PNAS

Supplementary Information for
Natural Evolution Provides Strong Hints about Laboratory Evolution of Designer Enzymes

Wen Jun Xie, Arieh Warshel

Email: xwj123@gmail.com; warshel@usc.edu

This PDF file includes:

Supplementary text
SI References
Figures S1 to S8
Tables S1 to S5

Supporting Information Text

Derivation of the Maximum Entropy Model

The maximum entropy (MaxEnt) model has been successfully applied to understanding a variety of phenomena, including protein structure prediction[1–3], protein fitness[4,5], neuron spiking[6], epigenomes[7], etc. In particular, we have recently connected enzyme evolution and catalysis for natural enzymes using the MaxEnt model in a quantitative way.

The MaxEnt model derived from the maximum entropy principle provides the least-biased model to describe the statistics of a protein family. The multiple sequence alignment (MSA) of the target sequence is first constructed to collect the statistics, including the single term (the probability of amino acid at any site $i, < S_i >_{MSA} = \sum_{\boldsymbol{S}} P_{MSA}(\boldsymbol{S}) S_i$) and double term (correlations between amino acids at any two different sites $i, j, < S_{ij} >_{MSA} = \sum_{\boldsymbol{S}} P_{MSA}(\boldsymbol{S}) S_i S_j$), where $P_{MSA}(\boldsymbol{S})$ is the probability of a sequence $\boldsymbol{S}$ in the MSA.

The information theory is then used to quantify the uncertainty by maximizing the information entropy subject to the statistics obtained from MSA. That is,

$$\text{maximize } L = -\sum_{\boldsymbol{S}} P(\boldsymbol{S}) \log P(\boldsymbol{S}) + h_i \left(\sum_{\boldsymbol{S}} P(\boldsymbol{S}) S_i - < S_i >_{MSA} \right) + J_{ij} \left(\sum_{\boldsymbol{S}} P(\boldsymbol{S}) S_i S_j - < S_i S_j >_{MSA} \right) + \alpha \left(\sum_{\boldsymbol{S}} P(\boldsymbol{S}) - 1 \right),$$

where $P(\boldsymbol{S})$ is the model distribution, h_i, J_{ij}, and α are Lagrange multiplier.

The solution of the above optimization problem leads to a Boltzmann distribution for protein sequences $P(\boldsymbol{S}) = e^{-E(\boldsymbol{S})}/Z$, where $E(\boldsymbol{S})$ is the statistical energy with effective temperature as unity and Z is the partition function:

$$E_{\text{MaxEnt}}(\boldsymbol{S}) = \sum_i h_i S_i + \sum_{i>j} J_{ij} S_i S_j.$$

The MaxEnt model can be parameterized by minimizing the cross-entropy $L(\theta)$ between the MSA distribution and the model distribution:

$$\text{minimize } L(\theta) = \sum_p F_p^{MSA} \theta_p - \log Z(\theta),$$

where $F_p^{MSA} = \{< S_i >_{MSA}, < S_i S_j >_{MSA}\}$, $\theta = \{h_i, J_{ij}\}$.

The parameters can be iteratively optimized using the steepest gradient descent optimization with the following expression

$$\theta_p^{t+1} = \theta_p^t - \alpha \left(\frac{\partial L}{\partial \theta_p} \right)_{\theta^t},$$

where $\frac{\partial L}{\partial \theta_p} = F_p^{MSA} - \frac{\partial \log Z(\theta)}{\partial \theta_p} = F_p^{MSA} - F_p^{model}$.

Parameterizing the Maximum Entropy Model

Parameterizing the MaxEnt model is a convex problem in principle. However, the optimization process is highly non-trivial in practice, partially due to the sampling noise in the MSA. We have developed a rigorous and efficient code for the parameterization employing replica exchange Markov chain Monte Carlo, momentum-assisted steepest gradient descent, and message passing interface for parallel computing.[7] The code has been successfully applied in previous studies on epigenome[7] and enzyme[8], where a detailed description of the algorithm is provided.

For Kemp eliminases studied in this manuscript, thirteen replicas with temperatures evenly distributed between 0.8 and 2.0 is adopted in replica-exchange Markov chain Monte Carlo. Each replica lasted for 10^5 steps before swapping. Sixteen processors were used in the message passing interface to obtain the F_p^{MSA}. The L_2-regularization is used to overcome the MSA sampling noise with the weight decay factor is 0.01. Although the pseudo-likelihood approximation[3,4] cannot reproduce the MSA statistics,[9] it is fast to converge. (Please refer to ref [3,4] for more details of pseudo-likelihood approximation.) Here we used our algorithm to fine-tune the parameters obtained from pseudo-likelihood approximation to get faster parameterization. We successfully reproduced and predicted the MSA statistics as shown in Fig S1.

Generating and Processing Multiple Sequence Alignment

We employed Alphafold (v2.0.0)[10] to construct the MSA for the target sequence and the default settings were used. Alphafold has collected the most comprehensive sequence databases till now, including BFD, Uniclust30, UniRef90, and clustered MGnify. The homologs from different databases were combined as the MSA. Large gaps in sequences will affect the downstream processing, in particular, it will increase the gappy region which is not considered in the model. Therefore, sequences with gaps larger than a threshold were disgarded before further processing (Table S1). Enough evolutionarily-related sequences provide converged MSA statistics to parameterize the MaxEnt model.

The MSA was then processed by excluding the sites with more than 30% gaps. The similarity between any two sequences was quantified by hamming distance, and each sequence is assigned a weight of 1/(number of sequences>80% identity) to down-weight similar sequences. Afterward, we calculated the statistics ($< S_i >_{exp}, < S_i S_j >_{exp}$) to constrain the model. It is possible individual single-site mutation in a design from directed evolution is excluded in the analysis if it occurs in the gappy region. Interestingly, we found that there are no such cases in the KE07 series.

The summary of the MSA is shown in Table S1.

SI References

1. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E1293–E1301 (2011).
2. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).
3. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15674–15679 (2013).
4. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).
5. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).
6. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007–1012 (2006).
7. Xie, W. J. & Zhang, B. Learning the formation mechanism of domain-level chromatin states with epigenomics data. *Biophys. J.* **116**, 2047–2056 (2019).
8. Xie, W. J., Asadi, M. & Warshel, A. Enhancing computational enzyme design by a maximum entropy strategy. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2122355119 (2022).
9. Figliuzzi, M., Barrat-Charlaix, P. & Weigt, M. How pairwise coevolutionary models capture the collective residue variability in proteins? *Mol. Biol. Evol.* **35**, 1018–1027 (2018).
10. Jumper, J. *et al.* Highly accurate protein structure prediction with Alphafold. *Nature* **596**, 583–589 (2021).

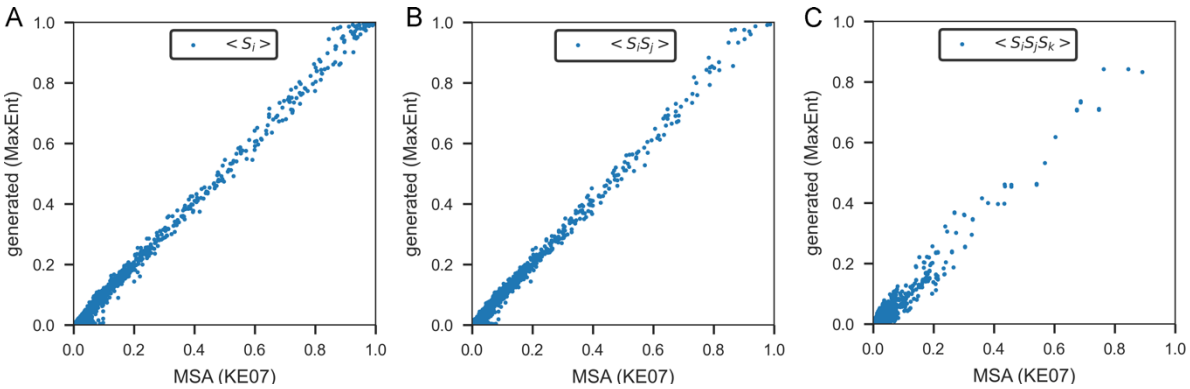

Fig. S1. Comparison of sequence statistics obtained from the MaxEnt model and natural MSA. (a) $< S_i >$; (b) $< S_i S_j >$; (c) $< S_i S_j S_k >$. The $< S_i >$ and $< S_i S_j >$ are explicitly used as constraints in the parameterization while $< S_i S_j S_k >$ does not. To avoid overplotting, 10E5 and 10E6 randomly sampled data are shown for both $< S_i S_j >$ and $< S_i S_j S_k >$, respectively. The excellent reproduction of $< S_i >$ and $< S_i S_j >$, and prediction of $< S_i S_j S_k >$ validates our implementation.

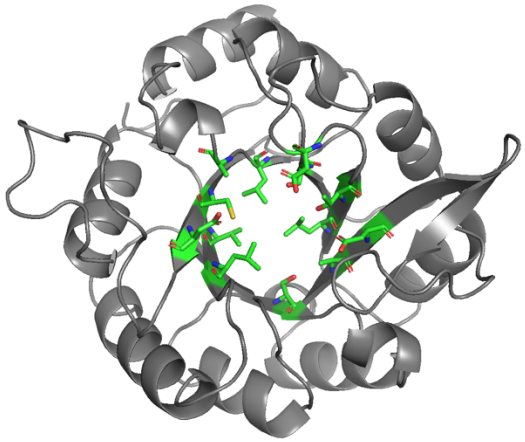

Fig. S2. The KE07 series start with thirteen mutations compared to the TIM barrel template (PDB ID: 1THF). The mutations are highlighted.

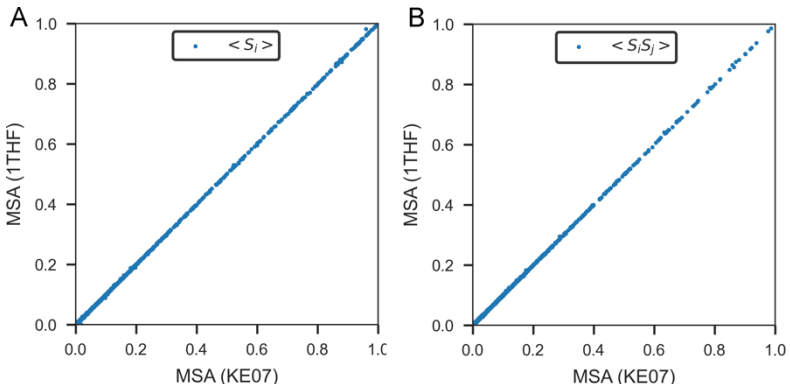

Fig. S3. Comparison between sequence statistics obtained from different MSA construction. The initial design of the KE07 and template (PDB ID: 1THF) were used as target sequences to construct MSA. (a) $< S_i >$; (b) $< S_iS_j >$. To avoid overplotting, 10E5 randomly sampled data is shown for $< S_iS_j >$.

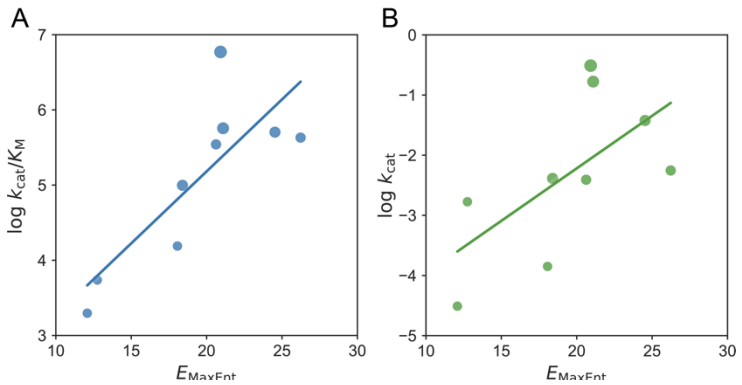

Fig. S4. The maximum entropy model for natural sequences is correlated with the catalytic power of designs in the KE07 directed evolution with the same number of mutations. Strong positive correlations between E_{MaxEnt} and $\log k_{\mathrm{cat}}/K_{\mathrm{M}}$ (A) and $\log k_{\mathrm{cat}}$ (B) for the KE07 directed evolution are shown with the least-squares regression lines included. The correlation values shown in panels A and B are 0.81 (*p*-value=0.0078) and 0.63 (*p*-value=0.069), respectively. Each dot represents a design with larger dot size indicating later rounds in directed evolution.

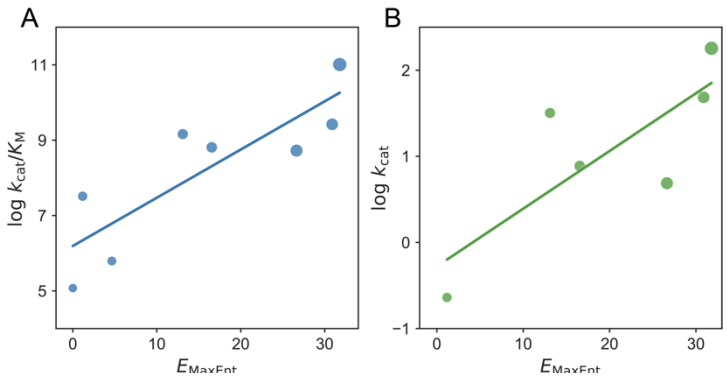

Fig. S5. The maximum entropy model for natural sequences is correlated with the catalytic power of designs in the KE59 directed evolution. Strong positive correlations between E_{MaxEnt} and $\log k_{\mathrm{cat}}/K_{\mathrm{M}}$ (A) and $\log k_{\mathrm{cat}}$ (B) for the KE59 directed evolution are shown with the least-squares regression lines included. The correlation values shown in panels A and B are 0.85 (*p*-value=0.0072) and 0.80 (*p*-value=0.058), respectively. Each dot represents a design with larger dot size indicating later rounds in directed evolution. In the KE59 series, we shifted E_{MaxEnt} by a constant so that the initial design has a zero value.

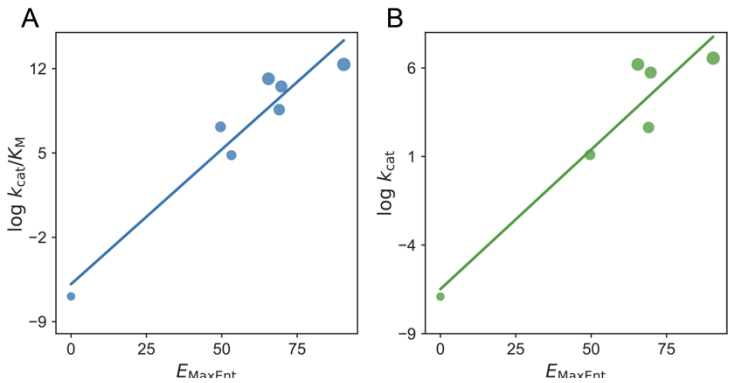

Fig. S6. The maximum entropy model for natural sequences is correlated with the catalytic power of designs in the HG directed evolution. Strong positive correlations between E_{MaxEnt} and $\log k_{\mathrm{cat}}/K_{\mathrm{M}}$ (A) and $\log k_{\mathrm{cat}}$ (B) for the HG directed evolution are shown with the least-squares regression lines included. The correlation values shown in panels A and B are 0.96 (p-value<0.001) and 0.95 (p-value=0.003), respectively. Each dot represents a design with larger dot size indicating later rounds in directed evolution. In the HG series, we shifted E_{MaxEnt} by a constant so that the enzyme template has a zero value.

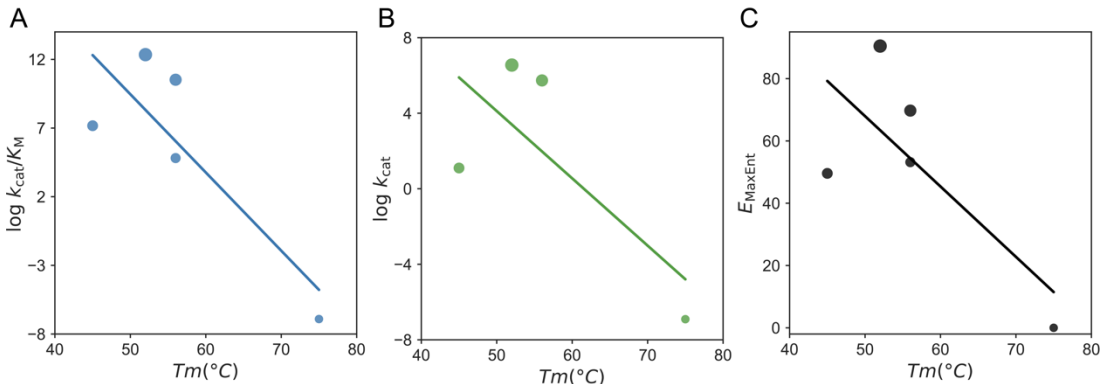

Fig. S7. Enzyme catalytic power and stability trade-off in the HG directed evolution. (A-B) Strong negative correlations between protein stability quantified by melting temperature (T_m) and enzyme catalytic power expressed by $\log k_{cat}/K_M$ (A) and $\log k_{cat}$ (B) for the HG directed evolution. (C) A strong negative correlation between T_m and E_{MaxEnt}. The correlation values shown in panels A, B and C are -0.84 (*p*-value=0.077), -0.74 (*p*-value=0.14) and -0.74 (*p*-value=0.26), respectively. Each dot represents a design with larger dot size indicating later rounds in directed evolution.

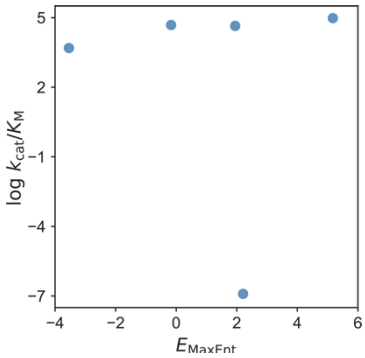

Fig. S8. The active sites do not correlate with the MaxEnt model for natural sequences. Data used are shown in Table S3.

Table S1. Summary of MSA information

System	Target sequence	Sequence number
KE07	KE07 initial design	15685 (gap<5)
KE07	PDB ID: 1THF	15822 (gap<5)
KE59	KE59 initial design	5174 (gap<40)
HG	PDB ID: 1GOR	5907 (gap<50)

The sequence of the KE07 initial design is
'MLAKRIIAALIVKDGRVVKGSNFENLRDSGDPVELGKFYSEIGIDELSFWDITASVEKRKTMLELVEKVAEQ
IDIPFTVGGGIHDFETASELILRGADKVEINTAAVENPSLITQIAQTFGSQAVVVYIAAKRVDGEFMVFTYSG
KKNTGILLRDWVVEVEKRGAGEIVLGSIDRLGTKSGYDTEMIRFVRPLTTLPIIAHGGAGKMEHFLEAFLAG
ADAAKANSVFHFREIDVRELKEYLKKHGVNVRLEGL';

The sequence of the PDB ID:1THF is
'MLAKRIIACLDVKDGRVVKGSNFENLRDSGDPVELGKFYSEIGIDELVFLDITASVEKRKTMLELVEKVAE
QIDIPFTVGGGIHDFETASELILRGADKVSINTAAVENPSLITQIAQTFGSQAVVVAIDAKRVDGEFMVFTYS
GKKNTGILLRDWVVEVEKRGAGEILLTSIDRDGTKSGYDTEMIRFVRPLTTLPIIASGGAGKMEHFLEAFLA
GADAALAASVFHFREIDVRELKEYLKKHGVNVRLEGL';

The sequence of the KE59 initial design is
'PRYLKGWLKDVVQLSLRRPSFRASRQRPIISLNERILEFNKRNITAIIAVYKRKSPSGLDVERDPIEYSKFM
ERYAVGLVILTEEKYFNGSYETLRKIASSVSIPILMWDFIVKESQIDDAYNLGADTVGLIVKILTERELESLLE
YARSYGMEPAIVINDENDLDIALRIGARFIIISSRDLETLEINKENQRKLISMIPSNVVKVAASGISERNEIEEL
RKLGVNAFEIGSSLMRNPEKIKEFIL';

The sequence of the PDB ID:1GOR is
'QAAQSVDQLIKARGKVYFGVATDQNRLTTGKNAAIIQADFGQVTPENSMKWDATEPSQGNFNFAGADYL
VNWAQQNGKLIRGHTLVWHSQLPSWVSSITDKNTLTNVMKNHITTLMTRYKGKIRAWDVVNEAFNEDGS
LRQTVFLNVIGEDYIPIAFQTARAADPNAKLYINDYNLDSASYPKTQAIVNRVKQWRAAGVPIDGIGSQTHL
SAGQGAGVLQALPLLASAGTPEVAITELDVAGASPTDYVNVVNACLNVQSCVGITVWGVADPDSWRAST
TPLLFDGNFNPKPAYNAIVQDLQQ'.

Table S2. KE07 directed evolution

variant	mutant	k_{cat} (s^{-1})	Km (mM)	k_{cat}/Km ($s^{-1}M^{-1}$)	Tm (°C)	E_{MaxEnt}
KE07 WT	WT	0.018	1.4	12.2	>95	0.00
Table1_1	E101A	0.0009	0.29	3.5		-3.59
Table1_2	K222A	0.03	1.3	22.7		-3.54
R2 11/10D	K19E/K146T/G202R/N224D/Q123R	0.0213	0.31	66	>95	18.07
R2 11/6F	K19E/I199T/N224D/I44T/L152P/F229S	0.0187	0.52	36		24.56
R2 11/9A	I7T/K146E/N224D/F86L/M207V	0.0624	1.48	42		12.75
R2 2/8B	K19E/K146T/N224D/F189S/V226A	0.011	0.41	27		12.08
R3 I3/10A	I7Q/K146T/G202R/N224D/F86L/F229S	0.206	0.48	425	~95	19.78
R3 I3/11B	I7H/K19E/G202R/N224D/M207I	0.105	0.38	279		26.23
R3 M7/4E	I7V/K19E/K146E/I199Q/N224D	0.09	0.36	255		20.62
R4 1A/7D	I7Q/K146E/G202R/N224D/F229S	0.241	0.8	300		24.53
R4 1E/11H	I7D/K146E/G202R/N224D	0.699	2.4	291	~86	18.48
R4 1H/8C	I7Q/K146E/I199V/N224D/F227L	0.092	0.625	148		18.40
R4 2C/11G	I7S/K19E/G202R/N224D	0.128	0.444	287		19.27
R4 2D/10F	I7V/K146E/I199Q/N224D/K162P/I173A/L176I/F229S	0.088	0.46	191		35.79
R4 2F/2G	I7T/K146T/I199Q/F86L/I173V/L176D/F227L	0.14	1.5	99		13.75
R5 10/3B	I7D/G202R/N224D/V12M	0.49	0.59	836	79	20.04
R5 9/11C	I7D/K146E/G202R/N224D/L47I	0.46	1.5	316		21.09
R6 3/7F	I7D/K19E/K146T/G202R/N224D	0.6	0.69	872	~87	20.92
R6 8/11D	I7D/G202R/N224D/V12M/A54D/G171A	0.48	0.58	827		29.70
R7 1/1F	I7D/K146T/G202R/N224D/V12M/F77I/H84Y/F229S	0.553	0.42	1310		32.97
R7 1/3H	I7D/K146T/G202R/N224D/V12L/F77I/H84Y/M207T/F229S	0.76	0.54	1414	76	36.63
R7 1/9F	I7D/K146T/G202R/N224D/V12M/L47I/F77I/G171A/F229S	0.66	0.44	1490		32.04
R7 10/11G	I7D/K146T/G202R/N224D/V12M/F77I/I102F/F229S	1.37	0.54	2590	76	33.91
R7 2/5B	I7D/G202R/N224D/F77I	1.2	0.86	1388		19.62

*References: Nature, 2008, 453, 190 (directed evolution); J Mol Biol, 2010, 396, 1025 (melting temperature)

*The melting temperature for some designs are provided as a range in the reference. The digits in the column of Tm were used in plotting Fig 3.

Table S3. KE07 catalytic-active remote region

variant	mutant	k_{cat} (s^{-1})	Km (mM)	k_{cat}/Km ($s^{-1}M^{-1}$)	E_{MaxEnt}	Note
R7 10/11G	WT	0.81	0.407	1990	0.00	catalytic-active remote region
R7_R16Q	R16Q	0.57	0.589	968	3.32	catalytic-active remote region
R7_N25S	N25S	0.58	0.479	1221	2.56	catalytic-active remote region
R7_I52A	I52A	0.51	0.514	992	5.85	catalytic-active remote region
R7_M62A	M62A	0.64	0.542	1181	2.87	catalytic-active remote region
R7_H84Y	H84Y	0.77	0.497	1549	3.91	catalytic-active remote region
R7_K132N	K132N	0.75	0.56	1339	4.64	catalytic-active remote region
R7_I199S	I199S	0.33	0.771	428	6.23	catalytic-active remote region
R7_I199F	I199F	0.26	0.564	461	9.42	catalytic-active remote region
R7_I199A	I199A	0.23	1.467	155	10.18	catalytic-active remote region
R7_S48N	S48N	0.1	0.6897	145	5.17	active site
R7_Y128F	Y128F			0	2.21	active site
R7_H201A	H201A			108	-0.17	active site
R7_H201K	H201K	0.562	5.4114	104	1.95	active site
R7_K222A	K222A			40	-3.54	active site

*References: Phys Chem Chem Phys, 2016, 18, 19386

*The inactive mutants were assigned a relatively small k_{cat} or k_{cat}/Km value (0.001) during log transformation to avoid singularity.

Table S4. KE59 directed evolution

variant	mutant	k_{cat} (s^{-1})	Km (mM)	k_{cat}/Km ($s^{-1}M^{-1}$)	E_{MaxEnt}
KE59	WT			160	0.00
R1-7/10H	W7A/F21L/N33R/S69A/T94A/N163E/F175I/F245L			328	4.65
R2-4/3D	K9E/L14R/F21V/N33K/S69A/T94D/E142K/N160H/V80A	0.528	0.29	1833	1.17
R4-5/11B	K9E/N33K/S69A/T94D/V80A/R181H/A208V/R222Y/L247Q	4.5	0.48	9524	13.10
R5-11/5F	K9E/F21L/N33K/S69A/T94D/I44N/V80A/R181H/A208V/R222Y/L247Q	2.43	0.36	6706	16.55
R8-2/7A	K9E/F21L/N33R/D60N/S69A/Y75G/T94D/E142K/L247G/R22H/I44N/A76V/V80A/L107M/F111I/R181H/I200V/N203D/A208V	5.4	0.44	12350	30.89
R9-1/4A	K9E/L14R/F21V/N33K/S69A/Y75G/T94D/Y151L/N160H/L16Q/I48M/A76V/V80A/F111I/R181H/A208V/R222Y/L247Q	1.99	0.32	6147	26.64
R13-3/11H	K9E/L14R/F21V/N33K/S69A/Y75G/T94D/Y151L/N160H/L16Q/I48M/A76V/V80A/I104V/F111I/S179T/R181H/K190N/A208V/R222Y/S233T/L247Q	9.53	0.16	60430	31.79

*References: PNAS, 2012, 109, 10358

Table S5. HG directed evolution

variant	mutant	k_{cat} (s^{-1})	Km (mM)	k_{cat}/Km ($s^{-1}M^{-1}$)	Tm (°C)	E_{MaxEnt}
PDB_ID: 1GOR	WT	0		0	75	0.00
HG-2	Q42M/T44W/R81G/H83G/T84M/N130G/N172M/A234S/T236L/E237M/T265S/W267F			123.2	56	53.19
HG3	Q42M/T44W/R81G/H83G/T84M/N130G/N172M/A234S/T236L/E237M/W267F	3	2.4	1300	45	49.58
HG3.3b	V6I/Q42M/T44W/K50H/R81G/H83G/T84C/S89R/Q90D/A125N/N130G/N172M/A234S/T236L/E237M/W267F	14	2.6	5400		69.03
HG3.7	V6I/Q37K/Q42M/T44W/K50Q/R81G/H83G/T84C/S89R/Q90H/A125N/N130G/N172M/A234S/T236L/E237M/W267F	310	8.3	37000	56	69.74
HG3.14	V6I/Q37K/Q42M/T44W/K50Q/R81G/G82A/H83G/T84C/Q90H/T105I/A125T/N130G/T142N/N172M/T208M/A234S/T236L/E237M/W267F/T279S/D300N	490	7	70000		65.47
HG3.17	V6I/Q37K/Q42M/T44W/N47E/K50Q/R81G/G82A/H83G/T84C/S89N/Q90F/T105I/A125T/N130G/T142N/N172M/T208M/A234S/T236L/E237M/W267M/W275A/R276F/T279S/D300N	700	3	230000	52	90.46

*References: PNAS, 2012 109, 3790 (activity for WT and HG-2, and their melting temperatures were obtained from Fig S1B); Nature, 2013, 503, 418 (activity for HG3~HG3.17; and three of them having measured melting temperature obtained from Fig 1D).

*The inactive mutants were assigned a relatively small k_{cat} or k_{cat}/Km value (0.001) during log transformation to avoid singularity.