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 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} s) 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.

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 energy¹⁰, 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 analogues¹¹. 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

61 reported in the literature and others of which are newly reported herein (see
62 Supplementary Table 1). Each entry of the database consists of the amino-acid
63 sequence and absorption wavelength $\lambda_{\max}$ of a rhodopsin. We introduce an ML
64 method for constructing a statistical model describing the relationship between
65 amino-acid sequences and absorption wavelengths. The goal of the present pa-
66 per is to demonstrate the effectiveness of ML-based data-driven approaches for
67 functional protein studies. By constructing a database based on past experi-
68 mental results and applying an ML method to the database, a statistical model
69 describing the relationship between amino-acid sequences and molecular prop-
70 erties can be constructed. In the context of microbial rhodopsin studies, we
71 illustrate the utility of such a statistical model by demonstrating that it can
72 be effectively used for understanding the colour tuning rules and predicting the
73 absorption wavelength (see Fig. 2).

74 We consider the following hypothetical scenario for the purpose of demon-
75 stration. The database is divided into two sets: a target protein set and a
76 training protein set. The target set contains KR2 wild-type rhodopsin and its
77 variants (which, in the present study, are assumed to be uninvestigated as of
78 yet), whereas the training set contains the remaining proteins in the database.
79 We constructed an ML model using only the proteins in the training set. The
80 constructed model was then applied to the proteins in the target set for pre-
81 dicting the absorption wavelengths of KR2 and its variants. This scenario is
82 interpreted as a hypothetical situation where a researcher is interested in pre-
83 dicting the absorption wavelengths of a new group of rhodopsin proteins based
84 on previously reported data on other groups of rhodopsin proteins.

85 Among the various available ML methods, we used a *group-wise sparse learn-*
86 *ing* approach^{12,13,14}. The advantages of group-wise sparse learning approaches
87 are not only predictability but also interpretability of the constructed models.
88 As we report later herein, by using a group-wise sparse learning approach, the
89 absorption wavelengths of KR2 and its variants could be predicted from their
90 amino-acid sequences with an average error of ± 7.8 nm. The residues affecting
91 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 $\lambda_{\max}$ 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 study¹⁵, 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

121 model was then applied to the rhodopsin proteins in the target set in order
122 to predict their absorption wavelengths. This scenario assumes a hypotheti-
123 cal situation in which a researcher is interested in investigating a new group
124 of rhodopsin proteins based on previously reported data on other groups of
125 rhodopsin proteins.

126 **Machine learning method** In order to handle amino-acid sequences in the
127 ML framework, we introduced a binary representation, as depicted in Fig. 4(a).
128 Let $M = 20$ be the number of different amino-acid species, and let $N = 210$ be
129 the number of residues considered herein. Then, an amino-acid sequence is rep-
130 resented by $M \times N = 4,200$ binary variables, which we denote as $\mathbf{x} \in \{0, 1\}^{MN}$.
131 We consider a linear model for such MN -dimensional variables with an intercept
132 parameter β_0 and MN coefficient parameters $\beta_{i,j}, i = 1, \dots, M, j = 1, \dots, N$
133 (see Fig. 4(b)). These $1 + MN$ parameters are fitted based on the training set
134 so that the output of the model $f(\mathbf{x})$ can predict the absorption wavelength of
135 the rhodopsin protein for which the amino-acid sequence is coded as $\mathbf{x}$. Since
136 this model has so many parameters, it is difficult to interpret the fitted model
137 if we simply use conventional methods such as the least-squares method. We
138 thus introduced the *group-wise sparsity mechanism* (See the Method section and
139 the Supplemental information for details). Using this mechanism, the fitted co-
140 efficient parameters $\beta_{i,j}$ have *residue-wise sparsity*. Here, $M = 20$ coefficient
141 parameters corresponding to the choice of an amino-acid species in each residue
142 is considered as a group. After we fitted the model, in many groups, all of the
143 M coefficient parameters become zero, indicating that the choice of an amino-
144 acid species in these residues does not affect the colour tuning property. On the
145 other hand, a small number of residues at which the coefficient parameters are
146 NOT zero are called *active residues*, i.e., the choice of the amino-acid species
147 in these residues is expected to play an important role in colour tuning. Figure
148 4(c) illustrates the fitted coefficient parameters using the group-wise-sparsity
149 mechanism. If a parameter $\beta_{i,j}$ is positive/negative, then the i -th amino-acid
150 species in the j -th residue has a red-shifting/blue-shifting effect on the light

151 absorption properties of rhodopsin proteins.

152 **Understanding colour tuning rules** By applying the above ML method
 153 to the training set containing pairs of the amino-acid sequence and absorption
 154 wavelength for 677 rhodopsin proteins, we fitted a linear model with $1 + MN =$
 155 4,201 parameters. A complete list of the fitted parameters is presented in Sup-
 156plementary Table 2. Figure 5 shows the fitted coefficient parameters at 20 active
 157 residues in decreasing order of $s_j := \sqrt{\sum_{i=1}^M \beta_{i,j}^2}$, $j = 1, \dots, N$, where the score
 158 s_j quantifies the *activeness* of the j -th residue. Here, red and blue indicate that
 159 the corresponding parameters are positive and negative, respectively, whereas
 160 grey indicates that the parameters were zero. In other words, red and blue sug-
 161gest that having the amino-acid species in the residue would have a red-shifting
 162 and a blue-shifting effect, respectively. The results in Supplementary Table 2
 163 and Fig. 5 can be interpreted as a comprehensive statistical description of the
 164 colour tuning rules of rhodopsin proteins based on previously investigated ex-
 165perimental results for 677 rhodopsin proteins (Supplementary Figure 1 shows
 166 the same results obtained using all 796 rhodopsin proteins, including those in
 167 the KR2 group).

168 **Predicting absorption wavelengths of KR2 rhodopsin and its variants**

169 Using the statistical model fitted based on the training set (containing all of the
 170 rhodopsin proteins except for the KR2 group), the absorption wavelengths of
 171 the 136 rhodopsin proteins in the target set (containing KR2 group rhodopsin
 172 proteins) were predicted. Figures 6(a) and 6(b) show examples of predicted
 173 (green lines) and observed (blue lines) wavelengths for red-shifted KR2 mu-
 174tants. For the KR2 NTQ/F72G mutant (Fig. 6(a)), the difference between
 175 the predicted (546.44 nm) and experimentally observed (543 nm) wavelengths
 176 is only 3.44 nm. In contrast, we observed a larger discrepancy (8.51 nm) for
 177 the predicted (556.49 nm) and experimentally observed (565 nm) wavelengths
 178 for KR2 D116N. This means that the precision of ML prediction differs for each
 179 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 $\beta_{i,j}$, $i = 1, \dots, M, j = 1, \dots, N$. 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

205 is $\beta_{i_2,j} - \beta_{i_1,j}$. Figure 7 shows a portion of the amino-acid sequences of KR2
 206 wildtype and its variants along with their observed and predicted absorption
 207 wavelengths. In Fig. 7, red and blue indicate red-shifting and blue-shifting ef-
 208 fects, respectively, in Eq. (1) estimated by the trained statistical model. Figure
 209 7(a) suggests that point mutation at BR residue number 89 would have red-
 210 shifting effects. On the other hand, Fig. 7(b) suggests that point mutation at
 211 BR residues 85 and 122 would have blue-shifting effects. These results indicate
 212 that the estimated colour-shifting effects are consistent with the actual observed
 213 wavelength shifts caused by the mutation.

214 3 Discussion

215 **Colour tuning rules in the estimated statistical models by ML** Ten
 216 residues showing the highest β -values were overlaid on the X-ray crystallo-
 217 graphic structure of BR (PDB code: 1BM1) (see Fig. 8). Eight of these
 218 residues are located around retinal within $< 5 \text{ \AA}$ (BR Thr89, Ala215, Gly122,
 219 Leu93, Asp85, Asp212, Met118, and Trp86 in the order of degree of activeness).
 220 Thr89 showed the highest degree of activeness. This is a member of the DTD-
 221 motif, which represents the type of functional determining three residues in the
 222 third transmembrane helix (helix-C) for each ion-pump rhodopsin. The DTD-
 223 motif is typical for the outward H^+ pump and is composed of Asp85, Thr89,
 224 and Asp96 for BR¹⁶. While this threonine is conserved among most microbial
 225 rhodopsins, it is replaced with an aspartate for sodium pump rhodopsin (NaR),
 226 which has the NDQ-motif rather than the DTD-motif^{17,16,18}. The position of
 227 BR Thr89 is close to RSB (the distance between BR Thr89C γ and the nitrogen
 228 atom of RSB is 3.4 $\text{\AA}$). The third and seventh active residues are BR Gly122
 229 and Met118, respectively. These residues are highly conserved among various
 230 microbial rhodopsins. Their mutation causes the rotation of the C6-C7 bond of
 231 retinal and the shortening of the π -electron conjugation between the β -ionone
 232 ring and the polyene chain^{19,20}. The largest coefficient parameters are obtained
 233 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 (ppR, which is also referred to as *pharaonis* sensory rhodopsin II) has a serine or threonine at this position²¹. The difference of coefficient parameter values is approximately 11.8, which is close to the reported $\lambda_{\max}$ shift of ppR T204A (8-nm red-shift)²¹ and the BR homolog of *Haloquadratum walsbyi* (HwBR) 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 environment²³. Ozaki et al. reported that mutations to valine or bulky residues (lysine, phenylalanine, tyrosine, and tryptophan) cause a large red-shift²⁴ 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

265 and 3.5 Å, respectively.)

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

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

293 These results imply the usefulness of ML analysis in identifying active residues
294 located far from retinal, which are generally of less concern in experimental re-
295 search on the colour tuning mechanism from a structural point of view. The

296 effects on the absorption wavelength by the mutation of these residues have not
297 yet been reported. However, we expect that they will be experimentally verified
298 in the near future.

299 **Toward Experimental Design** The fitted linear model parameters $\beta_{i,j}$,
300 $i = 1, \dots, M, j = 1, \dots, N$ can be also used as a guide for new functional
301 protein design. For example, suppose that a researcher wants to construct a
302 rhodopsin mutant, the absorption wavelength of which is as long as possible for
303 opt-genetics application. Note that positive/negative coefficient parameter val-
304 ues indicate that the amino-acid species at the residue have a red-shifting/blue-
305 shifting effect, respectively, on the light-absorption properties of rhodopsin pro-
306 teins. Consider a residue j at which there exists i_1 and i_2 such that $\beta_{i_1,j} < \beta_{i_2,j}$.
307 If there exists a rhodopsin protein having the i_1 -th amino-acid species at the
308 j -th residue, by replacing this species with the i_2 -th amino-acid species, the new
309 protein is expected to have a longer wavelength than the original protein. This
310 means that, the basic experimental design strategy for the above-mentioned re-
311 searcher would be to replace the amino-acid species having a smaller coefficient
312 parameter with that having a larger coefficient parameter. Although many other
313 factors, such as protein stability and functionality, must be taken into account in
314 new functional protein design, the above discussion suggests that the ML-based
315 data-driven approach enables systematic design of experiments without relying
316 on the intuition or heuristics of researchers.

317 4 Methods

318 **Construction of a dataset of amino-acid sequences and $\lambda_{\max}$ s** For ML
319 analysis, we constructed a database (Supplementary Table 1) composed of the
320 amino-acid sequences and the previously and newly reported $\lambda_{\max}$ s of microbial
321 rhodopsins and their variants. Previously reported $\lambda_{\max}$ s were collected from
322 102 reports (listed in Supplementary Information 2). Newly reported $\lambda_{\max}$ s
323 were experimentally determined in our group by the hydroxylamine bleaching

method for *E. coli* membrane expressing rhodopsins²⁶ or purified protein by Ni- or Co-NTA chromatography¹⁷, 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 method¹⁴ 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 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:

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

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 $j = 1, \dots, N$, 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.

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.

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

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)	$\lambda_{\max}$ / nm	
Rhodopsins other than KR2 677 proteins	BR	T G R P E ... R Y A D W L F T T P L L L L D L ... D V S A K ... I F G	560	Source
	AR3	L G L G D ... R Y A D W L F T T P L L L L D L ... D V T A K ... A I L	552	
	NpHR	P L L A S ... R Y L T W A L S T P M I L L A L ... D I V A K ... T S N	577	
	.	.	.	
	.	.	.	
	.	.	.	
KR2 wildtype KR2 mutants 118 proteins	KR2	F S E I A ... R Y L N W L I D V P M L L F Q I ... D V S S K ... T L S	524	Target
	KR2 D116N	F S E I A ... R Y L N W L I N V P M L L F Q I ... D V S S K ... T L S	565	
	.	.	.	
	.	.	.	

Figure 3: **Structure of the database used in the present study.** The database is composed of the sequences and $\lambda_{\max}$ 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.

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 $x_{i,j}, i = 1, \dots, M, j = 1, \dots, N$, 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: (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.

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 $\sum_{i=1}^M \beta_{i,j} \left(x_{i,j}^{(\text{Var})} - x_{i,j}^{(\text{KR2})} \right)$, where $x_{i,j}^{(\text{KR2})}$ and $x_{i,j}^{(\text{Var})}$ 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: 1BM1²⁷) with their numbers in the case of BR.

361 **Acknowledgements**

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

373 **Author contributions**

374 M.K. analyzed the data by machine learning. K.I. constructed the database and
 375 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] Bj, 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).

- 403 [11] Ganapathy, S. *et al.* Retinal-based proton pumping in the near infrared.
404 *J. Am. Chem. Soc.* **139**, 2338–2344 (2017).
- 405 [12] Hastie, T., Tibshirani, R., & Wainwright, M. *Statistical Learning with*
406 *Sparsity: The Lasso and Generalizations*. CBC Press, (2015).
- 407 [13] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Royal.*
408 *Statist. Soc. B* **58**, 267–288 (1996).
- 409 [14] Yuan, M. & Lin, Y. Model selection and estimation in regression with
410 grouped variables. *Journal of the Royal Statistical Society: Series B (Sta-*
411 *tistical Methodology)* **68**, 49–67 (2006).
- 412 [15] Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using
413 failed experiments. *Nature* **533**, 73–76 (2016).
- 414 [16] Béjà, O. & Lanyi, J. K. Nature’s toolkit for microbial rhodopsin ion pumps.
415 *Proc. Natl. Acad. Sci. USA* **111**, 6538–6539 (2014).
- 416 [17] Inoue, K. *et al.* A light-driven sodium ion pump in marine bacteria. *Nat.*
417 *Commun.* **4**, 1678 (2013).
- 418 [18] Inoue, K., Konno, M., Abe-Yoshizumi, R., & Kandori, H. The role of
419 the ndq motif in sodium-pumping rhodopsins. *Angew. Chem. Int. Ed.* **54**,
420 11536–11539 (2015).
- 421 [19] Kato, H. E. *et al.* Atomistic design of microbial opsin-based blue-shifted
422 optogenetics tools. *Nat. Commun.* **6**, 7177 (2015).
- 423 [20] Inoue, K., Nomura, Y., & Kandori, H. Asymmetric functional conversion of
424 eubacterial light-driven ion pumps. *J. Biol. Chem* **291**, 9883–9893 (2016).
- 425 [21] Shimono, K., Iwamoto, M., Sumi, M., & Kamo, N. Effects of three charac-
426 teristic amino acid residues of pharaonis phoborhodopsin on the absorption
427 maximum. *Photochem. Photobiol.* **72**, 141–145 (2000).

- 428 [22] Sudo, Y. *et al.* A blue-shifted light-driven proton pump for neural silencing.
429 *J. Biol. Chem* **288**, 20624–20632 (2013).
- 430 [23] Man, D. *et al.* Diversification and spectral tuning in marine prote-
431 orhodopsins. *EMBO J.* **22**, 1725–1731 (2003).
- 432 [24] Ozaki, Y., Kawashima, T., Abe-Yoshizumi, R., & Kandori, H. A color-
433 determining amino acid residue of proteorhodopsin. *Biochemistry* **53**, 6032–
434 6040 (2014).
- 435 [25] Fujimoto, K., Hayashi, S., Hasegawa, J. Y., & Nakatsuji, H. Theoretical
436 studies on the color-tuning mechanism in retinal proteins. *J. Chem. Theory*
437 *Comput.* **3**, 605–618 (2007).
- 438 [26] Abe-Yoshizumi, R., Inoue, K., Kato, H. E., Nureki, O., & Kandori, H. Role
439 of asn112 in a light-driven sodium ion-pumping rhodopsin. *Biochemistry*
440 **55**, 5790–5797 (2016).
- 441 [27] Sato, H. *et al.* Specific lipid-protein interactions in a novel honeycomb
442 lattice structure of bacteriorhodopsin. *Acta Crystallogr. D Biol. Crystallogr*
443 **55**, 1251–1256 (1999).