Supplementary Information

Application of machine learning in the quantitative analysis of surface characteristics of highly abundant cytoplasmic proteins: Toward AI-based biomimetics

Joo A Moon[1], Guanghao Hu[1], and Tomohiro Hayashi[1,2]*

[1] Department of Materials Science and Engineering, School of Materials and Chemical Technology, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8502, Japan

[2] The Institute for Solid State Physics, The University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba 277-0882, Japan

*corresponding author: tomo@mac.titech.ac.jp

Contents:

Table S1: Hyperparameters of machine learning models in a single-shot trial

Model	Full Name	Hyperparameters
KNN	k-Nearest Neighbors	{'algorithm': 'auto', 'metric': 'manhattan', 'n_neighbors': 11, 'weights': 'distance'}
LR	Logistic Regression	{'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
RF	Random Forest	{'max_depth': 10, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 1000}
SVM	Supported Vector Machine	{'C': 1.0, 'gamma': 'scale', 'kernel': 'rbf'}

Table S2: Statistical summary of surface descriptors for all protein data set.

	s_phobic_avg	s_neg_area_avg	s_pos_area_avg	norm_s_b	FD	s_bs	s_do
Extracellular Proteins							
count	331						
mean	-1.101	0.137	0.193	0.231	2.192	13.009	52.988
std	0.366	0.043	0.048	0.102	0.060	9.479	12.690
min	-1.998	0.024	0.061	-0.049	2.066	0	7.23
25%	-1.3365	0.1135	0.1635	0.159	2.144	5.694	45.2715
50%	-1.148	0.137	0.19	0.251	2.190	11.609	54.704
75%	-0.899	0.1605	0.221	0.298	2.235	19.968	60.874
max	0.719	0.305	0.364	0.491	2.372	47.337	88.165
Highly Abundant Cytoplasmic (HAC) Proteins							
count	337						
mean	-1.569	0.170	0.257	0.185	2.180	8.691	50.151
std	0.452	0.076	0.090	0.098	0.069	8.643	16.337
min	-2.74	0.022	0.031	-0.02	2.045	0	1
25%	-1.854	0.112	0.195	0.126	2.117	2.578	39.872
50%	-1.6	0.169	0.236	0.192	2.178	6.733	48.05
75%	-1.385	0.22	0.309	0.26	2.232	12.685	58.326
max	0.294	0.555	0.495	0.512	2.325	57.663	99.187

*Full name of each descriptor:

s_phobic_avg: Average Surface Hydrophobicity

s_neg_area_avg: Average Negatively Charged Surface Area

a_pos_area_avg: Average Positively Charged Surface Area

norm_s_b: Average Normalized Surface b-factor

FD: Average Surface Roughness

s_bs: Surface Beta Strands Composition

s_do: Surface Disordered Regions Composition

Table S3: pKa Values used to calculate net charge[1]

1-letter code	Amino Acid	pK_a Values		
		pK_{a1}(-COOH)	pK_{a2}(-NH_3^+)	pK_{aR}(R group)
A	Alanine	2.35	9.87	
C	Cysteine	2.05	10.25	8.00
D	Aspartic Acid	2.10	9.82	3.86
E	Glutamic Acid	2.10	9.47	4.07
F	Phenylalanine	2.58	9.24	
G	Glycine	2.35	9.78	
H	Histidine	1.77	9.18	6.10
I	Isoleucine	2.32	9.76	
K	Lysine	2.18	8.95	10.53
L	Leucine	2.33	9.74	
M	Methionine	2.28	9.21	
N	Asparagine	2.02	8.80	
P	Proline	2.00	10.60	
Q	Glutamine	2.17	9.13	
R	Arginine	2.01	9.04	12.48
S	Serine	2.21	9.15	
T	Threonine	2.09	9.10	
V	Valine	2.29	9.72	
W	Tryptophan	2.38	9.39	
Y	Tyrosine	2.20	9.11	10.07

Surface net charge was calculated using Table S3 under physiological conditions (pH=7). General expressions derived from Henderson-Hasselbalch equation calculated negative (Q^-) and positive (Q^+) charge as follows[2]:

$$Q^- = \frac{-1}{1 + 10^{-(pH-pKa)}}$$

$$Q^+ = \frac{1}{1 + 10^{+(pH-pKa)}}$$

Thus, surface net charge of protein was calculated for all the surface residues with the following general expression:

$$Q_{protein} = \sum Q^- + \sum Q^+$$

Where Q^- includes C-terminus and negatively charged R groups, and Q^+ includes N-terminus and positively charged R groups.

Reference:

(1) Miclotte, G.; Martens, K.; Fostier, J. Computational assessment of the feasibility of protonation-based protein sequencing. *PLoS One* **2020**, *15* (9), e0238625. DOI: 10.1371/journal.pone.0238625.
(2) Moore, D. S. Amino acid and peptide net charges: A simple calculational procedure. *Biochemical Education* **1985**, *13* (1), 10-11. DOI: https://doi.org/10.1016/0307-4412(85)90114-1.