Supplementary Figures

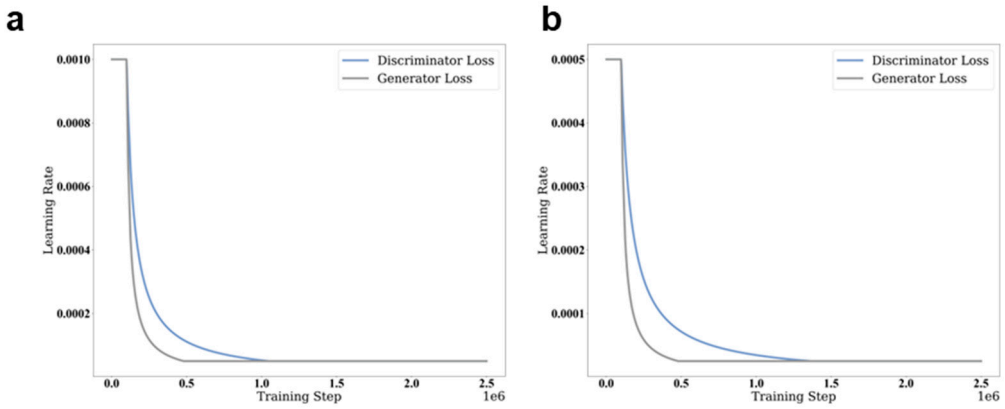

Figure S1: Learning rate schedules during the training process. (a) Learning rate schedule during training of MDH model. **(b)** Learning rate schedule during training of FBA model.

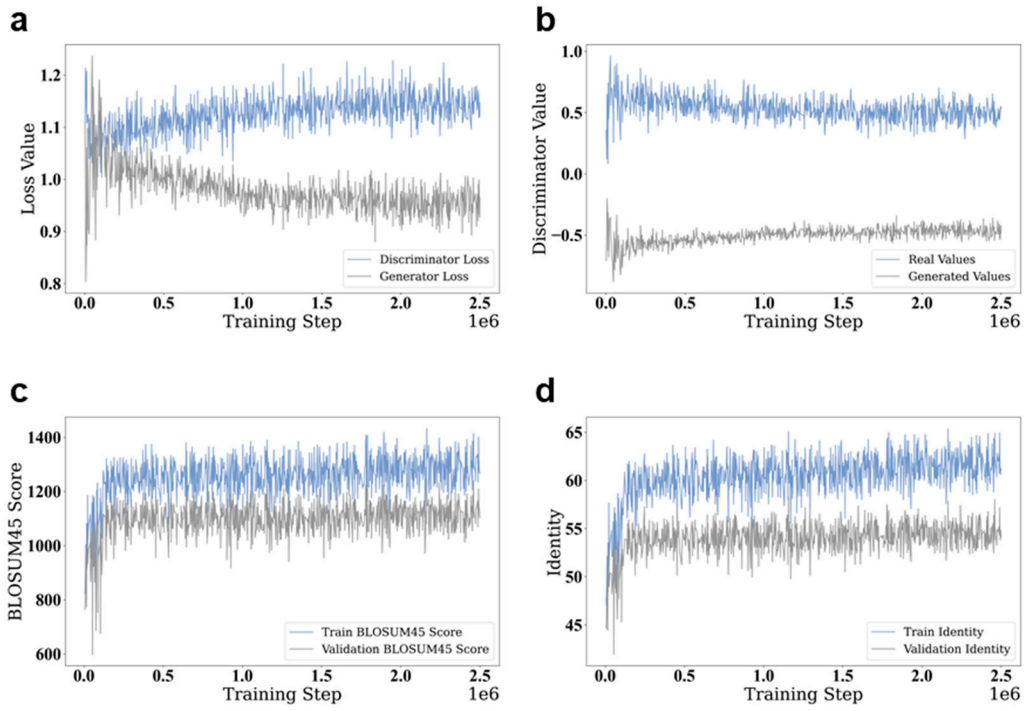

Figure S2: Training process of the selected MDH model. (a) Loss curves of generator and discriminator during the training process. **(b)** Model discriminator scores during the training process. **(c)** BLOSUM45 scores of generated sequences during the training process. **(d)** Identity trends of train and validation set during the training process.

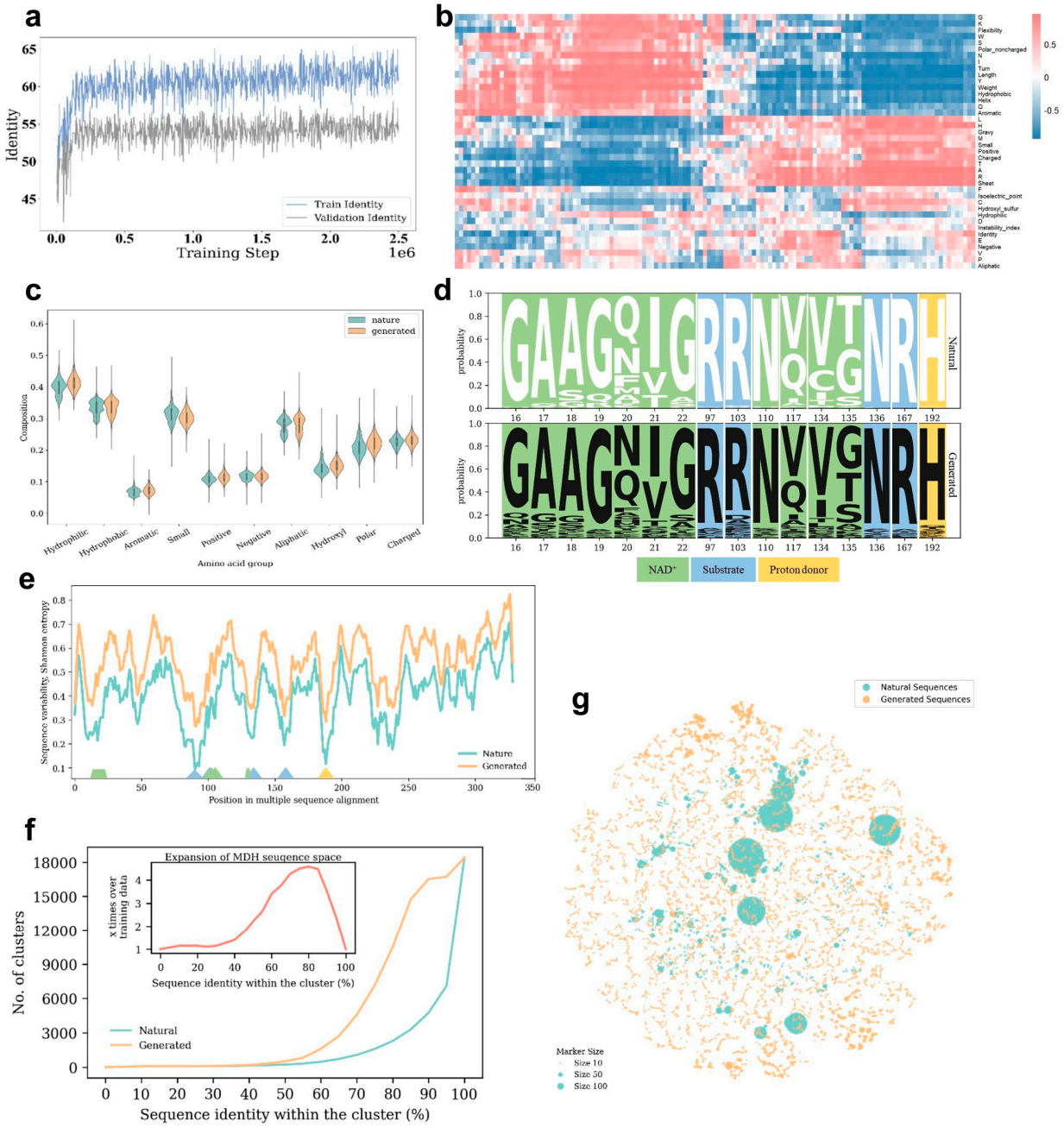

Figure S3: Evaluation of MDH training process and generated sequences. (a) The identity of generated sequences to training and validation sequences was monitored during the training process. A regression fit using a weighted sum of linear and exponential terms was applied, depicted by solid lines. **(b)** Interpolation results were obtained through correlating the latent space vectors with protein properties calculated through the interpolation of each variable dimension. **(c)** Amino acids were grouped on the basis of physicochemical properties and box plots of the percentage amino acid composition of the output and natural sequences were plotted. This analysis provides insight into the differential distribution of amino acid composition between the generated and natural sequences. **(d)** A sequence logo was created to illustrate the key conserved positions within the multiple sequence alignment. This visualization helped identify important residues or motifs that were preserved in the generated sequences, indicating their potential functional significance. **(e)** Shannon entropies were calculated to estimate the sequence variability for both the generated and training sequences, based on the multiple sequence alignment. This analysis provided insights into the diversity and conservation of amino acids at different positions within the sequences. **(f)** Evaluating the sequence diversity between the sequences we generated and the MDH training dataset. **(g)** A t-SNE visualization was performed to visualize the natural and generated MDH sequences. The dot sizes represented the cluster size based on 80% identity for each representative sequence. This analysis provided a visual representation of the distribution and clustering of the generated sequences compared to the natural MDH sequences.

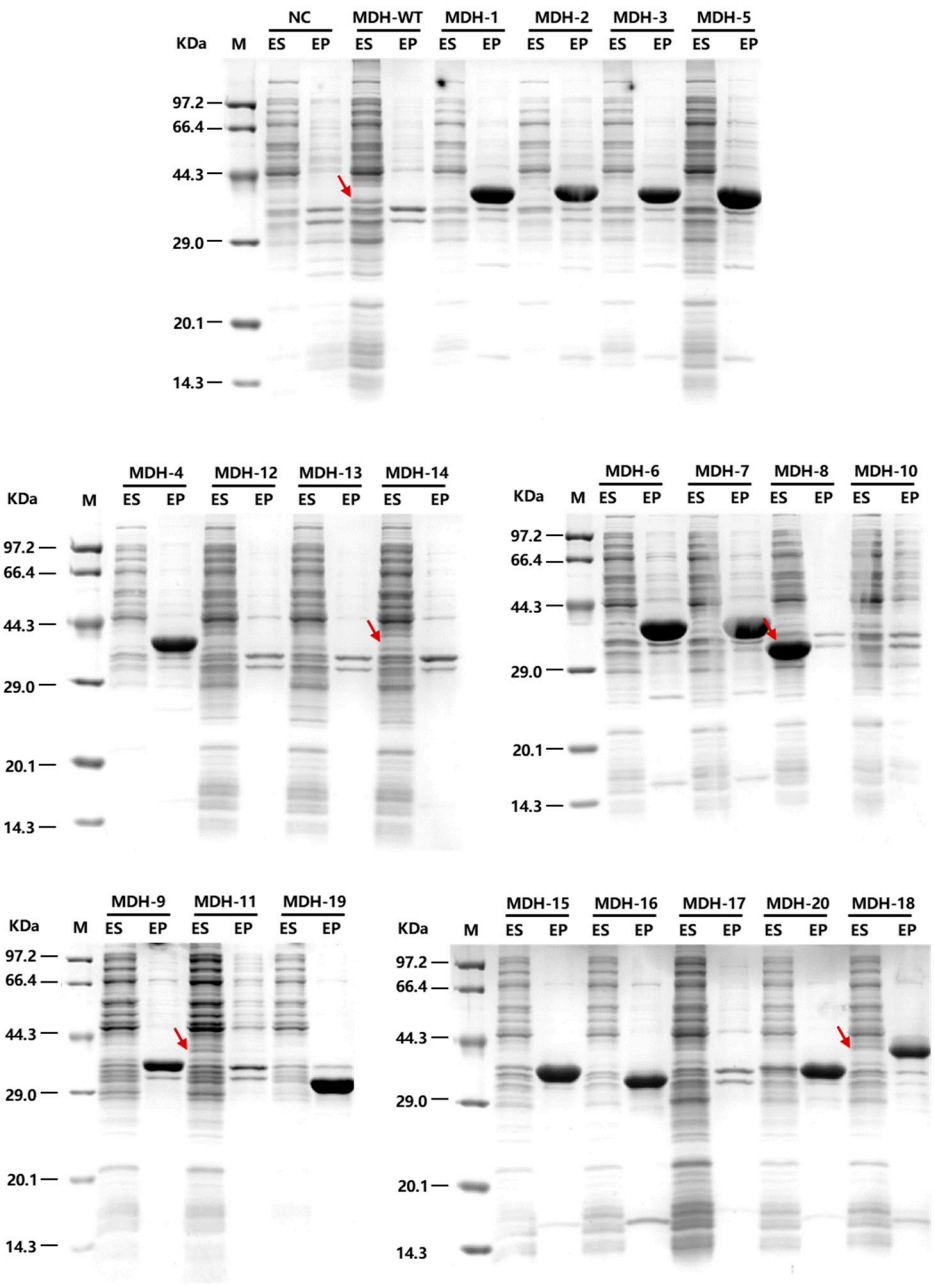

Figure S4: SDS-PAGE results for expression of recombinant MDHs. SDS-PAGE results of the expression of 20 generated MDHs (MDH-1 to MDH-20) in *E. coli*. Lane M, protein molecular weight marker; lane ES, expression supernatant; lane EP, expression precipitate; lane NC, negative control using pET32a vector; lane MDH-WT, positive control using MDH-WT. The soluble expressed target protein bands are indicated by a red arrow.

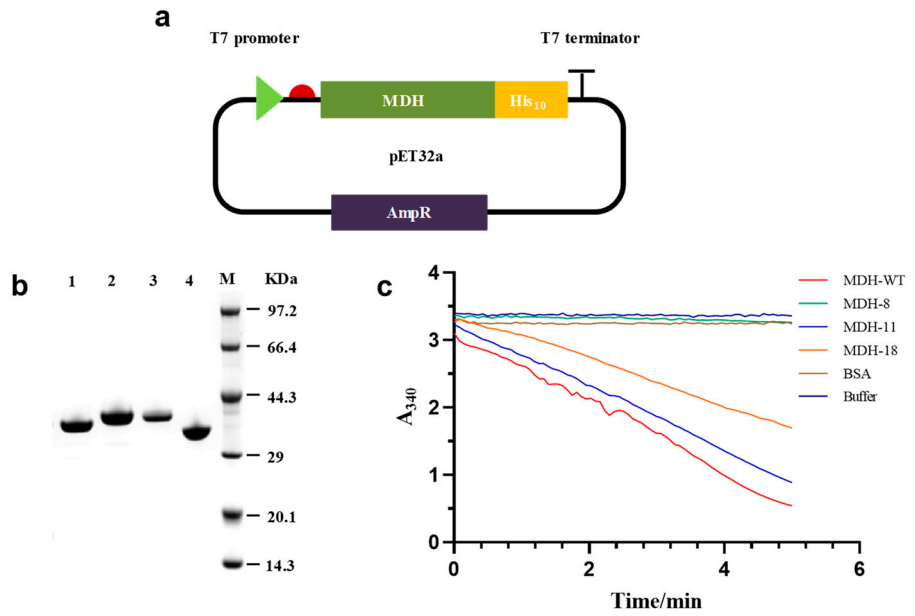

Figure S5: Purification and activity verification of soluble generated MDHs. (a) Schematic diagram of MDH expression vector. **(b)** SDS-PAGE results of the purified soluble MDHs. Lane M, protein molecular weight marker; lane 1, MDH-WT; lane 2, MDH-11; lane 3, MDH-18. **(c)** MDH activity measured by fluorescently monitoring NADH consumption (Supplementary Methods). Bovine serum albumin (BSA) was used as a negative control. MDH-WT was used as a positive control.

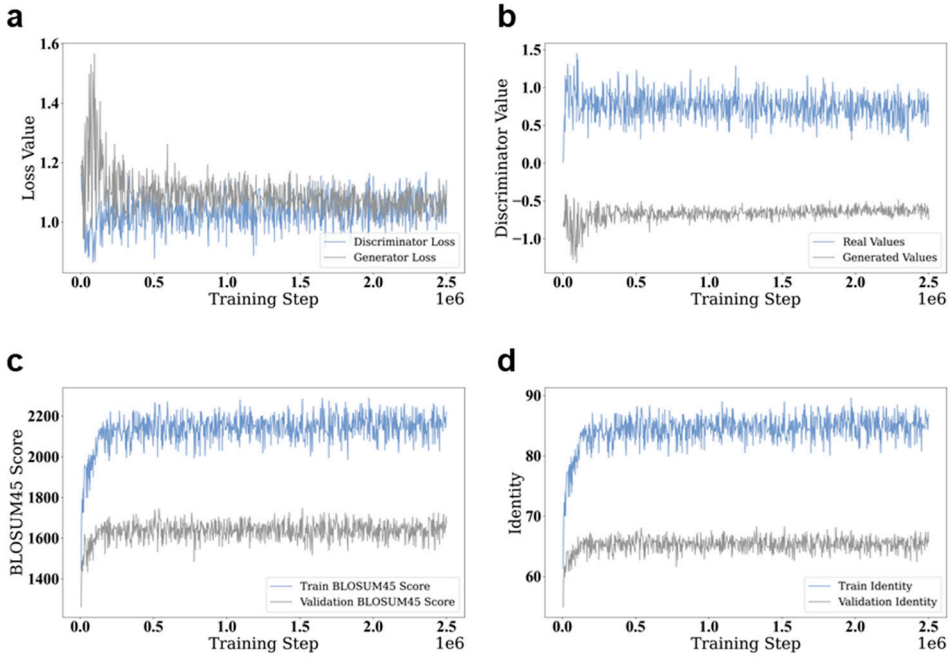

Figure S6: Training process of the selected FBA model. (a) Loss curves of generator and discriminator during the training process. **(b)** Model discriminator scores during the training process. **(c)** BLOSUM45 scores of generated sequences during the training process; **(d)** Identity trends of train and validation set during the training process.

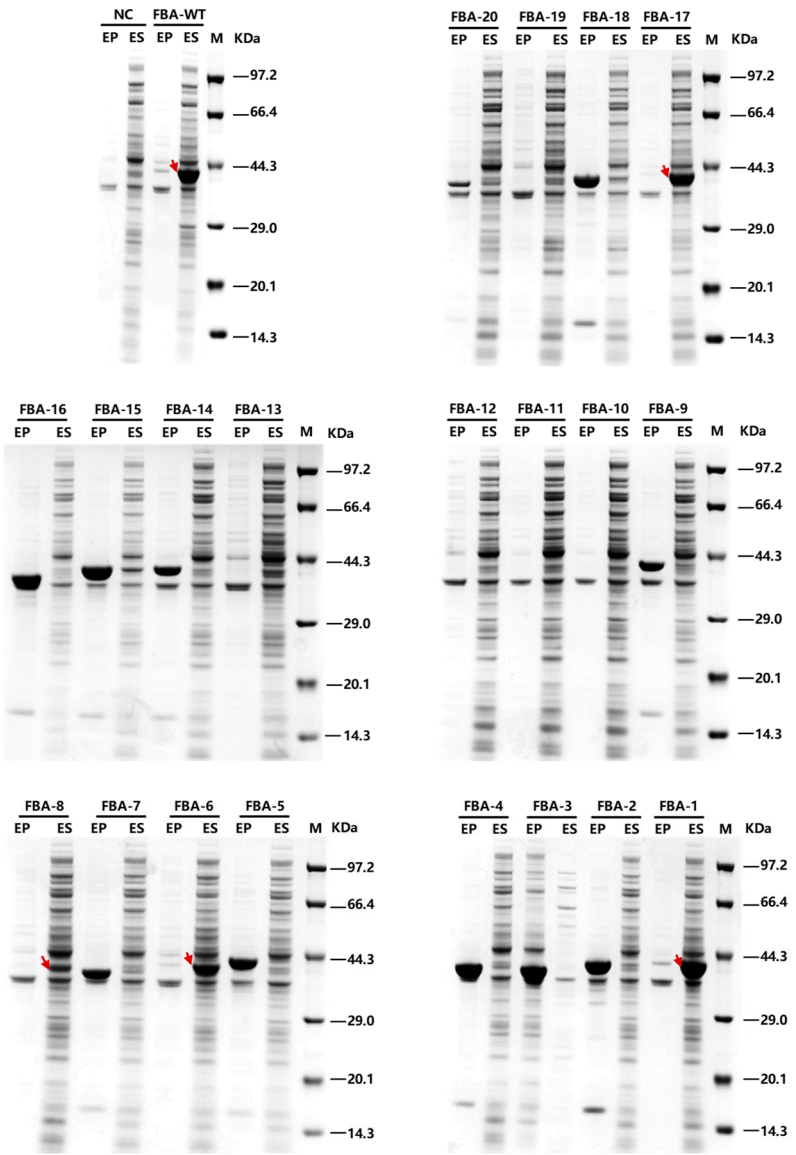

Figure S7: SDS-PAGE results for expression of recombinant FBAs (FBA-1 to FBA-20). SDS-PAGE results of the expression of 20 generated FBAs in *E. coli*. Lane M, protein molecular weight marker; lane ES, expression supernatant; lane EP, expression precipitate; lane NC, negative control using pET32a vector; lane FBA-WT, positive control using FBA-WT. The soluble expressed target protein bands are indicated by a red arrow.

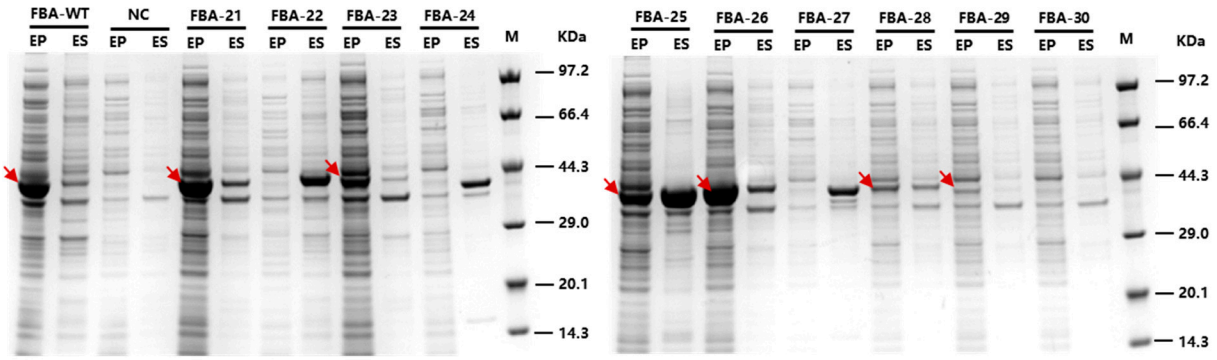

Figure S8: SDS-PAGE results for expression of recombinant reselected FBAs (FBA-21 to FBA-30). SDS-PAGE results of the expression of 10 reselected FBAs in *E. coli*. Lane M, protein molecular weight marker; lane ES, expression supernatant; lane EP, expression precipitate; lane NC, negative control using pET32a vector; lane FBA-WT, positive control using FBA-WT. The soluble expressed target protein bands are indicated by a red arrow.

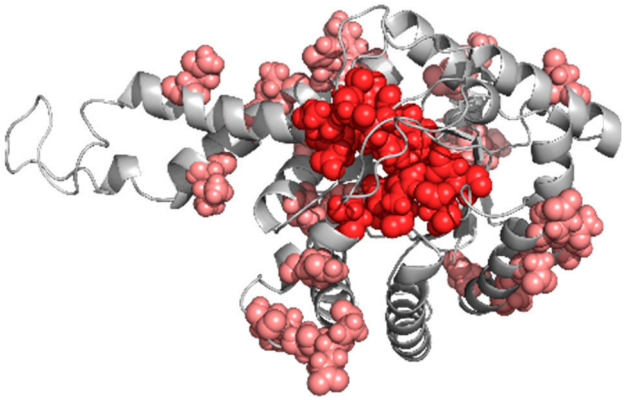

Figure S9: The distribution of differential sites in FBA-23 relative to FBA-WT. Red indicates retained critical residues, salmon indicates mutated residues.

Supplementary Tables

Table S1. Sequence length distribution of malate dehydrogenase dataset.

Length (aa)	Number of Sequences
(0, 100]	30
(100, 200]	60
(200, 300]	119
(300, 400]	16,481
(400, 500]	15
(500, 512]	1
Total	16,706

Table S2. Up-sampling factor for MDH dataset.

Up-sampling Factor[a]	Cluster Sequence Number Range[b]	Number of Sequences[c]
1	(1,000, 3,000]	6,863
2	(500, 1,000]	929
3	(400, 500]	935
5	(200, 400]	1,189
10	(110, 200]	756
20	(70, 110]	1,145
30	(50, 70]	552
40	(40, 50]	325
50	(35, 40]	237
60	(29, 35]	480
70	(25, 29]	174
80	(22, 25]	133
90	(3, 22]	2,988
Validation	(0,3]	192

[a] Indicating how many times the number of samples has increased relative to the original number of samples. [b] Range of original sequence number in the cluster. [c] Number of sequences in all clusters distributed in the range.

Table S3. Sequence identity, solubility, and activity information of the selected 20 generated MDHs.

ID	Length (aa)	Percent identity to closest natural MDH sequence	Number of mutations (aa)	Percent identity _to_ MDH-WT	Predicted solubility [a]	Solubility	Specific activity (U/mg)
MDH-WT	331	-		-	0.41	√	141.60
MDH-1	325	98.15%	6	23.97%	0.45	×	
MDH-2	324	97.84%	7	24.32%	0.43	×	
MDH-3	325	97.84%	8	24.32%	0.45	×	
MDH-4	325	97.54%	8	23.97%	0.45	×	
MDH-5	325	97.54%	8	24.32%	0.44	×	
MDH-6	325	97.53%	9	24.32%	0.44	×	
MDH-7	325	96.62%	11	24.32%	0.43	×	
MDH-8	306	94.14%	17	28.29%	0.75	√	6.39
MDH-9	306	93.49%	19	27.13%	0.77	×	
MDH-10	328	92.07%	26	60.99%	0.72	Not expressed	
MDH-11	328	91.46%	28	60.68%	0.76	√	129.36
MDH-12	328	88.15%	37	63.27%	0.63	Not expressed	
MDH-13	328	88.11%	39	60.99%	0.59	Not expressed	
MDH-14	308	87.62%	39	24.36%	0.73	√	×
MDH-15	309	86.73%	41	25.24%	0.58	×	
MDH-16	309	86.65%	43	25.22%	0.61	×	
MDH-17	328	86.63%	44	62.96%	0.71	Not expressed	
MDH-18	328	86.59%	44	60.12%	0.56	√	86.87
MDH-19	308	86.51%	45	26.50%	0.59	×	
MDH-20	308	85.07%	46	25.53%	0.61	×	

[a] Solubility was predicted using Protein-Sol [19], and predictions greater than 0.45 were predicted to have higher solubility than the average soluble *E. coli* protein in the experimental solubility dataset.

Table S4. Sequence alignment results of MDH-8, MDH-11, and MDH-18 with 13 functional sequences generated by Repecka et al.

ID	Percent identity to MDH-8	Percent identity to MDH-11	Percent identity to MDH-18
pGAN 9	46%	30%	29%
pGAN 22b	47%	25%	24%
pGAN 24	56%	26%	27%
pGAN 25	46%	24%	25%
pGAN 28	56%	27%	26%
pGAN 29	47%	25%	25%
pGAN 31	25%	70%	71%
pGAN 34	24%	68%	70%
pGAN 35	47%	28%	26%
pGAN 37	52%	26%	26%
pGAN 39	26%	53%	50%
pGAN 59	26%	58%	57%
pGAN 60	46%	24%	23%

Table S5. Up-sampling factor for class II FBA dataset.

Up-sampling Factor[a]	Cluster Sequence Number Range[b]	Number of Sequences
1	(2,000, 3000]	5,109
2	(1,000, 2,000]	5,705
4	(500, 1,000]	3,477
10	(200, 500]	1,819
20	(100, 200]	617
30	(67, 100]	371
40	(50, 67]	294
50	(40, 50]	225
60	(30, 40]	206
70	(20, 30]	198
80	(10, 20]	196
90	(0, 10]	347
Validation	(0,3]	160

[a] Indicating how many times the number of samples has increased relative to the original number of samples. [b] Range of original sequence number in the cluster. [c] Number of sequences in all clusters distributed in the range.

Table S6. Sequence length distribution of class II FBA dataset.

Length (aa)	Number of Sequences
(300, 400]	18,550
(400, 500]	13
(500, 512]	1
Total	18,564

Table S7. Evaluation and comparison of the 12 FBA models using interpolation methods. As calculated through interpolating each variable dimension, latent space vectors correlate with protein properties. The table shows the matrix statistics of correlation coefficients between 40 protein properties (columns) and 128 latent vectors (rows).

Model	Parameters			Correlation Coefficient Matrix Statistical Results			
	Steps	Learning Rate	Decay Ratio	Row-Column numbers [a]	Row-Column Average [b]	Column-Row numbers [c]	Column-Row Average[d]
Paper (MDH)	250w	0.001	1:1	101	0.86	28	0.83
Our (MDH)	250w	0.001	5:2	104	0.86	25	0.81
Class II FBA-1	250w	0.001	5:2	116	0.86	25	0.81
				122	0.87	35	0.86
				107	0.85	34	0.84
Class II FBA-2	300w	0.001	5:2	105	0.84	33	0.85
				104	0.83	25	0.81
				110	0.88	30	0.84
Class II FBA-3	250w	0.0005	5:2	86	0.82	26	0.79
				120	0.88	28	0.83
				93	0.83	29	0.81
Class II FBA-4	250w	0.0005	3:2	125	0.91	36	0.88
				102	0.83	28	0.82
				83	0.82	27	0.81

[a] Calculate the maximum absolute value of each column in the matrix, and then count the number of values that are greater than 0.8 among these maximum absolute values. [b] Calculate the maximum absolute value of each column in the matrix, and then calculate the average of these values. [c] Calculate the maximum absolute value of each row in the matrix, and then count the number of values that are greater than 0.8 among these maximum absolute values. [d] Calculate the maximum absolute value of each row in the matrix, and then calculate the average of these values.
The final selected model parameters are highlighted in green in the table.

Table S8. Sequence identity, solubility, and activity information of the selected 20 generated FBAs.

ID	Length (aa)	Percent identity to closest natural FBA sequence	Number of mutations (aa)	Percent identity to FBA-WT	Predicted solubility[a]	Solubility	Specific activity (U/mg)
FBA-WT	360	-		-	0.42	√	4.00
FBA-1	360	99.44%	2	70.03%	0.51	√	×
FBA-2	340	99.12%	3	40.18%	0.69	×	
FBA-3	340	98.53%	5	39.77%	0.67	×	
FBA-4	345	97.97%	7	24.79%	0.48	×	
FBA-5	359	97.49%	9	51.41%	0.69	×	
FBA-6	355	96.90%	11	68.18%	0.52	√	0.62
FBA-7	345	95.94%	13	23.50%	0.50	×	
FBA-8	354	94.92%	18	25.34%	0.51	√	×
FBA-9	359	93.04%	24	66.38%	0.53	×	
FBA-10	354	92.94%	24	25.14%	0.52	Not expressed	
FBA-11	354	91.50%	31	24.43%	0.50	Not expressed	
FBA-12	359	91.04%	32	24.93%	0.54	Not expressed	
FBA-13	359	90.53%	33	23.89%	0.52	×	
FBA-14	359	89.92%	38	24.04%	0.48	×	
FBA-15	343	89.50%	36	39.89%	0.73	×	
FBA-16	343	88.92%	37	40.17%	0.57	×	
FBA-17	358	88.55%	41	79.11%	0.64	√	0.56
FBA-18	340	87.02%	44	38.69%	0.76	×	
FBA-19	354	86.53%	52	24.12%	0.52	Not expressed	
FBA-20	341	85.92%	48	41.07%	0.71	×	

[a] Solubility was predicted using Protein-Sol [19], and predictions greater than 0.45 were predicted to have higher solubility than the average soluble *E. coli* protein in the experimental solubility dataset.

Table S9. Sequence identity, solubility, and activity information of the 10 reselected FBAs.

ID	Length (aa)	Percent identity to closest natural FBA sequence	Number of mutations (aa)	Percent identity to FBA-WT	Predicted solubility [a]	Solubility	Specific activity (U/mg)
FBA-21	359	94.15%	21	93.04%	0.42	√	0.37
FBA-22	358	93.32%	25	90.25%	0.44	×	
FBA-23	359	92.98%	28	92.72%	0.43	√	6.74
FBA-24	345	91.01%	31	83.33%	0.43	×	
FBA-25	359	90.25%	35	86.35%	0.50	√	2.20
FBA-26	359	89.14%	39	88.30%	0.42	√	2.57
FBA-27	354	88.70%	41	83.33%	0.50	×	
FBA-28	358	87.43%	45	81.34%	0.45	√	×
FBA-29	357	86.87%	47	86.63%	0.46	√	2.59
FBA-30	359	85.52%	52	83.01%	0.48	Not expressed	

[a] Solubility was predicted using Protein-Sol [19], and predictions greater than 0.45 were predicted to have higher solubility than the average soluble *E. coli* protein in the experimental solubility dataset.

Supplementary Methods

a. Hyper-parameter setting of training MDH

In the de novo design of apple acid dehydrogenase, the hyper-parameters were set as follows: the ratio of generator and discriminator training steps was 1:1; Adam optimizer parameters were set to 0.0, 0.9; the initial learning rate of the model encoder and discriminator networks was both set to 1×10^{-3}, and dynamically decreased at different rates from step 100,000, with the final learning rate stabilized at 5×10^{-5}; the convolution kernel size was 3×3; batch size was set to 64; every 1200 steps, the Basic Local Alignment Search Tool (BLAST) was used to compare the generated sequences with the training and validation sets, using the BLOSUM45 similarity matrix to calculate sequence consistency and BLOSUM45 matrix scores; the number of training steps was set to 2,500,000; the training time was approximately 10 days.

b. MDH experiment methods.

The wild-type MDH gene (GenBank: KJS04758.1) from *Gammaproteobacteria bacterium* was synthesized as a natural control, referred to as MDH-WT. The sequences generated by ProteinGAN (MDH-1-20) were codon-optimized and synthesized by Beijing Ruiboxingke Biotechnology Co., Ltd. (http://www.ruibiotech.com/). An AAALE linker and four histidine residues were added to the C-terminus of the synthetic sequences, and a 10×His tag (including six histidine residues from the expression vector) was added for cloning into the pET32a expression vector. The construct was transformed into *E. coli* Origami B (DE3) expression strain and the transformed cells were inoculated into 25 mL of Luria Broth (LB) medium containing 50 μg/mL kanamycin, 10 μg/mL tetracycline, and 50 μg/mL ampicillin and cultured overnight at 37°C and 220 rpm. The overnight culture was transferred to 1 L (1:40) of fresh LB medium with the same resistance and cultured at 37°C for 2 h until the cell density reached 0.6-0.8 at 600 nm. Then, 0.2 mM IPTG was added to the culture medium, and the culture was induced overnight at 18°C and 200 rpm.

The cells were collected via centrifugation at 4°C and 4,000 g for 10 min and resuspended in binding buffer (0.1 M sodium phosphate buffer, 0.5 M NaCl, 30 mM imidazole, pH 7.4) at a concentration of 50 OD/mL. The cells were subjected to sonication on ice with a 30% amplitude for a total of 15 min (3 s on/off, 30% power) in a 50 mL centrifuge tube. The cell debris was removed via centrifugation at 15,000 g and 4°C for 20 min, and the supernatant was filtered through a 0.22 μm low protein-binding membrane. The soluble recombinant MDH mutants were purified using a HisTrapTM HP 5 mL affinity column (Cytiva). The column was washed with binding buffer, and the protein was eluted with a gradient of elution buffer (0.1 M sodium phosphate buffer, 0.5 M NaCl, 500 mM imidazole, pH 7.4). The eluted protein was dialyzed against 0.1 M potassium phosphate buffer (pH 7.4). The purified protein was analyzed using sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE) and quantified using BSA standards for further characterization.

MDH activity was measured in 96-well UV-transparent plates (UV-Star microplates, Greiner Bio-One) at 25°C. The reaction mixture (final volume of 200 μL) contained equal amounts of purified protein, freshly prepared 1.2 mM NADH, and 1.6 mM oxaloacetate in 0.1 M potassium phosphate buffer (pH 7.4). The absorbance at 340 nm was continuously monitored for 5 min using an Infinite M200 Pro multimode microplate reader (TECAN). The extinction coefficient for NADH at 340 nm was 6.22 $mM\cdot cm^{-1}$ (εM), and the path length (l) in the microplate was set to 0.5 for calculation. One unit of enzyme activity was defined as the amount of enzyme required to consume 1 μmol of NADH per minute under these conditions.