Supporting Information

Machine-Learning-Guided Mutagenesis for Directed Evolution of Fluorescent Proteins

Yutaka Saito,[1,2]† Misaki Oikawa,[3]† Hikaru Nakazawa,[3] Teppei Niide,[3] Tomoshi Kameda,[1] Koji Tsuda,[4,5,6]* Mitsuo Umetsu[3,5]*

[1]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan.

[2]Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), National Institute of Advanced Industrial Science and Technology (AIST), 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan.

[3]Department of Biomolecular Engineering, Graduate School of Engineering, Tohoku University, 6-6-11 Aoba, Aramaki, Aoba-ku, Sendai 980-8579, Japan.

[4]Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan.

[5]Center for Advanced Intelligence Project, RIKEN, 1-4-1 Nihombashi, Chuo-ku, Tokyo, 103-0027, Japan.

[6]Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, 1-2-1 Sengen, Tsukuba, Ibaraki 305-0047, Japan.

† These authors contributed equally to this work.

* Joint corresponding authors:

Koji Tsuda (tsuda@k.u-tokyo.ac.jp)

Mitsuo Umetsu (mitsuo@tohoku.ac.jp)

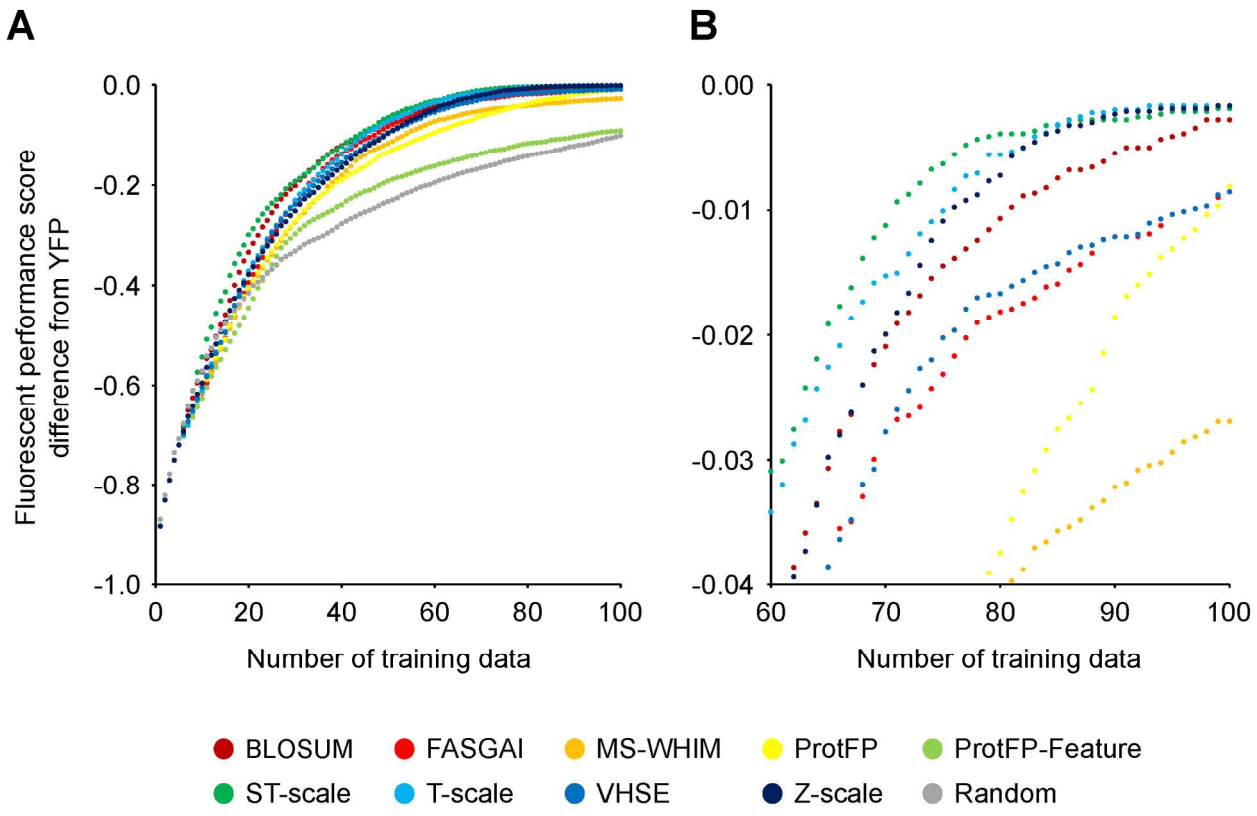

Fig. S1. Benchmark of amino acid descriptors. For each amino acid descriptor, COMBO was trained by an Bayesian optimization procedure. The fluorescence performance score of the current best example is plotted for a given number of training data. Z-scale, T-scale, ST-scale showed better performance than the other descriptors. (**B**) is a magnified version of the late stage of training in (**A**).

Table S1. The reference GFP and YFP used in this study. The amino acid sequences are shown with the four mutations colored in red (S65G, S72A, H77Y, and T203F).

Reference GFP

```
1         10          20          30          40          50          60
MSKGEELFTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTLKFICTT GKLPVPWPTL

61        70          80          90         100         110         120
VTTFSYGVQC FSRYPDHMKR HDFFKSAMPE GYVQERTISF KDDGNYKTRA EVKFEGDTLV

121      130         140         150         160         170         180
NRIELKGIDF KEDGNILGHK LEYNYNSHNV YITADKQKNG IKANFKIRHN IEDGSVQLAD

181      190         200         210         220         230        238
HYQQNTPIGD GPVLLPDNHY LSTQSALSKD PNEKRDHMVL LEFVTAAGIT HGMDELYK
```

Reference YFP

```
1         10          20          30          40          50          60
MSKGEELFTG VVPILVELDG DVNGHKFSVS GEGEGDATYG KLTLKFICTT GKLPVPWPTL

61        70          80          90         100         110         120
VTTFGYGVQC FARYPDYMKR HDFFKSAMPE GYVQERTISF KDDGNYKTRA EVKFEGDTLV

121      130         140         150         160         170         180
NRIELKGIDF KEDGNILGHK LEYNYNSHNV YITADKQKNG IKANFKIRHN IEDGSVQLAD
```

181 190 200 210 220 230 238

HYQQNTPIGD GPVLLPDNHY LS**F**QSALSKD PNEKRDHMVL LEFVTAAGIT HGMDELYK

Table S2. Sequences of 22-c trick primers for point saturation and site-directed random mutagenesis in GFP. XXX is NDT, VHG, or TGG. For PCR, the forward primers with NDT, VHG, and TGG was mixed at the ratio of 12:9:1, respectively. YYY is AHN, CDB, or CCA. For PCR, the forward primers with AHN, CDB, and CCA was mixed at the ratio of 12:9:1, respectively.

Primer	Sequence
External forward	GTGACGGTCTTCCCCTCTAG
External reverse	GAGTGCGGCCGCTTTGTAGAGCTCATCCATGCCATGTGTAAT
S65-forward	CTTGTCACTACTTTC**XXX**TATGGTGTTCAATGC
S65-reverse	GCATTGAACACCATA**YYY**GAAAGTAGTGACAAG
S72-forward	GGTGTTCAATGCTTT**XXX**CGTTATCCGG
S72-reverse	CCGGATAACG**YYY**AAAGCATTGAACACC
H77- forward	CCCGTTATCCGGAT**XXX**ATGAAACGGC
H77- reverse	GCCGTTTCAT**YYY**ATCCGGATAACGGG
T203- forward	CATTACCTGTCG**XXX**CAATCTGCCCTT
T203-reverse	AAGGGCAGATTG**YYY**CGACAGGTAATG

Table S3. Sequences of mutagenesis primers for preparing the top 78 candidate variants proposed by machine learning. Red, blue and orange characters are the mutagenesis sites for the residues 65, 72, and 77, respectively.

Primer	Sequence
External forward	GTGACGGTCTTCCCCTCTAG
External reverse	GAGTGCGGCCGCTTTGTAGAGCTCATCCATGCCATGTGTAAT
Group A	
S65S72H77-forward 1	ACTACTTTCGSGTATGGTGTTCAATGCTTTGSGCGTTATCCG
	GATTWTATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 1	GCCGTTTCATAWAATCCGGATAACGCSCAAAGCATTGAACAC
	CATACSCGAAAGTAGTGACAAGTGTTGGCC
S65S72H77-forward 2	ACTACTTTCGSGTATGGTGTTCAATGCTTTGSGCGTTATCCG
	GATCATATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 2	GCCGTTTCATATGATCCGGATAACGCSCAAAGCATTGAACAC
	CATACSCGAAAGTAGTGACAAGTGTTGGCC
Group B	
S65S72H77-forward 3	ACTACTTTCGGGTATGGTGTTCAATGCTTTTSTCGTTATCCG
	GATTWTATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 3	GCCGTTTCATAWAATCCGGATAACGASAAAAGCATTGAACAC
	CATACCCGAAAGTAGTGACAAGTGTTGGCC
S65S72H77-forward 4	ACTACTTTCGGGTATGGTGTTCAATGCTTTTSTCGTTATCCG
	GATCATATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 4	GCCGTTTCATATGATCCGGATAACGASAAAAGCATTGAACAC
	CATACCCGAAAGTAGTGACAAGTGTTGGCC
Group C	
S65S72H77-forward 5	ACTACTTTCGCGTATGGTGTTCAATGCTTTTSTCGTTATCCG
	GATTWTATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 5	GCCGTTTCATAWAATCCGGATAACGASAAAAGCATTGAACAC

	CATACGCGAAAGTAGTGACAAGTGTTGGCC
S65S72H77-forward 6	ACTACTTTCGCGTATGGTGTTCAATGCTTTTSTCGTTATCCG
	GATCATATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 6	GCCGTTTCATATGATCCGGATAACGASAAAGCATTGAACAC
	CATACGCGAAAGTAGTGACAAGTGTTGGCC
Group D	
S65S72H77-forward 7	ACTACTTTCGGGTATGGTGTTCAATGCTTTGSGCGTTATCCG
	GATTGGATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 7	GCCGTTTCATCCAATCCGGATAACGCSCAAAGCATTGAACAC
	CATACCCGAAAGTAGTGACAAGTGTTGGCC
Group E	
S65S72H77-forward 8	ACTACTTTCGGGTATGGTGTTCAATGCTTTTCTCGTTATCCG
	GATTATATGAAACGGCATGACTTTTTCAAGAG
S65S72H77-reverse 8	GCCGTTTCATATAATCCGGATAACGAGAAAAGCATTGAACAC
	CATACCCGAAAGTAGTGACAAGTGTTGGCC

Table S4. Detected GFP variants from the top 78 GFP variants. The 63 detected variants are red-colored in the table.

Ranking	Amino acid residue				Ranking	Amino acid residue			
	65	72	77	203		65	72	77	203
1	G	A	Y	Y	40	A	A	Y	W
2	G	A	H	Y	41	G	S	Y	H
3	G	A	F	Y	42	A	A	Y	H
4	G	A	H	H	43	A	A	F	W
5	G	G	H	Y	44	A	A	F	H
6	G	G	Y	Y	45	G	C	H	W
7	G	A	Y	H	46	G	C	H	H
8	G	G	F	Y	47	G	S	H	W
9	G	A	F	H	48	G	A	W	H
10	G	G	H	H	49	G	C	Y	H
11	G	A	H	W	50	G	C	Y	W
12	G	A	Y	W	51	A	G	H	H
13	G	G	Y	H	52	G	S	F	H
14	G	A	H	F	53	G	S	Y	W
15	A	A	H	Y	54	G	C	F	H
16	G	G	F	H	55	G	G	W	Y
17	A	A	Y	Y	56	A	G	Y	H
18	G	G	Y	F	57	A	A	Y	F
19	G	A	F	W	58	A	G	F	H
20	G	A	F	F	59	A	A	H	F
21	A	A	F	Y	60	G	C	F	W
22	G	S	Y	Y	61	A	A	F	F
23	G	G	H	F	62	A	G	H	W
24	G	S	H	Y	63	G	S	F	W
25	G	C	H	Y	64	A	C	F	Y
26	G	G	F	F	65	G	A	W	F
27	G	G	H	W	66	A	C	H	Y
28	G	C	Y	Y	67	A	G	Y	F

29	G	G	Y	W	68	G	S	Y	F
30	G	S	F	Y	69	G	A	W	W
31	A	G	H	Y	70	A	G	Y	W
32	G	C	F	Y	71	A	G	F	W
33	A	G	Y	Y	72	A	C	Y	Y
34	G	A	W	Y	73	A	S	Y	Y
35	A	G	F	Y	74	A	G	H	F
36	A	A	H	W	75	G	G	W	H
37	G	G	F	W	76	A	G	F	F
38	A	A	H	H	77	A	S	H	Y
39	G	S	H	H	78	A	S	F	Y

Table S5. Benchmark of amino acid descriptors. ST + T + Z: feature vectors of ST-scale, T-scale, and Z-scale are concatenated into a single vector and used in machine learning. All: feature vectors of all descriptors are concatenated into a single vector and used in machine learning.

Descriptor	Rank of the reference YFP
ST-scale	11724
T-scale	1313
Z-scale	10280
ST + T + Z	4715
All	10292

Table S6. Sequence similarity between the predicted variants and the training data.

For the 12 predicted variants better than the reference YFP, the yellow fluorescent variants in the training data with the most similar sequences are shown. Also shown is the number of differed residues out of the four mutated positions.

12 variants better than the reference YFP	Most similar variant(s) in the training data	#Different residues
GAYY	GAYF	1
GAHY	GAYF, SSHY	2
GAFY	GAYF	2
GGHY	SSHY	2
GGYY	GAYF	2
GGFY	GAYF, SSHY, GVTH	3
GCHY	SSHY	2
GCYY	GAYF	2
GSFY	SSHY	2
GCFY	GAYF	3
GAWY	GAYF	2
GGWY	GAYF, SSHY, GVTH	3

Data file S1. Complete list of all unknown variants ranked by machine learning.

(given as a separate Excel file)