


SCIENTIFIC REPORTS



OPEN

Understanding Colour Tuning Rules and Predicting Absorption Wavelengths of Microbial Rhodopsins by Data-Driven Machine-Learning Approach

Masayuki Karasuyama^{1,2,3}, Keiichi Inoue^{4,2,5,6}, Ryoko Nakamura⁷, Hideki Kandori^{4,5,6}  & Ichiro Takeuchi^{1,3,7}

The light-dependent ion-transport function of microbial rhodopsin has been widely used in optogenetics for optical control of neural activity. In order to increase the variety of rhodopsin proteins having a wide range of absorption wavelengths, the light absorption properties of various wild-type rhodopsins and their artificially mutated variants were investigated in the literature. Here, we demonstrate that a machine-learning-based (ML-based) data-driven approach is useful for understanding and predicting the light-absorption properties of microbial rhodopsin proteins. We constructed a database of 796 proteins consisting of microbial rhodopsin wildtypes and their variants. We then proposed an ML method that produces a statistical model describing the relationship between amino-acid sequences and absorption wavelengths and demonstrated that the fitted statistical model is useful for understanding colour tuning rules and predicting absorption wavelengths. By applying the ML method to the database, two residues that were not considered in previous studies are newly identified to be important to colour shift.

Microbial rhodopsin is a photoreceptive membrane protein of microbial species, such as eubacteria, archaea, fungi, and algae. The functions of microbial rhodopsin are very diverse. Light-driven ion (proton, chloride, sodium, and so on) pumps, light-gated cation and anion channels, photochromatic gene regulator and light-regulated enzymes have been reported for various species¹. The light-dependent ion-transport function of microbial rhodopsin is widely used in optogenetics for optical control of neural activity in the brain network². Most microbial rhodopsins bind a common chromophore, all-*trans* retinal, via a protonated Schiff-base linkage in the center of the hepta-transmembrane scaffold (Fig. 1). Each microbial rhodopsin exhibits a variety of specific visible absorption wavelengths of their retinal. While the protonated all-*trans* retinal Schiff-base shows the absorption peak at ~450 nm in organic solvents³, the wavelengths of absorption maxima of retinal (λ_{\max}) in microbial rhodopsin range from 436 nm of channel-rhodopsin from *Tetraselmis striata* (TsChR)⁴ to 587 nm of sensory rhodopsin I⁵. This wide-range colour tuning of the retinal in rhodopsin is considered to be achieved by optimizing the steric and/or electrostatic interaction with surrounding amino-acid residues.

¹Department of Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, Aichi, 466-8555, Japan. ²PRESTO, Japan Science and Technological Agency (JST), 4-1-8 Honcho, Kawaguchi, Saitama, 332-0012, Japan. ³Center for Materials research by Information Integration, National Institute for Materials Science (NIMS), Tsukuba, 305-0047, Japan. ⁴Department of Life Science and Applied Chemistry, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, Aichi, 466-8555, Japan. ⁵OptoBioTechnology Research Center, Gokiso, Showa-ku, Nagoya, Aichi, 466-8555, Japan. ⁶The Institute for Solid State Physics, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba, 277-8581, Japan. ⁷RIKEN Center for Advanced Intelligence Project, Nihonbashi 1-chome, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan. Masayuki Karasuyama and Keiichi Inoue contributed equally. Correspondence and requests for materials should be addressed to H.K. (email: kandori@nitech.ac.jp) or I.T. (email: takeuchi.ichiro@nitech.ac.jp)

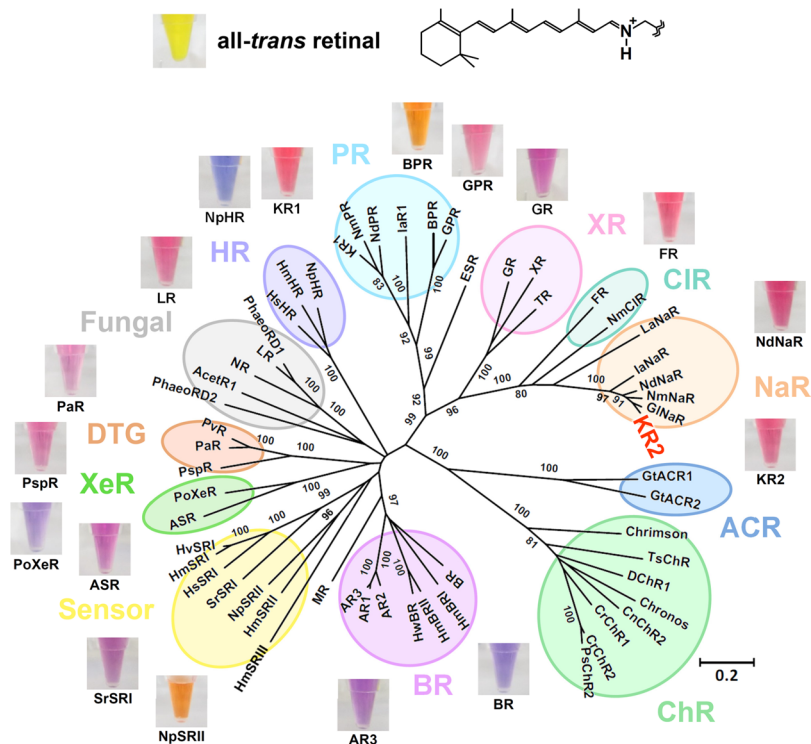


Figure 1. The chemical structure of all-*trans* retinal (upper) and phylogenetic tree of microbial rhodopsins (lower). The bootstrap values >80% are shown for the corresponding branches. The photographs of the DMSO solution of all-*trans* retinal and detergent solubilized rhodopsins were aligned to show representative colours. The abbreviations of rhodopsin proteins are listed in Supplementary Information 1. In the present paper, we construct a machine-learning-based (ML-based) statistical model that describes the relationship between amino-acid sequences and absorption wavelengths of microbial rhodopsins based on past experimental data.

Increasing the variety of absorption wavelengths enables simultaneous optical control by different colours of light. Furthermore, the microbial rhodopsin having highly red-shifted absorption maximum is strongly demanded for optogenetic application, because of the lower phototoxicity and higher tissue-penetration length of longer-wavelength light⁴. As such, various rhodopsin genes have been screened in order to find additional colour-shifted proteins^{4,6}. While many blue-absorbing rhodopsin at $\lambda < 500$ nm have been reported⁷ and even applied to optogenetics⁴, the longer absorption maxima are limited in < 600 nm. Thus, further artificial molecular modifications of protein were needed in order to achieve greater red-shifted absorption. Random and/or semi-empirical point mutations identify the types of amino-acid mutation that are effective for colour tuning^{8,9}. Although numerous mutations causing bathochromic shift without disrupting protein function were identified in this way, the degree of shift is insufficient for application, and comprehensive screening is difficult because of the large number of possible mutations ($> 20^{200}$). Although more rational molecular design is expected for quantum chemical calculation to estimate the absorption energy^{10–15}, its high calculation cost makes application to wide-range screening difficult. An alternative technique for expanding the absorption range is the incorporation of natural or artificial retinal analogues¹⁶. For optogenetic application, however, a tissue-directed delivery method of these analogues must be developed.

In the present paper, we report the results of a data-driven approach for studying the light-absorption properties of microbial rhodopsin proteins by machine learning (ML). We constructed a database of 796 proteins consisting of microbial rhodopsin wildtypes and their variants, some of which were previously reported in the literature and others of which are newly reported herein (see Supplementary Table 1). Each entry of the database consists of the amino-acid sequence and absorption wavelength λ_{\max} of a rhodopsin. We introduce an ML method for constructing a statistical model describing the relationship between amino-acid sequences and absorption wavelengths. The goal of the present paper is to demonstrate the effectiveness of ML-based data-driven approaches for functional protein studies. By constructing a database based on past experimental results and applying an ML method to the database, a statistical model describing the relationship between amino-acid sequences and molecular properties can be constructed. In the context of microbial rhodopsin studies, we illustrate the utility of such a statistical model by demonstrating that it can be effectively used for understanding the colour tuning rules and predicting the absorption wavelength (see Fig. 2).

We consider the following hypothetical scenario for the purpose of demonstration. The database is divided into two sets: a target protein set and a training protein set. The target set contains KR2 wild-type rhodopsin and its variants (which, in the present study, are assumed to be uninvestigated as of yet), whereas the training set contains the remaining proteins in the database. We constructed an ML model using only the proteins in the training set. The constructed model was then applied to the proteins in the target set for predicting the absorption

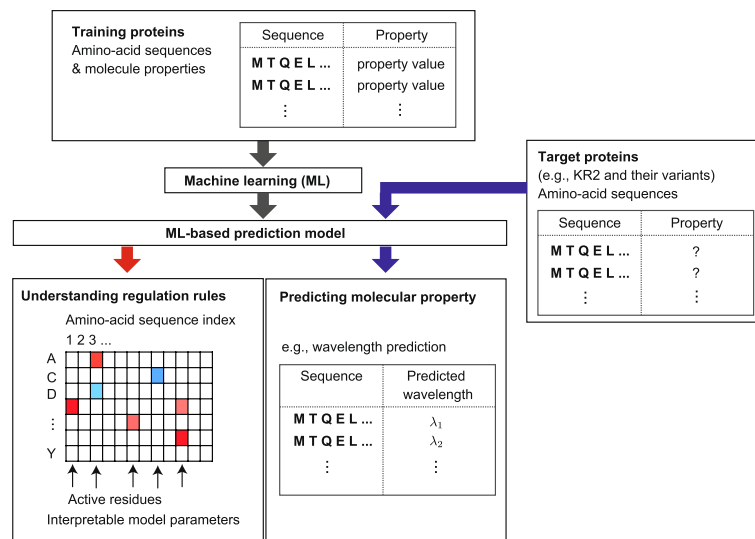


Figure 2. An overview of the machine-learning-based (ML-based) data-driven approach introduced in the present paper for functional protein studies. Using past experimental data, a *training protein set* containing pairs of amino-acid sequence and molecular properties is first constructed. Then, an ML method is applied to the training set, and an ML-based statistical model is constructed. The obtained ML model can be used in understanding the relationship between amino-acid sequences and molecular properties, such as the colour tuning rules in the case of microbial rhodopsins. The ML model can also be used to predict the molecular properties of new uninvestigated proteins. We refer to the set of new proteins as the *target protein set*. In the present paper, for the purpose of demonstration, we regard KR2 wildtype and its 118 variants as target proteins and other 677 rhodopsin proteins in the database as the training proteins.

wavelengths of KR2 and its variants. This scenario is interpreted as a hypothetical situation where a researcher is interested in predicting the absorption wavelengths of a new group of rhodopsin proteins based on previously reported data on other groups of rhodopsin proteins.

Among the various available ML methods, we used a *group-wise sparse learning* approach^{17–19}. The advantages of group-wise sparse learning approaches are not only predictability but also interpretability of the constructed models. As we report later herein, by using a group-wise sparse learning approach, the absorption wavelengths of KR2 and its variants could be predicted from their amino-acid sequences with an average error of ± 7.8 nm. The residues affecting the absorption wavelength were also identified, and their strength for colour shift and the effect of mutation were quantitatively investigated. Through this analysis, the positions of BR Glu161 and Ala126, the effects for colour shift of which were not reported in previous studies, were newly shown to significantly affect the absorption wavelengths. Furthermore, the model constructed by a group-wise sparsity learning approach enables the identification of *active residues*, i.e., residues for which the choice of the amino-acid species has a great influence on the absorption wavelength. Although we herein focus on the prediction of absorption wavelengths of rhodopsin proteins, the same ML approach can be used to predict other molecular properties in other types of functional proteins.

Results

Microbial rhodopsin database. In order to demonstrate the effectiveness of ML-based data-driven approaches for microbial rhodopsin studies, we constructed a database. The database is composed of amino-acid sequences and absorption wavelengths λ_{\max} s of 519 proteins previously reported in the literature and 277 proteins investigated by our group without previous report (see Supplementary Table 1). Noteworthy, some of reported value of λ_{\max} of BR mutants regenerated from extracted crude membrane with a mixture of organic solvent from *E. coli*²⁰ showed significant discrepancy from that observed in purple membrane obtained from *Halobacterium salinarum*²¹. Hence, we considered the results of regenerated BR may not properly show absorption wavelength and did not included in training data in machine learning. As reported in a previous study²², for data-driven approaches such as the present study, it is important to construct a database containing not only reported experimental results but also unreported results. We applied alignment algorithm ClustalW to these amino-acid sequences and obtained aligned sequences of 475 residues, among which we extracted the transmembrane region, resulting in 210 residues. For the purpose of demonstration, we divided the dataset into a *target protein set* and a *training protein set* (see Fig. 3).

The target set consists of 119 rhodopsin proteins in the KR2 group (KR2 wildtype and its 118 variants), whereas the training set consists of the remaining 677 rhodopsin proteins (see Figs 1 and 3). We applied an ML method to the training set and constructed a statistical model describing the relationship between the amino-acid sequences and absorption wavelengths. The statistical model was then applied to the rhodopsin proteins in the target set in order to predict their absorption wavelengths. This scenario assumes a hypothetical situation in which a researcher is interested in investigating a new group of rhodopsin proteins based on previously reported data on other groups of rhodopsin proteins.

Protein	Amino-acid Sequence (transmembrane region, N = 210)	λ_{\max} / nm
BR	TGRPE...RYADWLF...FTPL...LLLDL...DVSAK...IFG	560
AR3	LGLGD...RYADWLF...FTPL...LLLDL...DVTAK...AIL	552
NpHR	PLLAS...RYLTWALSTP...MILLAL...DIVAK...TSN	577
Rhodopsins other than KR2 677 proteins	.	.
	.	.
	.	.
	.	.
	.	.
	.	.
KR2 wildtype	KR2 FSEIA...RYLNW...LIDVP...MLLFQ...DVSSK...TLS	524
KR2 mutants 118 proteins	KR2 D116N FSEIA...RYLNW...LINVP...MLLFQ...DVSSK...TLS	565
.	.	.
.	.	.
.	.	.

Figure 3. Structure of the database used in the present study. The database is composed of the sequences and λ_{\max} s of 519 previously reported proteins and 277 newly reported proteins. We used 677 rhodopsin proteins other than KR2 and their variants as the training proteins (red rectangle) and 119 proteins in KR2 group as the target proteins (blue rectangle), respectively.

Machine learning method. In order to handle amino-acid sequences in the ML framework, we introduced a binary representation, as depicted in Fig. 4(a). Let $M = 20$ be the number of different amino-acid species, and let $N = 210$ be the number of residues considered herein. Then, an amino-acid sequence is represented by $M \times N = 4,200$ binary variables, which we denote as $\mathbf{x} \in \{0, 1\}^{MN}$. We consider a linear model for such MN -dimensional variables with an intercept parameter β_0 and MN coefficient parameters $\beta_{i,j}$, $i = 1, \dots, M$, $j = 1, \dots, N$ (see Fig. 4(b)). These $1 + MN$ parameters are fitted based on the training set so that the output of the model $f(\mathbf{x})$ can predict the absorption wavelength of the rhodopsin protein for which the amino-acid sequence is coded as \mathbf{x} . Since this model has so many parameters, it is difficult to interpret the fitted model if we simply use conventional methods such as the least-squares method. We thus introduced the *group-wise sparsity mechanism* (See the Method section and the Supplemental Information for details). Using this mechanism, the fitted coefficient parameters $\beta_{i,j}$ have *residue-wise sparsity*. Here, $M = 20$ coefficient parameters corresponding to the choice of an amino-acid species in each residue is considered as a group. After we fitted the model, in many groups, all of the M coefficient parameters become zero, indicating that the choice of an amino-acid species in these residues does not affect the colour tuning property. On the other hand, a small number of residues at which the coefficient parameters are NOT zero are called *active residues*, i.e., the choice of the amino-acid species in these residues is expected to play an important role in colour tuning. Figure 4(c) illustrates the fitted coefficient parameters using the group-wise-sparsity mechanism. If a parameter $\beta_{i,j}$ is positive/negative, then the i -th amino-acid species in the j -th residue has a red-shifting/blue-shifting effect on the light absorption properties of rhodopsin proteins.

Understanding colour tuning rules. By applying the above ML method to the training set containing pairs of the amino-acid sequence and absorption wavelength for 677 rhodopsin proteins, we fitted a linear model with $1 + MN = 4,201$ parameters. A complete list of the fitted parameters is presented in Supplementary Table 2. Figure 5 shows the fitted coefficient parameters at 20 active residues in decreasing order of $s_j = \sqrt{\sum_{i=1}^M \beta_{i,j}^2}$, $j = 1, \dots, N$, where the score s_j quantifies the *activeness* of the j -th residue. Here, red and blue indicate that the corresponding parameters are positive and negative, respectively, whereas grey indicates that the parameters were zero. In other words, red and blue suggest that having the amino-acid species in the residue would have a red-shifting and a blue-shifting effect, respectively. The results in Supplementary Table 2 and Fig. 5 can be interpreted as a comprehensive statistical description of the colour tuning rules of rhodopsin proteins based on previously investigated experimental results for 677 rhodopsin proteins (Supplementary Fig. 1 shows the same results obtained using all 796 rhodopsin proteins, including those in the KR2 group).

Predicting absorption wavelengths of KR2 rhodopsin and its variants. Using the statistical model fitted based on the training set (containing all of the rhodopsin proteins except for the KR2 group), the absorption wavelengths of the 119 rhodopsin proteins in the target set (containing KR2 group rhodopsin proteins) were predicted. Figure 6(a,b) show examples of predicted (green lines) and observed (blue lines) wavelengths for red-shifted KR2 mutants. For the KR2 NTQ/F72G mutant (Fig. 6(a)), the difference between the predicted (546.44 nm) and experimentally observed (543 nm) wavelengths is only 3.44 nm. In contrast, we observed a larger discrepancy (8.51 nm) for the predicted (556.49 nm) and experimentally observed (565 nm) wavelengths for KR2 D116N. This means that the precision of ML prediction differs for each type of mutation. Examples of blue-shifted mutants are shown in Fig. 6(c) (KR2 N112E) and Fig. 6(d) (KR2 DTD/D102N). The differences between the prediction and the observation were 7.34 and 19.92 nm for the former and latter, respectively. Figure 6(e) summarizes the prediction results for KR2 and all of its mutants, where the horizontal axis represents the *observed* absorption wavelengths measured in the experiments, whereas the vertical axis represents the *predicted* absorption wavelengths obtained by the ML model. The red points indicate the KR2 group rhodopsin proteins in the target set, whereas the black points indicate other rhodopsin proteins in the training set. Note that the prediction performance in the training set (black points) is slightly better than that in the target set (red points). This is because the former is used for fitting the ML model itself, whereas the latter is completely new to the fitted model. This phenomenon is known as *over-fitting* in the literature of machine learning. The absorption wavelengths of KR2 and its variants could be predicted from their amino-acid sequences with average errors of ± 7.8 nm. The

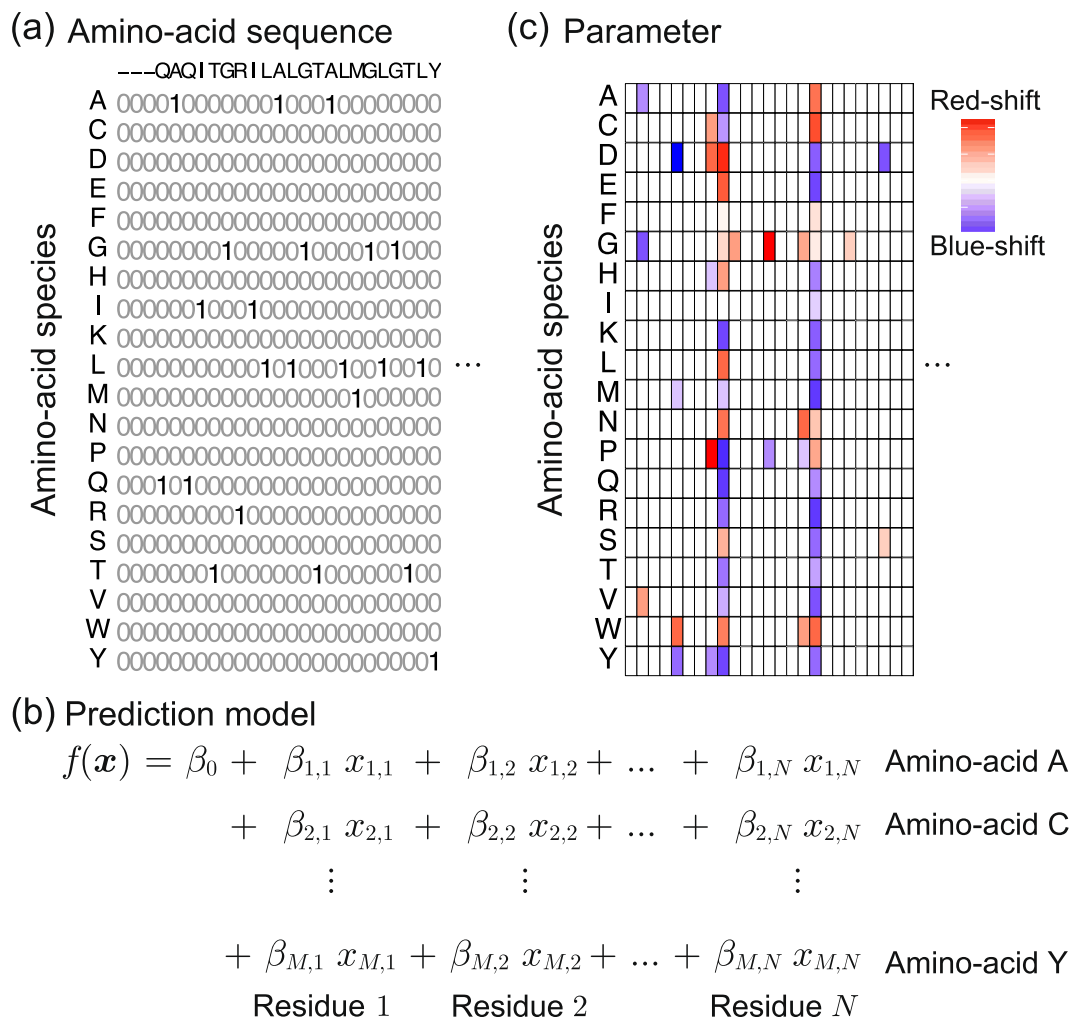


Figure 4. A schematic description of the ML method introduced in the present paper for functional protein studies. (a) Binary sequence representation of an amino-acid sequence. Let $M = 20$ be the number of amino-acid species, and let N be the number of residues considered in the present study. Then, the amino-acid sequence of a protein is represented by $M \times N$ binary variables, each of which represents the amino-acid species at each residue. (b) By writing the MN binary variables as $x_{i,j}$, $i = 1, \dots, M$, $j = 1, \dots, N$, we consider an MN -dimensional linear model. The linear model has an intercept parameter β_0 and MN coefficient parameters $\beta_{i,j}$, $i = 1, \dots, M$, $j = 1, \dots, N$. (c) When the linear model is fitted, a group-wise sparsity constraint is introduced. Then, in many residues, all of the corresponding M coefficients would be fitted to zero, and only a small number of residues have nonzero coefficient parameters. The latter residues are called *active residues*. The choice of amino-acid species in these active residues is expected to play an important role in determining molecular properties such as absorption wavelength.

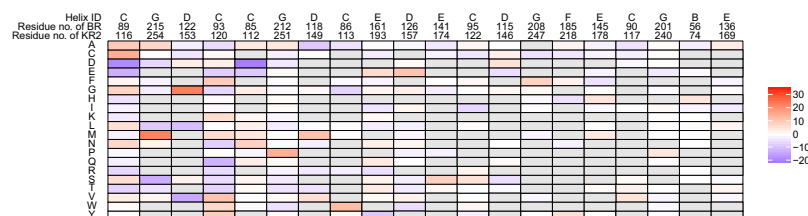


Figure 5. Coefficient parameters of the fitted statistical model. Coefficients for the top 20 active residues, where the activeness of each residue is defined as $s_j = \sqrt{\sum_{i=1}^M \beta_{i,j}^2}$, $j = 1, \dots, N$. Here, red and blue indicate that the corresponding parameters are positive and negative, respectively, whereas grey indicates that the amino-acid species did not exist in the training data. The figure can be interpreted such that, if the value of a coefficient parameter $\beta_{i,j}$ is positive/negative (i.e. red/blue), then the existence of the i -th amino-acid species at the j -th residue has a red-shifting/blue-shifting effect.

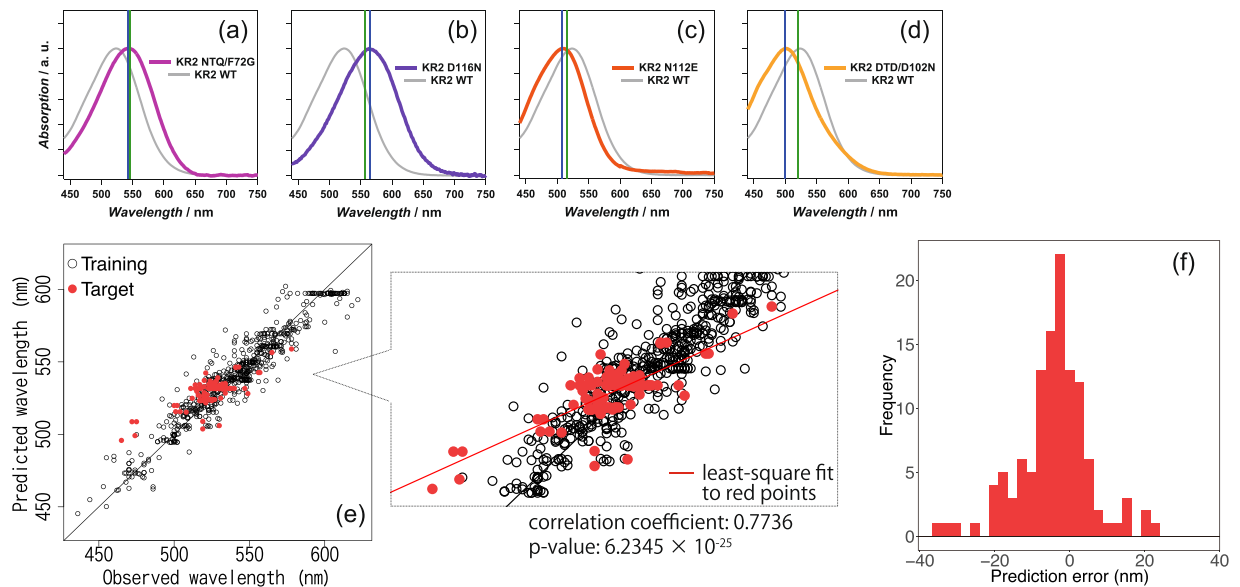


Figure 6. Absorption wavelength prediction results for KR2 wildtype and its 118 variants. (a–d) Absorption spectra of KR2 mutants (a) KR2 NTQ/F72G, (b) D116N, (c) N112E, and (d) DTD/D102N with their absorption maxima as predicted by ML analysis (green lines) and experimentally determined (blue lines). The spectrum of KR2 wildtype is indicated by the solid grey line. (e) The horizontal axis represents the experimentally observed absorption wavelengths, whereas the vertical axis represents the absorption wavelengths predicted by the ML model. The red points indicate the KR2 group rhodopsin proteins in the target set, whereas the black points indicate other rhodopsin proteins in the training set. (f) Histogram of the prediction errors for KR2 group proteins in the target set.

correlation coefficient ρ between the predicted and the observed wavelengths was $\rho = 0.7736$ (p -value for the null hypothesis $\rho = 0$ was 6.2345×10^{-25}). The histogram in Fig. 6(f) shows the distribution of the prediction errors in the KR2 group rhodopsin proteins in the target set. In this histogram, we observed the data highly deviated from the prediction for more than 25 nm. Interestingly, all of them were the mutants of KR2 Gly153 (KR2 G153I, G153F, G153V and G153L). KR2 Gly153 is homologous to BR Gly122 and directly interacting with β -ionone ring of retinal. The many previous study suggested that the mutation of this glycine to different types of amino acid uniformly blue-shifts various rhodopsins such as bacteriorhodopsin from *H. walsbyi* (*HwBR*), AR3, chimeric channel rhodopsin C1C2 and GR^{9,23,24}. The experimentally observed λ_{\max} for these KR2 mutants was showed larger hypsochromic shift than that predicted by ML. For the reason of blue shift by the mutation of this Gly, X-ray crystallographic structural analysis and computational simulation suggested that increased volume of side chain rotates β -ionone ring of retinal resulting in the shortening of π -conjugation system²⁴. The current results for KR2 implies the induced rotation of β -ionone ring is much larger compared with other rhodopsins.

Estimating the effect of point mutations. The effect of a point mutation on the absorption wavelength shift can be estimated based on the coefficient parameters $\beta_{i,j}$, $i = 1, \dots, M$, $j = 1, \dots, N$. Let $\mathbf{x}^{(\text{KR2})} \in \{0, 1\}^{MN}$ be the binary vector representation of the KR2 wild-type sequence. The difference in the predicted absorption wavelengths between KR2 wildtype and a variant having amino-acid sequence $\mathbf{x}^{(\text{Var})} \in \{0, 1\}^{MN}$ is written as

$$f(\mathbf{x}^{(\text{Var})}) - f(\mathbf{x}^{(\text{KR2})}) = \sum_{i=1}^M \sum_{j=1}^N \beta_{i,j} x_{i,j}^{(\text{Var})} - \sum_{i=1}^M \sum_{j=1}^N \beta_{i,j} x_{i,j}^{(\text{KR2})}.$$

The colour-shifting effect of point mutation at the j -th residue is written as

$$\sum_{i=1}^M \beta_{i,j} (x_{i,j}^{(\text{Var})} - x_{i,j}^{(\text{KR2})}). \quad (1)$$

For example, if the i_1 -th amino-acid species in the KR2 wildtype is replaced by the i_2 -th amino-acid species, the colour-shifting effect of the point mutation is $\beta_{i_2,j} - \beta_{i_1,j}$. Figure 7 shows a portion of the amino-acid sequences of KR2 wildtype and its variants along with their observed and predicted absorption wavelengths. In Fig. 7, red and blue indicate red-shifting and blue-shifting effects, respectively, in Eq. (1) estimated by the trained statistical model. Figure 7(a) suggests that point mutation at BR residue number 89 would have red-shifting effects. On the other hand, Fig. 7(b) suggests that point mutation at BR residues 85 and 122 would have blue-shifting effects. These results indicate that the estimated colour-shifting effects are consistent with the actual observed wavelength shifts caused by the mutation.

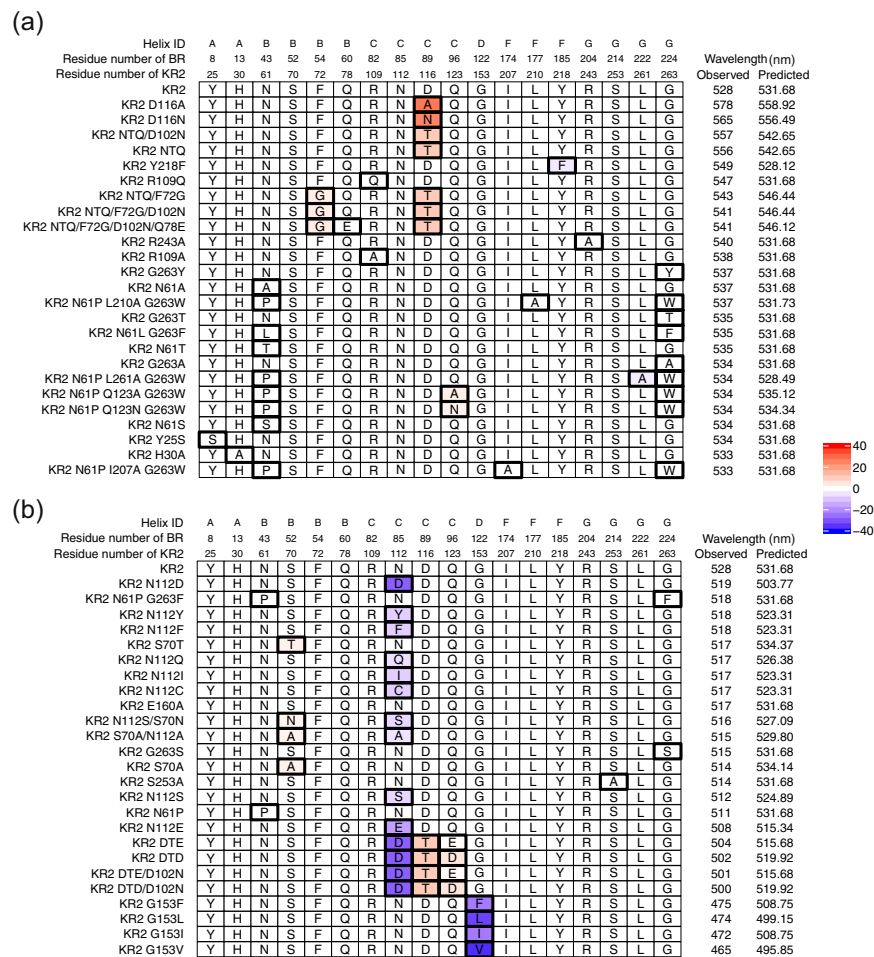


Figure 7. Lists of sequences for the KR2 wildtype and the variants with their observed and predicted absorption wavelengths. **(a)** KR2 and the 25 variants that have the longest observed wavelengths, and **(b)** KR2 and the 25 variants that have the shortest observed wavelengths. The residues shown here are replaced at least once among the 50 variants. Boxes with thick black lines indicate positions that have different amino-acid species from the KR2 wildtype. For these boxes, the colour indicates the wavelength change produced by the replacement of the j -th position, estimated by $\sum_{i=1}^M \beta_{i,j} (x_{i,j}^{(\text{Var})} - x_{i,j}^{(\text{KR2})})$, where $x_{i,j}^{(\text{KR2})}$ and $x_{i,j}^{(\text{Var})}$ are the binary representation the KR2 wildtype and a variant, respectively.

Discussion

Colour tuning rules in the estimated statistical models by ML.

Ten residues showing the highest β -values were overlaid on the X-ray crystallographic structure of BR (PDB code: 1BM1) (see Fig. 8). Eight of these residues are located around retinal within $<5 \text{ \AA}$ (BR Thr89, Ala215, Gly122, Leu93, Asp85, Asp212, Met118, and Trp86 in the order of activeness). Thr89 showed the highest degree of activeness. This is a member of the DTD-motif, which represents the type of functional determining three residues in the third transmembrane helix (helix-C) for each ion-pump rhodopsin. The DTD-motif is typical for the outward H^+ pump and is composed of Asp85, Thr89, and Asp96 for BR²⁵. While this threonine is conserved among most microbial rhodopsins, it is replaced with an aspartate for sodium pump rhodopsin (NaR), which has the NDQ-motif rather than the DTD-motif^{25–27}. The position of BR Thr89 is close to RSB (the distance between BR Thr89C γ and the nitrogen atom of RSB is 3.4 \AA). The third and seventh active residues are BR Gly122 and Met118, respectively. These residues are highly conserved among various microbial rhodopsins. Their mutation causes the rotation of the C6-C7 bond of retinal and the shortening of the π -electron conjugation between the β -ionone ring and the polyene chain^{23,24}. The largest coefficient parameters are obtained for glycine and methionine for the former and latter positions. This implies any type of mutation of these residues results in the blue-shift of λ_{max} and is consistent with previous experimental reports^{23,24}.

The residues of BR Ala215 and Leu93 exhibit the second and fourth highest degrees of activeness. Both BR Ala215 and Leu93 are well known to have a role in colour-tuning switching for various rhodopsins in nature. Shimono *et al.* reported that, whereas green-to-orange absorbing archeal rhodopsins (BR, halorhodopsin and sensory rhodopsin I) conserve an alanine at the position of BR Ala215, blue-absorbing rhodopsins, such as *pharaonis* phoborhodopsin (*ppR*, which is also referred to as *pharaonis* sensory rhodopsin II) has a serine or threonine at this position²⁸. The difference of coefficient parameter values is approximately 11.8, which is close to the reported

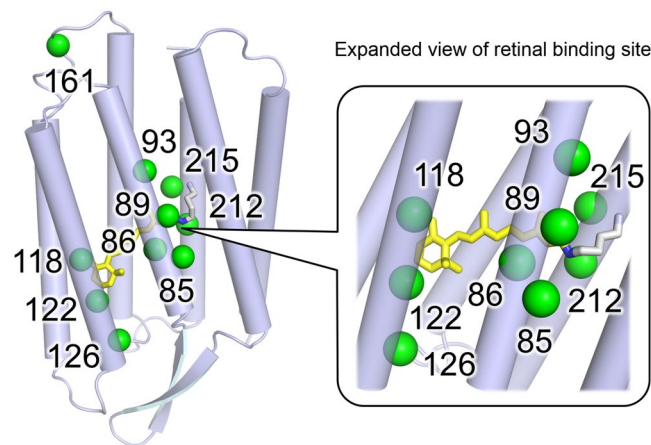


Figure 8. Top 10 active residues identified by the fitted statistical model. The positions of the active residues showing larger coefficient parameter values (green spheres) are mapped on the X-ray crystallographic structure of BR (blue, PDB code: 1BM1³³) with their numbers in the case of BR.

λ_{\max} shift of ppR T204A (8-nm red-shift)²⁸ and the BR homolog of *HwBR* A223T (13-nm blue-shift)²³. BR Leu93 corresponds to Leu120 of green-absorbing proteorhodopsin (GPR). This residue is replaced with a glutamine in blue-absorbing proteorhodopsin (BPR), and this type of colour regulation is known as “L/Q-switching”²⁹. The lowest coefficient parameter (−11.2) was obtained for a glutamine. This suggests that glutamine is most effective to achieve blue-shift absorption and is considered to be optimized in natural evolution in the deep-ocean environment²⁹. Ozaki *et al.* reported that mutations to valine or bulky residues (lysine, phenylalanine, tyrosine, and tryptophan) cause a large red-shift³⁰ of λ_{\max} . Their larger coefficient parameters are consistent with previous experimental results (Fig. 5).

BR Asp85 and Asp212 are generally deprotonated and work as counterions to protonated RSB. The electrostatic interaction between their negative charges and the π -electron of retinal destabilizes the energy level of the electronically excited state. This results in the blue-shift of λ_{\max} ³¹. Whereas the aspartate at the position of BR Asp85 has the second lowest coefficient value (−19.5) among all of the residues investigated herein, the value of the position of BR Asp212 is moderate (−3.2). This result suggests that the former has a much stronger effect on colour tuning, despite the symmetric location of these two residues relative to RSB. (The distances from Asp85 and Asp212 to the N atom of RSB are 3.4 and 3.5 Å, respectively.)

The eighth largest coefficient parameter was the position of BR Trp86. This tryptophan is one of the most highly conserved residues among microbial rhodopsins. It forms a part of the binding pocket by direct contact with the extracellular side of the polyene chain of retinal¹. This strong interaction with retinal is consistent with the high degree of activeness of this residue and the coefficient parameter of tryptophan is a large positive value (12.0). This suggests that this tryptophan has a role in shifting the absorption wavelength to be longer in many rhodopsins.

The positions of BR Glu161 and Ala126 are relatively far from retinal (having the 9-th and 10-th largest coefficient parameters). To our knowledge, there are no previous studies focused on the colour-tuning effects of these residues. For the position of BR Glu161, larger red- and blue shifts are expected for valine and tyrosine. In fact, sensory rhodopsin I (SRI), which is a positive phototactic sensor, has a valine at this position and exhibits relatively longer absorption maxima (e.g., the SRI of *Halobacterium salinarum* (HsSRI): 587 nm; SRI of *Haloarcula vallismortis* (HvSRI): 545 nm). In contrast, a tyrosine is conserved among various channelrhodopsins (ChRs), which generally have short absorption wavelengths (e.g., the ChR1 of *Chlamydomonas reinhardtii* (CrChR1): 453 nm; ChR1 of *Dunaliella salina* (DChR1): 475 nm; ChR2 of *Proteomonas sulcata* (PsChR2): 444 nm). The results of ML analysis suggest the position of BR Glu161 is important for the colour tuning of these rhodopsins in nature. The position of BR Ala126 exhibited a large coefficient value for glutamic acid (10.5). Actually, *Gloeobacter* rhodopsin (GR), the outward H⁺ pump rhodopsin of cyanobacterium, *Gloeobacter violaceus* PCC 7421, has a glutamic acid at this position (GR Glu166), and the mutation of this residue exhibited a blue-shift of 1 to 22 nm (Supplementary Table 1). Thus, GR Glu166 works as an active residue for the colour tuning in GR.

These results imply the usefulness of ML analysis in identifying active residues located far from retinal, which are generally of less concern in experimental research on the colour tuning mechanism from a structural point of view. The effects on the absorption wavelength by the mutation of these residues have not yet been reported. However, we expect that they will be experimentally verified in the near future.

Toward Experimental Design. The fitted linear model parameters β_{ij} , $i = 1, \dots, M$, $j = 1, \dots, N$ can be also used as a guide for new functional protein design. For example, suppose that a researcher wants to construct a rhodopsin mutant, the absorption wavelength of which is as long as possible for optogenetics application. Note that positive/negative coefficient parameter values indicate that the amino-acid species at the residue have a red-shifting/blue-shifting effect, respectively, on the light-absorption properties of rhodopsin proteins. Consider a residue j at which there exists i_1 and i_2 such that $\beta_{i_1j} < \beta_{i_2j}$. If there exists a rhodopsin protein having the i_1 -th amino-acid species at the j -th residue, by replacing this species with the i_2 -th amino-acid species, the new protein

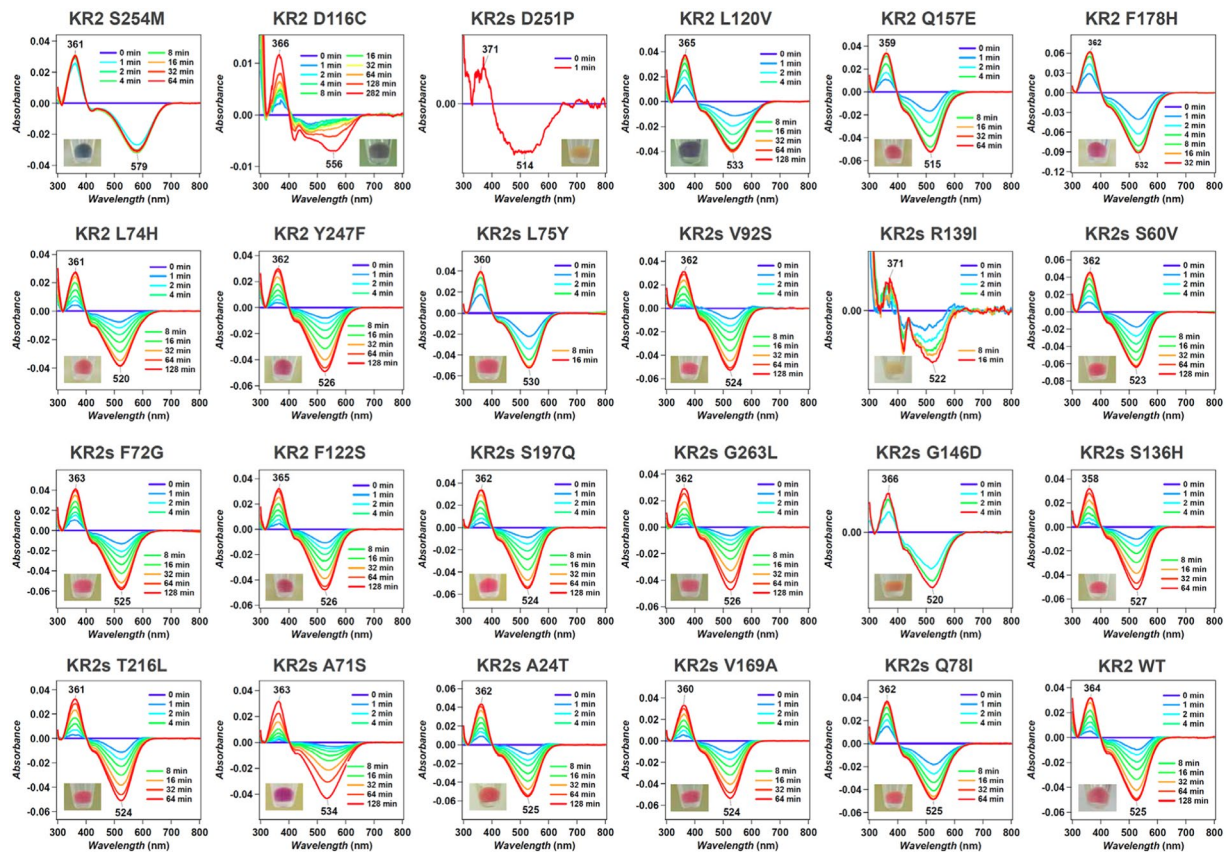


Figure 9. The difference spectra after and before the bleaching by hydroxylamine in 100 mM NaCl, 50 mM phosphate (pH 7.0), 2% DDM for KR2 mutants for which large red shift of the absorption was expected by ML. The times after the addition of hydroxylamine are shown in each panel.

is expected to have a longer wavelength than the original protein. This means that, the basic experimental design strategy for the above-mentioned researcher would be to replace the amino-acid species having a smaller coefficient parameter with that having a larger coefficient parameter. Although many other factors, such as protein stability and functionality, must be taken into account in new functional protein design, the above discussion suggests that the ML-based data-driven approach enables systematic design of experiments without relying on the intuition or heuristics of researchers.

In our result, the prediction error for the KR2 variants were surprisingly small (± 7.8 nm error on average) though the training dataset does not contain the KR2 variants (Supplementary Fig. 3 provides the same evaluation for the GR variants which also shows high prediction accuracy). This indicates that our model, which assumes an additive effect of the amino-acid sequence on the wavelength, provides useful estimation even for wildtypes (and their variants) which have not been investigated in the existing database. In order to estimate the precision of wavelength prediction by machine learning model, we actually constructed several KR2 mutants and measured their actual absorption wavelengths by hydroxylamine bleaching experiment (see Fig. 9). The errors between the predicted and the actual wavelengths are 8.04 nm in average, which is close to the average error 7.8 nm for target set. We believe that our result would be a significant case study for a data-driven analysis of functional proteins. We constructed the database of absorption wavelengths of microbial rhodopsins for the first time, and demonstrated that ML provides an accurate prediction model of the wavelength, which can be beneficial for optogenetic applications. Note that the machine-learning model employed in this paper is a simple approximation of complicated colour tuning rule of microbial rhodopsin. In particular, our machine-learning model only consider additive effect of each amino-acid. In fact, there is a trade-off between the database size and the model complexity. If we can increase the size of our database by collecting new experimental results in the future, we will consider more complicated model, e.g., which can incorporate combinatorial effects of multiple amino acids.

Methods

Construction of a dataset of amino-acid sequences and λ_{\max} s. For ML analysis, we constructed a database (Supplementary Table 1) composed of the amino-acid sequences and the previously and newly reported λ_{\max} s of microbial rhodopsins and their variants. Previously reported λ_{\max} s were collected from 102 reports (listed in Supplementary Information 2). Newly reported λ_{\max} s were experimentally determined in our group by the hydroxylamine bleaching method for *E. coli* membrane expressing rhodopsins³² or purified protein by Ni- or Co-NTA chromatography²⁶, as described previously. The method used to determine each rhodopsin is also listed in Supplementary Table 1.

Details of the ML method with group-wise sparsity regularization. Our data contains a larger number of variables (4,200 binary variables) than the number of instances (677 rhodopsin proteins). In this case, classical least-squares methods may cause over-fitting of the training data, which results in poor prediction accuracy for the target data. *Sparse modeling*^{17,18} is a standard approach to this problem setup so that only a small subset of coefficient parameters is automatically selected. In particular, we use a group-wise sparsity method¹⁹ to analyze the residue-wise effect on the absorption wavelength. Let $x_{i,j} \in \{0, 1\}$ be a binary variable that indicates the existence of the i -th amino-acid species in the j -th residue, where $i = 1, \dots, M$ and $j = 1, \dots, N$. Here, each $i = 1, \dots, M$ of $x_{i,j}$ corresponds to one of $M = 20$ amino-acid species.

We consider predicting the absorption wavelength based on a linear model:

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^M \sum_{j=1}^N \beta_{i,j} x_{i,j},$$

where β_0 and $\beta_{i,j}$ for $i = 1, \dots, M$ and $j = 1, \dots, N$ are parameters. Suppose that we have K pairs of an amino-acid sequence and its absorption wavelength $\{(\mathbf{x}^{(k)}, \lambda_{\max}^{(k)})\}_{k=1}^K$, where $\mathbf{x}^{(k)} \in \mathbb{R}^{MN}$ is the binned as a vector, and $\lambda_{\max}^{(k)} \in \mathbb{R}$ is the absorption wavelength of the k -th rhodopsin protein. The parameters are fitted by solving the following penalized least-squares problem:

$$\min_{\beta_0, \beta} \sum_{k=1}^K \left(\lambda_{\max}^{(k)} - \beta_0 - \sum_{i=1}^M \sum_{j=1}^N \beta_{i,j} x_{i,j}^{(k)} \right)^2 + \gamma \sum_{j=1}^N \sqrt{\sum_{i=1}^M \beta_{i,j}^2},$$

where $\gamma > 0$ is a tuning parameter. This formulation is called *group LASSO*¹⁹, in which the first term is the sum of the squared prediction errors, and the second term is the group-wise penalty for the parameters. For each residue $j = 1, \dots, N$, we define the $M = 20$ coefficient parameters $(\beta_{1,j}, \dots, \beta_{M,j})$ as a group. If the training set indicates that the choice of the amino-acid species at the j -th residue does not affect the colour tuning property, then the group-wise sparsity penalty forces all of the $M = 20$ parameters $(\beta_{1,j}, \dots, \beta_{M,j})$ to be exactly zero. We can easily identify a set of important residues for determining the absorption wavelength by this effect, called *group-wise sparsity*, because usually only a small subset of the residues have non-zero coefficient parameters. In our experiment, the parameter γ was objectively chosen by the cross-validation procedure within the training set.

Determination of λ_{\max} of new KR2 mutants. The λ_{\max} of new KR2 mutants shown in Fig. 9 was determined by a hydroxylamine (HA) bleaching experiment in 100 mM NaCl, 50 mM phosphate (pH 7.0), 2% DDM according to previous work³². The difference spectra between after and before the addition of 500 mM HA shows original absorption peak of KR2 mutant as a negative peak, and λ_{\max} was estimated from the peak position.

Code availability. Our program code of the group LASSO for wavelength prediction is available at <http://www-als.ics.nitech.ac.jp/~karasuyama/software/GrplassoSeq.zip>.

Data Availability

The database of the amino-acid sequences and their wavelengths is provided in Supplementary Table 1.

References

- Ernst, O. P. *et al.* Microbial and animal rhodopsins: structures, functions, and molecular mechanisms. *Chem. Rev.* **114**, 126–163 (2014).
- Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18**, 1213–1225 (2015).
- Blatz, P. E., Mohler, J. H. & Navangul, H. V. Anion-induced wavelength regulation of absorption maxima of schiff bases of retinal. *Biochem.* **11**, 848–855 (1972).
- Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
- Bogomolni, R. & Spudich, J. The photochemical reactions of bacterial sensory rhodopsin-I. flash photolysis study in the one microsecond to eight second time window. *Biophys. J.* **52**, 1071–1075 (1987).
- Lin, J. Y., Knutsen, P. M., Muller, A., Kleinfeld, D. & Tsien, R. Y. ReaChR: a red-shifted variant of channelrhodopsin enables deep transcranial optogenetic excitation. *Nat. Neurosci.* **16**, 1499–1508 (2013).
- Bejà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. Proteorhodopsin phototrophy in the ocean. *Nature* **411**, 786–789 (2001).
- Kim, S. Y., Waschuk, S. A., Brown, L. S. & Jung, K.-H. Screening and characterization of proteorhodopsin color-tuning mutations in *Escherichia coli* with endogenous retinal synthesis. *Biochim. Biophys. Acta* **1777**, 504–513 (2008).
- Engqvist, M. K. *et al.* Directed evolution of *Gloeobacter violaceus* rhodopsin spectral properties. *J. Mol. Biol.* **427**, 205–220 (2015).
- Nakanishi, K., Balogh-Nair, V., Arnaboldi, M., Tsujimoto, K. & Honig, B. An external point-charge model for bacteriorhodopsin to account for its purple color. *J. Am. Chem. Soc.* **102**, 7945–7947 (1980).
- Eichinger, M., Tavan, P., Hutter, J. & Parrinello, M. A hybrid method for solutes in complex solvents: Density functional theory combined with empirical force fields. *The J. Chem. Phys.* **110**, 10452–10467 (1999).
- Kloppmann, E., Becker, T. & Ullmann, G. M. Electrostatic potential at the retinal of three archaeal rhodopsins: Implications for their different absorption spectra. *Proteins: Struct. Funct. Bioinforma.* **61**, 953–965 (2005).
- Hoffmann, M. *et al.* Color tuning in rhodopsins: The mechanism for the spectral shift between bacteriorhodopsin and sensory rhodopsin II. *J. Am. Chem. Soc.* **128**, 10808–10818 (2006).
- Babitzki, G., Denschlag, R. & Tavan, P. Polarization effects stabilize bacteriorhodopsin's chromophore binding pocket: A molecular dynamics study. *The J. Phys. Chem. B* **113**, 10483–10495 (2009).
- Melaccio, F. *et al.* Toward automatic rhodopsin modeling as a tool for high-throughput computational photobiology. *J. Chem. Theory Comput.* **12**, 6020–6034 (2016).
- Ganapathy, S. *et al.* Retinal-based proton pumping in the near infrared. *J. Am. Chem. Soc.* **139**, 2338–2344 (2017).
- Hastie, T., Tibshirani, R. & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*. (CBC Press, 2015).
- Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal. Stat. Soc. B* **58**, 267–288 (1996).

19. Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *J. Royal Stat. Soc. Ser. B (Statistical Methodol.* **68**, 49–67 (2006).
20. Mogi, T., Marti, T. & Khorana, H. Structure-function studies on bacteriorhodopsin. IX. substitutions of tryptophan residues affect protein-retinal interactions in bacteriorhodopsin. *J. Biol. Chem.* **264**, 14197–14201 (1989).
21. Yamazaki, Y. *et al.* Interaction of the indole of tryptophan-182 with the 9-methyl group of the retinal in the L intermediate of bacteriorhodopsin. *Biochemistry* **34**, 577–582 (1995).
22. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
23. Sudo, Y. *et al.* A blue-shifted light-driven proton pump for neural silencing. *J. Biol. Chem.* **288**, 20624–20632 (2013).
24. Kato, H. E. *et al.* Atomistic design of microbial opsin-based blue-shifted optogenetics tools. *Nat. Commun.* **6**, 7177 (2015).
25. Béjà, O. & Lanyi, J. K. Nature's toolkit for microbial rhodopsin ion pumps. *Proc. Natl. Acad. Sci. USA* **111**, 6538–6539 (2014).
26. Inoue, K. *et al.* A light-driven sodium ion pump in marine bacteria. *Nat. Commun.* **4**, 1678 (2013).
27. Inoue, K., Konno, M., Abe-Yoshizumi, R. & Kandori, H. The role of the NDQ motif in sodium-pumping rhodopsins. *Angew. Chem. Int. Ed.* **54**, 11536–11539 (2015).
28. Shimono, K., Iwamoto, M., Sumi, M. & Kamo, N. Effects of three characteristic amino acid residues of pharaonis phoborhodopsin on the absorption maximum. *Photochem. Photobiol.* **72**, 141–145 (2000).
29. Man, D. *et al.* Diversification and spectral tuning in marine proteorhodopsins. *EMBO J.* **22**, 1725–1731 (2003).
30. Ozaki, Y., Kawashima, T., Abe-Yoshizumi, R. & Kandori, H. A color-determining amino acid residue of proteorhodopsin. *Biochemistry* **53**, 6032–6040 (2014).
31. 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).
32. Abe-Yoshizumi, R., Inoue, K., Kato, H. E., Nureki, O. & Kandori, H. Role of Asn112 in a light-driven sodium ion-pumping rhodopsin. *Biochemistry* **55**, 5790–5797 (2016).
33. Sato, H. *et al.* Specific lipid-protein interactions in a novel honeycomb lattice structure of bacteriorhodopsin. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 1251–1256 (1999).

Acknowledgements

We appreciate for the insightful discussion and data collection by Drs. R. Abe-Yoshizumi, M. Konno, Y. Kato, S. Ito, and Y. Inatsu. We appreciate fruitful advise by Dr. I. Fujiwara for ML application for the color tuning of rhodopsin. The present study was financially supported by grants from the Japanese Ministry of Education, Culture, Sports, Science and Technology to M.K. (16H06538 and 17H04694), K.I. (26708001, 26620005, and 17H03007), H.K. (25104009 and 15H02391), and I.T. (16H06538 and 17H00758), from JST PRESTO to M.K. (Grant Number JPMJPR15N2) and K.I. (Grant Numbers JPMJPR12A2 and JPMJPR15P2), from JST CREST to I.T. (Grant Numbers JPMJCR1302 and JPMJCR1502), from the RIKEN Center for Advanced Intelligence Project to I.T., and by the JST support program for starting up innovation-hub on materials research by information integration initiative to M.K. and I.T.

Author Contributions

M.K. analyzed the data by machine learning. K.I. constructed the database and interpreted the results. R.N. conducted hydroxylamine bleaching experiments. H.K. and I.T. designed the entire research study.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-33984-w>.

Competing Interests: The authors declare no competing interests.

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) 2018