHESIS OF GAMMA-HYDROXY AMINO ACIDS

(57) Abstract: The present disclosure is directed to UstD mutants capable of producing non-canonical amino acids that contain secondary or tertiary alcohols at the gamma position, and methods of using the mutants to produce the non-canonical amino acids. The mutants can use aldehyde-containing substrates and ketone-containing substrates to generate the secondary and tertiary alcohol groups, respectively.

WO 2025/230736 A1

ENGINEERED BIOCATALYSTS FOR THE SYNTHESIS OF GAMMA-HYDROXY AMINO ACIDS

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

This invention was made with government support under GM137417 awarded by the National Institutes of Health. The government has certain rights in the invention.

SEQUENCE LISTING

The instant application contains a Sequence Listing which has been submitted in an XML file with the USPTO through Patent Center and is hereby incorporated by reference in its entirety. The Sequence Listing XML, created on April 2, 2025, is named "SEQ_LIST--09824608-P240287WO01.xml" and is 98,314 bytes in size.

BACKGROUND

Amino acids are one of the fundamental building blocks of natural products and bioactive compounds. Modification of their chemical structure is one of the central themes of biosynthesis, and organic chemists have long sought to develop synthetic methodologies that can produce these valuable compounds. Amino acids that bear additional functional groups and stereocenters, however, have proven very difficult to synthesize efficiently. In particular, non-canonical amino acids (ncAAs) with a hydroxy group at the γ-position are common in natural products (**Fig. 1**), but they are so functionally complex so as to preclude routine synthesis for applications in medicinal chemistry.

UstD (Ustiloxin B biosynthesis protein D) is a decarboxylative aldolase discovered by Ye et al. as a unique tailoring enzyme in the biosynthesis of ustiloxin natural products (Ye et al. "Unveiling the Biosynthetic Pathway of the Ribosomally Synthesized and Post-Translationally Modified Peptide Ustiloxin B in Filamentous Fungi." Angewandte Chemie - International Edition 2016, 55, 8072-8075). A combination of classical and computationally guided directed evolution has previously been used for the synthesis of ncAAs with a secondary alcohol at the γ-position (Ellis et al. "Biocatalytic Synthesis of Non-Standard Amino Acids by a Decarboxylative Aldol Reaction." *Nat Catal* 2022, 5, 136-143). The resulting enzyme, UstD2.0, catalyzes the decarboxylative aldol addition with L-Asp and diverse aldehyde substrates.

Aldol additions into ketone electrophiles yield tertiary alcohols, but these reactions have distinct kinetic and thermodynamic challenges from their aldehyde counterparts.

Enzymes that produce ncAAs with a tertiary alcohol at the γ-position are needed.

SUMMARY

The present disclosure is directed to UstD mutants capable of producing non-canonical amino acids that contain secondary or tertiary alcohols at the gamma position, and methods of using the mutants to produce the non-canonical amino acids.

One aspect of the disclosure is directed to unnatural, mutant proteins.

In some versions, the unnatural, mutant protein comprises an amino acid sequence at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% identical to a UstD sequence selected from any one of SEQ ID NOs:1-6.

In some versions, the amino acid sequence comprises one or more of: a residue other than K at a position corresponding to position 2 of the UstD sequence; a residue other than V at a position corresponding to position 63 of the UstD sequence; a residue other than F at a position corresponding to position 75 of the UstD sequence; a residue other than P at a position corresponding to position 80 of the UstD sequence; a residue other than P at a position corresponding to position 82 of the UstD sequence; a residue other than P at a position corresponding to position 83 of the UstD sequence; a residue other than D at a position corresponding to position 86 of the UstD sequence; a residue other than Y at a position corresponding to position 96 of the UstD sequence; a residue other than G at a position corresponding to position 101 of the UstD sequence; a residue other than I at a position corresponding to position 141 of the UstD sequence; a residue other than H at a position corresponding to position 263 of the UstD sequence; a residue other than Y at a position corresponding to position 277 of the UstD sequence; a residue other than M at a position corresponding to position 299 of the UstD sequence; a residue other than V at a position corresponding to position 330 of the UstD sequence; a residue other than K at a position corresponding to position 342 of the UstD sequence; a residue other than G at a position corresponding to position 373 of the UstD sequence; a residue other than T at a position corresponding to position 388 of the UstD sequence; a residue other than I and T at a position corresponding to position 391 of the UstD sequence; a residue other than L at a position corresponding to position 392 of the UstD sequence; a residue other than L and M at a position corresponding to position 393 of the UstD sequence; a residue other than W at a position corresponding to position 399 of the UstD sequence; a residue other than S at a position corresponding to position 407 of the UstD sequence; a residue other than Y at a position corresponding to position 418 of the UstD sequence; and a residue other than L at a position corresponding to position 440 of the UstD sequence.

In some versions, the amino acid sequence comprises one or more, two or more, three or more, four or more, five or more, six or more, seven or more, eight or more, or each of: a residue other than F at a position corresponding to position 75 of the UstD sequence; a residue other than P at a position corresponding to position 82 of the UstD sequence; a residue other than D at a position corresponding to position 86 of the UstD sequence; a residue other than M at a position corresponding to position 299 of the UstD sequence; a residue other than V at a position corresponding to position 330 of the UstD sequence; a residue other than G at a position corresponding to position 373 of the UstD sequence; a residue other than I and T at a position corresponding to position 391 of the UstD sequence; a residue other than L and M at a position corresponding to position 393 of the UstD sequence; and a residue other than S at a position corresponding to position 407 of the UstD sequence.

In some versions, the amino acid sequence comprises one or more, two or more, or each of: a residue other than P at a position corresponding to position 82 of the UstD sequence; a residue other than V at a position corresponding to position 330 of the UstD sequence; and a residue other than G at a position corresponding to position 373 of the UstD sequence.

In some versions, the amino acid sequence comprises a residue other than P at a position corresponding to position 82 of the UstD sequence.

In some versions, the amino acid sequence comprises a residue other than V at a position corresponding to position 330 of the UstD sequence.

In some versions, the amino acid sequence comprises a residue other than G at a position corresponding to position 373 of the UstD sequence.

In some versions, the amino acid sequence comprises one, more than one, or each of: a residue other than P at a position corresponding to position 82 of the UstD sequence; and a residue other than G at a position corresponding to position 373 of the UstD sequence.

In some versions, the amino acid sequence comprises each of: a residue other than P at a position corresponding to position 82 of the UstD sequence; and a residue other than G at a position corresponding to position 373 of the UstD sequence.

In some versions, the amino acid sequence comprises one or more, two or more, three or more, or each of: a residue other than F at a position corresponding to position 75 of the UstD sequence; a residue other than D at a position corresponding to position 86 of the UstD sequence; a residue other than V at a position corresponding to position 330 of the UstD sequence; and a residue other than S at a position corresponding to position 407 of the UstD sequence.

In some versions, the amino acid sequence comprises each of: a residue other than F at a position corresponding to position 75 of the UstD sequence; a residue other than D at a position corresponding to position 86 of the UstD sequence; a residue other than V at a position corresponding to position 330 of the UstD sequence; and a residue other than S at a position corresponding to position 407 of the UstD sequence.

In some versions, the amino acid sequence comprises one or more, two or more, three or more, four or more, or each of: a residue other than M at a position corresponding to position 299 of the UstD sequence; a residue other than T at a position corresponding to position 388 of the UstD sequence; a residue other than I and T at a position corresponding to position 391 of the UstD sequence; a residue other than L at a position corresponding to position 392 of the UstD sequence; a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

In some versions, the amino acid sequence comprises one or more, two or more, or each of: a residue other than M at a position corresponding to position 299 of the UstD sequence; a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

In some versions, the amino acid sequence comprises each of: a residue other than M at a position corresponding to position 299 of the UstD sequence; a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

In some versions, the amino acid sequence comprises one or both of: a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

In some versions, the amino acid sequence comprises each of: a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

In some versions: the residue other than K at the position corresponding to position 2 of the UstD sequence, if present, is E or a conservative variant thereof; the residue other than V at the position corresponding to position 63 of the UstD sequence, if present, is I or a conservative variant thereof; the residue other than F at the position corresponding to position 75 of the UstD sequence, if present, is A, C, H, I, K, L, M, N, Q, R, S, T, V, W, Y, or a conservative variant of any of the foregoing; the residue other than P at the position corresponding to position 80 of the UstD sequence, if present, is G, L, R, or a conservative variant of any of the foregoing; the residue other than P at the position corresponding to

position 82 of the UstD sequence, if present, is G, Q, S, or a conservative variant of any of the foregoing; the residue other than P at the position corresponding to position 83 of the UstD sequence, if present, is G, T, R, V, Y, or a conservative variant of any of the foregoing; the residue other than D at the position corresponding to position 86 of the UstD sequence, if present, is I, N, V, or a conservative variant of any of the foregoing; the residue other than Y at the position corresponding to position 96 of the UstD sequence, if present, is C or a conservative variant thereof; the residue other than G at the position corresponding to position 101 of the UstD sequence, if present, is A, F, Q, R or a conservative variant of any of the foregoing; the residue other than I at the position corresponding to position 141 of the UstD sequence, if present, is M, V, or a conservative variant of any of the foregoing; the residue other than H at the position corresponding to position 263 of the UstD sequence, if present, is R or a conservative variant thereof; the residue other than Y at the position corresponding to position 277 of the UstD sequence, if present, is C, F, H, or a conservative variant of any of the foregoing; the residue other than M at the position corresponding to position 299 of the UstD sequence, if present, is L, V, or a conservative variant of any of the foregoing; the residue other than V at the position corresponding to position 330 of the UstD sequence, if present, is A, C, L, Q, R, or a conservative variant of any of the foregoing; the residue other than K at the position corresponding to position 342 of the UstD sequence, if present, is E or a conservative variant thereof; the residue other than G at the position corresponding to position 373 of the UstD sequence, if present, is E, R, or a conservative variant of any of the foregoing; the residue other than T at the position corresponding to position 388 of the UstD sequence, if present, is A, I, V, or a conservative variant of any of the foregoing; the residue other than I and T at the position corresponding to position 391 of the UstD sequence, if present, is F, S, or a conservative variant of any of the foregoing; the residue other than L at the position corresponding to position 392 of the UstD sequence, if present, is A or a conservative variant of thereof; the residue other than L and M at the position corresponding to position 393 of the UstD sequence, if present, is C, F, S, W, or a conservative variant of any of the foregoing; the residue other than W at the position corresponding to position 399 of the UstD sequence, if present, is C or a conservative variant of thereof; the residue other than S at the position corresponding to position 407 of the UstD sequence, if present, is A, E, N Q, T, or a conservative variant of any of the foregoing; the residue other than Y at the position corresponding to position 418 of the UstD sequence, if present, is H or a conservative variant of thereof; and the residue other than L at the position corresponding to position 440 of the UstD sequence, if present is P or a conservative variant of thereof.

In some versions: the residue other than F at the position corresponding to position 75 of the UstD sequence, if present, is A or a conservative variant thereof; the residue other than P at the position corresponding to position 82 of the UstD sequence, if present, is Q, S, or a conservative variant of any of the foregoing; the residue other than D at the position corresponding to position 86 of the UstD sequence, if present, is I or a conservative variant thereof; the residue other than M at the position corresponding to position 299 of the UstD sequence, if present, is V or a conservative variant thereof; the residue other than V at the position corresponding to position 330 of the UstD sequence, if present, is A, R, or a conservative variant of any of the foregoing; the residue other than G at the position corresponding to position 373 of the UstD sequence, if present, is E or a conservative variant thereof; the residue other than I and T at the position corresponding to position 391 of the UstD sequence, if present, is S or a conservative variant thereof; the residue other than L and M at the position corresponding to position 393 of the UstD sequence, if present, is F, W, or a conservative variant of any of the foregoing; and the residue other than S at the position corresponding to position 407 of the UstD sequence, if present, is Q or a conservative variant thereof.

In some versions, the protein has activity in generating a gamma-hydroxy amino acid from an amino acid and one or more of an aldehyde-containing substrate and a ketone-containing substrate.

Another aspect of the disclosure is directed to methods of making gamma-hydroxy amino acids.

In some versions, the methods comprising contacting an unnatural, modified protein disclosed herein with an amino acid and a substrate comprising one or more of an aldehyde-containing substrate and a ketone-containing substrate to yield the gamma-hydroxy amino acid.

In some versions, the substrate comprises a ketone-containing substrate.

In some versions, the ketone-containing substrate lacks vicinyl ketone groups.

In some versions, the unnatural, mutated protein comprises an amino acid sequence with at least 50% sequence identity but less than 100% sequence identity to a wild-type UstD protein as shown in SEQ ID NO: 1.

The objects and advantages of the disclosure will appear more fully from the following detailed description of the preferred embodiment of the disclosure made in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1. Natural products containing γ-hydroxy-α-amino acids. The decarboxylative aldol addition catalyzed by UstD with the aid of the pyridoxal phosphate (PLP) cofactor is given on the right. The native product of the enzyme is Ustiloxin B; the bond formed by UstD connects Cβ and Cγ, as indicated.

Fig. 2. Reactivity of UstD. A) Scheme of the general reaction catalyzed by UstD. B) Select aldehyde substrates that react with UstD and its engineered variants. C) Select non-aldehyde substrates that react with engineered variants.

Figs. 3A-3B. Application of SUMS. (Fig. 3A). General SUMS procedure. (Fig. 3B). UstD reaction scheme depicting the native reactivity with aldehyde electrophiles.

Figs. 4A-4B. Distal hotspots identified by globally random mutagenesis. (Fig. 4A). Combined results showing all promiscuity-shifting variants of interest from global random mutagenesis. Relative total activity for all products is represented by the dots. Product distribution is shown by the bars, with each bar segmented from top to bottom into compounds **3.1**, **3.2**, and **3.3**. The mutations found in each variant are displayed on the x-axis. Conditions: 50 mM l-asp, 5 mM thiophene-3-carboxaldehyde, 5 mM *o*-tolualdehyde, 40 mM (4-fluorophenyl)acetone, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, *E. coli* whole cells over-expressing $UstD^{2.0}$ variants, 37 °C, 200 rpm, 1 h reaction time. (Fig. 4B). $UstD^{2.0}$ structure showing the location of the distal residues in each monomer. The PLP cofactor is shown in the internal aldimine form.

Fig. 5. Combined results showing all promiscuity-shifting variants of interest from site saturation mutagenesis. Relative total activity for all products is represented by the dots. Product distribution is represented by the bars, with each bar segmented from top to bottom into compounds **3.1**, **3.2**, and **3.3**. The mutations found in each variant are displayed on the x-axis. Conditions: 50 mM l-asp, 5 mM thiophene-3-carboxaldehyde, 5 mM *o*-tolualdehyde, 40 mM (4-fluorophenyl)acetone, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, *E. coli* whole cells over-expressing $UstD^{2.0}$ variants, 37 °C, 200 rpm, 1 h reaction time.

Fig. 6. Single substrate reactions for F75X, P80X, P82X, and G373X at early timepoints. Reactions were performed in duplicate or triplicate, and activity was compared to $UstD^{2.0}$ reactions that were run simultaneously. The bars represent the average fold change of the replicates while the dots represent each individual measurement. The dotted line represents $UstD^{2.0}$ activity which is set to one. Conditions: 50 mM l-asp, 50 mM (4-fluorophenyl)acetone, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium

phosphate pH 7.0, 20 mg/mL *E. coli* whole cells over-expressing $UstD^{2.0}$ variants, 37 °C, 200 rpm, 1 h reaction time.

Fig. 7. Single substrate reactions of P82Q and G373E SSM libraries, and QE double mutant under challenging conditions. Reactions were performed in duplicate, and activity was compared to $UstD^{2.0}$ reactions that were run simultaneously. The bars represent the average fold change of the replicates while the dots represent each individual measurement. The dotted line represents $UstD^{2.0}$ activity which is set to one. Conditions: 50 mM l-asp, 50 mM (4-fluorophenyl)acetone, 5 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, 0.01 mol% cat (10,000 max turnovers), 37 °C, 16 h reaction time.

Fig. 8. Combined SSM results for libraries using QE as parent. All sequenced variants are displayed on the x-axis. Product distribution is shown by the bars, with each bar segmented from top to bottom into compounds **3.1**, **3.2**, **3.4**, and **3.5**. The black dots show the total fold change for all products. The total fold change for QE is represented by the red diamond and is set to one. The QE product distribution is the average QE product distribution across all library plates. Conditions: 50 mM l-asp 10 mM (4-fluorophenyl)acetone, 32.5 mM 4'-nitroacetophenone, 5 mM *o*-tolualdehyde, 2.5 mM 1,1,1-trifluoro-3-phenyl-2-propanone, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, *E. coli* whole cells over-expressing QE variants, 37 °C, 200 rpm, 1 h reaction time.

Fig. 9. Retention of function curves for overall activity on all products shown above the charts. Neither library shows any activated variants (white dots) compared to the QE parent enzyme (grey dots). Negative controls and the sterile well are shown in black.

Fig. 10. Subset of variants from SSM libraries re-screened in lysate against single substrates. Reactions were performed in at least duplicate, and activity was compared to QE reactions that were run simultaneously to determine fold change (y-axis). The bars represent the average fold change of the replicates while the dots represent each individual measurement. The dotted line represents QE activity which is set to one. Each group of bars corresponds to products **3.2**, **3.1**, **3.4**, and **3.5**, from left to right. Conditions: 50 mM l-asp, 50 mM electrophile, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, 40-50 mg/mL lysate, 37 °C, 16 h reaction time, QE parent enzyme.

Figs. 11A-11B. (Fig. 11A). All sequenced variants are displayed on the x-axis. The stacked bars depict the product distribution with each bar segmented from top to bottom into compounds **3.4**, **3.1**, **3.2**, and **3.5**. The black dots show the total fold change for all products. The total fold change for QE is represented by the red diamond and is set to one. The QE product distribution is the average QE product distribution across all library plates.

Conditions: 50 mM L-asp, 10 mM (4-fluorophenyl)acetone, 32.5 mM 4'-nitroacetophenone, 5 mM *o*-tolualdehyde, 2.5 mM 1,1,1-trifluoro-3-phenyl-2-propanone, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, *E. coli* whole cells over-expressing QE variants, 37 °C, 200 rpm, 8 h reaction time. (Fig. 11B). Subset of variants from recombination libraries re-screened against single substrates with purified protein. Reactions were performed in triplicate, and activity was compared to QE reactions that were run simultaneously to determine fold change (y-axis). The bars represent the average fold change of the replicates while the dots represent each individual measurement. The dotted line represents QE activity which is set to one. Each group of bars corresponds to products 3.2, 3.1, 3.4, and 3.5, from left to right. Conditions: 50 mM l-asp, 50 mM electrophile, 5 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, 0.01 mol% catalyst (10,000 Max turnovers), 37 °C, 16 h reaction time.

Fig. 12. UstD active site targeted for mutagenesis. (Panel A). Space filling view of overall UstD structure. (Panel B). Close up view of UstD active site.

Fig. 13. M299X SSM library data. All sequenced variants are displayed on the x-axis. The stacked bars depict the product distribution with each bar segmented from top to bottom into compounds **3.4**, **3.1**, **3.2**, and **3.5**. The black dots show the total fold change for all products. The total fold change for QE is represented by the red diamond and is set to one. Conditions: 50 mM L-asp, 10 mM (4-fluorophenyl)acetone, 32.5 mM 4'-nitroacetophenone, 5 mM *o*-tolualdehyde, 2.5 mM 1,1,1-trifluoro-3-phenyl-2-propanone, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, *E. coli* whole cells over-expressing QE variants, 37 °C, 200 rpm, 1 h reaction time.

Fig. 14. Combined library data of all sequenced active site recombination variants. All sequenced variants are displayed on the x-axis. The stacked bars depict the product distribution with each bar segmented from top to bottom into compounds **3.4**, **3.1**, **3.2**, and **3.5**. The distribution is shown starting at 40% to make visualization of the promiscuity shifts clearer for substrates with low activity. The black dots show the total fold change for all products. The total fold change for AIIRQ is represented by the red diamond and is set to one. The AIIRQ product distribution is the average product distribution across all library plates. Conditions: 50 mM L-asp, 10 mM (4-fluorophenyl)acetone, 32.5 mM 4'-nitroacetophenone, 5 mM *o*-tolualdehyde, 2.5 mM 1,1,1-trifluoro-3-phenyl-2-propanone, 50 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, *E. coli* whole cells over-expressing QE variants, 37 °C, 200 rpm, 6 h reaction time.

Fig. 15. Subset of variants from active-site recombination libraries re-screened against single substrates with purified protein. Reactions were performed in triplicate, and

activity was compared to AIIRQ reactions that were run simultaneously to determine fold change (y-axis). The bars represent the average fold activity of the replicates while the dots represent each individual measurement. The dotted line represents AIIRQ activity which is set to one. Each group of bars corresponds to products **3.2**, **3.1**, **3.4**, and **3.5**, from left to right. Conditions: 50 mM l-asp, 50 mM electrophile, 5 μM PLP, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, 0.01 mol% catalyst (10,000 Max turnovers), 37 °C, 16 h reaction time.

Fig. 16. Reaction condition optimization. Reaction condition optimization of 7G11. Initial reaction conditions are depicted in the top scheme. Conditions were optimized sequentially: (Panel A). 50 equivalents PLP relative to catalyst for (4-fluorophenyl)acetone. (Panel B). 5 equivalents of l-aspartate relative to (4-fluorophenyl)acetone. (Panel C). 10 equivalents PLP relative to catalyst for triFPhAT. (Panel D). 5 equivalents of l-aspartate relative to TriFPhAT.

Fig. 17. Lineage analysis of enzyme generality. Reactions were performed with single substrates in triplicate. The substrates are displayed to the right of the chart. The total turnover (TTN) number is displayed on the y-axis on a logarithmic scale. The bar represents the average of the technical replicates, and the dots represent the TTN measurement of each individual replicate. Each group of bars corresponds to products **3.6**, **3.5**, **3.7**, and **3.1**, from left to right. The mutations associated with each variant are displayed at the top of the graph with the connecting lines representing the relationship between them. Below the chart, the yield of protein per liter of culture is displayed for each variant. The ratio of TriFPhAT amino acid to the shunt pathway product, l-ala, is displayed below the chart to demonstrate that the ratio changes over the lineage.

Fig. 18. Chiral tertiary alcohol containing amino acids.

Figs. 19A-19C. Aldol reaction types. (Fig. 19A). Mukaiyama aldol using pre-formed enolates. (Fig. 19B). Decarboxylative aldol where the nucleophile is formed *in situ* upon decarboxylation of the pro-nucleophile. (Fig. 19C). Organocatalyst mediated aldol reactions where nucleophile is generated through covalent attachment to the catalyst.

Figs. 20A-20D. Biocatalytic methods for generating chiral tertiary alcohols.

Fig. 21. Analytical evaluation of substrate scope. All yields were quantified by Marfey's derivatization against a standard curve as described in the SI. For each substrate, the left stacked bar represents reactions with 7G11 and the right stacked bar represents reactions with 7B05. The upper segment of each stacked bar represents the amount of syn diastereomer formed with the assumption the anti-diastereomer is the major isomer. Each entry represents only a single reaction. Conditions: 250 mM L-asp, 50 mM electrophile, 10x

or 50x PLP compared to enzyme, 5% DMSO, 100 mM NaCl, 100 mM potassium phosphate pH 7.0, 7B05 or 7G11 (0.1 mol% cat, 1000 Max TON, 37 °C, 4 h reaction time.

Fig. 22. Preparative scale biocatalytic reactions.

Figs. 23A-23C. (Fig. 23A). A molecular drawing of the asymmetric unit in **4.4** shown with 50% probability ellipsoids. All H atoms (except those bound to N/O atoms or chiral centers) are omitted. The absolute configuration of the chiral atoms is C2–*S*, C4–*S*, C6–*R*, C2A–*S*, C4A–*S*, and C6A–*R*. (Fig. 23B). A molecular drawing overlaying the two symmetry independent molecules in **4.4** shown with 50% probability ellipsoids. All H atoms (except those bound to chiral centers) and solvent molecules are omitted. The second symmetry independent molecule is shown in green. The absolute configuration of the chiral atoms is C2–*S*, C4–*S*, C6–*R*, C2A–*S*, C4A–*S*, and C6A–*R*. (Fig. 23C). A molecular drawing of **4.6** shown with 50% probability ellipsoids. The two molecules are diastereomers.

DETAILED DESCRIPTION

One aspect of the disclosure is directed to mutant proteins. The mutant proteins comprise an amino acid sequence with at least one mutation with respect to a native amino acid sequence. "Native amino acid sequence" refers to the full amino acid sequence, or any contiguous portion thereof, of any protein found in nature.

The mutant proteins disclosed herein can comprise an amino acid sequence at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% to a UstD sequence. As used herein, "UstD sequence" refers to the amino acid sequence of a UstD protein. "UstD protein" as used herein refers to the UstD protein of *Aspergillus flavus* (SEQ ID NO:1) or variant forms thereof (*e.g.*, SEQ ID NOs:2-6). Accordingly, the mutant proteins disclosed herein may comprise an amino acid sequence at least about 80%, at least about 85%, at least about 90%, at least about 91%, at least about 92%, at least about 93%, at least about 94%, at least about 95%, at least about 96%, at least about 97%, at least about 98%, or at least about 99% identical to any one of SEQ ID NOs:1-6.

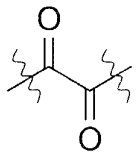
An exemplary codon-optimized coding sequence for the UstD protein of *Aspergillus flavus* is SEQ ID NO:7. Coding sequences for variants of the UstD protein can be generated by changing the codons to encode different amino acids according to the genetic code. Exemplary coding sequences for such variants are shown as SEQ ID NOs: 8-14.

The mutant proteins disclosed herein may have one or more substitutions at positions corresponding to particular positions of SEQ ID NOs:1-6. The one or more substitutions can include any substitutions at positions described or exemplified elsewhere herein.

In some versions, the UstD sequence can be fused to a heterologous amino acid sequence. The heterologous amino acid sequence can constitute a primary structure of any of a number of heterologous domains. Exemplary domains include linkers, affinity tags, or other catalytically active domains, among others. Linkers employed to fuse two heterologous polypeptides or domains to generate fusion proteins are well known in the art. See, *e.g.*, US Patent Nos. 5,525,491, 6,274,331, 6,479,626, 10,526,379, 10,752,965, and 11,123,438, among others. Exemplary linkers include linkers comprising glycine and serine, such as a -G-S- linker or a -G-S-G- linker. Exemplary linker lengths can be from 1-20 residues in length, such as from 1-20, 1-19, 1-18, 1-17, 1-16, 1-15, 1-14, 1-13, 1-12, 1-11, 1-10, 1-9, 1-8, 1-7, 1-6, 1-5, 1-4, 1-3, or 1-2 residues in length. Exemplary affinity tags include the His tag, the Strep II tag, the T7 tag, the FLAG tag, the S tag, the HA tag, the c-Myc tag, the dihydrofolate reductase (DHFR) tag, the chitin binding domain tag, the calmodulin binding domain tag, and the cellulose binding domain tag. The sequences of each of these tags are well-known in the art. Preferred affinity tags are those smaller than about 20 amino acids, such as the His tag, the Strep II tag, the T7 tag, the FLAG tag, the S tag, the HA tag, the c-Myc tag. Domains fused to the UstD sequence can be fused either directly or indirectly via a linker and, in various versions, can have a size less than 500 amino acids in length, less than 475 amino acids in length, less than 450 amino acids in length, less than 425 amino acids in length, less than 400 amino acids in length, less than 375 amino acids in length, less than 350 amino acids in length, less than 325 amino acids in length, less than 300 amino acids in length, less than 275 amino acids in length, less than 250 amino acids in length, less than 225 amino acids in length, less than 200 amino acids in length, less than 175 amino acids in length, less than 150 amino acids in length, less than 125 amino acids in length, less than 100 amino acids in length, less than 75 amino acids in length, less than 50 amino acids in length, or less than 25 amino acids in length. The heterologous amino acid sequence can be fused to the N-terminus or the C-terminus of the UstD sequence.

The mutant proteins disclosed herein can exhibit activity in generating a gamma-hydroxy amino acid from an amino acid and one or more of an aldehyde-containing substrate and a ketone-containing substrate. The amino acid can include any amino acid, whether natural or unnatural. In some versions, the amino acid is an alpha-amino acid. In some versions, the amino acid comprises any one or more of L-alanine (Ala/A), L-arginine (Arg/R), L-asparagine (Asn/N), L-aspartic acid (Asp/D), L-cysteine (Cys/C), L-glutamic

acid (Glu/E), L-glutamine (Gln/Q), L-glycine (Gly/G), L-histidine (His/H), L-isoleucine (Ile/I), L-leucine (Leu/L), L-lysine (Lys/K), L-methionine (Met/M), L-phenylalanine (Phe/F), L-proline (Pro/P), L-serine (Ser/S), L-threonine (Thr/T), L-tryptophan (Trp/W), L-tyrosine (Tyr/Y), L-valine (Val/V), L-selenocysteine (Sec/U), and L-pyrrolysine (Pyl/O). Exemplary aldehyde-containing substrates include those disclosed in Ellis et al. (2022), *supra*, and US Patent 11,591,626. Exemplary ketone-containing substrates are disclosed herein. In some versions, the ketone-containing substrate lacks vicinyl ketone groups:



Another aspect of the disclosure is a polynucleotide (or a gene) encoding a mutant protein disclosed herein. Another aspect of the disclosure is a vector comprising the polynucleotide (or the gene) according to the present disclosure. Vectors of the present disclosure can be transformed into suitable host cells to produce recombinant host cells.

Another aspect of the disclosure is a recombinant host cell comprising a polynucleotide encoding a mutant protein disclosed herein. In some versions, known genomic alteration or modification techniques can be employed to alter or modify the endogenous UstD protein of the host cell, effectuating one or more of the aforementioned mutations, such that at least one of the mutant UstD proteins has at least one altered property. In other versions, the recombinant host cell is engineered to include a plasmid comprising a polynucleotide encoding a mutant protein. In yet other versions, the recombinant host cell is engineered to include the polynucleotide encoding the mutant protein integrated into the chromosome of the host cell.

The recombinant host cell disclosed herein can be selected from any cell capable of expressing a recombinant gene construct, and can be selected from a microbial, plant or animal cell. In a particular embodiment, the host cell is bacterial, cyanobacterial, fungal, yeast, algal, human or mammalian in origin. In a particular embodiment, the host cell is selected from any of Gram positive bacterial species such as Actinomycetes; Bacillaceae, including *Bacillus alkalophilus, Bacillus subtilis, Bacillus licheniformis, Bacillus lentus, Bacillus brevis, Bacillus stearothermophilus, Bacillus alkalophilus, Bacillus amyloliquefaciens, Bacillus coagulans, Bacillus circulans, Bacillus lautus, Bacillus megaterium, B. thuringiensis; Brevibacteria* sp., including *Brevibacterium flavum,*

Brevibacterium lactofermentum, Brevibacterium ammoniagenes, Brevibacterium butanicum, Brevibacterium divaricatum, Brevibacterium healii, Brevibacterium ketoglutamicum, Brevibacterium ketosoreductum, Brevibacterium lactofermentum, Brevibacterium linens, Brevibacterium paraffinolyticum; Corynebacterium spp. such as *C. glutamicum* and *C. melassecola, Corynebacterium herculis, Corynebacterium lilium, Corynebactertium acetoacidophilum, Corynebacterium acetoglutamicum, Corynebacterium acetophilum, Corynebacterium ammoniagenes, Corynebacterium fujiokense, Corynebacterium nitrilophilus*; or lactic acid bacterial species including *Lactococcus* spp. such as *Lactococcus lactis; Lactobacillus* spp. including *Lactobacillus reuteri; Leuconostoc* spp.; *Pediococcus* spp.; *Serratia* spp. such as *Serratia marcescens; Streptomyces* species, such as *Streptomyces lividans, Streptomyces murinus, S. coelicolor* and *Streptococcus* spp. Alternatively, strains of a Gram negative bacterial species belonging to Enterobacteriaceae including *E. coli, Cellulomonas* spp.; or to Pseudomonadaceae including *Pseudomonas aeruginosa, Pseudomonas alcaligenes. Pseudomonas fluorescens, Pseudomonas putida, Pseudomonas syringae* and *Burkholderia cepacia, Salmonella* sp., *Stenotrophomonas* spp., and *Stenotrophomonas maltophilia.* Microorganisms such as *Rhodococcus* spp, *Rhodococcus* opacus, *Ralstonia* spp., and *Acetinobacter* spp. are useful as well. Furthermore, yeasts and filamentous fungal strains can be useful host cells, including *Absidia* spp.; *Acremonium* spp.; *Agaricus* spp.; *Anaeromyces* spp.; *Aspergillus* spp., including *A. aculeatus, A. awamori, A. flavus, A. foetidus, A. fumaricus, A. fumigatus, A. nidulans, A. niger, A. oryzae, A. terreus; A. tubingensis* and *A. versicolor; Aeurobasidium* spp.; *Cephalosporum* spp.; *Chaetomium* spp.; *Coprinus* spp.; *Dactyllum* spp.; *Fusarium* spp., including *F. conglomerans, F. decemcellulare, F. javanicum, F. lini, F. oxysporum* and *F. solani; Gliocladium* spp.; *Kluyveromyces* sp.; *Hansenula* sp.; *Humicola* spp., including *H. insolens* and *H. lanuginosa; Hypocrea* spp.; *Mucor* spp.; *Neurospora* spp., including *N. crassa* and *N. sitophila; Neocallimastix* spp.; *Orpinomyces* spp.; *Penicillium* spp.; *Phanerochaete* spp.; *Phlebia* spp.; *Pichia* sp.; *Piromyces* spp.; *Rhizopus* spp.; *Rhizomucor* species such as *Rhizomucor miehei; Saccaromyces* species such as *S. cerevisiae*, *S. pastorianus*, *S. eubayanus*, and *S. fragilis*; *Schizophyllum* spp.; *Schizosaccharomyces* such as, for example, *S. pombe* species; *chytalidium* sp., *Sulpholobus* sp., *Thermoplasma* sp., *Thermomyces* sp.; *Trametes* spp.; *Trichoderma* spp., including *T. reesei, T. reesei* (*longibrachiatum*) and *T. viride; Yarrowinia* sp.; and *Zygorhynchus* spp and in particular include oleaginous yeast just *Phafia* spp., *Rhorosporidium toruloides* Y4, *Rhodotorula Glutinis* and *Candida* 107.

In some versions, genes encoding mutant proteins and/or other recombinantly expressed genes in a recombinant host cell are modified to optimize at least one codon for expression in the recombinant host cell.

In some versions, a method is provided wherein the recombinant host cell according to the present disclosure is cultured under conditions that permit expression or overexpression of a mutant protein disclosed herein. The mutant protein can be recovered, and more preferably substantially purified, after the host cell is harvested and/or lysed.

Another aspect of the disclosure is directed to methods of generating a gamma-hydroxy amino acid with the mutant proteins disclosed herein. The method preferably comprises contacting a mutant protein disclosed herein with an amino acid and a substrate comprising one or more of an aldehyde-containing substrate and a ketone-containing substrate to yield a gamma-hydroxy amino acid. In some versions, the unnatural, mutated protein comprises an amino acid sequence with at least 50%, at least 55%, at least 60%, at least 65%, at least 70%, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% sequence identity but less than 100% sequence identity to a wild-type UstD protein as shown in SEQ ID NO: 1. In some versions, the substrate is a ketone-containing substrate. In some versions, the ketone-containing substrate contains vicinyl ketone groups. In some versions, the ketone-containing substrate lacks vicinyl ketone groups. In some versions, the contacting is performed in the presence of pyridoxal 5′-phosphate (PLP).

The term “altered property” refers to a modification in one or more properties of a mutant polynucleotide or mutant protein with reference to a corresponding polynucleotide or precursor protein.

The term “alignment” refers to a method of comparing two or more polynucleotides or polypeptide sequences for the purpose of determining their relationship to each other. Alignments are typically performed by computer programs that apply various algorithms, however it is also possible to perform an alignment by hand. Alignment programs typically iterate through potential alignments of sequences and score the alignments using substitution tables, employing a variety of strategies to reach a potential optimal alignment score. Commonly-used alignment algorithms include, but are not limited to, CLUSTALW, (see, Thompson J. D., Higgins D. G., Gibson T. J., CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Research 22: 4673-4680, 1994); CLUSTALV, (see, Larkin M. A., et al., CLUSTALW2, ClustalW and ClustalX version 2,

Bioinformatics 23(21): 2947-2948, 2007); Jotun-Hein, Muscle et al., MUSCLE: a multiple sequence alignment method with reduced time and space complexity, BMC Bioinformatics 5: 113, 2004); Mafft, Kalign, ProbCons, and T-Coffee (see Notredame et al., T-Coffee: A novel method for multiple sequence alignments, Journal of Molecular Biology 302: 205-217, 2000). Exemplary programs that implement one or more of the above algorithms include, but are not limited to MegAlign from DNAStar (DNAStar, Inc. 3801 Regent St. Madison, Wis. 53705), MUSCLE, T-Coffee, CLUSTALX, CLUSTALV, JalView, Phylip, and Discovery Studio from Accelrys (Accelrys, Inc., 10188 Telesis Ct, Suite 100, San Diego, Calif. 92121). In a non-limiting example, MegAlign is used to implement the CLUSTALW alignment algorithm with the following parameters: Gap Penalty 10, Gap Length Penalty 0.20, Delay Divergent Seqs (30%) DNA Transition Weight 0.50, Protein Weight matrix Gonnet Series, DNA Weight Matrix IUB.

The term "chromosomal integration" means the process whereby an incoming sequence is introduced into the chromosome of a host cell. The homologous regions of the transforming DNA align with homologous regions of the chromosome. Then, the sequence between the homology boxes can be replaced by the incoming sequence in a double crossover (i.e., homologous recombination). In some embodiments, homologous sections of an inactivating chromosomal segment of a DNA construct align with the flanking homologous regions of the indigenous chromosomal region of the microbial chromosome. Subsequently, the indigenous chromosomal region is deleted by the DNA construct in a double crossover.

The term "consensus sequence" or "canonical sequence" refers to an archetypical amino acid sequence against which all variants of a particular protein or sequence of interest are compared. Either term also refers to a sequence that sets forth the nucleotides that are most often present in a polynucleotide sequence of interest. For each position of a protein, the consensus sequence gives the amino acid that is most abundant in that position in the sequence alignment.

The term "conservative substitutions" or "conserved substitutions" refers to, for example, a substitution of an amino acid with a conservative variant.

"Conservative variant" refers to residues that are functionally similar to a given residue. Amino acids within the following groups are conservative variants of one another: glycine, alanine, serine, and proline (very small); alanine, isoleucine, leucine, methionine, phenylalanine, valine, proline, and glycine (hydrophobic); alanine, valine, leucine, isoleucine, methionine (aliphatic-like); cysteine, serine, threonine, asparagine, tyrosine, and glutamine (polar); phenylalanine, tryptophan, tyrosine (aromatic); lysine, arginine, and

histidine (basic); aspartate and glutamate (acidic); alanine and glycine; asparagine and glutamine; arginine and lysine; isoleucine, leucine, methionine, and valine; and serine and threonine.

The terms "corresponds to" and "corresponding to" used with reference to an amino acid residue or position refer to an amino acid residue or position in a first protein sequence being positionally equivalent to an amino acid residue or position in a second reference protein sequence by virtue of the fact that the residue or position in the first protein sequence aligns to the residue or position in the reference sequence using bioinformatic techniques, for example, using the methods described herein for preparing a sequence alignment. The corresponding residue in the first protein sequence is then assigned the position number in the second reference protein sequence.

The term "deletion," when used in the context of an amino acid sequence, means a deletion in or a removal of one or more residues from the amino acid sequence of a corresponding protein, resulting in a mutant protein having at least one less amino acid residue as compared to the corresponding protein. The term can also be used in the context of a nucleotide sequence, which means a deletion in or removal of a nucleotide from the polynucleotide sequence of a corresponding polynucleotide.

The term "DNA construct" and "transforming DNA" (wherein "transforming" is used as an adjective) are used interchangeably herein to refer to a DNA used to introduce sequences into a host cell or organism. Typically a DNA construct is generated *in vitro* by PCR or other suitable technique(s) known to those in the art. In certain embodiments, the DNA construct comprises a sequence of interest (e.g., an incoming sequence). In some embodiments, the sequence is operably linked to additional elements such as control elements (e.g., promoters, etc.). A DNA construct can further comprise a selectable marker. It can also comprise an incoming sequence flanked by homology targeting sequences. In a further embodiment, the DNA construct comprises other non-homologous sequences, added to the ends (e.g., stuffer sequences or flanks). In some embodiments, the ends of the incoming sequence are closed such that the DNA construct forms a closed circle. The transforming sequences may be wildtype, mutant or modified. In some embodiments, the DNA construct comprises sequences homologous to the host cell chromosome. In other embodiments, the DNA construct comprises non-homologous sequences. Once the DNA construct is assembled in vitro it may be used to: 1) insert heterologous sequences into a desired target sequence of a host cell; 2) mutagenize a region of the host cell chromosome (i.e., replace an endogenous sequence with a heterologous sequence); 3) delete target genes; and/or (4) introduce a replicating plasmid into the host.

A polynucleotide is said to "encode" an RNA or a polypeptide if, in its native state or when manipulated by methods known to those of skill in the art, it can be transcribed and/or translated to produce the RNA, the polypeptide, or a fragment thereof. The antisense strand of such a polynucleotide is also said to encode the RNA or polypeptide sequences. As is known in the art, a DNA can be transcribed by an RNA polymerase to produce an RNA, and an RNA can be reverse transcribed by reverse transcriptase to produce a DNA. Thus, a DNA can encode an RNA, and vice versa.

The term "expressed genes" refers to genes that are transcribed into messenger RNA (mRNA) and then translated into protein, as well as genes that are transcribed into types of RNA, such as transfer RNA (tRNA), ribosomal RNA (rRNA), and regulatory RNA, which are not translated into protein.

The terms "expression cassette" or "expression vector" refer to a polynucleotide construct generated recombinantly or synthetically, with a series of specified elements that permit transcription of a particular polynucleotide in a target cell. A recombinant expression cassette can be incorporated into a plasmid, chromosome, mitochondrial DNA, plasmid DNA, virus, or polynucleotide fragment. Typically, the recombinant expression cassette portion of an expression vector includes, among other sequences, a polynucleotide sequence to be transcribed and a promoter. In particular embodiments, expression vectors have the ability to incorporate and express heterologous polynucleotide fragments in a host cell. Many prokaryotic and eukaryotic expression vectors are commercially available. Selection of appropriate expression vectors is within the knowledge of those of skill in the art. The term "expression cassette" is also used interchangeably herein with "DNA construct," and their grammatical equivalents.

"Gene" refers to a polynucleotide (e.g., a DNA segment), which encodes a polypeptide, and may include regions preceding and following the coding regions as well as intervening sequences (introns) between individual coding segments (exons).

The term "homologous genes" refers to a pair of genes from different but related species, which correspond to each other and which are identical or similar to each other. The term encompasses genes that are separated by the speciation process during the development of new species) (e.g., orthologous genes), as well as genes that have been separated by genetic duplication (e.g., paralogous genes).

The term "endogenous protein" refers to a protein that is native or naturally occurring. "Endogenous polynucleotide" refers to a polynucleotide that is in the cell and was not introduced into the cell using recombinant engineering techniques; for example, a gene that was present in the cell when the cell was originally isolated from nature.

The term "heterologous" used with reference to a protein or a polynucleotide in a host cell refers to a protein or a polynucleotide that does not naturally occur in the host cell.

The term "heterologous" used to describe two different amino acid or nucleic acid sequences refers to two sequences that are not naturally present together in the same protein or nucleic acid. The term "heterologous" used to describe two different protein domains refers to two protein domains that are not naturally present together in the same protein. As used herein, "domain" refers to any portion of protein that confers a particular structural and/or functional characteristic to a protein. Exemplary protein domains include signal peptides, extracellular domains, transmembrane domains, cytoplasmic domains, catalytic domains, affinity tags, and linkers, among others.

The term "homologous recombination" refers to the exchange of DNA fragments between two DNA molecules or paired chromosomes at sites of identical or nearly identical nucleotide sequences. In certain embodiments, chromosomal integration is homologous recombination.

The term "homologous sequences" as used herein refers to a polynucleotide or polypeptide sequence having, for example, about 100%, about 99% or more, about 98% or more, about 97% or more, about 96% or more, about 95% or more, about 94% or more, about 93% or more, about 92% or more, about 91% or more, about 90% or more, about 88% or more, about 85% or more, about 80% or more, about 75% or more, about 70% or more, about 65% or more, about 60% or more, about 55% or more, about 50% or more, about 45% or more, or about 40% or more sequence identity to another polynucleotide or polypeptide sequence when optimally aligned for comparison. In particular embodiments, homologous sequences can retain the same type and/or level of a particular activity of interest. In some embodiments, homologous sequences have between 85% and 100% sequence identity, whereas in other embodiments there is between 90% and 100% sequence identity. In particular embodiments, there is 95% and 100% sequence identity.

"Homology" refers to sequence similarity or sequence identity. Homology is determined using standard techniques known in the art (see, e.g., Smith and Waterman, Adv. Appl. Math., 2:482, 1981; Needleman and Wunsch, J. Mol. Biol., 48:443, 1970; Pearson and Lipman, Proc. Natl. Acad. Sci. USA 85:2444, 1988; programs such as GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package (Genetics Computer Group, Madison, Wis.); and Devereux et al., Nucl. Acid Res., 12:387-395, 1984). A non-limiting example includes the use of the BLAST program (Altschul et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25:3389-3402, 1997) to identify sequences that can be said to be "homologous." A recent

version such as version 2.2.16, 2.2.17, 2.2.18, 2.2.19, or the latest version, including sub-programs such as blastp for protein-protein comparisons, blastn for nucleotide-nucleotide comparisons, tblastn for protein-nucleotide comparisons, or blastx for nucleotide-protein comparisons, and with parameters as follows: Maximum number of sequences returned 10,000 or 100,000: E-value (expectation value) of 1e-2 or 1e-5, word size 3, scoring matrix BLOSUM62, gap cost existence 11, gap cost extension 1, may be suitable. An E-value of 1e-5, for example, indicates that the chance of a homologous match occurring at random is about 1 in 10,000, thereby marking a high confidence of true homology.

The term "host strain" or "host cell" refers to a suitable host for an expression vector comprising a DNA disclosed herein. The host may comprise any organism, without limitation, capable of containing and expressing the nucleic acids or genes disclosed herein. The host may be prokaryotic or eukaryotic, single-celled or multicellular, including mammalian cells, plant cells, fungi, etc. Examples of single-celled hosts include cells of *Escherichia*, *Salmonella*, *Bacillus*, *Clostridium*, *Streptomyces*, *Staphyloccus*, *Neisseria*, *Lactobacillus*, *Shigella*, and *Mycoplasma*. Suitable *E. coli* strains (among a great many others) include BL21(DE3), C600, DH5αF′, HB101, JM83, JM101, JM103, JM105, JM107, JM109, JM110, MC1061, MC4100, MM294, NM522, NM554, TGI, χ1776, XL1-Blue, and Y1089+, all of which are commercially available.

The term "identical" (or "identity"), in the context of two polynucleotide or polypeptide sequences, means that the residues in the two sequences are the same when aligned for maximum correspondence, as measured using a sequence comparison or analysis algorithm such as those described herein. For example, if when properly aligned, the corresponding segments of two sequences have identical residues at 5 positions out of 10, it is said that the two sequences have a 50% identity. Most bioinformatic programs report percent identity over aligned sequence regions, which are typically not the entire molecules. If an alignment is long enough and contains enough identical residues, an expectation value can be calculated, which indicates that the level of identity in the alignment is unlikely to occur by random chance.

The term "increased," when used with respect to an increased activity, refers to an increase in activity over a baseline level of activity, regardless of whether the baseline level activity is a positive level of activity or a null level of activity.

The term "insertion," when used in the context of an amino acid sequence, refers to an insertion of an amino acid with respect to the amino acid sequence of a corresponding polypeptide, resulting in a mutant polypeptide having an amino acid that is inserted between two existing contiguous amino acids, i.e., adjacent amino acids residues, which are present

in the corresponding polypeptide. The term "insertion," when used in the context of a polynucleotide sequence, refers to an insertion of one or more nucleotides in the corresponding polynucleotide between two existing contiguous nucleotides, i.e., adjacent nucleotides, which are present in the corresponding polynucleotides.

The term "introduced" refers to, in the context of introducing a polynucleotide sequence into a cell, any method suitable for transferring the polynucleotide sequence into the cell. Such methods for introduction include but are not limited to protoplast fusion, transfection, transformation, conjugation, and transduction (see, e.g., Ferrari et al., Genetics, in Hardwood et al, (eds.), *Bacillus*, Plenum Publishing Corp., pp. 57-72, 1989).

The term "isolated" or "purified" means a material that is removed from its original environment, for example, the natural environment if it is naturally occurring, or a cultivation broth if it is produced in a recombinant host cell cultivation medium. A material is said to be "purified" when it is present in a particular composition in a higher concentration than the concentration that exists prior to the purification step(s). For example, with respect to a composition normally found in a naturally-occurring or wild type organism, such a composition is "purified" when the final composition does not include some material from the original matrix. As another example, where a composition is found in combination with other components in a recombinant host cell cultivation medium, that composition is purified when the cultivation medium is treated in a way to remove some component of the cultivation, for example, cell debris or other cultivation products, through, for example, centrifugation or distillation. As another example, a naturally-occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated, whether such process is through genetic engineering or mechanical separation. Such polynucleotides can be parts of vectors. Alternatively, such polynucleotides or polypeptides can be parts of compositions. Such polynucleotides or polypeptides can be considered "isolated" because the vectors or compositions comprising thereof are not part of their natural environments. In another example, a polynucleotide or protein is said to be purified if it gives rise to essentially one band in an electrophoretic gel or a blot.

The term "mutation" refers to, in the context of a polynucleotide, a modification to the polynucleotide sequence resulting in a change in the sequence of a polynucleotide with reference to a corresponding polynucleotide sequence. A mutation to a polynucleotide sequence can be an alteration that does not change the encoded amino acid sequence, for example, with regard to codon optimization for expression purposes, or that modifies a codon in such a way as to result in a modification of the encoded amino acid sequence.

Mutations can be introduced into a polynucleotide through any number of methods known to those of ordinary skill in the art, including random mutagenesis, site-specific mutagenesis, oligonucleotide directed mutagenesis, gene shuffling, directed evolution techniques, combinatorial mutagenesis, site saturation mutagenesis among others.

"Mutation" or "mutated" means, in the context of a protein, a modification to the amino acid sequence resulting in a change in the sequence of a protein with reference to a corresponding protein sequence. A mutation can refer to a substitution of one amino acid with another amino acid, an insertion of one or more amino acid residues, or a deletion of one or more amino acid residues. A mutation can include the replacement of an amino acid with a non-natural amino acid, or with a chemically-modified amino acid or like residues. A mutation can also be a truncation (e.g., a deletion or interruption) in a sequence or a subsequence from the corresponding sequence. A mutation can be made by modifying the DNA sequence corresponding to the corresponding protein. Mutations can be introduced into a protein sequence by known methods in the art, for example, by creating synthetic DNA sequences that encode the mutation with reference to corresponding proteins, or chemically altering the protein itself. A "mutant" as used herein is a protein comprising a mutation.

A "naturally occurring equivalent," in the context of the present disclosure, refers to a naturally-occurring UstD protein.

The term "operably linked," in the context of a polynucleotide sequence, refers to the placement of one polynucleotide sequence into a functional relationship with another polynucleotide sequence. For example, a DNA encoding a secretory leader (e.g., a signal peptide) is operably linked to a DNA encoding a polypeptide if it is expressed as a preprotein that participates in the secretion of the polypeptide. A promoter or an enhancer is operably linked to a coding sequence if it affects the transcription of the sequence. A ribosome binding site is operably linked to a coding sequence if it is positioned so as to facilitate translation. Generally, "operably linked" means that the DNA sequences being linked are contiguous, and, in the case of a secretory leader, contiguous and in the same reading frame.

The term "optimal alignment" refers to the alignment giving the highest overall alignment score.

"Overexpressed" or "overexpression" in a host cell occurs if the enzyme is expressed in the cell at a higher level than the level at which it is expressed in a corresponding wild-type cell.

The terms "percent sequence identity," "percent amino acid sequence identity," "percent gene sequence identity," and/or "percent polynucleotide sequence identity," with respect to two polypeptides, polynucleotides and/or gene sequences (as appropriate), refer to the percentage of residues that are identical in the two sequences when the sequences are optimally aligned. Thus, 80% amino acid sequence identity means that 80% of the amino acids in two optimally aligned polypeptide sequences are identical. The percent identities expressed herein with respect to a given named reference sequence are determined over the entire reference sequence, rather than only a portion thereof. Thus, an amino acid sequence at least about 80% identical to SEQ ID NO:1, for example, is at least about 80% identical to the entire sequence of SEQ ID NO:1, as opposed merely to subsequences thereof.

The term "plasmid" refers to a circular double-stranded (ds) DNA construct used as a cloning vector, and which forms an extrachromosomal self-replicating genetic element in some eukaryotes or prokaryotes, or integrates into the host chromosome.

A "production host" is a cell used to produce products. As disclosed herein, a production host is modified to express or overexpress selected genes, and/or to have attenuated expression of selected genes. Non-limiting examples of production hosts include plant, animal, human, bacteria, yeast, cyanobacteria, algae, and/or filamentous fungi cells.

A "promoter" is a polynucleotide sequence that functions to direct transcription of a downstream gene. In preferred embodiments, the promoter is appropriate to the host cell in which the target gene is being expressed. The promoter, together with other transcriptional and translational regulatory polynucleotide sequences (also termed "control sequences") is necessary to express a given gene. In general, the transcriptional and translational regulatory sequences include, but are not limited to, promoter sequences, ribosomal binding sites, transcriptional start and stop sequences, translational start and stop sequences, and enhancer or activator sequences.

The terms "protein" and "polypeptide" are used interchangeably herein. The 3-letter code as well as the 1-letter code for amino acid residues as defined in conformity with the IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) is used throughout this disclosure. It is also understood that a polypeptide may be coded for by more than one polynucleotide sequence due to the degeneracy of the genetic code. An enzyme is a protein. The terms "amino acid sequence" and "polypeptide sequence" are used interchangeably herein.

The terms "nucleic acid" and "polynucleotide" are used interchangeably herein.

The term "recombinant," when used to modify the term "cell" or "vector" herein, refers to a cell or a vector that has been modified by the introduction of a heterologous

polynucleotide sequence, or that the cell is derived from a cell so modified. Thus, for example, recombinant cells express genes that are not found in identical form within the native (non-recombinant) form of the cells or express, as a result of deliberate human intervention, native genes that are otherwise abnormally expressed, underexpressed, or not expressed at all. The terms "recombinant," used with respect to proteins and nucleic acids refers to mutant proteins and nucleic acids, respectively.

The terms "regulatory segment," "regulatory sequence," or "expression control sequence" refer to a polynucleotide sequence that is operatively linked with another polynucleotide sequence that encodes the amino acid sequence of a polypeptide chain to effect the expression of that encoded amino acid sequence. The regulatory sequence can inhibit, repress, promote, or even drive the expression of the operably-linked polynucleotide sequence encoding the amino acid sequence.

The term "substantially identical," in the context of two polynucleotides or two polypeptides refers to a polynucleotide or polypeptide that comprises at least 70% sequence identity, for example, at least 75%, at least 80%, at least 85%, at least 90%, at least 91%, at least 92%, at least 93%, at least 94%, at least 95%, at least 96%, at least 97%, at least 98%, or at least 99% sequence identity as compared to a reference sequence using the programs or algorithms (e.g., BLAST, ALIGN, CLUSTAL) using standard parameters..

"Substantially purified" means molecules that are at least about 60% free, preferably at least about 75% free, about 80% free, about 85% free, and more preferably at least about 90% free from other components with which they are naturally associated. As used herein, the term "purified" or "to purify" also refers to the removal of contaminants from a sample.

"Substitution" means replacing an amino acid in the sequence of a corresponding protein with another amino acid at a particular position, resulting in a mutant of the corresponding protein. The amino acid used as a substitute can be a naturally-occurring amino acid, or can be a synthetic or non-naturally-occurring amino acid.

The term "transformed" or "stably transformed" cell refers to a cell that has a non-native (heterologous) polynucleotide sequence integrated into its genome or as an episomal plasmid that is maintained for at least two generations.

"Vector" refers to a polynucleotide construct designed to introduce polynucleotides into one or more cell types. Vectors include cloning vectors, expression vectors, shuttle vectors, plasmids, cassettes and the like. In some embodiments, the polynucleotide construct comprises a polynucleotide sequence encoding a mutant protein that is operably linked to a suitable prosequence capable of effecting the expression of the polynucleotide or gene in a suitable host.

"Wild type" means, in the context of a gene or protein, a polynucleotide or protein sequence that occurs in nature. In some embodiments, the wild-type sequence refers to a sequence of interest that is a starting point for protein engineering. "Wild type" is used interchangeably with "native."

"Substituent" refers to a moiety other than hydrogen.

US Patent 11,591,626 is incorporated herein by reference. In case of conflict between the disclosures, the present disclosure controls.

The elements and method steps described herein can be used in any combination whether explicitly described or not.

All combinations of method steps as used herein can be performed in any order, unless otherwise specified or clearly implied to the contrary by the context in which the referenced combination is made.

The methods of the present disclosure can comprise, consist of, or consist essentially of the essential elements and limitations of the method described herein, as well as any additional or optional ingredients, components, or limitations described herein or otherwise useful in molecular biology, organic chemistry, and/or genetic engineering. The disclosure provided herein suitably may be practiced in the absence of any element which is not specifically disclosed herein.

All references to singular characteristics or limitations of the present disclosure shall include the corresponding plural characteristic or limitation, and vice-versa, unless otherwise specified or clearly implied to the contrary by the context in which the reference is made. The indefinite articles "a" and "an" mean "one or more." The word "or" is used inclusively and should be read "and/or."

Numerical ranges as used herein are intended to include every number and subset of numbers contained within that range, whether specifically disclosed or not. Further, these numerical ranges should be construed as providing support for a claim directed to any number or subset of numbers in that range. For example, a disclosure of from 1 to 10 should be construed as supporting a range of from 2 to 8, from 3 to 7, from 5 to 6, from 1 to 9, from 3.6 to 4.6, from 3.5 to 9.9, and so forth.

All patents, patent publications, and peer-reviewed publications (*i.e.*, "references") cited herein are expressly incorporated by reference to the same extent as if each individual reference were specifically and individually indicated as being incorporated by reference. In case of conflict between the present disclosure and the incorporated references, the present disclosure controls.

It is understood that the disclosure is not confined to the particular construction and arrangement of parts herein illustrated and described, but embraces such modified forms thereof as come within the scope of the claims.

EXAMPLES

The challenges of aldolase chemistry for synthesis of β-hydroxy-α-amino acids have been described in McDonald et al. ("Engineering Enzyme Substrate Scope Complementarity for Promiscuous Cascade Synthesis of 1,2-Amino Alcohols." *Angew Chem Int Ed Engl* 2022, 61, e202212637). Whether UstD could be engineered to overcome these limitations was unknown and even doubtful. Specifically, Zhang et al. reported on a UstD homolog from *Aspergillus pseudonomius* (ApUstD), highlighting the *inactivity* of the enzyme on other ketone electrophiles that lack a strong electron-withdrawing group in the form of the di-keto functionality ("Enzymatic Synthesis of Noncanonical α-Amino Acids Containing γ-Tertiary Alcohols." *Angew Chem Int Ed Engl* 2024, 63, e202318550).

We have performed extensive directed evolution to identify UstD variants with improved activity on both aldehyde and ketone substrates (**Fig. 2**). The resulting enzymes have higher expression titer in *E. coli*, up to 100 mg protein per L culture, which increases the utility of the enzymes for preparative scale biocatalysis. Variants within the UstD sequence space described herein enable preparative-scale reactions with *unactivated* ketone substrates, which was previously not known, to form amino acids with tertiary alcohol side chains. Reactions can operate in a whole-cell fashion and the catalysts may be used as either purified protein, clarified lysate, or *in vivo*.

Improved variants were identified through a promiscuity-guided evolutionary process using substrate-multiplexed screening. Here, enzymes were screened against not a single "representative" substrate, as is typical in directed evolution (Arnold. "Design by Directed Evolution." *Acc Chem Res* 1998, 31, 125-131; Arnold. "The Nature of Chemical Innovation: New Enzymes by Evolution." *Quarterly Reviews of Biophysics*. Cambridge University Press November 16, 2015, 404-410). Instead, evolution was done via screening against a mixture of substrates with the goal of identifying either variants with general increases in activity or variants with shifts in substrate promiscuity (McDonald et al., 2022, *supra*). We often identified mutations that cause decreases in activity but shift promiscuity. This information was used for targeted engineering to increase activity. The success of this strategy was shown by the high activity of a distal recombinant (UstD-ARIIQ), that bears no additional active site mutations but has a broad increase in reactivity with non-aldehyde substrates. The modest loss in activity on a model aldehyde compared to the parent enzyme, UstD2.0, is

partially offset by the improved soluble expression of the UstD-ARIIQ enzyme. Subsequently, active site mutagenesis at 5 positions identified 27 variants with increases in activity on at least one substrate of 1.2-fold in competition. Two active site variants were chosen for preparative scale reactions to validate the identity of the products and demonstrate the practical utility of these catalysts. **Table 1** summarizes the evolution rounds, including screening details and key variant outcomes.

Table 1. Summary of evolution.

Round	Description	Clones Screened	Variants with altered promiscuity and/or activity in screening	Best variant(s) subject to validation
1	**Random mutagenesis** of the entire gene	880	G373R, D86V, F75S, Y96C+G101R, P82S+V330A, P83R, Y277C+K342E+S407N, P80L, H263R, Y418H	**UstD$^{v2.0}$** All variants were generally deactivated but changed promiscuity. No mutations were fixed this round.
2	**Site saturation** of "hotspots" with UstD$^{v2.0}$: F75, P80, P82, G373	352	**F75:** A, C, H, I, K, L, M, N, Q, R, S, T, W, Y **P80:** G, R **P82:** G, Q **G373:** E, R	**P82Q** **G373E** **F75A**
3	**Double mutant** P82Q+G373E	1	P82Q+G373E	**QE** = P82Q+G373E
4	**Site saturation** of "hotspots" with QE as parent: P83, D86, Y96, G101, I141, Y277, V330, K342, S407, Y418,	880	**P83:** G, V, T, Y **D86:** I, V **G101:** F, Q, A+ΔH445 **I141:** V, M, **Y277:** H, F+W399C **V330:** Q, A, R, C **K342:** none **S407:** Q, T, A, E **Y418:** none	D86I and D86V. Mutational information from D86, I141, V330, and S407, along with F75 in the preceding round was used to design a subsequent recombination library
5	**Recombination** at sites F75, D86, I141, V330, S407	704	F75A+D86I+I141V+S407E F75A+D86I+V330R+S407Q F75V+D86N+V330V F75A+D86V+I141V+S407Q F75A+D86V+I141M+S407E F75V+D86V+S407Q F75A+D86V+V330L+S407E F75A+D86I+V330A+S407Q F75V+D86V+S407E F75A+D86I+V330A+S407Q K2E+F75V+D86I+V330V F75A+D86V+V330Q+S407E	**AIIRQ**= QE+ F75A+ D86I+ I141I(ATC→ATT)+ V330R+ S407Q
6	**Recombination** at active site residues M299,	968	M299L+T388V+T391F M299V+T388V M299V+T388V+T391S+L440	**7G11** = AIIRQ+ M299V+ T391S+ M393W

	T388, T391, L392, M393		P M299V+T388I+T391F+M393C M299L+T388I+T391S+M393C M299V+T388I T388V+T391F+M393S T388I+T391S T391S+M393W M299V+T391S+M393F M299L+T388A+T391F M299V+M393W M299V+T388V+M393W M299V+T388V+M393W T391S+M393F M299V+T388A+T391F+M393W M299V+T388I M299V+T391S+M393W M299V+T388V+T391F+M393F M299L+T391S+M393W M299V+T388I+T391S+M393F M299V+T388I+T391F M299V+T391S+L392A+M393W T388I V63I M299V+T391F+M393W M299V+T388I+M393F M299V+T388I+M393L M299V+T388A+T391F+M393W M299L T391F+M393W T391F+L392I+M393W T388A+T391F	**7B05** = AIIRQ+ T391S+ M393F

The engineering strategy delivers information on both the scope of the variants and demonstrates that the improvement in activity can extend to a wide sequence space. A collection of mutations are shown to be activating in a variety of combinations.

Detailed experimental methods, results, and a discussion of same are provided in Examples 1 and 2 below.

Example 1. Promiscuity-guided engineering of a decarboxylative aldolase

Introduction

Limitations of traditional protein engineering for identifying generalists

Directed evolution has found widespread use for developing biocatalysts that are applied in a myriad of industrial settings, including pharmaceuticals, fine chemicals, and bioremediation. The power of directed evolution lies in the iterative cycles of genetic diversification and selection, allowing researchers to identify mutations that enhance enzyme activity, either by increasing catalyst concentration, stability, or activity under researcher-defined conditions. The rapid identification of stabilizing mutations that do not compromise function still requires trial and error, but recent advances in computational design and machine learning are making rapid strides in this area. In contrast, strategies to identify mutations that alter catalytic activity are more limited. Mutagenesis of active site residues is a reliable strategy, but direct identification of distal sites that influence reactivity is a long-standing challenge in protein engineering because it requires in depth structural and mechanistic information.

In standard directed evolution workflows, a single model substrate is typically chosen for screening even when synthetic utility is the goal for enzyme evolution. While this approach is often quite successful, there are also many cases where directed evolution yields a catalyst that has high activity for the substrate under selection but struggles to react with substrate analogs (low substrate promiscuity). In some cases, intermediates along a directed evolution lineage are more promiscuous and evolution inadvertently limited the scope of the transformation. More broadly, the inability to efficiently track substrate promiscuity during the initial screening phase hinders the ability to engineer enzymes toward generality.

Substrate multiplexed screening (SUMS) as an alternative engineering approach

We advanced substrate multiplexed screening (SUMS) as an alternative screening strategy for directed evolution to aid in tracking changes to enzyme promiscuity while minimizing researcher intervention and screening time. By screening substrates in competition, total activity and the distribution of the products, representing promiscuity, are measured simultaneously **(Fig. 3A)**.

Here, we chose to explore the effectiveness of multi-generational, promiscuity-guided engineering in the context of a historically challenging reaction in protic solvents, aldol addition into ketones. The classical evolution of PLP-dependent decarboxylative aldolase, UstD, to react with benzaldehyde has been discussed in Ellis et al. "Biocatalytic synthesis of non-standard amino acids by a decarboxylative aldol reaction." *Nat. Catal.*

2022, 52, 136-143. This is a convergent C–C bond forming reaction yielding a γ-hydroxy non-canonical amino acid product (**Fig 3B**). The evolution, serendipitously, resulted in a "generalist" enzyme, UstD$^{2.0}$, capable of reacting with diverse aldehyde electrophiles to generate chiral secondary alcohols. Reactions with ketone electrophiles would yield chiral tertiary alcohols, a highly sought-after motif in medicinal chemistry. However, aldol-type reactions with ketones are deceptively challenging compared to aldehydes, as ketones are thermodynamically more stable and kinetically slower to react than aldehydes for both steric and electronic reasons. The challenging reactivity in combination with the synthetic relevance of the products makes the evolution of UstD$^{2.0}$ for aldol addition into ketones an excellent testing ground for exploring the SUMS method.

Practical considerations for implementing SUMS

To successfully implement SUMS, we considered the effect of several parameters, notably the composition of the substrate mixture and reaction time. We hypothesized that, by screening on a substrate mixture, changes in the relative rates of reactivity would aid in the identification of residues outside the active site that are influencing catalysis. We considered an initial substrate mixture of both aldehydes and ketones to provide UstD$^{2.0}$ with distinct steric and electronic challenges to reactivity. In direct competition, robust UPLC-MS signals for the secondary alcohol products were observed and indicated each aldehyde substrate reacted with similar efficiency. This indicated the selection of the substrate mixture is devoid of substrates that act as strong competitive inhibitors or skew the promiscuity profile due to high reactivity. However, the signal arising from ketone reactivity was trace under these conditions and a four-fold excess of ketone to aldehyde was deployed.

In a typical protein engineering assay where high yield is the goal and catalyst activity is limiting, reactions are quantified at the reaction endpoint because the greatest difference in enzyme activity would be visible at the end of a single substrate reaction. In contrast, the greatest differences in promiscuity are observed at early timepoints. If the promiscuity were measured under initial velocity conditions that would allow us to determine the specificity of each variant. However, maintaining initial velocity conditions would have required tedious optimization and monitoring of the assay. Therefore, we elected to only measure promiscuity for technical simplicity. Therefore, we assayed the variants after a one-hour reaction time, which provided a small but reproducible signal for the tertiary alcohol product. With resolved screening parameters, engineering towards ketone activity began using globally random mutagenesis.

Identification of distal "hotspots" from global random mutagenesis

We screened a small set of global random mutagenesis libraries prepared through error-prone PCR, comprising 880 random clones. From this screen we identified no variants with general boosts in activity on all substrates. Some variants, however, appeared to have changes in promiscuity (**Fig. 4A**). Diverse strategies to quantitate promiscuity have been developed. However, applications of these metrics to our data demonstrated they have limited utility in this context because activity on the ketone substrate is very low. Most variants had reduced activity relative to parent and the noise associated with low-intensity measurements was indistinguishable from a change in promiscuity. Nevertheless, visual inspection clearly showed that a range of mutational effects were observed with some variants displaying apparent shifts in promiscuity. We selected 7 such variants for sequencing from which we identified 12 mutations (**Fig. 4A**). Analysis of the $UstD^{2.0}$ crystal structure showed that no mutations were in the active site and the alpha carbon (Cα) of the mutations were an average 18 Å away from the cofactor (**Fig. 4B**). Consequently, we hypothesized that these mutations may be "hotspots" for altering activity. Although these specific mutations are deleterious, some other mutation at the same site may be beneficial for catalysis, which we tested with site saturation mutagenesis (SSM).

Mutation of distal promiscuity-shifting sites reveals activating mutations

We selected P82 and F75 as initial sites to test our allosteric site hypothesis. Residue P82 is 10.7 Å from the catalytic Lys, and the crystal structure of $UstD^{2.0}$ shows that it formed a cis-peptide bond. Mutation at P82 therefore has the potential to introduce more pervasive structural changes. Residue F75, is located 18 Å from the catalytic Lys and mutation at this residue during globally random mutagenesis displayed the largest apparent shift towards reactivity with a ketone. We also considered two sites from random mutagenesis that had a comparable decrease in total activity to the promiscuity-shifting sites, but no significant change in promiscuity, G373 and P80. The initial round of global random mutagenesis gave no suggestion that these mutations might impact reaction specificity. They therefore acted as control sites to the relative efficiency of SSM at a putative allosteric site relative to some other site in the protein. To our surprise, the SUMS profiles of the top variants from each SSM library suggested that the control site, G373, contained generally activating mutations. In contrast, mutation at P82 had similar total activity relative to parent, but with an altered reactivity profile suggestive of a more generalist enzyme (**Fig. 5**).

These variants were re-screened outside of a competition setting on just the desired ketone substrate, but we chose to mimic the screening conditions closely using whole-cell

catalyst and only allowing each reaction to run for one hour (**Fig. 6**). Under the whole-cell conditions, none of the P80 variants displayed increased activity for **3.1**, confirming our hypothesis that the site would not lead to activated variants. The other three sites all displayed increased activity for **3.1**. These results confirmed our allosteric site hypothesis as mutation at P82 and F75 significantly increased enzyme activity on a ketone substrate, as well as G373E that was discovered serendipitously.

We further tested the top two variants (P82Q and G373E) and assayed them under conditions that further challenge catalyst stability and turnover numbers (dilute catalyst, 16 h reaction time) to help differentiate between the variants. Under these more challenging conditions, G373E displayed a 1.2-fold boost in activity for **3.1** while P82Q had 9.6-fold higher activity for **3.1** compared to parent. (**Fig. 7**). To probe potential additive or cooperative effects between these distal sites, we generated a new variant combining the mutations P82Q and G373E to make a new variant, QE, with 11.7-fold improvement in **3.1** (**Fig. 7**). As the mutations are distal to the active site, it is difficult to even speculate on the molecular mechanisms through which these mutations operate. Nevertheless, this increase in activity is significant because activity with the desired ketone electrophile has crossed the threshold from stoichiometric turnover with $UstD^{2.0}$, to a modest 17 turnovers with QE.

Substrate space redesign to increase sensitivity to changes in reactivity

Our initial substrate mixture was chosen to minimize data complexity. Spurred by the success of the promiscuity-guided evolution above, we considered whether more information might be accessible from SUMS when using a substrate mixture with more diverse electrophiles. To maintain some continuity between the global random mutagenesis and the SSM screens both (4-fluorophenyl)acetone and *o*-tolualdehyde were retained in the substrate mixture, whereas the thiophene aldehyde was omitted. We introduced two new ketones, 1,1,1-trifluoro-3-phenyl-2-propanone (TFMK) to diversify ketone electrophilicity and 4'-nitroacetophenone (4NO2AP) to diversify structures in the substrate mixture. In simple mixtures with equimolar concentrations, the activated electrophiles dominated the overall reactivity and the signal for the tertiary alcohol products was again on the order of experimental noise. We therefore altered the substrate concentrations to ensure reproducible and robust signals for all products via UPLC-MS. The resulting mixture contained a 9:1 ratio of ketone to aldehyde substrates. In this straightforward way, substrate mixtures for multi-generational evolutionary campaigns can be re-tuned as reactivity is expanded.

We used QE as the parent enzyme for additional SSM at distal residues identified from globally random mutagenesis. SSM at eight out of twelve of these putative allosteric

sites led to variants with at least a 1.5-fold increase in total activity during screening (**Fig. 8**). We observed that mutations D86V and V330A both produced a 2-fold increase in activity. We also found, serendipitously, that synonymous codon changes at I141 and S371 led to boosts in whole-cell catalyst activity, presumably by increasing the soluble enzyme expression. We did not observe increases in activity from SSM at K342, a site that was previously identified in the context of a triple mutant. We again considered the possibility that SSM at distal sites might lead to improvements in total activity independent of any a priori knowledge about changes in promiscuity. We therefore screened a SSM library at another control site, Y418, where mutation to His was identified by global random mutagenesis as one that decreased activity without a significant shift in promiscuity. Unlike SSM at G373, this control site did not lead to increases in activity (**Fig. 9**).

We selected the top variants from five sites (16 variants) to re-screen outside of competition with longer reaction times (**Fig. 10**). Of the 16 variants four had reproducible increases in activity on at least one ketone substrate after long reaction times (16h). While only 25% of the identified variants were activating in this context, three of the five sites identified had at least one reproducible hit. Hence, 60% of the sites identified with SUMS promiscuity shift data resulted in a bonafide activated ketone variant. Notably, D86V displays 2.5 and 1.5-fold boosts in activity with **3.4** and **3.5** respectively, which matches well with the competition promiscuity profile.

We note that not all the promiscuity profiles matched well with the screening data. The differences between the competition promiscuity profiles (**Fig. 8**) and the single substrate activity (**Fig. 10**) could be due to differences in reaction time. The libraries were screened at early timepoints (1 h) to maximize differences in promiscuity, which has the potential to select for enzymes that are initially activating for ketones but are unstable and become inactive quickly. In contrast, validations were conducted overnight (16 h) to challenge the stability of the variants to the reaction conditions and identify variants that persist in solution. Taken together, these data demonstrate how SUMS can aid in the identification of variants with distal, cooperative interactions influencing reactivity. By further mutagenizing these sites, this two-step process allowed identification of activating mutations with no prior information about protein structure, dynamics, or evolution. While future studies may untangle the basis of these activating effects, the present investigation is focused on the development of a practical engineering approach that incorporates promiscuity information, which we continued by deploying another common step in enzyme evolution: recombination of activating mutations.

Identification of a cooperative mutational effects in a recombination library

There are many successful strategies for designing and screening recombination libraries. We considered a sequence space comprising five different positions (F75, D86, I141, V330, S407) that are distributed across the protein structure. We used screening data to identify degenerate codons that limit the inclusion of mutations that are deleterious when introduced independently. The resulting library consisted of ~2,800 possible variants (See **Table 2**). With this space, we again acknowledge a tradeoff between library size, fitness, and screening intensity. For this research, it is not necessary to exhaustively screen all possible combinations, but rather to use a carefully crafted library to efficiently traverse the largest recombination space without over-sampling inactive catalysts.

Table 2. Distal recombination library amino acid residues possible at each site.

site	F75	D86	I141	V330	S407
mutation	F	D	I	Q	S
	A	I	V	A	T
	V	V	M	R	Q
		N		E	A
				P	E
				L	
				G	
				V	
size	3	4	3	8	5

At this stage of evolution, we maintained the substrate mixture from the previous round and increased reaction time from one hour to eight hours to reduce the likelihood of selecting destabilizing mutations or those that accelerate decomposition of product through retro-aldol cleavage. We hypothesized that increasing the reaction time would also lead to greater consistency between competition activity and single substrate activity. We screened 704 clones, representing ~24% of the theoretical recombination space, and found many generally activated variants **(Fig. 11A)** The majority of the activated variants displayed a modest ~2-fold increase in total activity compared to QE. We validated a subset of the hits using purified enzyme in triplicate using single substrates **(Fig. 11B)**. As a control, D86V was included in the validation to ensure that the chosen recombination variants had higher ketone activity than a single point mutant. All hits displayed increased activity for multiple ketone electrophiles when validated. This result confirms our hypothesis that increasing

reaction time increases consistency between competition and single substrate activity. Of the proteins, three variants dubbed AVILE, AVIQE, and AIIRQ, showed a ~2-fold increase in activity with all three representative ketone electrophiles while slightly decreasing activity with *o*-tolualdehyde. Of these enzymes, the quadruple mutant AIIRQ had higher soluble expression (~80 mg protein/ L culture) compared to QE and the other variants (~50 mg protein/ L culture). Therefore, AIIRQ was chosen as the new parent enzyme for subsequent evolution.

Targeted mutagenesis in the active site reveals two mutants with distinct promiscuity profiles

In addition to distal hotspots, the active site provides a trove of sites capable of shifting activity. Previously, we identified a loop region in the UstD active site encompassing residues 391-393 which was key to unlocking high activity with challenging aldehydes (Ellis et al. 2022, *supra*). We hypothesized retargeting the 391-393 residues would reveal high activity ketone variants. Two additional residues, M299 and T388 were also considered because both appear to have side chains that protrude into the active site **(Fig. 12)**.

Simple modelling of potential mutations at T388 indicated that smaller residues would minimize the likelihood of blocking the active site. Residue M299 appears to be positioned in the substrate channel where the electrophile substrate approaches the nucleophilic enamine intermediate. We had previously screened a SSM library at this site with QE as the parent, which showed conservative mutations to hydrophobic residues were generally neutral to activating **(Fig. 13)**. With this information, we designed a focused recombination library of active site variants comprising ~2,300 variants **(Table 3)**. Because cooperative effects are particularly common in enzyme active sites, there is a higher propensity for multiple active site mutations to be deleterious, raising the specter of laborious screening of predominately inactive sequence space. We therefore biased our search to lower mutagenesis rates by constructing the library such that the parent codon is re-sampled at greater frequency (see methods for details). This strategy decreased the expected mutational load from 3.9 to 3.3 mutations per variant.

Table 3S.2. Active site recombination library amino acid residues possible at each site.

site	M299	T388	T391	L392	M393
mutation	M	T	T	L	M
	L	V	F	V	W

	V	I A	S L P I	A S	F L R C I S
size	3	4	6	4	8

We screened 968 clones within this focused library, comprising 42% of the theoretical sequence space. These results showed much greater diversity in the promiscuity profile from SUMS, consistent with the role of the active site in determining substrate specificity (**Fig. 14**). From these variants, we identified seven with distinct specificity profiles, each of which are activating with the ketone electrophiles, which was consistent in the single substrate validation (**Fig. 15**). Notably, these activated enzymes contained an average of 3.1 mutations, in line with our hypothesis that higher mutational rates would enrich the recombination space in inactive variants. Each of these seven active site variants have distinct sequence and activity profiles and may be stimulating sources of future inquiry. Nevertheless, our focus is on developing practical catalysts with high activity and the broadest possible substrate scope. For this reason, we selected two enzymes with distinct reactivity profiles, 7G11 (M299V, T391S, M393W) and 7B05 (T391S, M393F) for further exploration.

Lineage analysis of enzyme generality during promiscuity-guided evolution

Characterizing key variants in an evolutionary lineage is often used to show how successive rounds of mutagenesis affect activity. Such retrospective analysis is typically limited to the single transformation that was under selective pressure, with the notable exception of evolutions deploying a substrate walking strategy. We wanted to perform the same type of analysis but focus on characterizing how promiscuity changed through evolution. To accomplish this goal, activity of several different substrates was measured to understand how the promiscuity-guided evolutionary process led to the activity observed by the final variants.

In order to compare activity under optimized conditions, we performed a brief survey of reaction conditions using the final variant 7G11. Increasing the concentration of l-Asp from 50 to 250 mM increased yields with both unactivated and trifluoromethyl ketones (**Fig. 16**). While such high concentrations are not ideal, the amino acid is cheap and

commercially available. We further observed that additional equivalents of PLP relative to enzyme (between 10-50-fold excess) were beneficial for all substrates.

Using these optimized conditions, we found there were significant changes in activity for substrates that were under direct selective pressure (4-fluorophenylacetone and TFMK) and on those that were not (benzaldehyde and furylacetone). Cumulatively, these data for total activity with a single substrate show how evolution led toward general aldolase activity.

Measurement of the lineage with benzaldehyde, the model substrate used to engineer the parent enzyme UstD$^{2.0}$, provides insight into how activity with this substrate changed throughout evolution (**Fig. 17**). Benzaldehyde was too reactive with the enzyme for inclusion in the multiplexing mixtures, meaning it was not under direct selective pressure during evolution. UstD$^{2.0}$ performs 9,700 turnovers with benzaldehyde. This activity is in good agreement with previously reported turnover numbers (Ellis et al., 2022, *supra*). Activity on this substrate decreased through the evolution with AIIRQ performing 3,800 turnovers. The final variants had very different reactivity on benzaldehyde. Compared to AIIRQ, 7G11 decreased activity with only 1,800 turnovers performed whereas 7B05 increased activity with 4,900 turnovers performed. While data are plotted on a log scale, we emphasize that the changes in activity on benzaldehyde are statistically significant. Overall, through evolution, 7G11 and 7B05 display a ~5-fold and ~2-fold decrease in activity with benzaldehyde.

Throughout evolution, activity on the different ketones steadily increased. Beginning with the TFMK, UstD$^{2.0}$ performed a modest 480 turnovers with this substrate. The distal recombinant, AIIRQ, increased the TTN to 1,900 turnovers. The top TFMK variant, 7B05, performed 2,120 turnovers, representing ~4-fold increase in activity on this ketone. The UstD$^{2.0}$ activity with unactivated ketones was particularly low, 126 turnovers with furylacetone and trace reactivity with (4-fluorophenyl)acetone. AIIRQ performs 370 turnovers with furylacetone and 18 turnovers with (4-fluorophenyl)acetone. The final variants have even higher activity with these substrates; 7G11 performs 740 turnovers with furylacetone and 67 turnovers with (4-fluorophenyl)acetone while 7B05 performs 580 turnovers with furylacetone and 29 turnovers with (4-fluorophenyl)acetone. Through our evolution, we enhanced the reaction with (4-fluorophenyl)acetone from stoichiometric to catalytic representing a 67-fold increase in activity. Even furylacetone, which was never screened against our variants, increased activity ~6-fold through evolution. Most gratifyingly, these data show concretely how promiscuity-guided evolution is an effective strategy for developing generalists.

Materials and Methods

All chemicals and reagents were purchased from commercial suppliers (Sigma-Aldrich, VWR, Chem-Impex International, Alfa Aesar, Combi-blocks, Oakwood Products) at the highest quality available and used without further purification unless stated otherwise. Genes were purchased as gBlocks from Integrated DNA Technologies (IDT). *E. coli* cells were electroporated with an Eppendorf E-porator at 2500 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. UV-vis spectroscopic measurements were collected on a UV-2600 Shimadzu spectrophotometer. Optical density measurements were collected using an optical density reader (Amersham Biosciences). Ultra-high pressure liquid chromatography-mass spectrometry (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 aromatic substrates, or an Intrada Amino Acid column (Imtakt) for aliphatic substrates.

Plasmid Preparations

A 5-mL overnight culture of *E. coli* harboring the plasmid of interest was grown overnight at 37 °C with shaking at 200 rpm. The plasmid was isolated and purified using Zymo Plasmid Miniprep or Macherey-Nagel kits and sequenced through Functional Biosciences.

Protein and DNA sequences

The UstD protein of *Aspergillus flavus* and its variants are provided as SEQ ID NOs:1-6. Exemplary coding sequences of the UstD protein and its variants are provided as SEQ ID NOs:7-14. The constructs used in the experiments included a C-terminal 6×His tag with an LE linker (LEHHHHHH; SEQ ID NO:15). In SEQ ID NOs:3-6, mutations relative to $UstD^{2.0}$ are bolded and underlined. In SEQ ID NOs:9-14, mutations relative to $UstD^{2.0}$ are bolded and underlined.

SEQ ID NO:1

UstD (*Aspergillus flavus*)

MKSVATSSLDDVDKDSVPLGSSINGTAQAETPLENVIDVESVRSHFPVLGGETAAFNNASGTVVLKE
AIESTSNFMYSFPFPPPGVDAKSMEAITAYTGNKGKVAAFINALPDEITFGQSTTCLFRLLGLSLKPM
LNNDCEIVCSTLCHEAAASAWIHLSRELGITIKWWSPTTTPNSPDDPVLTTDSLKPLLSPKTRLVTC
NHVSNVVGTIHPIREIADVVHTIPGCMLIVDGVACVPHRPVDVKELDVDFYCFSWYKLFGPHLGTLY

ASRKAQDRYMTSINHYFVSSSSLDGKLALGMPSFELQLMCSPIVSYLQDTVGWDRIVRQETVLVTIL
LEYLLSKPSVYRVFGRRNSDPSQRVAIVTFEVVGRSSGDVAMRVNTRNRFRITSGICLAPRPTWDVL
KPKSSDGLVRVSFVHYNTVEEVRAFCSELDEIVTRDT

SEQ ID NO:2

$UstD^{2.0}$

MKSVATSSLDDVDKDSVPLGSSINGTAQAETPLENVIDVESVRSHFPVLGGETAAFNNASGTVVLKE
AIESTSNFMYSFPFPPGVDAKSMEAITAYTGNKGKVAAFINALPDEITFGQSTTALFRLLGLSLKPM
LNNDCEIVCSTLCHEAAASAWIHLSRELGITIKWWSPTTTPNSPDDPVLTTDSLKPLLSPKTRLVTC
NHVSNVVGTIHPIREIADVVHTIPGAMLIVDGVASVPHRPVDVKELDVDFYCFSWYKLFGPHLGTLY
ASRKAQDRYMTSINHYFVSSSSLDGKLALGMPSFELQLMCSPIVSYLQDTVGWDRIVRQETVLVTIL
LEYLLSKPSVYRVFGRRNSDPSQRVAIVTFEVVGRSSGDVAMRVNTRNRFRITSGTLMAPRPTWDVL
KPKSSDGLVRVSFVHYNTVEEVRAFCSELDEIVTRDT

SEQ ID NO:3

$UstD^{QE}$ ($UstD^{2.0}$ + P82Q + G373E)

MKSVATSSLDDVDKDSVPLGSSINGTAQAETPLENVIDVESVRSHFPVLGGETAAFNNASGTVVLKE
AIESTSNFMYSFPF**Q**PGVDAKSMEAITAYTGNKGKVAAFINALPDEITFGQSTTALFRLLGLSLKPM
LNNDCEIVCSTLCHEAAASAWIHLSRELGITIKWWSPTTTPNSPDDPVLTTDSLKPLLSPKTRLVTC
NHVSNVVGTIHPIREIADVVHTIPGAMLIVDGVASVPHRPVDVKELDVDFYCFSWYKLFGPHLGTLY
ASRKAQDRYMTSINHYFVSSSSLDGKLALGMPSFELQLMCSPIVSYLQDTVGWDRIVRQETVLVTIL
LEYLLSKPSVYRVFGRRNSDPSQRVAIVTFEVVGRSS**E**DVAMRVNTRNRFRITSGTLMAPRPTWDVL
KPKSSDGLVRVSFVHYNTVEEVRAFCSELDEIVTRDT

SEQ ID NO:4

$UstD^{AIIRQ}$ ($UstD^{QE}$ + F75A + D86I + V330R + S407Q)

MKSVATSSLDDVDKDSVPLGSSINGTAQAETPLENVIDVESVRSHFPVLGGETAAFNNASGTVVLKE
AIESTSN**A**MYSFPF**Q**PGV**I**AKSMEAITAYTGNKGKVAAFINALPDEITFGQSTTALFRLLGLSLKPM
LNNDCEIVCSTLCHEAAASAWIHLSRELGITIKWWSPTTTPNSPDDPVLTTDSLKPLLSPKTRLVTC
NHVSNVVGTIHPIREIADVVHTIPGAMLIVDGVASVPHRPVDVKELDVDFYCFSWYKLFGPHLGTLY
ASRKAQDRYMTSINHYFVSSSSLDGKLALGMPSFELQLMCSPIVSYLQDTVGWDRIVRQET**R**LVTIL
LEYLLSKPSVYRVFGRRNSDPSQRVAIVTFEVVGRSS**E**DVAMRVNTRNRFRITSGTLMAPRPTWDVL
KPKS**Q**DGLVRVSFVHYNTVEEVRAFCSELDEIVTRDT

SEQ ID NO:5

$UstD^{7G11}$ ($UstD^{AIIRQ}$ + M299V + T391S + M393W)

MKSVATSSLDDVDKDSVPLGSSINGTAQAETPLENVIDVESVRSHFPVLGGETAAFNNASGTVVLKE
AIESTSN**A**MYSFPF**Q**PGV**I**AKSMEAITAYTGNKGKVAAFINALPDEITFGQSTTALFRLLGLSLKPM

LNNDCEIVCSTLCHEAAASAWIHLSRELGITIKWWSPTTTPNSPDDPVLTTDSLKPLLSPKTRLVTC
NHVSNVVGTIHPIREIADVVHTIPGAMLIVDGVASVPHRPVDVKELDVDFYCFSWYKLFGPHLGTLY
ASRKAQDRYMTSINHYFVSSSSLDGKLALG**V**PSFELQLMCSPIVSYLQDTVGWDRIVRQET**R**LVTIL
LEYLLSKPSVYRVFGRRNSDPSQRVAIVTFEVVGRSS**E**DVAMRVNTRNRFRITSG**S**L**W**APRPTWDVL
KPKS**Q**DGLVRVSFVHYNTVEEVRAFCSELDEIVTRDT

SEQ ID NO:6

UstD^{7B05} (UstD^{AIIRQ} + T391S + M393F)

MKSVATSSLDDVDKDSVPLGSSINGTAQAETPLENVIDVESVRSHFPVLGGETAAFNNASGTVVLKE
AIESTSN**A**MYSFPF**Q**PGV**I**AKSMEAITAYTGNKGKVAAFINALPDEITFGQSTTALFRLLGLSLKPM
LNNDCEIVCSTLCHEAAASAWIHLSRELGITIKWWSPTTTPNSPDDPVLTTDSLKPLLSPKTRLVTC
NHVSNVVGTIHPIREIADVVHTIPGAMLIVDGVASVPHRPVDVKELDVDFYCFSWYKLFGPHLGTLY
ASRKAQDRYMTSINHYFVSSSSLDGKLALGMPSFELQLMCSPIVSYLQDTVGWDRIVRQET**R**LVTIL
LEYLLSKPSVYRVFGRRNSDPSQRVAIVTFEVVGRSS**E**DVAMRVNTRNRFRITSG**S**L**F**APRPTWDVL
KPKS**Q**DGLVRVSFVHYNTVEEVRAFCSELDEIVTRDT

SEQ ID NO:7

Codon-Optimized UstD Coding Sequence

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAATTTCATGTATAGCTTTCCTTTTCCCCCGGGTGTTGACGCTAAGTCAA
TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTTGTCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGTGAAATCGTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGATGCATGCTTATCGTTGACGGTGTGGCATGTGTCCCGCATCGTCCAGTTGATGTTAAAGA
ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGCATGCCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC
GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACTGTGCTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCCGGGGATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGAATTTGCCTGGCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGTAGCGACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACTAA

SEQ ID NO:8

DNA sequence of ustD$^{2.0}$

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAATTTCATGTATAGCTTTCCTTTTCCCCCGGGTGTTGACGCTAAGTCAA
TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTGCCCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGCGAAATCGTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGAGCGATGCTTATCGTTGACGGTGTGGCAAGCGTCCCGCATCGTCCAGTTGATGTTAAAGA
ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGCATGCCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC
GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACTGTGCTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCCGGGGATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGAACCTTAATGGCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGTAGCGACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACCCTCGAGCACCATCACCATCACCA
TTGA

SEQ ID NO:9

DNA sequence of ustD^{SA}

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAATTTCATGTATAGCTTTCCTTTT**TCC**CCGGGTGTTGACGCTAAGTCAA
TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTGCCCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGCGAAATCGTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGAGCCATGCTTATCGTTGACGGTGTGGCAAGCGTCCCGCATCGTCCAGTTGATGTTAAAGA

ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGCATGCCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC
GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACT**GCG**CTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCCGGGGATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGAACCTTAATGGCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGTAGCGACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACCCTCGAGCACCATCACCATCACCA
TTGA

SEQ ID NO: 10

DNA sequence of ustD^{Q}

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAATTTCATGTATAGCTTTCCTTTT**CAG**CCGGGTGTTGACGCTAAGTCAA
TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTGCCCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGCGAAATCGTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGAGCCATGCTTATCGTTGACGGTGTGGCAAGCGTCCCGCATCGTCCAGTTGATGTTAAAGA
ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGCATGCCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC
GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACTGTGCTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCCGGGGATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGAACCTTAATGGCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGTAGCGACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACCCTCGAGCACCATCACCATCACCA
TTGA

SEQ ID NO:11

DNA sequence of ustD^{QE}

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAATTTCATGTATAGCTTTCCTTTT**CAG**CCGGGTGTTGACGCTAAGTCAA
TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTGCCCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGCGAAATCGTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGAGCCATGCTTATCGTTGACGGTGTGGCAAGCGTCCCGCATCGTCCAGTTGATGTTAAAGA
ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGCATGCCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC
GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACTGTGCTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCC**GAG**GATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGAACCTTAATGGCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGTAGCGACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACCCTCGAGCACCATCACCATCACCA
TTGA

SEQ ID NO:12

DNA sequence of ustD^{AIIRQ}

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAAT**GCA**ATGTATAGCTTTCCTTTT**CAG**CCGGGTGTT**ATC**GCTAAGTCAA
TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTGCCCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGCGAA**ATT**GTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGAGCCATGCTTATCGTTGACGGTGTGGCAAGCGTCCCGCATCGTCCAGTTGATGTTAAAGA
ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGCATGCCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC

GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACT**CGA**CTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCC**GAG**GATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGAACCTTAATGGCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGT**CAA**GACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACCCTCGAGCACCATCACCATCACCA
TTGA

SEQ ID NO:13

DNA sequence of ustD^{7B05}

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAAT**GCA**ATGTATAGCTTTCCTTTT**CAG**CCGGGTGTT**ATC**GCTAAGTCAA
TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTGCCCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGCGAA**ATT**GTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGAGCCATGCTTATCGTTGACGGTGTGGCAAGCGTCCCGCATCGTCCAGTTGATGTTAAAGA
ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGCATGCCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC
GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACT**CGA**CTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCC**GAG**GATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGA**TCC**TTA**TTT**GCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGT**CAA**GACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACCCTCGAGCACCATCACCATCACCA
TTGA

SEQ ID NO:14

DNA sequence of ustD^{7G11}

ATGAAGAGCGTAGCGACGAGTTCCCTTGATGACGTAGATAAAGATTCCGTCCCCCTGGGCAGTTCGA
TCAATGGCACTGCACAAGCGGAAACTCCGCTGGAGAATGTGATCGACGTCGAATCAGTGCGCTCACA
TTTCCCGGTATTAGGGGGGGAAACGGCCGCGTTTAACAATGCATCAGGAACCGTAGTTTTGAAGGAG
GCAATTGAATCGACTTCAAAT**GCA**ATGTATAGCTTTCCTTTT**CAG**CCGGGTGTT**ATC**GCTAAGTCAA

TGGAGGCTATTACCGCATATACGGGGAATAAGGGCAAGGTTGCGGCATTTATCAATGCACTTCCTGA
TGAAATTACATTCGGGCAGTCCACAACTGCCCTGTTCCGTTTATTAGGTCTGTCGCTTAAACCTATG
CTGAATAACGATTGCGAA**ATT**GTATGCTCAACATTATGTCACGAAGCAGCAGCTTCCGCATGGATTC
ATTTAAGTCGCGAATTAGGAATTACCATTAAGTGGTGGAGCCCAACTACTACACCGAATAGTCCCGA
TGATCCAGTTCTGACGACTGACTCATTGAAGCCCTTGCTTAGTCCAAAAACGCGCCTTGTTACATGT
AATCACGTGTCGAATGTTGTAGGAACCATCCACCCTATTCGTGAGATTGCCGACGTGGTACATACCA
TTCCTGGAGCCATGCTTATCGTTGACGGTGTGGCAAGCGTCCCGCATCGTCCAGTTGATGTTAAAGA
ATTGGATGTAGATTTTTACTGCTTTTCCTGGTACAAGTTGTTCGGACCGCATCTTGGAACCCTGTAT
GCTTCCCGCAAAGCCCAAGACCGCTATATGACCTCAATTAACCATTACTTCGTCTCATCGTCGAGCC
TTGATGGTAAGCTGGCATTAGGC**GTG**CCGTCCTTTGAACTGCAGTTGATGTGCTCTCCAATTGTTTC
GTATTTGCAAGATACGGTGGGCTGGGACCGTATCGTGCGCCAAGAGACT**CGA**CTGGTAACTATTTTG
TTGGAGTATTTACTTAGCAAGCCATCTGTATATCGTGTGTTCGGACGTCGTAATTCTGATCCCAGTC
AGCGTGTAGCAATCGTAACTTTTGAAGTCGTGGGACGTAGTTCC**GAG**GATGTGGCAATGCGCGTAAA
TACGCGTAATCGCTTCCGCATTACCTCTGGA**TCC**TTA**TGG**GCACCGCGCCCGACATGGGACGTCTTG
AAACCGAAGAGT**CAA**GACGGACTTGTTCGCGTCAGCTTTGTACATTACAACACGGTTGAGGAAGTGC
GTGCGTTCTGCAGCGAGTTAGACGAGATTGTGACACGCGACACCCTCGAGCACCATCACCATCACCA
TTGA

Table 4. Primer Sequences. All primers were purchased from Integrated DNA Technologies.

Protein	Forward Primer (5' to 3')	Reverse Primer (5' to 3')
pET22b(+)-UstD$^{2.0}$	GAAATAATTTTGTTTAACTTTAAG AAGGAGATATACATATG (SEQ ID NO:16)	GCCGGATCTCAATGGTGATGG TGATGGTGCTCGAG (SEQ ID NO:17)
D86X	TTTCCTTTTCAGCCGGGTGTTNNN GCTAAGTCAATGGAGGC (SEQ ID NO:18)	AACACCCGGCTGAAAAGGAA (SEQ ID NO:19)
F75X	GGCAATTGAATCGACTTCAAATNN NATGTATAGCTTTCCTTTTCCCCC (SEQ ID NO:20)	ATTTGAAGTCGATTCAATTGC CTCCTTC (SEQ ID NO:21)
G101X	ACCGCATATACGGGGAATAAGNN NAAGGTTGCGGCATTTATCAATGC (SEQ ID NO:22)	CTTATTCCCCGTATATGCGGT AATAGCCTC (SEQ ID NO:23)
G373X	GAAGTCGTGGGACGTAGTTCCNNN GATGTGGCAATGCGCGTAAATACG (SEQ ID NO:24)	GGAACTACGTCCCACGACTTC AAAAGTTA (SEQ ID NO:25)

I141X	CTATGCTGAATAACGATTGCGAANNNGTATGCTCAACATTATGTCACGAAG (SEQ ID NO:26)	TTCGCAATCGTTATTCAGCATAGGTTTAA (SEQ ID NO:27)
K342X	GTTGGAGTATTTACTTAGCNNNCCATCTGTATATCGTGTGTTCGG (SEQ ID NO:28)	GCTAAGTAAATACTCCAACAAAATAGTTACC (SEQ ID NO:29)
M299X	GATGGTAAGCTGGCATTAGGCNNNCCGTCCTTTGAACTGCAGTTGA (SEQ ID NO:30)	GCCTAATGCCAGCTTACCATCAAGGC (SEQ ID NO:31)
P80X	CAAATTTCATGTATAGCTTTNNNTTTCCCCCGGGTGTTGAC (SEQ ID NO:32)	AAAGCTATACATGAAATTTGAAGTCGATTC (SEQ ID NO:33)
P82X	CATGTATAGCTTTCCTTTTNNNCCGGGTGTTGACGCTAAGTC (SEQ ID NO:34)	AAAAGGAAAGCTATACATGAAATTTGAAGTCG (SEQ ID NO:35)
P83X	TTTCATGTATAGCTTTCCTTTTCAGNNNGGTGTTGACGCTAAGTCAATGGAGGC (SEQ ID NO:36)	CTGAAAAGGAAAGCTATACATGAAATTTGAAGTCG (SEQ ID NO:37)
S407X	CGTCTTGAAACCGAAGAGTNNNGACGGACTTGTTCGCGTCAG (SEQ ID NO:38)	ACTCTTCGGTTTCAAGACGTCC (SEQ ID NO:39)
V330X	GTATCGTGCGCCAAGAGACTNNNCTGGTAACTATTTTGTTGGAG (SEQ ID NO:40)	AGTCTCTTGGCGCACGATACGGTCCC (SEQ ID NO:41)
Y277X	CCCGCAAAGCCCAAGACCGCNNNATGACCTCAATTAACCATTACTTCG (SEQ ID NO:42)	GCGGTCTTGGGCTTTGCGGG (SEQ ID NO:43)
Y418X	GTTCGCGTCAGCTTTGTACATNNNAACACGGTTGAGGAAGTGCG (SEQ ID NO:44)	ATGTACAAAGCTGACGCGAACAAG (SEQ ID NO:45)
Y96X	GTCAATGGAGGCTATTACCGCANNNACGGGGAATAAGGGCAAGGT (SEQ ID NO:46)	TGCGGTAATAGCCTCCATTG (SEQ ID NO:47)
F75[GYA]	GGCAATTGAATCGACTTCAAATGY	ATTTGAAGTCGATTCAATTGC

	AATGTATAGCTTTCCTTTTCAGCC (SEQ ID NO:48)	CTCCTTC (SEQ ID NO:49)
D86[RWC]	GCTTTCCTTTTCAGCCGGGTGTT**R** **WC**GCTAAGTCAATGGAGGCTATT AC (SEQ ID NO:50)	AACACCCGGCTGAAAAGGAA (SEQ ID NO:51)
I141[RTK]	CCTATGCTGAATAACGATTGCGAA **RTK**GTATGCTCAACATTATGTCAC GAAGC (SEQ ID NO:52)	TTCGCAATCGTTATTCAGCAT AGGTTTAA (SEQ ID NO:53)
V330[SNA]	CCGTATCGTGCGCCAAGAGACT**SN** **A**CTGGTAACTATTTTGTTGGAGTA TTTAC (SEQ ID NO:54)	AGTCTCTTGGCGCACGATACG GTCCC (SEQ ID NO:55)
S407[SAA]	CGTCTTGAAACCGAAGAGT**SAA**GA CGGACTTGTTCGCGTCAGC (SEQ ID NO:56)	ACTCTTCGGTTTCAAGACGTC C (SEQ ID NO:57)
S407[DCA]	CGTCTTGAAACCGAAGAGT**DCA**G ACGGACTTGTTCGCGTCAGC (SEQ ID NO:58)	ACTCTTCGGTTTCAAGACGTC C (SEQ ID NO:59)
TTLM_4 mutation	CGCGTAATCGCTTCCGCATT**RYA**T CTGGA**HYCDYAWKK**GCACCGCG CCCGACATGGGACG (SEQ ID NO:60)	AATGCGGAAGCGATTACGCGT ATTTACGC (SEQ ID NO:61)
TTLM_3 mutation_T388	CGCGTAATCGCTTCCGCATTACCT CTGGA**HYCDYAWKK**GCACCGCG CCCGACATGGGACG (SEQ ID NO:62)	AATGCGGAAGCGATTACGCGT ATTTACGC (SEQ ID NO:63)
TTLM_3 mutation_T391	CGCGTAATCGCTTCCGCATT**RYA**T CTGGAACC**DYAWKK**GCACCGCGC CCGACATGGGACG (SEQ ID NO:64)	AATGCGGAAGCGATTACGCGT ATTTACGC (SEQ ID NO:65)
TTLM_3 mutation_L392	CGCGTAATCGCTTCCGCATT**RYA**T CTGGA**HYC**TTA**WKK**GCACCGCGC CCGACATGGGACG (SEQ ID NO:66)	AATGCGGAAGCGATTACGCGT ATTTACGC (SEQ ID NO:67)
TTLM_3 mutation_M393	CGCGTAATCGCTTCCGCATT**RYA**T CTGGA**HYCDYA**ATGGCACCGCGC CCGACATGGGACG (SEQ ID NO:68)	AATGCGGAAGCGATTACGCGT ATTTACGC (SEQ ID NO:69)
M299 [DTG]	GATGGTAAGCTGGCATTAGGC**DT** **G**CCGTCCTTTGAACTGCAGTTG	GCCTAATGCCAGCTTACCATC (SEQ ID NO:71)

	(SEQ ID NO:70)	
H263X	CCTGGTACAAGTTGTTCGGACCGN NNCTTGGAACCCTGTATGCTTCCC GC (SEQ ID NO:72)	CGGTCCGAACAACTTGTACCA GGAAAAGC (SEQ ID NO:73)
H283X	CCAAGACCGCTATATGACCTCAAT TAACNNNTACTTCGTCTCATCGTC GAGCCTTG (SEQ ID NO:74)	
F285X	GCTATATGACCTCAATTAACCATT ACNNNGTCTCATCGTCGAGCCTTG ATGGTAAGCT (SEQ ID NO:75)	GTAATGGTTAATTGAGGTCAT ATAGCGG (SEQ ID NO:76)
P300X-7B05	GATGGTAAGCTGGCATTAGGCATG NNNTCCTTTGAACTGCAGTTGATG TGCTC (SEQ ID NO:77)	CATGCCTAATGCCAGCTTACC ATCA (SEQ ID NO:78)
P300X-7G11	GATGGTAAGCTGGCATTAGGCGTG NNNTCCTTTGAACTGCAGTTGATG TGCTC (SEQ ID NO:79)	CACGCCTAATGCCAGCTTACC ATCA (SEQ ID NO:80)
S389X-7B05	TACGCGTAATCGCTTCCGCATTAC CNNNGGATCCTTATTTGCACCGCG CCCG (SEQ ID NO:81)	GGTAATGCGGAAGCGATTACG CGTA (SEQ ID NO:82)
S389X-7G11	TACGCGTAATCGCTTCCGCATTAC CNNNGGATCCTTATGGGCACCGCG CCCG (SEQ ID NO:83)	GGTAATGCGGAAGCGATTACG CGTA (SEQ ID NO:84)

* **NNN** indicates a 22-codon library made as a mixture of 3 degenerate codon primers (NDT, VHG, TGG), as described by Kille et al. "Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis." *ACS Synth. Biol.* 2013, 2, 83-92.

DNA Isolation and Storage

DNA was purified via gel electrophoresis and isolated using a DNA gel extraction kit (Zymo Research or Macherey-Nagel). All isolated DNA was stored at -20 °C.

Protein Expression & Purification

Optimized Expression of UstD$^{2.0}$ and variants: An overnight culture of *E. coli* BL21(DE3) harboring a pET-22b(+) plasmid encoding a given UstD$^{2.0}$ variant was created by inoculating 10 mL of TB_{amp} media with a single colony. This culture was shaken at 37 °C

and 200 rpm for ~16 h. 10 mL of overnight culture was then used to inoculate 1 L of TB_{amp}, which was shaken at 37 °C and 200 rpm for approximately 1.5 h or until an optical density (OD) 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 IPTG. The cultures were allowed to continue to grow for an additional ~16 h at 20 °C and shaking at 200 rpm. Cells were then harvested by centrifugation (4 °C, 30 min, 4,000 x*g*), and the cell pellets were stored at -20 °C overnight.

Protein Purification: To purify UstD, cell pellets were thawed at room temperature and then resuspended in lysis buffer, comprised of enzyme storage buffer (100 mM potassium phosphate buffer pH 7.0, 100 mM sodium chloride) containing 20 mM imidazole, 1 mg/mL Hen Egg White Lysozyme (GoldBio), 0.2 mg/mL DNase (GoldBio), 1 mM $MgCl_2$, and 0.5 mg/g cell pellet pyridoxal 5`-phosphate (PLP). A ratio of 4 mL 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 (30 s per g cell pellet, 2 s on, 2 s off, 40% amplitude). The resulting lysate was then spun down at 48,384 x*g* to pellet cellular debris. Ni/NTA beads were pre-equilibrated in storage buffer containing 20 mM imidazole. 1-2 mL of resin were used per 50 mL of lysis. The flow-through was re-passed once to collect any remaining beads from the original vessel. The collected beads were washed with 5 column volumes each of storage buffer containing 20 mM, 40 mM, and 60 mM imidazole. Protein was eluted with 3 column volumes of storage buffer containing 250 mM imidazole and collecting the flow-through until the eluent was no longer yellow (color due to the enzymatically bound PLP cofactor). Imidazole was then removed using a PD10 salt exchange column.

Protein Characterization and Storage

Concentration measurement: Enzyme concentration was determined by Bradford assay, using bovine serum albumin for a standard concentration curve.

Gel Electrophoresis: Protein purity was analyzed by sodium dodecyl sulfate-polyacrylamide (SDS-PAGE) gel electrophoresis using 12% polyacrylamide gels.

Storage: Purified enzyme was flash frozen in pellet form by pipetting enzyme dropwise into a crystallization dish filled with liquid nitrogen. The enzyme was transferred to a plastic conical and stored at -80 °C until further use. Frozen pellets were thawed at room temperature and centrifuged before use.

Library Generation for Directed Evolution

Production of UstD random mutagenesis libraries: Random mutagenesis was carried out via error-prone PCR. Reaction conditions were optimized to generate 1-2 codon mutations per plasmid. Reactions were setup by adding the following to a PCR tube: 5 µL 10x Taq buffer (New England Biolabs), 1 µL 10 mM dNTP mix, 1 µL 10 µM 22b-intF, 1 µL 10 µM 22b-intR, 1 µL ~100 ng/µL parent plasmid, 5.5 µL 50 µM $MgCl_2$, 7.5 µL 100 µM $MnCl_2$, 1 µL DMSO, 0.5 µL Taq polymerase (New England Biolabs) and diluted to a total volume of 50 µL with milliQ H_2O. Reactions were carried out in a thermocycle according to the following scheme: (1) Step 1: 95 °C 2 min 30 s; (2) Step 2: 95 °C 15 s; (3) Step 3: 54 °C 20 s; (4) Step 4: 68 °C 1 min 45 s; (5) Step 5: 68 °C 5 min. Extension steps 2 – 4 were performed for 30 cycles.

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. BL21 (DE3) E. coli cells were subsequently transformed with the resulting cyclized DNA product via electroporation. After 45 min of recovery in Terrific Broth (TB) media at 37 °C, 200 rpm, cells were plated onto LB plates with 100 µg/mL Ampicillin (amp) and incubated overnight. Single colonies were used to inoculate 5 mL TB + 100 µg/mL amp (TB_{amp}), which were grown overnight at 37 °C, 200 rpm. Colonies were sequenced and there was an average of 2 coding mutations.

Production of UstD degenerate codon libraries: Primers containing degenerate codons were purchased from IDT and are listed above. Mutagenesis was carried out via overlap-extension PCR. Reactions were setup by adding the following to a PCR tube: 10 µL 5x HF buffer (New England Biolabs), 1 µL 10 mM dNTP mix, 1 µL 10 µM forward primer, 1 µL 10 µM reverse primer, 1 µL ~100 ng/µL parent plasmid, 1 µL DMSO, 1 µL Phusion polymerase (New England Biolabs) and diluted to a total volume of 50 µL with milliQ H_2O. Reactions were carried out in a thermocycle according to the following scheme: (1) Step 1: 98 °C 1 min; (2) Step 2: 98 °C 15 s; (3) Step 3: 54 °C 20 s; (4) Step 4: 72 °C 1 min; (5) Step 5: 72 °C 5 min. Extension steps 2 – 4 were performed for at least 30 cycles.

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. BL21 (DE3) E. coli cells were subsequently transformed with the resulting cyclized DNA product via electroporation. After 45 min of recovery in Terrific Broth (TB) media at 37 °C, 200 rpm, cells were plated onto LB plates with 100 µg/mL Ampicillin (amp) and incubated overnight. Single colonies were used to inoculate 5 mL TB + 100 µg/mL amp (TB_{amp}), which were grown overnight at 37 °C, 200 rpm. Colonies were sequenced to confirm correct incorporation of desired mutations.

Enzymatic Activity Experiments

General procedure for library generation and screening: Mutagenized plasmid DNA was generated Electrocompetent BL21(DE3) were transformed with mutagenized plasmid DNA and allowed to recover for 45 min in 800 µL of Terrific Broth (TB). After recovery, the cells were plated onto LB plates containing 100 µg/mL ampicillin (LB_{amp}) and incubated overnight. A 96-well plate containing 600 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, 200 rpm. Expression plates were prepared with 600-610 µL of TB_{amp} per well and inoculated with 6-20 µL of overnight culture. Glycerol stocks of each starter plate well were made using 150 µL of the remaining culture and 100 µL of 60% sterile glycerol to ensure the sequence of any mutants of interest could be determined. The expression plates were grown at 37 °C, 200 rpm for 3 h. Expression plates were then placed on ice for 30 min. Cultures were induced with a final concentration of 0.1 mM IPTG in 70 µL of fresh TB_{amp} and, if necessary, diluted to a final volume of 700 µL with fresh TB_{amp}. The expression culture was grown overnight at 20 °C, 200 rpm. Following overnight growth, the plate was centrifuged (4,000 x*g*, 30 min, 4 °C) and all media was removed by striking plates against a paper towel on a table. Expression plates were stored at -20 °C until further use.

A reaction master mix containing a final concentration of 50 mM L-asp, 50 µM PLP, and buffer (100 mM KPi + NaCl, pH 7.0) was added to the thawed expression pellets using an Opentrons OT-2 liquid handling robot. The pellets were resuspended by vortexing. Then, 50 mM final concentration of electrophile mix (substrates varied throughout evolution), was added to the reaction mixture by Opentrons OT-2 robot and reactions were allowed to incubate at 37 °C, 200 rpm for the desired reaction time (1-8 h). Subsequently, reactions were quenched with 300 µL (1 reaction volume) acetonitrile using Opentrons OT-2 robot and clarified at 4,000 x*g* for 30 min. The supernatant was transferred to a 0.2 µm centrifuge filter plate (PALL) and filtered at 1,500 rpm for 10 min into a clean Waters 96-well UPLC plate before being sealed prior to analysis by UPLC-MS.

Specific library generation and screening conditions for global random mutagenesis and P80X, F75X, G373X, and P82X site saturation libraries: Library generation and screening for these libraries follows the general procedure laid out above. The parent enzyme was $UstD^{v2.0}$ for these libraries. Expression plates were prepared with 600 µL of TB_{amp} per well and inoculated with 6 µL of overnight culture. After induction, all expression plate wells had a final volume of 700 µL. The enzymatic reaction time was 1 h at 37 °C, 200

rpm with a 50 mM final concentration electrophile mix consisting of 4.2 mM thiophene-3-carboxyaldehyde, 4.2 mM *o*-tolualdehyde, and 41.7 mM (4-fluorophenyl)acetone. The ketone:aldehyde ratio is 5:1 for these libraries.

Specific library generation and screening conditions for M299X, P83X, Y96X, Y277X, G101X, D86X, K342X, S407X, V330X, I141X, Y418X site saturation libraries: Library generation and screening for these libraries follows the general procedure laid out above. The parent enzyme was UstD^{QE} for these libraries. Expression plates were prepared with 600 µL of TB_{amp} per well and inoculated with 6 µL of overnight culture. After induction, all expression plate wells had a final volume of 700 µL. The enzymatic reaction time was 1 h at 37 °C, 200 rpm with a 50 mM final concentration electrophile mix consisting of 10 mM (4-fluorophenyl)acetone, 32.5 mM 4'-nitroacetophenone, 5 mM *o*-tolualdehyde, 2.5 mM 1,1,1-trifluoro-3-phenyl-2-propanone. The ketone:aldehyde ratio is 9:1 for these libraries.

Specific library generation and screening conditions for distal recombination library: Library generation and screening for these libraries follows the general procedure laid out above. The parent enzyme was UstD^{QE} for these libraries. Expression plates were prepared with 600 µL of TB_{amp} per well and inoculated with 6 µL of overnight culture. After induction, all expression plate wells had a final volume of 700 µL. The enzymatic reaction time was 8 h at 37 °C, 200 rpm with a 50 mM final concentration electrophile mix consisting of 10 mM (4-fluorophenyl)acetone, 32.5 mM 4'-nitroacetophenone, 5 mM *o*-tolualdehyde, 2.5 mM 1,1,1-trifluoro-3-phenyl-2-propanone. The ketone:aldehyde ratio is 9:1 for these libraries.

Specific library generation and screening conditions for active site mutagenesis libraries: Library generation and screening for these libraries follows the general procedure laid out above. For DNA generation, an equimolar primer mix of TTLM_4 mutation, TTLM_3 mutation_T388, TTLM_3 mutation_T391, TTLM_3 mutation_L392, TTLM_3 mutation_M393 was used as a doping strategy to lower the mutagenesis rate to 3.3 mutations per variant. All other primers were used in the standard fashion associated with overlap extension PCR. The parent enzyme was AIIRQ for these libraries. Expression plates were prepared with 610 µL of TB_{amp} per well and inoculated with 20 µL of overnight culture. After induction, all expression plate wells had a final volume of 700 µL. The enzymatic reaction time was 6 h at 37 °C, 200 rpm with a 50 mM final concentration electrophile mix consisting of 10 mM (4-fluorophenyl)acetone, 32.5 mM 4'-nitroacetophenone, 5 mM *o*-tolualdehyde, 2.5 mM 1,1,1-trifluoro-3-phenyl-2-propanone. The ketone:aldehyde ratio is 9:1 for these libraries.

Analysis of mutagenesis libraries

The relative amount of product formed in the reactions compared to the positive control reaction was measured by single ion retention mass analysis via UPLC/MS. Given the relatively high variability in the parent signal in this assay, wells typically require an apparent 1.5-fold increase in product compared to the parent to be carried forward for validation. Using the glycerol stocks from the starter culture plate (described above), wells of interest could be streaked onto a fresh LB_{amp} plate for subsequent sequencing and validation.

Validations of mutants of interest

Every validated mutant of interest was validated by heterologous expression in duplicate or triplicate reactions using whole cell, cell lysate or Ni-NTA purified enzyme reactions.

Validations of F75A: A reaction master mix (225 µL) containing a final concentration of 50 mM L-asp, 50 µM PLP, and buffer (100 mM KPi + NaCl, pH 7.0) was added to epi tubes. (4-fluorophenyl)acetone (15 µL, 50 mM final concentration, 5% DMSO in reaction mixture) was added to the epi tubes. Frozen cell stocks of *E. coli* cells expressing $UstD^{2.0}$+F75A were thawed to room temperature. The reaction was initiated upon addition of 60 µL of cell suspension (20 mg/mL final concentration). The reactions were allowed to incubate at 37 °C, 200 rpm for 1 h. Subsequently, reactions were quenched with 300 µL (1 reaction volume) acetonitrile and clarified at 16160 xg for 10 min. The supernatant was transferred to a UPLC vial and analyzed by UPLC-MS. The reactions were done in triplicate technical replicates at the same time as parent to determine the fold change most accurately.

Validations of G373E, P82Q, QE: A reaction master mix containing a final concentration of 50 mM L-asp, 5 µM PLP, and buffer (100 mM KPi + NaCl, pH 7.0) was added to epi tubes. The electrophile (4-fluorophenyl)acetone (30 µL, 50 mM final concentration, 5% DMSO in reaction mixture) was added to the epi tubes. The enzyme catalyst was thawed to room temperature and clarified at 16160 xg for 3 min. The reaction was initiated upon addition of enzyme (0.01 mol% catalyst, 10,000 Max TON). The reactions were allowed to incubate at 37 °C overnight. Subsequently, reactions were quenched with 300 µL (1 reaction volume) acetonitrile and clarified at 16160 xg for 10 min. The supernatant was transferred to a UPLC plate and analyzed by UPLC-MS. The reactions were done in duplicate technical replicates at the same time as parent to determine the fold change most accurately.

Validations of P83T, P83Y, P83V, P83G, D86I, D86V, Y96V, Y96F, G101R, G101Q, G101A+ΔH445, Y277H, Y277F+W399C: A reaction master mix containing a final concentration of 50 mM L-asp, 50 μM PLP, and buffer (100 mM KPi + NaCl, pH 7.0) was added to epi tubes. The electrophile (4-fluorophenyl)acetone (15 μL, 50 mM final concentration, 5% DMSO in reaction mixture) was added to the epi tubes. The reaction was initiated upon addition of lysate (60 μL, 40 - 50 mg/mL final concentration). The reactions were allowed to incubate at 37 °C overnight. Subsequently, reactions were quenched with 300 μL (1 reaction volume) acetonitrile and clarified at 4300 rpm for 15 min. The supernatant (250 μL) was transferred to a filter plate (0.2 μm filter) and centrifuged at 1000 rpm into a UPLC plate and analyzed by UPLC-MS. The reactions were done in duplicate or triplicate technical replicates at the same time as parent to determine the fold change most accurately.

Validations of distal recombination and active site variants: A reaction master mix containing a final concentration of 50 mM L-asp, 5 μM PLP, and buffer (100 mM KPi, pH 7.0, 100 mM NaCl) was added to epi tubes. The electrophile (10 μL, 50 mM final concentration, 5% DMSO in reaction mixture) was added to the epi tubes. The enzyme catalyst was thawed to room temperature and clarified at 16160 x*g* for 3 min. The reaction was initiated upon addition of enzyme (0.01 mol% catalyst, 10,000 Max TON). The reactions were allowed to incubate at 37 °C for 16 h. Subsequently, reactions were quenched with 100 μL (1 reaction volume) acetonitrile and clarified at 4000 x*g* for 15 min. The supernatant was diluted and filtered through a 0.2 μm PALL filter into a UPLC plate and analyzed by UPLC-MS. The reactions were done in triplicate technical replicates at the same time as parent to determine the fold change most accurately. Only a single electrophile was use per reaction, not a mixture. The electrophiles tested were *o*-tolualdehyde, (4-fluorophenyl)acetone, 4'-nitroacetophenone, and 1,1,1-trifluoro-3-phenyl-2-propanone. To analyze each reaction by UPLC, the enzymatic reactions were diluted different amounts. The *o*-tolualdehyde reactions were diluted 50x. (4-fluorophenyl)acetone reactions were diluted 3.3x. 4'-nitroacetophenone reactions were diluted 3.3x. 1,1,1-trifluoro-3-phenyl-2-propanone reactions were diluted 100x.

Discussion of Directed Evolution Strategy

Table 5. Directed evolution summary with the sequences of variants with altered promiscuity or activity.

Round	Description	Clones Screened	Variants with altered promiscuity and/or activity	Best Variant(s)

			in screening	
1	**Random mutagenesis** of the entire gene	880	G373R, D86V, F75S, Y96C+G101R, P82S+V330A, P83R, Y277C+K342E+S407N, P80L, H263R, Y418H	**UstD$^{v2.0}$** All variants were generally deactivated but changed promiscuity. No mutations were fixed this round.
2	**Site saturation** of 'hotspots'with UstD$^{v2.0}$: F75, P80, P82, G373	352	**F75:** A, C, H, I, K, L, M, N, Q, R, S, T, W, Y **P80:** G, R **P82:** G, Q **G373:** E, R	**P82Q** **G373E** **F75A**
3	**Double mutant** P82Q+G373E	1	P82Q+G373E	**QE** = P82Q+G373E
4	**Site saturation** of 'hotspots' with QE as parent: P83, D86, Y96, G101, I141, Y277, V330, K342, S407, Y418,	880	**P83:** G, V, T, Y **D86:** I, V **G101:** F, Q, A+ΔH445 **I141:** V, M, **Y277:** H, F+W399C **V330:** Q, A, R, C **K342:** none **S407:** Q, T, A, E **Y418:** none * silent mutations were fixed through primer design I141 (ATC→ ATT)	None. Used mutational information from D86, I141, V330, and S407 in subsequent library
5	**Recombination** at sites F75, D86, I141, V330, S407	704	F75A+D86I+I141V+S407E F75A+D86I+V330R+S407Q F75V+D86N+V330V F75A+D86V+I141V+S407Q F75A+D86V+I141M+S407E F75V+D86V+S407Q F75A+D86V+V330L+S407E F75A+D86I+V330A+S407Q F75V+D86V+S407E	**AIIRQ**= QE+ F75A+ D86I+ I141I(ATC→ATT)+ V330R+ S407Q

			F75A+D86I+V330A+S407Q K2E+F75V+D86I+V330V F75A+D86V+V330Q+S407E *synonymous mutations not included	
6	**Recombination** at active site residues M299, T388, T391, L392, M393	968	M299L+T388V+T391F M299V+T388V M299V+T388V+T391S+L440P M299V+T388I+T391F+M393C M299L+T388I+T391S+M393C M299V+T388I T388V+T391F+M393S T388I+T391S T391S+M393W C201C+M299V+T391S+M393F M299L+T388A+T391F M299V+M393W M299V+T388V+M393W M299V+T388V+M393W T391S+M393F M299V+T388A+T391F+M393W M299V+T388I M299V+T391S+M393W M299V+T388V+T391F+M393F M299L+T391S+M393W M299V+T388I+T391S+M393F M299V+T388I+T391F	**7G11** = AIIRQ+ M299V+ T391S+ M393W **7B05** = AIIRQ+ T391S+ M393F

			M299V+T391S+L392A+M393W T388I V63I M299V+T391F+M393W M299V+T388I+M393F M299V+T388I+M393L M299V+T388A+T391F+M393W M299L T391F+M393W T391F+L392I+M393W T388A+T391F	
7	**Site saturation** of various residues both in and out of active site: Y257, H263, F285, P300, S389, H283. Each library was generated with 7G11 and 7B05 as the parent enzyme.	440	none	No variants surpassed 7G11 or 7B05 activity, so no further engineering was pursued.

Reaction Condition Optimization

Before testing additional substrates, we first attempted to optimize reaction conditions by observing the progress curve of a **[4FPhAT]** reaction with 7G11 in varying conditions. It was known to us that PLP is degraded over the course of the reaction via off-pathway deamination of alanine which is formed as a shunt product via protonation of the nucleophilic enamine intermediate. We hypothesized that increasing the amount of PLP in the reaction would allow for higher yield of the product. Therefore, we observed the reaction progress at varying amounts of PLP (10x, 20x, 50x). While each reaction appeared to reach the same yield, the 50x PLP reaction achieved the max yield and then the product was

degraded in the reaction time. This indicated to us that there was enough PLP in the reaction to build-up and degrade the desired product indicating a sufficient excess for enzyme catalysis. Therefore, we elected to run reactions at 50x PLP for unactivated ketones. Next, we varied the concentration of L-asp (50 mM, 100 mM, 250 mM) in the reaction mixture to attempt to competitively inhibit product reentry and degradation. Any amount of additional L-aspartate increased the yield of the amino acid product and kept degradation from occurring.

When we repeated the same experiments using **[TriFPhAT]** as the substrate, we observed a worse yield with 50x PLP than 10x PLP. Therefore, we chose to use 10x PLP for highly activated ketone substrates. However, the reaction reached higher yield upon addition of 250 mM L-asp. Therefore, we determined our standard conditions to be 250 mM L-asp, 50 mM ketone, 10x PLP for activated substrates, 50x PLP for unactivated substrates, 4 hour reaction time, 37 °C, 5% cosolvent, 100 mM KPi pH 7.0, 100 mM NaCl.

Ketone lineage analysis

General Procedure: All reactions were done in triplicate on analytical scale (200 µL). The buffer used for all enzymatic reactions was 100 mM KPi, pH 7.0, 100 mM NaCl. PLP and l-aspartate stock solutions were made in water. Electrophile stock solutions were made in DMSO and each reaction only has a single electrophile. All UstD variant stock solutions concentrations were quantified by Bradford assay prior to setting up reactions. The difference enzyme concentration between variants was corrected for prior to addition into the reactions. All samples were analyzed following Marfey's derivatization by Waters Acquity UPLC-MS using a BEH C18 column (Waters). To correct for small deviations in injection volume, an internal derivatization standard was included (0.098 mM l-arginine). Derivatized amino acid product quantitation was performed by integrating chromatograph peaks at 340 nm and corrected by dividing by the internal standard peak area. To calculate product concentrations, a standard curve was generated by subjecting stock solutions of l-phenylalanine (50 mM – 0.4 mM) in water to the identical procedure used to process and derivatize enzymatic reaction solutions, in tripliciate, with internal standard. These curves were used to calculate the concentrations of UstD products in solution, and subsequently total turnover numbers after dilution factor correction.

Enzymatic reactions: Each well of a 96-well plate was charged with electrophile (10 µmol, 1 equiv., 50 mM final concentration, 5% DMSO final concentration). A reaction master mixture containing l-aspartate sodium salt monohydrate (50 µmol, 5 equiv., 250 mM final concentration), pyridoxal-5`-phosphate (10 or 50 molar equivalents to final enzyme

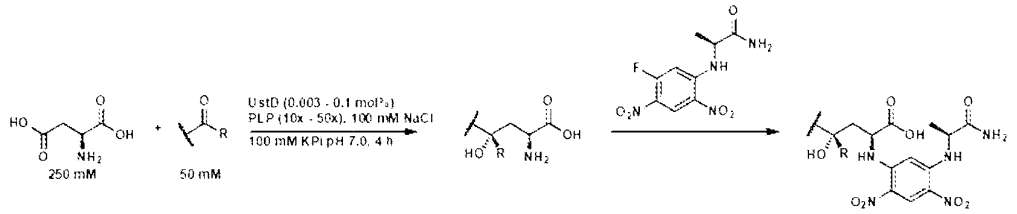
concentration, see **Table 6**) and buffer was aliquoted into each of the wells. The plate was vortexed gently to mix. Reactions were initiated by addition of UstD (0.1-0.003 mol% catalyst, 1,000-30,000 max turnover number, see **Table 6**) to bring the total reaction volume to 200 µL. The 96-well plate was sealed with a silicon lid and placed at 37 °C for 4 h. Reactions were quenched with 200 µL of acetonitrile (1 reaction volume) and diluted with 200 µL of 1:1 ACN:DI H_2O to homogenize reaction solutions. Denatured enzyme was removed by centrifugation at 2000 rpm for 10 min. The resulting supernatant was passed through a 0.2 µm PALL filter plate to remove any remaining particulates of enzyme prior to derivatization. Marfey's derivatization of the clarified enzymatic reactions were performed to quantify amino acid yield, results shown in **Table 7**.

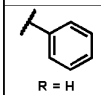
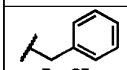
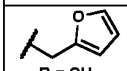
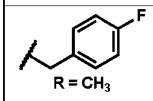
Marfey's derivatization procedure: To a fresh 96-well plate, 10 µL of quenched reaction mixture (2.7 mM total amines in reaction, 1 equiv., ~0.8 µmol total amines in reaction) 140 µL of 10.41 mM $NaHCO_3$ with 0.21 mM of l-arginine as an internal standard (9.7 mM $NaHCO_3$ final concentration, 3.5 equiv. base, 2.9 µmol $NaHCO_3$), and 10 mM l-FDAA (5 mM final concentration, 1.8 equiv., 1.5 µmol) were added in each well. The derivatization reaction was allowed to proceed at 37 °C for 18 h. The reactions were quenched with 300 µL of 60 mM HCl in acetonitrile (1 reaction volume) and analyzed via UPLC-MS no later than 24 h after quenching. Note that the amino acid products are susceptible to lactonization upon addition of the acid required to quench the reactions resulting in two product peaks with masses differing by 18 mass units. The linear product will display mass peaks for both the molecular ion and a dehydrated ion (-18 mass units), while the lactone will only display a mass signal associated with dehydration (-18 mass units). Turnover numbers were calculated based on the total integration of linear and lactonized amino acid product peaks at 340 nm.

Table 6. Lineage reaction conditions by electrophile.

Electrophile	**Molar equivalents PLP**	**Max TON**	**mol% catalyst**
benzaldehyde	10	30000	0.003
1,1,1-trifluoro-3-phenyl-2-propanone	10	5000	0.02
2-furylacetone	50	2500	0.04
(4-fluorophenyl)acetone	50	1000	0.1

Table 7. Turnover numbers for ketone lineage.



Variant	UstD$^{2.0}$	SA	Q	QE	AIIRQ	7G11	7B05
Product	Avg TTN	Avg TTN	Avg TTN	Avg TTN	Avg TTN	Avg TTN	Avg TTN
 R = H	9700 ± 500	3900 ± 200	4300 ± 100	4300 ± 300	3800 ± 200	1800 ± 100	4900 ± 200
 R = CF_3	480 ± 50	1110 ± 90	1000 ± 100	1070 ± 40	1900 ± 100	1560 ± 50	2120 ± 40
 R = CH_3	126 ± 6	160 ± 10	160 ± 10	200 ± 20	370 ± 30	740 ± 10	580 ± 10
 R = CH_3	trace	~ 10	~ 10	17 ± 1	18 ± 1	67 ± 6	29 ± 2

Example 2. Biocatalytic synthesis of chiral tertiary alcohol non-standard amino acids

Introduction

Chiral tertiary alcohols represent a class of industrially relevant compounds. Non-standard amino acids (nsAAs) bearing a chiral tertiary alcohol have shown to have a variety of industrial and biologically relevant properties (**Fig. 18**), but direct methods to generate this motif remain limited.

Generation of nsAAs bearing a chiral tertiary alcohol is challenging for two reasons. First, installation of both an amine and an alcohol in a stereoselective fashion in small molecules remains challenging. Synthetic methods to install both functional groups are challenging and tedious, requiring many steps and multiple protections and deprotections to maintain stereochemistry. While there are biocatalytic methods to generate amino acids, these often require enzyme cascades to sequentially incorporate both functional groups reducing their synthetic utility. Second, generation of chiral tertiary alcohols in general remains an unsolved synthetic and biocatalytic challenge. The most widely used and successful method for generating chiral tertiary alcohols is stereoselective aldol addition. Direct methods of performing these aldol reactions suffer from limitations whether traditional synthetic or biocatalytic approaches are pursued.

The three main synthetic approaches for direct, catalytic aldol addition into ketones are Mukaiyama-aldol, decarboxylative aldol, and organocatalyst mediated aldol (**Fig. 19A**). The Mukaiyama-aldol reaction requires pre-formation of the enolate. This method has many iterations, but only a handful demonstrated utility using unactivated ketones as electrophiles. Unfortunately, the synthetic utility of the reactions were hindered by long reaction times, harsh conditions, low stereoselectivity, or small scopes.

Decarboxylative aldol strategies generate the nucleophile *in situ* using metal catalyzed addition into activated diketones, such as isatin or α-ketoacid derivatives (**Fig. 19B**). Again, the reactions require high catalyst loadings and long reaction times usually over 24 h. Lastly, enamine catalysis using organocatalysts can be used as biomimetic methods to generate chiral tertiary alcohols (**Fig. 19C**). However, the organocatalyst methods are limited by limited reaction scope, low ee, long reaction times, and require activated ketones to work. Despite several decades of research, development of aldol reactions for the synthesis of chiral tertiary alcohols are still challenging and exploring alternative methods may provide solutions.

Enzymes are an attractive alternative to the synthetic methods as they work in mild conditions and have exquisite regio- and stereoselectivity. Similar to synthetic methods, biocatalytic strategies for generating chiral tertiary alcohols include aldol addition of a carbon nucleophile into a ketone electrophile. Direct addition of carbon nucleophiles can be mediated using aldolases and TPP dependent enzymes. The classical aldolases historically have struggled to produce synthetically relevant quantities of desired tertiary alcohol product because the reactions are highly reversible. Only a few examples exist in the literature demonstrating addition into ketones with these enzymes is possible, but not scalable. For example, the 4-hydroxy-4-methyl-2-oxoglutarate (HMG) aldolases catalyze the reversible aldol-cleavage of 4-carboxy-4-hydroxy-2-oxoadipate to pyruvate and oxaloacetate, but no scope has been explored (**Fig. 20A**). TPP dependent enzymes such as YerE and acetylacetoin synthase (AAS) have demonstrated addition of a carbon nucleophile into ketone substrates (**Fig. 20B**). In both cases the scope was limited with moderate product yields and low *ee*. A proteinase from *Aspergillus melleus* (AMP) catalyzes an asymmetric aldol reaction between β,γ-unsaturated α-keto esters and ketones affording tertiary chiral alcohols in moderate to good yield with moderate to poor stereoselectivity (**Fig. 20C**). In this case, both the electrophile and nucleophile could be modified, but only minor changes were tolerated limiting synthetic utility. In addition to the limited scope, all of the aforementioned methods only install the chiral tertiary alcohol functional group. Therefore, generation of the amino group to generate the amino acid would require additional steps to install.

Direct methods to generate nsAAs bearing a chiral tertiary alcohol remain extremely limited. Any method capable of providing this functionality in a streamlined fashion would represent a significant synthetic advance. While some C–H activation strategies exist using *cis*-4-proline hydroxylases (P4H) and isoleucine dioxygenase (IDO), these enzymes are not building the carbon framework and are better suited to late-stage functionalization. A recent report demonstrated a wild-type UstD enzyme from *Aspergillus pseudonomius*, a homolog of the UstD used in our work, was capable of catalyzing aldol addition into ketones, but it was limited to activated di-ketone substrates only **(Fig. 20D)** (Zhang et al. "Enzymatic Synthesis of Noncanonical α-Amino Acids Containing γ-Tertiary Alcohols." *Angew. Chemie Int. Ed.* 2024, 63, e202318550). Given the many challenges associated with making chiral tertiary alcohol nsAAs, our extensive engineering of $UstD^{2.0}$ using SUMS to increase activity on ketone electrophiles may provide access to amino acids not previously accessible. The two best ketone variants from our engineering, 7B05 and 7G11 were investigated to determine their utility for synthesis of chiral tertiary alcohol nsAAs.

Exploration of biocatalytic synthesis of chiral tertiary alcohols

Survey of substrate scope on analytical scale

To test synthetic utility of the enzymes, analytical reactions were conducted to probe the reaction scope under the optimized conditions reported in Example 1 **(Fig. 21)**. Yield for each reaction was quantified by Marfey's derivatization.

In general, both enzymes, 7B05 and 7G11 show broad activity with aromatic and aliphatic substrates. The diastereoselectivity varies substrate to substrate, but higher diastereoselectivity is observed when the substituents connected to the carbonyl are sterically distinct. Presumably, the more pronounced the steric differences in the carbonyl substituents the greater the discrimination between binding poses leading to higher diastereoselectivity for those substrates. This hypothesis is most cleanly supported by the trend in diastereoselectivity for the alkyl trifluoromethyl ketones bearing hexyl- butyl- and methyl-substituents. Both the hexyl and butyl substrates have high diastereoselectivity, but trifluoroacetone has no selectivity with 7G11 and a modest 5:2 d.r. with 7B05.

In general, 7B05 can convert substrates bearing a trifluoromethyl on the carbonyl with more efficiency than 7G11, albeit the difference is modest. Meanwhile, 7G11 can convert unactivated ketones more efficiently. Structurally, the four and five-membered rings outperform the six-membered rings. Neither enzyme tolerates a pyridine ring. Despite extensive engineering, the two unactivated ketones (4-FPhAT) and (4NO2AP) have not reached synthetically useful levels of product formation. Nevertheless, we wanted to

demonstrate synthetic utility of the enzymes. Therefore, we chose to isolate and characterize a subset of nsAAs on preparative scale. Of the listed entries, substrates were selected based on chemical diversity to demonstrate the widest breadth of amino acids that can be synthesized.

Preparative scale biocatalytic synthesis of chiral tertiary alcohols

Efficient isolation of aromatic containing amino acids was possible using reverse-phase flash chromatography (**Fig. 22**). Beginning with **4.1**, a good yield of 55% and high diastereoselectivity (>20:1) was observed (**Fig. 22**). Amino acid, **4.2** which was one of the products observed during screening, was isolated at 46% yield with excellent d.r. (>20:1). Excitingly, **4.3** was isolated at 39% yield and excellent d.r. (>20:1). Product **4.4** was isolated at 31% yield and high d.r. (18:1). The high diastereoselectivity was particularly exciting in this case as there are three stereocenters present in this product. A crystal structure of the isolated product confirmed the absolute configuration of the major isomer to be *S,S,R*, with the anti-configuration of the α-amine to the γ-hydroxy groups conserved (**Figs. 23A-23B, Table 8**). A similar ring motif to **4.4** is found in **4.5**, which was isolated in excellent yield (96%) as a 1:1 mixture of diastereomers. The ketone substrate for **4.5** is a synthetic fragrance molecule known as calone or "watermelon ketone" and is used commercially to convey a sea breeze with floral overtones. Unfortunately, the fragrance properties of calone are abolished during the reaction leaving the amino acid odorless.

Table 8. Crystal data and structure refinement for 4.4.

Empirical formula	$C_{13}H_{17}NO_5 \cdot H_2O$
Formula weight	285.29
Temperature/K	100.00
Crystal system	triclinic
Space group	*P*1
a/Å	5.7856(6)
b/Å	6.2026(6)
c/Å	20.699(2)
α/°	94.081(6)
β/°	90.513(6)
γ/°	115.481(6)
Volume/Å^3	668.19(13)

Z	2
$\rho_{calc}g/cm^3$	1.418
μ/mm^{-1}	0.951
F(000)	304.0
Crystal size/mm^3	0.055 × 0.041 × 0.011
Radiation	Cu Kα (λ = 1.54178)
2Θ range for data collection/°	8.574 to 144.67
Index ranges	$-7 \leq h \leq 7$, $-7 \leq k \leq 7$, $-25 \leq l \leq 25$
Reflections collected	22465
Independent reflections	5087 [R_{int} = 0.0300, R_{sigma} = 0.0235]
Data/restraints/parameters	5087/56/399
Goodness-of-fit on F^2	1.060
Final R indexes [I>=2σ (I)]	R_1 = 0.0260, wR_2 = 0.0651
Final R indexes [all data]	R_1 = 0.0266, wR_2 = 0.0656
Largest diff. peak/hole / e $Å^{-3}$	0.20/-0.21
Flack parameter	-0.02(6)

The aliphatic nsAA products were isolated either by reverse-phase chromatography or protected and purified using normal-phase chromatography. The amino acid, **4.6** was isolated with good yield (59%) as a 1:1 mixture of diastereomers. While both diastereomers were structurally characterized by small molecule crystallography, only the anti-isomer is displayed here for clarity (**Figs. 23C, Table 9**). The amino acid, **4.7**, was isolated as an Fmoc protected product in 62% yield. The sugar-like amino acid **4.8**, derived from reaction with 1,3-dihydroxyacetone, was isolated in 71% yield. For **4.8**, the enzymatic reaction was run at a higher catalyst loading (250 Max TON, 0.4 mol% catalyst) to demonstrate that increased catalyst loading can improve yield.

Table 9. Crystal data and structure refinement for 4.6

Empirical formula	$C_{14}H_{26}N_2O_6S_2$
Formula weight	382.49
Temperature/K	100.00
Crystal system	triclinic
Space group	P1
a/Å	5.4808(7)

b/Å	5.6991(7)
c/Å	13.8474(14)
α/°	80.745(8)
β/°	87.828(6)
γ/°	79.666(9)
Volume/Å^3	419.96(9)
Z	1
ρ_{calc}g/cm^3	1.512
μ/mm^{-1}	3.187
F(000)	204.0
Crystal size/mm^3	0.08 × 0.02 × 0.02
Radiation	CuKα (λ = 1.54178)
2Θ range for data collection/°	6.468 to 160.754
Index ranges	$-6 \leq h \leq 6$, $-7 \leq k \leq 7$, $-17 \leq l \leq 17$
Reflections collected	15945
Independent reflections	3413 [R_{int} = 0.0636, R_{sigma} = 0.0514]
Data/restraints/parameters	3413/20/241
Goodness-of-fit on F^2	1.090
Final R indexes [I>=2σ (I)]	R_1 = 0.0423, wR_2 = 0.1111
Final R indexes [all data]	R_1 = 0.0436, wR_2 = 0.1121
Largest diff. peak/hole / e Å^{-3}	0.32/-0.38
Flack parameter	0.014(16)

Conclusions

Here, we have explored the scope of 7B05 and 7G11 for aldol addition into ketones. We isolated a small set of chemically diverse amino acid products in moderate to good yield using a single enzyme system. While there are examples in the literature of biocatalytic C–C bond formation into ketone electrophiles, many of these enzymes have limited scopes making practical synthesis difficult. Here, the engineered enzymes 7G11 and 7B05 are capable of catalyzing aldol addition into many ketones, providing access to previously inaccessible amino acids. Additionally, the enzymes react with many of these ketones in a stereoselective fashion affording many of the products as single diastereomers. This streamlined single enzyme system represents a significant advance towards developing biocatalytic platforms for C–C bond formation. Future work will encompass mechanistic

investigations into the specific molecular determinants of promiscuity for each enzyme to aid in future development of new synthetic applications of UstD and its engineered variants.

Materials and Methods

Analytical substrate scope

General Procedure

All reactions were done only once on analytical scale (200 µL). The buffer used for all enzymatic reactions was 100 mM KPi, pH 7.0, 100 mM NaCl. PLP and l-aspartate stock solutions were made in water. Electrophile stock solutions were made in DMSO and each reaction only has a single electrophile. All samples were analyzed following Marfey's derivatization by Waters Acquity UPLC-MS using a BEH C18 column (Waters). To correct for small deviations in injection volume, an internal derivatization standard was included (0.098 mM l-arginine). Derivatized amino acid product quantitation was performed by integrating chromatograph peaks at 340 nm and corrected by dividing by the internal standard peak area. To calculate product concentrations, a standard curve was generated by subjecting stock solutions of l-phenylalanine (50 mM – 0.4 mM) in water to the identical procedure used to process and derivatize enzymatic reaction solutions, in triplicate, with internal standard. These curves were used to calculate the concentrations of UstD products in solution, and subsequently total turnover numbers after dilution factor correction.

Enzymatic reactions

An Eppendorf tube was charged with electrophile (10 µmol, 1 equiv., 50 mM final concentration, 5% DMSO final concentration). Then, l-aspartate sodium salt monohydrate (50 µmol, 5 equiv., 250 mM final concentration), pyridoxal-5'-phosphate (trifluoromethyl ketones reactions used 10- molar equivalents all others used 50- molar equivalents to final enzyme concentration,) and buffer were aliquoted into the tube. Reactions were initiated by addition of UstD (0.1 mol% catalyst, 1,000 max turnover number) to bring the total reaction volume to 200 µL. The reactions were placed at 37 °C for 4 h. Reactions were quenched with 200 µL of acetonitrile (1 reaction volume) and diluted with 200 µL of 1:1 ACN:DI H_2O to homogenize reaction solutions. Denatured enzyme was removed by centrifugation at 16160 *xg* for 5 min. Marfey's derivatization of the clarified enzymatic reactions were performed to quantify amino acid yield.

Marfey's derivatization procedure

To a fresh 96-well plate, 10 µL of quenched reaction mixture (2.7 mM total amines in reaction, 1 equiv., ~0.8 µmol total amines in reaction) 140 µL of 10.41 mM $NaHCO_3$ with 0.21 mM of l-arginine as an internal standard (9.7 mM $NaHCO_3$ final concentration, 3.5 equiv. base, 2.9 µmol $NaHCO_3$), and 10 mM l-FDAA (5 mM final concentration, 1.8 equiv., 1.5 µmol) were added in each well. The derivatization reaction was allowed to proceed at 37 °C for 18 h. The reactions were quenched with 300 µL of 60 mM HCl in acetonitrile (1 reaction volume) and analyzed via UPLC-MS no later than 24 h after quenching. Note that the amino acid products are susceptible to lactonization upon addition of the acid required to quench the reactions resulting in two product peaks with masses differing by 18 mass units. The linear product will display mass peaks for both the molecular ion and a dehydrated ion (-18 mass units), while the lactone will only display a mass signal associated with dehydration (-18 mass units). Turnover numbers were calculated based on the total integration of linear and lactonized amino acid product peaks at 340 nm.

Preparative Scale *in vitro* Biocatalytic Reactions

Procedure P1: *Preparative scale production of unprotected γ-hydroxy amino acids*

A 100-mL round bottom flask was charged with a given ketone (0.5 mmol, 1.0 equiv, 50 mM final concentration), which was then dissolved in an appropriate amount of MeOH (5% v/v final concentration). This solution was then diluted with 100 mM potassium phosphate buffer (pH 7.0) containing 100 mM sodium chloride. l-aspartate sodium salt monohydrate (2.5 mmol, 5.0 equiv, 250 mM final concentration) and 10-50 molar equivalents of pyridoxal-5'-phosphate (PLP) relative to final enzyme concentration were then added, followed by addition of 7G11 or 7B05 (0.1% mol cat). The total reaction volume was 10 mL. The reaction flask was placed in the dark at 37 °C for 4 h. Product formation was monitored by UPLC-MS. After reaction completion, the reaction mixture was quenched with an equivalent volume of acetonitrile (ACN) and centrifuged (4,000 rpm, 15 min) to remove aggregated protein. The decanted supernatant was then concentrated to ~2 mL by rotary evaporation and loaded onto a preparative reverse-phase C18 column pre-equilibrated at 1% methanol:water. Purification was performed via gradient elution on an Isolera One Flash Purification system (Biotage). Fractions bearing product (confirmed by UPLC-MS sampling of fraction tubes) were pooled and dried by rotary evaporation. The product was then resuspended in a minimal quantity of water, transferred to a pre-weighed 20 mL vial, frozen, and lyophilized.

Procedure P2: *Preparative scale production of Fmoc-protected XX (oxetanone)*

A 100-mL round bottom flask was charged with oxetanone (0.5 mmol, 1.0 equiv, 50 mM final dimer concentration), which was then dissolved in an appropriate amount of MeOH (5% v/v final concentration). This solution was then diluted with 100 mM potassium phosphate buffer (pH 7.0) containing 100 mM sodium chloride. l-aspartate sodium salt monohydrate (2.5 mmol, 5.0 equiv, 250 mM final concentration) and 50 molar equivalents of pyridoxal-5'-phosphate (PLP) relative to final enzyme concentration were then added, followed by addition of 7G11 (0.1% mol cat). The total reaction volume was 10 mL. The reaction flask was placed in the dark at 37 °C for 4 h. Product formation was monitored by UPLC-MS. After reaction completion, the reaction mixture was quenched with an equivalent volume of acetonitrile (ACN) and centrifuged (4,000 rpm, 15 min) to remove aggregated protein. The supernatant was collected in a 250-mL round bottom and basified to pH ~10 using 6 M NaOH. Then Fmoc-Cl (3.7 mmol, 1.5 equiv of total amines) was dissolved in 10 mL ACN and added to the round bottom. The reaction was allowed to stir at room temperature for 4 h. The ACN was removed from the reaction by rotary evaporation. The resulting aqueous layer was acidified to pH ~ 3 using HCl. An extraction of the resulting solution was performed with EtOAc (3x25 mL). The combined organic layers were washed with saturated $NaHCO_3$ (2x25 mL), then saturated NaCl (2x25 mL), and finally dried over $MgSO_4$. The organic layer was gravity filtered then concentrated to dryness and redissolved in a minimum amount of DCM. The resulting solution was loaded onto a preparative normal-phase silica column. Purification was performed via gradient elution (hexane:EtOAc) on an Isolera One Flash Purification system (Biotage). Fractions bearing product (confirmed by UPLC-MS sampling of fraction tubes) were pooled and dried by rotary evaporation. The product was transferred to a pre-weighed vial, and dried on high vacuum.

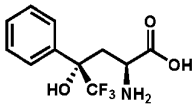
Procedure P3: *Preparative scale production of Fmoc-protected XX (diOhAT)*

A 100-mL round bottom flask was charged with 1,3-dihydroxyacetone dimer (0.5 mmol, 1.0 equiv, 25 mM final dimer concentration), which was then dissolved in an appropriate amount of MeOH (5% v/v final concentration). This solution was then diluted with 100 mM potassium phosphate buffer (pH 7.0) containing 100 mM sodium chloride. l-aspartate sodium salt monohydrate (5 mmol, 10.0 equiv, 250 mM final concentration) and 50 molar equivalents of pyridoxal-5'-phosphate (PLP) relative to final enzyme concentration were then added, followed by addition of 7G11 (0.4% mol cat). The total reaction volume was 20 mL. The reaction flask was placed in the dark at 37 °C for 4 h. Product formation was monitored by UPLC-MS. After reaction completion, the reaction mixture was quenched

with an equivalent volume of acetonitrile (ACN) and centrifuged (4,000 rpm, 15 min) to remove aggregated protein. The supernatant was collected in a 250-mL round bottom and basified to pH ~10 using 300 µL of 6 M NaOH. Then Fmoc-Cl (3.7 mmol, 0.74 equiv of total amines, 93.7 mM final concentration) was added and the reaction was allowed to stir at room temperature for 4 h, final pH = 6. The ACN was removed from the reaction by rotary evaporation. The resulting aqueous layer was acidified to pH ~ 4 using HCl which caused the formation of a gel-like solid. Addition of EtOAc (100 mL) was added to the heterogeneous solution and the mixture was stirred vigorously until the solid dissolved completely. An extraction of the resulting solution was performed with EtOAc (3x50 mL). The combined organic layers were concentrated to dryness and redissolved in a minimum amount of DCM. The resulting solution was loaded onto a preparative normal-phase silica column. Purification was performed via gradient elution (hexane:EtOAc) on an Isolera One Flash Purification system (Biotage). Fractions bearing product (confirmed by UPLC-MS sampling of fraction tubes) were pooled and dried by rotary evaporation. The product was transferred to a pre-weighed vial, and dried on high vacuum.

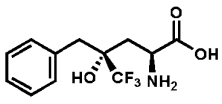
Characterization of γ-hydroxy amino acid products

4.1 **– Synthesis of (2*S*,4*S*)-2-amino-5,5,5-trifluoro-4-hydroxy-4-phenylpentanoic acid**



Prepared from 2,2,2-trifluoroacetophenone using procedure P1. Enzyme: 7G11 Isolated yield: 55% ^{1}H NMR (500 MHz, Deuterium Oxide:MeOH-d^4) δ 7.69 (d, J = 7.6 Hz, 2H), 7.53 – 7.41 (m, 3H), 3.14 (d, J = 11.2 Hz, 1H), 2.70 (d, J = 14.5 Hz, 1H), 2.19 (dd, J = 14.4, 11.3 Hz, 1H). ^{13}C NMR (126 MHz, Deuterium Oxide:MeOH-d^4) δ 181.6, 140.0z, 131.5, 131.4, 129.8, 128.1 (q, J = 284.4 Hz), 80.70 (q, J = 28.1 Hz) 55.4, 38.6. ^{19}F NMR (377 MHz, Deuterium Oxide:MeOH-d^4) δ -80.84. HRMS (ESI): [M-H]$^-$ calcd. for $C_{12}H_{14}F_3NO_3$, 276.0853; found, 276.0855.

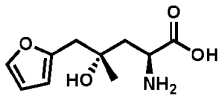
4.2 **– Synthesis of (2*S*,4*R*)-2-amino-4-benzyl-5,5,5-trifluoro-4-hydroxypentanoic acid**



Prepared from 1,1,1-trifluoro-3-phenyl-2-propanone using procedure P1. Enzyme: 7B05 Isolated yield: 46% ^{1}H NMR (500 MHz, Deuterium Oxide) δ 7.45 – 7.35 (m, 5H), 3.56 (dd, J = 7.7, 6.0 Hz, 1H), 3.19 (d, J = 14.1 Hz, 1H), 3.08 (d, J = 14.1 Hz, 1H), 2.17 (dd, J = 15.1, 6.3 Hz, 1H), 2.03 (dd, J = 15.2, 7.7 Hz, 1H).^{13}C NMR (126 MHz, Deuterium Oxide) δ

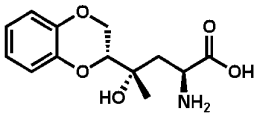
178.9, 135.0, 131.1, 128.4, 127.3, 126.8 (q, J = 286.3 Hz), 75.1 (q, J = 26.6 Hz), 51.7, 40.2, 34.4.^{19}F NMR (377 MHz, D_2O) δ -79.58. HRMS (ESI): $[M-H]^-$ calcd. for $C_{11}H_{12}F_3NO_3$, 262.0697; found, 262.0696.

4.3 **– Synthesis of (2*S*,4*S*)-2-amino-5-(furan-2-yl)-4-hydroxy-4-methylpentanoic acid**



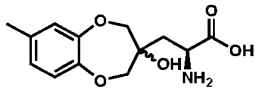
Prepared from 2-furylacetone using procedure P1. Enzyme: 7G11 Isolated yield: 39% ^{1}H NMR (500 MHz, Deuterium Oxide) δ 7.50 (d, J = 1.2 Hz, 1H), 6.47 (dd, J = 3.2, 1.9 Hz, 1H), 6.30 (d, J = 3.1 Hz, 1H), 3.83 (dd, J = 8.3, 4.8 Hz, 1H), 2.96 (dd, J = 16.9, 2.1 Hz, 2H), 2.13 (dd, J = 15.0, 4.8 Hz, 1H), 1.86 (dd, J = 15.0, 8.4 Hz, 1H), 1.29 (s, 3H). ^{13}C NMR (126 MHz, D_2O) δ 177.9, 152.0, 142.1, 110.6, 108.5, 72.7, 52.6, 41.9, 39.1, 26.7. HRMS (ESI): $[M-H]^-$ calcd. for $C_{10}H_{15}NO_4$, 212.0928; found, 212.0925.

4.4 **– Synthesis of (2*S*,4*S*)-2-amino-4-((*R*)-2,3-dihydrobenzo[*b*][1,4]dioxin-2-yl)-4-hydroxypentanoic acid**



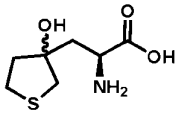
Prepared from 1-(2,3-dihydro-1,4-benzodioxin-2-yl)ethanone using procedure P1. Enzyme: 7G11 Isolated yield: 31% ^{1}H NMR (500 MHz, Deuterium Oxide:MeOH-d^4) δ 7.02 – 6.96 (m, 1H), 6.95 – 6.87 (m, 3H), 4.53 (dd, J = 10.8, 1.6 Hz, 1H), 4.14 – 4.01 (m, 2H), 3.76 (dd, J = 8.2, 5.2 Hz, 1H), 2.33 (dd, J = 14.9, 5.2 Hz, 1H), 1.81 (dd, J = 14.9, 8.2 Hz, 1H), 1.32 (s, 3H). ^{13}C NMR (126 MHz, Deuterium Oxide:MeOH-d^4) δ 179.7, 144.3, 143.8, 123.0, 122.7, 118.3, 117.8, 78.8, 73.2, 65.6, 53.1, 40.7, 23.4. HRMS (ESI): $[M-H]^-$ calcd. for $C_{13}H_{17}NO_5$, 266.1034; found, 266.1034.

4.5 **– Synthesis of (2*S*)-2-amino-3-(3-hydroxy-7-methyl-3,4-dihydro-2*H*-benzo[*b*][1,4]dioxepin-3-yl)propanoic acid**



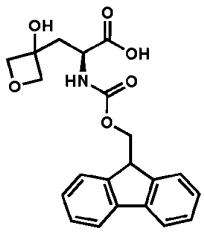
Prepared from calone using procedure P1. Enzyme: 7G11 Isolated yield: 96%, d.r. = 1:1 ^{1}H NMR (500 MHz, Deuterium Oxide:MeOH-d^4) δ 6.90 (dd, J = 8.0, 3.9 Hz, 1H), 6.86 – 6.80 (m, 2H), 4.18 – 4.02 (m, 4H), 3.69 (dd, J = 8.7, 4.8 Hz, 1H), 2.24 (s, 3H), 2.21 (dd, J = 15.0, 4.8 Hz, 1H), 1.87 (dd, J = 15.0, 8.5 Hz, 1H). ^{13}C NMR (126 MHz, Deuterium Oxide:MeOH-d^4) δ 181.0, 152.38, 152.36, 150.44, 150.41, 137.0, 127.2, 123.94, 123.92, 123.36, 123.34, 80.7, 80.5, 79.7, 79.5, 76.7, 54.6, 39.4, 22.4. HRMS (ESI): $[M-H]^-$ calcd. for $C_{13}H_{17}NO_5$, 266.1034; found, 266.1033.

4.6 **– Synthesis of (2*S*)-2-amino-3-(3-hydroxytetrahydrothiophen-3-yl)propanoic acid**



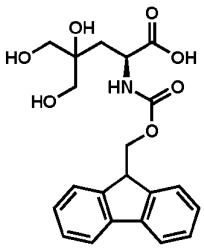
Prepared from 4,5-dihydro-3(*2H*)-thiophenone using procedure P1. Enzyme: 7G11 Isolated yield: 59%, d.r. = 1:1 ^{1}H NMR (500 MHz, Deuterium Oxide) δ 3.89 – 3.80 (m, 1H), 3.07 – 2.86 (m, 4H), 2.39 – 2.09 (m, 3H), 1.95 (dddd, *J* = 12.9, 10.1, 8.1, 2.0 Hz, 1H). ^{13}C NMR (126 MHz, D_2O) δ 176.8, 176.7, 83.2, 82.9, 53.76, 53.71, 42.3, 42.1, 41.0, 40.8, 39.6, 28.25, 28.21. HSQC: note a high-resolution HSQC is included to differentiate ^{13}C signals for each diastereomer. HRMS (ESI): [M-H]$^-$ calcd. for $C_7H_{13}NO_3S$, 192.0689; found, 192.0687.

4.7 **– Synthesis of ((*S*)-2-((((9*H*-fluoren-9-yl)methoxy)carbonyl)amino)-3-(3-hydroxyoxetan-3-yl)propanoic acid**



Prepared from 3-oxetanone using procedure P2. Enzyme: 7G11 Isolated yield: 62% ^{1}H NMR (500 MHz, DMSO-d_6) δ 7.93 – 7.86 (m, 3H), 7.68 (d, *J* = 7.5 Hz, 2H), 7.42 (t, *J* = 7.5 Hz, 2H), 7.34 (t, *J* = 7.4 Hz, 2H), 4.75 (d, *J* = 7.8 Hz, 1H), 4.68 – 4.62 (m, 2H), 4.60 (d, *J* = 7.8 Hz, 1H), 4.45 (dd, *J* = 18.1, 9.4 Hz, 1H), 4.39 – 4.31 (m, 2H), 4.23 (t, *J* = 6.7 Hz, 1H), 2.86 (dd, *J* = 13.0, 9.5 Hz, 1H), 2.32 (dd, *J* = 13.0, 10.0 Hz, 1H). ^{13}C NMR (126 MHz, DMSO) δ 173.9, 155.6, 143.72, 143.67, 140.7, 127.6, 127.1, 125.09, 125.05, 120.1, 82.0, 80.9, 80.7, 65.7, 49.7, 46.6, 35.9. HRMS (ESI): [M-H]$^-$ calcd. for $C_{21}H_{21}NO_6$, 382.1296; found, 382.1299.

4.8 **– (*S*)-2-((((9*H*-fluoren-9-yl)methoxy)carbonyl)amino)-4,5-dihydroxy-4-(hydroxymethyl)pentanoic acid**



Prepared from 1,3-dihydroxyacetone using procedure P3. Enzyme: 7G11 Isolated yield: 71% ^{1}H NMR (500 MHz, Methanol-d_4) δ 7.79 (d, *J* = 7.5 Hz, 2H), 7.65 (d, *J* = 7.5 Hz,

2H), 7.39 (t, J = 7.5 Hz, 2H), 7.31 (t, J = 7.1 Hz, 2H), 4.66 (t, J = 10.2 Hz, 1H), 4.41 – 4.32 (m, 2H), 4.22 (t, J = 7.0 Hz, 1H), 3.75 (d, J = 12.0 Hz, 1H), 3.68 – 3.58 (m, 3H), 2.52 (dd, J = 12.7, 10.0 Hz, 1H), 2.13 (dd, J = 12.7, 10.5 Hz, 1H). ^{13}C NMR (126 MHz, MeOD) δ 177.4, 158.3, 145.22, 145.20, 142.6, 128.8, 128.2, 126.22, 126.21, 120.9, 87.9, 68.1, 65.7, 65.5, 52.6, 48.3, 33.4. HRMS (ESI): $[M-H_2O+NH_4]^+$ calcd. for $C_{21}H_{23}NO_7$, 401.1707; found, 401.1700. $[M-H_2O+CH_3CO_2]^-$ calcd. for $C_{21}H_{23}NO_7$, 442.1507; found, 442.1508. Note: ammonium acetate was added to samples during preparation for collecting HRMS data, the adducts detected reflect this addition.

CLAIMS

What is claimed is:

1. An unnatural, mutant protein comprising an amino acid sequence at least 80% identical to a UstD sequence selected from any one of SEQ ID NOs:1-6, wherein the amino acid sequence comprises one or more of:

a residue other than K at a position corresponding to position 2 of the UstD sequence;

a residue other than V at a position corresponding to position 63 of the UstD sequence;

a residue other than F at a position corresponding to position 75 of the UstD sequence;

a residue other than P at a position corresponding to position 80 of the UstD sequence;

a residue other than P at a position corresponding to position 82 of the UstD sequence;

a residue other than P at a position corresponding to position 83 of the UstD sequence;

a residue other than D at a position corresponding to position 86 of the UstD sequence;

a residue other than Y at a position corresponding to position 96 of the UstD sequence;

a residue other than G at a position corresponding to position 101 of the UstD sequence;

a residue other than I at a position corresponding to position 141 of the UstD sequence;

a residue other than H at a position corresponding to position 263 of the UstD sequence;

a residue other than Y at a position corresponding to position 277 of the UstD sequence;

a residue other than M at a position corresponding to position 299 of the UstD sequence;

a residue other than V at a position corresponding to position 330 of the UstD sequence;

a residue other than K at a position corresponding to position 342 of the UstD sequence;

a residue other than G at a position corresponding to position 373 of the UstD sequence;

a residue other than T at a position corresponding to position 388 of the UstD sequence;

a residue other than I and T at a position corresponding to position 391 of the UstD sequence;

a residue other than L at a position corresponding to position 392 of the UstD sequence;

a residue other than L and M at a position corresponding to position 393 of the UstD sequence;

a residue other than W at a position corresponding to position 399 of the UstD sequence;

a residue other than S at a position corresponding to position 407 of the UstD sequence;

a residue other than Y at a position corresponding to position 418 of the UstD sequence; and

a residue other than L at a position corresponding to position 440 of the UstD sequence.

2. The protein of claim 1, wherein the amino acid sequence comprises one or more, two or more, three or more, four or more, five or more, six or more, seven or more, eight or more, or each of:

a residue other than F at a position corresponding to position 75 of the UstD sequence;

a residue other than P at a position corresponding to position 82 of the UstD sequence;

a residue other than D at a position corresponding to position 86 of the UstD sequence;

a residue other than M at a position corresponding to position 299 of the UstD sequence;

a residue other than V at a position corresponding to position 330 of the UstD sequence;

a residue other than G at a position corresponding to position 373 of the UstD sequence;

a residue other than I and T at a position corresponding to position 391 of the UstD sequence;

a residue other than L and M at a position corresponding to position 393 of the UstD sequence; and

a residue other than S at a position corresponding to position 407 of the UstD sequence.

3. The protein of any prior claim, wherein the amino acid sequence comprises one or more, two or more, or each of:

a residue other than P at a position corresponding to position 82 of the UstD sequence;

a residue other than V at a position corresponding to position 330 of the UstD sequence; and

a residue other than G at a position corresponding to position 373 of the UstD sequence.

4. The protein of any prior claim, wherein the amino acid sequence comprises a residue other than P at a position corresponding to position 82 of the UstD sequence.

5. The protein of any prior claim, wherein the amino acid sequence comprises a residue other than V at a position corresponding to position 330 of the UstD sequence.

6. The protein of any prior claim, wherein the amino acid sequence comprises a residue other than G at a position corresponding to position 373 of the UstD sequence.

7. The protein of any prior claim, wherein the amino acid sequence comprises one, more than one, or each of:

a residue other than P at a position corresponding to position 82 of the UstD sequence; and

a residue other than G at a position corresponding to position 373 of the UstD sequence.

8. The protein of any prior claim, wherein the amino acid sequence comprises each of:

a residue other than P at a position corresponding to position 82 of the UstD sequence; and

a residue other than G at a position corresponding to position 373 of the UstD sequence.

9. The protein of any prior claim, wherein the amino acid sequence comprises one or more, two or more, three or more, or each of:

a residue other than F at a position corresponding to position 75 of the UstD sequence;

a residue other than D at a position corresponding to position 86 of the UstD sequence;

a residue other than V at a position corresponding to position 330 of the UstD sequence; and

a residue other than S at a position corresponding to position 407 of the UstD sequence.

10. The protein of any prior claim, wherein the amino acid sequence comprises each of:

a residue other than F at a position corresponding to position 75 of the UstD sequence;

a residue other than D at a position corresponding to position 86 of the UstD sequence;

a residue other than V at a position corresponding to position 330 of the UstD sequence; and

a residue other than S at a position corresponding to position 407 of the UstD sequence.

11. The protein of any prior claim, wherein the amino acid sequence comprises one or more, two or more, three or more, four or more, or each of:

a residue other than M at a position corresponding to position 299 of the UstD sequence;

a residue other than T at a position corresponding to position 388 of the UstD sequence;

a residue other than I and T at a position corresponding to position 391 of the UstD sequence;

a residue other than L at a position corresponding to position 392 of the UstD sequence;

a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

12. The protein of any prior claim, wherein the amino acid sequence comprises one or more, two or more, or each of:

a residue other than M at a position corresponding to position 299 of the UstD sequence;

a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and

a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

13. The protein of any prior claim, wherein the amino acid sequence comprises each of:

a residue other than M at a position corresponding to position 299 of the UstD sequence;

a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and

a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

14. The protein of any prior claim, wherein the amino acid sequence comprises one or both of:

a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and

a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

15. The protein of any prior claim, wherein the amino acid sequence comprises each of:

a residue other than I and T at a position corresponding to position 391 of the UstD sequence; and

a residue other than L and M at a position corresponding to position 393 of the UstD sequence.

16. The protein of any prior claim, wherein:

the residue other than K at the position corresponding to position 2 of the UstD sequence, if present, is E or a conservative variant thereof;

the residue other than V at the position corresponding to position 63 of the UstD sequence, if present, is I or a conservative variant thereof;

the residue other than F at the position corresponding to position 75 of the UstD sequence, if present, is A, C, H, I, K, L, M, N, Q, R, S, T, V, W, Y, or a conservative variant of any of the foregoing;

the residue other than P at the position corresponding to position 80 of the UstD sequence, if present, is G, L, R, or a conservative variant of any of the foregoing;

the residue other than P at the position corresponding to position 82 of the UstD sequence, if present, is G, Q, S, or a conservative variant of any of the foregoing;

the residue other than P at the position corresponding to position 83 of the UstD sequence, if present, is G, T, R, V, Y, or a conservative variant of any of the foregoing;

the residue other than D at the position corresponding to position 86 of the UstD sequence, if present, is I, N, V, or a conservative variant of any of the foregoing;

the residue other than Y at the position corresponding to position 96 of the UstD sequence, if present, is C or a conservative variant thereof;

the residue other than G at the position corresponding to position 101 of the UstD sequence, if present, is A, F, Q, R or a conservative variant of any of the foregoing;

the residue other than I at the position corresponding to position 141 of the UstD sequence, if present, is M, V, or a conservative variant of any of the foregoing;

the residue other than H at the position corresponding to position 263 of the UstD sequence, if present, is R or a conservative variant thereof;

the residue other than Y at the position corresponding to position 277 of the UstD sequence, if present, is C, F, H, or a conservative variant of any of the foregoing;

the residue other than M at the position corresponding to position 299 of the UstD sequence, if present, is L, V, or a conservative variant of any of the foregoing;

the residue other than V at the position corresponding to position 330 of the UstD sequence, if present, is A, C, L, Q, R, or a conservative variant of any of the foregoing;

the residue other than K at the position corresponding to position 342 of the UstD sequence, if present, is E or a conservative variant thereof;

the residue other than G at the position corresponding to position 373 of the UstD sequence, if present, is E, R, or a conservative variant of any of the foregoing;

the residue other than T at the position corresponding to position 388 of the UstD sequence, if present, is A, I, V, or a conservative variant of any of the foregoing;

the residue other than I and T at the position corresponding to position 391 of the UstD sequence, if present, is F, S, or a conservative variant of any of the foregoing;

the residue other than L at the position corresponding to position 392 of the UstD sequence, if present, is A or a conservative variant of thereof;

the residue other than L and M at the position corresponding to position 393 of the UstD sequence, if present, is C, F, S, W, or a conservative variant of any of the foregoing;

the residue other than W at the position corresponding to position 399 of the UstD sequence, if present, is C or a conservative variant of thereof;

the residue other than S at the position corresponding to position 407 of the UstD sequence, if present, is A, E, N Q, T, or a conservative variant of any of the foregoing;

the residue other than Y at the position corresponding to position 418 of the UstD sequence, if present, is H or a conservative variant of thereof; and

the residue other than L at the position corresponding to position 440 of the UstD sequence, if present is P or a conservative variant of thereof.

17. The protein of any prior claims, wherein:

the residue other than F at the position corresponding to position 75 of the UstD sequence, if present, is A or a conservative variant thereof;

the residue other than P at the position corresponding to position 82 of the UstD sequence, if present, is Q, S, or a conservative variant of any of the foregoing;

the residue other than D at the position corresponding to position 86 of the UstD sequence, if present, is I or a conservative variant thereof;

the residue other than M at the position corresponding to position 299 of the UstD sequence, if present, is V or a conservative variant thereof;

the residue other than V at the position corresponding to position 330 of the UstD sequence, if present, is A, R, or a conservative variant of any of the foregoing;

the residue other than G at the position corresponding to position 373 of the UstD sequence, if present, is E or a conservative variant thereof;

the residue other than I and T at the position corresponding to position 391 of the UstD sequence, if present, is S or a conservative variant thereof;

the residue other than L and M at the position corresponding to position 393 of the UstD sequence, if present, is F, W, or a conservative variant of any of the foregoing; and

the residue other than S at the position corresponding to position 407 of the UstD sequence, if present, is Q or a conservative variant thereof.

18. The protein of any prior claim, wherein the protein has activity in generating a gamma-hydroxy amino acid from an amino acid and one or more of an aldehyde-containing substrate and a ketone-containing substrate.

19. A method of making a gamma-hydroxy amino acid, the method comprising contacting the protein of any one of claims 1-18 with an amino acid and a substrate comprising one or more of an aldehyde-containing substrate and a ketone-containing substrate to yield the gamma-hydroxy amino acid.

20. The method of claim 19, wherein the substrate comprises a ketone-containing substrate.

21. The method of claim 19, wherein the ketone-containing substrate lacks vicinyl ketone groups.

22. A method of making a gamma-hydroxy amino acid, the method comprising contacting an unnatural, mutated protein with an amino acid and a ketone-containing substrate to yield the gamma-hydroxy amino acid, wherein the unnatural, mutated protein comprises an amino acid sequence with at least 50% sequence identity but less than 100% sequence identity to a wild-type UstD protein as shown in SEQ ID NO:1.

23. The method of claim 22, wherein the ketone-containing substrate lacks vicinyl ketone groups.

24. The method of any one of claims 22-23, wherein the unnatural, mutated protein is the protein of any one of claims 1-18.

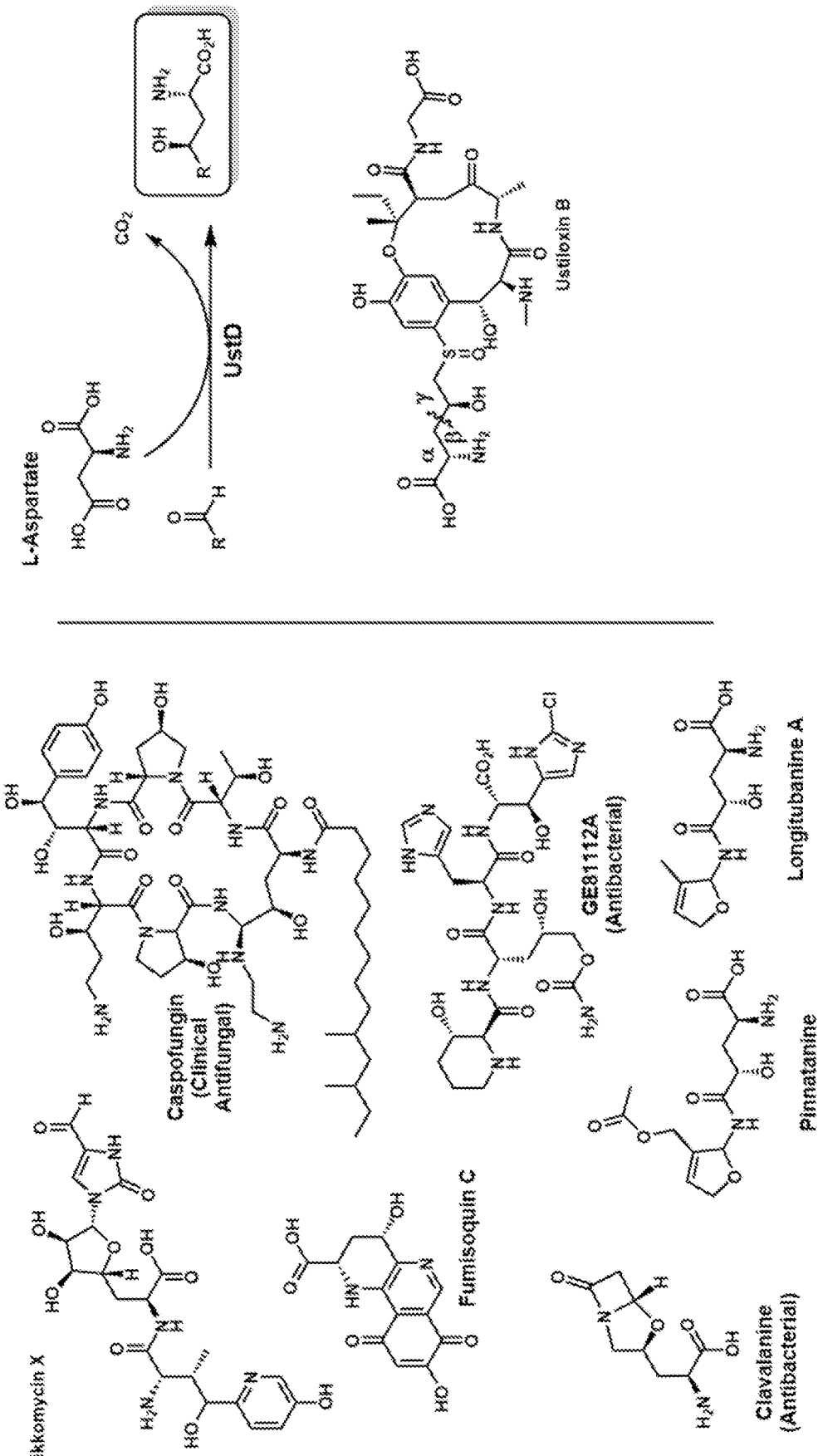
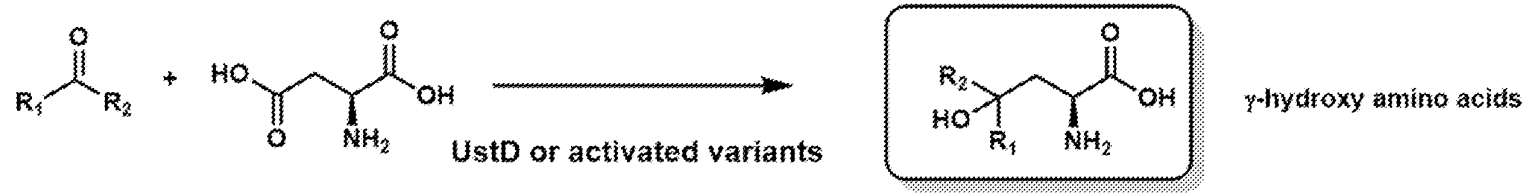
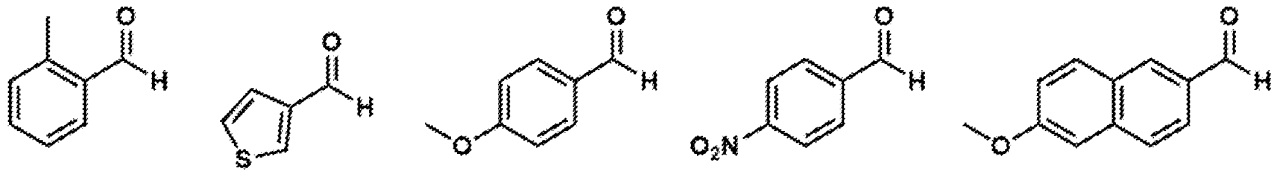

L-Aspartate
UstD
CO2
Ustiloxin B
Nikkomycin X
Caspofungin (Clinical Antifungal)
Fumisoquin C
GE81112A (Antibacterial)
Longitubanine A
Pinnatanine
Clavalanine (Antibacterial)

Fig. 1

A) General (non-native) reaction catalyzed by UstD



B) Representative aldehyde substrates with which engineered UstD variants been experimentally shown to react



C) Representative non-aldehyde substrates with which engineered UstD reacts

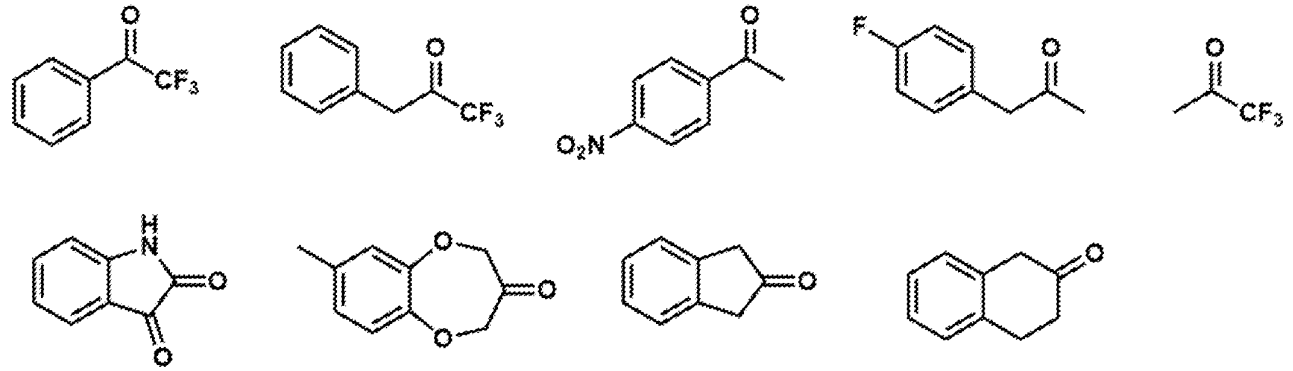


Fig. 2

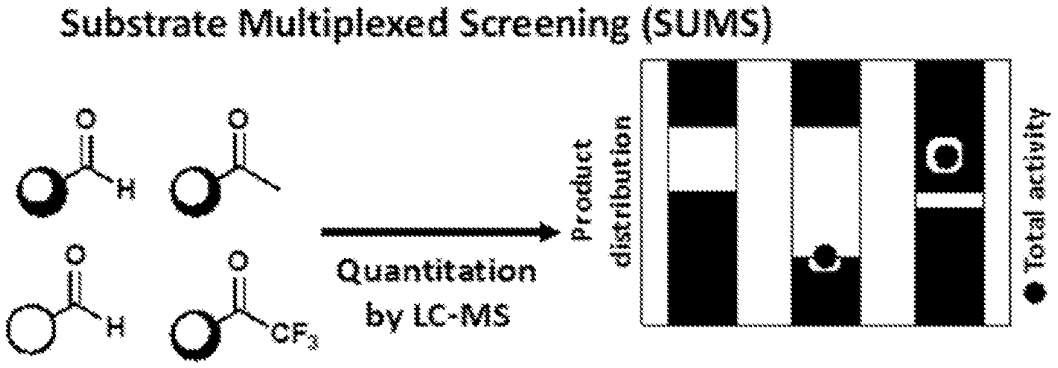

Substrate Multiplexed Screening (SUMS)
Quantitation by LC-MS
Product distribution
Total activity

Fig. 3A

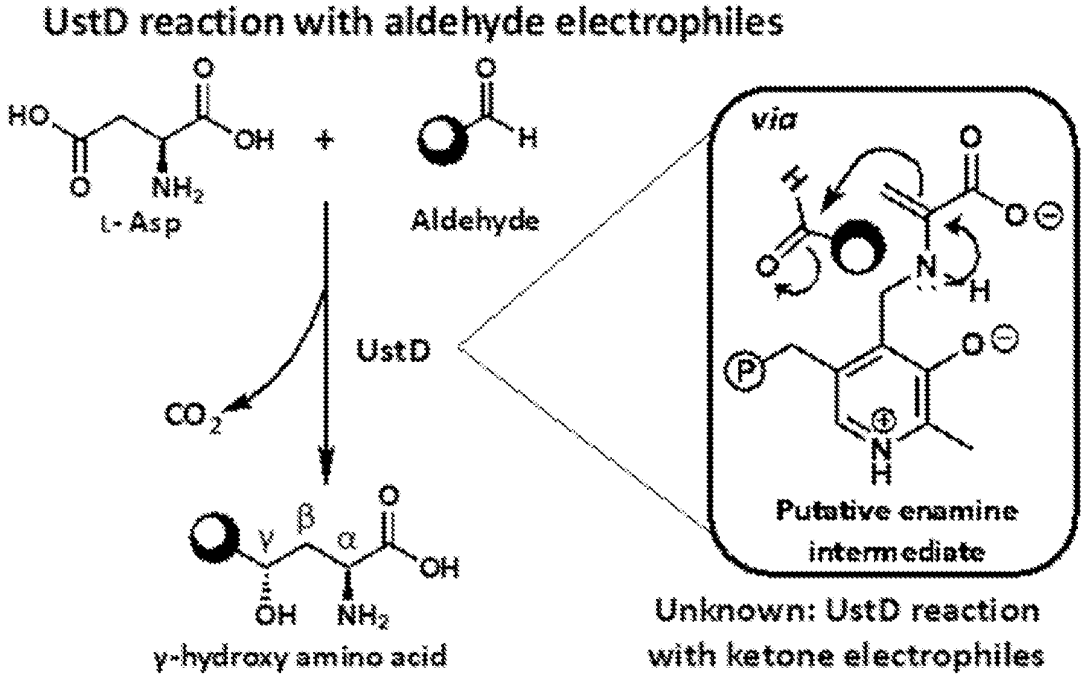

UstD reaction with aldehyde electrophiles
L-Asp
Aldehyde
UstD
via
Putative enamine intermediate
γ-hydroxy amino acid
Unknown: UstD reaction with ketone electrophiles

Fig. 3B

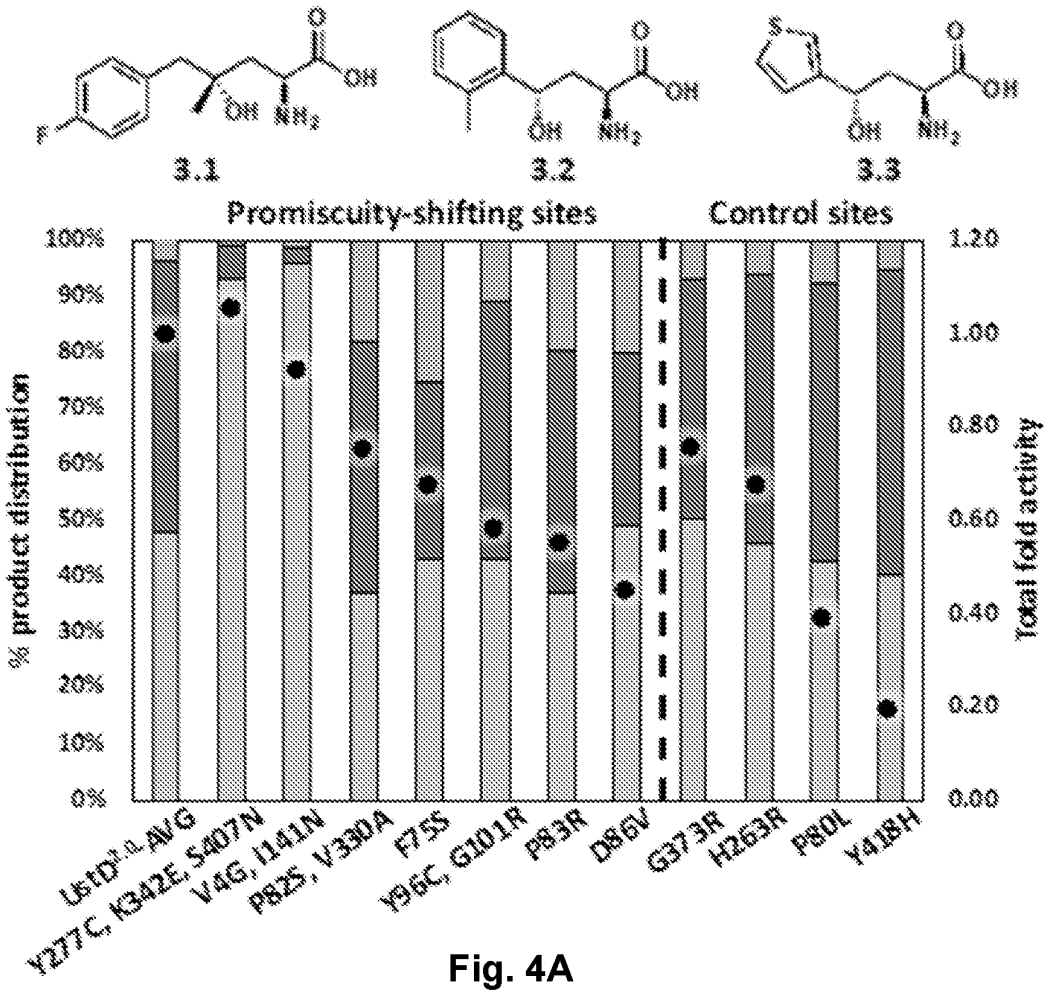

3.1
3.2
3.3
Promiscuity-shifting sites
Control sites
% product distribution
Total fold activity
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
1.20
1.00
0.80
0.60
0.40
0.20
0.00
UstD^n.d.-AVG
Y277C, K342E, S407N
V4G, I141N
P82S, V330A
F75S
Y96C, G101R
P83R
D86V
G373R
H263R
P80L
Y418H

Fig. 4A

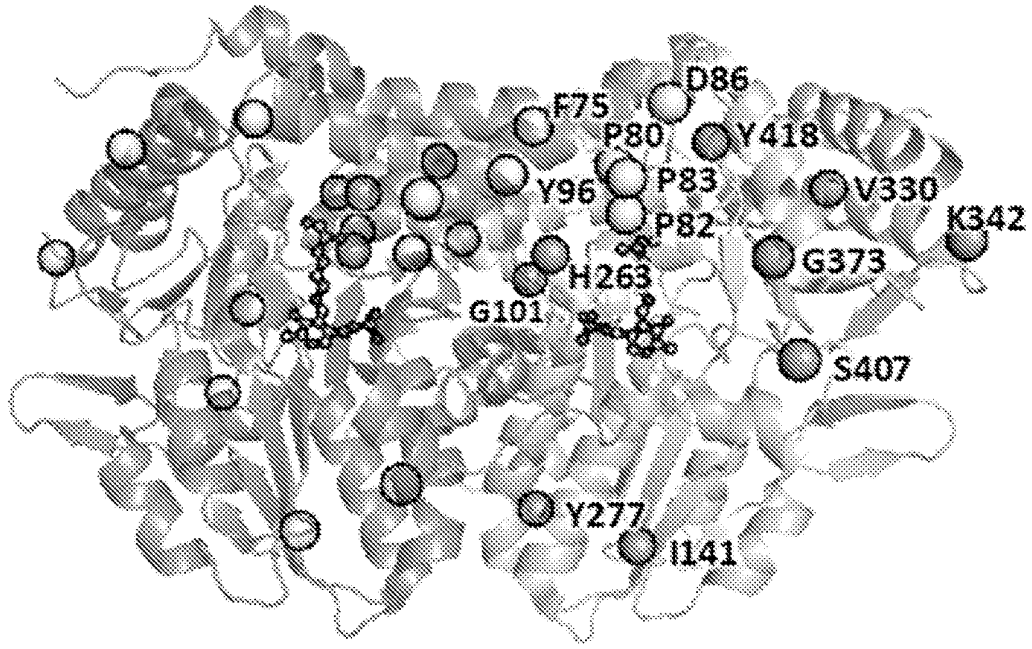

D86
F75
P80
Y418
Y96
P83
V330
K342
P82
G373
H263
G101
S407
Y277
I141

Fig. 4B

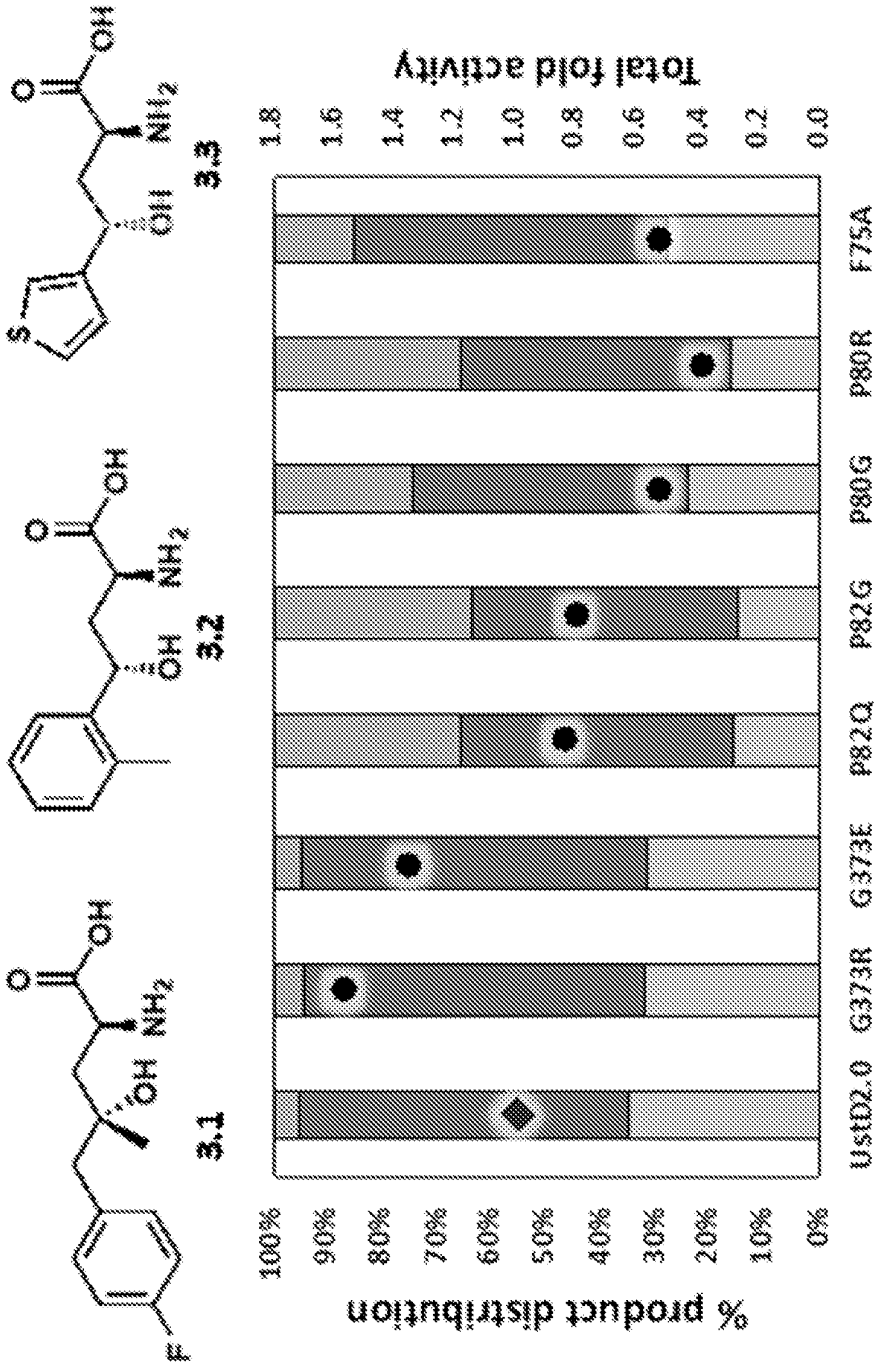

3.1
3.2
3.3
% product distribution
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Total fold activity
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
UstD2.0
G373R
G373E
P82Q
P82G
P80G
P80R
F75A

Fig. 5

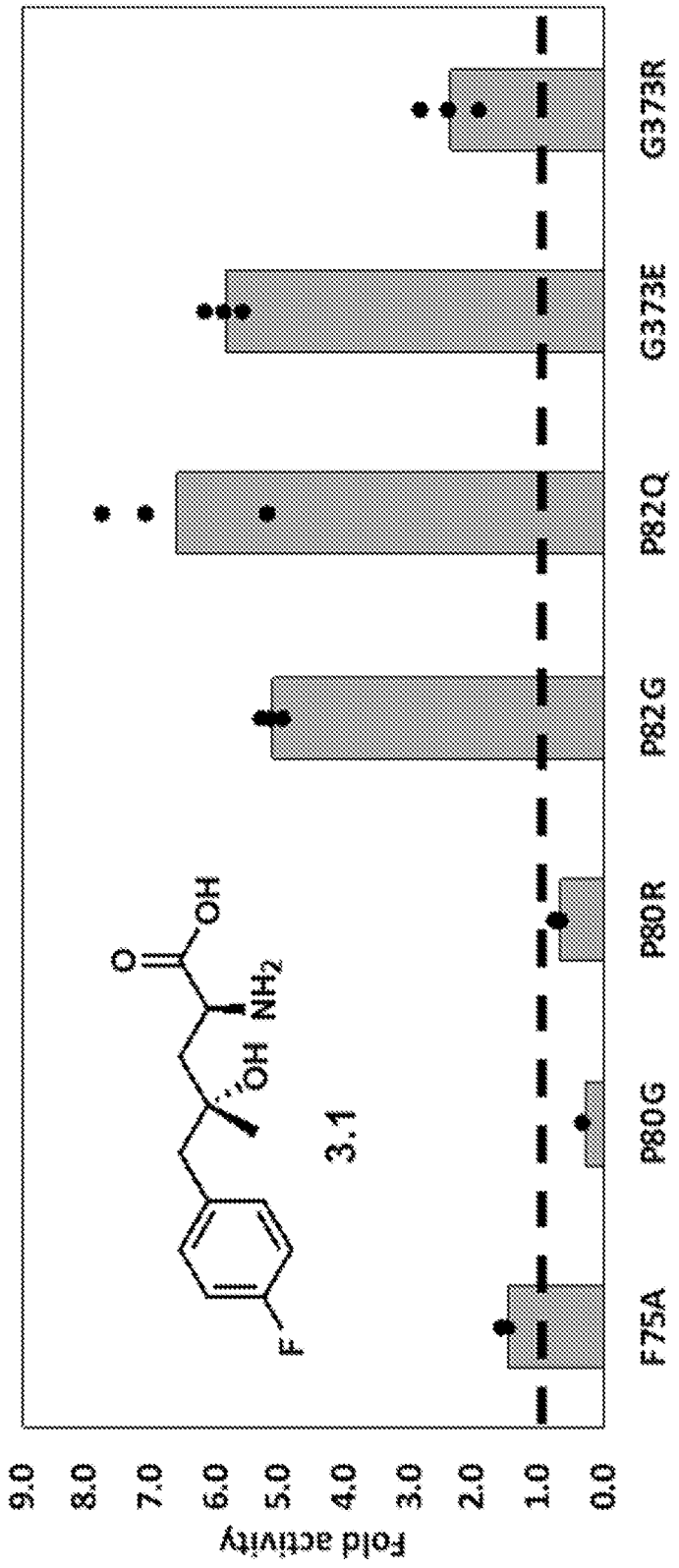

Fold activity
9.0
8.0
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
F75A
P80G
P80R
P82G
P82Q
G373E
G373R
3.1
OH
O
NH2
OH
F

Fig. 6

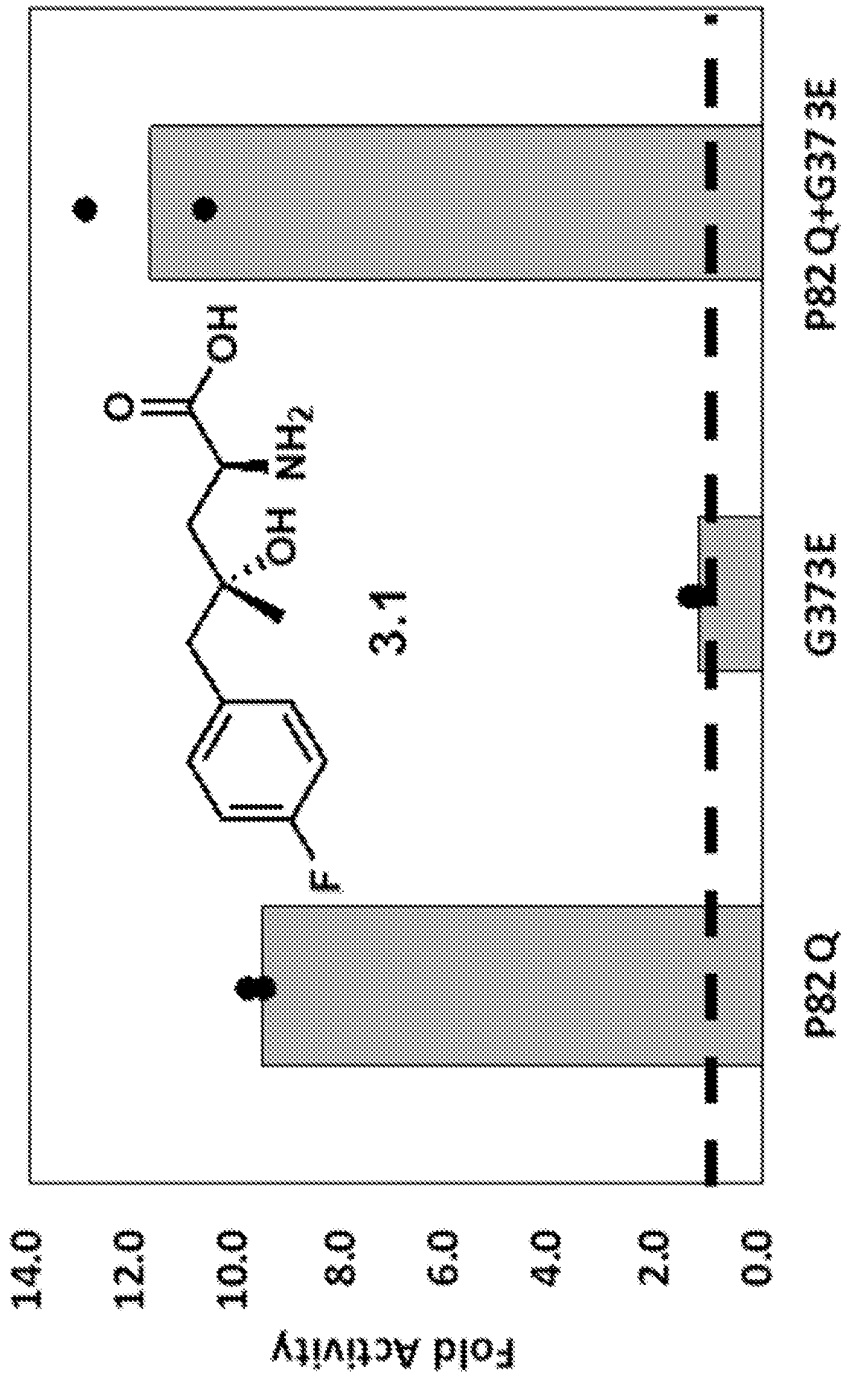

Fold Activity
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
P82Q
G373E
P82Q+G373E
3.1
OH
O
NH2
OH
F

Fig. 7

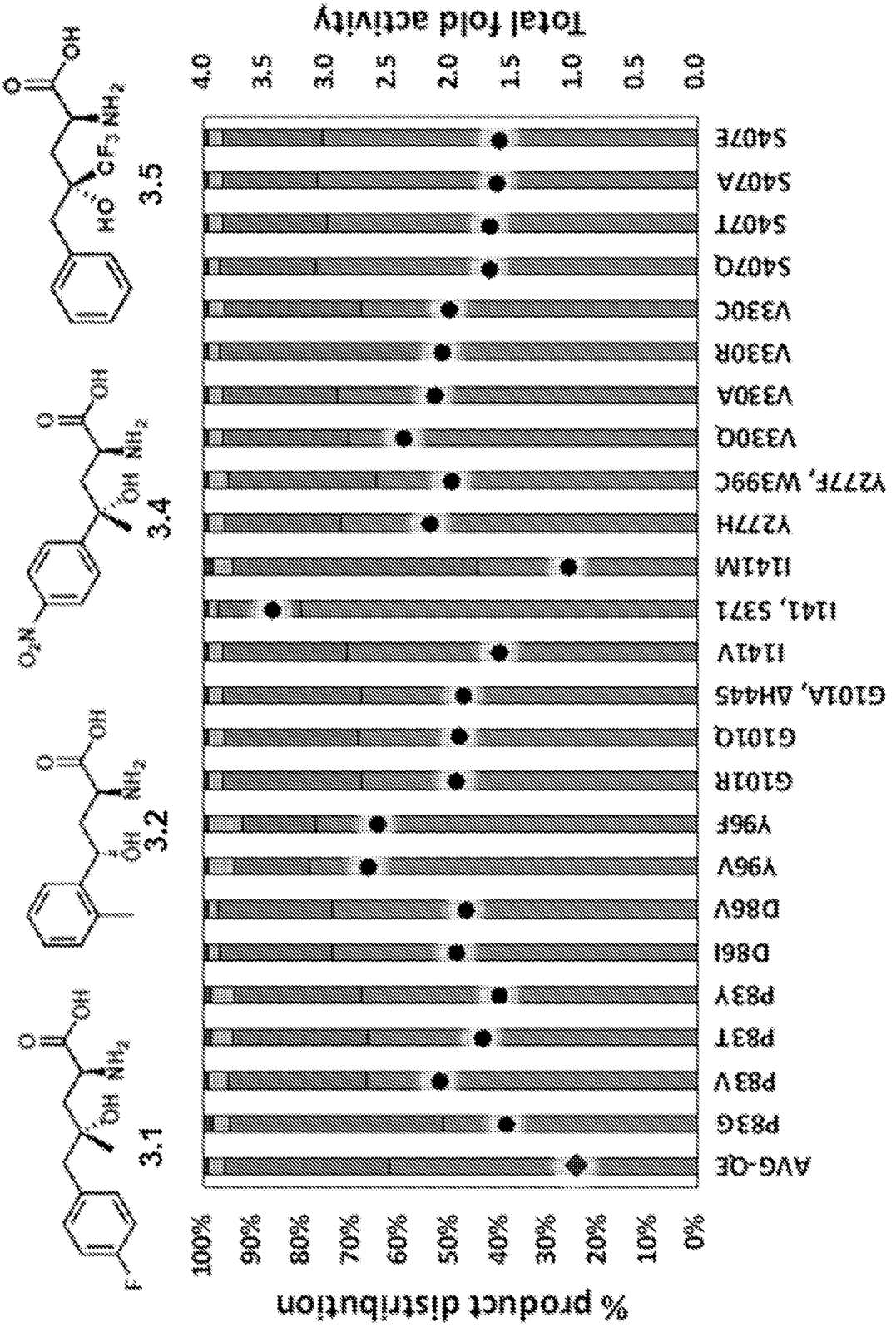

Total fold activity
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
% product distribution
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
AVG-QE
P83G
P83V
P83T
P83Y
D86I
D86V
Y96V
Y96F
G101R
G101Q
G101A, ΔH445
I141V
I141, S371
I141M
Y277H
Y277F, W399C
V330Q
V330A
V330R
V330C
S407Q
S407T
S407A
S407E
3.1
3.2
3.4
3.5

Fig. 8

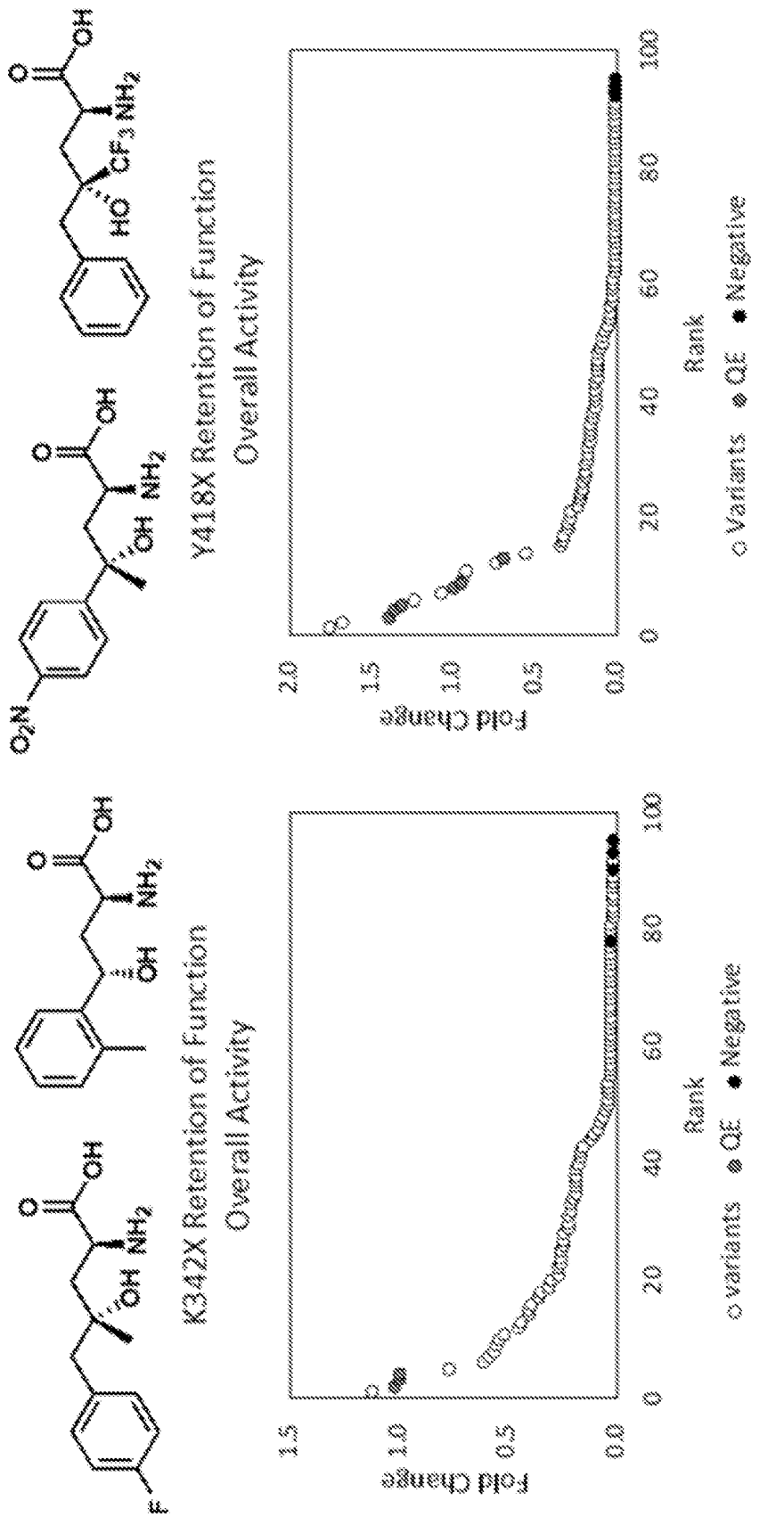

K342X Retention of Function
Overall Activity
Fold Change
Rank
variants
QE
Negative
Y418X Retention of Function
Overall Activity
Fold Change
Rank
Variants
QE
Negative

Fig. 9

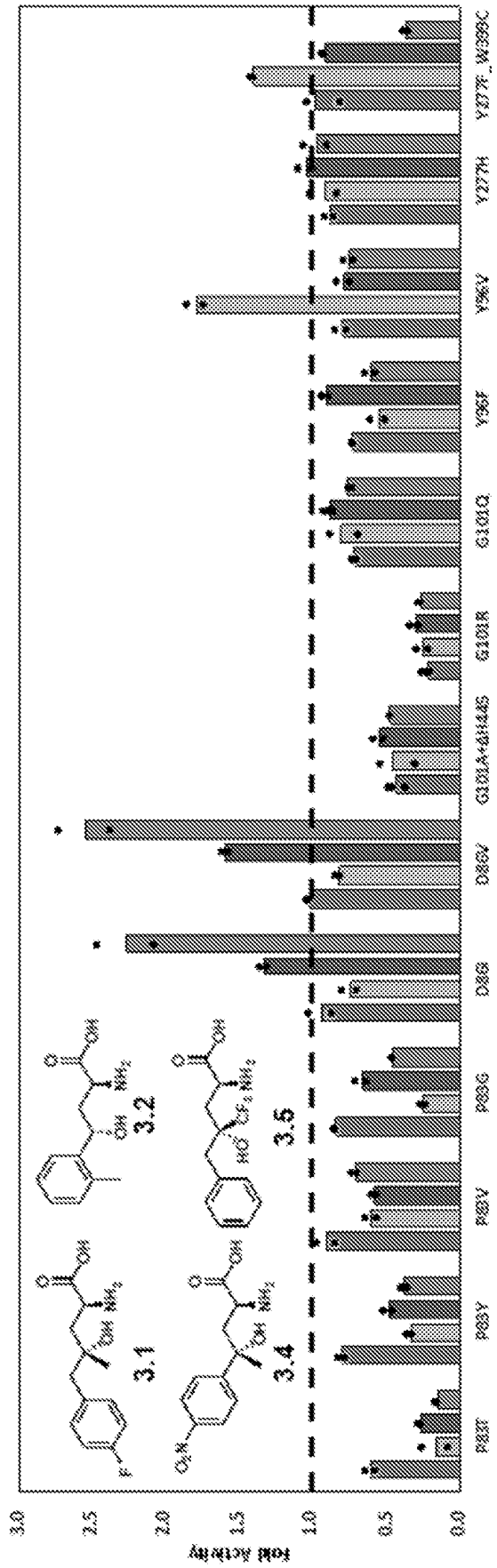

fold activity
3.0
2.5
2.0
1.5
1.0
0.5
0.0
3.1
3.2
3.4
3.5
P83T
P83Y
P83V
P83G
D86S
D86V
G101A+H445
G101R
G101Q
Y96F
Y96V
Y277H
Y277K_W339C

Fig. 10

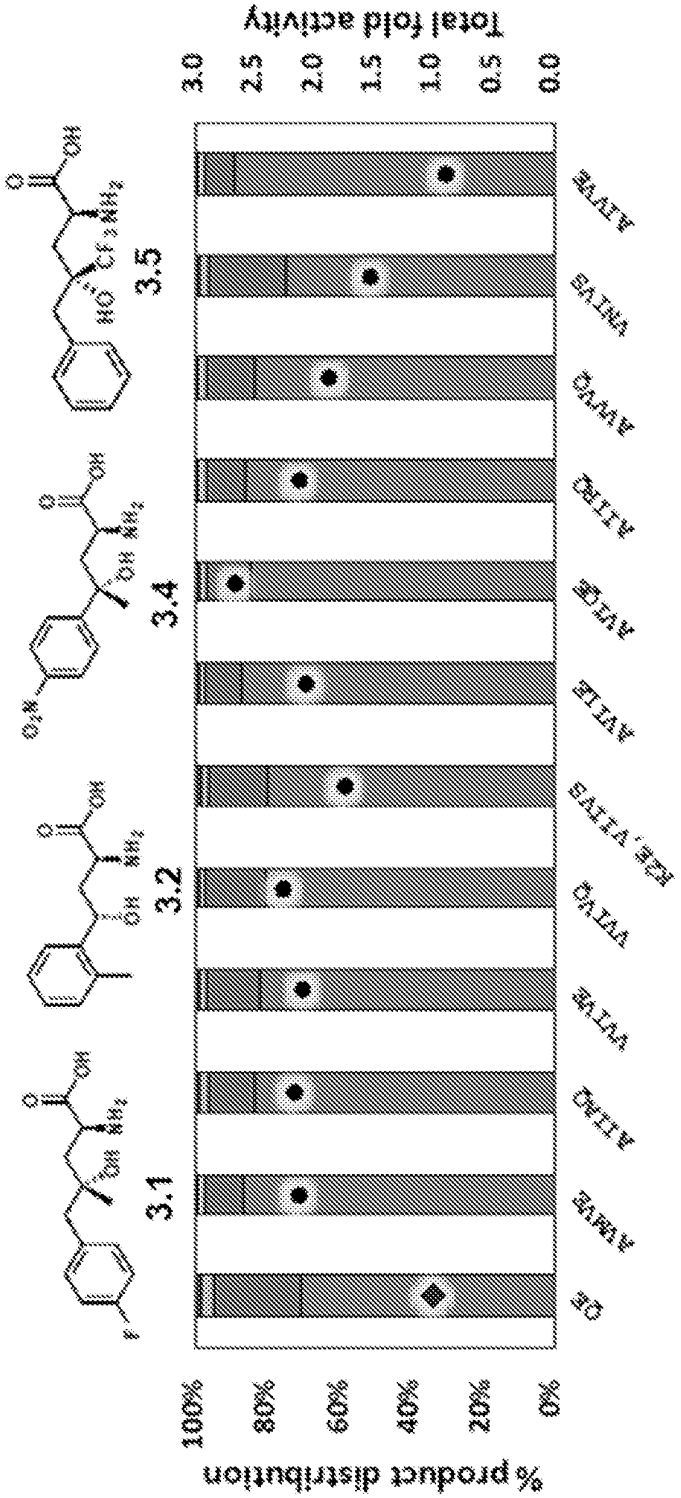


Fig. 11A

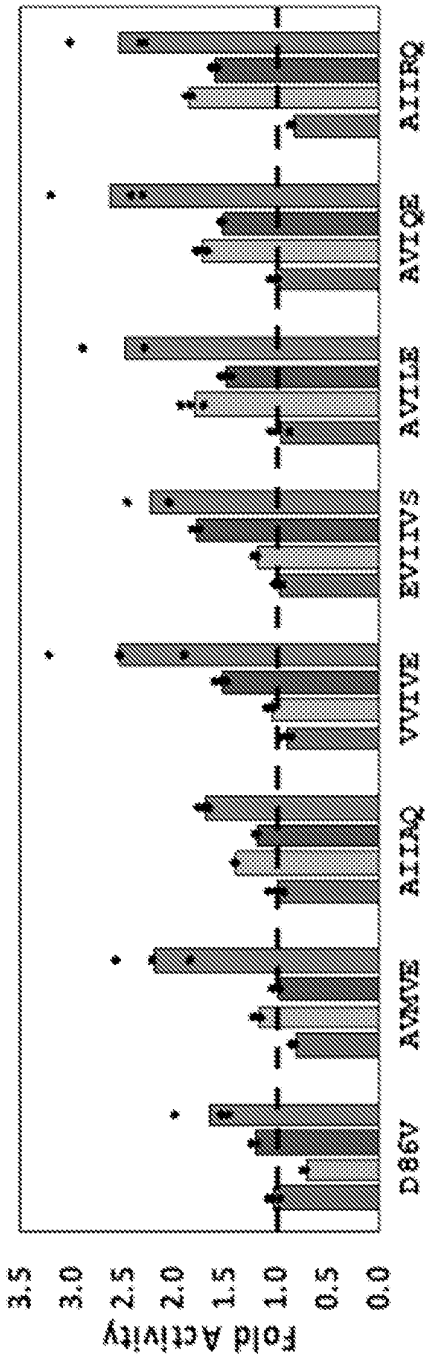


Fig. 11B

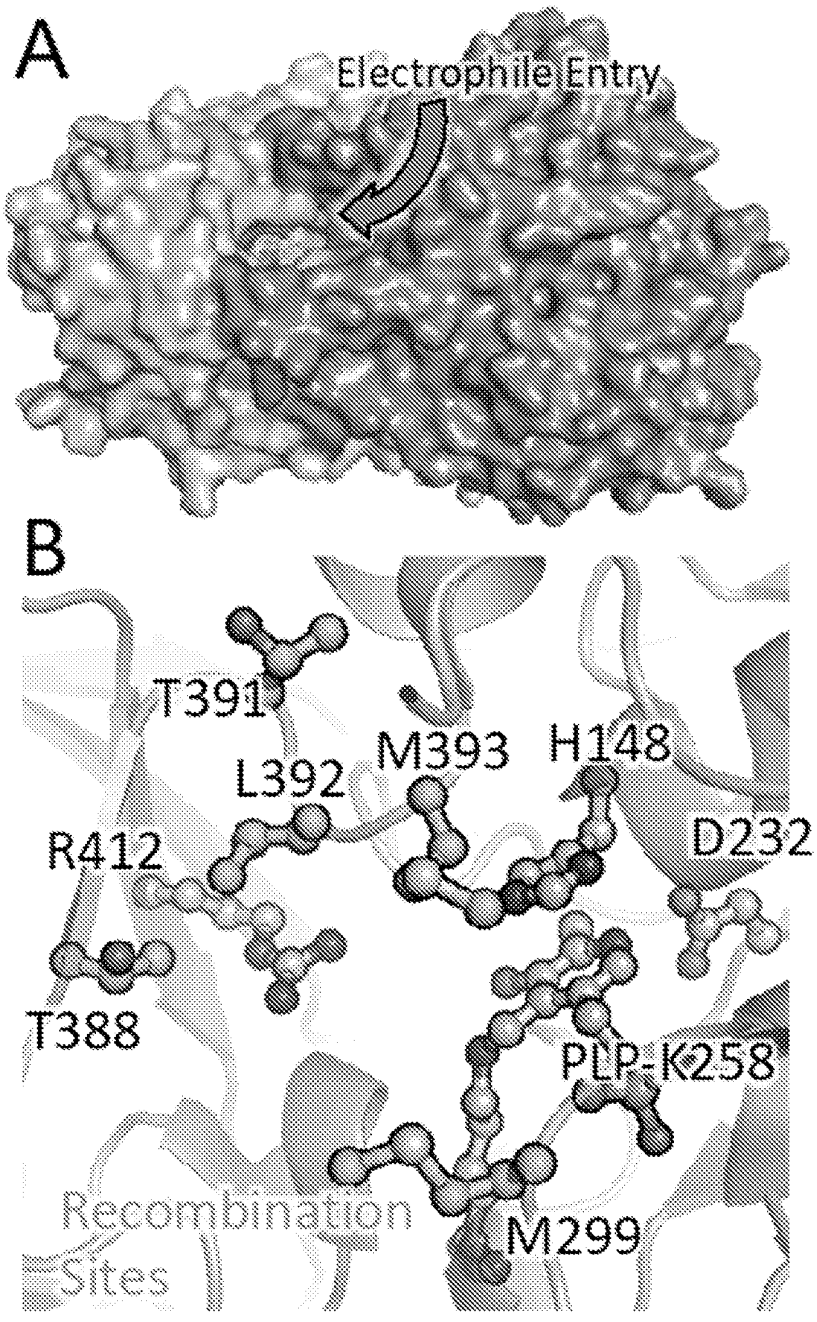

A
Electrophile Entry
B
T391
M393
H148
L392
R412
D232
T388
PLP-K258
Recombination Sites
M299

Fig. 12

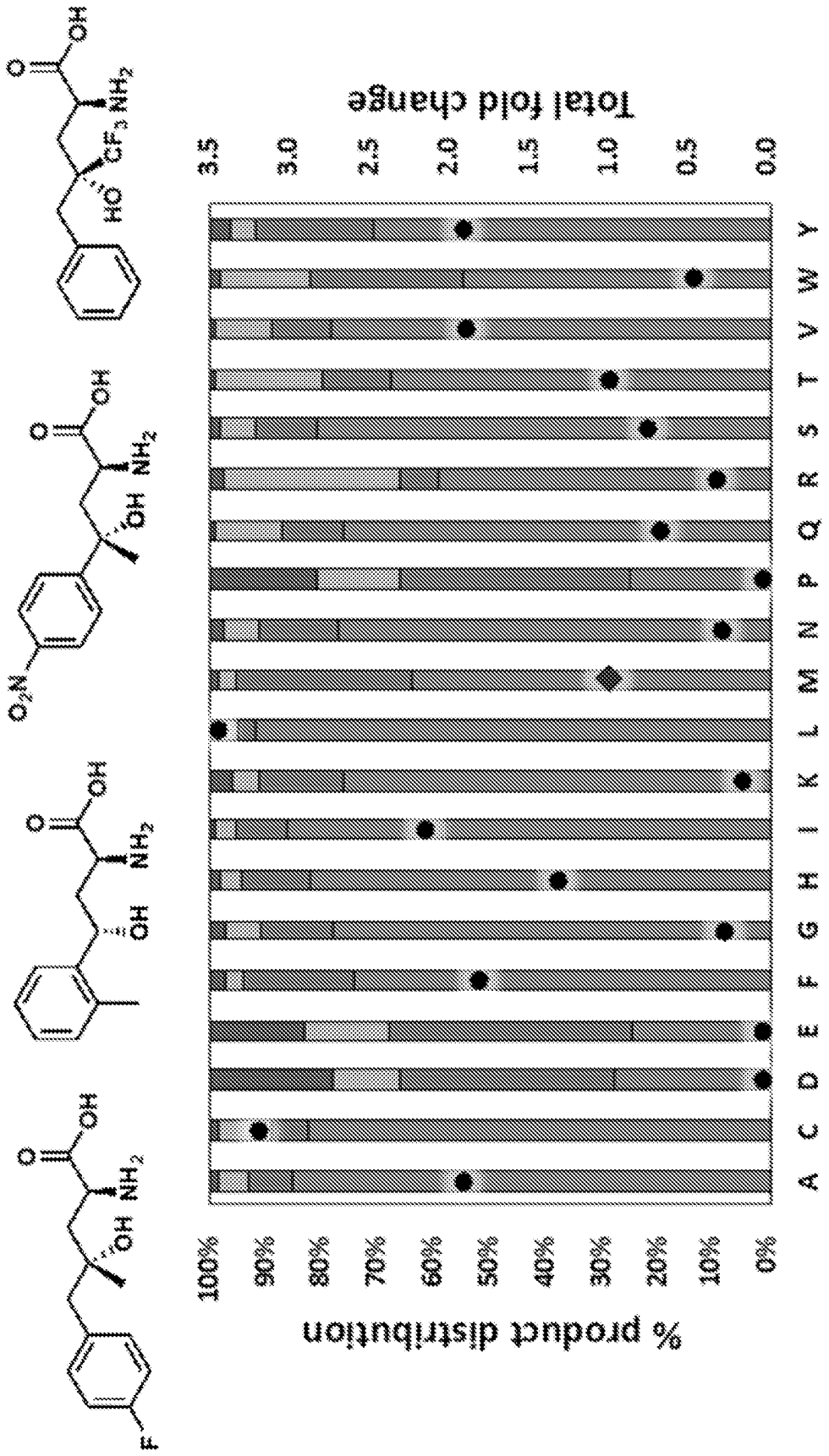

% product distribution
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Total fold change
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
A C D E F G H I K L M N P Q R S T V W Y
OH
O
NH2
HO
CF3
OH
O
NH2
OH
O2N
OH
O
NH2
OH
OH
O
NH2
OH
F

Fig. 13

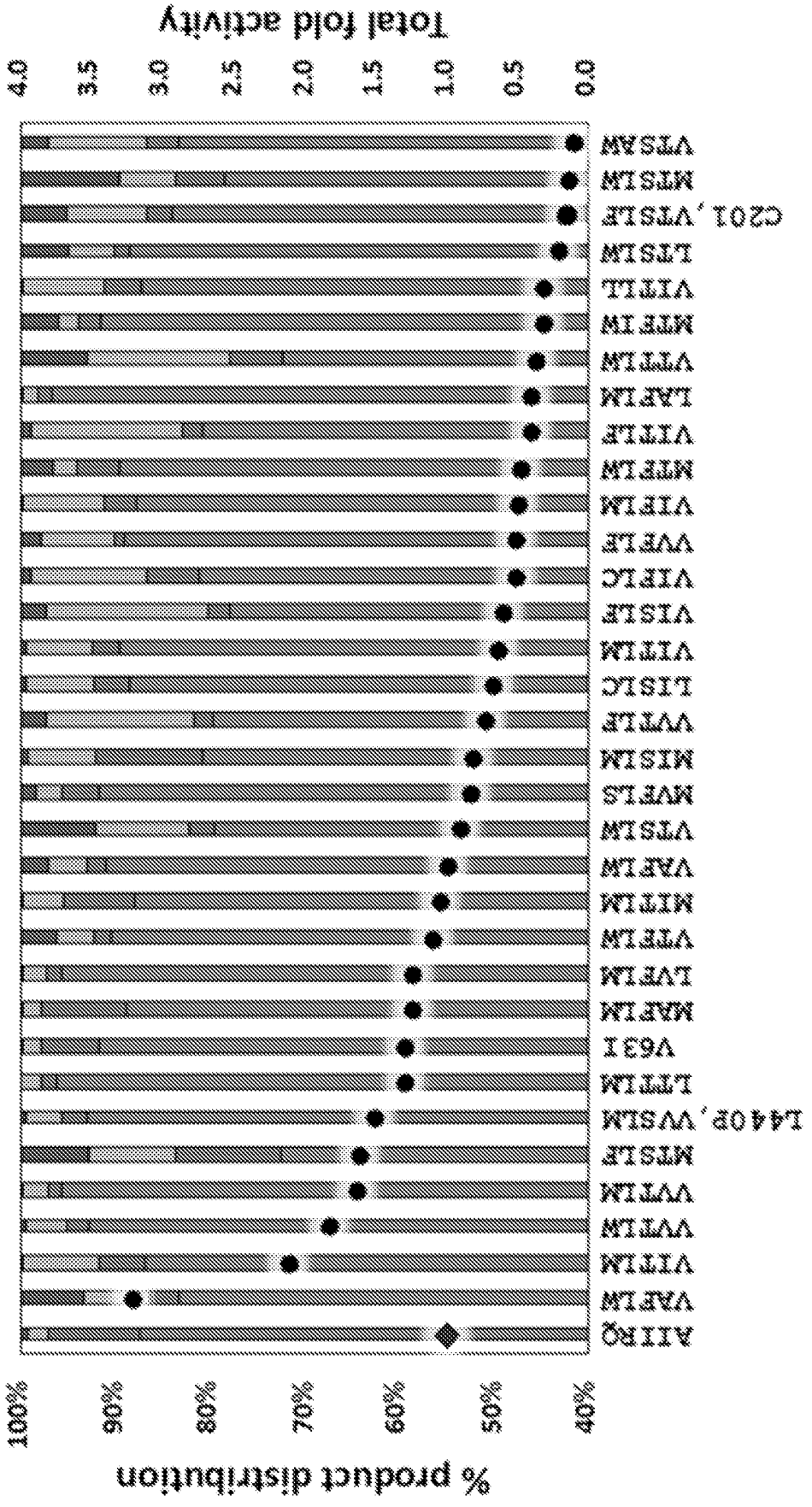

Total fold activity
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
VTSAW
MTSIW
C201,VTSLF
LTSIW
VITIL
MTFIW
VTTIW
LAFIM
VITLF
MTFLW
VIFIM
VVFLF
VIFLC
VISLF
VITIM
LISLC
VVTLF
MISIM
MVFLS
VTSIW
VAFIW
MITIM
VTFIW
LVFIM
MAFIM
V631
LTTIM
I440P,VVSIM
MTSLF
VVTIM
VVTIW
VITIM
VAFIW
AIIRQ
100%
90%
80%
70%
60%
50%
40%
% product distribution

Fig. 14

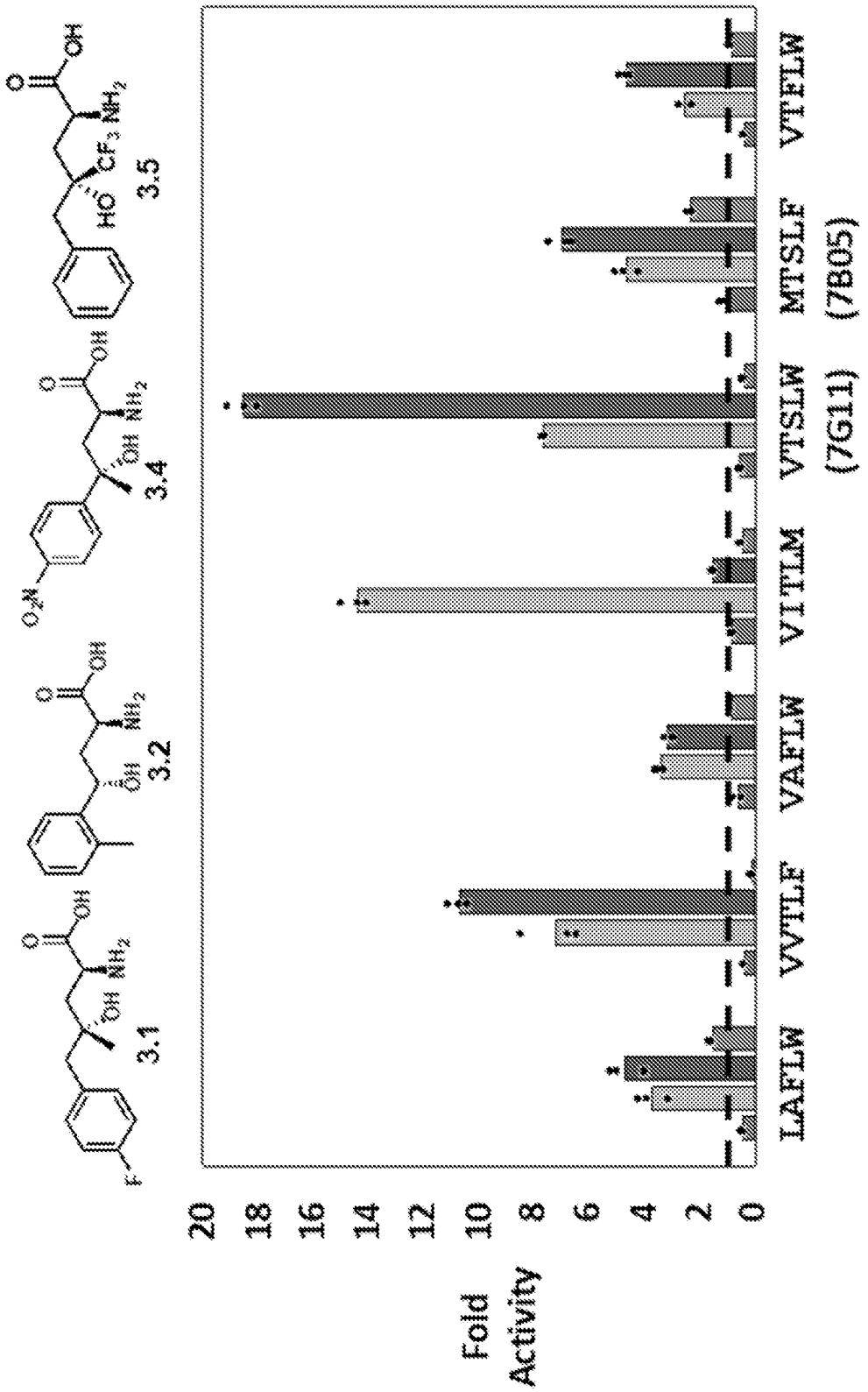

3.1
3.2
3.4
3.5
Fold Activity
0
2
4
6
8
10
12
14
16
18
20
LAFLW
VVTLF
VAFLW
VITLM
VTSLW
MTSLF
VTFLW
(7G11)
(7B05)

Fig. 15

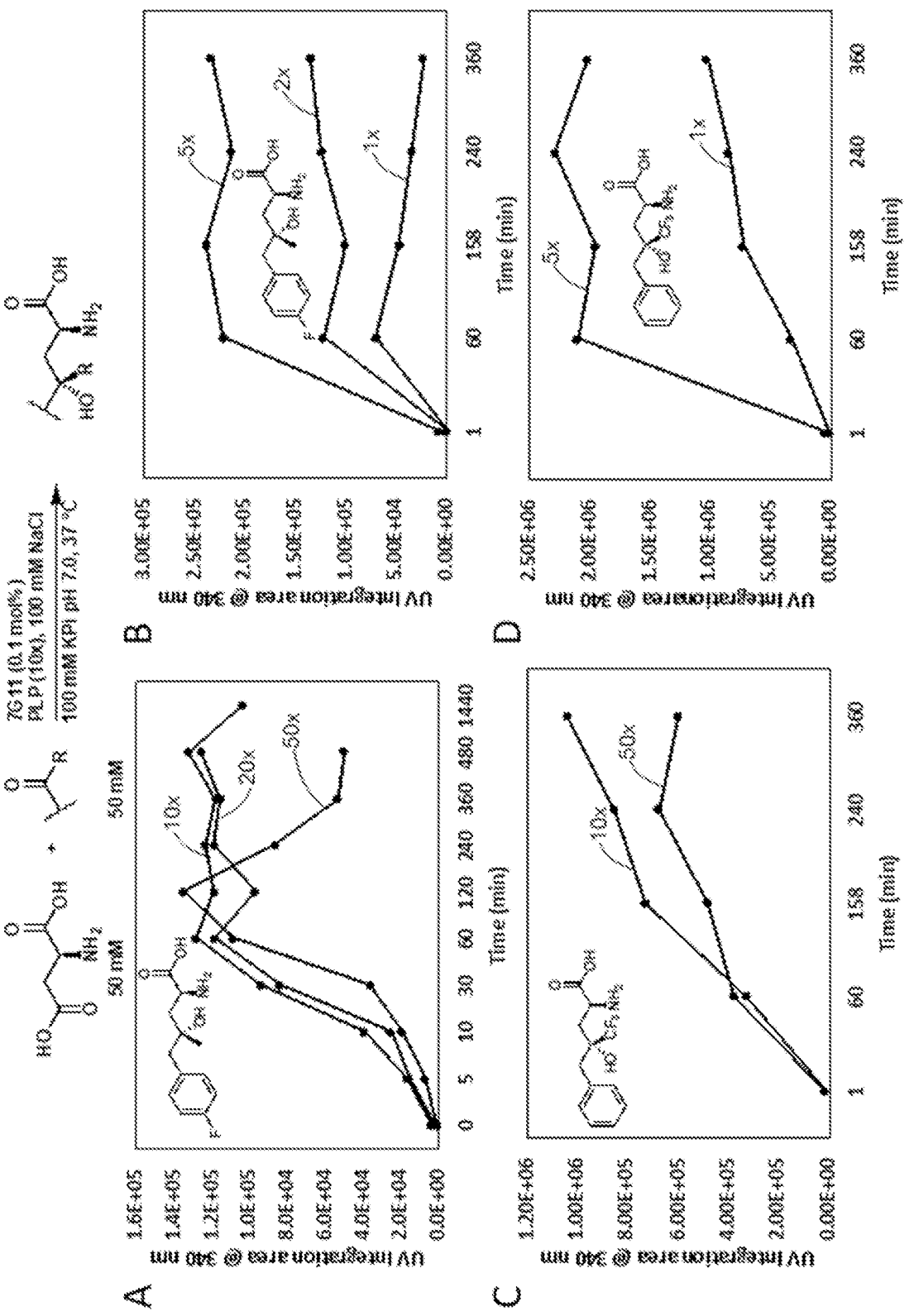

7G11 (0.1 mol%)
PLP (10x), 100 mM NaCl
100 mM KPi pH 7.0, 37 °C
50 mM
50 mM
A
B
C
D
UV integration area @ 340 nm
Time (min)
10x
20x
50x
5x
2x
1x

Fig. 16

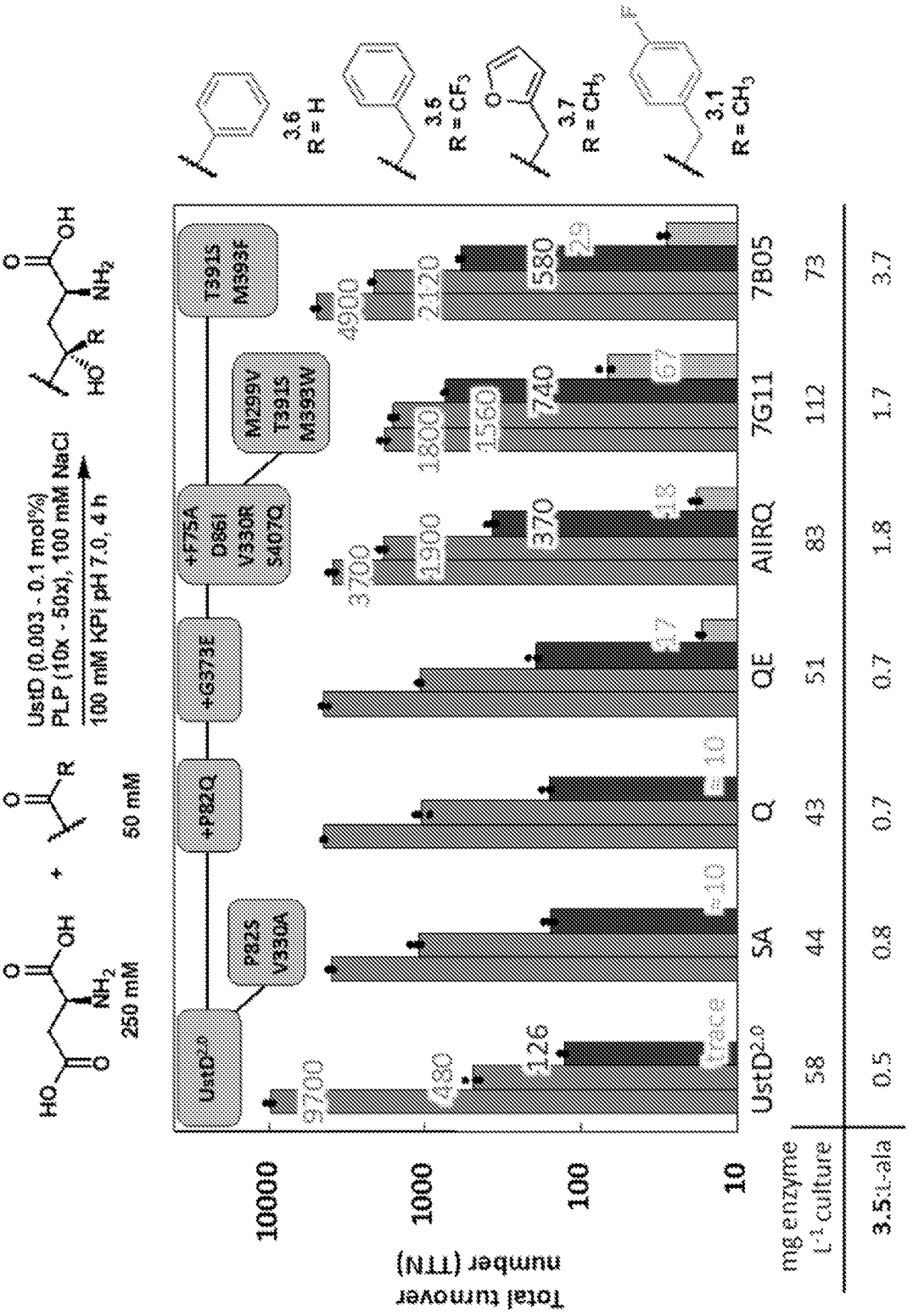

UstD (0.003 - 0.1 mol%)
PLP (10x - 50x), 100 mM NaCl
100 mM KPi pH 7.0, 4 h
250 mM
50 mM
3.6 R = H
3.5 R = CF3
3.7 R = CH3
3.1 R = CH3
Total turnover number (TTN)
10000
1000
100
10
UstD2.0
P82S V330A
+P82Q
+G373E
+F75A D86I V330R S407Q
M296V T391S M393W
T391S M393F
9700
480
126
trace
3700
1900
370
48
1800
1560
740
67
4900
2120
580
20
UstD2.0
5A
Q
QE
AllRQ
7G11
7B05
mg enzyme L-1 culture
58
44
43
51
83
112
73
3.5i-ala
0.5
0.8
0.7
0.7
1.8
1.7
3.7

Fig. 17

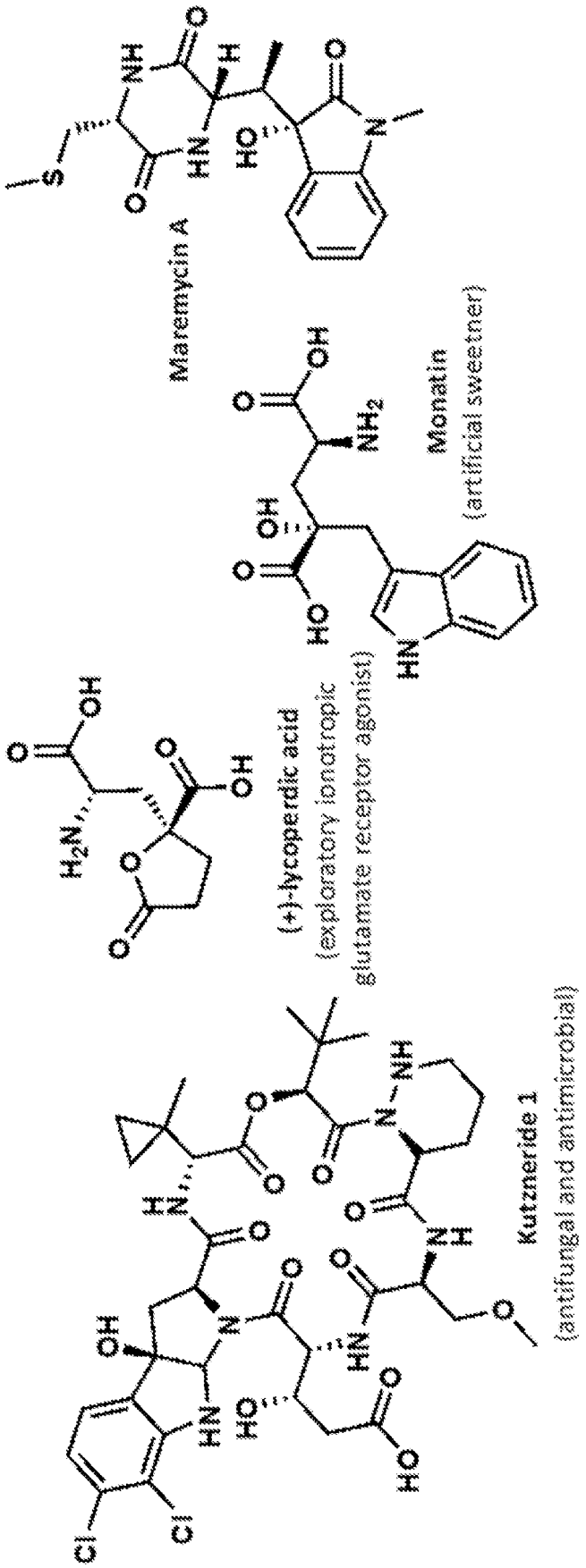

Maremycin A
Monatin
(artificial sweetner)
(+)-lycoperdic acid
(exploratory ionotropic
glutamate receptor agonist)
Kutzneride 1
(antifungal and antimicrobial)

Fig. 18

Fig. 19A Mukaiyama-aldol

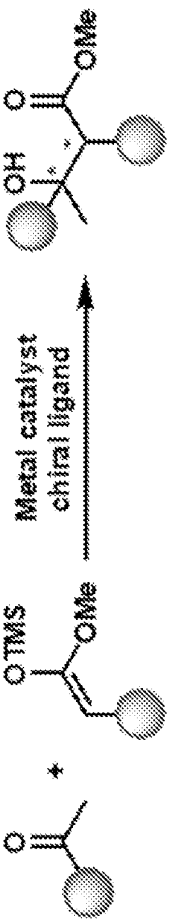


Fig. 19B Decarboxylative aldol

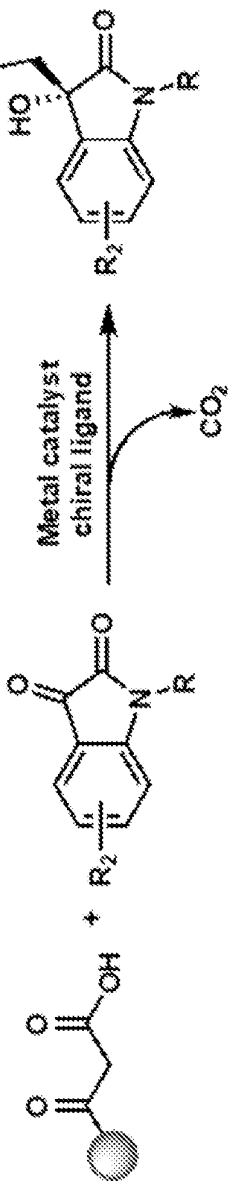


Fig. 19C Organocatalyst aldol

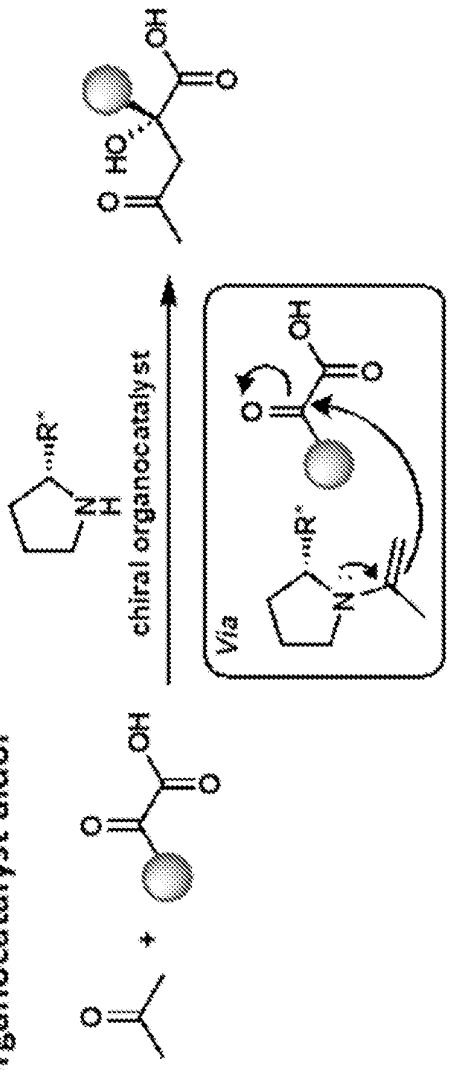


Fig. 20A HMG aldolase

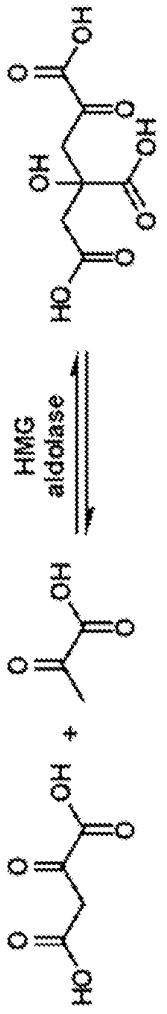


Fig. 20B TPP dependent YerE

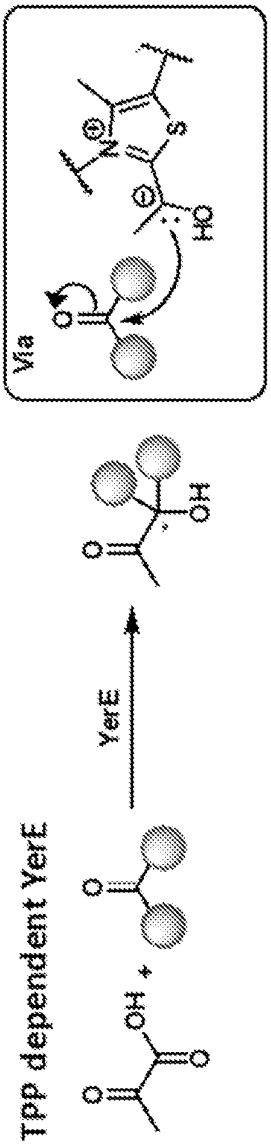


Fig. 20C Proteinase AMP

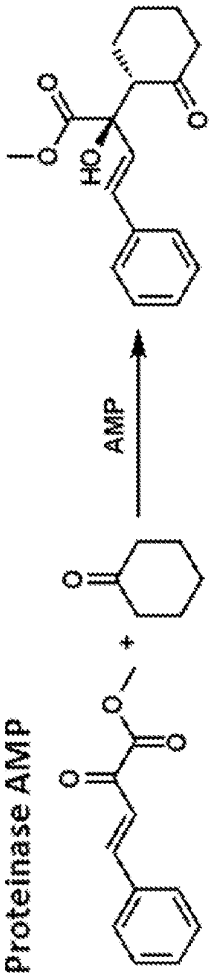
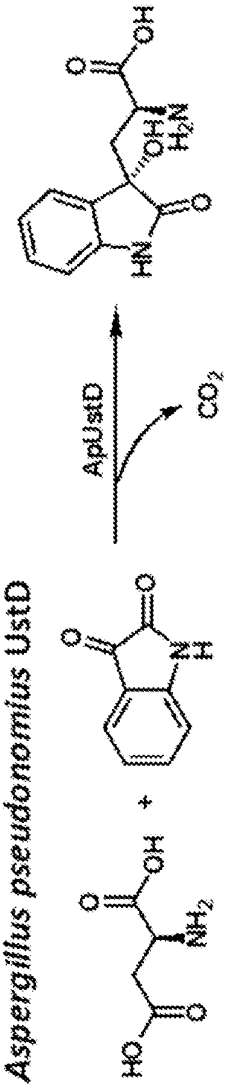


Fig. 20D *Aspergillus pseudonomius* UstD



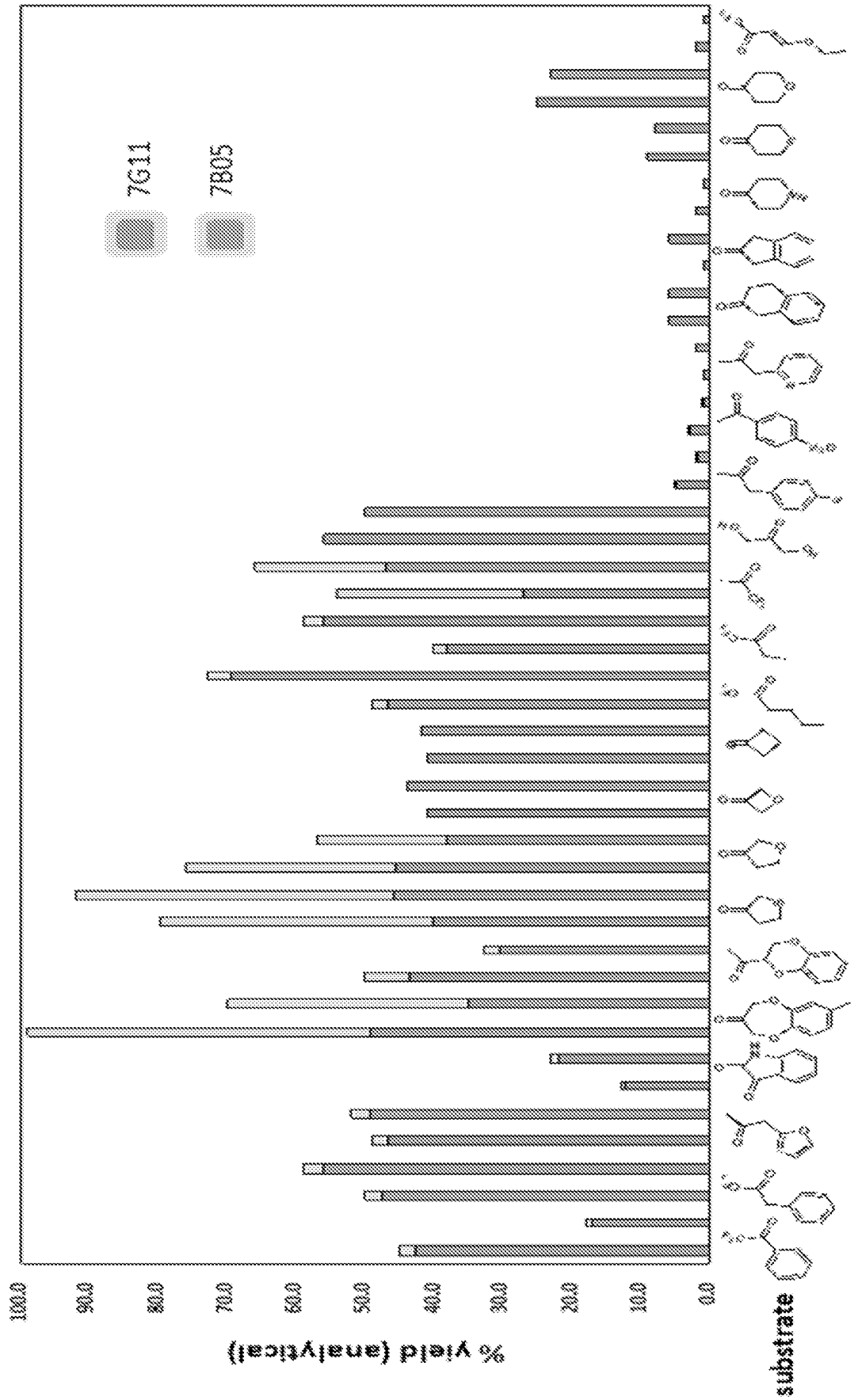
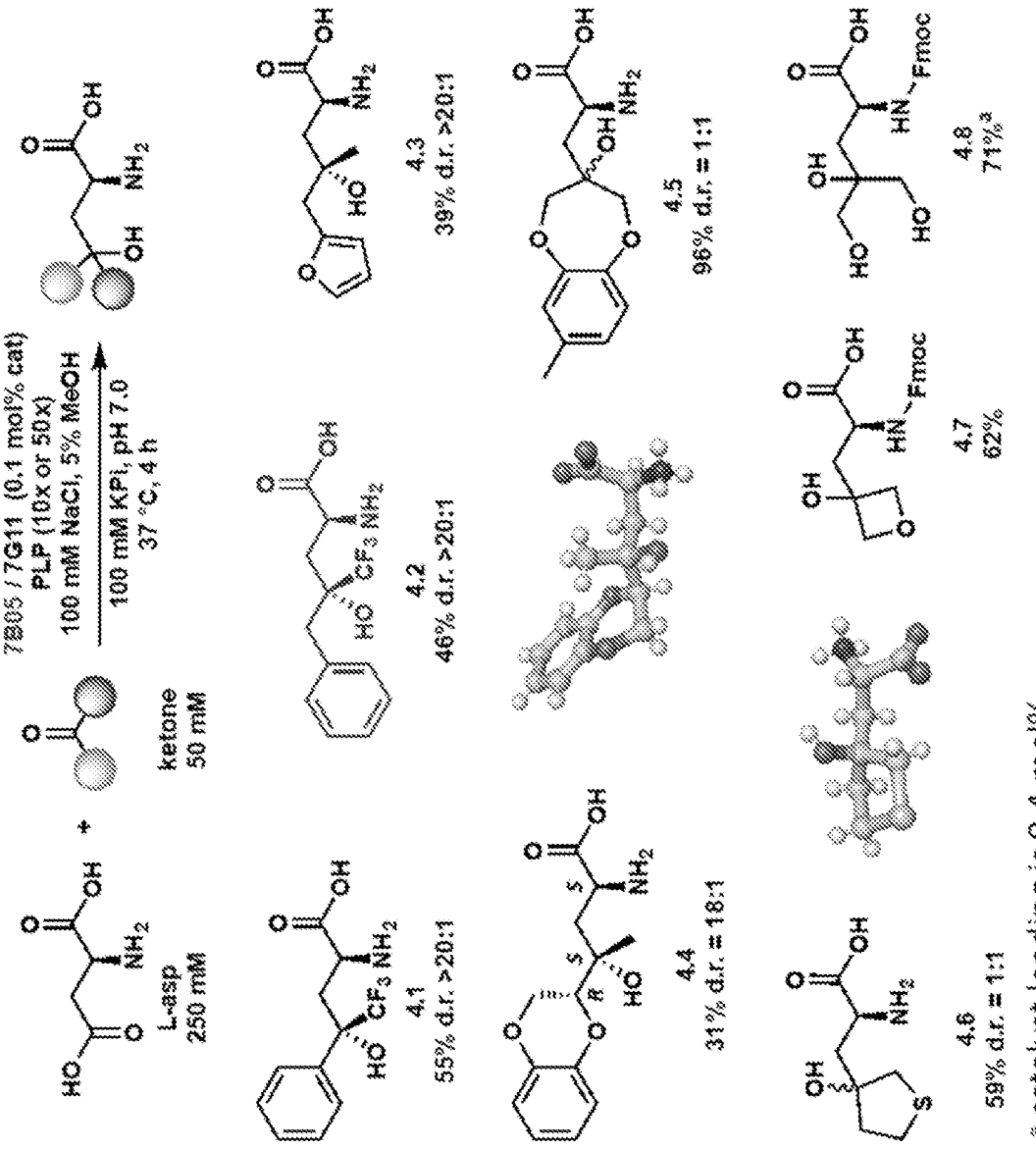

7G11
7B05
100.0
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
% yield (analytical)
substrate

Fig. 21



[a] catalyst loading is 0.4 mol%

Fig. 22

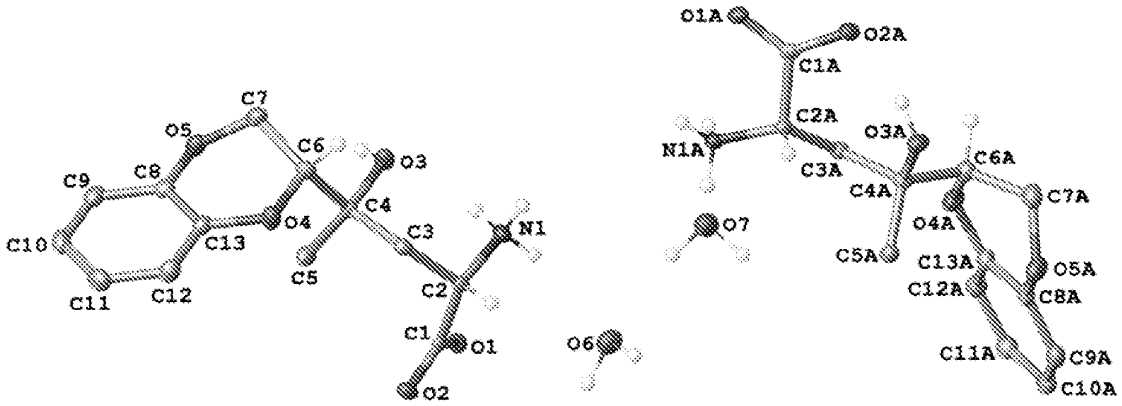


Fig. 23A

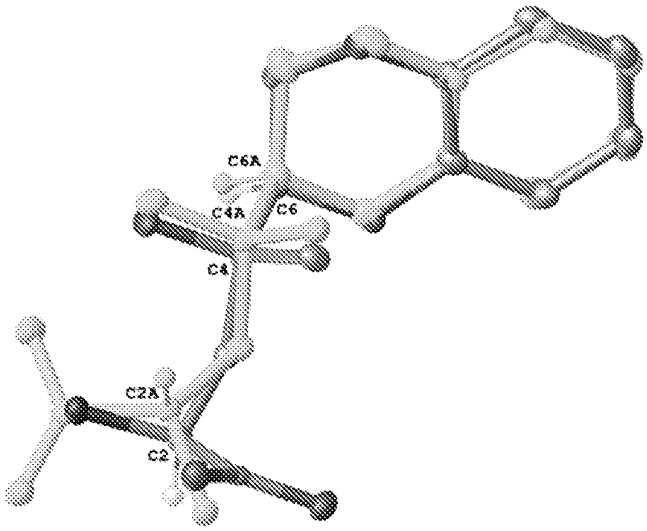


Fig. 23B

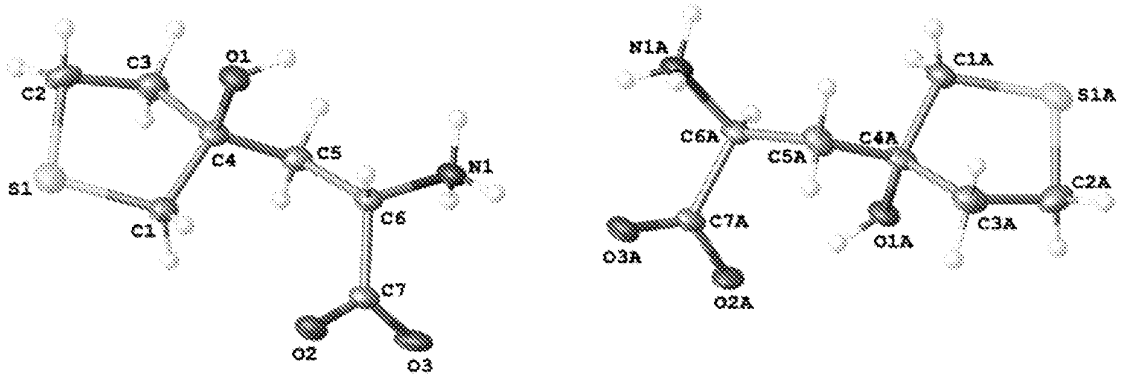


Fig. 23C

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2025/025073

A. CLASSIFICATION OF SUBJECT MATTER

IPC: ***C12N 9/10*** (2025.01); ***C12P 13/12*** (2025.01)

CPC: ***C12N 9/13; C12P 13/12;*** *C12N 9/88; C12P 13/04; C12Y 401/01*

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

See Search History Document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

See Search History Document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

See Search History Document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2021/0115480 A1 (WISCONSIN ALUMNI RESEARCH FOUNDATION) 22 April 2021 (22.04.2021) para [0008]; SEQ ID NO: 1	1
A	ELLIS et al. Biocatalytic synthesis of non-standard amino acids by a decarboxylative aldol reaction. Nature Catalysis, 21 February 2022, Vol. 5, No. 2, Pgs. 136-143 pg 139, col 1, para 4 – pg 139, col 2, para 1	1

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"D" document cited by the applicant in the international application

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
26 June 2025 (26.06.2025)	**04 August 2025 (04.08.2025)**

Name and mailing address of the ISA/US	Authorized officer
COMMISSIONER FOR PATENTS **MAIL STOP PCT, ATTN: ISA/US** **P.O. Box 1450** **Alexandria, VA 22313-1450** **UNITED STATES OF AMERICA**	**HARRY KIM**
Facsimile No. **571-273-8300**	Telephone No. **PCT Help Desk: 571-272-4300**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2025/025073

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing:
   a. ☑ forming part of the international application as filed.
   b. ☐ furnished subsequent to the international filing date for the purposes of international search (Rule 13*ter*.1(a)),
      ☐ accompanied by a statement to the effect that the sequence listing does not go beyond the disclosure in the international application as filed.

2. ☐ With regard to any nucleotide and/or amino acid sequence disclosed in the international application, this report has been established to the extent that a meaningful search could be carried out without a WIPO Standard ST.26 compliant sequence listing.

3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2025/025073

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☑ Claims Nos.: **4-21, 24**
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1.

Group I+, claims 1-3, directed to an unnatural, mutant protein comprising an amino acid sequence at least 80% identical to a UstD sequence, wherein the amino acid sequence comprises one or more mutations. The UstD mutant protein will be searched to the extent that the UstD mutant protein sequence encompasses SEQ ID NO: 1 with a residue other than K at a position corresponding to position 2. The first named invention was determined based on this being the first listed UstD mutant protein sequence and mutant residue (claim 1). This first named invention has been selected based on the guidance set forth in section 10.54 of the PCT International Search and Preliminary Examination Guidelines. It is believed that claim 1 encompass this first named invention, and thus these claims will be searched without fee to the extent that the UstD mutant protein sequence comprises SEQ ID NO: 1 with a residue other than K at a position corresponding to position 2. Additional UstD mutant protein sequence(s) or mutant residue(s) will be searched upon the payment of additional fees. Applicants must specify the claims that encompass any additionally elected UstD mutant protein sequence(s) or mutant residue(s). Applicants must further indicate, if applicable, the claims which encompass the first named invention, if different than what was indicated above for this group. Failure to clearly identify how any paid additional invention fees are to be applied to the "+" group(s) will result in only the first claimed invention to be searched. An exemplary election would be where the UstD mutant protein sequence comprises SEQ ID NO: 1 with a residue other than V at a position corresponding to position 63, (claim 1).

Group II, claims 22-23, directed to a method of making a gamma-hydroxy amino acid, the method comprising contacting an unnatural, mutated protein with an amino acid and a ketone-containing substrate to yield the gamma-hydroxy amino acid.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2025/025073

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☑ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.: **1, limited to SEQ ID NO: 1 with a residue other than K at a position corresponding to position 2**

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.