Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine learning-guided dual optimization of the substrate specificity and thermostability of yeast alcohol dehydrogenase

Yuki Ogawa
RIKEN
Akihiko Kondo
RIKEN
Tomokazu Shirai
tomokazu.shirai@riken.jp

RIKEN

Research Article

Keywords: Machine-learning, Yeast alcohol dehydrogenase, Thermostability, Enzyme engineering, Substrate specificity

Posted Date: June 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9708759/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Yeast alcohol dehydrogenase I (YADH) catalyzes the reversible reduction of aldehydes to primary alcohols. Thermostable YADH variants with broad substrate specificities at elevated temperatures are desirable for industrial biocatalysis. In this study, we aimed to alter the substrate specificity of YADH to efficiently catalyze the reduction of the non-native substrate, *trans*-2-nonenal, to *trans*-2-nonenol, while simultaneously enhancing thermostability. Residues around the substrate-binding site were identified by docking simulation and subjected to saturation and random mutagenesis. Measured activity and thermostability were used to train a machine-learning model to predict YADH variants with improved activity and thermostability.

Results

The machine learning-guided approach identified variants with improved properties. Four variants showed increased specific activity toward *trans*-2-nonenal compared with the wild type, and one variant also exhibited improved thermostability. The best-performing variant showed a more than twofold increase in specific activity along with increased thermal resistance. Variants that lost activity toward the native substrate acetaldehyde while gaining higher activity toward *trans*-2-nonenal were also obtained, indicating altered substrate specificity. These results demonstrate efficient identification of beneficial mutations from a limited dataset.

Conclusions

Overall, this study presents a machine learning–guided framework that integrates docking and multi-model evaluation for simultaneously optimizing substrate specificity and thermostability in YADH.

Background

Enzymes are highly specific biocatalysts widely applied in the food, animal feed, pharmaceutical, medical application, bioremediation, biofuel, and detergent industries [1]. However, the catalytic properties of natural enzymes, such as stability, activity, and substrate specificity, often require optimization to meet industrial demands [1]. Various protein engineering strategies have thus been developed to obtain the desired enzymes.

Enzyme thermostability is crucial from an industrial perspective because improved stability allows processes to be conducted at high temperatures [2]. However, natural enzymes typically exhibit poor thermostability [3], and enhanced stability is frequently accompanied by reduced activity [4, 5].

Importantly, appropriate mutations in or near the active pocket and substrate-binding regions can simultaneously improve both thermostability and activity in industrial enzymes [3].

In recent years, enzyme engineering has increasingly incorporated machine learning (ML) to explore sequence–function relationships. ML models trained on experimental data can efficiently identify beneficial mutations and reduce screening efforts [6–11].

Although many studies have focused on optimizing single properties such as activity or stability, efforts toward multi-objective optimization are emerging. Simultaneous improvement of properties such as activity and stability has been explored using composite objectives or strategies accounting for trade-offs between objectives [12–14]. However, the general design principles of constructing ML models for multi-objective optimization are still evolving, and key aspects such as objective formulation and balancing remain insufficiently explored. From an application perspective, simultaneously improving multiple enzyme properties is often desirable, particularly for enzymes required to function under non-native conditions. Although traditional approaches such as directed evolution have been successful, extending them to multiple objectives typically requires additional experimental effort [15].

Machine learning (ML) provides a promising framework to address this challenge by integrating multiple properties into a unified optimization process. Therefore, in this study, we applied an ML-based approach to simultaneously optimize enzymatic activity and thermostability. We further examined strategies for constructing predictive models for multi-objective optimization, aiming to contribute to the development of general design principles in ML-guided enzyme engineering.

Yeast alcohol dehydrogenase I (YADH) is an NAD+-dependent enzyme catalyzing the reversible reduction of aldehydes to their corresponding primary alcohols [16]. It is commercially available and highly active with ethanol as a substrate [17]; its activity declines with increasing carbon chain length of the alcohol [18]. YADH is a tetramer composed of four identical subunits, each containing 347 amino acid residues [19]. Its kinetic, chemical, and physical properties have been extensively studied [16, 18–22]. Despite its favorable catalytic properties, its industrial application is limited by poor stability [23–25]. Enhancing its robustness could make YADH a promising candidate for diverse biotechnological and industrial applications.

Alcohol dehydrogenases (ADHs) are widely used for producing value-added compounds, including long-chain fatty alcohols with applications in the pharmaceutical and healthcare industries, as well as intermediates relevant to hydrocarbon fuels [23, 26, 27]; their activity toward long-chain substrates can be enhanced by strategies such as immobilization [23]. However, most ADHs intrinsically prefer short-chain substrates such as acetaldehyde, which limits their applicability to medium- and long-chain aldehydes. Therefore, shifting substrate specificity toward longer-chain substrates may help reduce competition with native cellular reactions and improve their utility in producing value-added compounds.

In this study, we aimed to alter the substrate specificity of YADH by engineering it to catalyze the conversion of *trans*-2-nonenal to *trans*-2-nonenol, while simultaneously enhancing thermostability. *Trans*-

2-nonenal is a volatile unsaturated aldehyde responsible for age-related body odor in humans, resulting from the oxidative degradation of fatty acids in the skin [28]. It has drawn considerable attention in the cosmetic and healthcare industries as a biomarker of aging and oxidative stress [29]. Its reduced form, *trans*-2-nonenol, is less volatile and toxic and has potential applications as a fragrance ingredient [28]. Engineering YADH to catalyze this transformation provides a useful model for expanding substrate range and offers potential value in industrial bioconversion.

Methods

Strains, media, chemicals, and other materials

Escherichia coli strains NovaBlue (Novagen, Madison, WI, USA) and C41(DE3) (Sigma-Aldrich Co. LLC, St. Louis, Missouri, USA) were used for DNA manipulation and protein expression analysis, respectively. Luria-Bertani (LB) broth (Difco; Franklin Lakes, New Jersey, USA) and Terrific broth (TB; Sigma-Aldrich; St. Louis, Missouri, USA) were used as growth media. Chemicals were purchased from Sigma-Aldrich (St. Louis, MO, USA), Tokyo Chemical Industry (Tokyo, Japan), and Nacalai Tesque Inc. (Kyoto, Japan). Polymerase chain reaction (PCR) was performed using PrimeSTAR Max (Takara Bio, Shiga, Japan). Gene assembly with the respective plasmids was performed using the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs, Ipswich, MA, USA). Primers were synthesized by Eurofins Genomics (Tokyo, Japan), and their sequences are listed in Table S1. DNA sequences were analyzed by Eurofins Genomics. Synthetic genes were purchased from Thermo Fisher Scientific (Life Technologies, Carlsbad CA, USA).

Construction of pCold-YADH

The synthetic *yadh* gene (UniProtKB entry P00330 [30]) was cloned into the NdeI and XhoI sites of pColdI to yield pCold-YADH. The product was then used to transform *E. coli* JM109 cells.

Site-directed mutagenesis of YADH

Site-specific mutations were introduced using PrimeSTAR Max with primers designed for PCR (Table S1). The PCR mixture (20 µL) contained 0.4 µL primers, 0.4 µL template plasmids, 8.8 µL distilled water, and 10 µL PrimeStarMax solution. The cycling conditions were as follows: 98°C for 1 min; 30 cycles of 98°C for 10 s, 65°C for 5 s, and 72°C for 30 s; and a final extension at 72°C for 7 min. The resulting PCR products were incubated with DpnI to remove the template DNA. The product was used to transform *E. coli* JM109 cells. The mutations were confirmed by DNA sequencing. Recombination mutagenesis was performed using a single-point mutant plasmid as the template.

Purification of recombinant YADH

Escherichia coli C41(DE3) harboring pColdI-YADH was pre-cultured overnight in LB broth containing 100 μg/mL ampicillin at 37°C. The culture was transferred to TB containing 100 μg/mL ampicillin and incubated at 37°C until the OD_{600} reached 0.4. The cells were cooled on ice, and gene expression was induced using 0.1 mM isopropyl β-D-thiogalactopyranoside at 15°C for 24 h. Cells were harvested by centrifugation, lysed using B-PER™ Bacterial Cell Lysis Reagent (Thermo Fisher Scientific) with DNase (5 U/mL) and lysozyme (10 mg/mL), and cell debris was removed by centrifugation. The supernatant was purified using TALON affinity resin (Takara, Shiga, Japan). Recombinant YADH was eluted using lysis buffer (20 mM Tris, 200 mM NaCl, and 20% [v/v] glycerol; pH 8.0) containing 500 mM imidazole, dialyzed, and analyzed using sodium dodecyl sulfate (SDS)-sulfate-polyacrylamide gel electrophoresis (PAGE) (Figure S1).

Enzyme assay of alanine-substituted YADH-variants

The dehydrogenation activity of YADH was assayed at 30°C. The reaction mixture (0.1 mL) contained 83 mM potassium phosphate buffer (pH 7.3), 40 mM KCl, 0.25 mM EDTA, 2 mM NADH, and 20 mM *trans*-2-nonenal. The reaction was initiated by adding YADH (6 μM), and the decrease in NADH absorbance was monitored at 340 nm wavelength.

Enzyme assay of the YADH-library and obtained YADH-variants

The dehydrogenation activity of YADH was assayed at 30°C. The reaction mixture (0.1 mL) contained 83 mM potassium phosphate buffer (pH 7.3), 40 mM KCl, 0.25 mM EDTA, 4 mM NADH, and 10 mM *trans*-2-nonenal. The reaction was initiated by adding YADH (100 nM), and the decrease in NADH absorbance was monitored at 340 nm wavelength.

Kinetics study

The steady-state kinetic parameters of YADH were calculated by measuring the initial velocity of the decrease in NADH absorbance. The reaction mixture (0.1 mL) containing 83 mM potassium phosphate buffer (pH 7.3), 40 mM KCl, 0.25 mM EDTA, 0.5 mM NADH, and various concentrations of *trans*-2-nonenal (0, 0.25, 0.5, 1, 2, 4, 8, 12, 20, 40, or 60 mM) was incubated for 10 min at 30°C. The reaction was initiated by adding an appropriate concentration of YADH (100 nM), and the decrease in NADH absorbance was monitored at 340 nm wavelength. The kinetic parameters were determined by nonlinear regression fitting of the initial velocity versus substrate concentration data to a substrate inhibition model:

$v = V_{max}\,[S] / (K_M + [S]\,(1 + [S]/K_i))$,

where Vmax is the maximum velocity, K_M is the Michaelis constant, and K_i is the substrate inhibition constant. Curve fitting was performed using nonlinear least-squares regression implemented in Python.

Substrate specificity assay

YADH dehydrogenation activity was assayed at 30°C. The reaction mixture (0.1 mL) contained 83 mM potassium phosphate buffer (pH 7.3), 40 mM KCl, 0.25 mM EDTA, 4 mM NADH, and 10 mM acetaldehyde, propionaldehyde, butylaldehyde, valeraldehyde, hexanal, heptanal, octanal, *trans*-2-nonenal, or *trans*-2-decenal. The reaction was initiated by adding YADH (100 nM), and the decrease in NADH absorbance was monitored at 340 nm wavelength.

Thermal shift assay

A thermal shift assay was performed using a LightCycler 96 RT-PCR instrument (Roche Diagnostics K.K., Basel, Switzerland) to monitor protein thermal unfolding in the presence of a fluorescent probe from the Protein Thermal Shift™ Dye Kit (Thermo Fisher Scientific). Briefly, 10 µL of the protein–dye solution was dispensed into a 96-well microplate. The final concentrations in each well were 3 µg of YADH and 20× dye. The excitation and emission filters were set at 577 and 620 nm, respectively. The assay was conducted over a temperature range of 37–98°C, at a linear heating rate of 0.07°C/s. Melting temperatures (T_M values) were determined using the LightCycler® 96 Application Software (Roche Diagnostics).

Machine Learning Models for Predicting Activity and Stability

Amino acid variants containing five mutation sites were encoded using physicochemical descriptors (z-scale [31], VHSE [32], T-scale [33], ST-scale [34], MS-WHIM [35], FASGAI [36], BLOSUM [37], and ProtFP [38]). Descriptor vectors for each position were concatenated into a single feature vector and standardized prior to model training.

Relative activity was defined as the ratio of the specific activity of each variant to that of WT-YADH. Variants with relative activity below the detection limit (0.01) were excluded from regression model training and treated as inactive in a separate classification model. The regression target was defined as:

$y = \log_2(\{\text{relative activity}\} + 0.01)$

Multiple machine-learning algorithms, including Random Forest, LightGBM, and support vector regression (SVR), were evaluated for activity prediction. Model performance was assessed using Spearman's rank correlation coefficient and ranking-based metrics. Hyperparameters were optimized using grid search using five-fold cross-validation, with model selection based primarily on Spearman's rank correlation coefficient. Regression performance was evaluated using Spearman's rank correlation coefficient (ρ), together with ranking-based metrics reflecting experimental screening utility: (i) the

median experimentally measured log2-transformed relative activity of the top 10% of variants ranked by predicted score, compared with the overall dataset median, and (ii) the fraction of the top 10% of variant exhibiting activity greater than WT, as an indicator of enrichment for beneficial variants.

For activity classification, variants with relative activity ≥ 0.01 were labeled as active and those below the detection limit as inactive. A support vector classifier (SVC) with a radial basis function kernel was trained to estimate the probability of activity. Classification performance was evaluated using the area under the receiver operating characteristic curve (ROC-AUC) and top-k (10, 20, and 30%) recall, defined as the proportion of experimentally active variants recovered among the top-k highest-confidence predictions.

Thermal stability was quantified as:

$\Delta T_M = T_{M,\,variant} - T_{M,\,WT}$

and modeled using SVR.

redictions from the activity regression, activity classification, and thermostability models were integrated into a single ranking score:

Final Score = $P_{active} \times A \times \exp(\Delta T_M/\tau)$

where P_{active} is the probability of activity predicted by the classification model, A is the predicted activity from the regression model, and ΔT_M is the predicted thermal stability difference. The exponential term acts as a penalty for destabilizing mutations, with τ controlling the tolerance to thermal stability loss.

Thermostability assays

The protein half-inactivation temperature (T_{50}) was defined as the temperature at which the enzyme lost 50% of its initial activity after heating. The T_{50} values were measured in the range of 55–75°C. The enzyme concentration was adjusted to 100 nM, and the purified enzymes were incubated at 55, 60, 65, 70, or 75°C for 10 min and cooled on ice for 5 min. Residual activity was determined as described above and compared with that of the untreated enzyme, which was set at 100%. The T_{50}-value was calculated by curve fitting. To determine enzyme decay at elevated temperatures, the enzyme solution was incubated at 60°C for various time intervals, and the remaining activity was measured at each time point. The half-life ($t_{1/2,\,60°C}$) was derived from the inactivation curve, defined as the time at which YADH lost 50% of its activity at a given temperature.

Docking simulation of YADH and trans-2-nonenal

The three-dimensional protein structure of YADH (PDB ID: 4w6z [19]) was obtained from the RCSB Protein Data Bank. The ligand structure was generated using ChemSketch and OpenBabel [39], and polar

hydrogen atoms were added using AutoDock Tools suite. Docking simulations were performed using the general docking tool in AutoDock 4.2.6 [40].

Docking simulation of the modeled WT-YADH, YADH-Y119F, YADH-M270G, and YADH-LWFMY

The three-dimensional protein structures of YADH variants were modeled using the local ColabFold environment [41]. The ligand structure was generated using ChemSketch and OpenBabel [39], and polar hydrogen atoms were added using the AutoDock Tools suite. Docking simulations were performed using the general docking tool in AutoDock4.2.6 [40]. The Grid box parameters were set as center_x = 88.453, center_y = -8.546, center_z = 44.035, size_x = 30, size_y = 30, and size_z = 44.

Results

Selecting sites for mutagenesis by docking simulation

Yeast alcohol dehydrogenase (YADH) was selected as the template enzyme for *trans*-2-nonenal reduction. YADH-mediated dehydrogenation requires both NADH and Zn^{2+}. YADH substrate specificity was determined based on the residues surrounding the substrate-binding pocket. The NICOTINAMIDE-8-IODO-ADENINE-DINUCLEOTIDE (8ID)-bound YADH structure (PDB ID: 4w6z [19]) was used. A model of YADH bound to *trans*-2-nonenal was generated using AutoDock4.2.6 [40]. In the constructed *trans*-2-nonenal-bound YADH model, residues T45, W54, L56, H66, W92, Y119, C153, M270, Y294, and V295 were located within 4 Å of *trans*-2-nonenal (Fig. 1). To select the amino acid residues for mutagenesis, YADH variants were constructed with alanine substitutions at these positions (except for H66 and C153, which are coordinated with catalytic zinc and are thus essential for activity [15]). Their relative activities compared with that of wild type YADH (WT-YADH) were measured by monitoring NADH depletion in the presence of *trans*-2-nonenal. The relative activities of YADH were defined by the relative absorbance (= [(absorbance value of YADH-variant)/(absorbance value of WT-YADH)]). YADH-T45A and YADH-V295A showed lower relative activities than that of WT-YADH, whereas YADH-W54A, YADH-W92A, YADH-Y119A, YADH-M270A, and YADH-Y294A showed relatively higher relative activities (Fig. 2). Five residues with the highest relative activity after alanine substitution (W54, W92, Y119, M270, and Y294) were selected for further analysis.

The amino acid residues surrounding the substrate are shown as sticks. The amino acid residues selected for mutagenesis are shown in red. The sphere indicates the zinc ion. *Trans*-2-nonenal and nicotinamide-8-iodo-adenine-dinucleotide (8ID) are shown in light blue and yellow, respectively.

Relative activities of YADH variants with alanine substitution at the substrate pocket residues were analyzed. Data are presented as the mean ± standard deviation of three independent experiments.

Preparation of training data

For training data acquisition, we constructed a library of 190 YADH variants with substitution at five amino acid residues (W54, W92, Y119, M270, and Y294). This included all single amino acid-substituted YADH variants at five residues (19 amino acid substitution × five residues = 95 variants) and YADH variants with multiple amino acid substitutions at five residues. Hereafter, we describe the substitutions of W54, W92, Y119, M270, and Y294 as five letters. For instance, a variant with W54A, W92A, Y119A, M270A, and Y294A is described as YADH-AAAAA. Dehydrogenase activity toward *trans*-2-nonenal was evaluated by monitoring NADH consumption as a decrease in absorbance at 340 nm. The relative activity of YADH was defined as:

Relative specific activity = (Specific activity of YADH variant) / (Specific activity of WT-YADH) (1)

The relative thermal stabilities of YADH variants were defined as:

ΔT_M = (T_M value of YADH variant) – (T_M value of WT-YADH). (2)

To minimize inter-experimental variability in the enzyme assays, the substrate concentration was fixed at 10 mM, corresponding to approximately 5–10 times the Km value of the WT-YADH [14]. Further, the WT-YADH was included as a control in each experiment, and variant activities were normalized by the specific activity of WT-YADH measured under the same conditions.

We identified 18 YADH variants with higher activities and three with higher thermostabilities than those of WT-YADH. When W54 was substituted, YADH-W54F, -W54G, -W54H, -W54I, -W54L, -W54M, -W54Q, -W54S, -W54V, and -W54Y showed higher activity but lower thermostability than that of WT-YADH (Figure S2A). When W92 was substituted, YADH-W92G exhibited higher activity but lower thermostability than that of the WT-YADH (Figure S2B). Notably, we obtained YADH-Y119F, which showed modest improvements in both activity and thermostability (Figure S2C). YADH-Y119W showed increased thermostability though it showed lower activity than that of WT-YADH. Although YADH -M270G, -M270N, -M270Q, -M270S, and -M270V showed higher activity, only YADH-M270L showed increased thermostability, whereas the other variants showed reduced thermostability compared with that of WT-YADH (Figure S2D). Substituting Y294 with other amino acids resulted in the reduction of both activity and thermostability (Figure S2E). We also obtained two YADH variants with multiple substitutions that showed increased relative activity, but most YADH variants with multiple substitutions showed reduced activity or complete loss of activity (Figure S2F-J).

In total, 19 YADH variants exhibited improved activity toward *trans*-2-nonenal, while three YADH variants showed enhanced thermostability (Fig. 3, Figure S2).

The relative activities and thermostabilities of 190 YADH variants with single or multiple amino acid substitution were analyzed.

Construction of a machine-learning model for the dual optimization of YADH activity and thermostability

To construct a machine-learning framework for the simultaneous optimization of catalytic activity and thermostability, multiple combinations of amino acid descriptor sets and machine-learning algorithms were evaluated. Among the tested combinations, Support Vector Regression (SVR) with z-scale descriptors achieved the highest predictive performance for activity prediction and was therefore selected for further optimization (Figure S3(a)). Hyperparameter tuning further improved Spearman's rank correlation coefficient, indicating enhanced preservation of the relative ordering of variants. It also improved ranking-based metrics directly relevant to experimental screening efficiency, including the median activity among the top 10% of model-ranked variants and the enrichment of variants outperforming WT within this subset (Figure S3(b) and Table S3(a)).

To account for variants with activities below the experimental detection limit, we additionally trained a binary Support Vector Classifier (SVC) to distinguish active from inactive variants. Performance was assessed using ROC-AUC and top-k recall, the latter quantifying the proportion of experimentally active variants recovered among the highest-confidence predictions (Figure S3(c) and Table S3(b)).

An additional SVR model was trained to predict changes in thermal stability (ΔT_M). Although most mutations reduced thermostability, incorporating predicted ΔT_M as a penalty term enabled balanced prioritization of variants with improved activity while avoiding severely destabilizing substitutions (Figure S3(d)).

The activity regression, activity classification, and thermostability prediction models were integrated to rank all variants in the unexplored combinatorial sequence space (20^5 combinations). Based on the integrated ranking score, the top 22 candidate variants were selected for experimental validation (Data S1).

Experimental evaluation of YADH variants predicted to have high activities

The top 22 YADH variants predicted to have activity comparable to or higher than that of WT-YADH (Data S1) were experimentally evaluated for relative activity and thermal stability (Fig. 4 (a)). YADH-LWMMY was excluded from further analysis because no soluble protein was detected. Among the analyzed YADH variants, four YADH variants (YADH-FWMMY, YADH-FWYGY, YADH-LWFMY, and YADH-YWYGY) demonstrated higher activity than that of WT-YADH, and one of them (YADH-LWFMY) also showed improved thermal stability (Fig. 4 (b)).

Kinetic study of the obtained YADH variants

Subsequently, kinetic and substrate specificity analyses were performed for the selected variants, as well as for variants containing the corresponding single amino acid substitutions. Further, comprehensive thermal stability characterization was performed for the variant exhibiting both enhanced activity and improved stability.

The kinetic parameters of the purified YADH variants with respect to *trans*-2-nonenal were determined at 30°C (Table 1). Compared with the kinetic parameters of WT-YADH (K_M = 1.5 ± 0.3 mM; k_{cat} = 1479 ± 231 (min^{-1})), other YADH variants showed similar kinetic parameters except for YADH-LWFMY (K_M = 13.0 ± 3.3 mM; k_{cat} = 25187 ± 6012 (min^{-1})) (Table 1, Figure S4)).

	WT	FWMMY	FWYGY	LWFMY	YWYGY
K_M (mM)	1.5±0.3	2.9±0.2	2.3±0.3	13.0±3.3	1.2±0.1
K_i (mM)	41.3±11.1	30.1±2.9	40.1±9.7	2.3±0.7	59.2±5.6
k_{cat} (min^{-1})	1479±231	4747±182	6010±390	25187±6012	3296±60
k_{cat}/K_M ($min^{-1}mM^{-1}$)	1007±247	1657±117	2569±369	1934±669	2735±133

	W54F	W54Y	W54L	Y119F	Y119M	M270G
K_M (mM)	1.3±0.1	0.6±0.1	2.0±0.1	2.4±0.1	2.0±0.1	1.6±0.1
K_i (mM)	53.5±5.9	69.6±25.8	34.9±6.1	48.7±6.0	67.2±5.3	41.1±6.3
k_{cat} (min^{-1})	5737±269	847±96	6670±337	1509±57	2523±58	2726±86
k_{cat}/K_M ($min^{-1}mM^{-1}$)	4559±451	1419±333	3304±283	629±41	1254±53	1670±90

Table1. Kinetic parameters of YADH variants. Data are presented as the mean ± standard deviation of three independent experiments.

Substrate specificity analyses of YADH variants

The substrate specificities of YADH-variants were evaluated using a series of aldehydes with chain lengths ranging from C2 to C10 (acetaldehyde, propionaldehyde, butylaldehyde, valeraldehyde, hexanal, heptanal, octanal, *trans*-2-nonenal, and *trans*-2-decenal respectively) (Fig. 5). WT-YADH exhibited high specificity toward short-chain aldehydes (C2–C4). In contrast, YADH-FWMMY, YADH-W54F, and YADH-W54L showed a broadened substrate specificity toward longer-chain aldehydes (C5–C6), accompanied by an overall increase in relative specific activity. YADH-Y119F and YADH-Y119M exhibited substrate specificity profiles similar to those of WT-YADH; however, their activities toward short-chain aldehydes were notably enhanced.

Interestingly, YADH-FWYGY and YADH-YWYGY had lost activity toward the native substrates (C2–C4) and instead displayed enhanced activity exclusively toward longer-chain aldehydes (especially toward heptanal and *trans*-2-nonenal), suggesting a shift in substrate specificity toward longer-chain substrates.

Among the corresponding single amino acid substitutions, YADH-M270G exhibited activity only toward C7 and C9 aldehydes.

Thermostability analysis of YADH variants

The thermostabilities of WT-YADH, YADH-Y119F, and YADH-LWFMY were compared. After a 10-min incubation at 55–75°C, the activity of WT-YADH decreased sharply relative to that of YADH-Y119F and YADH-LWFMY. The T_{50} values of YADH-Y119F and YADH-LWFMY were 59.8 ± 1.3°C, and 60.7 ± 0.4°C respectively, which were approximately 3°C higher than that of WT-YADH (57.0 ± 0.3˚C) (Fig. 6(a)). At 60°C, the $t_{(1/2, 60°C)}$ of YADH-Y119F and YADH-LWFMY were 29 min and 19 min respectively, which were more than 1.4 ~ 2.2 times higher than that of WT-YADH (13 min) (Fig. 6(b)). These results indicate that the Y119F substitution enhanced YADH thermostability.

Comparison of the structures of WT-YADH, YADH-Y119F, YADH-M270G, and YADH-LWFMY

To investigate the mechanism underlying the increased activity of YADH in the catalysis of *trans*-2-nonenal, we compared the three-dimensional structures of WT-YADH, YADH-Y119F, YADH-M270G, and YADH-LWFMY (Fig. 7) and performed docking simulation of each YADH variant with *trans*-2-nonenal. Comparison of WT-YADH with YADH-Y119F revealed minimal structural differences (Figure. 7 (a)), and the minimum binding energy was largely unchanged (wild type: -4.23 kcal/mol; YADH-Y119F: -4.19 kcal/mol). In contrast, comparison with YADH-M270G showed that substituting methionine with glycine enlarged the substrate-binding pocket (Fig. 7 (b)). The minimum binding energy was largely unchanged (YADH-M270G: -4.29 kcal/mol). For YADH-LWFMY, the substrate entrance appeared slightly widened relative to that of the wild type (Figure. 7 (c)), accompanied by a slight decrease in the minimum binding energy (YADH-LWFMY: -4.34 kcal/mol).

Discussion

In this study, machine learning models were constructed to predict YADH variants with enhanced activity toward the non-native substrate *trans*-2-nonenal and with enhanced thermostability. Although not all predicted variants showed improved activity experimentally, several variants exhibited substantially higher activity than WT-YADH, demonstrating that the model was able to efficiently enrich beneficial variants from a relatively small candidate set despite the limited and imbalanced training data.

The discrepancy between predicted and experimental activities likely reflects the scarcity of highly active variants in the training dataset. Because most multi-substituted variants in the dataset exhibited reduced activity and stability, prediction of rare high-performance variants remained challenging. Nevertheless, compared with exhaustive combinatorial approaches, the present strategy substantially reduced the experimental search space while still enabling identification of improved variants. Future performance may be improved by iterative incorporation of experimentally validated high-activity variants to progressively rebalance the training distribution and refine model generalization.

Interestingly, several variants also exhibited altered substrate specificity toward longer-chain aldehydes. In particular, YADH-FWYGY and YADH-YWYGY lost substantial activity toward native short-chain substrates while showing enhanced activity toward heptanal and *trans*-2-nonenal. Structural analyses

suggested that substitutions such as M270G enlarged the substrate-binding pocket or substrate entrance, potentially facilitating accommodation of bulky long-chain aldehydes. Since long-chain aldehydes are important intermediates in microbial production of fatty alcohols and hydrocarbon-related compounds, such substrate specificity shifts may improve compatibility with industrial bioproduction systems.

The kinetic analyses further suggested that the screening conditions strongly influenced the properties of the selected variants. Because the training data were generated using relatively high substrate concentrations (10 mM), variants with improved performance under high-substrate conditions were likely preferentially enriched. Consistent with this possibility, YADH-LWFMY exhibited markedly increased kcat together with elevated Km values. Although the increased Km indicates reduced substrate affinity at low substrate concentrations, such a shift toward higher substrate concentrations may be advantageous under industrial reaction conditions, where high substrate loading is often used to maximize productivity.

Notably, several predicted variants retained substantial thermostability despite containing multiple substitutions, and YADH-LWFMY exhibited simultaneous improvement of both activity and thermostability. Because enhancement of catalytic activity is frequently accompanied by reduced protein stability, this result suggests that machine learning–guided approaches may help identify mutational combinations that partially overcome the conventional activity–stability tradeoff.

Conclusions

In this study, we applied a machine learning–based approach to predict enzyme variants with improved activity and thermostability. Despite limited predictive accuracy, likely because of constraints in the training data, the model enabled efficient identification of variants with relatively high activity from a small candidate set. Notably, despite the generally deleterious effects of multiple mutations on activity and thermostability, the predicted variants largely retained both properties. Remarkably, we also obtained YADH variants in which substrate specificity was completely shifted from the native C2 substrate, ethanol, to a non-native C9 *trans*-2-nonenal. Structural analyses suggest that substitutions of larger amino acids with smaller amino acids may expand the substrate-binding pocket or entrance, contributing to these functional changes. Overall, this approach provides a useful framework for identifying variants that balance activity and stability, although improvements in training data are important for further enhancing predictive performance.

Abbreviations

Luria-Bertani (LB)

polyacrylamide gel electrophoresis (PAGE)

polymerase chain reaction (PCR)

protein half-inactivation temperature (T_{50})

sodium dodecyl sulfate (SDS)

wild-type YADH (WT-YADH)

Yeast alcohol dehydrogenase I (YADH)

Support Vector Regression (SVR)

Support Vector Classifier (SVC)

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competinginterests

The authors declare no conflicts of interest.

Acknowledgements

The authors would like to thank Editage (www.editage.jp) for the English language review of this manuscript.

Authorcontributions

Y.O.: Writing-review and editing, Writing-original draft, Visualization, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, and Conceptualization. **A.K.**: Writing-review and editing, Methodology, Formal analysis, Data curation, Resources. **T.S.**: Writing – review and editing, Methodology, Formal analysis, Data curation, and Resources.

Funding

This study was supported by the RIKEN Center for Sustainable Resource Science (Y.O., T.S., and A.K.).

This study was supported by JSPS KAKENHI (Grant Number JP21H05230; Bottom-Up Biotech).

Data availability

All data generated or analyzed in this study are included in this published article.

Authordetails

[1]Center for Sustainable Resource Science, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [2]Graduate School of Science, Technology, and Innovation, Kobe University, 1-1 Rokkodai, Nada, Kobe 657-8501, Japan

ORCID ID

Yuki Ogawa: https://orcid.org/0000-0001-5798-7895

Akihiko Kondo: https://orcid.org/0000-0003-1527-5288

Tomokazu Shirai: https://orcid.org/0000-0002-8565-5895

References

1. Victorino da Silva Amatto I, Gonsales da Rosa-Garzon N, Antônio de Oliveira Simões F, Santiago F, Pereira da Silva Leite N, Raspante Martins J, et al. Enzyme engineering and its industrial applications. Biotechnol Appl Biochem. 2022;69:389–409.
2. Rigoldi F, Donini S, Redaelli A, Parisini E, Gautieri A. Review: Engineering of thermostable enzymes for industrial applications. APL Bioeng. 2018;2:011501.
3. Nezhad NG, Rahman RNZRA, Normi YM, Oslan SN, Shariff FM, Leow TC. Recent advances in simultaneous thermostability-activity improvement of industrial enzymes through structure modification. Int J Biol Macromol. 2023;232:123440.
4. Yang W, Yang Y, Zhang L, Xu H, Guo X, Yang X, et al. Improved thermostability of an acidic xylanase from Aspergillus sulphureus by combined disulphide bridge introduction and proline residue substitution. Sci Rep. Nature Publishing Group; 2017;7:1587.
5. Zhou Z, Wang X. Improve thermostability of Bacillus sp. TS chitosanase through structure-based alignment. Sci Rep. Nature Publishing Group; 2021;11:15846.
6. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad Sci. Proceedings of the National Academy of Sciences; 2019;116:8852–8.
7. Ogawa Y, Saito Y, Yamaguchi H, Katsuyama Y, Ohnishi Y. Engineering the Substrate Specificity of Toluene Degrading Enzyme XylM Using Biosensor XylS and Machine Learning. ACS Synth Biol. 2023;12:572–82.
8. Saito Y, Oikawa M, Sato T, Nakazawa H, Ito T, Kameda T, et al. Machine-Learning-Guided Library Design Cycle for Directed Evolution of Enzymes: The Effects of Training Data Composition on Sequence Space Exploration. ACS Catal. American Chemical Society; 2021;11:14615–24.
9. Huang C, Zhang L, Tang T, Wang H, Jiang Y, Ren H, et al. Application of Directed Evolution and Machine Learning to Enhance the Diastereoselectivity of Ketoreductase for Dihydrotetrabenazine

Synthesis. JACS Au. American Chemical Society; 2024;4:2547–56.
10. Thomas N, Belanger D, Xu C, Lee H, Hirano K, Iwai K, et al. Engineering highly active nuclease enzymes with machine learning and high-throughput screening. Cell Syst. 2025;16:101236.
11. Landwehr GM, Bogart JW, Magalhaes C, Hammarlund EG, Karim AS, Jewett MC. Accelerated enzyme engineering by machine-learning guided cell-free expression. Nat Commun. Nature Publishing Group; 2025;16:865.
12. Ding K, Chin M, Zhao Y, Huang W, Mai BK, Wang H, et al. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. Nat Commun. Nature Publishing Group; 2024;15:6392.
13. Yi X, Yu H, Yu H, Ye L. Enhancing Activity and Stability of Transaminase through Integrated Machine Learning, Rational Design, and Directed Evolution Approaches. ACS Catal. American Chemical Society; 2025;15:14150–60.
14. Zhang F, Naeem M, Yu B, Liu F, Ju J. Improving the enzymatic activity and stability of N-carbamoyl hydrolase using deep learning approach. Microb Cell Factories. 2024;23:164.
15. Li G, Zhang H, Sun Z, Liu X, Reetz MT. Multiparameter Optimization in Directed Evolution: Engineering Thermostability, Enantioselectivity, and Activity of an Epoxide Hydrolase. ACS Catal. American Chemical Society; 2016;6:3679–87.
16. Nealon CM, Musa MM, Patel JM, Phillips RS. Controlling Substrate Specificity and Stereospecificity of Alcohol Dehydrogenases. ACS Catal. 2015;5:2100–14.
17. Madhusudhan MC, Raghavarao KSMS, Nene S. Integrated process for extraction and purification of alcohol dehydrogenase from Baker's yeast involving precipitation and aqueous two phase extraction. Biochem Eng J. 2008;38:414–20.
18. Green DW, Sun HW, Plapp BV. Inversion of the substrate specificity of yeast alcohol dehydrogenase. J Biol Chem. 1993;268:7792–8.
19. Raj SB, Ramaswamy S, Plapp BV. Yeast Alcohol Dehydrogenase Structure and Catalysis. Biochemistry. 2014;53:5791–803.
20. Gould RM, Plapp BV. Substitution of arginine for histidine-47 in the coenzyme binding site of yeast alcohol dehydrogenase I. Biochemistry. 1990;29:5463–8.
21. Pal S, Plapp BV. The Thr45Gly substitution in yeast alcohol dehydrogenase substantially decreases catalysis, alters pH dependencies, and disrupts the proton relay system. Chem Biol Interact. 2021;349:109650.
22. Weinhold EG, Benner SA. Engineering yeast alcohol dehydrogenase. Replacing Trp54 by Leu broadens substrate specificity. Protein Eng. 1995;8:457–61.
23. Ottone C, Bernal C, Serna N, Illanes A, Wilson L. Enhanced long-chain fatty alcohol oxidation by immobilization of alcohol dehydrogenase from S. cerevisiae. Appl Microbiol Biotechnol. 2018;102:237–47.

24. Larreta-Garde V, Xu And ZF, Thomas D. Behavior of Enzymes in the Presence of Additives Influence of Alcohóls, Polyols, and Sugars on Activity and Stability of Yeast Alcohol Dehydrogenase. Ann N Y Acad Sci. 1988;542:294–8.
25. Markossian KA, Golub NV, Khanova HA, Levitsky DI, Poliansky NB, Muranov KO, et al. Mechanism of thermal aggregation of yeast alcohol dehydrogenase I: Role of intramolecular chaperone. Biochim Biophys Acta BBA - Proteins Proteomics. 2008;1784:1286–93.
26. Zheng Y-G, Yin H-H, Yu D-F, Chen X, Tang X-L, Zhang X-J, et al. Recent advances in biotechnological applications of alcohol dehydrogenases. Appl Microbiol Biotechnol. 2017;101:987–1001.
27. Schirmer A, Rude MA, Li X, Popova E, del Cardayre SB. Microbial Biosynthesis of Alkanes. Science. American Association for the Advancement of Science; 2010;329:559–62.
28. Haze S, Gozu Y, Nakamura S, Kohno Y, Sawano K, Ohta H, et al. 2-Nonenal Newly Found in Human Body Odor Tends to Increase with Aging. J Invest Dermatol. 2001;116:520–4.
29. Liguori I, Russo G, Curcio F, Bulli G, Aran L, Della-Morte D, et al. Oxidative stress, aging, and diseases. Clin Interv Aging. 2018;13:757–72.
30. Bennetzen JL, Hall BD. The primary structure of the Saccharomyces cerevisiae gene for alcohol dehydrogenase. J Biol Chem. 1982;257:3018–25.
31. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. J Med Chem. American Chemical Society; 1998;41:2481–91.
32. Mei H, Liao ZH, Zhou Y, Li SZ. A new set of amino acid descriptors and its application in peptide QSARs. Pept Sci. 2005;80:775–86.
33. Tian F, Zhou P, Li Z. T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides. J Mol Struct. 2007;830:106–15.
34. Yang L, Shu M, Ma K, Mei H, Jiang Y, Li Z. ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues. Amino Acids. 2010;38:805–16.
35. Zaliani A, Gancia E. MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. J Chem Inf Comput Sci. American Chemical Society; 1999;39:525–33.
36. Liang G, Li Z. Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. QSAR Comb Sci. 2007;26:754–63.
37. Georgiev AG. Interpretable Numerical Descriptors of Amino Acid Space. J Comput Biol. Mary Ann Liebert, Inc., publishers; 2009;16:703–23.
38. van Westen GJ, Swier RF, Cortes-Ciriano I, Wegner JK, Overington JP, IJzerman AP, et al. Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. J Cheminformatics. 2013;5:42.
39. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminformatics. 2011;3:33.

40. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30:2785–91.
41. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. Nature Publishing Group; 2022;19:679–82.

Figures

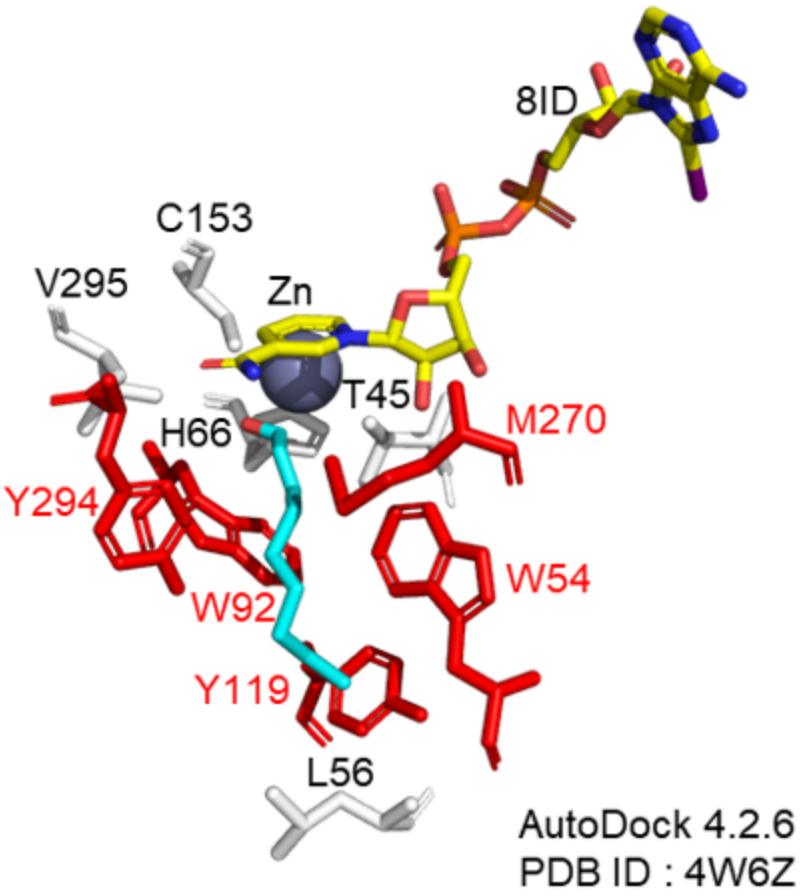


Figure 1

The substrate pocket of YADH (PDB ID:4W6Z)

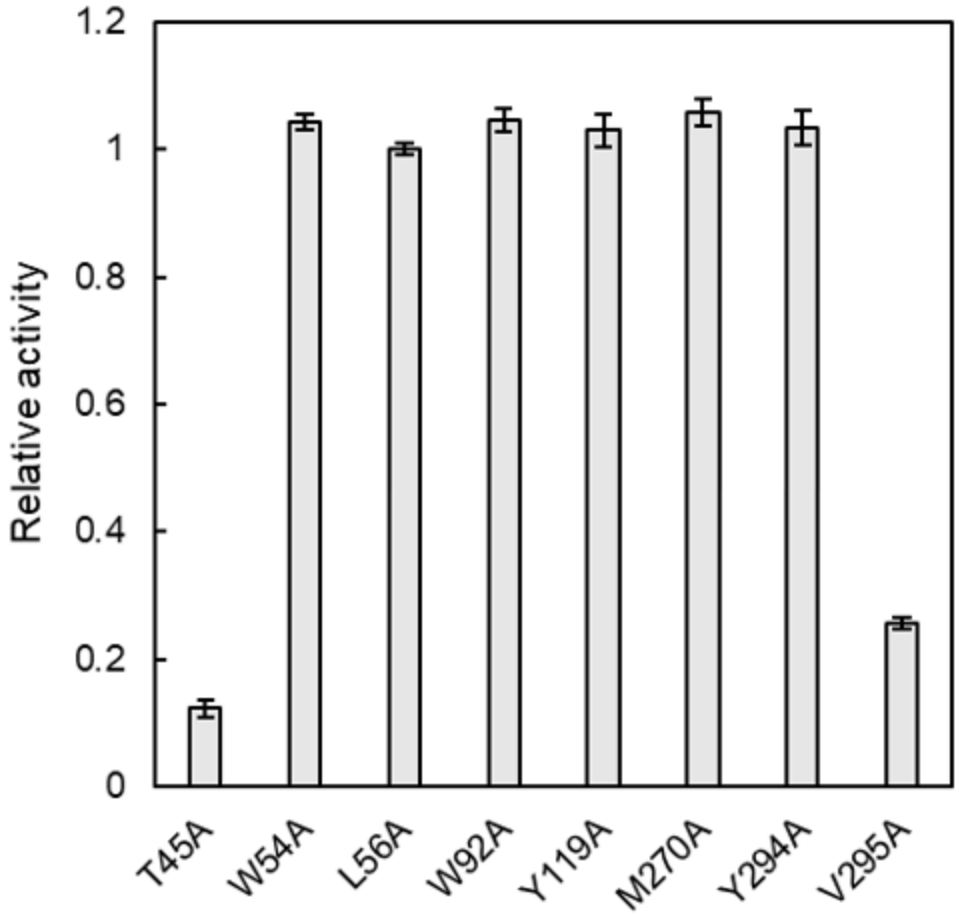


Figure 2

Relative activities of alanine-substituted YADH variants

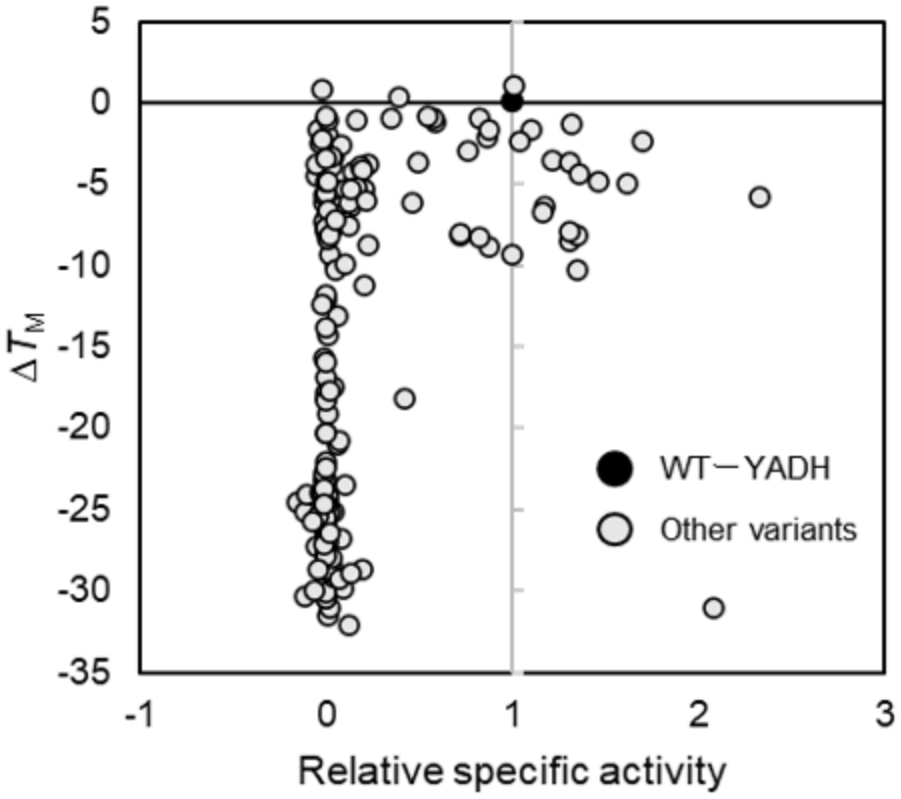


Figure 3

Relative specific activities and thermostabilities of YADH variants

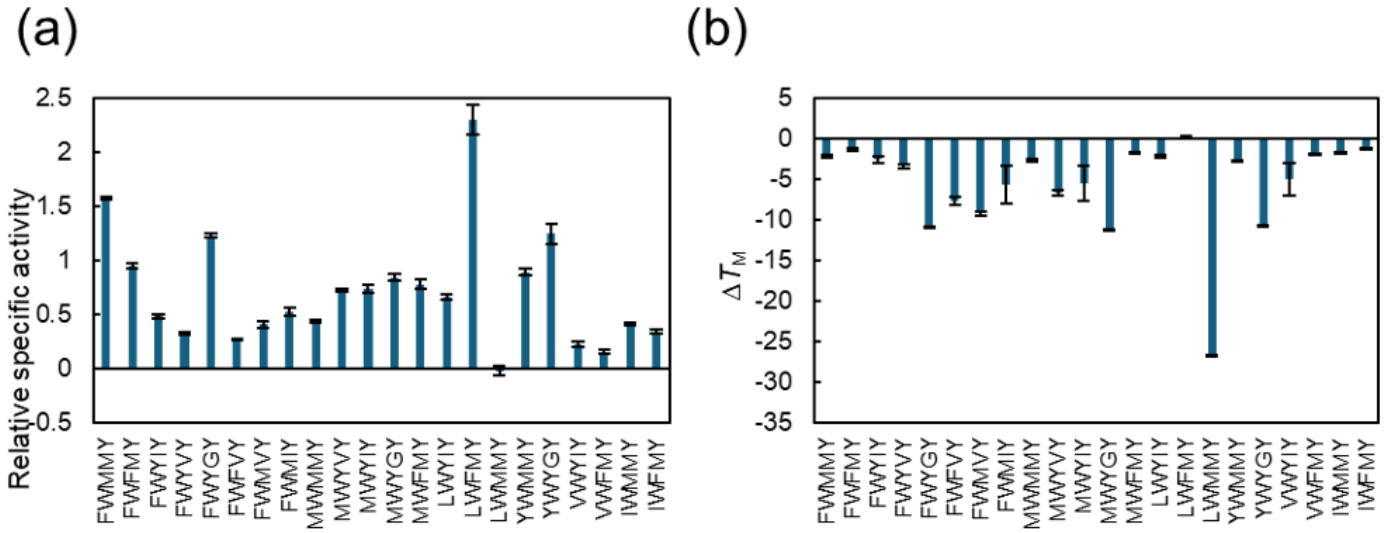


Figure 4

Experimental (a) relative specific activity and (b) thermostability of YADH variants predicted to be highly active. Data are presented as the mean ± standard deviation of three independent experiments.

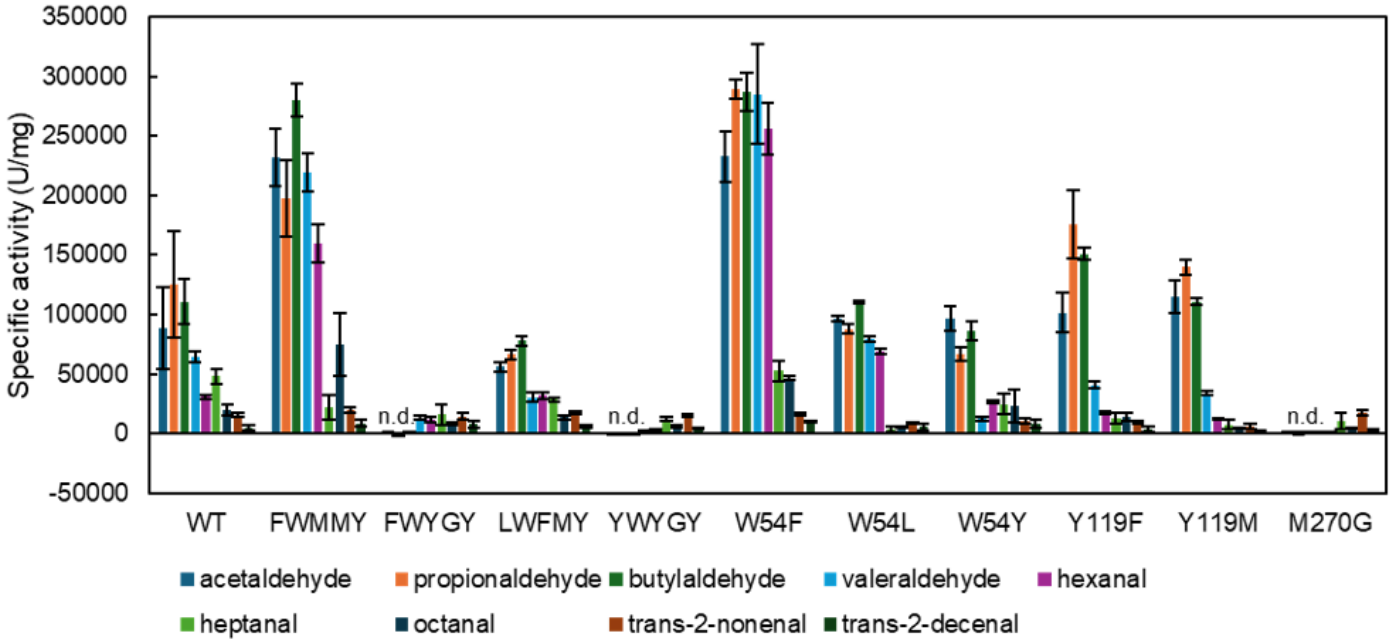


Figure 5

Substrate specificity of the top four YADH-variants and YADH-variants with the corresponding single amino acid substitutions. Data are presented as the mean ± standard deviation of three independent experiments.

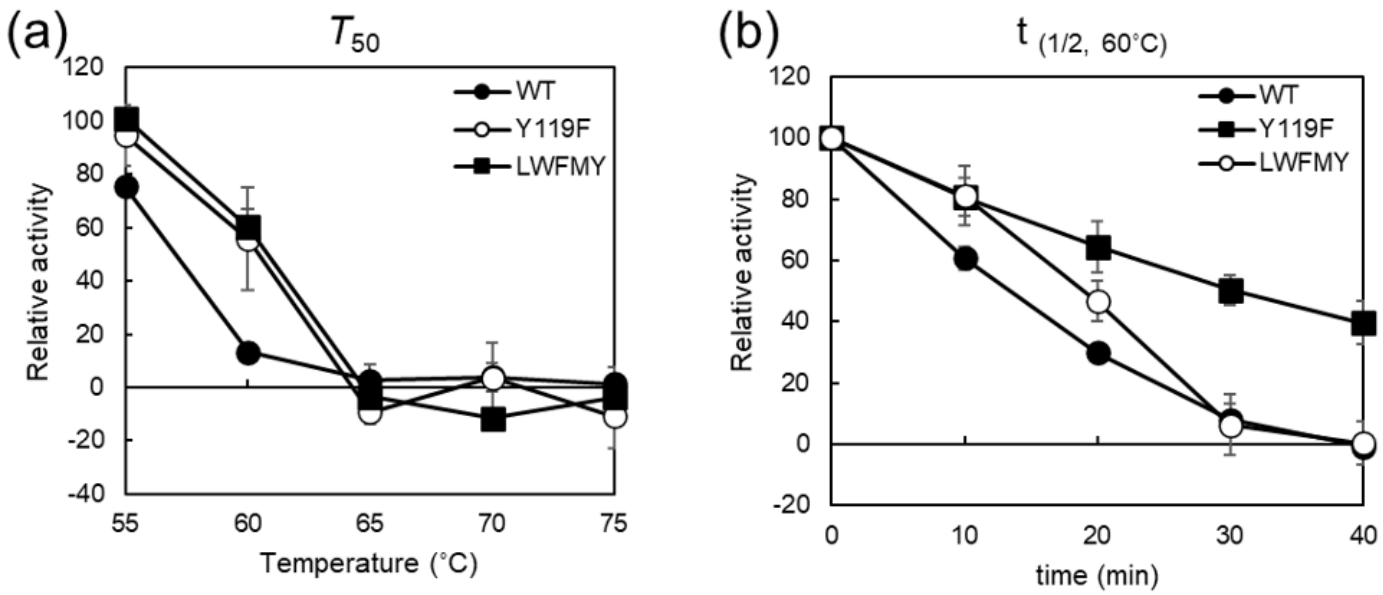


Figure 6

Thermostability of WT-YADH and YADH-Y119F

(a) T_M values of WT-YADH and YADH-Y119F. (b) $t_{(1/2, 70°C)}$ values of WT-YADH and YADH-Y119F. Data are presented as the mean ± standard deviation of three independent experiments.

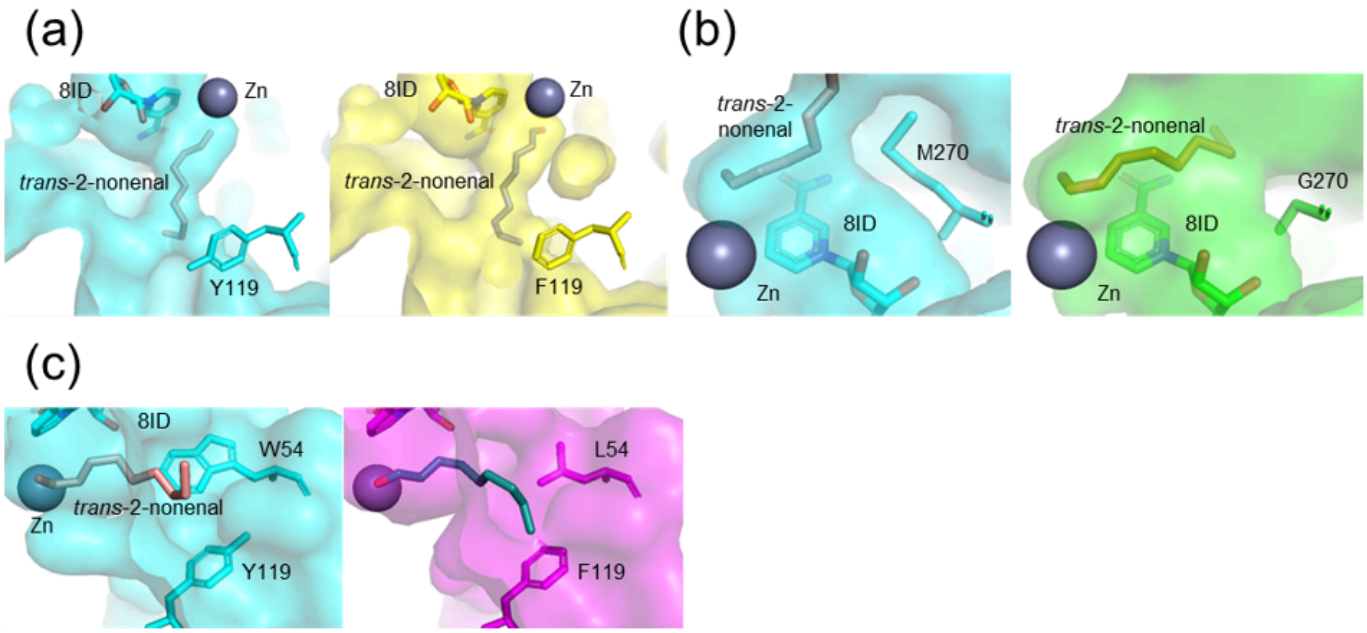


Figure 7

Structural comparison of (a) WT-YADH (blue) and YADH-Y119F (yellow), (b) WT-YADH (blue) and YADH-M270G (green), and (c) WT-YADH (blue) and YADH-LWFMY (magenta), highlighting mutation-induced local structural rearrangements around the substrate-binding pocket.

Supplementary Files

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

- yukiogawasupplementary.docx
- DataS1.xlsx