Machine learning discovery of missing links that mediate alternative branches to plant alkaloids

Christopher Vavricka (chrisv@people.kobe-u.ac.jp)
Kobe University https://orcid.org/0000-0003-2101-4359
Shunsuke Takahashi
Tokyo Denki University
Naoki Watanabe
Kobe University
Musashi Takenaka
Kobe University
Mami Matsuda
Kobe University
Takanobu Yoshida
Kobe University
Hiromasa Kiyota
Okayama University
Hiromichi Minami
Ishikawa Prefectural University
Jun Ishii
Kobe University https://orcid.org/0000-0003-2568-515X
Kenji Tsuge
Kobe University
Michihiro Araki
Kobe University
Akihiko Kondo
Kobe University https://orcid.org/0000-0003-1527-5288
Tomohisa Hasunuma
Kobe University https://orcid.org/0000-0002-8382-2362

Article

Keywords:

Posted Date: April 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-184114/v1

Version of Record: A version of this preprint was published at Nature Communications on March 16th, 2022. See the published version at https://doi.org/10.1038/s41467-022-28883-8.

Abstract

Engineering the microbial production of secondary metabolites is limited by the known reactions of correctly annotated enzymes in sequence databases. To expand the range of biosynthesis pathways, machine learning is herein demonstrated for the discovery of missing link enzymes, using benzylisoquinoline alkaloid production as a model application with potential to revolutionize the paradigm of sustainable biomanufacturing. Bacterial studies utilize a tetrahydropapaveroline pathway, whereas plants are reported to contain a more stable norcoclaurine pathway, which is exploited in yeast. However, committed aromatic precursors are currently produced by microbial enzymes that remain elusive in plants, and additional downstream missing links are hidden within highly-duplicated plant gene families. Accordingly, the machine learning enzyme selection algorithm is first applied to predict the missing links from diverse candidate sequences. Metabolomics-based characterization of selected sequences reveals distinct oxidases and carboxy-lyases in reconstructed plant gene-only benzylisoquinoline alkaloid pathways from tyrosine. Synergistic application of aryl acetaldehyde producing enzymes results in enhanced production through hybrid norcoclaurine and tetrahydropapaveroline pathways. Transplantation of features into homologous enzyme templates leads to the highest levels of bacterial norcoclaurine and N-methylcoclaurine. Mechanism-directed isotope tracing patterns confirm alternative flux branches from aromatic precursors to alkaloids. This machine learning-driven workflow can be adapted to numerous pathways.

Introduction

Strain engineering is now a reliable approach to scale up production of target metabolites by integrating known genes, and applying simple yet effective metabolic engineering strategies[1]. But engineering the microbial production of secondary metabolites reaches the limitation of the characterized enzymes present in sequence databases, where many annotations are incorrect. In reality, there are millions of enzyme variants to choose from for each desired reaction, and a great abundance of variations are still hidden in nature with unknown sequence and function. In this way the evolution of nature over millions of years can be viewed as a highly diverse screening resource for synthetic biologists. Accordingly, the rational discovery of natural enzymes with novel functions is a powerful and inevitable approach to improve microbial bioproduction pathways[2-6].

The current study clarifies how machine learning can reveal novel enzyme functions with potential for sustainable biosynthesis. In our previous study, aromatic acetaldehyde synthase (AAS) was predicted with the enzyme selection software M-path to improve production of valuable alkaloids[7]. However, only EC digits could be predicted with M-path and the actual selection of candidate sequences had to be performed by *human intuition*. This issue is addressed by developing a support vector machine (SVM) algorithm[8] to automatically select

specific enzymes sequences: an upgrade that enables *computer automated* Design, Build, Test and Learn (DBTL) cycles.

To prove the concept of machine learning enzyme prediction, conversion of tyrosine to benzylisoquinoline alkaloid (BIA) is selected as the target pathway for optimization (Fig. 1). BIAs are precursors to opioid analgesic medications that are currently mass-produced by industrially grown *Papaver somniferum* plants, which are a historical target for human-directed evolution of natural product production. While opioid misuse is a global problem, natural and semi-synthetic opioids derived from the BIA reticuline actually result in fewer deaths than less expensive and overly potent synthetic opioids (CDC Opioid Data Analysis and Resources). With diverse potential, natural BIAs have been shown to inhibit coronavirus[9], and the BIA norcoclaurine is a β2-adrenergic receptor agonist that is present in edible plants, medicinal herbs and sports supplements[10,11].

BIA production in *Escherichia coli* has utilized bacterial monoamine oxidase and insect DHPAAS to generate toxic 3,4-dihydroxyphenylacetaldehyde (DHPAA)[12]. However, the DHPAA containing pathways result in rapid loss of unstable catechol containing intermediates[7,12-14]. Other reports suggest that plants favor the 4-hydroxyphenylacetaldehyde (4HPAA) pathway to norcoclaurine (Fig. 1a), which may be more stable due to lack of a catechol group in early intermediates. Therefore, plant 4HPAA pathways offer potential to prevent loss of BIA intermediates in *E. coli*. Furthermore, the combination of 4HPAA and DHPAA pathways may also improve utilization of tyrosine and aryl acetaldehydes. Despite success with the 4HPAA pathway in yeast[1,4,15,16], and many discussions on the expected phenylpyruvate decarboxylase (PPDC, EC 4.1.1.43) and AAS (EC 4.1.107-9) activities in plants, no enzymes to produce aryl acetaldehydes 4HPAA or DHPAA have been characterized from high alkaloid producing poppy plants[17]. Moreover, no plant sequence annotated as phenylpyruvate decarboxylase can be found from public databases, and numerous *P. somniferum* cytochrome P450 (CYP450) monooxygenases (EC 1.14.14) require complex clarification. This serious limitation in known enzymes is addressed by applying machine learning to predict the essential missing links in plant alkaloid pathways shown as dotted arrows in Figure 1.

To guide the selection of sequences from over 100 candidates present throughout highly duplicated carboxy-lyase and oxidase gene families, 8 refined SVM models are built from training sequences classified using structure-based rules. Then, to verify the machine learning prediction, 50 strains expressing combinations of candidate sequences and analogous templates are screened using liquid chromatography–mass spectrometry (LC-MS)-, capillary electrophoresis–MS (CE-MS)- and gas chromatography–MS (GC-MS)-based metabolomics. As a result, AAS, PPDC, *N*-methylcoclaurine 3-hydroxylase (NMCH) and CYP450 reductase (CPR) enzymes with distinct features are identified as missing links that mediate uncharacterized branches of the *Papaver somniferum* alkaloid pathway. Synergistic combination of predicted

sequences together with homologous microbial templates affords 96.7 mg/L norcoclaurine, 71.8 mg/L *N*-methylcoclaurine (NMC) and 24.6 mg/L reticuline, without using any strain engineering. The alternative branches of flux from tyrosine to downstream alkaloids are confirmed using dynamic metabolic profiling[5] with mechanism-directed deuterium labeling patterns.

Prediction and discovery of *P. somniferum* TyDC1 as a missing link to plant arly acetaldehydes

DHPAA and THP are more easily oxidized and more toxic than their corresponding 4-hydroxyphenyl analogues[12]. Therefore, the 4HPAA pathway to norcoclaurine is explored as a first example of machine learning enzyme selection to construct an improved metabolic pathway (Fig. 2). Our previous M-path analysis identified 4-hydroxyphneylacetaldehyde synthase (4HPAAS, 4.1.1.108) to mediate 4HPAA production from tyrosine; however specific 4HPAAS sequences are incompletely annotated throughout most databases. In this study the term AAS is used to cover plant-type AAS enzymes 4HPAAS and phenylacetaldehyde synthase (PAAS, 4.1.1.109) as well as insect 3,4-dihydroxyphenylacetaldehyde synthase (DHPAAS, EC 4.1.1.107), because substrate specificities are often mixed throughout these groups.

Unclear variations within the plant-type AAS group, which may act upon a wide range of substrates including phenylalanine, tyrosine, 3,4-dihydroxy-L-phenylalanine (L-DOPA), tryptophan and histidine, further complicates the selection of a correct sequence based on phylogenetic and structural analyses alone. Accordingly, no AAS enzyme from *P. somniferum* has been clearly established[17]. To overcome this challenge in prediction, our SVM-based algorithm[8] is first applied to select AAS from *P. somniferum* homologs annotated as tyrosine/DOPA decarboxylase (TyDC) (Fig. 2b and Supplementary Table 1).

AAS often exhibits mixed carboxy-lyase and oxidative deamination activities, so separate SVM models for aromatic amino acid decarboxylase (AAAD) and AAS (Supplementary Table 1) were trained using sequences classified as described in the methods. According to database annotations and previous reports, *P. somniferum* TyDC (PsTyDC) proteins should be expected to catalyze the decarboxylation of tyrosine to form tyramine, and possibly L-DOPA conversion to dopamine[20,21]. In contrast, SVM-based models show that while most of the 8 full length PsTyDC sequences have high potential for AAAD activity, PsTyDC1-8 also appear in AAS prediction space, with PsTyDC1, PsTyDC2 and PsTyDC6 scoring highest for positive AAS prediction (Fig. 2b and Supplementary Table 1). PsTyDC6 contains AAAD-like active site residues Y98, F99, H205, Y350 and S372, and scores high for AAAD prediction, suggesting PsTyDC1 as a better candidate. PsTyDC2 scores high for AAS and low for AAAD and should be explored as AAS in future studies, but it was suggested that PsTyDC2, which also contains AAAD-like residues Y100, F101, H203, Y348 and S370, mediates typical tyrosine decarboxylase activity according to the original report[21]. In addition to positive AAS prediction

scores, PsTyDC1 and PsTyDC3 contain unique active site residues L205 and I370 (serine in all other TyDC sequences), respectively, further suggesting atypical activities of these test sequences.

In accordance with the SVM prediction, expression of PsTyDC1 in *E. coli* using various plasmids leads to *in vivo* production of norcoclaurine, with significantly low production of tyramine (Fig. 2c), indicating that PsTyDC1 has specific 4HPAAS activity with low bifunctional AAAD activity. As a positive AAS control, PsTyDC1-Y98F-F99Y-L205N with engineered active site residues transplanted from insect DHPAAS, is also presented with similar results to that of wild-type PsTyDC1. After substitution of PsTyDC1-L205 to a histidine residue found in typical AAAD, the decarboxylation product tyramine increases dramatically (Fig. 2c and Supplementary Fig. 1). PsTyDC1 mediated production of norcoclaurine is further confirmed in strains with additional variations in the alkaloid pathway (Supplementary Fig. 2 and Table 1). Consistent with lower AAS prediction scores, wild-type PsTyDC3 produces lower ratios of AAS product norcoclaurine to AAAD product tyramine, compared to ratios of wild-type PsTyDC1.

Table 1 | Aromatic producing strains of this study

Strain	Genotype	Conditions	Products
BL21(DE3)	F– ompT gal dcm lon hsdSB(rB–mB–) λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS)	-	-
BL21-AI	F– ompT gal dcm lon hsdSB(rB–mB–) [malB+]K-12(λS) araB::T7RNAP-tetA	-	-
T1-01-DE3	pCDFD-TfNCS-PsTyDC1	LB-Tyr+DA	NC
T1-02-DE3	pCDFD-TfNCS-PsTyDC1-S	LB-Tyr+DA	Tyramine
T1-03-DE3	pCDFD-TfNCS-PsTyDC1-T	LB-Tyr+DA	NC
T3-01-DE3	pCDFD-TfNCS-PsTyDC3	LB-Tyr+DA	trace NC
T3-02-DE3	pCDFD-TfNCS-PsTyDC3-S	LB-Tyr+DA	Tyramine
T3-03-DE3	pCDFD-TfNCS-PsTyDC3-T	LB-Tyr+DA	NC
T1-04-DE3	pCDFD-PsONCS3-PsTyDC1	LB-Tyr+DA	trace NC
T1-05-DE3	pCDFD-PsONCS3-PsTyDC1-S	LB-Tyr+DA	Tyramine
T1-06-DE3	pCDFD-PsONCS3-PsTyDC1-T	LB-Tyr+DA	trace NC
T3-04-DE3	pCDFD-PsONCS3-PsTyDC3	LB-Tyr+DA	trace NC
T3-05-DE3	pCDFD-PsONCS3-PsTyDC3-S	LB-Tyr+DA	Tyramine
T3-06-DE3	pCDFD-PsONCS3-PsTyDC3-T	LB-Tyr+DA	trace NC
T1-07-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-TfNCS-PsTyDC1	M9-Tyr+DA	Reticuline
T1-08-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-TfNCS-PsTyDC1-S	M9-Tyr+DA	Reticuline
T1-09-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-TfNCS-PsTyDC1-T	M9-Tyr+DA	Reticuline
T1-10-DE3	pACYC-3CjMTs-PpDDC, pCDFD-PsONCS3-PsTyDC1	**TB-DOPA***; TB-Tyr+DOPA	**1 μM Reticuline**; NC
P1-01-AI	pBAD-PsPDC1	M9-4HPP	Tyrosol
P2-01-AI	pBAD-PsPDC2	M9-4HPP	
P3-01-AI	pBAD-Ps2HCLL	M9-4HPP	
P1-02-AI	pACYC-3CjMTs-PpDDC, pCDFD-PsONCS3, pBAD-PsPDC1	M9-Tyr+DA; **TB-DOPA***; TB-Tyr*+DOPA*	NC; **1.5 μM Reticuline**; NMC, Reticuline
P1-03-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-PsONCS3-PsTyDC1, pBAD-PsPDC1	TB-Tyr+DA	NC
P1-04-AI	pACYC-3CjMTs-PpDDC, pCDFD-PsONCS3-PsTyDC1, pBAD-PsPDC1	**TB-DOPA**; TB-Tyr*+DOPA*	**1.5 μM Reticuline**; NMC, Reticuline
N1-01-DE3	pET23a-3PsMTs, pCOLAD-PsNMCH-PsCPR	TB-NC	27.8 μM Reticuline
N1-02-DE3	pET23a-3PsMTs, pCOLAD-PsNMCH-H203Y-PsCPR	TB-NC	15.8 μM Reticuline
N1-03-DE3	pET23a-3PsMTs, pCOLAD-PsNMCH-AtATR2	TB-NC	15.9 μM Reticuline
N1-04-DE3	pET23a-3PsMTs, pCOLAD-PsNMCH-H203Y-AtATR2	TB-NC	8.0 μM Reticuline
N2-01-DE3	pET23a-3PsMTs, pCOLAD-EcNMCH-AtATR2	TB-NC	4.9 μM Reticuline
N2-02-DE3	pET23a-3PsMTs, pCOLAD-EcNMCH-Y202H-AtATR2	TB-NC	8.4 μM Reticuline
N2-03-DE3	pET23a-3PsMTs, pCOLAD-EcNMCH-PsCPR	TB-NC	3.7 μM Reticuline
N2-04-DE3	pET23a-3PsMTs, pCOLAD-EcNMCH-Y202H-PsCPR	TB-NC	3.6 μM Reticuline
DS-01-DE3	pACYC-3CjMTs-PpDDC-S, pCDFD-PsONCS3	M9-DOPA	
DT-01-DE3	pACYC-3CjMTs-PpDDC-T, pCDFD-TfNCS	M9-DOPA	16.8 μM THP, 2.5 μM Reticuline
DT-02-DE3	pACYC-3CjMTs-PpDDC-T, pCDFD-PsONCS3	M9-DOPA	34 μM THP, 6.4 μM Reticuline
DS-02-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-PpDDC-S	M9-Tyr+DA	NC
DD-01-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-PpDDC-D	M9-Tyr+DA	
DT-03-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-PpDDC-T	M9-Tyr+DA; **M9-DOPA***	**2.1 μM THP, 0.26 μM Reticuline**
DQ-01-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-PpDDC-Q	M9-Tyr*+DA	
DS-03-DE3	pACYC-3CjMTs-PpDDC, pCDFD-CjNCS-PpDDC-S, pET23a-EcHpaBC	M9-Tyr	NC
DD-02-DE3	pACYC-3CjMTs-PpDDC, pCDFD-CjNCS-PpDDC-D, pET23a-EcHpaBC	M9-Tyr	
DT-04-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-PpDDC-T, pET23a-EcHpaBC	M9-Tyr	DA, THP
T1-11-DE3	pACYC-3CjMTs-PpDDC, pCDFD-PsONCS3-PsTyDC1, pET23a-EcHpaBC	LB-Tyr	DA
A1-01-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-ARO10	**TB-Tyr*+DA***;	**356 μM NC**, **240 mM NMC**; 1 μM Reticuline

		TB-DOPA*+DA*	
A1-02-DE3	pACYC-3CjMTs-PpDDC, pCDFD-CjNCS-ARO10		
A1-03-DE3	pACYC-3CjMTs-PpDDC, pCDFD-CjNCS-ARO10, pET23a-EcHpaBC	M9-Tyr*; **TB-Tyr***	4HPP, DA, NC; **THP**, **NMC**
DS-04-DE3	pACYC-3CjMTs-PpDDC, pCDFD-CjNCS-PpDDC-S, pET23a-EcHpaBC, pE-BmDHPAAS		
A1-05-DE3	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-ARO10, pTrc-BmDHPAAS-T	TB-Tyr+DOPA	DA
A1-06-AI	pACYC-3CjMTs-PpDDC-T, pCDFD-CjNCS-ARO10, pTXB1-PsTyDC1	TB-Tyr+DOPA; **TB-Tyr+DA**	74.9 µM Reticuline; 112 µM NMC
P1-05-AI	pACYC-3CjMTs-PsNMCH, pCDFD-CjNCS-PpDDC-T, pBAD-PsPDC1	TB-Tyr+DOPA	
P1-06-DE3	pACYC-3CjMTs-PpDDC-T, pCDFD-PsONCS3, pBAD-PsPDC1	TB-Tyr+DOPA	3.7 µM Reticuline
P1-07-AI	pACYC-3CjMTs-PpDDC-T, pCDFD-PsONCS3, pBAD-PsPDC1	TB-Tyr+DOPA	61.7 µM Reticuline

Strains with names ending with "DE3" are derived from BL21(DE3), and strains ending with "AI" are derived from BL21-AI. Plasmid details are given in Supplementary Table 6. The last two columns list successfully tested conditions (growth medium and added substrate) and produced BIAs or BIA precursors. Concentrations of extracted NMC and reticuline per culture volume are listed for AI-01-DE3, A1-06-AI, P1-06-DE3 and P1-07-AI; all other listed concentrations represent titers in filtered culture medium. Only product titers precisely quantified above 1 µM are listed. Matched substrates and corresponding products are indicated by bold font. Substrates marked with * include isotopes (tyrosine-^{13}C, tyrosine-d_4, L-DOPA-d_3 or dopamine-d_2). Abbreviations: PsONCS3 - *P. somniferum* multi-domain NCS, CjNCS - *Coptis japonica* NCS, 2CHLL - 2-hydroxyacyl-CoA ligase-like, S - single variant, D - double variant, T - triple variant, Q - quadruple variant, 3CjMTs - *C. japonica* 6OMT, CNMT and 4OMT, 3PsMTs - *P. somniferum* 6OMT, CNMT and 4OMT, Tyr - tyrosine, 4HPP - 4-hydroxyphenylpyruvate, DA - dopamine, DOPA - 3,4-dihydroxy-L-phenylalanine, NC - norcoclaurine, THP - tetrahydropapaveroline, NMC - *N*-methylcoclaurine.

P. somniferum PDC1 decarboxylates 4-hydroxyphenylpyruvate in an alternative 4HPAA bypass pathway

Phenylpyruvate decarboxylase (PPDC) is an alternative to AAS for production of aryl acetaldehyde intermediates 4HPAA and DHPAA (Fig. 2, d-f). Previous reports hypothesize that *P. somniferum* should contain PPDC with specificity towards 4-hydroxyphenylpyruvate (4HPP)[17]; however no plant protein accessions are found with the annotation of phenylpyruvate decarboxylase. In comparison to the known enzymes with PPDC activity, including *Azospirillum brasilense* ipdC[22], *Lactococcus lactis* KdcA[23], and yeast ARO10[24], the PsPDC1 active site more closely resembles that of typical pyruvate decarboxylases[25] (Fig. 2d). Yet, in SVM prediction models constructed according to the methods section, *P. somniferum* PDC1 (PsPDC1) scores high for PPDC activity, relative to other homologous sequences (Fig. 2e, Supplementary Fig. 3a and Supplementary Table 2). Two additional test candidates, PsPDC2 and 2-hydroxyacyl-CoA ligase-like sequences, are predicted with lower scores, and found to exhibit lower levels of 4HPP decarboxylase activity, compared to that of PsPDC1 (Fig. 2, e - f, and Supplementary Fig. 3).

In vivo screenings with PsPDC1 reveal the alternative alkaloid route through 4HPP, and this PPDC bypass is distinct from the direct aromatic amino acid branch mediated by PsTyDC1 (Fig. 2f). Application of PsPDC1 for conversion of tyrosine through the 4HPP and 4HPAA containing

pathway results in improvement in norcoclaurine titers to the >10 µM range (Fig 2f), compared the 100-200 nM range of PsTyDC1 (Fig. 2c and Supplementary Figure 2).

Paired prediction of NMCH and CPR missing links extend the 4HPAA pathway

After constructing the 4HPPA pathway to noroclaurine, *P. somniferum* cytochrome P450 (CYP450) homologs of NMCH are next considered to extend this pathway to reticuline (Fig. 3). Currently all yeast alkaloid production studies utilize characterized *Eschscholzia californica* NMCH (EcNMCH) for conversion of NMC to 3-hydroxy-*N*-methylcoclaurine (3HNMC) within the conversion of norcoclaurine to reticuline[1,15,26]. There are several promising *P. somniferum* CYP450 sequences annotated as NMCH based on gene expression analyses in plants[27]; however these reports lack clear *in vivo* or *in vitro* characterization of the exact *P. somniferum* NMCH (PsNMCH) sequence[28]. Furthermore, the presence of many additional CYP450 homologs in the *P. somniferum* genome further complicates the selection of the best candidate sequence. To clarify the selection of this important missing link, a SVM model was trained using plant CYP80B sequences annotated as "N-methylcoclaurine hydroxylase" as positive examples. 100 *P. somniferum* CYP450 sequences were then tested against this model to assist the selection of an optimal candidate (Fig. 3b, Supplementary Table 3). As a result, PsNMCH Isoform 1 (PsNMCH-I1) scored high against the model and was selected.

A CPR redox partner for PsNMCH was also selected using the same workflow. While a CPR sequence has been characterized from *P. somniferum*[29], the referenced sequence AAC05021.1 is annotated as "NADPH:ferrihemoprotein oxidoreductase", confusing the rapid database selection of this sequence as CPR. Moreover there are at least 8 other unique *P. somniferum* sequences with high CPR homology that have not been characterized. After testing the 8 additional *P. somniferum* candidates against the CPR SVM model, XP_026404029.1 is selected as a high scoring sequence (Fig 3c and Supplementary Table 4), and observed to exhibit CPR activity (Fig. 3d). This new CPR sequence is annotated as "NADPH--cytochrome P450 reductase-like", and accordingly it is referred to as PsCPR-L in this manuscript.
NMCH activity is evaluated by converting norcoclaurine to stable reticuline using NMCH and CPR variants expressed together with norcoclaurine 6-*O*-methyltransferase (6OMT), coclaurine *N*-methyltransferase (CNMT) and 3-hydroxy-*N*-methylcoclaurine 4-*O*-methyltransferase (4OMT) (Fig. 3d and Table 1). NMC accumulates much more than other intermediates in this system, and therefore reticuline titers should reflect the activity of the NMCH bottleneck. In this system, PsNMCH-I1 affords higher amounts of reticuline than that of EcNMCH, when paired with either PsCPR-L or AtATR2 (Fig. 3d). PsNMCH-I1 pairs best with PsCPR-L from the same species, resulting in the highest amount of reticuline. On the other hand, reticuline production with EcNMCH is best with AtATR2 pairing, with no improvement from PsCPR-L pairing.

Just one residue difference is observed when comparing the binding pockets of PsNMCH and EcNMCH: PsNMCH-H203 versus EcNMCH-Y202 (Fig. 3a). SVM prediction of PsNMCH-H203Y

and EcNMCH-Y202H sequences results in lower and higher SVM scores, respectively (Supplementary Table 3 and Fig 3c), indicating that the SVM model is able to identify this key residue as an important feature. Consistent with this prediction, transplantation of EcNMCH-Y202 into engineered PsNMCH-H203Y results in lower reticuline, and transplantation of PsNMCH-H203 into engineered EcNMCH-Y202H results in higher conversion of norcoclaurine to reticuline when paired with AtATR2. The improvement in reticuline with EcNMCH-Y202H could be replicated in a second independent test of the same strains (data not shown).

Early *in vivo* tests of PsNMCH-I1 without a CPR redox partner in *E. coli* did not result in detectable NMCH activity, but L-DOPA production from tyrosine was detected (Supplementary Fig. 2). This led us to hypothesize that PsNMCH-I1 might also have potential tyrosine 3-monoxygenase activity; however the observed L-DOPA production is probably more likely to be mediated by native *E. coli* HpaBC. To further clarify this important missing link in *P. somniferum*, the candidate CYP450 monooxygenase sequences are also explored as potential tyrosine 3-monooxygenase templates (Supplementary Table 5). Here, the candidate sequences are tested against a plant CYP76AD SVM model and a combined SVM model trained with plant CYP76AD, CYP98A3 and CYP199A2 sequences that hydroxylate tyrosine and structurally similar compound coumaric acid[30-32]. CYP98A2-like (XP_026403623.1), geraniol 8-hydroxylase-like (XP_026409442.1) and flavonoid 3',5'-hydroxylase 1-like (XP_026378021.1) sequences appear as prime targets with relatively high scores in the positive prediction space of both high-dimensional models of Supplementary Table 5. Completion of the discovery and testing of the tyrosine 3-monooxygenase candidate sequences, which are toxic and difficult to clone, is currently being addressed in a parallel study.

Expansion of PsTyDC1 and PsPDC1 routes into dual pathways

Docking of L-DOPA to PsTyDC1 (Fig. 2a) shows stronger binding affinity than that of tyrosine, suggesting that PsTyDC1 may also possess DHPAAS activity. Accordingly, co-expression of PsTyDC1 with PsNMCH-I1, 6OMT, CNMT and 4OMT, results in a plant-gene only dual pathway through 4HPAA and DHPAA to norcoclaurine and reticuline (Supplementary Fig. 2). Discovery of DHPAAS activity mediated by PsTyDC1 is confirmed in other strains, and also with isotope tracing from L-DOPA-d_3 to downstream deuterium labeled BIA (Table 1). Based on these findings, the potential DHPAAS activity of PsTyDC1 is further explored to combine norcoclaurine and THP pathways (Fig. 4a).

After incorporating L-DOPA decarboxylase (DDC) from *Pseudomonas putida* (PpDDC) for *in vivo* dopamine production and optimization in Terrific Broth (TB), PsPDC1 and PsTyDC1 containing strains produce reticuline from L-DOPA via the DHPAA containing pathway, with titers reaching the µM range (Fig. 4b). This result indicates that PsPDC1 can also produce DHPAA from 3,4-dihydroxyphenylpyruvic acid (DHPP) that is supplied by L-DOPA transamination. Previously, a single strain containing DHPAAS, 6OMT, CNMT and 4OMT only produced

reticuline titers of 0.2 µM from L-DOPA[7]. Under optimized conditions, PsPDC1 performs better than PsTyDC1 up to 44 hours. However, 61 hours after addition of L-DOPA substrate, PsPDC1-mediated reticuline titers decline slightly, likely due to oxidative degradation, and PsTyDC1 mediates a similar titer to that of PsPDC1. PsPDC1 works synergistically with PsTyDC1 at later production times to maintain higher reticuline. Accordingly, combinations of PPDC and AAS are next explored to improve BIA titers.

Switching missing link templates to homologous microbial sequences for improved production

Natural plant enzymes PsTyDC1 and PsPDC1 exhibit desired specificities, but their expression and activity in *E. coli* is sensitive, contributing to limited titers. Therefore, to better apply missing link functions, analogous microbial sequences are explored as enzyme engineering templates (Fig. 5).

AAS activity analogous to that of PsTyDC1 could be engineered into the bacterial PpDDC template by transplanting *Bombyx mori* DHPAAS (BmDHPAAS) catalytic residues F79, Y80 and N181. Rationally engineered PpDDC-Y79F-F80Y-H181N mediates improved THP production in *E. coli* (Fig. 5a). Switching from PsPDC1 to a *Saccharomyces cerevisiae* ARO10 template confers improved *in vivo* PPDC activity towards both DHPP (Fig. 5b) and 4HPP (Fig. 6), in comparison to corresponding strains containing PsPDC1. However, the high activity of ARO10 may come at a specificity tradeoff, as production of additional aromatic keto acid derived alkaloids result from ARO10 expression (Fig 5d).

Combinations of natural and analogous enzyme templates result in improved *E. coli* BIA production (Fig. 6a and Table 1). Expression of PpDDC- Y79F-F80Y-H181N together with PsPDC1 in strain P1-07-AI selectively promotes the DHPAA pathway in the presence of tyrosine and L-DOPA to produce 20.3 mg/L reticuline, while the application of ARO10 in strain A1-01-DE3 selectively favors the 4HPAA pathway in the presence of tyrosine and dopamine to produce 96.7 mg/L norcoclaurine and 71.8 mg/L NMC. A dual pathway from tyrosine and dopamine to 33.6 mg/L NMC and 24.6 mg/L reticuline is promoted through the combination of PpDDC-Y79F-F80Y-H181N, ARO10 and PsTyDC1 in strain A1-06-AI.

Dynamic metabolomic profiling of AAS and PPDC branch pathways

Isotope tracing enables analysis of flux through bioproduction pathways[5,33,34]. While multiple reaction monitoring (MRM) with LC-MS is sensitive, this method does not readily detect isotope-labeled intermediates. After improving BIA titers to µM levels suitable for quantification with high-resolution CE-MS, isotope tracing experiments could be performed. Combinations of PsPDC1, ARO10, PsTyDC1 and PpDDC produce various labeling patterns: tyrosine-^{13}C to BIA-$^{13}C_2$, L-DOPA-d_3 with tyrosine-d_4 to d_6-labeled BIA, L-DOPA-d_3 to

d_5-labeled BIA, L-DOPA-d_3 with dopamine-d_2 to d_5-labeled BIA, tyrosine-d_4 with dopamine-d_2 to d_6-labeled BIA, and tyrosine-d_4 with dopamine to d_4-labeled BIA (Fig. 6 and Supplementary Fig. 4). The loss of a ring deuterium atom during NCS-mediated condensation of aryl acetaldehydes with ring-labeled dopamine is consistent with the reported NCS mechanism (Fig. 6b and Supplementary Fig. 4d)[35,36]: this kind of mechanism-directed deuterium labeling pattern has not been reported for the tracing of BIA[37-39]. Isotope tracing from L-DOPA-d_3 to d_5-labeled BIA supports the bifunctional decarboxylase and oxidative deamination activities of PpDDC-Y79F-F80Y-H181N (Supplementary Fig. 4d). Improvement of NMC-d_6 and reticuline-d_5 production via PsTyDC1 in addition to PsPDC1 again demonstrates the synergistic combination of these distinct aryl acetaldehyde producing enzymes (Fig. 6b). Moreover, amounts of NMC-d_6 and reticuline-d_5 relative to their respective precursors norcoclaurine-d_6 and THP-d_5 (Supplementary Fig. 4, b and c) show the bottleneck of the *S*-adenosylmethionine (SAM)-dependent methylation of deuterium-labeled BIA. Furthermore, isotope tracing from tyrosine-^{13}C supports that PsPDC1 and ARO10 are converting isotope labeled 4-hydroxyphenylpyruvate (4HPP) to downstream BIA (Supplementary Fig. 4a).

Monitoring time course turnovers of isotope labeled metabolite fractions enables direct observations of metabolic flux[5,33,34]. Mixed fractions of unlabeled and labeled BIA, could be clearly quantified in the case of high-titer production of d_4-labeled norcoclaurine and NMC (Fig. 6c). In this case, a higher fraction of d_4-labeled norcoclaurine relative to d_4-labeled NMC is consistent with the SAM-dependent methyltransferase bottleneck observed previously[1,7].

Discussion

This report demonstrates that machine learning can uncover missing link enzymes with direct applications to biomanufacturing. While previous studies have also reported machine learning for enzyme prediction, these examples were never applied to the discovery of uncharacterized enzymes[40-43]. In the report by Li *et al.*, prediction of active glutaminase and aurora kinases B were used as examples to verify their algorithm, however this test data was obtained from previous publications[42]. The current study not only demonstrates the prediction and verification of 4 distinct plant enzymes, but also the possibility to engineer artificial enzymes as demonstrated by prediction of PsNMCH-H203Y and EcNMCH-Y202H with scores in agreement with *in vivo* test results (Fig. 3). Moreover, all of the prediction results show the value of machine learning prediction when added on top of structure and homology analyses, especially for inferring the function of enzymes of which many homologs are present in target species. PsPDC1 prediction demonstrates SVM prediction of enzymes that could not be easily distinguished using only sequence homology comparison and structural analysis. The PPDC prediction models (Supplementary Table 2) open the door to the machine learning-enabled discovery of additional high-scoring plant PPDC sequences.

PsPDC1 exhibits PPDC activity and contains active site residue Y332, which is also present in ZmPDC that is known to only convert small non-aromatic substrates. This active site tyrosine is substituted with smaller residues in characterized yeast and bacterial PPDC enzymes (Fig. 2d), and therefore the structural basis of plant PPDC substrate recognition appears to be unique. Species-by-species variation in functional residues is also seen with the evolution of AAS variants throughout plants and insects. In plants, tyrosine is commonly substituted with a more hydrophobic phenylalanine (residue 346 of *Petroselinum crispum* 4HPAAS, Fig. 2a) to switch AAAD to AAS activity. *In vivo* tests of PsTyDC suggest that hydrophobic PsTyDC1-L205 and PsTyDC3-I370 offer choices at two additional positions with potential to lower AAAD activity while promoting AAS activity. In contrast, insects have evolved a histidine to asparagine active site switch, corresponding to residue 192 of BmDHPAAS, to promote AAS activity essential for their survival[7,12]. These points illustrate that plant PPDC and AAS have evolved independent mechanisms to promote their unique characteristics that are distinct from typical PDC and AAAD functions, respectively. Furthermore, these insights establish additional structure-based rules for selection and engineering of plant carboxy-lyases (EC 4.1.1.X) and CYP450 monooxygenases (EC 1.14.14).

Transplantation of discovered functional residues into high-activity microbial templates is an effective strategy for improving bioproduction, as demonstrated by PpDDC-Y79F-F80Y-H181N with transplanted BmDHPAAS active site residues. Improved protein stability, removal of regulation/inhibition, and higher bacterial expression are additional factors that might contribute to improved templates. Transplantation of PsTyDC1-specific L205 into PpDDC also results in detected AAS activity, however in this case BIA production from tyrosine and dopamine does not improve (Table 1). On the other hand, the PpDDC-H181L variant and natural PsTyDC1 may exert higher AAS specificity than that of PpDDC-Y79F-F80Y-H181N (Fig. 5c). Both PsPDC1 and ARO10 were able to convert 4HPP to 4HPAA and DHPP to DHPAA, but PsPDC1 favors the DHPAA containing pathway (Fig. 6, a and b). Furthermore, PsTyDC1 is capable of mediating DHPAA and 4HPAA containing pathways, while PpDDC-Y79F-F80Y-H181N favors the DHPAA pathway. Therefore, the selection of flux through the norcoclaurine pathway versus the THP pathway may be controlled by the choice of ARO10 versus PpDDC-Y79F-F80Y-H181N or PsPDC1, respectively.

Application of natural plant enzymes PsTyDC1 and PsPDC1 results in a dual pathway to norcoclaurine and THP. The dual pathway offers advantages for utilization of tyrosine, and for improving amounts of unstable aryl acetaldehydes relative to dopamine. Accordingly, increased aryl acetaldehyde production by synergistic expression of PPDC together with AAS results in increased reticuline through an enhanced THP pathway (Fig. 6, a and b). Furthermore, the enabled isotopic labeling methods offer additional options to identify new bottleneck targets

for increasing metabolic flux to alkaloids. In conclusion, machine learning-based selection of enzyme sequences can directly improve microbial production pathways for valuable chemicals.

Materials And Methods

SVM machine learning prediction

Support vector machine (SVM) Enzyme-models were built based on the methods of our previous study[8] with modifications. Aromatic amino acid decarbloxylase (AAAD), aromatic acetaldehyde synthase (AAS, previously referred to as aromatic aldehyde synthase) and phenylpyruvate decarboxylase (PPDC) prediction models were trained with vectors generated by PROFEAT[18]. AAAD positive training sequences include DDC and other typical PLP-dependent carboxy-lyases that decarboxylate aromatic amino acids. The AAAD model is trained with positive examples based on only typical AAAD sequences that contain a catalytic histidine, corresponding to H181 of PpDDC (Fig. 2a). Characterized PsTyDC9[20] is included as a positive AAAD training example to ensure there is no bias towards AAS prediction. For AAS models, the positive training examples consist of sequences with homology to known plant-type and insect-type AAS enzymes, including *Petroselinum crispum* 4HPAAS (Pc4HPAAS) and BmDHPAAS. Insect-type AAS sequences are classified based on the presence of N192 (BmDHPAAS numbering), and plant-type AAS enzymes are classified based on the presence of F346 or V346 (Pc4HPAAS numbering). For PPDC prediction models, positive training vectors included sequences annotated as PPDC and indolepyruvate decarboxylase. Since all current database sequences annotated as phenylpyruvate decarbloxylase are from bacteria and fungi (plus 1 from Archaea), PDC sequences also had to be included in the first prediction model (Supplementary Table 2, upper table). After discovering PsPDC1, a rose PPDC sequence was found from continuous literature searches, although its protein accession (BAU70033.1) is annotated as "pyruvate decarboxylase"[44]. A second PPDC specific SVM model was therefore built by training with 19 homologous plant sequences in the same phylogenetic clade as rose PPDC.

CYP450 prediction models were trained with vectors generated by ProtVec[19]. To clarify potential NMCH activities of CYP450 monooxygenases, SVM models were trained with CYP80B positive sequences. To clarify potential tyrosine 3-monooxygenase activities, SVM models were trained with positive sequences related to CYP76AD, CYP98A3 and CYP199A2, which are reported to mediate aromatic hydroxylation of tyrosine as well as similarly sized substrate coumaric acid[30-32].

Prediction models were first built with high-dimensional vectors. Cross validation of all high-dimensional models resulted in F-scores above 0.96. Candidate sequences were selected based on high-dimensional scores. Two-dimensional and three-dimensional plots were used for visual representation of data in Figures. For two-dimensional plots, high-dimensional vectors were compressed to 2 dimensions using principal component analysis (PCA). 2-dimensional SVM models were then built derived from the PCA compressed vectors. SVM and PCA from the scikit-learn library were used[45]. The three-dimensional SVM plot in Fig 2e was adopted from an SVM illustration by Dr. Saptashwa Bhattacharyya (https://towardsdatascience.com/visualizing-support-vector-machine-decision-boundary-69e7591dacea). Compressed two-dimensional decision scores from the combined model (Supplementary Table 2, upper table) are used as the third dimension of Fig. 2e.

Training sequences, cross validation F-scores and additional parameters for high-dimensional models are available in the supplementary data file "M-Links SVM Training.xlsx".

Protein structural modeling and docking analysis

Homology models were built using Modeller as described in the previous study[7]. Multimeric structures and ligands were first prepared in Pymol. Structures were refined and prepared for docking analysis using Molecular Operating Environment as described previously[7].

Materials and reagents

KOD -Plus- and Ex-Taq HS DNA polymerases were purchased from Toyobo (Tokyo, Japan) and Takara (Tokyo, Japan), respectively. A DNA ligation kit and JM109 chemical competent cells were purchased from Takara (Tokyo, Japan). BL21-AI competent cells were purchased from Thermo Fisher Scientific (Waltham, MA, USA). All restriction endonucleases were purchased from New England Biolabs (Ipswich, MA, USA). Antibiotics were purchased from Nacalai Tesque (Kyoto, Japan), Sigma-Aldrich (St. Louis, MO, USA) and FUJIFILM Wako Pure Chemical (Osaka, Japan). Growth medium components were purchased from BD (Franklin Lakes, NJ, USA) and Nacalai Tesque. 3-(3,4-dihydroxyphenyl)-L-alanine (L-DOPA), and 3-hydroxytyramine hydrochloride (dopamine) were purchased from TCI (Tokyo, Japan). 4-hydroxyphenylpyruvic acid was from Sigma-Aldrich. L-Tyrosine and L-ascorbic acid sodium salt were obtained from Nacalai Tesque. Analytical standards and isotopes were purchased from Santa Cruz Biotechnology (Dallas, TX, USA), Toronto Research Chemicals (New York, ON, Canada), ALB Technology (Kuala Lumpur, Malaysia), Sigma-Aldrich and Cambridge Isotope Laboratories (Tewksbury, MA, USA).

Preparation of plasmids

Constructed plasmids (Supplementary Table 6) were transformed into JM109 chemically competent *E. coli* (Takara). Transformants were grown on LB-agar plates supplemented with the appropriate antibiotics at 30-37°C. Positive clones were screened using colony PCR and target plasmids were purified using a QIAprep Miniprep Kit (Qiagen). Plasmids were then sequenced using primers listed in Supplementary Table 7, a BigDye Terminator v3.1 cycle-sequencing kit, and an Applied Biosystems 3500xL Genetic Analyzer (Foster City, CA, USA).

Preparation of candidate missing link genes

Full length *P. somniferum TyDC1* and *TyDC3* native coding sequences were synthesized by Integrated DNA Technologies (IDT). Codon optimization of *PsONCS3* and *TfNCS* nucleotide sequences[46] for expression in *E. coli* was assisted by Codon Optimization OnLine (COOL)[47], resulting in the coding sequences shown in Supplementary Table 8, and the selected sequences were synthesized by IDT. The native sequence of full length *P. somniferum NMCH isoform 1 (PsNMCH-I1)* was also synthesized by IDT.

Native coding sequences of full length *PsPDC1*, full length *Ps2HCLL*, and N-terminal truncated *PsPDC2* were synthesized and cloned into pBAD-DEST49 (LifeSensors Inc.) via the Gateway cloning system by GeneArt (Invitrogen). Native coding sequences of full length *EcNMCH, AtATR2*, and *P. somniferum CPR-like (PsCPR-L)* were synthesized and subcloned into pMA vector (Invitrogen) by GeneArt (Invitrogen).

Construction of pACYC-3CjMTs-DDC vectors

The pACYC184-derived vectors containing *Coptis japonica 4OMT*, *CNMT*, *6OMT* (pACYC184-Cj4OMT-CjCNMT-Cj6OMT), and *PpDDC* (pACYC184-Cj4OMT-CjCNMT-PpDDC-Cj6OMT) were constructed in previous reports[13,14]. Active site mutations were introduced into *PpDDC* in pACYC184, by way of site-directed mutagenesis using PCR with primers shown in Supplementary Table 7.

Construction of subcloning vectors and mutations

To construct subcloning vectors for synthetic genes (*PsONCS3*, *TfNCS*, *PsTyDC1*, *PsTyDC3*, *CjNCS*, *PsNMCH-I1*, *EcNMCH, PsCPR-L*, *AtATR2*), and PCR amplified *PpDDC* and *ARO10* (amplified from pGK424-ARO10[48]), 3’ end A-protrusions were added to each DNA fragment using A-attachment Mix (Toyobo).

Gene mutations were generated using site directed mutagenesis by PCR with primers listed in Supplementary Table 7. *TyDC1* mutations (L205H and Y98F-F99Y-H205N) and *TyDC3* mutations (I370S and Y100F-F101Y-H203N) were generated in subcloning vectors by PCR. *PpDDC* mutations (H181L, H181L-G344S, Y79F-F80Y-H181N, Y79F-F80Y-H181N-G344S) were generated by PCR. *EcNMCH* mutation (Y202H) and *PsNMCH-I1* mutation (Y203H) were generated in subcloning vectors by PCR.

Construction of alkaloid production vectors

A *PsONCS3*[46] containing DNA fragment was obtained from NcoI and BamHI digestion of the *PsONCS3* subcloning vector, and then cloned into pCDFDuet-1 via the NcoI and BamHI restriction sites to produce pCDFD-PsONCS3. A *TfNCS* containing DNA fragment was obtained from NcoI and BamHI digestion of the TfNCS subcloning vector, and then cloned into pCDFDuet-1 via the NcoI and BamHI restriction sites to produce pCDFD-TfNCS.

DNA fragments of *TyDC1* were obtained from NdeI and XhoI digestion of *TyDC1* subcloning vectors, and then cloned into pCDFDuet-1-PsONCS3 via NdeI and XhoI restriction sites to produce pCDFD-PsONCS3-TyDC1. The *TyDC1* containing gene fragments were also cloned into pCDFDuet-1-TfNCS via NdeI and XhoI sites to produce pCDFD-TfNCS-TyDC1. DNA fragments of *TyDC3* were obtained from NdeI and XhoI digestion of *TyDC3* subcloning vectors, and then cloned into pCDFDuet-1-PsONCS3 via NdeI and XhoI restriction sites to produce pCDFD-PsONCS3-TyDC3. The *TyDC3* containing gene fragments were also coned into pCDFDuet-1-TfNCS via NdeI and XhoI sites to produce pCDFD-TfNCS-TyDC3.

EcNMCH and *EcNMCH-Y202H* gene fragments were digested with SalI and NotI in subcloning vectors and then cloned into pCOLADuet-1 via the SalI and NotI restriction sites. *AtATR2 and PsCPR-L* fragments were next digested from the subcloning vectors using NdeI and XhoI, and then cloned into pCOLAD-EcNMCH and pCOLAD-EcNMCH-Y202H via the NdeI and XhoI restriction sites to produce pCOLAD-EcNMCH-AtATR2, pCOLAD-EcNMCH-Y202H-AtATR2, pCOLAD-EcNMCH-PsCPR and pCOLAD-EcNMCH-Y202H-PsCPR.

The DNA fragment encoding *PsNMCH-I1* with a truncated N-terminal, was digested by NotI and XhoI from the subcloning vector and then cloned into a pACYC184 derived vector containing *C. japonica 4OMT*, *CNMT*, and *6OMT* via Not I and Xho I restriction sites to produce pACYC-3CjMTs-PsNMCH. Truncated *PsNMCH-I1* and truncated *PsNMCH-Y203H* gene fragments were PCR amplified from subcloning vectors and then cloned into pCOLAD-EcNMCH-PsCPR digested with BamHI and NotI via Gibson assembly (NEB)[49] to produce pCOLAD-PsNMCH-PsCPR and pCOLAD-PsNMCH-Y203H-PsCPR. Truncated *PsNMCH-I1* and truncated *PsNMCH-Y203H* gene fragments were also digested with NcoI and NotI and cloned into pCOLAD-EcNMCH-AtATR2 digested with NcoI and NotI to produce pCOLAD-PsNMCH-AtATR2 and pCOLAD-PsNMCH-Y203H-AtATR2.

DNA fragments of *PpDDC-H181L*, *PpDDC-H181L-G344S*, *PpDDC-Y79F-F80Y-H181N* and *PpDDC-Y79F-F80Y-H181N-G344S* were PCR amplified from subcloning vectors and then cloned into pCDFDuet-1 digested with NcoI and BamHI via Gibson assembly (NEB). A *CjNCS* DNA fragment was obtained from NdeI and XhoI digestion of the *CjNCS* subcloning vector, and then cloned into pCDFDuet-1-PpDDC vectors via NdeI and XhoI sites to produce pCDFD-CjNCS-DDC. A *S. cerevisiae ARO10* gene fragment was digested with NcoI and NotI in the *ARO10* subcloning vector and then cloned into pCDFDuet-1 via NcoI and NotI restriction sites. A

CjNCS gene fragment was next digested from the subcloning vector using NdeI and XhoI, and then cloned into pCDFDuet-1-ARO10 via the NdeI and XhoI restriction sites to produce pCDFD-CjNCS-ARO10. *E. coli HpaBC* containing gene fragments were PCR amplified from *E. coli* using the Gibson assembly primers shown in Supplementary Table 7. The PCR product was cleaned using a conventional column-based kit, and then cloned into XhoI-digested pET23 via Gibson assembly (NEB) (pET23-HpaBC).

In vivo production of BIA

BL21(DE3) and BL21-AI competent *E. coli* cells were transformed with various combinations of plasmids from Supplementary Table 6, resulting in the strains shown in Table 1. Strains were tested in M9, LB or TB, supplemented with various substrates according to Table 1. Expression of recombinant genes in expression vectors containing the T7 promoter system was induced by addition of 0.5 - 1.5 mM IPTG (isopropyl β-D-1-thiogalactopyranoside) to BL21(DE3) cultures. When using BL21-AI cells 0.08 - 0.4% arabinose was included. Expression of PsPDC1, PsPDC2 and Ps2HCLL in pBAD-DEST49 was also induced by addition of 0.08 - 0.4% arabinose.

For reported titers, A1-01-DE3 (3 CjMTs, PsNMCH, CjNCS and ARO10), P1-02-AI (3CjMTs, PpDDC, PsONCS3 and PsPDC1), P1-04-AI (3CjMTs, PpDDC, PsONCS3, PsTyDC1 and PsPDC1), P1-06-DE3 (3CjMTs, PpDDC-Y79F-F80Y-H181N, PsONCS3 and PsPDC1), P1-07-AI (3CjMTs, PpDDC-Y79F-F80Y-H181N, PsONCS3 and PsPDC1), A1-06-AI (3 CjMTs, PpDDC-Y79F-F80Y-H181N, CjNCS, ARO10 and PsTyDC1) and T1-10-DE3 (3CjMTs, PpDDC, PsONCS3 and TyDC1) (Fig. 4 - 6, Table 1, and Supplementary Fig. 4, b-d) were grown using 3.5 mL teriffic broth (TB) supplemented with sodium ascorbate and appropriate antibiotics, in plastic culture tubes at 34-37°C with shaking at 180-190 rpm. After reaching late log phase, inducing agent (IPTG or arabinose) and substrates (>8 mM tyrosine, >8 mM L-DOPA, >9 mM tyrosine-^{13}C, >3 mM tyrosine-d_4, >11 mM L-DOPA-d_3, >7 mM dopamine-d_2) were added. When tyrosine was used as a substrate, sometimes dopamine was included as indicated in Table 1 (>3.5 mM dopamine, >7 mM dopamine-*d*2). Cultures were then incubated at 25°C with shaking at 180-200 rpm.

DT-01-DE3 (3CjMTs, PpDDC-Y79F-F80Y-H181N and TfNCS), DS-02-DE3 (3CjMTs, PsNMCH, CjNCS and PpDDC-H181L), DD-01-DE3 (3CjMTs, PsNMCH, CjNCS and PpDDC-H181L-G344S), DQ-01-DE3 (3CjMTs, PsNMCH, CjNCS and PpDDC-Y79F-F80Y-H181N-G344S), DT-02-DE3 (3CjMTs, PpDDC-Y79F-F80Y-H181N and PsONCS3), DT-03-DE3 (3CjMTs, PsNMCH, CjNCS and PpDDC-Y79F-F80Y-H181N) and A1-03-DE3 (3CjMTs, PpDDC, CjNCS, ARO10 and EcHpaBC) (Table 1, Fig. 5a, and Supplementary Fig. 4, a and d) were tested in 3 - 4.8 mL M9 supplemented with ascorbate and appropriate antibiotics. After reaching log phase in plastic culture tubes at 36-37°C, IPTG and substrates (>4.5 mM tyrosine, >2 mM L-DOPA, >5 mM tyrosine-^{13}C, >4 mM L-DOPA-d_3) were added. When tyrosine was used as a substrate, sometimes dopamine was included as indicated in Table 1 (>3.5 mM dopamine). Cultures were then incubated at 20-25°C with shaking at 180 rpm. Additional ascorbate was added as needed to prevent oxidative degradation of target compounds and melanization.

Conversion of norcoclaurine to reticuline was mediated by NMCH and CPR containing strains N1-01-DE3, N1-02-DE3, N1-03-DE3, N1-04-DE3, N2-01-DE3, N2-02-DE3, N2-03-DE3 and N2-04-DE3 (Table 1). Here, strains first grown in LB medium were used to inoculate TB medium to a starting OD_{600} of 0.02 in 3 mL, with appropriate antibiotics. After four hours at 37°C with shaking at 200 rpm, recombinant protein expression was induced with 0.68 mM IPTG and the temperature was lowered to 20°C. After 5.5 hours, cells were spun down and re-suspended in 1.5 mL TB supplemented with 1.2 mM norcoclaurine, 5.1 mM sodium ascorbate and 0.2 mM IPTG. After 1.5 days at 25°C with shaking at 200 rpm, BIA titers were measured with LC-MS.

Additional bioproduction conditions are given in the legends of Fig. 2c, Fig. 2f, Fig. 6a, as well as Supplementary Figures 1, 2 and 4. Bioproduction times are defined based on the addition of substrate.

Quantitative analysis of BIA pathway intermediates with LC-MS, CE-MS and GC-MS

Culture medium was filtered with Amicon Ultra 0.5 mL centrifugal filters with a molecular weight cut-off of 3,000 Da (Millipore). Filtrates were kept on ice and immediately processed for analysis, or stored at -30^{o}C or -80^{o}C before use.

For LC-MS analysis with multiple-reaction monitoring (MRM), filtered culture medium was diluted in a solution of camphor sulfonic acid, and then loaded onto a Shimadzu LCMS-8050 system according to the methods described in Vavricka *et al.*[7]. Over 100 metabolites could be monitored with MRM detection.

For CE-MS analysis, filtered samples were diluted in a methionine sulfone solution when using positive ion mode, or in a PIPES solution for negative ion mode. CE-MS analysis was performed according to methods of Hasunuma *et al.*[33,34]. Quantification of isotopes in Fig. 6 and Supplementary Fig. 4 was based on standard curves of non-labeled compounds. CE-MS peak areas in relation to internal standard peak areas were used to quantify all compounds except for 4HPP (Supplementary Fig. 4a), which was quantified based on its own peak intensity.

For GC–MS analysis, the filtered samples were dried under vacuum and then derivatized with BSTFA (Wako) and TMS-Cl (Alfa Aesar) before analysis. Derivatized aromatic compounds were analyzed on a GCMS-QP2010 Plus (Shimadzu) with a DB-5 capillary column (Agilent).

Extraction of aromatic compounds for GC-MS quantification

A solution of ammonium carbonate was added to culture samples, followed by addition of EtOAc. After vortexing, the organic layer was removed and evaporated under vacuum. The dried extracts were then derivatized in a mixture of BSTFA (Wako), TMS-Cl (Alfa Aesar) and EtOAc. Quantitative standard curves were produced by extracting alkaloid standards from equivalent TB solutions, followed by TMS-derivatization in equivalent volumes. The TMS-derivatized samples were analyzed with GC-MS.

Declarations

Funding: The authors are grateful for support from NEDO projects P16009 (Development of production techniques for highly functional biomaterials using plant and other organism smart cells) and P20011 (Development of bio-derived product production technology that accelerates the realization of carbon recycling). CJV was further supported by Kato Memorial Bioscience Foundation (2017M-014) and JSPS KAKENHI (V18K065770) while working on this study.

Competing interests: The authors declare no competing interests.

Data and materials availability: All necessary data and materials will be made available upon acceptance of the manuscript.

Code availability: Computer code for SVM-based enzyme prediction will be made available upon acceptance of the manuscript.

Author contributions: CJV, MA, AK and TH designed the research strategies and concepts. CJV performed enzyme engineering and bioproduction experiments, analyzed all data, and wrote

the manuscript. MA and NW developed the machine learning methods. ST and TK managed DNA construction. CJV, TH, MM and TY managed metabolomics analyses. HK, JI and HM provided other specialized technical assistance.
This work is dedicated to Natalie Chanier.

References

. Pyne, M. E. et al. A yeast platform for high-level synthesis of tetrahydroisoquinoline alkaloids. *Nat. Commun.* **11**, 3337 (2020).

2. Luo, X. et al. Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature* **567**,123-126. (2019)

3. Srinivasan, P. & Smolke, C.D. Biosynthesis of medicinal tropane alkaloids in yeast. *Nature* **585**, 614-619 (2020).

4. Gu, Y. & Xu, P. Synthetic yeast brews neuroactive compounds. *Nat. Chem. Biol.* **17**, 8-9 (2021)

5. Vavricka, C. J., Hasunuma, T. & Kondo, A. Dynamic metabolomics for engineering biology: accelerating learning cycles for bioproduction. *Trends Biotechnol.* **38**, 68-82 (2020).

6. Nett, R.S., Lau, W. & Sattely, E.S. Discovery and engineering of colchicine alkaloid biosynthesis. *Nature* **584**, 148-153 (2020)

7. Vavricka, C. J. et al. Mechanism-based tuning of insect 3,4-dihydroxyphenylacetaldehyde synthase for synthetic bioproduction of benzylisoquinoline alkaloids. *Nat. Commun.* **10**, 2015 (2019).

8. Watanabe, N. et al. Exploration and evaluation of machine learning-based models for predicting enzymatic reactions. *J. Chem. Inf. Model.* **60**, 1833-1843 (2020).

9. He, C. L. et al. Identification of bis-benzylisoquinoline alkaloids as SARS-CoV-2 entry inhibitors from a library of natural products. *Sig. Transduct. Target Ther.* **6,** 131 (2021).

0. Bai, G. et al. Identification of higenamine in Radix Aconiti *Lateralis Preparata* as a beta2-adrenergic receptor agonist1. *Acta Pharmacol. Sin.* **29**, 1187-1194 (2008).

. Cohen, P. A., Travis, J. C., Keizers, P. H. J., Boyer, F. E. & Venhuis, B. J. The stimulant higenamine in weight loss and sports supplements. *Clin. Toxicol.* **57**, 125-130 (2019).

2. Vavricka, C. et al. From L-dopa to dihydroxyphenylacetaldehyde: a toxic biochemical pathway plays a vital physiological function in insects. *PLoS One* **6**, e16124 (2011).

3. Matsumura, E. et al. Microbial production of novel sulphated alkaloids for drug discovery. *Sci. Rep.* **8**, 7980 (2018).

4. Matsumura, E. et al. Laboratory-scale production of (*S*)-reticuline, an important intermediate of benzylisoquinoline alkaloids, using a bacterial-based method. *Biosci. Biotechnol. Biochem.* **81**, 396-402 (2017).

5. Galanie, S., Thodey, K., Trenchard, I. J., Filsinger Interrante, M. & Smolke, C. D. Complete biosynthesis of opioids in yeast. *Science* **349**, 1095-1100 (2015).
6. Grewal, P.S., Samson, J.A., Baker, J.J. Choi, B. & Dueber, J.E. Peroxisome compartmentalization of a toxic enzyme improves alkaloid production. *Nat. Chem. Biol.* **17,** 96–103 (2021).
7. Labanca, F., Ovesna, J. & Milella, L. *Papaver somniferum* L. taxonomy, uses and new insight in poppy alkaloid pathways. *Phytochem. Rev.* **17**, 853–871 (2018).
8. Li, Z. R. et al. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res*. **34**, W32-W37 (2006).
9. Asgari, E. & Mofrad, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLOS One* **10**, e0141287 (2015).
0. Torrens-Spence, M. P., Lazear, M., von Guggenberg, R., Ding, H. & Li. J. Investigation of a substrate-specifying residue within *Papaver somniferum* and *Catharanthus roseus* aromatic amino acid decarboxylases. *Phytochemistry* **106**, 37-43 (2014).
. Facchini, P. J. & De Luca, V. Expression in *Escherichia coli* and partial characterization of two tyrosine/dopa decarboxylases from opium poppy. *Phytochemistry* **38**, 1119-26 (1995).
2. Versées, W. et al. Molecular mechanism of allosteric substrate activation in a thiamine diphosphate-dependent decarboxylase. *J. Biol. Chem.* **282**, 35269-35278 (2007).
3. Berthold, C. L. et al. Structure of the branched-chain keto acid decarboxylase (KdcA) from *Lactococcus lactis* provides insights into the structural basis for the chemoselective and enantioselective carboligation reaction. *Acta Crystallogr. D Biol. Crystallogr.* **63**, 1217-1224 (2007).
4. Kneen, M. M. et al. Characterization of a thiamin diphosphate-dependent phenylpyruvate decarboxylase from *Saccharomyces cerevisiae*. *FEBS J.* **278**,1842-53 (2011).
5. Sun, S., Duggleby, R. G. & Schowen, R. L. Linkage of catalysis and regulation in enzyme action, Carbon isotope effects, solvent isotope effects, and proton inventories for the unregulated pyruvate decarboxylase of *Zymomonas mobilis*. *J. Am. Chem. Soc.* **117**, 7317-7322 (1995).
6. Pauli, H. H. & Kutchan, T. M. Molecular cloning and functional heterologous expression of two alleles encoding (S)-N-methylcoclaurine 3'-hydroxylase (CYP80B1), a new methyl jasmonate-inducible cytochrome P-450-dependent mono-oxygenase of benzylisoquinoline alkaloid biosynthesis. *Plant J.* **13**, 793-801 (1998).
7. Desgagné-Penix, I. & Facchini, P. J. Systematic silencing of benzylisoquinoline alkaloid biosynthetic genes reveals the major route to papaverine in opium poppy. *Plant J.* **72**, 331-344 (2012).
8. Frick, S., Kramell, R. & Kutchan, T. M. Metabolic engineering with a morphine biosynthetic P450 in opium poppy surpasses breeding. *Metab. Eng.* **9**, 169-176 (2007).

). Rosco, A., Pauli, H. H., Priesner, W. & Kutchan, T. M. Cloning and heterologous expression of NADPH-cytochrome P450 reductases from the Papaveraceae. *Arch. Biochem. Biophys.* **348**, 369-377 (1997)

). Sunnadeniya, R. et al. Tyrosine hydroxylation in betalain pigment biosynthesis is performed by cytochrome P450 enzymes in beets (*Beta vulgaris*). *PLoS One* **11**, e0149417 (2016).

. Nair, R. B. et al. Arabidopsis CYP98A3 mediating aromatic 3-hydroxylation. Developmental regulation of the gene, and expression in yeast. *Plant Physiol.* **130**, 210–220 (2002).

2. Furuya, T. Arai, Y. & Kino, K. Biotechnological production of caffeic acid by cacterial cytochrome P450 CYP199A2. *Appl. Environ. Microbiol.* **78**, 6087-6094 (2012).

3. Hasunuma, T. et al. Dynamic metabolic profiling of cyanobacterial glycogen biosynthesis under conditions of nitrate depletion. *J. Exp. Bot.* **64**, 2943-2954 (2013).

4. Hasunuma, T., Matsuda, M., Kato, Y., Vavricka, C. J. & Kondo, A. Temperature enhanced succinate production concurrent with increased central metabolism turnover in the cyanobacterium *Synechocystis* sp. PCC 6803. *Metab. Eng.* **48**, 109-120 (2018)

5. Bonamore, A., Barba, M., Botta, B., Boffi, A. & Macone, A. Norcoclaurine synthase: mechanism of an enantioselective Pictet-Spengler catalyzing enzyme. *Molecules* **15**, 2070-2078 (2010).

6. Lechner, H., Pressnitz, D. & Kroutil. W. Biocatalysts for the formation of three- to six-membered carbo- and heterocycles. *Biotechnol. Adv.* **33** 457–480 (2015).

7. Poeaknapo, C., Schmidt, J., Brandsch, M., Dräger, B. & Zenk, M. H. Endogenous formation of morphine in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 14091-14096 (2004).

8. Grobe, N. et al. Urinary excretion of morphine and biosynthetic precursors in mice. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 8147-8152 (2010).

9. Nakabayashi, R. et al. Metabolomics with ^{15}N Labeling for Characterizing Missing Monoterpene Indole Alkaloids in Plants. *Anal Chem.* **92**, 5670-5675 (2020).

0. Che, Y., Ju, Y., Xuan, P., Long, R. & Xing F. Identification of multi-functional enzyme with multi-label classifier. *PLoS One* **11**, 1–13 (2016).

. Dalkiran, A. et al. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics* **19**, 334 (2018).

2. Li, Y. et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics* **34**, 760–769 (2018).

3. Kulmanov, M. & Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).

4. Hirata, H. et al. Seasonal induction of alternative principal pathway for rose flower scent. *Sci. Rep.* 6, 1-9 (2016).

5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. & Thirion, B. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

6. Li, J., Lee, E. J., Chang, L. & Facchini, P. J. Genes encoding norcoclaurine synthase occur as tandem fusions in the Papaveraceae. *Sci. Rep.* **6**, 39256 (2016).
7. Chin, J. X., Chung, B. K. S. & Lee, D. Y. Codon Optimization OnLine (COOL): a web-based multi-objective optimization platform for synthetic gene design. *Bioinformatics* **30**, 2210-2212 (2014).
8. Kondo, T. et al. Genetic engineering to enhance the Ehrlich pathway and alter carbon flux for increased isobutanol production from glucose by *Saccharomyces cerevisiae*. *J. Biotechnol.* **159**, 32-37 (2012).
9. Gibson, D. G. et al. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343-345 (2009).

Figures

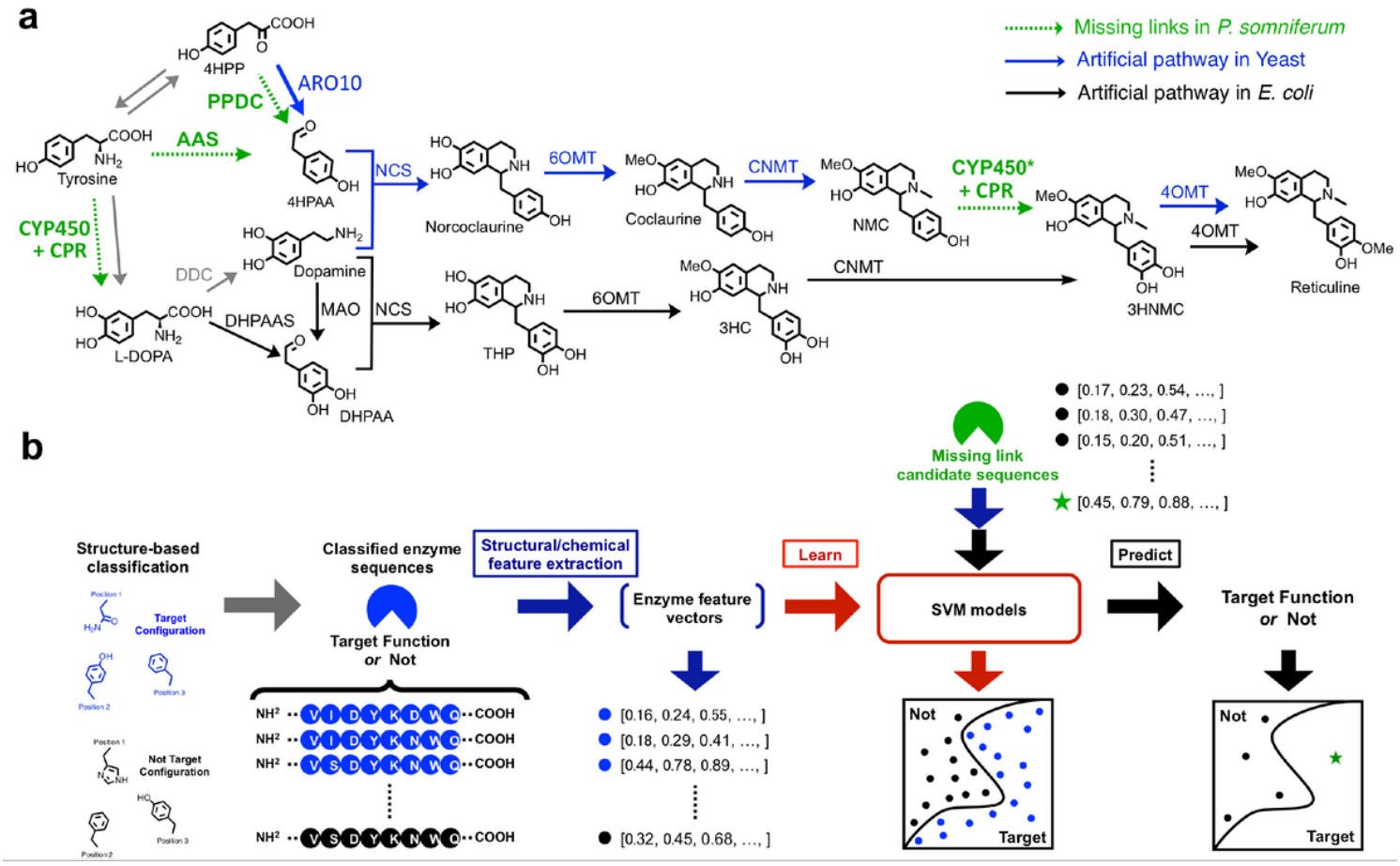

Figure 1

Uncovering missing links in Papaver somniferum as alternative branches to alkaloids. a, Reconstructed pathways to reticuline in yeast and E. coli are shown with blue or black arrows, respectively. Steps mediated by unclear P. somniferum enzymes are shown as green dotted arrows. Metabolite abbreviations: 4HPP - 4-hydroxyphenylpyruvic acid, 4HPAA - 4-hydroxyphenylacetaldehyde, L-DOPA - 3,4-

dihydroxy-L-phenylalanine, DHPAA - 3,4-dihydroxyphenylacetaldehyde, NMC - N-methylcoclaurine, THP - tetrahydropapaveroline, 3HC - 3-hydroxycoclaurine, 3HNMC 3-hydroxy-N-methylcoclaurine. Enzyme abbreviations: AAS - aromatic acetaldehyde synthase, DHPAAS - 3,4-dihydroxyphenylacetaldehyde synthase, PPDC - phenylpyruvate decarboxylase, DDC - L-DOPA decarboxylase, CYP450 - cytochrome P450, CPR - CYP450 reductase, ARO10 - Saccharomyces cerevisiae transaminated amino acid decarboxylase, NCS - norcoclaurine synthase, 6OMT - norcoclaurine 6-O-methyltransferase, CNMT - coclaurine N-methyltransferase, NMCH - N-methylcoclaurine 3-hydroxylase, 4OMT - 3-hydroxy-N-methylcoclaurine 4-O-methyltransferase. b, Missing link candidate sequences are predicted and ranked based on high-dimensional support vector machine (SVM) models8. Structure-based rules are first determined to classify training sequences as those of the target function and to differentiate related sequences with a different function. Structural and chemical features are then extracted from classified training and test sequences18,19, resulting in enzyme feature vectors. SVM models are built with the training vectors8, and vectors from missing link candidate sequences are then tested against the models to predict their novel function.

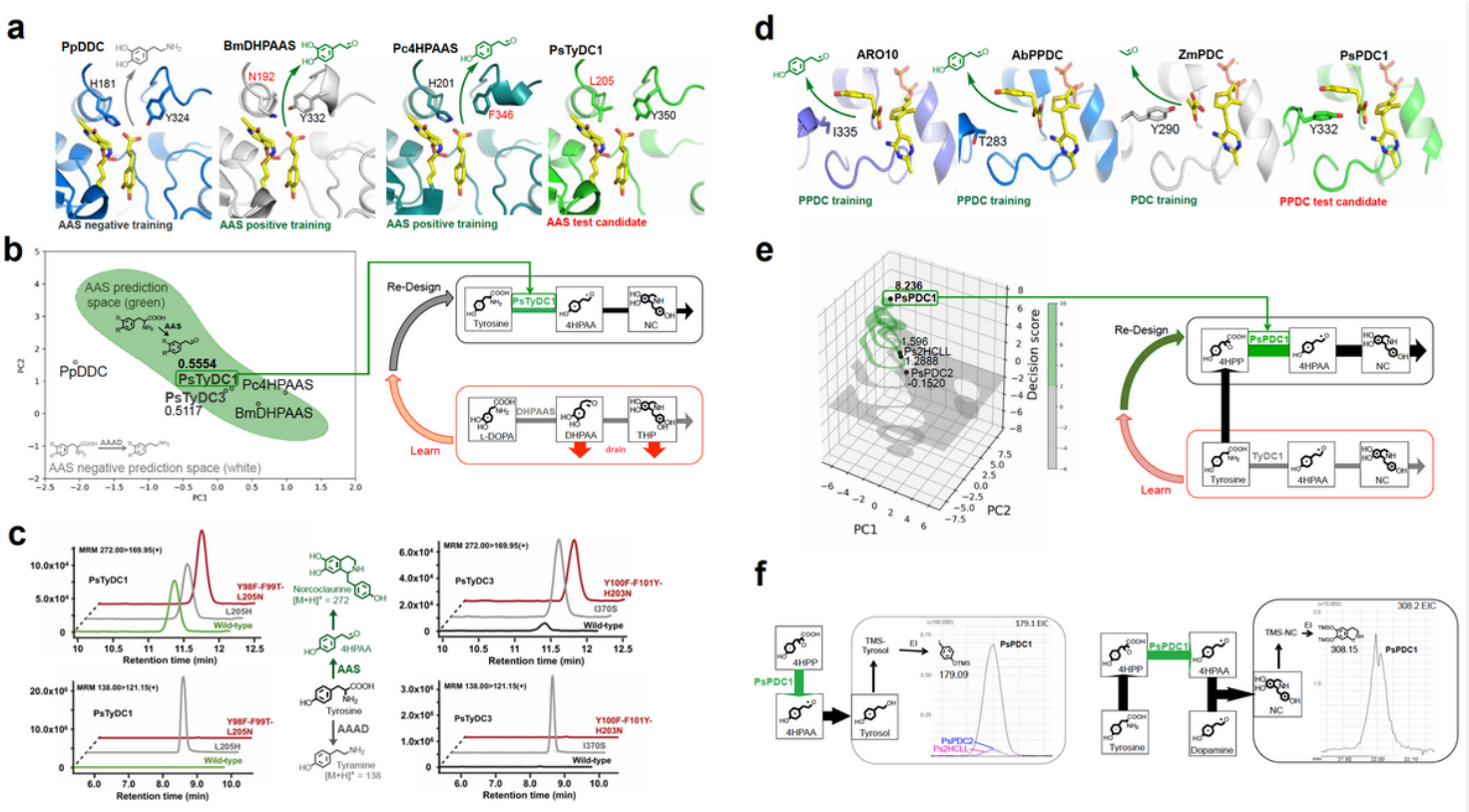

Figure 2

PsTyDC1 and PsPDC1 mediate alternative branches to 4HPAA for norcoclaurine production. a, Structural classification of typical AAAD, insect type AAS and plant type AAS based on the active site configurations of Pseudomonas putida DDC (PpDDC), Bombyx mori DHPAAS (BmDHPAAS) and Petroselinum crispum 4HPAAS (Pc4HPAAS). AAS candidate PsTyDC1 has a unique active site. In these models, PLP is included as an internal aldimine. Docking of L-DOPA into the PsTyDC1 active site results in lower predicted binding energy compared to that of L-tyrosine. b, For visual representation, a two-dimensional plot of AAS SVM-

based prediction is shown, with positive and negative prediction spaces colored green and white, respectively (b - left side). Principal component analysis (PCA) is used to compress multi-dimensional data into two dimensions (PC1 and PC2) for a visual representation. Corresponding high-dimensional SVM results are detailed in Supplementary Table 1, and high-dimensional SVM decision scores for tested sequences PsTyDC1 and PsTyDC3 are listed. Decision scores represent the distance from the SVM prediction boundary. Enzyme selection-based re-design of the BIA pathway from an unstable THP route to a more stable norcoclaurine (NC) route (b - right side). Red arrows represent the rapid degradation of unstable intermediates DHPAA and THP. c, LC-MS detection of products from Thalictrum flavum norcoclaurine synthase (TfNCS) containing strains T1-01-DE3 (wild-type PsTyDC1 + TfNCS), T1-02-DE3 (PsTyDC1-L205H + TfNCS), T1-03-DE3 (PsTyDC1-Y98F-F99Y-L205N + TfNCS), T3-01-DE3 (wild-type PsTyDC3 + TfNCS), T3-02-DE3 (PsTyDC3-I370S + TfNCS) and T3-03-DE3 (PsTyDC3-Y100F-F101Y-H203N + TfNCS) (Table 1) grown in LB supplemented with 1 mM tyrosine and 0.5 mM dopamine, at 28°C with 180 rpm shaking for 51 hours. The downstream AAS product norcoclaurine is selectively produced by wild-type PsTyDC1, as well as triple variants of PsTyDC1 and PsTyDC3 with engineered active sites based on that of insect DHPAAS. Tyramine is the major product of PsTyDC1-L205H and PsTyDC3-I370S, which contain engineered active sites based on canonical aromatic amino acid decarboxylase (AAAD). Similar results are replicated in strains T1-04-DE3 (wild-type PsTyDC1 + PsONCS3), T1-05-DE3 (PsTyDC1-L205H + PsONCS3), T1-06-DE3 (PsTyDC1-Y98F-F99Y-L205N + PsONCS3), T3-04-DE3 (wild-type PsTyDC3 + PsONCS3), T3-05-DE3 (PsTyDC3-I370S + PsONCS3) and T3-06-DE3 (PsTyDC3-Y100F-F101Y-H203N + PsONCS3) (Supplementary Fig. 1), as well as strains T1-07-DE3, T1-08-DE3 and T1-09-DE3 (Supplementary Fig. 2). d, Structural classification of known phenylpyruvate decarboxylase (PPDC) enzymes ARO10, Azospirillum brasilense PPDC (AbPPDC, PDB ID: 2Q5O) in comparison to typical Zymomonas mobilis PDC (ZmPDC, PDB ID: 2WVA) and candidate PPDC sequence PsPDC1. The modeled PsPDC1 active site contains Y332, which is also present in typical pyruvate decarboxylase (PDC) enzymes which decarboxylate pyruvate. In this respect, the PsPDC1 active site is distinct from microbial PPDCs, which all contain smaller residues in place of Y332 (Lactococcus lactis KdcA contains S286 corresponding to Y332). Yet, the presence of Y332 in PsPDC1 does not interfere with docking of tyrosine into the PsPDC1 active site. e, SVM-based prediction of putative PPDC sequences (e - left side), visualized in three dimensions by compressing high-dimensional data (Supplementary Table 2, upper table) into two dimensions (PC1 and PC2) and plotting them together with two-dimensional decision scores. Prediction score trends for PsPDC1, PsPDC2 and Ps2HCLL (2-hydroxyacyl-CoA ligase-like) are similar in high dimensional models (Supplementary Table 2 and Supplementary Fig. 2a). Enzyme selection-based design of a PPDC-dependent pathway to norcococlaurine (NC) bypasses sensitive AAS activity (e - right side). f, PsPDC1 mediates in vivo production of 4-hydroxyphenylethanol (tyrosol) through a 4HPAA intermediate, in M9 medium supplemented with 1.2 mM 4HPP at 25°C with 180 rpm shaking. Strain P1-01-AI, which contains PsPDC1, mediates higher tyrosol production than that of strains P2-01-AI and P3-01-AI, which contain PsPDC2 and Ps2HCLL, respectively. PsPDC1 mediates downstream production of norcoclaurine (NC) from LB supplemented with 5 mM tyrosine and 3.77 mM dopamine in strain P1-02-AI, at 20-25°C with 180 rpm shaking. Here, tyrosol is detected after 71 hours, and norcoclaurine is detected

after 41 hours, from filtered and dried culture medium as TMS-derivatives using GC-MS. PsPDC1 products are replicated in Supplementary Figure 3b.

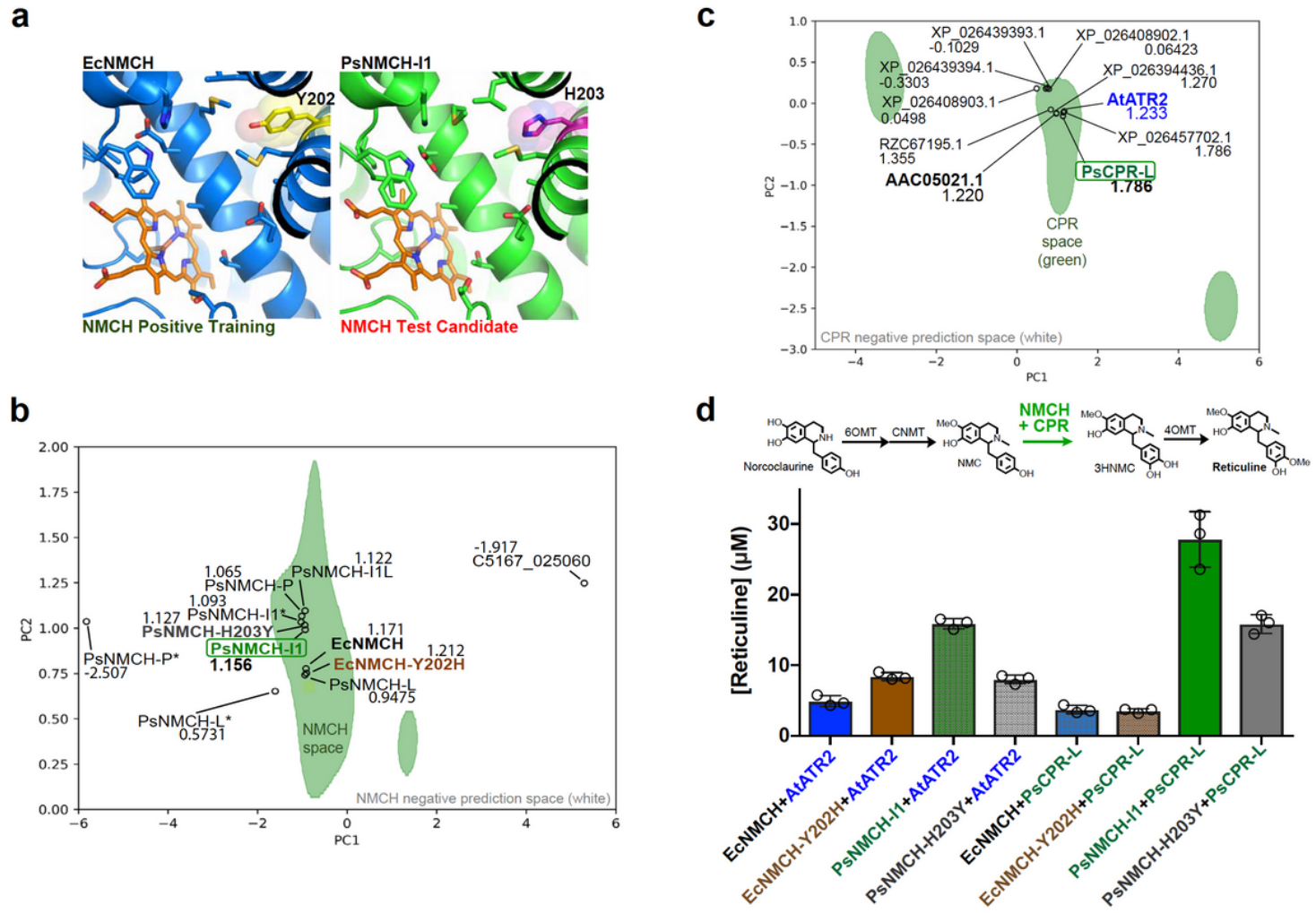

Figure 3

Prediction and discovery of P. somniferum NMCH and CPR for improved reticuline production. a, Comparison of active site configurations of positive training example EcNMCH with that of candidate sequence PsNMCH Isoform 1 (PsNMCH-I1). b, For visual representation, two-dimensional SVM-based prediction of NMCH sequences is shown, with positive and negative prediction spaces colored green and white, respectively. With exception to EcNMCH and EcNMCH-Y202H, all points represent P. somniferum sequences. PCA is used to compress multi-dimensional data into two dimensions (PC1 and PC2) for the visual representation. Corresponding high-dimensional SVM results are detailed in Supplementary Table 3, and high-dimensional SVM decision scores are listed. c, Two-dimensional SVM-based prediction of CPR sequences, with positive and negative prediction spaces colored green and white, respectively. With exception to positive training sequence AtATR2, all points represent tested P. somniferum sequences. PCA is used to compress multi-dimensional data into two dimensions (PC1 and PC2) for the visual representation. Corresponding high-dimensional SVM results are detailed in Supplementary Table 4, and high-dimensional SVM decision scores are listed. d, Conversion of 1.2 mM norcoclaurine to reticuline, mediated by various combinations of NMCH and CPR, together with Ps6OMT, PsCNMT and Ps4OMT, in

strains N1-01-DE3, N1-02-DE3, N1-03-DE3, N1-04-DE3, N2-01-DE3, N2-02-DE3, N2-03-DE3, N2-04-DE3 (Table 1). Here, individual samples were analyzed 3 times to generate bar graphs in Prism 7 with error bars representing standard deviation (n=3).

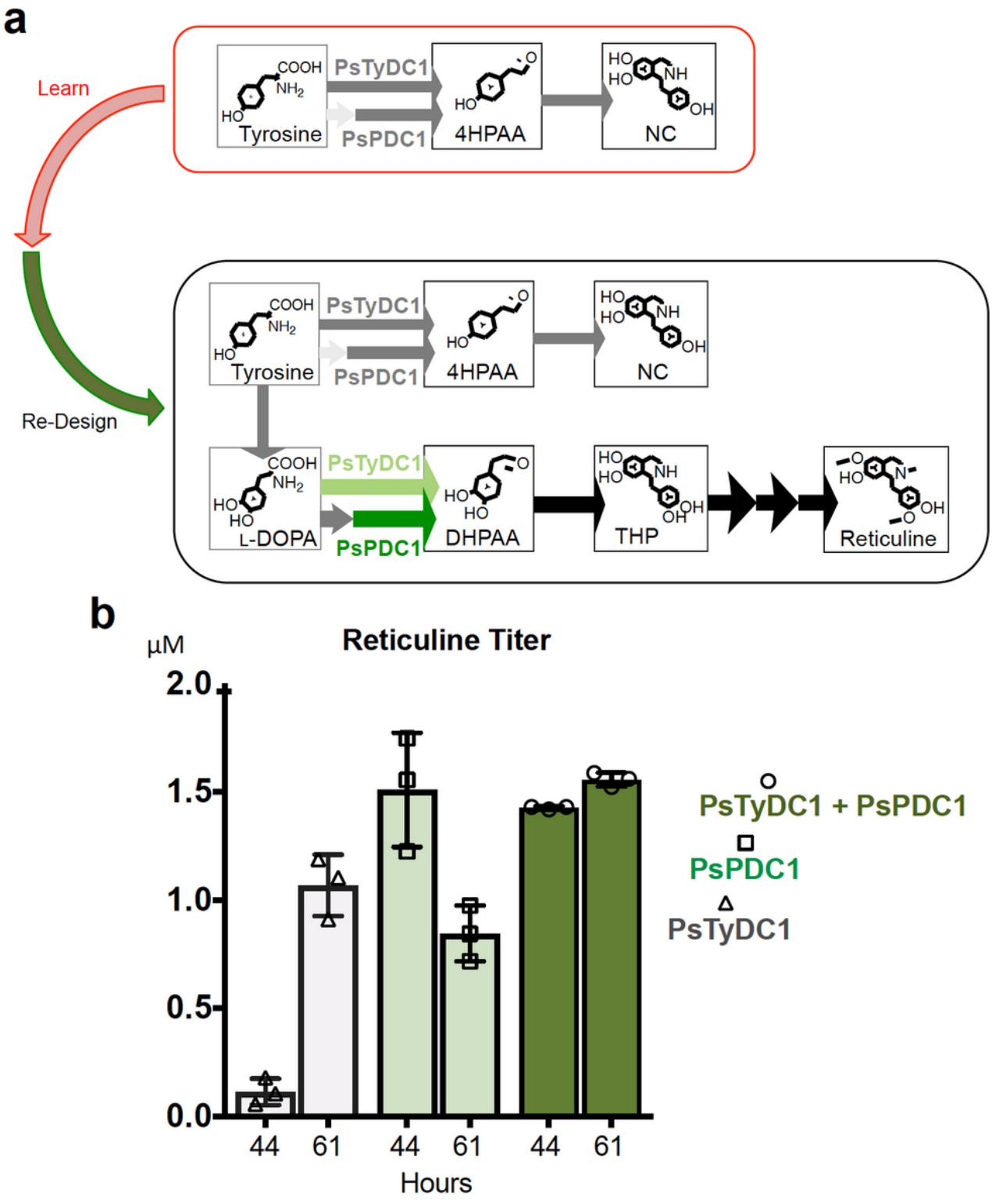

Figure 4

Optimized reticuline titers from L-DOPA mediated by PsPDC1 and PsTyDC1. a, Pathway expansion of the P. somniferum 4HPAA pathway to a dual norcoclaurine (NC) and THP pathway. b, Strains T1-10-DE3, P1-

02-AI and P1-04-AI contain PpDDC, PsONCS3, 6OMT, CNMT and 4OMT in addition to PsTyDC1 (light grey), PsPDC1 (light green) and PsTyDC1+PsPDC1 (dark green), respectively. Cultures were grown to high density in TB before addition of inducing agent, L-DOPA and ascorbate according to the methods section. Replicate samples of filtered culture medium were analyzed with CE-MS (n= 3). Here, 3 samples from individual cultures were analyzed to generate bar graphs in Prism 7 with error bars representing standard deviation (n=3).

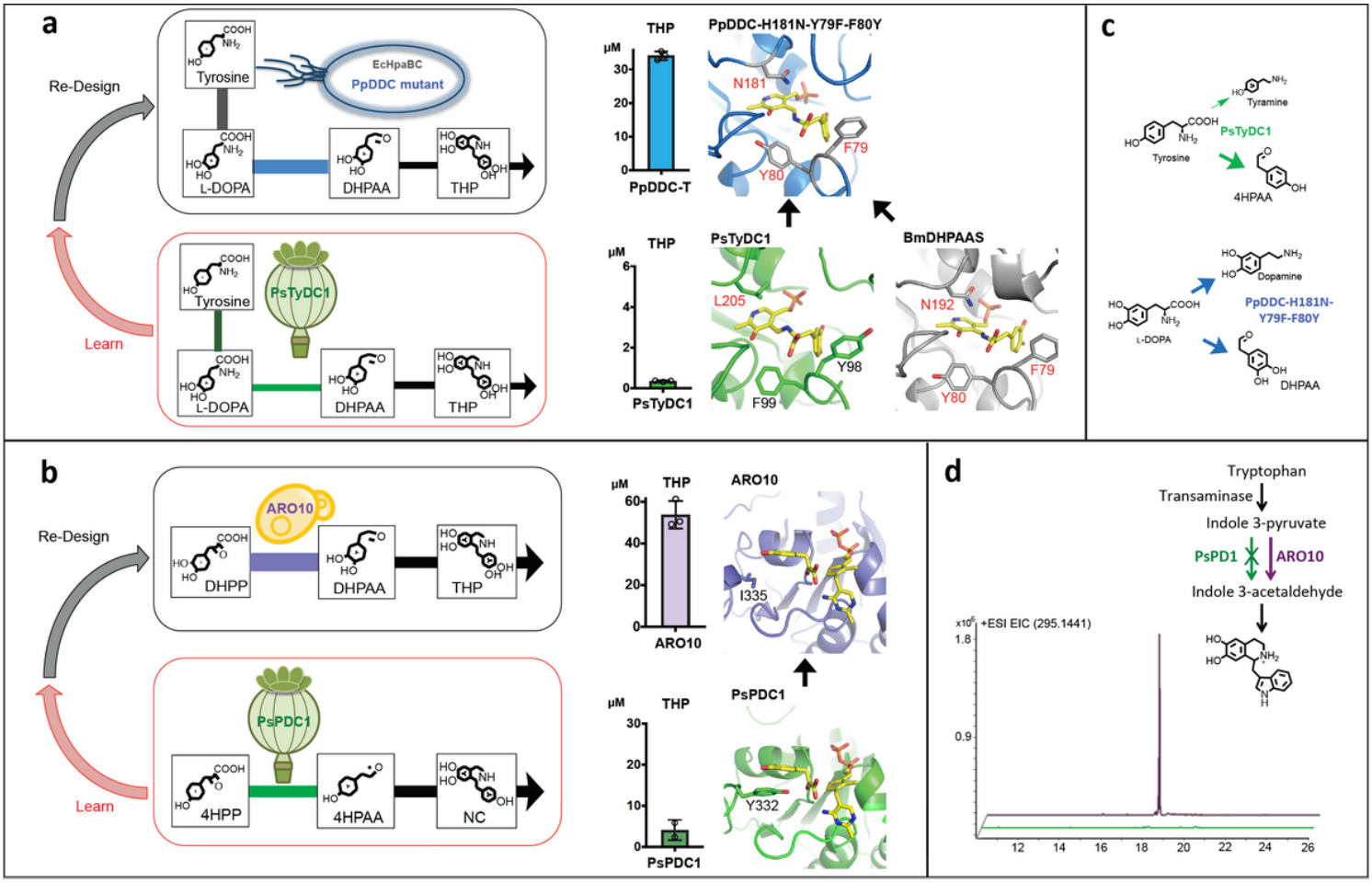

Figure 5

Improved alkaloid production by switching to microbial templates. a, PsTyDC1 is exchanged with an engineered PpDDC template via three active site gain-of-function substitutions, Y79F, F80Y and H181N, to promote AAS activity. THP production from L-DOPA was compared using PsTyDC1 in strain T1-10-DE3 (t = 44 hours, n = 3) and PpDDC-Y79F-F80Y-H181N (DDC-T) in strain DT-02-DE3 (t = 40.5 hours, n = 2). Culture conditions for each strain are described in the methods section. b, PsPDC1 is exchanged with S. cerevisiae ARO10 for higher PPDC activity. Production of THP from L-DOPA by PsPDC1 in strain P1-02-AI is shown (t = 44 hours, n = 3). Production of THP from L-DOPA and dopamine by ARO10 in strain A1-01-DE3 is compared (t = 44 hours, n = 3). Culture conditions are described in the methods section. For panels a and b, 2 or 3 samples from individual cultures were analyzed to generated bar graphs in Prism 7, with error bars representing standard deviation (n=2 or n=3). c, The triple variant PpDDC-T exhibits independent decarboxylase and oxidative deamination activities to produce high levels of both dopamine

and DHPAA, whereas PsTyDC1 favors aldehyde production according to Fig. 2c results. The smaller arrow to and smaller structure of tyramine represent the lower relative production of tyramine mediated by PsTyDC1. d, Strain A1-01-DE3 containing ARO10 metabolizes tryptophan in TB medium to produce an indole-3-acetaldehyde derived indole alkaloid as a non-targeted byproduct during BIA production (t = 61 hours). Strain P1-02-AI containing PsPDC1 did not readily convert indole 3-pyruvate to indole 3-acetaldehyde, as indicated by no detectable indole alkaloid byproduct (t = 61 hours).

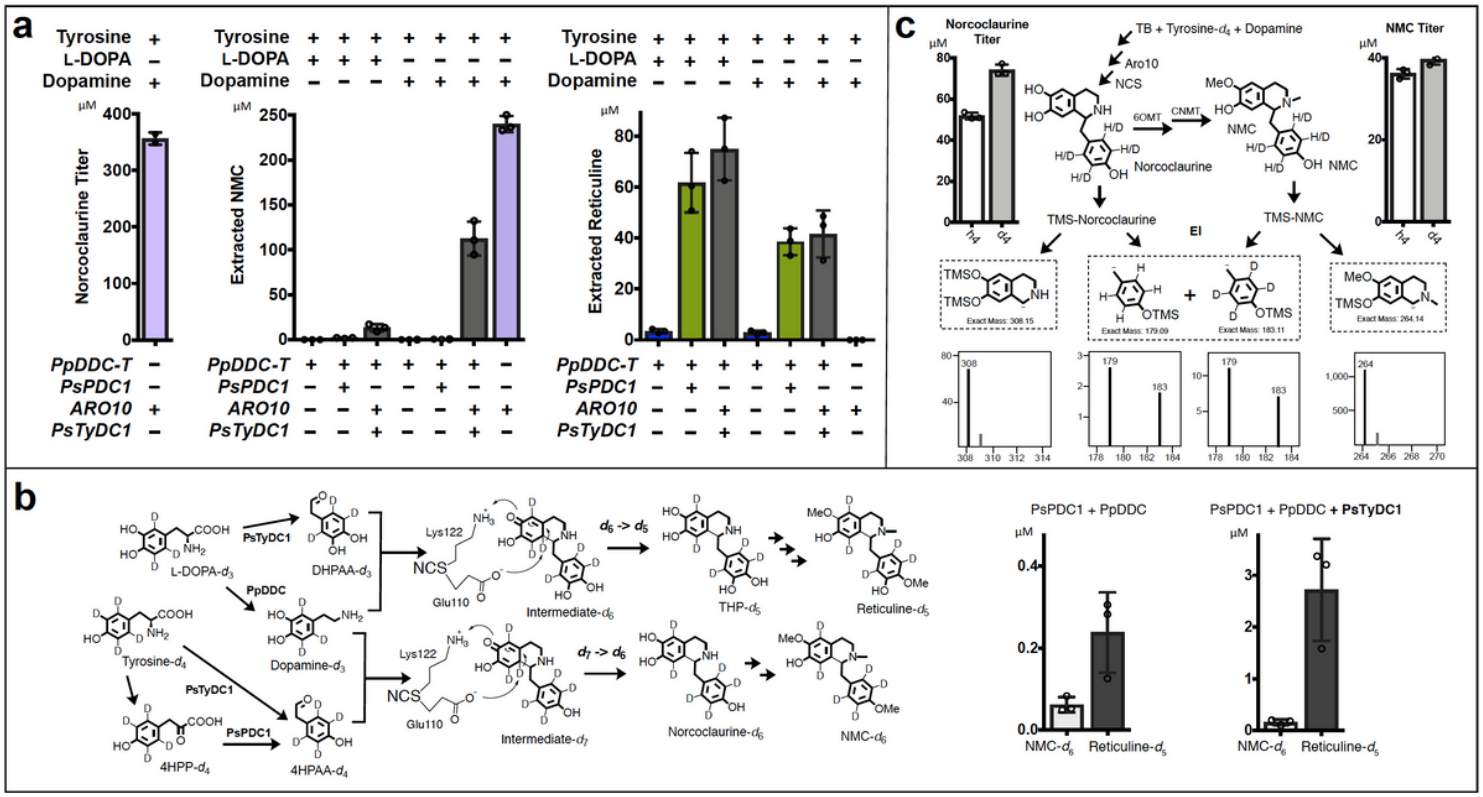

Figure 6

Optimization of norcocluarine, reticuline, and NMC production for analysis of flux through hybrid pathways. a, PpDDC-Y79F-F80Y-H181N (PpDDC-T) and PsPDC1 containing strain P1-07-AI (olive green) prefers the THP containing pathway. Combination of PpDDC-T, ARO10 and TyDC1 in strain A1-06-AI (dark grey) promotes both norcoclaurine and THP containing pathways. ARO10 expressing strain A1-01-DE3 (light purple) converts tyrosine and dopamine to norcoclaurine and NMC. NMC and reticuline were extracted with EtOAc from cultures 40 hours after addition of tyrosine together with L-DOPA or dopamine. Tested strains P1-06-DE3 (blue), P1-07-AI (olive green), A1-06-AI (dark grey) and A1-01-DE3 (light purple) each contain Cj6OMT, CjCNMT, Cj4OMT, NCS, plus the indicated genes of the bottom 4 rows. P1-06-DE3 (blue) and P1-07-AI (olive green) contain the same genes, but P1-06-DE3 was induced with only IPTG, without including arabinose for PsPDC1 expression. Cultures containing PpDDC-T and L-DOPA were supplemented with additional sodium ascorbate. The BL21(AI) derived strain P1-07-AI was induced with IPTG and arabinose. For improved NMC production, A1-01-DE3 was supplemented with the aldehyde reductase/dehydrogenase inhibitor gossypol. Additional culture conditions are described in the methods section. Extracted NMC and reticuline were TMS-derivatized and analyzed with GC-MS (t = 40 hours, n =

3). After extraction, cultures were stored at 4 °C and stable norcoclaurine titers from culture medium were analyzed with LC-MS (n = 2). b, Isotope profiling of strains P1-02-AI (expressing PsPDC1) and P1-04-AI (expressing PsPDC1 and PsTyDC1), which produce NMC-d6 (P1-02-AI - 62 nM, P1-04-AI 160 nM) and reticuline-d5 from tyrosine-d4 and L-DOPA-d3 (t = 61 hours, n = 3). There is a synergistic improvement in BIA production when combining PsPDC1 and TyDC1. Here, NCS catalyzes the loss of a deuterium from in vivo-generated dopamine-d3. c, For tracing aromatic isotope flux from tyrosine to norcoclaurine and NMC, alkaloids were extracted with EtOAc from the A1-01-DE3 culture 40 hours after addition of tyrosine-d3 and dopamine, according to the methods section. Extracted alkaloids were TMS-derivatized and analyzed with GC-MS (n = 3). After extraction, cultures were stored at 4°C and stable BIA titers from culture medium were analyzed with CE-MS (n = 3); the fraction of labeled BIA-d4 and unlabeled BIA from natural tyrosine in the rich TB broth can be quantified. With exception to unlabeled norcoclaurine (n=2), all other individual samples were analyzes 3 times (n=3) to generate bar graphs in Prism 7 with error bars representing standard deviation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GraphicalAbstract.jpg
- MLinksSINCB210402.docx