Toward Machine Learning-based Data-driven Functional Protein Studies: Understanding Colour Tuning Rules and Predicting the Absorption Wavelengths of Microbial Rhodopsins

Masayuki Karasuyama[†]

Nagoya Institute of Technology / JST PRESTO /
National Institute for Materials Science

karasuyama@nitech.ac.jp

Keiichi Inoue[†]

Nagoya Institute of Technology / OptoBioTechnology Research Center /
JST PRESTO

inoue.keiichi@nitech.ac.jp

Hideki Kandori[*]

Nagoya Institute of Technology / OptoBioTechnology Research Center

kandori@nitech.ac.jp

Ichiro Takeuchi[*]

Nagoya Institute of Technology / RIKEN /
National Institute for Materials Science

takeuchi.ichiro@nitech.ac.jp

November 29, 2017

[†] Equally contributed

[*] Corresponding author

Abstract

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.

1 Introduction

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 species[1]. The light-dependent ion-transport function of microbial rhodopsin is widely used in optogenetics for optical control of neural activity in the brain network[2]. 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 solvents[3], the wavelengths of absorption maxima of retinal (λ_{max}s) in microbial rhodopsin range from 436 nm of channel-rhodopsin from *Tetraselmis striata* (*Ts*ChR)[4] to 587 nm of sensory rhodopsin I[5]. 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.

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 light[4]. 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 reported[7] and even applied to optogenetics[4], 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], 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 analogues[11]. 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 $\lambda_{\max}$ 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 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[12,13,14]. 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.

2 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). As reported in a previous study[15], 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.

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 $\boldsymbol{x} \in \{0,1\}^{MN}$. We consider a linear model for such MN-dimensional variables with an intercept parameter β_0 and MN coefficient parameters $\beta_{i,j}, i = 1, \ldots, M, j = 1, \ldots, N$ (see Fig. 4(b)). These $1 + MN$ parameters are fitted based on the training set so that the output of the model $f(\boldsymbol{x})$ can predict the absorption wavelength of the rhodopsin protein for which the amino-acid sequence is coded as $\boldsymbol{x}$. 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 $s_j := \sqrt{\sum_{i=1}^{M} \beta_{i,j}^2}, j = 1, \ldots, N$, 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 Figure 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 136 rhodopsin proteins in the target set (containing KR2 group rhodopsin proteins) were predicted. Figures 6(a) and 6(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 Figs. 6(c)

(KR2 N112E) and 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 histogram in Fig. 6(b) shows the distribution of the prediction errors in the KR2 group rhodopsin proteins in the target set.

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 $\beta_{i,j}$, $i = 1, \ldots, M, j = 1, \ldots, N$. Let $\boldsymbol{x}^{(\mathrm{KR2})} \in \{0,1\}^{MN}$ 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 $\boldsymbol{x}^{(\mathrm{Var})} \in \{0,1\}^{MN}$ is written as

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

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

$$\sum_{i=1}^{M} \beta_{i,j} \left(x_{i,j}^{(\mathrm{Var})} - x_{i,j}^{(\mathrm{KR2})} \right). \tag{1}$$

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.

3 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 Å(BR Thr89, Ala215, Gly122, Leu93, Asp85, Asp212, Met118, and Trp86 in the order of degree 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 BR[16]. 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[17,16,18]. The position of BR Thr89 is close to RSB (the distance between BR Thr89Cγ and the nitrogen atom of RSB is 3.4 Å). 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[19,20]. 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 $\lambda_{\max}$ and is consistent with previous experimental reports[18,19].

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 (*p*pR, which is also referred to as *pharaonis* sensory rhodopsin II) has a serine or threonine at this position[21]. The difference of coefficient parameter values is approximately 11.8, which is close to the reported $\lambda_{\max}$ shift of *p*pR T204A (8-nm red-shift)[21] and the BR homolog of *Haloquadratum walsbyi* (*Hw*BR) A223T (13-nm blue-shift)[22]. 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"[23]. 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 environment[23]. Ozaki et al. reported that mutations to valine or bulky residues (lysine, phenylalanine, tyrosine, and tryptophan) cause a large red-shift[24] of $\lambda_{\max}$. 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 $\lambda_{\max}$[25]. 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* (*Hs*SRI): 587 nm; SRI of *Haloarcula vallismortis* (*Hv*SRI): 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* (*Cr*ChR1): 453 nm; ChR1 of *Dunaliella salina* (*D*ChR1): 475 nm; ChR2 of *Proteomonas sulcata* (*Ps*ChR2): 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 $\beta_{i,j}$, $i = 1, \ldots, M, j = 1, \ldots, N$ 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 opt-genetics 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 $\beta_{i_1,j} < \beta_{i_2,j}$. 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 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.

4 Methods

Construction of a dataset of amino-acid sequences and $\lambda_{\max}$s For ML analysis, we constructed a database (Supplementary Table 1) composed of the amino-acid sequences and the previously and newly reported $\lambda_{\max}$s of microbial rhodopsins and their variants. Previously reported $\lambda_{\max}$s were collected from 102 reports (listed in Supplementary Information 2). Newly reported $\lambda_{\max}$s were experimentally determined in our group by the hydroxylamine bleaching method for *E. coli* membrane expressing rhodopsins[26] or purified protein by Ni- or Co-NTA chromatography[17], 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*[12,13] 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 method[14] 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 $i = 1, \ldots, M$ and $j = 1, \ldots, N$. Here, each $i = 1, \ldots, M$ of $x_{i,j}$ corresponds to one of $M = 20$ amino-acid species.

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

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

where β_0 and $\beta_{i,j}$ for $i = 1, \ldots, M$ and $j = 1, \ldots, N$ are parameters. Suppose that we have K pairs of an amino-acid sequence and its absorption wavelength $\{(\boldsymbol{x}^{(k)}, \lambda_{\max}^{(k)})\}_{k=1}^{K}$, where $\boldsymbol{x}^{(k)} \in \mathbb{R}^{MN}$ is the binary representation of the amino-acid sequence aligned 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:

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

where $\gamma > 0$ is a tuning parameter. This formulation is called *group LASSO*[14], 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 $j = 1, \ldots, N$, we define the $M = 20$ coefficient parameters $(\beta_{1,j}, \ldots, \beta_{M,j})$ 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 $(\beta_{1,j}, \ldots, \beta_{M,j})$ 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.

Code availability Our program code of the group LASSO for wavelength prediction is available at `http://...`[1]

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

[1] The site will be public after acceptance. The code is attached to our submission.

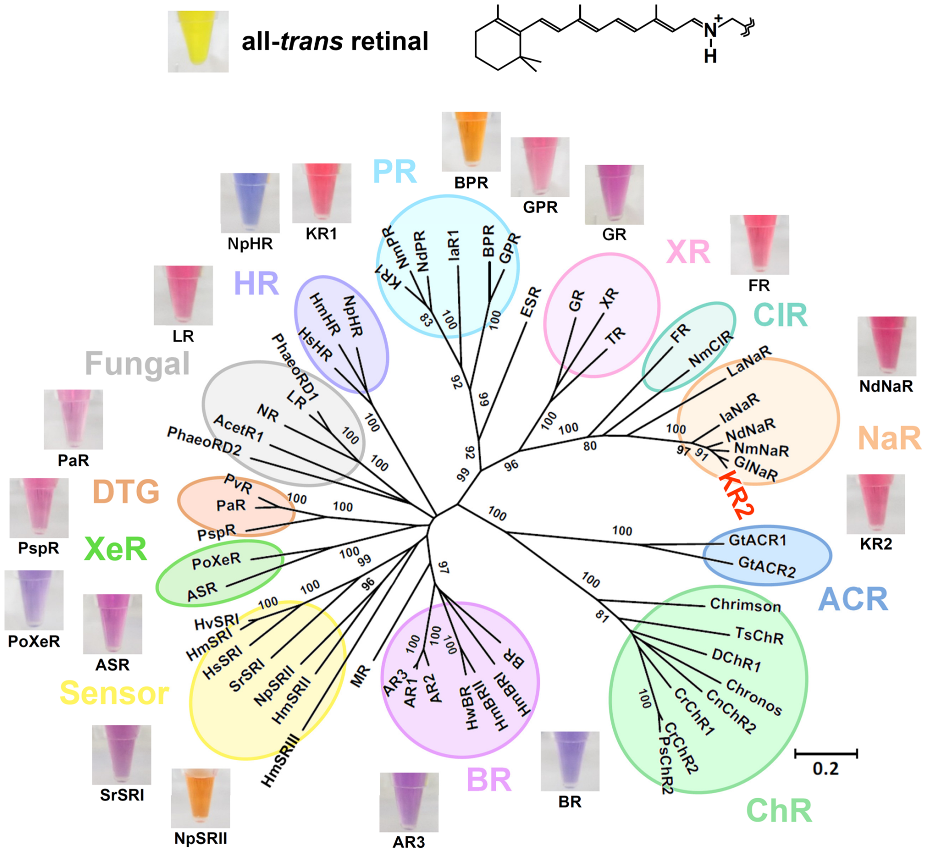


Figure 1: **The chemical structure of all-*trans* retinal (upper) and phylogenetic tree of microbial rhodopsins (lower).** The bootstrap values $> 80\%$ are shown for the corresponding branches. The photographs of the DMSO solution of all-*trans* retinal and detergent solubilized rhodopsins were aligned 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.

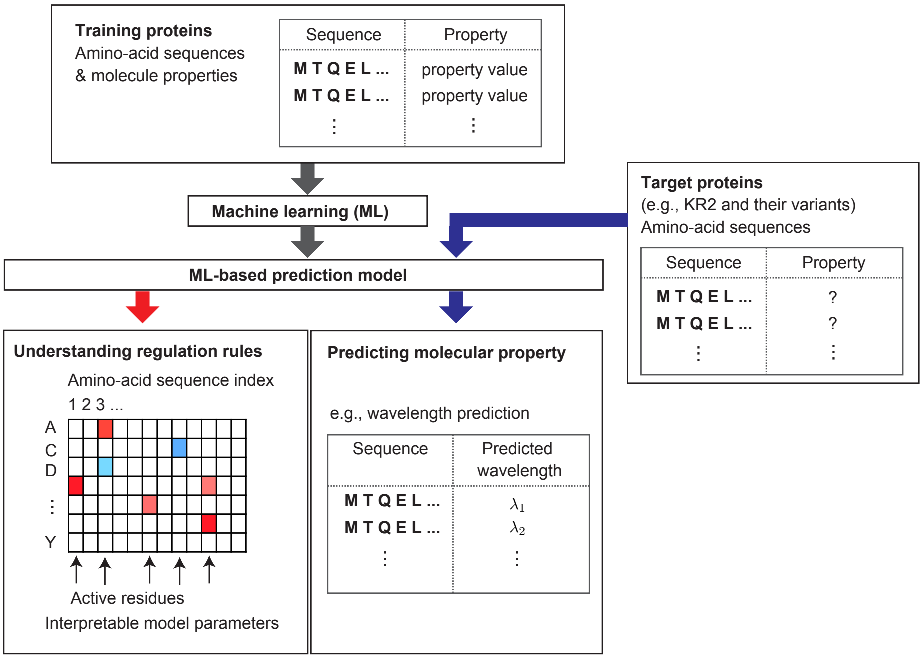
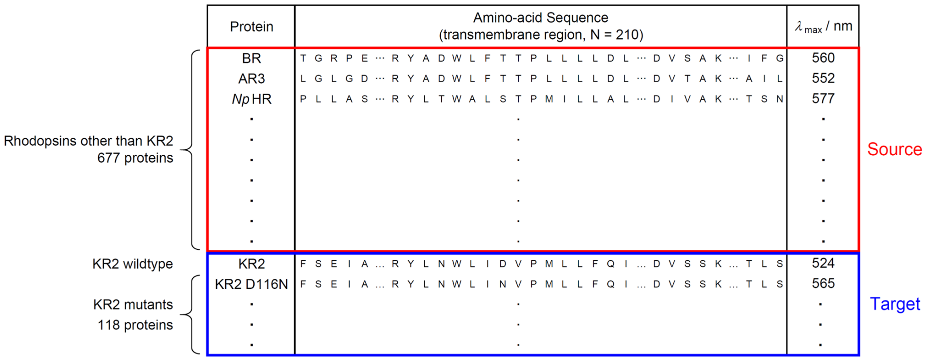


Figure 2: **An overview of the machine-learning-based (ML-based) data-driven approach introduced in the present paper for functional protein studies.** Using past experimental data, a *training protein set* containing pairs of amino-acid sequence and molecular properties is first constructed. Then, an 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.



Protein	Amino-acid Sequence (transmembrane region, N = 210)	λ_{max} / nm
BR	TGRPE···RYADWLFTTPLLLLDL···DVSAK···IFG	560
AR3	LGLGD···RYADWLFTTPLLLLDL···DVTAK···AIL	552
Np HR	PLLAS···RYLTWALSTPMILLAL···DIVAK···TSN	577
⋮	⋮	⋮
KR2	FSEIA...RYLNWLIDVPMLLFQI...DVSSK...TLS	524
KR2 D116N	FSEIA...RYLNWLINVPMLLFQI...DVSSK...TLS	565
⋮	⋮	⋮

Rhodopsins other than KR2
677 proteins
KR2 wildtype
KR2 mutants
118 proteins
Source
Target

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.

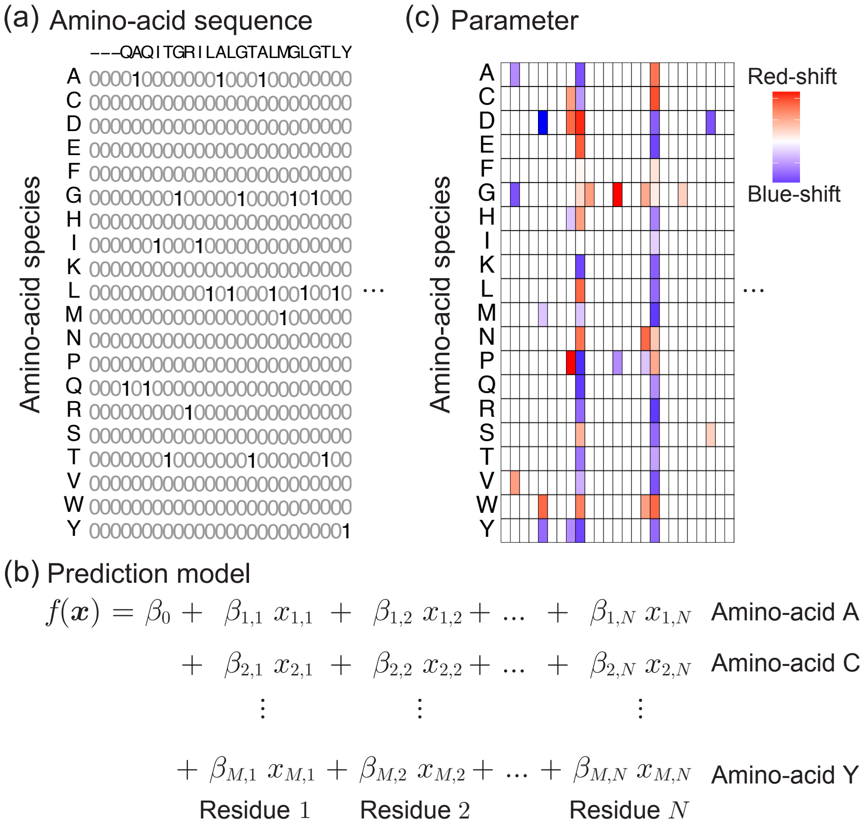

(a) Amino-acid sequence
---QAQITGRILALGTALMGLGTLY
Amino-acid species
A 0000100000001000100000000
C 0000000000000000000000000
D 0000000000000000000000000
E 0000000000000000000000000
F 0000000000000000000000000
G 0000000010000010000101000
H 0000000000000000000000000
I 0000001000100000000000000
K 0000000000000000000000000
L 0000000000010100010010010 ···
M 0000000000000000001000000
N 0000000000000000000000000
P 0000000000000000000000000
Q 0001010000000000000000000
R 0000000001000000000000000
S 0000000000000000000000000
T 0000000100000001000000100
V 0000000000000000000000000
W 0000000000000000000000000
Y 0000000000000000000000001
(c) Parameter
Amino-acid species
A C D E F G H I K L M N P Q R S T V W Y
Red-shift
Blue-shift
···
(b) Prediction model
$f(\boldsymbol{x}) = \beta_0 + \beta_{1,1}\, x_{1,1} + \beta_{1,2}\, x_{1,2} + \ldots + \beta_{1,N}\, x_{1,N}$ Amino-acid A
$+ \beta_{2,1}\, x_{2,1} + \beta_{2,2}\, x_{2,2} + \ldots + \beta_{2,N}\, x_{2,N}$ Amino-acid C
$\vdots \quad \vdots \quad \vdots$
$+ \beta_{M,1}\, x_{M,1} + \beta_{M,2}\, x_{M,2} + \ldots + \beta_{M,N}\, x_{M,N}$ Amino-acid Y
Residue 1 Residue 2 Residue N

Figure 4: (See next page for the caption)

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 $x_{i,j}, i = 1, \ldots, M, j = 1, \ldots, N$, we consider an MN-dimensional linear model. The linear model has an intercept parameter β_0 and MN coefficient parameters $\beta_{i,j}, i = 1, \ldots, M, j = 1, \ldots, N$. (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.

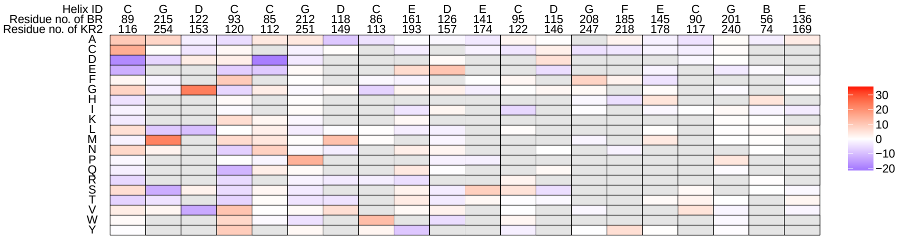


Figure 5: **Coefficient parameters of the fitted statistical model**. Coefficients for the top 20 active residues, where the activeness of each residue is defined as $s_j := \sqrt{\sum_{i=1}^{M} \beta_{i,j}^2}, j = 1, \ldots, N$. Here, red and blue indicate that the corresponding parameters are positive 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.

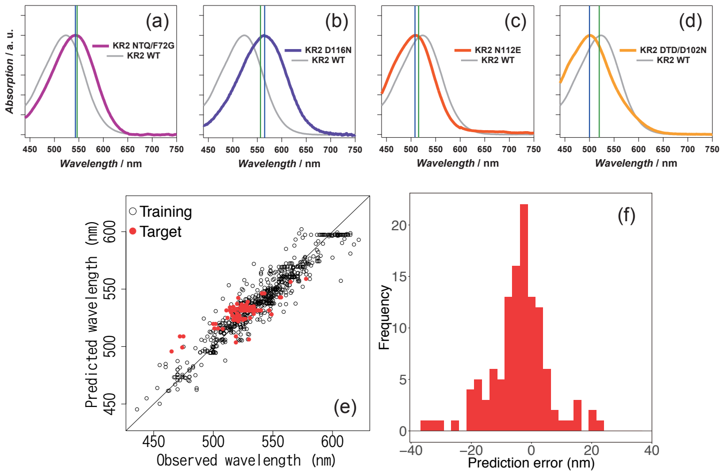
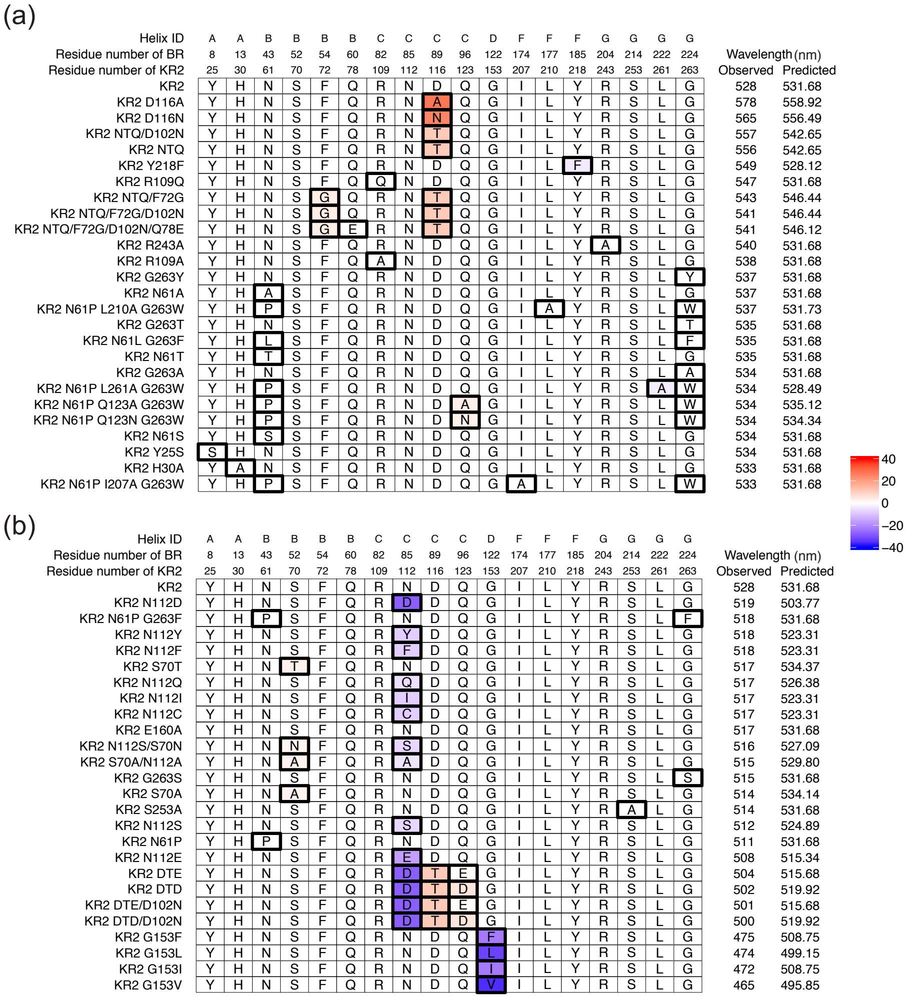


Figure 6: (See next page for the caption)

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.



(a)

Helix ID	A	A	B	B	B	B	C	C	C	C	D	F	F	F	G	G	G	G		
Residue number of BR	8	13	43	52	54	60	82	85	89	96	122	174	177	185	204	214	222	224	Wavelength (nm)	
Residue number of KR2	25	30	61	70	72	78	109	112	116	123	153	207	210	218	243	253	261	263	Observed	Predicted
KR2	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	528	531.68
KR2 D116A	Y	H	N	S	F	Q	R	N	A	Q	G	I	L	Y	R	S	L	G	578	558.92
KR2 D116N	Y	H	N	S	F	Q	R	N	N	Q	G	I	L	Y	R	S	L	G	565	556.49
KR2 NTQ/D102N	Y	H	N	S	F	Q	R	N	T	Q	G	I	L	Y	R	S	L	G	557	542.65
KR2 NTQ	Y	H	N	S	F	Q	R	N	T	Q	G	I	L	Y	R	S	L	G	556	542.65
KR2 Y218F	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	F	R	S	L	G	549	528.12
KR2 R109Q	Y	H	N	S	F	Q	Q	N	D	Q	G	I	L	Y	R	S	L	G	547	531.68
KR2 NTQ/F72G	Y	H	N	S	G	Q	R	N	T	Q	G	I	L	Y	R	S	L	G	543	546.44
KR2 NTQ/F72G/D102N	Y	H	N	S	G	Q	R	N	T	Q	G	I	L	Y	R	S	L	G	541	546.44
KR2 NTQ/F72G/D102N/Q78E	Y	H	N	S	G	E	R	N	T	Q	G	I	L	Y	R	S	L	G	541	546.12
KR2 R243A	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	A	S	L	G	540	531.68
KR2 R109A	Y	H	N	S	F	Q	A	N	D	Q	G	I	L	Y	R	S	L	G	538	531.68
KR2 G263Y	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	Y	537	531.68
KR2 N61A	Y	H	A	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	537	531.68
KR2 N61P L210A G263W	Y	H	P	S	F	Q	R	N	D	Q	G	I	A	Y	R	S	L	W	537	531.73
KR2 G263T	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	T	535	531.68
KR2 N61L G263F	Y	H	L	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	F	535	531.68
KR2 N61T	Y	H	T	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	535	531.68
KR2 G263A	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	A	534	531.68
KR2 N61P L261A G263W	Y	H	P	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	A	W	534	528.49
KR2 N61P Q123A G263W	Y	H	P	S	F	Q	R	N	D	A	G	I	L	Y	R	S	L	W	534	535.12
KR2 N61P Q123N G263W	Y	H	P	S	F	Q	R	N	D	N	G	I	L	Y	R	S	L	W	534	534.34
KR2 N61S	Y	H	S	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	534	531.68
KR2 Y25S	S	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	534	531.68
KR2 H30A	Y	A	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	533	531.68
KR2 N61P I207A G263W	Y	H	P	S	F	Q	R	N	D	Q	G	A	L	Y	R	S	L	W	533	531.68

(b)

Helix ID	A	A	B	B	B	B	C	C	C	C	D	F	F	F	G	G	G	G		
Residue number of BR	8	13	43	52	54	60	82	85	89	96	122	174	177	185	204	214	222	224	Wavelength (nm)	
Residue number of KR2	25	30	61	70	72	78	109	112	116	123	153	207	210	218	243	253	261	263	Observed	Predicted
KR2	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	528	531.68
KR2 N112D	Y	H	N	S	F	Q	R	D	D	Q	G	I	L	Y	R	S	L	G	519	503.77
KR2 N61P G263F	Y	H	P	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	F	518	531.68
KR2 N112Y	Y	H	N	S	F	Q	R	Y	D	Q	G	I	L	Y	R	S	L	G	518	523.31
KR2 N112F	Y	H	N	S	F	Q	R	F	D	Q	G	I	L	Y	R	S	L	G	518	523.31
KR2 S70T	Y	H	N	T	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	517	534.37
KR2 N112Q	Y	H	N	S	F	Q	R	Q	D	Q	G	I	L	Y	R	S	L	G	517	526.38
KR2 N112I	Y	H	N	S	F	Q	R	I	D	Q	G	I	L	Y	R	S	L	G	517	523.31
KR2 N112C	Y	H	N	S	F	Q	R	C	D	Q	G	I	L	Y	R	S	L	G	517	523.31
KR2 E160A	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	517	531.68
KR2 N112S/S70N	Y	H	N	N	F	Q	R	S	D	Q	G	I	L	Y	R	S	L	G	516	527.09
KR2 S70A/N112A	Y	H	N	A	F	Q	R	A	D	Q	G	I	L	Y	R	S	L	G	515	529.80
KR2 G263S	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	S	515	531.68
KR2 S70A	Y	H	N	A	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	514	534.14
KR2 S253A	Y	H	N	S	F	Q	R	N	D	Q	G	I	L	Y	R	A	L	G	514	531.68
KR2 N112S	Y	H	N	S	F	Q	R	S	D	Q	G	I	L	Y	R	S	L	G	512	524.89
KR2 N61P	Y	H	P	S	F	Q	R	N	D	Q	G	I	L	Y	R	S	L	G	511	531.68
KR2 N112E	Y	H	N	S	F	Q	R	E	D	Q	G	I	L	Y	R	S	L	G	508	515.34
KR2 DTE	Y	H	N	S	F	Q	R	D	T	E	G	I	L	Y	R	S	L	G	504	515.68
KR2 DTD	Y	H	N	S	F	Q	R	D	T	D	G	I	L	Y	R	S	L	G	502	519.92
KR2 DTE/D102N	Y	H	N	S	F	Q	R	D	T	E	G	I	L	Y	R	S	L	G	501	515.68
KR2 DTD/D102N	Y	H	N	S	F	Q	R	D	T	D	G	I	L	Y	R	S	L	G	500	519.92
KR2 G153F	Y	H	N	S	F	Q	R	N	D	Q	F	I	L	Y	R	S	L	G	475	508.75
KR2 G153L	Y	H	N	S	F	Q	R	N	D	Q	L	I	L	Y	R	S	L	G	474	499.15
KR2 G153I	Y	H	N	S	F	Q	R	N	D	Q	I	I	L	Y	R	S	L	G	472	508.75
KR2 G153V	Y	H	N	S	F	Q	R	N	D	Q	V	I	L	Y	R	S	L	G	465	495.85

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 $\sum_{i=1}^{M} \beta_{i,j} \left(x_{i,j}^{(\mathrm{Var})} - x_{i,j}^{(\mathrm{KR2})} \right)$, where $x_{i,j}^{(\mathrm{KR2})}$ and $x_{i,j}^{(\mathrm{Var})}$ are the binary representation the KR2 wildtype and a variant, respectively.

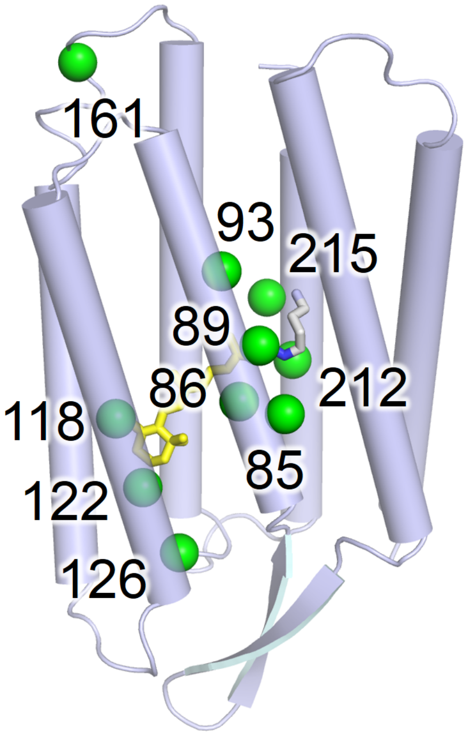


Figure 8: **Top 10 active residues identified by the fitted statistical model**. The positions of the active residues showing larger coefficient parameter values (green spheres) are mapped on the X-ray crystallographic structure of BR (blue, PDB code: 1BM1[27]) with their numbers in the case of BR.

Acknowledgements

We appreciate for the insightful discussion and data collection by Drs. R. Abe-Yoshizumi, M. Konno, Y. Kato, S. Ito, and Y. Inatsu. 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. H.K. and I.T. designed the entire research study.

References

[1] Ernst, O. P. *et al.* Microbial and animal rhodopsins: structures, functions, and molecular mechanisms. *Chem. Rev.* **114**, 126–163 (2014).

[2] Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18**, 1213–1225 (2015).

[3] Blatz, P. E., Mohler, J. H., & Navangul, H. V. Anion-induced wavelength regulation of absorption maxima of schiff bases of retinal. *Biochemistry* **11**, 848–855 (1972).

[4] Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).

[5] Bogomolni, R. & Spudich, J. The photochemical reactions of bacterial sensory rhodopsin-i. flash photolysis study in the one microsecond to eight second time window. *Biophysical Journal* **52**, 1071–1075 (1987).

[6] 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).

[7] Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M., & DeLong, E. F. Proteorhodopsin phototrophy in the ocean. *Nature* **411**, 786–789 (2001).

[8] 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).

[9] Engqvist, M. K. *et al.* Directed evolution of *Gloeobacter violaceus* rhodopsin spectral properties. *J. Mol. Biol.* **427**, 205–220 (2015).

[10] Melaccio, F. *et al.* Toward automatic rhodopsin modeling as a tool for high-throughput computational photobiology. *J. Chem. Theory Comput.* **12**, 6020–6034 (2016).

[11] Ganapathy, S. *et al.* Retinal-based proton pumping in the near infrared. *J. Am. Chem. Soc.* **139**, 2338–2344 (2017).

[12] Hastie, T., Tibshirani, R., & Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations.* CBC Press, (2015).

[13] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B* **58**, 267–288 (1996).

[14] Yuan, M. & Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67 (2006).

[15] Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).

[16] Béjà, O. & Lanyi, J. K. Nature's toolkit for microbial rhodopsin ion pumps. *Proc. Natl. Acad. Sci. USA* **111**, 6538–6539 (2014).

[17] Inoue, K. *et al.* A light-driven sodium ion pump in marine bacteria. *Nat. Commun.* **4**, 1678 (2013).

[18] 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).

[19] Kato, H. E. *et al.* Atomistic design of microbial opsin-based blue-shifted optogenetics tools. *Nat. Commun.* **6**, 7177 (2015).

[20] Inoue, K., Nomura, Y., & Kandori, H. Asymmetric functional conversion of eubacterial light-driven ion pumps. *J. Biol. Chem* **291**, 9883–9893 (2016).

[21] 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).

[22] Sudo, Y. *et al.* A blue-shifted light-driven proton pump for neural silencing. *J. Biol. Chem* **288**, 20624–20632 (2013).

[23] Man, D. *et al.* Diversification and spectral tuning in marine proteorhodopsins. *EMBO J.* **22**, 1725–1731 (2003).

[24] Ozaki, Y., Kawashima, T., Abe-Yoshizumi, R., & Kandori, H. A color-determining amino acid residue of proteorhodopsin. *Biochemistry* **53**, 6032–6040 (2014).

[25] 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).

[26] 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).

[27] 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).