ACS Catalysis

Open Access
This article is licensed under CC-BY-NC-ND 4.0

pubs.acs.org/acscatalysis
Research Article

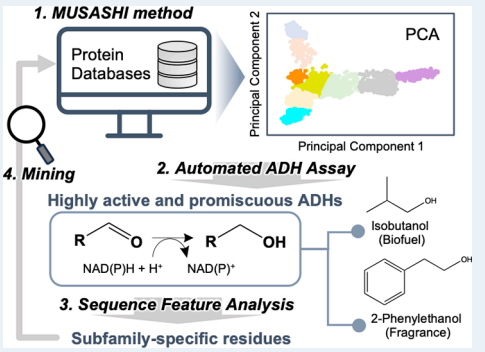
Identification of Subfamily Specific Residues within Highly Active and Promiscuous Alcohol Dehydrogenases

Ryota Hidese, Kanae Sakai, Musashi Takenaka, Keiji Fushimi, Hisashi Kudo, Kenya Tanaka, Ryo Nasuno, Christopher J. Vavricka, Akihiko Kondo, and Tomohisa Hasunuma*

Cite This: ACS Catal. 2025, 15, 11931−11943
Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Enzyme selection is an essential process in the biobased production of chemicals. It is essential to develop a method to extract yet unknown useful enzymes from protein databases. Enzymes that exhibit substrate promiscuity and high activity hold the potential to access unknown reactions and mediate known reactions with a higher performance. Herein, we propose and validate a principal component analysis (PCA)-based classification method, termed MUSASHI (**MU**ltiple-**S**equence **A**lignment-based protein **S**election via clustering using **HI**gh-dimensional analysis), to identify subfamily-specific residues that are highly conserved among promiscuous alcohol dehydrogenase (ADH). Specifically, zinc-dependent ADH homologues retrieved from the protein database were classified into 9 groups, and according to PCA-based clustering, the activities of 18 ADHs, with representative enzymes from each group, were characterized. As a result, we identified two promiscuous ADH groups: Group 1 ADH, efficient with short-chain and aromatic aldehydes, and Group 3 ADH, efficient with aliphatic and aromatic ketones. Sequence feature analysis then revealed subfamily-specific residues, which are highly conserved only in promiscuous ADH Groups 1 and 3, with the potential to biosynthesize a wide spectrum of target compounds. *Tatumella ptyseos* ADH, identified from Group 1 of this study, showed higher isobutanol and 2-phenylethanol bioconversions than that of a conventional ADH (Ahr). These results indicate that the MUSASHI method for subfamily-specific residue identification can enable optimal enzyme selection from protein databases.



KEYWORDS: *alcohol dehydrogenase, principal component analysis, subfamily specific residues, substrate promiscuity, automated enzyme assay system*

Downloaded via CHINESE ACADEMY AGRICULTURAL SCI on July 1, 2025 at 08:31:10 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Fuel and chemical production by microorganisms from sustainable biomass or carbon dioxide has been attracting attention due to concerns about global environmental changes and the depletion of fossil resources. Recent innovations in biotechnology are enabling the biobased production of various compounds, such as alcohols, amines, and carboxylic acids, that are currently produced from petroleum, to meet industrial requirements such as productivity, cost performance, and environmental sustainability.[1,2] Metabolic engineering, which relies upon enzyme selection, metabolic pathway optimization, and modification of gene regulatory networks, is necessary to increase carbon uptake and flux to the desired end-product.[3−8]

Enzyme selection for pathway optimization is one of the most important steps that determines the productivity of end-product production.[9,10] When biosynthetic pathways for natural or artificial products are constructed, enzymes are conventionally selected on the basis of reported experimental characterizations. On the other hand, the use of yet-unknown natural enzymes offers potential to further improve productivity or expand to additional target compounds through metabolic pathway construction.[11] A protein database, UniProt, which is one of the most widely used protein information resources of protein sequence and functional information, contains entries for over 227 million sequence records[12] and includes 3D structure prediction by AlphaFold[13] for more than 85% of all entries. However, the annotated functions of most sequences in databases such as UniProt are inferred based on sequence homology with experimentally validated enzymes. Therefore, mining novel enzymes retrieved from a protein database often requires time-consuming and labor-intensive processes to experimentally verify the enzymatic properties.[14] Accordingly, the high-quality prediction of protein function is still a great challenge in metabolic engineering.

It is well-known that some enzymes show substrate promiscuity, i.e., catalysis of the same chemical reaction for a

Received: April 22, 2025
Revised: May 29, 2025
Accepted: June 5, 2025

© XXXX The Authors. Published by American Chemical Society

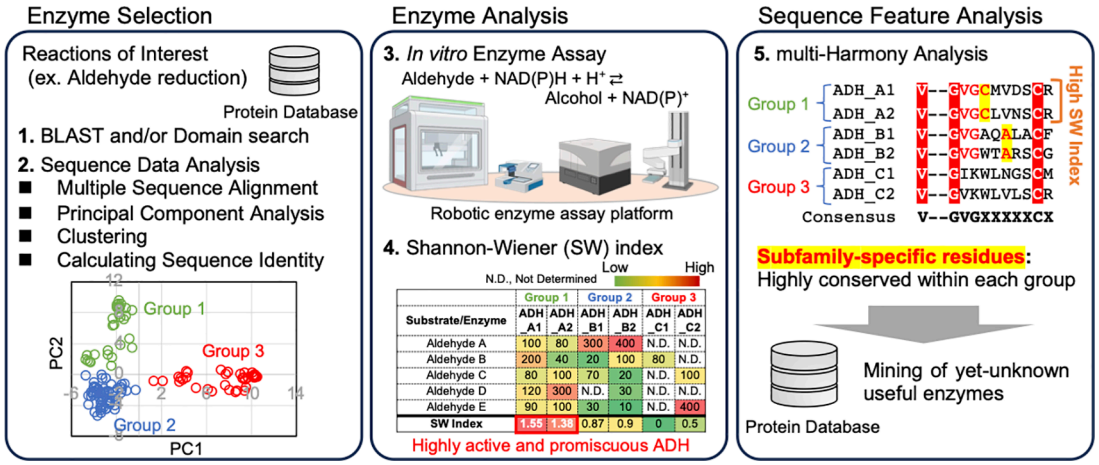


Figure 1. Workflow for the identification of highly active and promiscuous enzymes from protein databases. 1. Enzymes with common reaction specificity (i.e., ADH, alcohol dehydrogenase) are selected as queries and protein sequences are comprehensively extracted from protein databases. 2. Large-scale multiple sequence alignments are converted to binary vectors for principal component analysis (PCA). PCA plots are grouped by *k*-means clustering. Then, representative sequences are selected based on sequence identity score among each group of the resulting PCA analysis. 3. Enzyme activity data of a purified selected enzyme (in this case ADH) toward a variety of candidate substrates, including unknown substrates, are obtained by an automated enzyme assay system. 4. Enzymes in each group are ranked based on the breadth of substrate specificity and high activity by Shannon−Wiener (SW) analysis. SW index is calculated based on specific activities for all substrates for a particular ADH. 5. multi-Harmony analysis is then applied to identify subfamily-specific residues that are highly conserved only within each group, leading to prediction of useful functions of unexplored enzymes in protein databases.

range of different substrates including both physiologically relevant and irrelevant substrates, such as biologically harmful compounds.[15,16] Alcohol dehydrogenase (ADH) (EC 1.1.1.1) catalyzes the reversible reduction of aldehydes/ketones, using nicotinamide adenine dinucleotide (NADH) or nicotinamide adenine dinucleotide phosphate (NADPH) as a coenzyme.[17,18] ADHs including YjgB (Ahr), YahK, and YqhD exhibit promiscuous substrate recognition with high activity and can be used to produce alcohols.[19] The microbial production of industrially relevant alcohols, such as ethanol, *n*-butanol,[20] isobutanol,[21,22] 1,3-propanediol,[23] and 2-phenylethanol[24] has been extensively studied; and in the current reports, ADH is considered as a major bottleneck. Despite the great potential of promiscuous ADH, it is difficult to predict the substrate promiscuity by phylogenetic analysis alone because the sequence features of the promiscuous ADHs have not been systematically identified.

To detect "subfamily-specific residues", which are highly conserved only within a homologous protein group, among more than thousands of proteins from a protein database, a large sequence data set from multiple sequence alignments must be processed to minimize the loss of multidimensional information and computational load for analysis. Multivariate analysis is a promising approach to classify functionally specific subfamilies and to identify specifically conserved features related to functional specificity within a protein family.[25−29] Principal component analysis (PCA), an unsupervised multivariate technique, can readily be applied to identify possible functional residues from multiple sequence alignments.[25,28−30] Therefore, PCA is an ideal method to extract features from a large number of sequence alignment sets.

Here, we report comprehensive ADH enzyme classification by way of PCA analysis of original binary vector matrices, where each amino acid residue of a multiple sequence alignment is converted into a minimal descriptor. Based on the PCA analysis, we selected a total of 18 ADHs from 9 groups and screened the activities of selected enzymes using a library of aldehyde, ketone, and alcohol substrates. We then successfully identified "subfamily-specific residues" within highly active and promiscuous ADHs by analysis of sequence features using the PCA-based clustering, a method termed MUSASHI (**MU**ltiple-**S**equence **A**lignment-based protein **S**election via clustering using **HI**gh-dimensional analysis). This study establishes the importance of subfamily-specific residue identification to improve the prediction of unexplored enzymes from protein databases (Figure 1).

RESULTS

ADH Classification Based on PCA Clustering of Sequence Alignments. NAD(P)$^+$-dependent ADH is classified into three nonhomologous families based on sequence length:[17] zinc-containing medium-chain ADH superfamily, including YahK and Ahr (Type I), short-chain ADH superfamily (Type II), and iron-containing long-chain ADH superfamily, including YqhD (Type III). Of these ADH superfamilies, the enzymatic properties of type I ADH, which includes *Saccharomyces cerevisiae* ADH1-6, are well-studied.[18] We retrieved type I ADH sequences from the UniProt database. Two proteins that are more than 80% identical with conserved active sites often share the same function;[31] therefore, we used Cluster Database at High Identity with Tolerance (CD-HIT) to remove redundant sequences that share more than 90% identity.[32] The resulting set of 6727 sequences was aligned using the multiple alignment using fast Fourier transform (MAFFT) algorithm.[33,34]

PCA can process multidimensional information, including multiple sequence alignments. For PCA analysis, the multiple sequence alignment must be converted to numerical representation schemes that highly correlate with multidimensional information. Therefore, we applied binary vector profiling of sequence patterns to maximize variance in a mean-centered variance/covariance matrix in PCA. One-letter symbols and gaps in the multiple sequence alignment were converted to binary representation (5 rows × 21 columns were

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

used to represent each amino acid type and a sequence gap as shown in Table S1). PCA, based on the covariance matrix, was then used to project the multidimensional ADH sequence information onto 2 dimensions (Data set 1). The vertical and horizontal axes of Figure 2 are derived as principal components

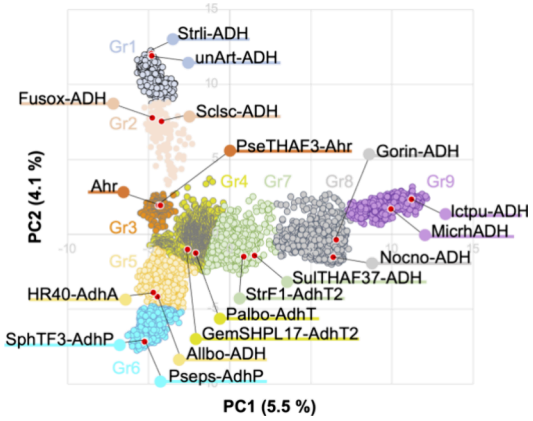


Figure 2. PCA of the ADH data set. A 6727-sequence multiple sequence alignment was projected to 2 dimensions of PCA, principal component 1 and 2 with eigenvalue. The plot of PCA scores (PC1 vs PC2) with *k*-means clustering is visualized in 9 different colors, corresponding to each group. Specific ADHs are abbreviated based on their natural host microorganisms (see Materials and Methods).

1 (eigenvalue: 5.5%) and 2 (eigenvalue: 4.1%), respectively, which represent the amount of variance that can be explained by the principal component. The PCA score plots were classified into 9 groups using a *k*-means clustering algorithm with the elbow method, which is an unsupervised algorithm that assigns data points into different clusters[35,36] (Figure S1). The elbow method indicated the k = 2.5−5 range of the inflection point in the plot of inertia (within-cluster sum of squares) versus k (Figure S1). In the range of k = 6−9, the decrease in inertia became more gradual. Although the reduction in inertia at k = 9 was not steep, we determined that the clustering quality improved sufficiently at this point. The identity score between sequences in each group was calculated using the sequence demarcation tool (SDT) version 1.3, which displays the percentage pairwise identity.[37] Because sequences with high total scores are considered to represent the average sequence of all sequences within the group, sequences with higher scores were selected as representative sequences in each group (Data set 2). Across all groups, the enzyme selection was further based on solubility and recombinant protein expression levels in *E. coli* expression system. For Groups 1 to 8, the two enzymes were selected as follows: group 1, unArt-ADH and Strli-ADH; group 2, Fusox-ADH and Sclsc-ADH; group 3, Ahr and PseTHAF3-Ahr; group 4, GemSHPL17-AdhT2 and Palbo-AdhT; group 5, HR40-AdhA and Allbo-ADH; group 6, Pseps-AdhP and SphTF3-AdhP; group 7, StrF1-AdhT2 and SulTHAF37-ADH; group 8, Gorin-ADH and Nocno-ADH. For Group 9, the enzyme was selected (MicrhADH) along with the enzyme derived from bacteria (Ictpu-ADH). Protein sequence homology was assessed using the NCBI BLAST program.[38,39] As a result, ADHs that exhibited significant homology clustered together in the PCA score plot (Figures S2 and 2). Therefore, our original binary vector profiling method is considered to be effective in transforming major features of multidimensional information in the multiple sequence alignment into the PCA score plots.

Activity Screening of Representative ADHs of Each PCA Group. To investigate the substrate specificity and activity of 18 ADHs selected as described above, 41 commercially available substrates, which are short- and long-chain aliphatic aldehydes, aromatic aldehydes, keto acids, aliphatic and aromatic ketones, and alcohols, were selected for the assay (Figure 3A). Since Ahr (*E. coli*) and YahK (*E. coli*) have been reported to have wide substrate specificity,[19] we selected these enzymes as benchmarks in the present assay system. Recombinant ADH was expressed in *E. coli* BL21 (DE3) and purified to homogeneity (Figure S3). Both ADHs from Groups 4 and 8 were partially purified due to their low protein expression levels in the current *E. coli* expression system. All enzyme assays with each substrate were measured in the presence of $NAD(P)^+$/NAD(P)H cofactors under neutral pH and at 37 °C, with the final goal of producing useful substances by widely used microbes such as *E. coli*. The cofactor dependency of each ADH was determined through sequence analysis[40] and by assessing activity toward 41 substrates in the presence of $NAD(P)^+$/NAD(P)H cofactors. The substrate concentration of each reaction is shown in Data set 3A. After testing each of the 18 ADHs by an automated enzyme assay system, Group 1 ADHs (Strli-ADH and unArt-ADH), Group 2 ADHs (Fusox-ADH and Sclsc-ADH), and Group 3 ADHs (Ahr and PseTHAF3-Ahr) showed wide substrate specificity with high activity toward both aldehydes and ketones. In contrast, ADHs in Groups 4, 5, 6, 7, 8, and 9 showed lower activity toward substrates tested when compared to those in Groups 1, 2, and 3. As previously reported,[19] Ahr in Group 3 showed the highest activities toward various aldehyde substrates, such as acetaldehyde and caproaldehyde, relative to enzymes of the other groups (Figure 3B). The activity of YahK toward butyraldehyde (15.9 μmol min^{-1} mg^{-1}) obtained in the present assay system was comparable to that previously reported (YahK for butyraldehyde: approximately 20 μmol min^{-1} mg^{-1}),[19] suggesting the validity of our assay system for activity measurement. In contrast, Group 1 ADH showed higher activity toward cinnamaldehyde and phenylacetaldehyde than ADH of the other groups. The spectra of specific activities toward various substrates showed similar trends among the enzymes of each group. To rank ADH groups based on both wide substrate specificity and high activity, the Shannon−Wiener index (H′) was calculated for each of the 18 enzymes, based on specific activities toward 41 substrates (Figure 3B). The Shannon−Wiener (SW) index was originally developed to estimate ecological species diversity[41] and was first adopted to quantitatively describe the substrate spectra of lipases/esterases,[42] where high specific activity across many substrates will result in a high index value. Because Ahr exhibited particularly high activity toward specific substrates, including benzaldehyde and butyraldehyde, the index value is lower based on the applied formula. As a result, Group 1 ADH showed higher SW indexes compared to ADH of the other groups, indicating that Group 1 ADH exhibits wide substrate specificity and high activity (Figure 3C). To analyze the common characteristics observed in the Group 1 and Group 3 ADHs (i.e., broad substrate specificity and high catalytic activity), we arbitrarily extracted enzyme sequences from Group 1 (Strsc-AdhC2, YahK, and TptADH) and Group 3 (Phyma-ADH, SphSMR4y-ADH, and TguADH) that showed moderate or low SDT total scores (Data set 2). We then

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

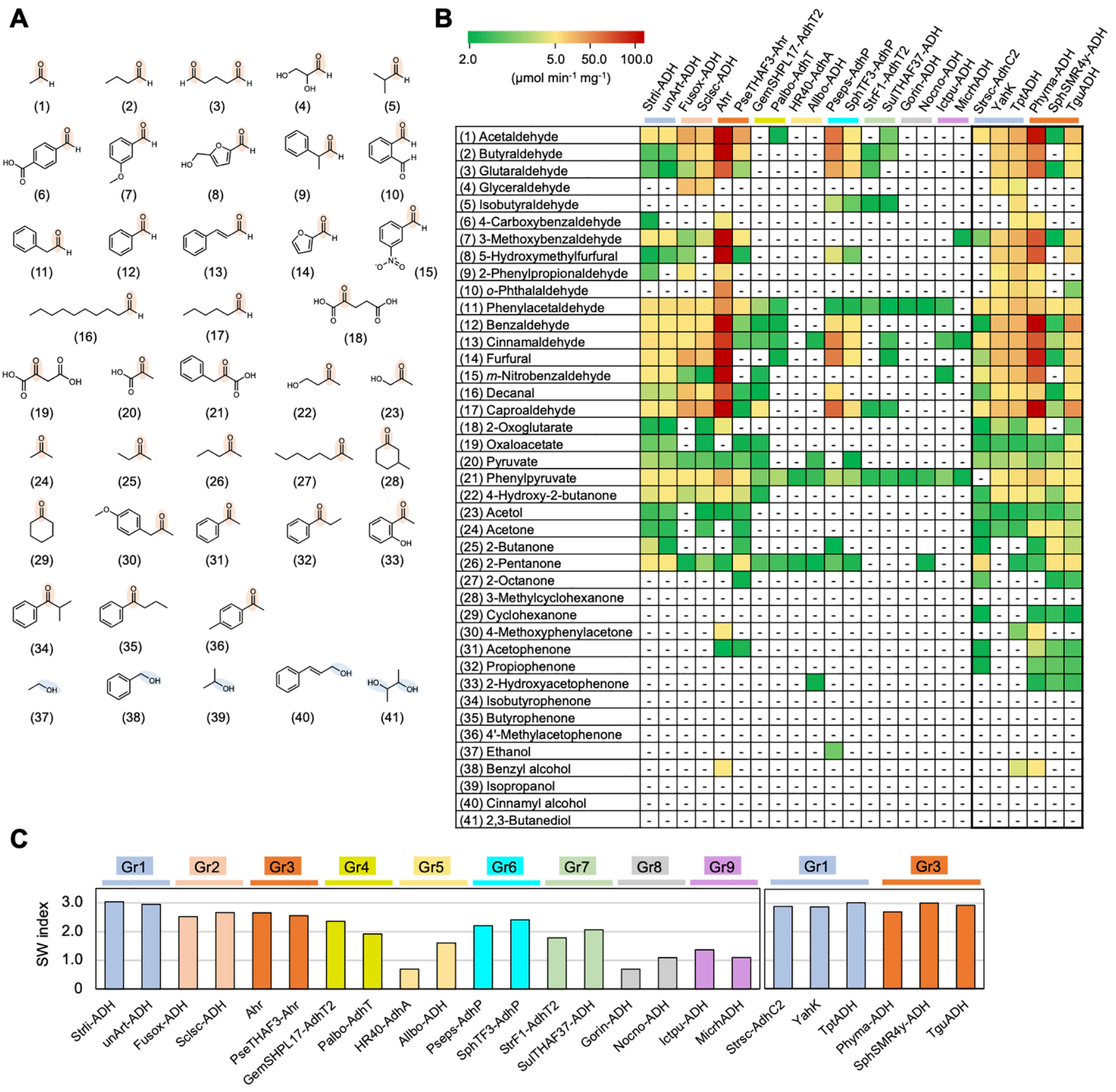


Figure 3. Enzymatic assay of the 18 ADHs and SW analysis. (A) Chemical structures of the substrates for ADHs. Short-aliphatic aldehydes,[1−5] aromatic aldehydes,[6−15] long-aliphatic aldehydes,[16,17] keto acids,[18−21] aliphatic ketones,[22−29] aromatic ketones,[30−36] and alcohols[37−41]. (B) Specific activities of the purified ADHs toward 36 aldehyde and 5 alcohol substrates. Heat map of the specific activity values obtained by each reaction with 18 ADHs and 41 substrates. The hyphen means that the activity was below 2.0 μmol min^{-1} mg^{-1} or was not detected. All data are presented as the mean of four technical replicates. (C) SW index, calculated based on specific activities (more than 2.0 μmol min^{-1} mg^{-1}) of each ADH for 41 substrates, were shown.

measured their enzymatic activity against 41 kinds of substrates. As a result, the tested enzymes exhibited broad substrate specificity and high catalytic efficiency. These in vitro results show that the PCA-based clustering of sequence features can be used to group enzymes with similar functions.

The enzymatic screening data in Figure 3B show that TptADH (Group 1 ADH) and TguADH (Group 3 ADH) exhibit high activity against a wide range of substrates, revealing the catalytic efficiency of TptADH and TguADH for the first time (Table 1 and Figure S4). The K_m values of the reduction reactions toward the different substrates varied acetaldehyde) for TptADH and 0.02 (acetone) and 10.99 mM (2-octanone) for TguADH. It is known that YahK in Group 1 shows lower K_m values than Ahr in Group 3 for short-aliphatic aldehydes, such as acetaldehyde and isobutyraldehyde.[19] TptADH showed high affinities for short-aliphatic aldehydes, especially the branched short-chain aldehyde isobutyraldehyde, resulting in significantly higher catalytic efficiency than TguADH (Table 1). On the other hand, TguADH showed broad substrate specificity compared to TptADH and showed high affinities for aliphatic and aromatic ketones, including acetone and acetophenone, respectively. The enzymatic

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

Table 1. Kinetic Parameters of TptADH and TguADH for Different Substrates[a]

	TptADH (group 1)		TguADH (group 3)	
reagents	K_m (mM)	k_{cat}/K_m (s^{-1} mM^{-1})	K_m (mM)	k_{cat}/K_m (s^{-1} mM^{-1})
short-aliphatic aldehydes				
(1) acetaldehyde	2.12 ± 0.62	6.42 ± 1.93	8.52 ± 2.35	2.36 ± 0.71
(2) butyraldehyde	0.08 ± 0.02	179.07 ± 47.76	0.47 ± 0.09	138.01 ± 31.87
(3) glutaraldehyde	0.24 ± 0.00	100.09 ± 5.01	0.50 ± 0.03	144.11 ± 4.95
(4) glyceraldehyde	0.44 ± 0.08	15.84 ± 0.99	0.06 ± 0.02	68.26 ± 14.50
(5) isobutyraldehyde	0.15 ± 0.01	135.09 ± 21.97	4.04 ± 0.35	1.75 ± 0.19
aromatic aldehydes				
(6) 4-carboxybenzaldehyde	N. D.	N. D.	0.07 ± 0.01	40.23 ± 2.39
(7) 3-methoxybenzaldehyde	0.07 ± 0.01	242.20 ± 25.73	0.21 ± 0.03	201.62 ± 24.68
(8) 5-hydroxymethylfurfural	0.08 ± 0.02	198.72 ± 12.71	0.16 ± 0.01	355.76 ± 26.57
(9) 2-phenylpropionaldehyde	0.08 ± 0.01	163.28 ± 6.87	N. D.	N. D.
(10) *o*-phthalaldehyde	0.29 ± 0.06	65.67 ± 5.08	0.20 ± 0.01	201.59 ± 5.16
(11) phenylacetaldehyde	0.11 ± 0.02	133.40 ± 5.80	0.61 ± 0.11	27.01 ± 3.32
(12) benzaldehyde	0.08 ± 0.01	254.41 ± 18.60	0.19 ± 0.03	382.40 ± 47.81
(13) cinnamaldehyde	0.10 ± 0.01	178.87 ± 13.44	$K_{0.5}$ = 0.04 ± 0.00 (n_{Hill} = 1.9)	$k_{cat}/K_{0.5}$ = 852.75 ± 24.08
(14) furfural	0.04 ± 0.00	479.10 ± 25.75	0.10 ± 0.01	448.75 ± 55.59
(15) *m*-nitrobenzaldehyde	0.02 ± 0.00	615.45 ± 29.59	0.06 ± 0.01	792.84 ± 71.83
long-aliphatic aldehydes				
(16) decanal	K_m = 0.48 ± 0.12 (K_i = 1.13)	k_{cat}/K_m = 28.39 ± 9.41	1.32 ± 0.56	32.31 ± 6.30
(17) caproaldehyde	0.26 ± 0.06	105.29 ± 15.83	0.84 ± 0.33	125.98 ± 33.26
Keto Acids				
(18) 2-oxoglutarate	K_m = 40.19 ± 2.33 (K_i = 42.75)	k_{cat}/K_m = 10.43 ± 2.75	$K_{0.5}$ = 16.57 ± 0.38 (n_{Hill} = 2.3)	$k_{cat}/K_{0.5}$ = 10.08 ± 0.51
(19) oxaloacetate	K_m = 22.51 ± 0.27 (K_i = 115.65)	k_{cat}/K_m = 12.45 ± 0.91	$K_{0.5}$ = 16.97 ± 0.58 (n_{Hill} = 2.4)	$k_{cat}/K_{0.5}$ = 11.23 ± 0.07
(20) pyruvate	K_m = 28.95 ± 2.56 (K_i = 82.53)	k_{cat}/K_m = 10.81 ± 1.83	$K_{0.5}$ = 18.14 ± 0.80 (n_{Hill} = 3.8)	$k_{cat}/K_{0.5}$ = 9.30 ± 0.42
(21) phenylpyruvate	$K_{0.5}$ = 20.37 ± 2.22 (n_{Hill} = 1.9)	$k_{cat}/K_{0.5}$ = 4.27 ± 0.25	N. D.	N. D.
ketones				
(22) 4-hydroxy-2-butanone	0.01 ± 0.00	213.83 ± 74.58	K_m = 0.26 ± 0.08 (K_i = 5.19)	k_{cat}/K_m = 24.80 ± 12.86
(23) acetol	N. D.	N. D.	0.98 ± 0.61	4.07 ± 2.13
(24) acetone	0.06 ± 0.03	27.00 ± 7.53	0.02 ± 0.01	112.54 ± 38.21
(25) 2-butanone	N. D.	N. D.	0.08 ± 0.01	35.34 ± 7.90
(26) 2-pentanone	N. D.	N. D.	N. D.	N. D.
(27) 2-octanone	N. D.	N. D.	10.99 ± 3.41	0.50 ± 0.06
(28) 3-methylcyclohexanone	N. D.	N. D.	N. D.	N. D.
(29) cyclohexanone	N. D.	N. D.	0.01 ± 0.00	203.09 ± 48.44
(30) 4-methoxyphenylacetone	N. D.	N. D.	N. D.	N. D.
(31) acetophenone	N. D.	N. D.	0.04 ± 0.01	86.38 ± 21.46
(32) propiophenone	N. D.	N. D.	0.03 ± 0.01	126.81 ± 45.19
(33) 2-hydroxyacetophenone	N. D.	N. D.	0.02 ± 0.0	205.61 ± 44.17
(34) isobutyrophenone	N. D.	N. D.	N. D.	N. D.
(35) butyrophenone	N. D.	N. D.	N. D.	N. D.
(36) 4′-methylacetophenone	N. D.	N. D.	N. D.	N. D.
alcohols				
(37) ethanol	N. D.	N. D.	N. D.	N. D.
(38) benzyl alcohol	20.44 ± 0.42	0.14 ± 0.00	N. D.	N. D.
(39) isopropanol	N. D.	N. D.	N. D.	N. D.
(40) cinnamyl alcohol	1.26 ± 0.04	1.25 ± 0.04	N. D.	N. D.
(41) 2,3-butanediol	N. D.	N. D.	N. D.	N. D.

[a]N.D., Not determined; n_{Hill}, Hill coefficient; K_i, substrate inhibitory constant; $K_{0.5}$, substrate concentration at half-maximum velocity.

activities with decanal and keto acids (2-oxoglutarate, oxaloacetate, and pyruvate) for TptADH and with 4-hydroxy-2-butanone for TguADH decrease at high substrate concentrations after reaching V_{max}. The experimental data fit well with the equation describing *substrate inhibition* (eq 1). While a sigmoidal curve when plotted against the concentrations of phenylpyruvate for TptADH (Figure S4A) and cinnamaldehyde and keto acids (2-oxoglutarate, oxaloacetate, and pyruvate) for TguADH (Figure S4A,B), the plots were well-fitted to the Hill equation (eq 2). In contrast to the reduction reaction, TptADH and TguADH seem to have relatively low catalytic efficiency for alcohol substrates.

Identification of Groups 1 and 3 ADH Sequence Features. TptADH showed higher affinities toward short-chain aliphatic aldehydes and aromatic aldehydes than TguADH; while the activities of TptADH toward these substrates were lower than those of TguADH, even though TptADH and TguADH have promiscuous features. Conserved amino acid residues among proteins that share functional similarity are defined as subfamily-specific residues, and these

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

A

	YahK Sequence	Gly87	Cys88	Val90	Ser92	Ser165	Ala246	Ile294	Tyr338
	Ahr Sequence	Gly90	Trp91	Ala93	Ser95	Lys159	Asn240	Thr287	Tyr331
	Score	1.14	1.19	1.11	1.08	1.25	1.09	1.24	1.09
Gr1	YahK	G	C	V	S	S	A	I	Y
Gr1	Tatpt-ADH	G	C	V	S	S	A	I	Y
Gr1	unArt-ADH	G	C	V	S	S	S	I	Y
Gr1	Strli-ADH	G	C	V	S	S	S	I	Y
Gr1	Strsc-AdhC2	G	C	V	S	S	S	I	Y
Gr2	Fusox-ADH	G	A	V	S	S	T	L	Y
Gr2	Sclsc-ADH	G	A	S	S	S	S	L	Y
Gr3	Tragu-ADH	G	W	A	S	K	N	T	Y
Gr3	Ahr	G	W	A	S	K	N	T	Y
Gr3	SphSMR4y-ADH	G	G	S	Y	N	D	V	Y
Gr3	PseTHAF3-Ahr	G	W	S	Y	N	N	V	Y
Gr3	Phyma-ADH	G	W	A	S	K	N	T	Y
Gr4	GemSHPL17-AdhT2	C	F	P	–	N	G	V	G
Gr4	Palbo-AdhT	N	F	P	–	H	G	V	G
Gr5	HR40-AdhA	P	W	G	T	R	P	N	G
Gr5	Allbo-ADH	P	W	G	T	R	P	N	G
Gr6	Pseps-AdhP	P	W	Y	T	K	V	V	G
Gr6	SphTF3-AdhP	P	W	Y	A	K	V	V	G
Gr7	StrF1-AdhT2	P	–	V	A	R	G	G	G
Gr7	SulTHAF37-ADH	P	–	I	A	H	G	G	G
Gr8	Gorin-ADH	S	F	P	–	G	D	F	–
Gr8	Nocno-ADH	A	F	P	–	G	G	F	–
Gr9	Ictpu-ADH	V	Y	P	–	G	G	F	–
Gr9	MicrhADH	L	Y	P	–	G	G	F	–

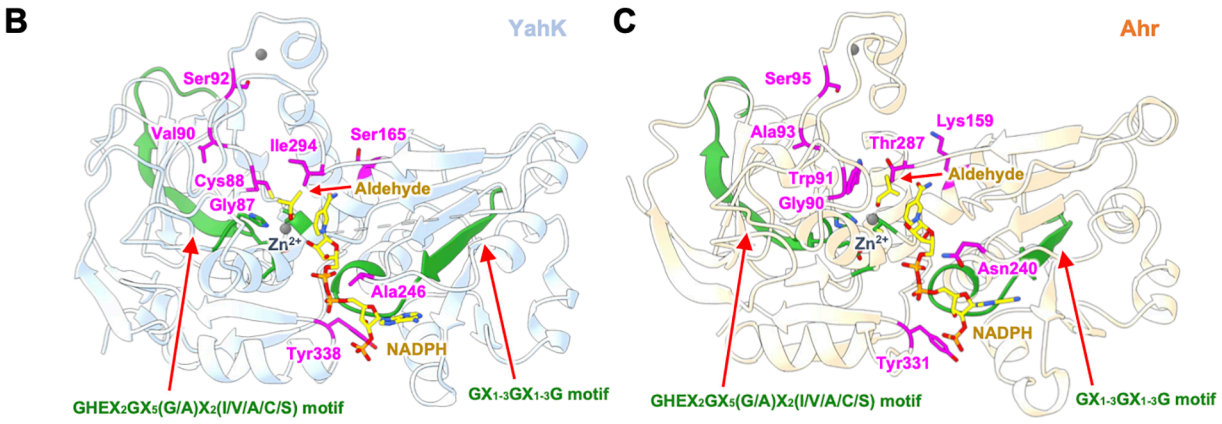


Figure 4. Subfamily-specific residues conserved in Group 1 and 3 ADHs. (A) Subfamily-specific residues of Group 1 ADH (corresponding to eight residues: Gly87, Cys88, Val90, Ser92, Ser165, Ala246, Ile294, and Tyr338 on YahK or Group 3 ADH (corresponding to eight residues: Gly90, Trp91, Ala93, Ser95, Lys159, Asn240, Thr287, and Tyr331 on Ahr) are highlighted in red. (B,C) Structural positions of subfamily-specific residues are shown in magenta. The ADH structures docked with NADPH and isobutyraldehyde of Group 1 ADH YahK (PDB ID: 1UUF) (B) and Group 3 ADH Ahr (PDB ID: 7BU2) (C) as modeled in the MOE are presented. Two distinctive conserved motifs ($GHEX_2GX_5(G/A)X_2(I/V/A/C/S)$ and $GX_{1-3}GX_{1-3}G$ motifs) in the ADH groups are highlighted as green. Three conserved residues near the catalytic Zn^{2+}-binding site (Cys40, His62, and Cys158 on YahK; Cys41, His63, and Cys152 on Ahr) are highlighted in yellow.

residues determine the functional specialization of proteins. To find subfamily-specific residues contributing to the functional specialization of Group 1 and Group 3 ADHs, we aligned sequences of ADHs from all groups using the multi-Harmony server, which combines Sequence Harmony (SH) and multi-Relief (mR) methods.[43,44] A lower SH, ranging from 0 to 1, indicates a more nonoverlapping residue composition among groups, while a higher mR weight, ranging from −1 to 1, indicates identical residues within groups. Thus, a residue that shows a lower SH score and a higher mR weight is a subfamily-specific residue candidate for a particular group. Among a total of 6727 ADH sequences, 73 ADHs from Group 3, 174 ADHs from Group 1, and 840 ADHs from other groups (including 24 experimentally evaluated ADHs) were selected as input sequences for the multi-Harmony web server, while minimizing the extraction of biased plots as much as possible. A total of 1087 ADHs were aligned using MAFFT, and total scores at each amino acid position were calculated according to the equation total score = (1-SH) + mR (Data set 4). As a result, Ahr Cys41 and Cys152, which are responsible for Zn^{2+}-binding and are highly conserved among type I ADH, exhibited low total scores of multi-Harmony (Data set 4). Also, because type I ADH possesses $GHEX_2GX_5(G,A)X_2(I,V,A,C,S)$ and the $GX_{1-3}GX_{1-3}G$ motifs located at the zinc-binding site and the NADP-binding site,[45] respectively, these residues showed low total scores of multi-Harmony (Data set 4). Thus, the highly conserved amino acid residues among all ADHs are not regarded as subfamily-specific residues. As for Group 1 ADH, eight amino acid residues (Gly87, Cys88, Val90, Ser92, Ser165, Ala246, Ile294, and Tyr338) in YahK were selected to be subfamily-specific residues because these residues showed higher total scores than those of other residues (Figure 4A). On the other hand, eight amino acid residues (Gly90, Trp91, Ala93, Ser95, Lys159, Asn240, Thr287, and Tyr331) in Ahr were selected as subfamily-specific residues among Group 3 ADHs (Figure 4A).

According to MOE docking simulations with YahK and Ahr, it is observed that the subfamily-specific residues in Group 1

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

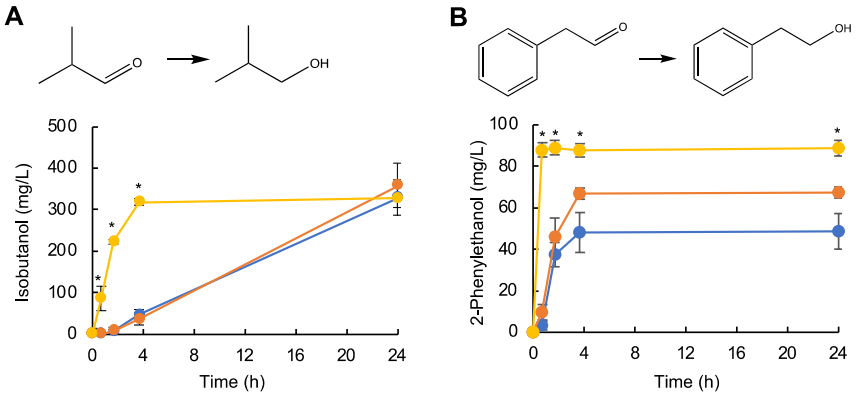


Figure 5. ADH-mediated bioconversion of isobutyraldehyde and phenylacetaldehyde. Each recombinant cell expressing Ahr (orange) or TptADH (yellow) or the cell with the control plasmid (blue) was incubated with 1 g/L of isobutyraldehyde and phenylacetaldehyde for 1, 2, 4, and 24 h, respectively. Extracellular concentrations of isobutanol (A) and 2-phenylethanol (B) were quantified by high-performance liquid chromatography (HPLC). All data are presented as means ± standard deviations (n = 3 independent biological experiments). Statistical significance of TptADH against Ahr and the control was determined using the Tukey–Kramer test (*p < 0.01).

and Group 3 are not always located near the ligands, except for Cys88 and Ile294 in YahK (Figure 4B) and Trp91 and Thr297 in Ahr (Figure 4C). To find the structural features that underlie catalytic efficiency, we compared catalytically active sites of isobutyraldehyde- or phenylacetaldehyde-docked model structures of YahK (Group 1) and Ahr (Group 3) (Figure S5). To validate the docking results, we performed an additional in silico docking simulation using AutoDock Vina.[46] The top-ranked binding poses from AutoDock Vina exhibited substrate RMSD values that were comparable to those obtained from the MOE docking simulations (2.385 Å for isobutyraldehyde-docked YahK; 1.905 Å for phenylacetaldehyde-docked YahK; 0.750 Å for isobutyraldehyde-docked Ahr; 0.957 Å for phenylacetaldehyde-docked Ahr) (Figure S5), thereby supporting the reliability of the MOE docking results. Distances among the Zn^{2+}-binding residues and Zn^{2+} ion of YahK or Ahr are almost comparable to those in the active site geometry of *Saccharomyces cerevisiae* ADH1.[47] The orientation of the side chain of substrates (isobutyraldehyde and phenylacetaldehyde) in the docked model structure varies between YahK and Ahr. The conformation of YahK Cys88, which corresponds to Trp91 of Ahr, probably affects the interaction environment of the Zn^{2+}-binding residues. On the other hand, Ahr Trp91 appears to play a crucial role in substrate binding, because Trp could form perpendicular anion-π stacking with the oxygen moiety of the aldehyde substrate.[48] YahK Ile294 forms a substrate pocket and would play a role in increasing hydrophobicity, which contributes to interaction with the side chain of substrates. On the other hand, the corresponding amino acid residue, Ahr Thr287, has slightly lower hydrophobicity than Ile. This difference may serve as a clue to explain the difference in substrate specificity. As observed in the catalytic efficiency of TguADH, the high K_m values for aromatic aldehydes and aromatic ketones suggest that Trp91 is essential for substrate binding. In conclusion, the subfamily-specific residues Cys88 and Ile294 in YahK and Trp91 and Thr287 in Ahr participate in the substrate coordination environment, indicating that these residues may influence the catalytic properties of ADH. As subfamily-specific residues, Gly87, Ser92, and Tyr338 are highly conserved in Group 1, Group 2, and Group 3 ADHs (Figure 4). The conservation of these amino acid residues likely reflects these characteristics, which are broad substrate specificity and high catalytic efficiency.

Bioconversion of Isobutyraldehyde and Phenylacetaldehyde to Isobutanol and 2-Phenylethanol. To evaluate the potential of TptADH for valuable alcohol production, bioconversions were performed using *E. coli* whole cell biocatalyst, expressing *TptADH* or *Ahr* as a benchmark.[49] Isobutanol, a short aliphatic alcohol, is a basic building block for coating resins and paint thinners and has been approved as a blending component for gasoline. Isobutanol is also considered an alternative fuel with an energy content that is approximately 82% of the energy content of gasoline.[50] 2-phenylethanol, an aromatic alcohol, is widely used as a fragrance for perfumes and cosmetics and as an organoleptic enhancer in food.[51] The cultivated recombinant cells (OD_{600} = 2.0) were incubated with 1 g/L isobutyraldehyde or phenylacetaldehyde in 50 mM phosphate buffer (pH 7.2) containing 5% (v/v) glycerol. Figure 5A shows the time courses of isobutanol production in culture supernatants of *E. coli* BL21(DE) with Ahr (Ahr cells), TptADH (TptADH cells), or control plasmid (CT). Because *E. coli* naturally expresses various kinds of ADHs, including Ahr and YahK,[52] some background aldehyde conversion is always observed. When using TptADH, isobutanol production peaked at 317 mg/L after 4 h of incubation (Figure 5A), probably due to the consumption of NADPH as reducing power. The productivity of TptADH cells was 10-fold greater than that of the other recombinant cells; however, the productivity of Ahr was almost comparable to that of CT cells. The expression level of TptADH was similar to that of Ahr throughout the incubation (Figure S6). Therefore, lower production of isobutanol in Ahr cells can be attributed to the low catalytic efficiency of Ahr toward isobutyraldehyde. The 2-phenylethanol production (90 mg/L) in TptADH cells peaked at 30 min incubation (Figure 5B). The productivity of 2-phenylethanol was 10 times higher than that of Ahr and CT cells. Production of 2-phenylethanol in Ahr and CT cells reached 70 and 50 mg/L, respectively. These results indicate that TptADH is useful for isobutyl alcohol and 2-phenylethanol production.

DISCUSSION

In the present study, 6727 nonredundant ADH candidates retrieved from UniProt were classified into 9 Groups based on

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

PCA-based clustering. Two ADHs from each of the nine groups (a total of 18 ADHs) were selected as representative sequences from each group (Figure 2). Using the laboratory automation system, we obtained activity data for 18 enzymes and 41 substrates (Figure 3B), as well as the catalytic efficiency of TptADH and TguADH toward various substrates (Table 1 and Figure S4A,B). TptADH and TguADH exhibited high affinities for short-aliphatic aldehydes and aliphatic and aromatic ketones, respectively. Although the reactions toward each substrate examined in this study are already known, we have identified a novel enzyme, TptADH, that demonstrates significantly higher catalytic efficiency toward isobutyraldehyde and phenylacetaldehyde than ADH, which has been previously reported as a benchmark.[53,54] Our laboratory automation system enables the acquisition of highly accurate data with high throughput (>2000 assays per day). The comprehensive acquisition of catalytic efficiency data for promiscuous enzymes toward diverse substrates led to the discovery of novel aspects of catalytic function, such as substrate inhibition and allosteric regulation. This automation system is expected to provide fundamental data for not only the findings of the catalytic function of promiscuous enzymes but also the development of diverse engineered enzymes and the construction of enzyme function prediction models.

In our PCA-based clustering, residues that are highly conserved or not conserved at all result in minimum and maximum possible variances, respectively. When residues or gaps are completely identical in a multiple sequence alignment, their positions do not significantly contribute to the variance of the PCA plot. Our method involves converting single letters in sequence alignment to numerical representations based on an original binary vector (Table S1). Additionally, sequence alignment gaps are assigned specific numerical representations, contributing to the variance. Consequently, sequences that share high homology, including gaps, are located close to each other on the PCA plot, whereas sequences with lower homology are positioned farther apart on the PCA plot (Figures 2 and S2). Therefore, this PCA-based approach, termed the MUSASHI method, could be applied to understand differences in sequence features within data sets. On the other hand, phylogenetic trees based on sequence alignment, illustrating lines of evolutionary descent of different proteins from a common ancestor, have traditionally been used to classify sequences.[55,56] In contrast to methods for inferring phylogeny, such as the maximum likelihood method[55] and distance-matrix methods,[56] the PCA-based approach disregards information about nodes and branch lengths, representing evolutionary points and distances in the phylogenetic tree, respectively. This allows for the feature visualization based on sequence conservation across data sets containing more than thousands of diverse protein sequences (Figure 2).

As a method for numerically representing enzyme sequences, the position-specific scoring matrix (PSSM) is widely used to capture the evolutionary information on protein sequences.[57,58] PSSM profiles are typically generated using the PSI-BLAST program to search the nonredundant database. However, when PSSMs are derived from a multiple sequence alignment (MSA) of 6727 ADH amino acid sequences retrieved from UniProt using keyword searches, the inclusion of sequences from different families can lead to dispersed conservation patterns, which may obscure meaningful features. In contrast, the MUSASHI method can extract features even from such an MSA of 6727 sequences, making it applicable to data sets that do not rely on PSI-BLAST. One limitation of the MUSASHI method, however, is the low separability observed in some clusters, which appears to result from constraints in the binary variables used in PCA. For example, Groups 4, 5, and 7 in Figure 3 were not clearly separated, and the SW indexes differ significantly (Figure 3C). To improve classification accuracy in future work, it is necessary to refine the clustering approach. Specifically, incorporating evaluation metrics, such as the Matthews Correlation Coefficient,[59,60] and optimizing the binary variables used in classification, might lead to more reliable clustering outcomes.

To collect sequence data sets, sequence similarity searches are effective for extracting proteins of interest with desired functions from protein databases, represented by successful heuristic methods such as NCBI-BLAST.[38,39] Importantly, the variance of the PCA score plot depends on the sequence data sets in multiple sequence alignment. Similarly to phylogenetic tree estimation with multiple sequence alignment, it is crucial to note that an alignment step accompanied by substantial errors can lead to biased results in the PCA-score plot. Local sequence alignment makes it particularly useful for identifying the most similar subsequences, such as functional domains and substrate-binding sequences, within a set of sequences.[61] A PCA based on the alignment of these subsequences might allow for more precise classification within each group. Our data showed that substrate specificities and activities vary among different ADHs within a group (Figure 3B). For example, it is evident that Ahr does not exhibit activity toward acetone, whereas other ADHs in Group 3 do (Figure 3B). Therefore, analyzing specific subsequences within a group might help further clarify functional differences within each group. The MUSASHI method, which can rapidly process large amounts of data, is expected to be an extremely useful clustering tool as genomic data continues to expand. The present enzyme selection workflow enables the identification of highly conserved residues within each group, allowing for the selection of groups with the desired properties without omission. We have identified subfamily specific residues in Group1 ADHs (Gly87, Cys88, Val90, Ser92, Ser165, Ala246, Ile294, and Tyr338 in YahK) and Group3 ADHs (Gly90, Trp91, Ala93, Ser95, Lys159, Asn240, Thr287, and Tyr331 in Ahr) (Figure 4A). These residues are highly conserved within each group. The functional roles of these residues in catalytic activity are still unknown; however, the protein groups associated with these specific residues would exhibit similar properties in terms of substrate specificity and activity.

NADPH-dependent primary and secondary ADHs are often employed in the production of isobutyl alcohol and 2-phenylethanol. Reported catalytic efficiencies (k_{cat}/K_m) of these ADHs are as follows: YqhD (0.90 s^{-1} mM^{-1}) and YugJ (6.12 s^{-1} mM^{-1}) for 2-phenylethanol production;[54] AdhA (0.8 s^{-1} mM^{-1}) and YqhD (0.7 s^{-1} mM^{-1}) for isobutanol production.[53] Additionally, NADPH-dependent ADHs from *Clostridium autoethanogenum* (k_{cat}/K_m = 86 s^{-1} mM^{-1})[62] or *Clostridium beijerinckii* (k_{cat}/K_m = 141 s^{-1} mM^{-1})[63,64] are overexpressed to produce isopropanol from acetone. The catalytic efficiencies of TptADH for these substrates were found to be higher than or comparable to those of previously reported enzymes; however, we cannot conclusively determine the predominance of TptADH due to differences in the assay conditions. Nonetheless, TptADH exhibited higher affinities for aliphatic short-chain and aromatic aldehydes compared to Ahr.[19] It is evident that the bioconversion efficiencies of

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

TptADH in isobutanol and 2-phenylethanol production are superior to those of Ahr, widely used for alcohol production, suggesting the utility of TptADH in producing various alcohols.

To utilize promiscuous enzymes for future metabolic engineering, it will be essential to perform enzyme engineering to ensure their activity, which is specific to the desired substrates. Promiscuous activity offers new opportunities for neofunctionalization during evolutionary adaptation without compromising function.[65,66] Enzymes exhibiting substrate promiscuity serve as starting points for enzyme engineering, harboring the potential to yield useful enzymes. Future studies should explore the possibilities and limitations of enzyme engineering based on promiscuous enzymes. Our enzyme selection approach using subfamily-specific residues as a clue will provide a foundational basis in the fields of enzyme engineering and metabolic engineering.

CONCLUSIONS

In summary, we demonstrated a strategy for identifying highly active and promiscuous ADHs from protein databases. To convert large-scale multiple sequence alignments into numerical representations that effectively capture multidimensional information, we applied binary vector profiling of sequence patterns to maximize variance in a mean-centered variance−covariance matrix used for PCA. This PCA was then used to project the multidimensional ADH sequence information onto two dimensions. Two representative sequences were selected from 9 groups formed via *k*-means clustering of the PCA plots based on SDT total scores. The enzyme selection process was termed the MUSASHI method. To evaluate substrate specificity and catalytic activity, a total of 18 ADHs were assayed against 41 commercially available substrates. Enzyme activities were measured using an automated enzyme assay system. To rank the ADH groups in terms of both broad substrate specificity and high activity, the Shannon−Wiener index was calculated for each of the 18 enzymes. Two promiscuous ADH groups were identified: Group 1 ADHs, which exhibited high activity toward short-chain and aromatic aldehydes, and Group 3 ADHs, which showed strong activity against aliphatic and aromatic ketones. In Group 1 ADHs, eight amino acid residues (Gly87, Cys88, Val90, Ser92, Ser165, Ala246, Ile294, and Tyr338) in YahK were identified as subfamily-specific based on high total scores in multi-Harmony analysis. Similarly, in Group 3 ADHs, eight residues (Gly90, Trp91, Ala93, Ser95, Lys159, Asn240, Thr287, and Tyr331) in Ahr were identified as subfamily-specific. In the production of isobutanol and 2-phenylethanol, TptADH from Group 1 showed higher productivity than a conventional ADH (Ahr), highlighting its potential as a promising biocatalyst for valuable alcohol production. The MUSASHI method enables high-throughput processing of large-scale sequence data and is expected to become a powerful tool for enzyme clustering as genomic databases continue to expand. Furthermore, this method facilitates the identification of highly conserved residues within each enzyme group, allowing for the efficient selection of enzymes with desired functional properties.

MATERIALS AND METHODS

Principal Component Analysis of Aligned ADH Sequences. ADH amino acid sequences were retrieved from UniProt using keyword searches, resulting in 13,992 and 3948 hits for EC 1.1.1.1, EC1.1.1.2, respectively. Multiple keywords, "NAD(P)-dependent alcohol dehydrogenase", "ADH_N", and "ADH_zinc_N", were used for the search by using a boolean operator of "AND". "ADH_N" and "ADH_zinc_N" represent short names of Pfam IDs PF08240 (the catalytic domain of alcohol dehydrogenases) and PF00107 (the cofactor-binding domain of zinc-containing alcohol dehydrogenases), respectively. A total of 17,940 sequences were applied to the cluster database at high identity with tolerance (CD-HIT)[32] to remove redundant sequences with a threshold value of 0.9, resulting in 6727 nonredundant sequences. The 6727 amino acid sequences were aligned using multiple alignment using fast Fourier transform (MAFFT) [Strategy, FFT-NS-1; Scoring matrix, BLOSUM62; Gap opening penalty, 1.53; Offset value, 0.0].[33,34] To maximize variance in a mean-centered variance/covariance matrix used for principal component analysis (PCA), a well-known tool in multivariate statistics, the 20 amino acid residues and gaps, marked by a hyphen, in the sequence alignment are represented as a binary vector matrix (5 rows × 21 columns), as shown in Table S1. After applying sequence alignment to the matrix, PCA was performed to reduce the data dimension. *k*-means clustering, which is a commonly used data clustering for unsupervised learning tasks,[35] was performed to analyze the resulting PCA scores plot. The elbow method was employed to calculate the optimal cluster.[36] Original PCA and *k*-means, and Elbow programs were written in Python (http://www.python.org). The pairwise identity score between sequences (in a FASTA format) in each group was calculated with the sequence demarcation tool (SDT) version 1.3 in Python.[37] The total identity score was calculated by summing each pairwise identity score. The top five sequences with the highest total score were selected as representative sequences within each group.

Cloning, Protein Production, and Protein Purification. *Escherichia coli* strains used in this work are as follows: *E. coli* DH5α (Takara Bio, Tokyo, Japan) and BL21(DE3) (Agilent Technologies, Santa Clara, CA). The UniProt accession numbers for the selected ADH sequences with organisms through SDT calculations are as follows: Strli-ADH from *Streptomyces lividans* 1326 (A0A7U9DXW2), unArt-ADH from uncultured *Arthrobacter sp.* (A0A6J4JFR6), Fusox-ADH from *Fusarium oxysporum* (A0A0D2XEJ8), Sclsc-ADH from *Sclerotinia sclerotiorum* (A7EP86), Ahr from *E. coli* K12 (P27250), and PseTHAF3-Ahr from *Pseudoalteromonas sp.* THAF3 (A0A5P9J280), GemSHPL17-AdhT2 from *Gemmata sp.* HPL17 (A0A142XHS6), Palbo-AdhT from *Paludisphaera borealis* (A0A1U7CM70), HR40-AdhA from bacterium HR40 (A0A2H6AXG7), Allbo-ADH from *Allorhizobium borbori*, (A0A7W6NZU6), Pseps-AdhP from *Pseudomonas oryzihabitans* (A0A1G5PE21), SphTF3-AdhP from *Sphingomonas sp.* TF3 (A0A432VF00), and StrF1-AdhT2 from *Streptomyces sp.* F-1 (A0A1K2FT97), and SulTHAF37-ADH from *Sulfitobacter sp.* THAF37 (A0A5P9FAJ9), Gorin-ADH from *Gordonia insulae* (A0A3G8JR04), Nocno-ADH from *Nocardia nova* (A0A2S6A1U9), Ictpu-ADH from *Ictalurus punctatus* (Channel catfish) (A0A2D0RJM7), MicrhADH from *Microbulbifer rhizosphaerae* (A0A7W4W956), Strsc-AdhC2 from *Streptomyces scabiei* (A0A124C5I1), YahK from *E. coli* K12 (P75691), TptADH from *Tatumella ptyseos* ATCC 33301 (A0A085JKH3), Phyma-ADH from *Phytobacter massiliensis* (A0A6N3F1L1), and SphSMR4y-ADH from *Sphingorhabdus sp.* SMR4y (A0A220WB65), and TguADH from *Trabulsiella*

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

guamensis ATCC 49490 (A0A084ZMM2). The ADH genes were synthesized and cloned into the multicloning site of pET28a (Novagen) encoding an N-terminal His_6-tag by Gene Synthesis Service from GenScript Biotech Corporation (Piscataway, NJ) with codon optimization for *E. coli*. DH5α was used as a host for plasmid amplification.

Each recombinant ADH was produced in *E. coli* BL21 (DE3) cells, which were grown in Luria−Bertani (LB) medium at 37 °C for 3 h. The overnight precultured cells were inoculated into LB medium with 100 μg mL^{-1} ampicillin and cultivated at 37 °C until the OD_{600} value reached 0.5. Protein expression was induced by adding 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG), and the induced cells were cultivated overnight at 18 °C. After harvesting, cells were suspended in buffer A (20 mM Tris-HCl (pH 8.0), 100 mM NaCl, 5 mM Imidazole) and disrupted using sonication. The cell extract was centrifuged at 23,000 × *g* for 60 min. The resulting supernatant was applied to a 5 mL TALON metal affinity resin (TAKARA Bio) and eluted with a stepwise imidazole concentration gradient (0−500 mM) in buffer A. The collected fraction was applied to a PD-10 desalting column (Cytiva, Marlborough, MA) equilibrated with buffer A. The concentration of purified ADH was measured using the Quick Start Bradford Protein Assay (Bio-Rad Laboratories, Inc., Hercules, CA), and purity was confirmed by SDS-PAGE. The purified proteins were stored as 50% glycerol stocks at −30 °C until use.

Automation of ADH Enzymatic Assay. A general-purpose robotic system was constructed to automate the microplate reader-based enzyme assay. This automated system features a SCARA (Selective and Compliance Articulated Robot Arm)-type arm, the ThermoFisher Spinnaker (ThermoFisher Scientific, Waltham, MA), which transfers microplates among three instruments installed on the platform. The robotic system for the enzyme assay consists of a Beckman Biomek i5 liquid handling workstation (Beckman Coulter, Brea, CA), a Thermo Fisher Multidrop Combi reagent dispenser (ThermoFisher Scientific), and a Tecan Spark multimode microplate reader (Tecan, Männedorf, Switzerland).

For the enzyme assay, reaction premixtures containing 50 mM MOPS (pH 7.0), 0.1 mM NAD(P)H, and a substrate at an appropriate concentration (without enzyme) were incubated at 37 °C for 10 min. UV-transparent 96-well plates (Funakoshi Co., Ltd., Tokyo, Japan) were used for the enzyme assay. The automated process for the enzyme assay included plate transfer, enzyme and reaction premixture solution distribution, and absorbance monitoring, scheduled with ThermoFisher Momentum software. The programmed workflow was as follows: A 96-well microplate from a microplate stacker was transferred to the stage of the Biomek i5. Then, 20 μL of each enzyme solution (at an appropriate concentration of each enzyme stock solution) was dispensed into each well of the 96-well microplate according to the programmed pipetting protocol of Beckman Biomek software. Afterward, the microplate on the Biomek i5 stage was transferred to the Multidrop Combi reagent dispenser, where 180 μL of the reaction premixture was dispensed into all wells. Finally, the microplate was transferred to the Tecan Spark multimode microplate reader, and the decrease in absorbance at 340 nm, corresponding to the oxidation of NAD(P)H, was immediately monitored at 37 °C for 10 min with a 30 s interval. In the case of alcohol oxidation reactions, NAD(P)$^+$ was added to the pre-reaction mixture, and the increase in absorbance at 340 nm accompanying the production of NAD(P)H was monitored.

For the analysis of catalytic efficiency of TptADH or TguADH, reaction premixtures containing 50 mM MOPS (pH 7.0), 0.1 mM NAD(P)H, and either TptADH or TguADH at an appropriate concentration (without substrate) were incubated at 37 °C for 10 min. The substrate solution (20 μL) was dispensed into each well of a 96-well microplate with a serial dilution by Biomek i5, according to the programmed pipetting protocol of the Beckman Biomek software. The plate transfer and absorbance monitoring were followed by the automated system described above. The initial rates of the enzyme reaction were measured by varying the concentration of one substrate while keeping 100 μM NADPH.

The data were fitted to eqs 1 and 2, which are the standard equations for substrate inhibition and allosteric regulation, respectively.

$$v = \frac{V_{max}[S]}{K_m + [S]\left(1 + \frac{[S]}{K_i}\right)} \quad (1)$$

$$v = \frac{V_{max}[S]^n}{K_{0.5}{}^n + [S]^n} \quad (2)$$

One unit of enzymatic activity is defined as the amount of enzyme that consumes 1 μmol of a substrate per min. Kinetic parameters were determined by monitoring the enzymatic activities with varying substrate concentrations. The kinetic parameters and Hill coefficient were calculated from nonlinear least-squares fits of the data with Kaleida Graph software (Adelbeck Software, Reading, PA). Kinetic measurements were performed in triplicate as independent technical replicates. Enzymatic assay reagents were obtained from Nacalai Tesque, Inc. (Kyoto, Japan), Tokyo Chemical Industry Co., Ltd. (Tokyo, Japan), and FUJIFILM Wako Chemicals Co. (Osaka, Japan).

Shannon−Wiener Index Calculation. The Shannon−Wiener index was calculated according to eq 3. where r is the number of substrates (r = 41), n_i is the specific activity (i), and N is the sum of specific activities of an enzyme for all substrates.

$$H = -\sum_{i=1}^{r} Pi \times \ln Pi,\ Pi = \frac{n_i}{N} \quad (3)$$

ADH Structure Modeling. Two structure models (YahK in Group 1 ADH and Ahr in Group 3 ADH) were constructed using Molecular Operating Environment 2022 (MOE) (MOLSIS Inc., Tokyo, Japan).[67] The modeled structures with cofactors (NADPH and Zn^{2+} ion) and substrates (isobutyraldehyde and phenylacetaldehyde) were built based on the conformations of cofactor and substrate in *Saccharomyces cerevisiae* ADH1 (SceADH1) as a reference (PDB ID: 4W6Z).[47] Small molecules were docked in the apoprotein structures (YahK, PDB ID: 1UUF and Ahr, PDB ID: 7BU2). To preserve spatial geometry relevant to catalytic activity, distance and angle constraints were applied using the MOE's "Constraint Builder" tool. A distance constraint (2.0−3.0 Å) was set between the Zn^{2+} ion, the oxygen atom of the substrate's formyl group, and the nicotinamide ring of NADPH to mimic metal coordination in SceADH1. An angle constraint was introduced between the nicotinamide C4 atom, the formyl carbon of the substrate, and its oxygen atom to reproduce the

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

hydride transfer geometry observed in SceADH1. After the constraints were set, energy minimization was performed with the Amber10:EHT force field. Only docking poses that satisfied the constraints throughout the simulation were retained for further analysis. The two models were protonated using the Protonate3D tool in MOE at pH 7.0 and a temperature of 300 K. Subsequently, energy minimization was performed using the AMBER10:EHT force field (gradient = 0.01 RMS kcal mol^{-1} $Å^{-2}$). For docking simulation, the force field of AMBER10:EHT and the implicit solvation model of the reaction field (R-field 1:80; cutoff 8, 10) were selected. The docking simulations were performed using the general dock tool in MOE with the following settings: site, ligand atoms; ligand, aldehydes; placement, triangle Matcher method with London ΔG scoring; refinement, induced Fit with GBVI/WSA ΔG scoring. Specifically, the following residues, located within 6 Å of the substrate, were set as flexible by using the "Induced Fit" option in MOE: Ser40, Asp41, Leu42, Ile62, Val88, Cys105, Ser114, Thr161, Tyr162, His186, Ile292, and Val339 in YahK; and Cys41, Ser43, Trp52, His63, Glu64, Trp91, Thr92, Ile108, Leu151, Cys152, Thr156, Thr287, and Arg332 in Ahr.

To validate the MOE docking results, molecular docking was also performed using AutoDock Vina.[46] Ligands were treated as fully flexible, and selected receptor side chains within 6 Å of the substrate were designated as flexible using AutoDockTools (ver. 1.5.7). The flexible residues were the same as those defined using the "Induced Fit" option in the MOE. The receptor was divided into rigid and flexible parts accordingly. The docking grid box was centered on the active site with the following dimensions and coordinates: for YahK with isobutyraldehyde or phenylacetaldehyde, 15 × 15 × 15 Å (center: x = 89.715, y = −13.807, z = 46.829); for Ahr with isobutyraldehyde, 30 × 30 × 30 Å (center: x = −18.087, y = 35.855, z = 19.714); and for Ahr with phenylacetaldehyde, 25 × 25 × 25 Å (center: x = −21.717, y = 34.791, z = 18.238). Docking was conducted with the following exhaustiveness settings: YahK_isobutyraldehyde, 20; YahK_phenylacetaldehyde, 100; Ahr_isobutyraldehyde, 8; Ahr_phenylacetaldehyde, 8. The binding pose with the lowest predicted binding free energy was selected for further analysis. The substrate RMSD was calculated by PyMOL (The PyMOL Molecular Graphics System, Version 2.0, Schrödinger, LLC). Visualization of structures and calculation of distances between atoms in ADH active sites were performed using UCSF ChimeraX.[68]

Cell-Based Production of 2-Phenylethanol and Isobutanol. BL21(DE3) was separately transformed with Ahr in pET-Ahr and TptADH in pET-TptADH. For alcohol production, each strain was initially grown at 37 °C in 40 mL of LB supplemented with 100 μg/mL ampicillin in 50 mL conical plastic tubes (Falcon, Corning, MA). After reaching OD_{600} = 0.5, IPTG was added to a final concentration of 1.0 mM, and then, the temperature was lowered to 20 °C. After 15 h of cultivation, cells were collected by centrifugation at 4000 g and washed with a 0.85% NaCl solution. The pellet was resuspended in 50 mM phosphate buffer (pH 7.2) containing 5% (v/v) glycerol to a final OD_{600} = 2.0. Isobutyraldehyde or phenylacetaldehyde was added to 4 mL of the cell suspension at a final concentration of 1.0 g/L and incubated at 20 °C in an anaerobic chamber (COY Laboratory Products, MI) with <1 ppm of O_2. A 500 μL aliquot of the culture solution was collected at each sampling point (1, 2, 4, and 24 h) of the incubation and stored at −80 °C.

HPLC Analysis of 2-Phenylethanol and Isobutanol. Cell suspensions were centrifuged at 23,000 × g for 10 min, and the resulting supernatant was collected. The mixture was centrifuged, and the supernatant was filtered with a 0.45-μm-pore-size Millex-LH filter (Merck-Millipore, MA, Burlington). Each supernatant (10 μL) was analyzed by HPLC (HPLC prominence liquid chromatograph system; Shimadzu, Kyoto, Japan) on a Shim-Pack SCR-101P column (7.9 mm [inner diameter] by 300 mm; Shimadzu). The column was equilibrated with ultrapure water, produced by a Milli-Q EQ 7000 ultrapure water system (Merck-Millipore), at a flow rate of 0.6 mL min^{-1} at 80 °C. The eluted alcohols were monitored with a refractive index detector (RID-20A; Shimadzu, Kyoto, Japan).

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscatal.5c02764.

k-means clustering of the principal-component analysis score plot; percent identity matrix; SDS-PAGE analysis of the purified alcohol dehydrogenase; Michaelis−Menten kinetics; molecular docking; and expression level of recombinant alcohol dehydrogenase (PDF)

PCA of the ADH dataset; sequence demarcation tool (SDT) total scores of ADHs within each group; final concentration (mM) of substrates in each reaction mixture; and multi-harmony analysis of the ADH dataset (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Tomohisa Hasunuma − *Engineering Biology Research Center and Graduate School of Science, Innovation and Technology, Kobe University, Kobe 657-8501, Japan; Research Center for Sustainable Resource Science, RIKEN, Yokohama, Kanagawa 230-0045, Japan;* orcid.org/0000-0002-8382-2362; Email: hasunuma@port.kobe-u.ac.jp

Authors

Ryota Hidese − *Engineering Biology Research Center, Kobe University, Kobe 657-8501, Japan*

Kanae Sakai − *Graduate School of Science, Innovation and Technology, Kobe University, Kobe 657-8501, Japan*

Musashi Takenaka − *Graduate School of Science, Innovation and Technology, Kobe University, Kobe 657-8501, Japan*

Keiji Fushimi − *Engineering Biology Research Center, Kobe University, Kobe 657-8501, Japan*

Hisashi Kudo − *Engineering Biology Research Center, Kobe University, Kobe 657-8501, Japan*

Kenya Tanaka − *Graduate School of Science, Innovation and Technology, Kobe University, Kobe 657-8501, Japan;* orcid.org/0000-0002-6506-3788

Ryo Nasuno − *Graduate School of Science, Innovation and Technology, Kobe University, Kobe 657-8501, Japan;* orcid.org/0000-0002-4456-2801

Christopher J. Vavricka − *Department of Biotechnology and Life Science, Graduate School of Engineering, Tokyo University of Agriculture and Technology, Tokyo 184-8588, Japan;* orcid.org/0000-0003-2101-4359

Akihiko Kondo − *Engineering Biology Research Center and Graduate School of Science, Innovation and Technology,*

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943

Kobe University, Kobe 657-8501, Japan; Research Center for Sustainable Resource Science, RIKEN, Yokohama, Kanagawa 230-0045, Japan; orcid.org/0000-0003-1527-5288

Complete contact information is available at: https://pubs.acs.org/10.1021/acscatal.5c02764

Author Contributions

R.H. and K.S. contributed equally. R.H., K.S., A.K., and T.H. conceived and designed the research. R.H., K.S., and K.T. performed experiments. R.H., K.S., M.T., H.K., and K.F. performed bioinformatic analysis. R.H., K.S., H.K., R.N., and K.T. analyzed data. R.H. and K.S. wrote the original draft. C.J.V. and T.H. reviewed and edited the manuscript. All authors approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Ms. Miyuki Yamaguchi for her technical assistance. This work was supported by NEDO Projects P16009 (Development of production techniques for highly functional biomaterials using plant and other organism smart cells) and P20011 (Development of bioderived product production technology that accelerates the realization of carbon recycling). It was also supported by the Program for Forming Japan's Peak Research Universities (J-PEAKS) from the Japan Society for the Promotion of Science (JSPS).

■ REFERENCES

(1) Hasunuma, T.; Okazaki, F.; Okai, N.; Hara, K. Y.; Ishii, J.; Kondo, A. A review of enzymes and microbes for lignocellulosic biorefinery and the possibility of their application to consolidated bioprocessing technology. *Bioresour. Technol.* **2013**, *135*, 513−522.

(2) Becker, J.; Wittmann, C. Advanced biotechnology: metabolically engineered cells for the bio-based production of chemicals and fuels, materials, and health-care products. *Angew. Chem., Int. Ed. Engl.* **2015**, *54*, 3328−3350.

(3) Vavricka, C. J.; Hasunuma, T.; Kondo, A. Dynamic Metabolomics for Engineering Biology: Accelerating Learning Cycles for Bioproduction. *Trends Biotechnol.* **2020**, *38*, 68−82.

(4) Liu, Y.; Nielsen, J. Recent trends in metabolic engineering of microbial chemical factories. *Curr. Opin. Biotechnol.* **2019**, *60*, 188−197.

(5) Volk, M. J.; Tran, V. G.; Tan, S. I.; Mishra, S.; Fatma, Z.; Boob, A.; Li, H.; Xue, P.; Martin, T. A.; Zhao, H. Metabolic Engineering: Methodologies and Applications. *Chem. Rev.* **2023**, *123*, 5521−5570.

(6) Yuan, Y.; Shi, C.; Zhao, H. Machine Learning-Enabled Genome Mining and Bioactivity Prediction of Natural Products. *ACS Synth. Biol.* **2023**, *12*, 2650−2662.

(7) Nishida, K.; Kondo, A. CRISPR-derived genome editing technologies for metabolic engineering. *Metab. Eng.* **2021**, *63*, 141−147.

(8) Tanaka, K.; Bamba, T.; Kondo, A.; Hasunuma, T. Metabolomics-based development of bioproduction processes toward industrial-scale production. *Curr. Opin. Biotechnol.* **2024**, *85*, No. 103057.

(9) Wiltschi, B.; et al. Enzymes revolutionize the bioproduction of value-added compounds: From enzyme discovery to special applications. *Biotechnol. Adv.* **2020**, *40*, No. 107520.

(10) Intasian, P.; Prakinee, K.; Phintha, A.; Trisrivirat, D.; Weeranoppanant, N.; Wongnate, T.; Chaiyen, P. Enzymes, *In Vivo* Biocatalysis, and Metabolic Engineering for Enabling a Circular Economy and Sustainability. *Chem. Rev.* **2021**, *121*, 10367−10451.

(11) Orsi, E.; Claassens, N. J.; Nikel, P. I.; Lindner, S. N. Optimizing microbial networks through metabolic bypasses. *Biotechnol. Adv.* **2022**, *60*, No. 108035.

(12) UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D523−D531.

(13) Jumper.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

(14) Kroll, A.; Ranjan, S.; Engqvist, M. K. M.; Lercher, M. J. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nat. Commun.* **2023**, *14*, 2787.

(15) Khersonsky, O.; Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **2010**, *79*, 471−505.

(16) Babtie, A.; Tokuriki, N.; Hollfelder, F. What makes an enzyme promiscuous? *Curr. Opin. Chem. Biol.* **2010**, *14*, 200−207.

(17) Reid, M. F.; Fewson, C. A. Molecular characterization of microbial alcohol dehydrogenases. *Crit. Rev. Microbiol.* **1994**, *20*, 13−56.

(18) de Smidt, O.; du Preez, J. C.; Albertyn, J. The alcohol dehydrogenases of *Saccharomyces cerevisiae*: a comprehensive review. *FEMS Yeast Res.* **2008**, *8*, 967−978.

(19) Pick, A.; Rühmann, B.; Schmid, J.; Sieber, V. Novel CAD-like enzymes from *Escherichia coli* K-12 as additional tools in chemical production. *Appl. Microbiol. Biotechnol.* **2013**, *97*, 5815−5824.

(20) Martin, J. P.; Rasor, B. J.; DeBonis, J.; Karim, A. S.; Jewett, M. C.; Tyo, K. E. J.; Broadbelt, L. J. A dynamic kinetic model captures cell-free metabolism for improved butanol production. *Metab. Eng.* **2023**, *76*, 133−145.

(21) Gupta, M.; Wong, M.; Jawed, K.; Gedeon, K.; Barrett, H.; Bassalo, M.; Morrison, C.; Eqbal, D.; Yazdani, S. S.; Gill, R. T.; Huang, J.; Douaisi, M.; Dordick, J.; Belfort, G.; Koffas, M. A. G. Isobutanol production by combined *in vivo* and *in vitro* metabolic engineering. *Metab. Eng. Commun.* **2022**, *15*, No. e00210.

(22) Kondo, T.; Tezuka, H.; Ishii, J.; Matsuda, F.; Ogino, C.; Kondo, A. Genetic engineering to enhance the Ehrlich pathway and alter carbon flux for increased isobutanol production from glucose by *Saccharomyces cerevisiae*. *J. Biotechnol.* **2012**, *159*, 32−37.

(23) Zhong, W.; Zhang, Y.; Wu, W.; Liu, D.; Chen, Z. Metabolic Engineering of a Homoserine-Derived Non-Natural Pathway for the De Novo Production of 1,3-Propanediol from Glucose. *ACS Synth. Biol.* **2019**, *8*, 587−595.

(24) Hassing, E. J.; de Groot, P. A.; Marquenie, V. R.; Pronk, J. T.; Daran, J. G. Connecting central carbon and aromatic amino acid metabolisms to improve de novo 2-phenylethanol production in *Saccharomyces cerevisiae*. *Metab. Eng.* **2019**, *56*, 165−180.

(25) Casari, G.; Sander, C.; Valencia, A. A method to predict functional residues in proteins. *Nat. Struct. Biol.* **1995**, *2*, 171−178.

(26) Atchley, W.; Zhao, J.; Fernandes, A.; Druke, T. Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6395−6400.

(27) Wallace, I.; Higgins, D. Supervised multivariate analysis of sequence groups to identify specificity determining residues. *BMC Bioinf.* **2007**, *8*, 135.

(28) Wang, B.; Kennedy, M. A. Principal components analysis of protein sequence clusters. *J. Struct. Funct. Genomics.* **2014**, *15*, 1−11.

(29) Shafee, T.; Anderson, M. A. A quantitative map of protein sequence space for the cis-defensin superfamily. *Bioinformatics.* **2019**, *35*, 743−752.

(30) Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. A Math Phys. Eng. Sci.* **2016**, *374*, 20150202.

(31) Pearson, W. R. An introduction to sequence similarity ("Homology") searching. *Curr. Protoc. Bioinf.* **2013**, *Chapter 3*, 3.1.1−3.1.8.

(32) Li, W.; Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* **2006**, *22*, 1658−1659.

(33) Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **2002**, *30*, 3059−3066.
(34) Katoh, K.; Rozewicki, J.; Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* **2019**, *20*, 1160−1166.
(35) Hartigan, J. A.; Wong, M. A. A *K*-Means Clustering Algorithm. *J. R. Stat. Soc., C: Appl. Stat.* **1979**, *28*, 100−108.
(36) Thorndike, R. L. Who belongs in the family? *Psychometrika.* **1953**, *18*, 267−276.
(37) Muhire, B. M.; Varsani, A.; Martin, D. P. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PLoS One.* **2014**, *9*, No. e108277.
(38) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(39) Johnson, M.; Zaretskaya, I.; Raytselis, Y.; Merezhuk, Y.; McGinnis, S.; Madden, T. L. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **2008**, *36*, W5−W9.
(40) Sun, H. W.; Plapp, B. V. Progressive sequence alignment and molecular evolution of the Zn-containing alcohol dehydrogenase family. *J. Mol. Evol.* **1992**, *34*, 522−535.
(41) Barnes, B. V.; Zak, D. R.; Denton, S. R.; Spurr, S. H. *Forest ecology*, 4th ed.; John Wiley and Sons, Inc: NJ, 1998; 773 p.
(42) Liu, J.-Y.; Zheng, G.-W.; Li, C.-X.; Yu, H.-L.; Pan, J.; Xu, J.-H. Multi-substrate fingerprinting of esterolytic enzymes with a group of acetylated alcohols and statistic analysis of substrate spectrum. *J. Mol. Catal., B Enzym.* **2013**, *89*, 41−47.
(43) Ye, K.; Feenstra, K. A.; Heringa, J.; Ijzerman, A. P.; Marchiori, E. Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting. *Bioinformatics* **2008**, *24*, 18−25.
(44) Brandt, B. W.; Feenstra, K. A.; Heringa, J. Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res.* **2010**, *38*, W35−40.
(45) Persson, B.; Hallborn, J.; Walfridsson, M.; Hahn-Hägerdal, B.; Keränen, S.; Penttilä, M.; Jörnvall, H. Dual relationships of xylitol and alcohol dehydrogenases in families of two protein types. *FEBS Lett.* **1993**, *324*, 9−14.
(46) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455−461.
(47) Raj, S. B.; Ramaswamy, S.; Plapp, B. V. Yeast alcohol dehydrogenase structure and catalysis. *Biochemistry.* **2014**, *53*, 5791−5803.
(48) Nguyen, G. T.; Kim, Y. G.; Ahn, J. W.; Chang, J. H. Structural Basis for Broad Substrate Selectivity of Alcohol Dehydrogenase YjgB from *Escherichia coli*. *Molecules.* **2020**, *25*, 2404.
(49) Guo, D.; Zhang, L.; Kong, S.; Liu, Z.; Li, X.; Pan, H. Metabolic Engineering of *Escherichia coli* for Production of 2-Phenylethanol and 2-Phenylethyl Acetate from Glucose. *J. Agric. Food Chem.* **2018**, *66*, 5886−5891.
(50) Yusoff, M. N. A. M.; Zulkifli, N. W. M.; Masum, B. M.; Masjuki, H. H. Feasibility of bioethanol and biobutanol as transportation fuel in spark-ignition engine: A review. *RSC Adv.* **2015**, *5*, 100184−100211.
(51) Scognamiglio, J.; Jones, L.; Letizia, C. S.; Api, A. M. Fragrance material review on phenylethyl alcohol. *Food Chem. Toxicol.* **2012**, *50*, S224−S239.
(52) Pugh, S.; McKenna, R.; Halloum, I.; Nielsen, D. R. Engineering *Escherichia coli* for renewable benzyl alcohol production. *Metab. Eng. Commun.* **2015**, *2*, 39−45.
(53) Atsumi, S.; Wu, T. Y.; Eckl, E. M.; Hawkins, S. D.; Buelter, T.; Liao, J. C. Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes. *Appl. Microbiol. Biotechnol.* **2010**, *85*, 651−657.
(54) Zhan, Y.; Shi, J.; Xiao, Y.; Zhou, F.; Wang, H.; Xu, H.; Li, Z.; Yang, S.; Cai, D.; Chen, S. Multilevel metabolic engineering of *Bacillus licheniformis* for de novo biosynthesis of 2-phenylethanol. *Metab. Eng.* **2022**, *70*, 43−54.
(55) Huelsenbeck, J. P.; Crandall, K. A. Phylogeny Estimation and Hypothesis Testing Using Maximum Likelihood. *Annu. Rev. Ecol. Evol. Syst.* **1997**, *28*, 437−466.
(56) Yang, Z. *Computational Molecular Evolution*; Oxford University Press: Oxford, 2006, p 376.
(57) Xu, R.; Zhou, J.; Wang, H.; He, Y.; Wang, X.; Liu, B. Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation. *BMC Syst. Biol.* **2015**, *9* (Suppl 1), S10.
(58) Ruan, X.; Xia, S.; Li, S.; Su, Z.; Yang, J. Hybrid framework for membrane protein type prediction based on the PSSM. *Sci. Rep.* **2024**, *14* (1), 17156.
(59) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.
(60) Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, *21*, 6.
(61) Smith, T. F.; Waterman, M. S. The identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195−197.
(62) Köpke, M.; Gerth, M. L.; Maddock, D. J.; Mueller, A. P.; Liew, F.; Simpson, S. D.; Patrick, W. M. Reconstruction of an acetogenic 2,3-butanediol pathway involving a novel NADPH-dependent primary-secondary alcohol dehydrogenase. *Appl. Environ. Microbiol.* **2014**, *80*, 3394−3403.
(63) Ismaiel, A. A.; Zhu, C. X.; Colby, G. D.; Chen, J. S. Purification and characterization of a primary-secondary alcohol dehydrogenase from two strains of *Clostridium beijerinckii*. *J. Bacteriol.* **1993**, *175*, 5097−5105.
(64) Hanai, T.; Atsumi, S.; Liao, J. C. Engineered synthetic pathway for isopropanol production in *Escherichia coli*. *Appl. Environ. Microbiol.* **2007**, *73*, 7814−7818.
(65) Miton, C. M.; Jonas, S.; Fischer, G.; Duarte, F.; Mohamed, M. F.; van Loo, B.; Kintses, B.; Kamerlin, S. C. L.; Tokuriki, N.; Hyvönen, M.; Hollfelder, F. Evolutionary repurposing of a sulfatase: A new Michaelis complex leads to efficient transition state charge offset. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E7293−E7302.
(66) Baier, F.; Hong, N.; Yang, G.; Pabis, A.; Miton, C. M.; Barrozo, A.; Carr, P. D.; Kamerlin, S. C.; Jackson, C. J.; Tokuriki, N. Cryptic genetic variation shapes the adaptive evolutionary potential of enzymes. *Elife.* **2019**, *8*, No. e40789.
(67) *Molecular Operating Environment (MOE)*; 2022.02 Chemical Computing Group ULC: 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7, Canada, 2023.
(68) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Meng, E. C.; Couch, G. S.; Croll, T. I.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **2021**, *30*, 70−82.

https://doi.org/10.1021/acscatal.5c02764
ACS Catal. 2025, 15, 11931−11943