










<https://doi.org/10.1038/s42003-021-01878-9>

OPEN

Exploration of natural red-shifted rhodopsins using a machine learning-based Bayesian experimental design

Keiichi Inoue ^{1,2,3,4,5,11}, Masayuki Karasuyama^{5,6,11}, Ryoko Nakamura³, Masae Konno³, Daichi Yamada³, Kentaro Mannen¹, Takashi Nagata^{1,5}, Yu Inatsu⁶, Hiromu Yawo ¹, Kei Yura ^{7,8,9}, Oded Béjà¹⁰, Hideki Kandori ^{2,3,4} & Ichiro Takeuchi ^{2,4,6}

Microbial rhodopsins are photoreceptive membrane proteins, which are used as molecular tools in optogenetics. Here, a machine learning (ML)-based experimental design method is introduced for screening rhodopsins that are likely to be red-shifted from representative rhodopsins in the same subfamily. Among 3,022 ion-pumping rhodopsins that were suggested by a protein BLAST search in several protein databases, the ML-based method selected 65 candidate rhodopsins. The wavelengths of 39 of them were able to be experimentally determined by expressing proteins with the *Escherichia coli* system, and 32 (82%, $p = 7.025 \times 10^{-5}$) actually showed red-shift gains. In addition, four showed red-shift gains >20 nm, and two were found to have desirable ion-transporting properties, indicating that they would be potentially useful in optogenetics. These findings suggest that data-driven ML-based approaches play effective roles in the experimental design of rhodopsin and other photobiological studies. (141/150 words).

¹The Institute for Solid State Physics, The University of Tokyo, Kashiwa, Japan. ²RIKEN Center for Advanced Intelligence Project, Tokyo, Japan.

³Department of Life Science and Applied Chemistry, Nagoya Institute of Technology, Nagoya, Japan. ⁴OptoBioTechnology Research Center, Nagoya Institute of Technology, Nagoya, Japan. ⁵PRESTO, Japan Science and Technology Agency, Kawaguchi, Japan. ⁶Department of Computer Science, Nagoya Institute of Technology, Nagoya, Japan. ⁷Graduate School of Humanities and Sciences, Ochanomizu University, Tokyo, Japan. ⁸Center for Interdisciplinary AI and Data Science, Ochanomizu University, Tokyo, Japan. ⁹School of Advanced Science and Engineering, Waseda University, Tokyo, Japan. ¹⁰Faculty of Biology, Technion-Israel Institute of Technology, Haifa, Israel. ¹¹These authors contributed equally: Keiichi Inoue, Masayuki Karasuyama.

email: inoue@issp.u-tokyo.ac.jp; takeuchi.ichiro@nitech.ac.jp

Microbial rhodopsins are photoreceptive membrane proteins widely distributed in bacteria, archaea, unicellular eukaryotes, and giant viruses^{1,2}. They consist of seven transmembrane (TM) α helices, with a retinal chromophore bound to a conserved lysine residue in the seventh helix (Fig. 1a). The first microbial rhodopsin, bacteriorhodopsin (BR), was discovered in the plasma membrane of the halophilic archaea *Halobacterium salinarum* (formerly called *H. halobium*)³. BR forms a purple-colored patch in the plasma membrane called purple membrane, which outwardly transports H⁺ using sunlight energy⁴. After the discovery of BR, various types of microbial rhodopsins were reported from diverse microorganisms, and recent progress in genome sequencing techniques has uncovered several thousand microbial rhodopsin genes^{1,5–7}. These microbial rhodopsins show various types of biological functions upon light absorption, leading to all-*trans*-to-13-*cis* retinal isomerization. Among them, ion transporters, including light-driven ion pumps and light-gated ion channels, are the most ubiquitous (Fig. 1b). Ion-transporting rhodopsins can transport several types of cations and anions, including H⁺, Na⁺, K⁺, halides (Cl⁻, Br⁻, I⁻), NO₃⁻, and SO₄²⁻,^{8–10}. The molecular mechanisms of ion-transporting rhodopsins have been detailed in numerous biophysical, structural, and theoretical studies^{1,2}.

In recent years, many ion-transporting rhodopsins have been used as molecular tools in optogenetics to control the activity of animal neurons optically *in vivo* by heterologous expression¹¹, and optogenetics has revealed various new insights regarding the neural network relevant to memory, movement, and emotional behavior^{12–15}. However, strong light scattering by biological tissues and the cellular toxicity of shorter wavelength light make precise optical control difficult. To circumvent this difficulty, new molecular optogenetics tools based on red-shifted rhodopsins, which can be controlled by weak scattering and low toxicity longer-wavelength light are urgently needed. Therefore, many approaches to obtain red-shifted rhodopsins have been reported, including gene screening, amino acid mutation based on biophysical and structural insights, and the introduction of retinal analogs^{16–18}. The insights obtained in these experimental studies, and further theoretical and computational studies^{19–22} revealed basic physical principle regulating absorption maximum wavelengths (λ_{\max}) of rhodopsins (also called spectral or color-tuning rule) in which the distortion of retinal polyene chain induced by steric interactions with surrounding residues, electrostatic interaction between protonated retinal Schiff base and counterion(s), and polarizability of the retinal binding pocket play essential

role²³. The λ_{\max} of several rhodopsins could be red-shifted by 20–40 nm without impairing the ion-transport function based on these physicochemical insights^{17,24,25}. These are successful examples of knowledge-driven experimental approach. Recently, a new method using a chimeric rhodopsin vector and functional assay was reported to screen the λ_{\max} and proton transport activities of several microbial rhodopsins that are present in specific environments²⁶. This method identified partial sequences of red-shifted yellow (560–570 nm)-absorbing proteorhodopsin (PR), the most abundant outward H⁺-pumping bacterial rhodopsin subfamily, from the marine environment. These works identified several red-shifted rhodopsins^{15,16,18,27}. Especially, most successful optogenetic tools are red-shifted channel rhodopsins such as Chrimson^{27,28} and RubyACR²⁹ which can induce and inhibit neural firing by absorbing 590 and 610-nm light, respectively. The rational amino acid mutation based on the structural insight further red-shifted the λ_{\max} of Chrimson to 608 nm²⁷. The development of next-generation sequencing technology is expected to continue to more rapidly identify a large number of new rhodopsin genes, including proteins with even longer wavelength-shifted absorption. However, screening of all of them either by experimental or theoretical methods would be very costly. Therefore, a less expensive and more efficient approach to screen red-shifted rhodopsins is needed, and data-driven study is expected as the third class of approach to investigate the color-tuning rule of rhodopsins at low cost.

To estimate the λ_{\max} of rhodopsins, we recently introduced a data-driven approach³⁰. In this previous study, we investigated the statistical relationship between the amino acid types at each position of the seven TM helices and the absorption wavelength of rhodopsins. We constructed a database containing 796 wild-type (WT) rhodopsins and their variants, the λ_{\max} of which had been reported in earlier studies. Then, we evaluated the strength of the relationship with a data-splitting approach, i.e., the data set was divided into a training set and a test set; the former was used to construct the predictive model, and the latter was used to estimate the predictive ability. The results of this “proof-of-concept” study suggested that the λ_{\max} of an unknown family of rhodopsins could be predicted with an average error of ± 7.8 nm, which is comparable to the mean absolute error of λ_{\max} estimated by the hybrid quantum mechanics/molecular mechanics (QM/MM)²¹ method. Considering the computational cost of both approaches, the data-driven approach was found to be much more efficient than the QM/MM approach, while the latter provides insights on the physical origin controlling λ_{\max} .

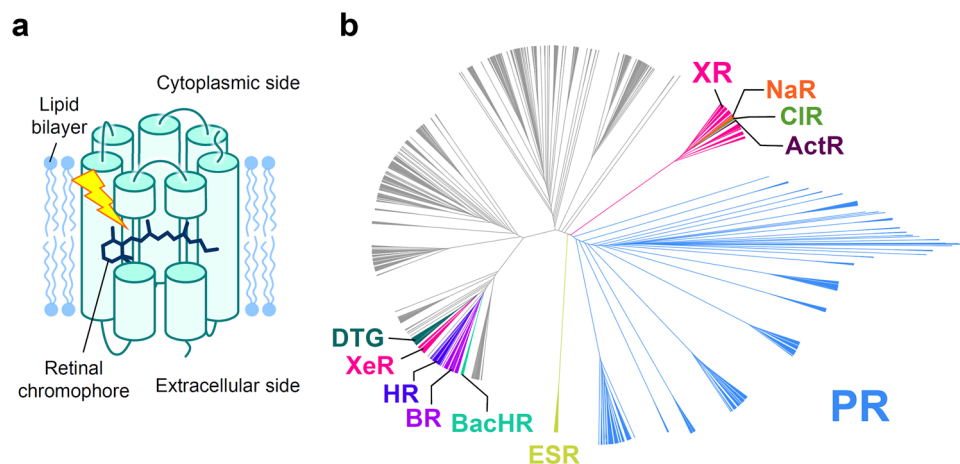


Fig. 1 Structure and phylogenetic tree of microbial rhodopsins. **a** Schematic structure of microbial rhodopsins. **b** Phylogenetic tree of microbial rhodopsins. The subfamilies of light-driven ion-pump rhodopsins targeted in this study are differently colored; non-ion-pump microbial rhodopsins and ion-pumping microbial rhodopsins from eukaryotic and giant viral origins are shown in gray.

Encouraged by this result, in this study, we introduced a machine-learning (ML)-based experimental design method which enables us screening more efficiently the candidates of rhodopsins that are likely to have red-shift gains with data-driven assist compared to the random or knowledge-driven screening. For this aim, we constructed a new dataset of 3022 wild-type putative ion-pump rhodopsins which were collected from public gene databases (NCBI non-redundant protein sequences, and metagenomic proteins³¹ and the *Tara* Oceans microbiome and virome database³²) and for which λ_{\max} have not been experimentally investigated yet to explore new red-shifted rhodopsins. The goal of the present study was to identify rhodopsins with λ_{\max} longer than the wavelengths of the representative rhodopsins in each subfamily of microbial rhodopsins for which the λ_{\max} has already been reported (base wavelengths). Here, we call the degrees of red-shift of the wavelength from the base wavelength the “red-shift gain”. We focus on rhodopsins with large red-shift gains because this would lead to the identification of amino acid types and residue positions that play important roles in red-shifting absorption wavelengths. Also, it is practically important in optogenetics applications to have a wide variety of ion-pumping rhodopsins from each subfamily to construct a new basis for rhodopsin toolboxes with red-shifted absorption and various types of ion species that can be transported. We constructed the ML-based experimental design method so that it could properly predict the expected red-shift gains, and applied this new method to 3022 putative ion-pumping rhodopsins derived from archaeal and bacterial origins that can be easily expressed in *Escherichia coli* (Fig. 1b).

We conducted experiments by introducing the synthesized rhodopsin genes into *E. coli* to measure the absorption wavelengths of 65 candidates for which the ML-based experimental design method predicted that the expected gains were >10 nm. Of these 65 selected candidates, 39 showed substantial coloring in *E. coli* cells, 32 showed actual red-shift gains, 6 showed blue-shifts, and 1 showed no change, i.e., 82% ($=32/39$, 7.025×10^{-5}) of the selected candidates showed actual red-shift gains. We then investigated the ion-transportation properties of the rhodopsins for which the red-shift gains were >20 nm, and found that some actually had desirable ion-transporting properties, suggesting that they (and their variants) could potentially be used as new optogenetics tools. Furthermore, the differences in the amino acid sequences of the newly examined rhodopsins and the representative ones in the same subfamily could be used for further investigation of the red-shifting mechanisms. This result suggests that it should be possible to find rhodopsins that have desired properties without conducting exhaustive biological experiments, and suggests that data-driven ML-based approaches should play effective roles in the experimental design of rhodopsin and other photobiological studies.

Results

Construction of an ML-based experimental design method for predicting expected red-shift gain. To screen rhodopsins that would have large red-shift gains, it is necessary to consider the uncertainty of prediction in the form of “predictive distributions”³³. By using predictive distributions, it is possible to consider appropriately the “exploration–exploitation trade-off” in screening processes^{34,35}, where exploration indicates an approach that prefers candidates with larger predictive variances, and exploitation indicates an approach that prefers candidates with longer predictive mean wavelengths (Fig. 2). Here, the term “exploration–exploitation” is a technical term used in the fields of active learning and experimental design, and “explorations” in the title of this paper is used in a broader sense and is not directly related to the former technical terminology. We employed a Bayesian modeling framework to compute the predictive

distributions of candidate rhodopsin red-shift gains. We then consider an exploration–exploitation trade-off by selecting candidate rhodopsins based on a criterion called “expected red-shift gains”.

To compute the expected red-shift gains of a wide variety of rhodopsins, we developed ML-based experimental design method based on the statistical analysis in our previous study³⁰. Figure 3 shows a schematic illustration of the ML-based experimental design method. First, we added 88 WT microbial rhodopsins and their variants for which the λ_{\max} had recently been reported in the literature or determined by our experiments, to a previously reported data set³⁰. In other words, the new training data set consisted of the amino acid sequences and λ_{\max} of 884 WT microbial rhodopsins and their variants (Supplementary Data 1). Second, the new ML model used only $N=24$ residues located around the retinal chromophore (Supplementary Fig. 1) because our previous study³⁰ indicated that amino acid residues at these 24 positions play significant roles in predicting absorption wavelengths (Fig. 3a). Third, $M=18$ amino acid physicochemical features (Supplementary Data 2) were used as inputs in the ML model, as opposed to the amino acid types used in the previous statistical analysis. This enabled us to predict the absorption wavelengths of a wide range of target rhodopsins that contain unexplored amino acid types in the training data at certain positions. Therefore, an amino acid sequence is transformed into an $M \times N=432$ dimensional feature vector $\mathbf{x} \in \mathbb{R}^{MN}$ by concatenating $x_{i,j}$, the j -th feature of the i -th residue (Fig. 3b). We consider a linear prediction model $f(\mathbf{x}) = \mu + \sum_{i=1}^N \sum_{j=1}^M \beta_{i,j} x_{i,j}$, where $\beta_{i,j}$ is the parameter for the j -th feature of the i -th residue, and μ is the intercept term.

Finally, to consider the exploration–exploitation trade-off appropriately in the screening process, we introduce a Bayesian modeling framework, which allows us to compute the predictive distributions of red-shift gains. Specifically, we employed Bayesian sparse modeling called BLASSO³⁶ (see the Methods section for details). This enables us to provide not only the mean, but also the variance of the predicted wavelengths. Unlike classical regression analysis, BLASSO regards the model parameters $\beta_{i,j}$ and μ as random variables generated from underlying distributions, as illustrated in Fig. 3c. Therefore, the wavelength prediction $f(\mathbf{x})$ is also represented as a distribution. The red-shift gain is defined as $\text{gain} = \max(f(\mathbf{x}) - \lambda_{\text{base}}, 0)$, where λ_{base} is the wavelength of the representative rhodopsin in the same subfamily whose λ_{\max} has been experimentally determined and reported in the literature (Supplementary Data 3). Note that the red-shift gain is positive if $f(\mathbf{x})$ is greater than λ_{base} ; otherwise, it takes the value of zero. Since $f(\mathbf{x})$ is regarded as a random variable in BLASSO, the red-shift gain is also regarded as a random variable. Therefore, we employ the expected value of the red-shift gain, denoted by $\mathbb{E}[\text{gain}]$, as the screening criterion where \mathbb{E} represents the expectation of a random variable. Illustrative examples of $\mathbb{E}[\text{gain}]$ are shown in Fig. 3d. Unlike the simple expectation of the wavelength prediction $\mathbb{E}[f(\mathbf{x})]$, $\mathbb{E}[\text{gain}]$ depends on the variance of the predictive distribution (For example, $\mathbb{E}[\text{gain}]$ of target #4 is larger than #1 in Fig. 2f though $\mathbb{E}[f(\mathbf{x})] - \lambda_{\text{base}}$ of #4 is smaller than #1 in Fig. 2e). This encourages the exploration of rhodopsin candidates having large uncertainty (for exploration), as opposed to only those having longer wavelengths with high confidence (for exploitation).

Screening potential red-shifted microbial rhodopsins based on expected red-shift gains. The target data set to explore red-shifted microbial rhodopsins was constructed with putative microbial rhodopsin genes collected by a protein BLAST (blastp) search³⁷ of the NCBI non-redundant protein and metagenome

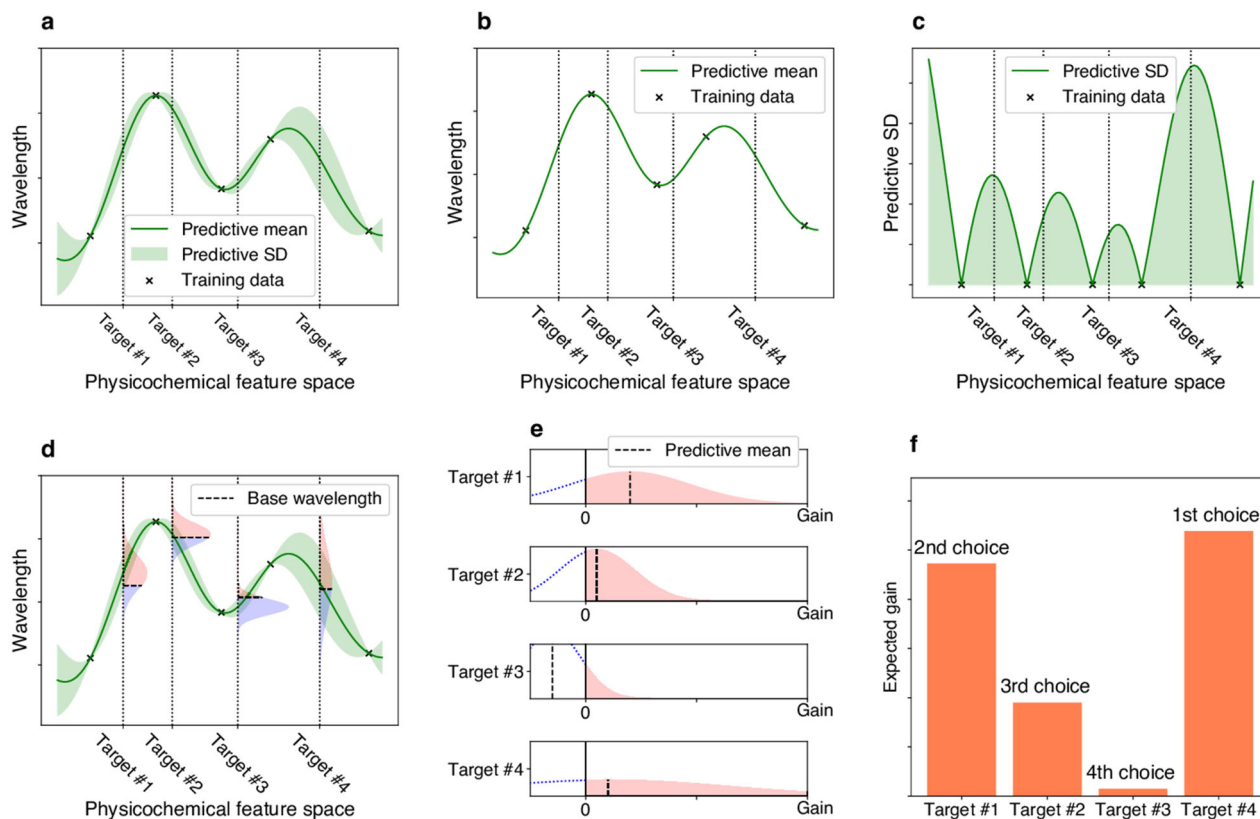


Fig. 2 Illustrations of exploration-exploitation for screening rhodopsins with red-shift gain. **a** Bayesian prediction model constructed using the current training data (black crosses). The prediction model is represented by the predictive mean and predictive standard deviation (SD). The horizontal axis schematically illustrates the space of proteins defined through physicochemical features. The four vertical dotted lines indicate target proteins (candidates to synthesize). **b** Predictive mean. This function is defined as the expected value of the probabilistic prediction by the Bayesian model. **c** Predictive SD. Since the predictive SD represents the uncertainty of the prediction, it has a larger value when the training data points do not exist nearby. **d** The distributions on the vertical dotted lines represent the predictive distributions, and the horizontal dashed lines are the base wavelengths of the target points. The base wavelength is different for each target point because it depends on the subfamily of the protein. **e** The density of the predictive distribution of each target protein on its red-shift gain value. The gain is defined as the predicted wavelength subtracted by the base wavelength, and if it is negative, the value is truncated as 0. This can be seen as a “benefit” that can be obtained by observing the target protein. **f** Expected value of the red-shift gain. This provides a ranking list from which the next candidates to be experimentally investigated can be determined. Target #4 has the largest expected gain, although target #1 has the largest increase in the predictive mean compared with base wavelength in **e**. Because of its larger SD (as shown in **a**, **c**, **d**, and **e**), target #4 is probabilistically expected to have a larger gain than the other targets.

databases³¹, as well as the *Tara* Oceans microbiome and virome databases³². As a result, we obtained a non-redundant data set of 5558 microbial rhodopsin genes (Fig. 1b). The sequences were aligned by ClustalW and categorized to subfamilies of microbial rhodopsins based on the phylogenetic distances, as reported previously³⁸. Among these, 3022 rhodopsin genes, which did not have identical sequences in the training data and from bacterial and archaeal origins, were extracted because their λ_{\max} can be easily measured by expressing in *E. coli* cells. We calculated the $\mathbb{E}[\text{gain}]$ of these 3022 genes (Supplementary Data 4), and then selected 65 genes of putative light-driven ion-pump rhodopsins showing an $\mathbb{E}[\text{gain}] > 10$ nm for further experimental evaluation, as ion-pump rhodopsins can be used as new optogenetics tools.

Experimental measurement of the absorption wavelengths of microbial rhodopsins showing high red-shift gains. We synthesized the selected 65 genes that showed an $\mathbb{E}[\text{gain}] > 10$ nm. These were then introduced into *E. coli* cells, and the proteins expressed in the presence of 10 μM all-*trans* retinal. As a result, 39 *E. coli* cells showed substantial coloring, indicating high expression of folded protein, and their λ_{\max} were determined by observing ultraviolet (UV)-visible absorption changes upon

bleaching of the expressed rhodopsins through a hydrolysis reaction of their retinal with hydroxylamine, as previously reported³⁰ (Fig. 4). The observed gains were compared with the $\mathbb{E}[\text{gain}]$ shown in Table 1. A full list of unexpressed genes is shown in Supplementary Data 5. In total, 32 out of 39 genes showed a longer wavelength than their base wavelength (that is, positive red-shift gain; Fig. 5), suggesting that our ML-based model can significantly improve the efficiency of screening to explore new red-shifted microbial rhodopsins compared with random sampling ($p = 7.025 \times 10^{-5}$ by a binomial test assuming that the probability of red-shift gain for random choice is 50%).

Ion-transport function of red-shifted microbial rhodopsins.

Overall, 4 of the 39 rhodopsins showed red-shifted absorption ≥ 20 nm compared with the base wavelengths (Table 1): three were halorhodopsins (HRs) from bacterial species^{10,39,40} (to distinguish classical HRs from archaeal species, these are hereafter referred to as bacterial-halorhodopsins [BacHRs]), and one was a PR⁴¹. Their ion-transport activities were then investigated by expressing in *E. coli* cells and observing the pH change in external solvent whose pH was initially set to 7 (Fig. 6a). Upon light illumination, BacHRs from *Rubrivirga marina* and *Myxosarcina*

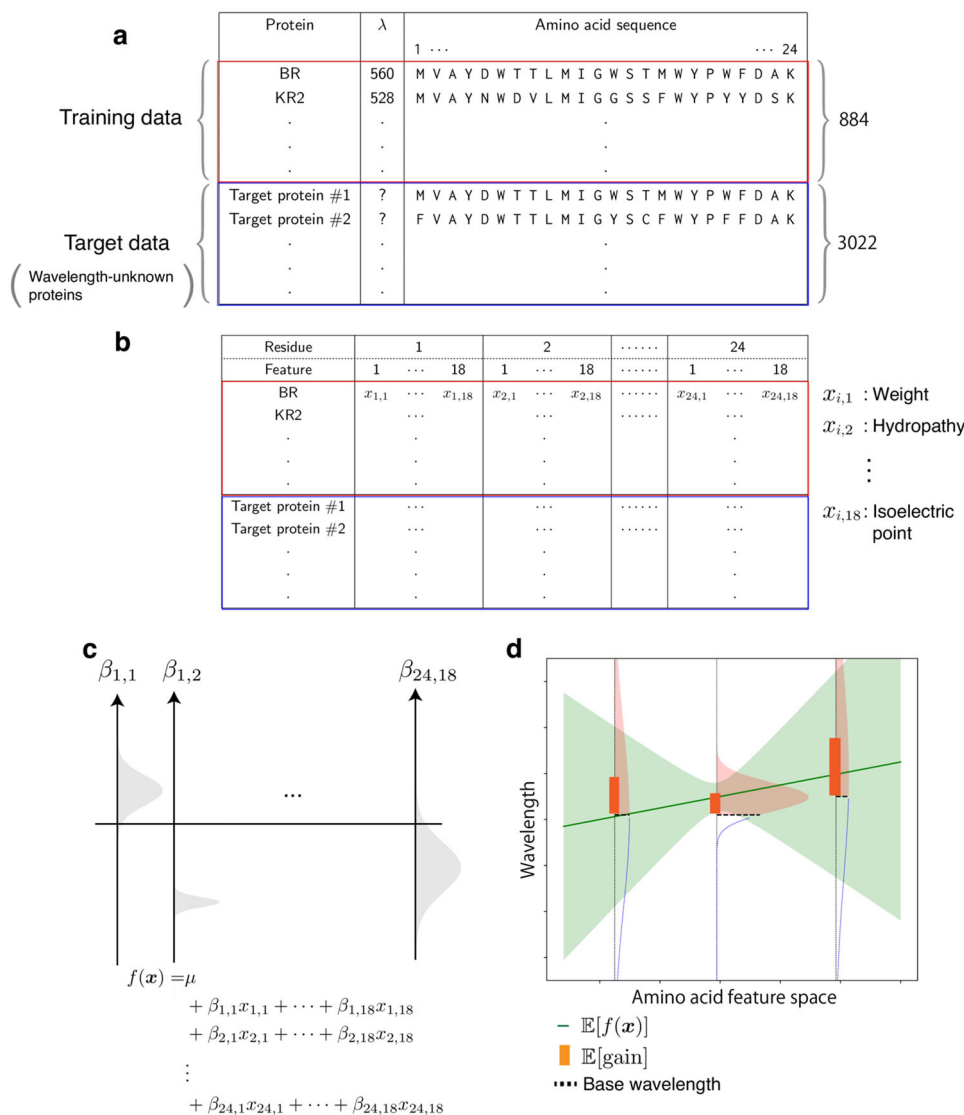


Fig. 3 Overview of the ML-based exploration of natural red-shifted rhodopsins. a Using existing experimental data, a training data set consisting of pairs of a wavelength λ_{max} and an amino acid sequence was constructed. A particular focus was placed on the 24 amino acid residues around the retinal chromophore to build an ML-based prediction model. A set of protein sequences with no known wavelength was also collected as target proteins. **b** All amino acid sequences were transformed into physicochemical features, leading to $24 \times 18 = 432$ dimensional numerical representations of each protein. **c** A linear regression model was constructed using the Bayesian approach. Each regression coefficient $\beta_{i,j}$ was estimated as a distribution (shown as a gray region). The broadness of these distributions represent the uncertainty of the current estimation. **d** The expected red-shift gain values were evaluated for the target proteins. The green region is the standard deviation of the prediction. The red shaded region in the vertical distribution corresponds to the probability that the wavelength is larger than the base wavelength (dashed line), which is determined by the subfamily of the microbial rhodopsin. The bar represents the expected red-shift gain, defined by the expected value of the increase from the base wavelength.

sp. GII showed alkalization of external solvent, which was enhanced by addition of the protonophore (CCCP), which increases the H^+ permeability of the cell membrane, and the light-dependent alkalizations disappeared when anions were exchanged from Cl^- to NO_3^- , indicating that these were light-driven Cl^- pumps, similar to other rhodopsins in the same BacHR subfamily^{10,39}. By contrast, *Cyanothece* sp. PCC 7425 did not show any substantial transport. While no transporting function can be attributed to the heterologous expression in *E. coli*, it would have considerably different molecular properties from other BacHRs. PR from a metagenome sequence (ECV93033.1) showed acidification of external solvent that was abolished by the addition of CCCP and was independent from ionic species in the solvent. Hence, this was a new red-shifted outward H^+ pump compared with typical PRs whose λ_{max} are present at ca. 520

nm⁴¹. Furthermore, these rhodopsins are needed to be functional in mammalian cells for their optogenetic applications. To verify this issue, we carried out electrophysiological experiment to measure the photocurrent of BacHRs from *Rubrivirga marina* and PR from a metagenome sequence (ECV93033.1) in mammalian cells (ND7/23; Fig. 6b). Both of them showed substantial photocurrent even in the mammalian cells. These light-driven ion-pumping rhodopsins with red-shifted λ_{max} have the potential to be applied as new optogenetics tools, and thus, warrant further study in the near future.

Discussion

Microbial rhodopsins show a wide variety of λ_{max} by changing steric and electrostatic interactions between all-*trans* retinal

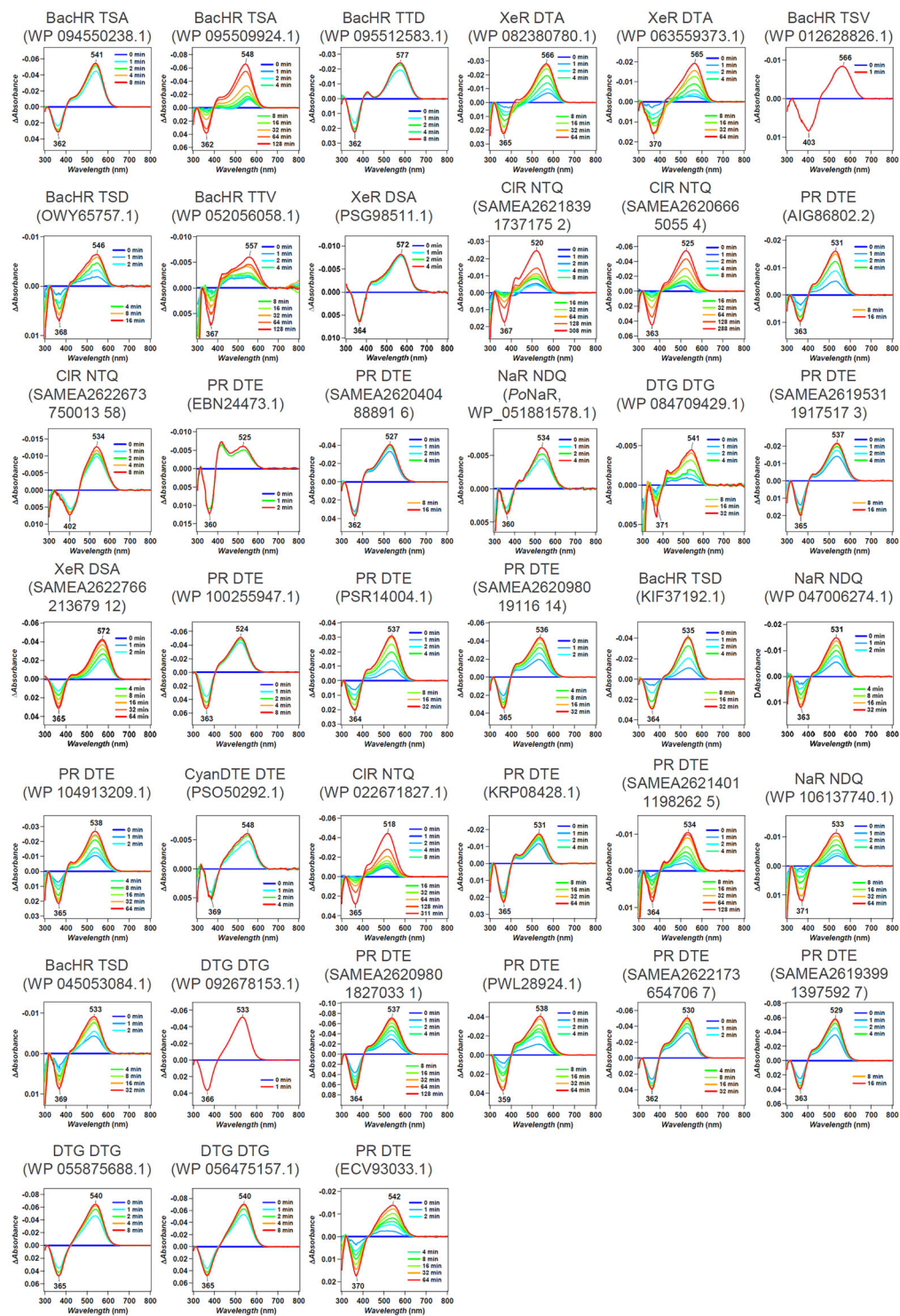


Fig. 4 λ_{\max} of 39 microbial rhodopsins in solubilized *E. coli* membrane observed upon hydroxylamine bleach reaction. The difference absorption spectra between before and after hydroxylamine bleaching reaction of microbial rhodopsins in solubilized *E. coli* membrane. The λ_{\max} of each rhodopsin was determined by the peak positions of the absorption spectra of the original proteins, and the absorption of retinal oxime produced by the reaction of retinal Schiff base and hydroxylamine was observed as a negative peak at ~360–370 nm.

chromophores and surrounding amino acid residues. An understanding of the color-tuning rule enables more efficient screening and the design of new red-shifted rhodopsins that have value as optogenetics tools, and our ML-based data-driven approach therefore provides a new basis to identify color-regulating factors without assumptions.

We previously demonstrated that an ML-based model based on ~800 experimental results could predict the λ_{\max} of microbial rhodopsins with an average error of ± 7.8 nm. Encouraged by this result, in the present study, we constructed a new ML-based model to compute expected red-shift gains for a wide range of unknown families of microbial rhodopsins. As a result,

Table 1 Predicted and observed gains of 39 microbial rhodopsins expressed in *E. coli*.

Origin	Accession	Subfamily	Motif	Base wavelength/nm	\mathbb{E} [gain]	Observed wavelength/nm	(Observed wavelength)-(base wavelength)/nm
<i>Rubricoccus marinus</i>	WP 094550238.1	BacHR	TSA	537	40.7	541	4
<i>Rubrivirga marina</i>	WP 095509924.1	BacHR	TSA	537	39.8	548	11
<i>Rubrivirga marina</i>	WP 095512583.1	BacHR	TTD	537	35.5	577	40
<i>Bacillus</i> sp. CHD6a	WP 082380780.1	XeR	DTA	565	35.3	566	1
<i>Bacillus horikoshii</i>	WP 063559373.1	XeR	DTA	565	35.3	565	0
<i>Cyanotheca</i> sp. PCC 7425	WP 012628826.1	BacHR	TSV	537	32.9	566	29
<i>Cyanobacterium</i> TDX16	OWY65757.1	BacHR	TSD	537	32.9	546	9
<i>Myxosarcina</i> sp. GII	WP 052056058.1	BacHR	TTV	537	31.2	557	20
<i>Nanohaloarchaea</i> archaeon SW 7 43 1	PSG98511.1	XeR	DSA	565	29.2	572	7
Metagenome sequence	SAMEA2621839 1737175 2	CIR	NTQ	530	25.7	520	-10
Metagenome sequence	SAMEA2620666 5055 4	CIR	NTQ	530	25.1	525	-5
<i>Nonlabens</i> sp. YIK11	AIG86802.2	PR	DTE	520	21.5	531	11
Metagenome sequence	SAMEA2622673 750013 58	CIR	NTQ	530	21.4	534	4
Metagenome sequence	EBN24473.1	PR	DTE	520	20.0	525	5
Metagenome sequence	SAMEA2620404 88891 6	PR	DTE	520	20.0	527	7
<i>Parvularcula oceani</i>	WP_051881578.1	NaR	NDQ	525	19.7	534	9
<i>Rubrobacter</i> aplysinae	WP 084709429.1	DTG	DTG	535	19.5	541	6
Metagenome sequence	SAMEA2619531 1917517 3	PR	DTE	520	18.0	537	17
Metagenome sequence	SAMEA2622766 213679 12	XeR	DSA	565	17.8	572	7
<i>Reinekea forsetii</i>	WP 100255947.1	PR	DTE	520	17.1	524	4
<i>Bacteroidetes</i> bacterium	PSR14004.1	PR	DTE	520	15.4	537	17
Metagenome sequence	SAMEA2620980 19116 14	PR	DTE	520	15.4	536	16
<i>Hassallia byssoidea</i> VB512170	KIF37192.1	BacHR	TSD	537	15.1	535	-2
<i>Erythrobacter gangjinensis</i>	WP 047006274.1	NaR	NDQ	525	13.7	531	6
<i>Pontimonas salivibrio</i>	WP 104913209.1	PR	DTE	520	12.2	538	18
<i>Cyanobacteria</i> bacterium QH 1 48 107	PSO50292.1	CyanDTE	DTD	545	12.0	548	3
<i>Sphingopyxis baekryungensis</i>	WP 022671827.1	CIR	NTQ	530	11.0	518	-12
<i>Sphingobacteriales</i> bacterium BACL12 MAG120802bin5	KRP08428.1	PR	DTE	520	10.9	531	11
Metagenome sequence	SAMEA2621401 1198262 5	PR	DTE	520	10.9	534	14
<i>Spirosoma oryzae</i>	WP 106137740.1	NaR	NDQ	525	10.8	533	8
<i>Aliterella atlantica</i>	WP 045053084.1	BacHR	TSD	537	10.8	533	-4
<i>Rosenbergiella nectarea</i>	WP 092678153.1	DTG	DTG	535	10.8	533	-2
Metagenome sequence	SAMEA2620980 1827033 1	PR	DTE	520	10.4	537	17
<i>Fluviicola</i> sp. XM24bin1	PWL28924.1	PR	DTE	520	10.4	538	18
Metagenome sequence	SAMEA2622173 654706 7	PR	DTE	520	10.4	530	10
Metagenome sequence	SAMEA2619399 1397592 7	PR	DTE	520	10.4	529	9
<i>Sphingomonas</i> sp. Leaf34	WP 055875688.1	DTG	DTG	535	10.3	540	5
<i>Sphingomonas</i> sp. Leaf38	WP 056475157.1	DTG	DTG	535	10.3	540	5
Metagenome sequence	ECV93033.1	PR	DTE	520	10.3	542	22

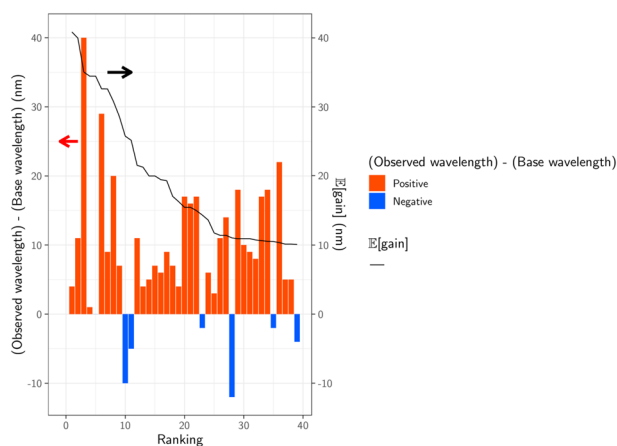


Fig. 5 Observed wavelengths and expected red-shift gains. The predicted and observed red-shift (and blue-shift) gains for the 39 candidate rhodopsins that showed substantial coloring in *E. coli* cells. Differences between observed and base wavelengths are shown by the bars. The red bars indicate red-shift from the base wavelength, while the blue bars indicate observed wavelengths that were shorter than the base wavelengths. Proteins are sorted in the descending order by \mathbb{E} [gain], as shown by the black line. Among the 39 candidates, 32 (82%) showed red-shift gains, suggesting that the proposed ML-based model can screen red-shifted rhodopsins more efficiently than random choice.

32 out of 39 microbial rhodopsins were found to have red-shifted absorption compared with the base wavelengths of each subfamily of microbial rhodopsins (Table 1), suggesting that our data-driven ML approach can screen red-shifted

microbial rhodopsin genes more efficiently than random choice ($p = 7.025 \times 10^{-5}$).

By considering the exploration–exploitation trade-off, that is, to consider not only the expected value of the prediction, but also the uncertainty, it was possible to construct a red-shift protein screening process, as shown in Fig. 7. Figure 7a shows the relationships between the prediction uncertainty (as measured by the standard deviation) and the observed red-shift gains. It can be seen that rhodopsins with red-shift gain are found in areas of not only low (small standard deviation), but also high prediction uncertainty (large standard deviation). Figure 7b shows the two-dimensional projection of the $d = 432$ dimensional feature space by principal component analysis. It can be seen that red-shift gains (red) are found for target proteins not only close to training proteins (green), but also far from training proteins. Figure 8 shows that the observed wavelengths and red-shift gains tend to be smaller than the predicted ones. We conjecture that these differences between the observed and predicted wavelengths and red-shift gains are due to modeling errors, possibly caused by a lack of sufficient information (e.g., three-dimensional structures) and modeling flexibility (e.g., nonlinear effects); in other words, rhodopsins having high prediction values partly by modeling errors have a high chance of being selected. Therefore, it would be valuable to develop a statistical methodology to eliminate selection bias due to modeling errors.

Four rhodopsins showed red-shifted absorption ≥ 20 nm than the base wavelength, three of which showed light-driven ion-transport function. Interestingly, while one BacHR from *Rubrivirga marina* (accession No.: WP 095512583.1) showed a 40-nm longer λ_{max} (577 nm) than the base wavelength, another 11-nm red-shifted BacHR (WP 095509924.1) was also identified from the

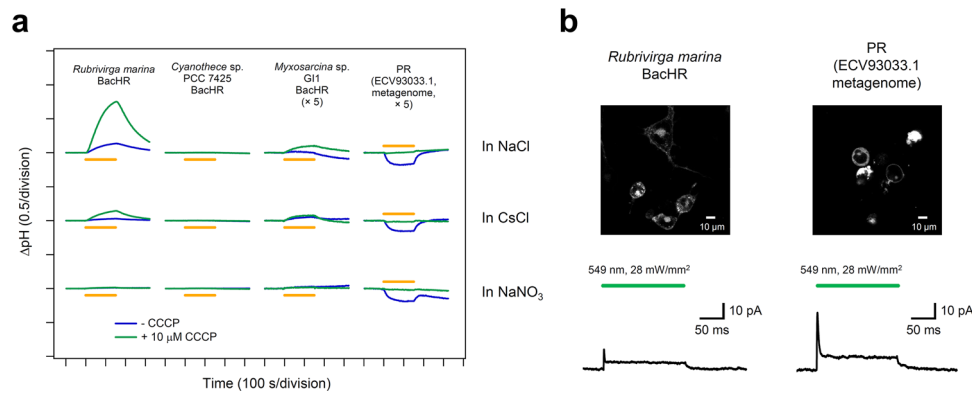


Fig. 6 Light-driven ion-transport activities of microbial rhodopsins showed longer λ_{\max} . **a** The light-induced pH change in the external solvent of *E. coli* cells expressing four microbial rhodopsins that showed a $\lambda_{\max} \geq 20$ nm longer than the base wavelength of the subfamily. The data obtained without and with 10 μ M CCCP are indicated by the blue and green lines, respectively, in 100 mM NaCl, CsCl, and NaNO₃. Light was illuminated for 150 s (yellow solid lines). **b** *Rubrivirga marina* BacHR or PR (ECV93033.1 metagenome) were expressed in the membrane of ND7/23 cells (top image) and generated positive photocurrent in response to a green light pulse (200 ms, 549 nm, 28 mW/mm²). The traces in the bottom are typical records at a holding potential of 0 mV.

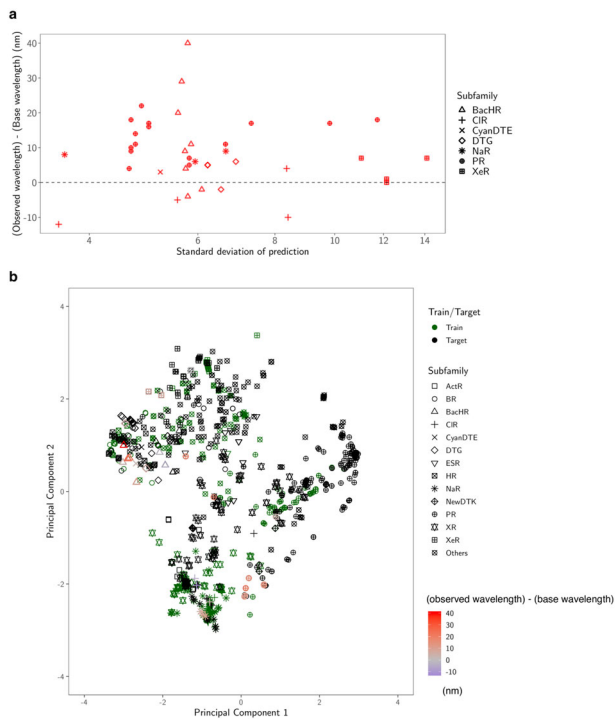


Fig. 7 Diversity of the selected proteins. **a** Predicted standard deviation (horizontal axis) vs. observed gain (vertical axis). The marker shape represents the subfamily of each protein. **b** Two-dimensional projection created by principal component analysis. The original $d = 432$ dimensional feature space is projected onto the first two principal component directions. The first component (horizontal axis) explains 33% of the total variance of the original space, and the second (vertical axis) explains 17%. The green markers are the training data, and the black markers are the target data. For the synthesized proteins, differences in the observed and base wavelengths are shown by the color map. The results indicate that, by considering the exploration-exploitation trade-off, it was possible to make a red-shift protein screening process that considered not only the expected value of the prediction, but also the uncertainty.

same bacteria (Table 1). These BacHRs are highly similar to each other (55.2% identity and 70.6% similarity), and only four of 24 amino acid residues around the retinal chromophore differ. Hence, *R. marina* evolved two BacHRs with 29-nm different λ_{\max} by

a small number of amino acid replacements; the amino acid residue(s) responsible for this color-tuning should be investigated in the future.

The differences in amino acids in three of 24 retinal-surrounding residues are known to play a color-tuning role in natural rhodopsins without affecting their biological function. These correspond to positions 93, 186, and 215 in BR (BR Leu93, Pro186, and Ala215, respectively)¹⁷. Position 93 is known to be diversified in the PR family (the well-known position 105 in PRs). Green-light-absorbing PRs (GPRs) have leucine as a BR, whereas glutamine is conserved in blue-light-absorbing PRs^{5,26}. This color-tuning effect by the difference between leucine and glutamine is known as the “L/Q-switch”⁴². Interestingly, while 29.8% of 3022 candidate genes have glutamine at this position, all 39 genes whose large red-shift gains were suggested by our ML-based model have amino acids other than glutamine, which suggests that our ML-based model avoided the genes having glutamine at position 93. Especially, 12 (37.5%) of 32 genes that actually showed red-shifted absorption compared with the base wavelengths had methionine at this position (Supplementary Data 6), which is substantially higher than the proportion of methionine-conserving genes in the 3022 candidates (16.1%). The red-shifting effect of the L-to-M mutation of this residue in GPRs previously reported⁴² and the current result imply that many rhodopsins have evolved methionine to absorb light with longer wavelengths. Position 215 in BR is also known to have a color-tuning role. The mutation from alanine to threonine or serine (A/TS switch) has a blue-shifting effect of 9–20 nm^{17,43–45}. Five of six genes that showed blue-shifted λ_{\max} compared with the base wavelengths have threonine or serine at this position, suggesting that these types of genes should be avoided to explore red-shifted rhodopsins. By contrast, asparagine was conserved in more than half (58.4%) of the 3022 candidate genes, especially in those belonging to the PR subfamily. A substantial portion (37.5%) of the genes with red-shifted absorption compared with the base wavelengths also had asparagine at this position (Supplementary Data 6). The A-to-N mutation at this position had a smaller effect (4–7 nm)^{30,44} than that of the A-to-S/T mutation; thus, the difference between alanine and asparagine is not so critical to explore red-shifted rhodopsins. Position 186 in BR is proline in most microbial rhodopsins (in 98.7% of the 3022 candidate genes), and the mutation to non-proline amino acids induces red-shift of absorption¹⁷. We identified sodium pump rhodopsin (NaR) from *Parvularcula oceani*, which also has a threonine at this position, and showed 10-nm longer absorption than the base

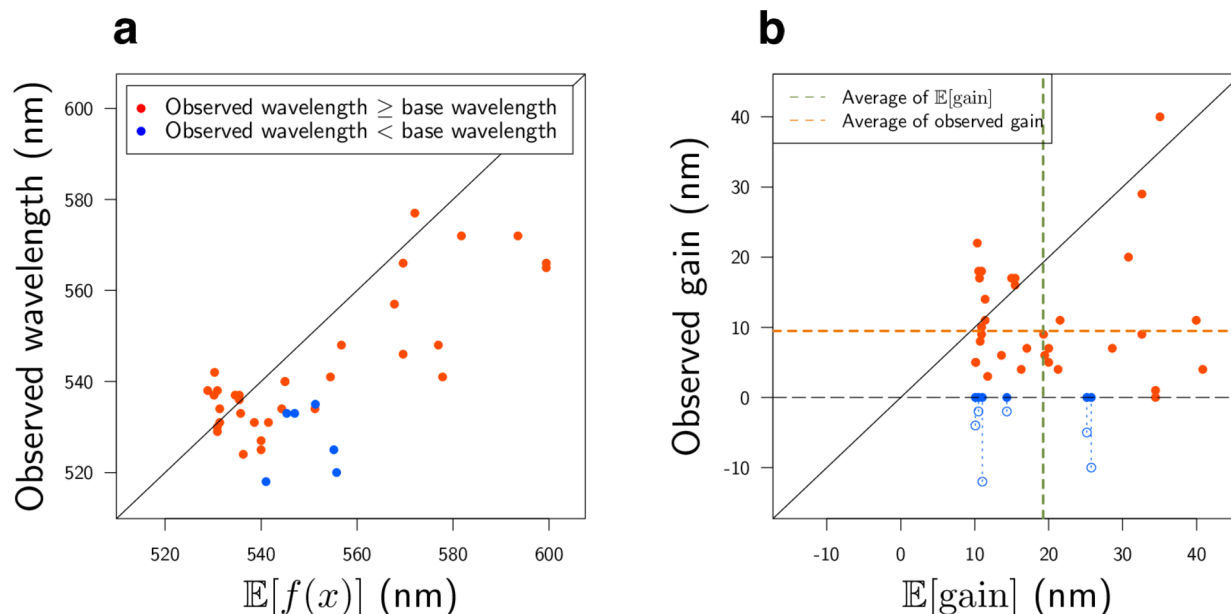


Fig. 8 Comparisons of experimental observations and ML predictions. In these two plots, the red points have longer observed wavelengths than the base wavelength λ_{base} , while the blue points have shorter observed wavelengths than λ_{base} . **a** ML-based prediction of λ_{max} (horizontal axis) vs. experimentally observed λ_{max} (vertical axis). **b** Expected red-shift gain (horizontal axis) vs. observed gain (vertical axis). Since we selected rhodopsins having expected red-shift gains of ≥ 10 nm, all the points on the horizontal axis are ≥ 10 nm. The observed gain, defined by $\max(\lambda_{\text{max}} - \lambda_{\text{base}}, 0)$, is nonnegative by definition. For blue points whose observed gain is equal to 0, the value of $\lambda_{\text{max}} - \lambda_{\text{base}}$ is also shown as blue outlined circles. The green and orange dashed lines are the averages of the horizontal and vertical axes (19.2 nm and 9.5 nm), respectively. The results indicate that the observed wavelengths and red-shift gains tended to be smaller than the predicted ones. We conjecture that these differences between the observed and predicted wavelengths are due to modeling errors (see the Discussion for details).

wavelength. Although genes having non-proline amino acids are rare in nature, it would be beneficial to identify new red-shifted rhodopsins. These results indicate that ML-based modeling can provide insights for identifying new functional tuning rules for proteins based on specific amino acid residues.

The number of reported microbial rhodopsin genes is rapidly increasing due to the development of next-generation sequencing techniques and microbe culturing methods. New microbial rhodopsins with molecular characteristics suitable for optogenetics applications are expected to be included in upcoming genomic data. Data-driven approaches would be able to efficiently suggest promising rhodopsins which should be investigated preferentially. Although the absorption of the most red-shifted rhodopsin found in this study (BacHR from *Rubrivirga marina*, $\lambda_{\text{max}} = 577$ nm) is shorter than the peak activation wavelength of eNpHR3.0 (590 nm) which is extensively used in optogenetic studies⁴⁶, our ML-based model could be expected to reduce the costs associated with identifying red-shifted rhodopsins from upcoming genomic data. Especially, we expect that our ML-based model could be applied to ion channel and enzymatic rhodopsins, which were not a focus of this study because of their eukaryotic origins; however, their use in optogenetics research could help identify more useful optogenetics tools with red-shifted absorption in the future.

Methods

Experimental design. The objective of this study was to introduce and demonstrate the effectiveness of a data-driven experimental design method to screen candidates for rhodopsin proteins with desired properties from more than several thousand candidates identified in various microbial species. To this end, we constructed a training dataset for developing a ML model and a target dataset for screening targets (Construction of training and target data sets). A machine learning model was constructed using the training dataset (ML modeling), which was used to select the 65 candidates from 3022 in the target dataset. The protein expressions of selected candidates were induced (Protein expression), and the

absorption spectra and λ_{max} of the selected rhodopsins were measured (Measurement of the absorption spectra and λ_{max} of rhodopsins by bleaching with hydroxylamine). Furthermore, we investigated the ion-transportation properties of the rhodopsins that showed large red-shift gains (Ion-transport assay of rhodopsins in *E. coli* cells). Statistical significance of the effectiveness of the data-driven experimental design method was assessed by a binomial test.

Construction of training and target data sets. In this study, we constructed a new training data set (Supplementary Data 1) by adding 88 genes for which the λ_{max} had recently been reported in the literature or determined by our experiments, to a previously reported data set³⁰. The sequences were aligned using ClustalW⁴⁷ and the results were manually checked to avoid improper gaps and/or shifts in the TM parts. The aligned sequences were then used for ML-based modeling.

To collect microbial rhodopsin genes for the training data set, BR⁴⁸ and heliorhodopsin 48C12⁴⁹ sequences were used as queries for searching homologous amino acid sequences in NCBI non-redundant protein sequences and metagenomic proteins³¹ and the *Tara* Oceans microbiome and virome database³². Protein BLAST (blastp)³⁷ was used for the homology search, with the threshold E-value set at < 10 by default, and sequences with > 180 amino acid residues were collected. All sequences were aligned using ClustalW⁴⁷. The highly diversified C-terminal 15-residue region behind the retinal binding Lys (BR Lys216) and long loop of HeR between helices A and B were removed from the sequences to avoid unnecessary gaps in the alignment. The successful alignment of the TM helical regions, especially the 3rd and 7th helices, was checked manually. The phylogenetic tree was drawn using the neighbor-joining method⁵⁰, and the microbial rhodopsin subfamilies were categorized based on the phylogenetic distances, as reported previously³⁸. Based on the phylogenetic tree, 3022 putative ion-pumping rhodopsin genes from bacterial and archaeal origins were extracted, and their aligned sequences were used as the training data set for the prediction of λ_{max} . The original training and test sets are provided in Supplementary Data 1 and Table 1, respectively, and the entire transformed datasets with physicochemical features (see Supplementary Data 2) are provided in Supplementary Data 7.

ML modeling. Suppose that we have K pairs of an amino acid sequence and an absorption wavelength $\{(\mathbf{x}^{(k)}, \lambda_{\text{max}}^{(k)})\}_{k=1}^K$, where $\mathbf{x}^{(k)} \in \mathbb{R}^{MN}$ is the feature vector of the k -th amino acid sequence and $\lambda_{\text{max}}^{(k)} \in \mathbb{R}$ is the absorption wavelength of the k -th rhodopsin protein. The least-absolute shrinkage selection operator (LASSO) is a standard regression model in which important regression coefficients can be

automatically selected by the penalty on the absolute value of the coefficient, as follows:

$$\min_{\mu, \beta} \sum_{k=1}^K \left(\lambda_{\max}^{(k)} - \mu - \sum_{i=1}^M \sum_{j=1}^N \beta_{ij} x_{ij}^{(k)} \right)^2 + \gamma \sum_{i=1}^M \sum_{j=1}^N |\beta_{ij}|,$$

where $\beta \in \mathbb{R}^{MN}$ is a vector of β_{ij} and $\gamma > 0$ is the regularization parameter. BLASSO is a Bayesian extension of LASSO for which the model is defined through the following random variables:

$$\lambda_{\max}^{(k)} \sim N(\mu + \beta^T x^{(k)}, \sigma^2), \beta \sim \pi(\beta | \sigma^2),$$

where $N(\mu, s^2)$ is a Gaussian distribution with mean μ and variance s^2 , and $\pi(\beta | \sigma^2) = \prod_{i=1}^M \prod_{j=1}^N \frac{\gamma}{2\sqrt{\sigma^2}} e^{-\gamma|\beta_{ij}|/\sqrt{\sigma^2}}$ is the conditional Laplace prior. In this model, the maximum of the conditional distribution of the parameter $\beta | \{x^{(k)}, \lambda_{\max}^{(k)}\}_{k=1}^K, \lambda, \sigma$ is equivalent to the LASSO⁵¹ estimator. For γ , a hyper-prior is set through the gamma distribution prior on γ^2 , and the inverse gamma prior is assumed for σ^2 . For the computational details, see the original paper³⁶. We used the “monomvn” package of R in our implementation. The prediction $f(x)$ was sampled through the Gibbs sampler of β and μ . The number of samplings was set as $T = 10,000$ times. For each candidate x , we approximately obtain $\mathbb{E}[\text{gain}]$ by

$$\mathbb{E}[\text{gain}] \approx \frac{1}{T} \sum_{t=1}^T \max(\mu^{(t)} + \beta^{(t)T} x - \lambda_{\text{base}}, 0),$$

where $\mu^{(t)}$ and $\beta^{(t)}$ are the t -th sampled parameters. The parameters of the trained model is provided in Supplementary Data 8.

Protein expression. The synthesized genes of microbial rhodopsins codon-optimized for *E. coli* (Genscript, NJ) were incorporated into the multi-cloning site in the pET21a(+) vector (Novagen, Merck KGaA, Germany). The plasmids carrying the microbial rhodopsin genes were transformed into the *E. coli* C43(DE3) strain (Lucigen, WI). Protein expression was induced by 1 mM isopropyl β -D-1-thiogalactopyranoside (IPTG) in the presence of 10 μ M all-*trans* retinal for 4 h.

Measurement of the absorption spectra and λ_{\max} of rhodopsins by bleaching with hydroxylamine. *E. coli* cells expressing rhodopsins were washed three times with a solution containing 100 mM NaCl and 50 mM Na_2HPO_4 (pH 7). The washed cells were treated with 1 mM lysozyme for 1 h and then disrupted by sonication for 5 min (VP-300N; TAITEC, Japan). To solubilize the rhodopsins, 3% *n*-dodecyl- β -maltoside (DDM, Anatrace, OH) was added, and the samples were stirred for overnight at 4 °C. The rhodopsins were bleached with 500 mM hydroxylamine and subjected to yellow light illumination ($\lambda > 500$ nm) from the output of a 1-kW tungsten-halogen projector lamp (Master HILUX-HR; Rikagaku) through colored glass (Y-52; AGC Techno Glass, Japan) and heat-absorbing filters (HAF-50S-15H; SIGMA KOKI, Japan). The absorption change upon bleaching was measured by a UV-visible spectrometer (V-730; JASCO, Japan).

Ion-transport assay of rhodopsins in *E. coli* cells. To assay the ion-transport activity in *E. coli* cells, the cells carrying expressed rhodopsin were washed three times and resuspended in unbuffered 100 mM NaCl. A cell suspension of 7.5 mL at $\text{OD}_{660} = 2$ was placed in the dark in a glass cell at 20 °C and illuminated at $\lambda > 500$ nm from the output of a 1-kW tungsten-halogen projector lamp (Rikagaku, Japan) through a long-pass filter (Y-52; AGC Techno Glass, Japan) and a heat-absorbing filter (HAF-50S-50H; SIGMA KOKI, Japan). The light-induced pH changes were measured using a pH electrode (9618S-10D; HORIBA, Japan). All measurements were repeated under the same conditions after the addition of 10 μ M CCCP.

Imaging and electrophysiological assays. For heterologous expression in mammalian cultured cells, the synthesized rhodopsin genes were inserted into the cloning site between the CMV promoter and eYFP in pHKR2-3.0-EYFP⁵² using EcoRI and BamHI. All experiments were carried out using ND7/23 cells, lined hybrid cells derived from neonatal rat dorsal root ganglion neurons fused with the mouse neuroblastoma, which were transfected with plasmids as previously described⁵³. EYFP fluorescence (543 nm) in the ND7/23 cells expressing the rhodopsins were imaged under a confocal laser scanning microscopy (LSM510, Carl Zeiss, Oberkochen, Germany) at 512 \times 512 pixels using a water-immersion objective ($\times 63/0.95$, Achromplan, Carl Zeiss) and Ar laser (514 nm). Currents were recorded using an EPC-8 amplifier (HEKA Electronic, Lambrecht, Germany) under a whole-cell patch clamp configuration while a 200 ms pulse illuminations at 549 \pm 15 nm, $>90\%$ of the maximum) and 28 $\text{mW}\cdot\text{mm}^{-2}$ was given at 0.1 Hz using a SpectraX light engine (Lumencor Inc., Beaverton, OR). The internal pipette solution contained (in mM) 121.2 KOH, 90.9 glutamate, 5 Na_2EGTA , 49.2 HEPES, 2.53 MgCl_2 , 2.5 MgATP , 0.0025 ATR (pH 7.4 adjusted with HCl). The extracellular Tyrode's solution contained (in mM): 138 NaCl, 3 KCl, 2.5 CaCl_2 , 1 MgCl_2 , 10 HEPES, 4 NaOH, and 11 glucose (pH 7.4 adjusted with HCl).

Statistical analysis. We assessed the effectiveness of the data-driven experimental design method by comparing it with random selection in terms of the proportions

of observing red-shift gains in the selected rhodopsins. The statistical significance of the effectiveness was quantified by comparing the red-shift gain proportions 0.82 ($=32/39$, $p = 7.025 \times 10^{-5}$) with the probability of observing red-shift gains from randomly selected rhodopsins, i.e., 0.50, based on a binomial test. Since we set the base wavelength of each subfamily to the λ_{\max} of rhodopsin which was studied in detail in previous work and equal or longer than the empirical median of the λ_{\max} in each subfamily (Supplementary Fig. 2), it is reasonable to assume that the probability of observing red-shift gains from randomly selected rhodopsins must be smaller than or equal to 0.50. For statistical analysis of the ML model building and the evaluation of its performance, see the ML modeling section above.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data shown in main figures were deposited in Supplementary Data 9. Data supporting the findings are available from the corresponding authors upon reasonable request.

Code availability

The computational code of this manuscript is available at <http://www-als.ics.nitech.ac.jp/~karasuyama/BLASSO-for-Rhodopsins/>.

Received: 20 July 2020; Accepted: 19 February 2021;

Published online: 19 March 2021

References

- Ernst, O. P. et al. Microbial and animal rhodopsins: Structures, functions, and molecular mechanisms. *Chem. Rev.* **114**, 126–163 (2014).
- Govorunova, E. G., Sineshchekov, O. A., Li, H. & Spudich, J. L. Microbial rhodopsins: diversity, mechanisms, and optogenetic applications. *Annu. Rev. Biochem.* **86**, 845–872 (2017).
- Oesterheld, D. & Stoeckenius, W. Rhodopsin-like protein from the purple membrane of *Halobacterium halobium*. *Nat. New Biol.* **233**, 149–152 (1971).
- Oesterheld, D. & Stoeckenius, W. Functions of a new photoreceptor membrane. *Proc. Natl Acad. Sci. USA* **70**, 2853–2857 (1973).
- Man, D. et al. Diversification and spectral tuning in marine proteorhodopsins. *EMBO J.* **22**, 1725–1731 (2003).
- Venter, J. C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74 (2004).
- Inoue, K., Kato, Y. & Kandori, H. Light-driven ion-translocating rhodopsins in marine bacteria. *Trends Microbiol.* **23**, 91–98 (2014).
- Inoue, K. et al. A light-driven sodium ion pump in marine bacteria. *Nat. Commun.* **4**, 1678 (2013).
- Nagel, G. et al. Channelrhodopsin-1: a light-gated proton channel in green algae. *Science* **296**, 2395–2398 (2002).
- Niho, A. et al. Demonstration of a light-driven SO_4^{2-} transporter and its spectroscopic characteristics. *J. Am. Chem. Soc.* **139**, 4376–4389 (2017).
- Deisseroth, K. Optogenetics 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18**, 1213–1225 (2015).
- Liu, X. et al. Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature* **484**, 381–385 (2012).
- Ramirez, S. et al. Creating a false memory in the hippocampus. *Science* **341**, 387–391 (2013).
- Yizhar, O. et al. Neocortical excitation/inhibition balance in information processing and social dysfunction. *Nature* **477**, 171–178 (2011).
- Marshall, J. H. et al. Cortical layer-specific critical dynamics triggering perception. *Science* **365**, eaaw5202 (2019).
- Schneider, F., Grimm, C. & Hegemann, P. Biophysics of channelrhodopsin. *Annu. Rev. Biophys.* **44**, 167–186 (2015).
- Inoue, K. et al. Red-shifting mutation of light-driven sodium-pump rhodopsin. *Nat. Commun.* **10**, 1993 (2019).
- Ganapathy, S. et al. Retinal-based proton pumping in the near infrared. *J. Am. Chem. Soc.* **139**, 2338–2344 (2017).
- Hayashi, S. et al. Structural determinants of spectral tuning in retinal proteins-bacteriorhodopsin vs sensory rhodopsin II. *J. Phys. Chem. B* **105**, 10124–10131 (2001).
- Fujimoto, K., Hayashi, S., Hasegawa, J. Y. & Nakatsuji, H. Theoretical studies on the color-tuning mechanism in retinal proteins. *J. Chem. Theory Comput.* **3**, 605–618 (2007).
- Pedraza-González, L., De Vico, L., Marí, N. M., Fanelli, F. & Olivucci, M. a-ARM: automatic rhodopsin modeling with chromophore cavity generation, ionization state selection, and external counterion placement. *J. Chem. Theory Comput.* **15**, 3134–3152 (2019).

22. Tsujimura, M. et al. Mechanism of absorption wavelength shifts in anion channelrhodopsin-1 mutants. *Biochim. Biophys. Acta Bioenerg.* **1862**, 148349 (2021).
23. Katayama, K. & Sekharan, S. S. Y. *Optogenetics* (eds Yawo, H., Kandori, H. & Koizumi, A.) Ch. 7, 89–107 (Springer, 2015).
24. Engqvist, M. K. et al. Directed evolution of *Gloeobacter violaceus* rhodopsin spectral properties. *J. Mol. Biol.* **427**, 205–220 (2015).
25. Kojima, K. et al. Green-sensitive, long-lived, step-functional anion channelrhodopsin-2 variant as a high-potential neural silencing tool. *J. Phys. Chem. Lett.* **11**, 6214–6218 (2020).
26. Pushkarev, A. et al. The use of a chimeric rhodopsin vector for the detection of new proteorhodopsins based on color. *Front. Microbiol.* **9**, 439 (2018).
27. Oda, K. et al. Crystal structure of the red light-activated channelrhodopsin Chrimson. *Nat. Commun.* **9**, 3949 (2018).
28. Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
29. Govorunova, E. G. et al. RubyACRs, nonalgal anion channelrhodopsins with highly red-shifted absorption. *Proc. Natl Acad. Sci. USA* **117**, 22833–22840 (2020).
30. Karasuyama, M., Inoue, K., Nakamura, R., Kandori, H. & Takeuchi, I. Understanding colour tuning rules and predicting absorption wavelengths of microbial rhodopsins by data-driven machine-learning approach. *Sci. Rep.* **8**, 15580 (2018).
31. Brown, G. R. et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, D36–D42 (2015).
32. Sunagawa, S. et al. Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
33. Bishop, C. M. *Pattern Recognition And Machine Learning* (Springer, 2006).
34. Snoek, J., Larochelle, H. & Adams, R. P. *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. (eds Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 2951–2959 (Curran Associates, Inc., 2012).
35. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & Freitas, N. D. in *Proceedings of the IEEE*. 148–175 (IEEE, 2016).
36. Park, T. & Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **103**, 681–686 (2008).
37. Johnson, M. et al. Ncbi blast: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
38. Yamauchi, Y. et al. Engineered functional recovery of microbial rhodopsin without retinal-binding lysine. *Photochem. Photobiol.* **95**, 1116–1121 (2019).
39. Hasemi, T., Kikukawa, T., Kamo, N. & Demura, M. Characterization of a cyanobacterial chloride-pumping rhodopsin and its conversion into a proton pump. *J. Biol. Chem.* **291**, 355–362 (2016).
40. Harris, A. et al. Molecular details of the unique mechanism of chloride transport by a cyanobacterial rhodopsin. *Phys. Chem. Chem. Phys.* **20**, 3184–3199 (2018).
41. Béjà, O. et al. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* **289**, 1902–1906 (2000).
42. Ozaki, Y., Kawashima, T., Abe-Yoshizumi, R. & Kandori, H. A color-determining amino acid residue of proteorhodopsin. *Biochemistry* **53**, 6032–6040 (2014).
43. Shimono, K., Ikeura, Y., Sudo, Y., Iwamoto, M. & Kamo, N. Environment around the chromophore in *pharaonis* phoborhodopsin: Mutation analysis of the retinal binding site. *Biochim. Biophys. Acta* **1515**, 92–100 (2001).
44. Sudo, Y. et al. A blue-shifted light-driven proton pump for neural silencing. *J. Biol. Chem.* **288**, 20624–20632 (2013).
45. Inoue, K. et al. Converting a light-driven proton pump into a light-gated proton channel. *J. Am. Chem. Soc.* **137**, 3291–3299 (2015).
46. Fenno, L., Yizhar, O. & Deisseroth, K. The development and application of optogenetics. *Annu. Rev. Neurosci.* **34**, 389–412 (2011).
47. Thompson, J. D., Higgins, D. G. & Gibson, T. J. Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
48. Khorana, H. G. et al. Amino acid sequence of bacteriorhodopsin. *Proc. Natl Acad. Sci. USA* **76**, 5046–5050 (1979).
49. Pushkarev, A. et al. A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. *Nature* **558**, 595–599 (2018).
50. Saitou, N. & Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
51. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
52. Kato, H. E. et al. Structural basis for Na⁺ transport mechanism by a light-driven Na⁺ pump. *Nature* **521**, 48–53 (2015).
53. Nagasaka, Y. et al. Gate-keeper of ion transport—a highly conserved helix-3 tryptophan in a channelrhodopsin chimera, C1C2/ChRWR. *Biophys. Physicobiol.* **17**, 59–70 (2020).

Acknowledgements

This work was supported by Grants-in-Aid from the Japan Society for the Promotion of Science (JSPS) for Scientific Research (KAKENHI grant Nos. 17H03007 to K.I., 17H04694 and 16H06538 to M.Karasuyama, 19H04959 to H.K., and 16H06538, 17H00758, and 20H00601 to I.T.), the Japan Science and Technology Agency (JST), PRESTO, Japan (grant Nos. JPMJPR15P2 to K.I. and JPMJPR15N2 to M.Karasuyama), and CREST, Japan (grant No. JPMJCR1502) to I.T.; K.I., H.K., and I.T. received support from RIKEN AIP; O.B. received support from the Louis and Lyra Richmond Memorial Chair in Life Sciences.

Author contributions

K.I., M.Karasuyama, H.K., and I.T. contributed to the study design. K.I., D.Y., K.Y., and O.B. conducted the phylogenetic analysis of rhodopsins and the construction of training data. M.Karasuyama, Y.I., and I.T. constructed the ML model and calculated E_{gain} . K.I., R.N., K.M., and T.N. constructed the DNA plasmids of rhodopsin genes and introduced them into *E. coli* and mammalian cells. R.N. and K.M. measured λ_{max} of rhodopsins by bleaching proteins with hydroxylamine. M.Konno conducted the pump activity assay of rhodopsins in *E. coli* cells. H.Y. conducted the electrophysiological measurement of rhodopsins in mammalian cells. K.I., M. Karasuyama, H.K., and I.T. wrote the paper. All authors discussed and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42003-021-01878-9>.

Correspondence and requests for materials should be addressed to K.I. or I.T.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021