Automatic Rhodopsin Modeling with Multiple Protonation Microstates

Gustavo Cárdenas,[†] Vincent Ledentu,[†] Miquel Huix-Rotllant,[†] Massimo Olivucci,[‡,¶] and Nicolas Ferré[*,†]

[†]*Aix-Marseille Univ, CNRS, ICR, Marseille, 13013, France*

[‡]*Department of Chemistry, Bowling Green State University, Bowling Green, Ohio 43403, United States*

[¶]*Dipartimento di Biotecnologie, Chimica e Farmacia, Università degli Studi di Siena, Siena, 53100, Italy*

E-mail: nicolas.ferre@univ-amu.fr

Abstract

Automatic Rhodopsin Modeling (ARM) is a simulation protocol providing QM/MM models of rhodopsins, capable of reproducing experimental electronic absorption and emission trends. Currently, ARM is restricted to a single protonation microstate for each rhodopsin model. Herein, we incorporate an extension of the minimal electrostatic model (MEM) into the ARM protocol to account for all relevant protonation microstates at a given pH. The new ARM+MEM protocol determines the most important microstates contributing to the description of the absorption spectrum. As a test case, we have applied this methodology to simulate the pH-dependent absorption spectrum of a toy model, showing that the single microstate picture breaks down at certain pH values. Subsequently, we applied ARM+MEM to *Anabaena* Sensory Rhodopsin, confirming an improved description of its absorption spectrum when the titration of several key residues is considered.

1 Introduction

Rhodopsins are a class of photoresponsive transmembrane proteins that are involved in different biological functions, such as vision, ion-gating, and ion-pumping, among others.[1–5] Despite their wide range of functions, rhodopsins share a common topology consisting of seven α helices enclosing a retinal moiety which is covalently bound *via* a Schiff base linkage with a lysine amino acid.[6,7] For the most part, rhodopsin activity is triggered by the absorption of visible light by the retinal protonated Schiff base (rPSB) and its subsequent photoisomerization.[4,8,9] The functionalities of rhodopsins depend on different factors such as the nature of their protein sequence, the arrangement of the side chains, the structure of the interfacial regions, among others.[4] In particular, regarding the nature of the protein sequence, the variations in the absorption or emission properties of the protein due to variations in the amino acid sequence is a constant object of study[10–17] since, remarkably, the absorption maximum wavelength ($\lambda_{\max}$), which is related to the energy required to photo-

activate rhodopsins, varies between 420 and 644 nm.[18,19] The study of the λ_{max} dependence on the amino acid sequence is not limited to the field of photobiology itself,[17,20–26] but is also important for the design of artificial molecular devices[27–29] and in the field of optogenetics,[30–36] whereby selected wild type or mutant microbial rhodopsins are expressed in neurons to trigger, silence or monitor their activity using specific light wavelengths. Another variable that impacts λ_{max} is the pH. Microscopically, the change of protonation state of aspartic (ASP)/glutamic (GLU) acid, histidine (HIS), lysine (LYS), arginine (ARG), ... which have titratable side chains (in what follows they will be referred to as titratable amino acids, for simplicity), modifies the electrostatic potential experienced by the retinal chromophore and can, thus, tune its light absorption properties.[37–39]

The study of the dependence of the rhodopsin function/property on its structure has profited enormously from the usage of computational methodologies. In fact, the construction of suitable computational models of the protein would, in principle, assist the experimental search for rhodopsin mutants with wanted properties. At the same, such models provide insights into the structure-property relationship at an atomistic level[40–42] with unmatched detail.

Building an accurate molecular model for any biomolecule is always based on a compromise between its inherent complexity (*i.e.*, the number of input parameters) and the typical timescale of the property of interest. When one is interested in spectroscopic or reactive processes occurring in photoactive proteins like rhodopsins, it is then possible to resort to the so-called QM/MM[43] partitioning of the molecular system: the chromophore is treated at the quantum mechanical (QM) level while its surroundings are modeled using molecular mechanics (MM). On the one hand, the **accuracy** of QM/MM models depends on the choice of the QM and MM methods, on their interactions, on the treatment of the frontier between the two subsystems, on the choice of the initial structure, etc.[44–48] For example, the interaction between the QM and the MM regions can be treated by means of electrostatic[49–51] or polarizable[52–54] embedding schemes: whereas the latter considers the mutual

polarization between the QM and the MM regions, the former is the most widely employed for its excellent compromise between accuracy and computational cost. On the other hand, the **reproducibility** of such QM/MM models depends on the definition of a well-established workflow.

The Automatic Rhodopsin Modeling (ARM) simulation protocol[55,56] aims at automatizing the construction of QM/MM models for rhodopsin proteins. In practice, ARM only requires a crystallographic structure and the pH value as input. Following a well-established workflow (Figure 1) involving a minimal user input, ARM can produce absorption or emission $\lambda_{\max}$ values with an established accuracy of ca. 3.0 kcal/mol (in terms of the mean absolute error - MAE).[55] However, in some occasions, the deviation from experimental data is much larger (*e.g.*, 20.7 kcal/mol for a particular mutant of channelrhodopsin-2, 15.5 kcal/mol for *Krokinobacter eikastus* rhodopsin 2).[55] In that case, the user has to make an educated guess of the origin of this discrepancy and modify the model accordingly. Often, it appears that a wrong selection of the protein protonation microstate (defined as the ensemble of the titratable amino acid protonation states) is responsible for the large deviation from the experimental data.[55] The automatic determination of this particular protonation microstate (MS) is based on one parameter in input, the pH value, and on a predicted pK_a value for each titratable amino acid. In ARM, these pK_a values are obtained using the PROPKA[57,58] software, a fast and empirical pK_a prediction tool. Such prediction may occasionally fail, especially when interactions (*i.e.*, correlations) between titrated residues are strong, *e.g.*, when they are spatially close to part of the same hydrogen bond network.

Assuming that pK_a predictions are qualitatively correct, we insist here that the selection of a single protonation microstate may not suffice. Actually, when the system features a very large number of protonation microstates (at least 2^N where N is the number of ASP, GLU, HIS, LYS, and ARG amino acids), it is virtually impossible to decide if one and only one protonation microstate is sufficient for building an effective model capable of reproducing the target property.

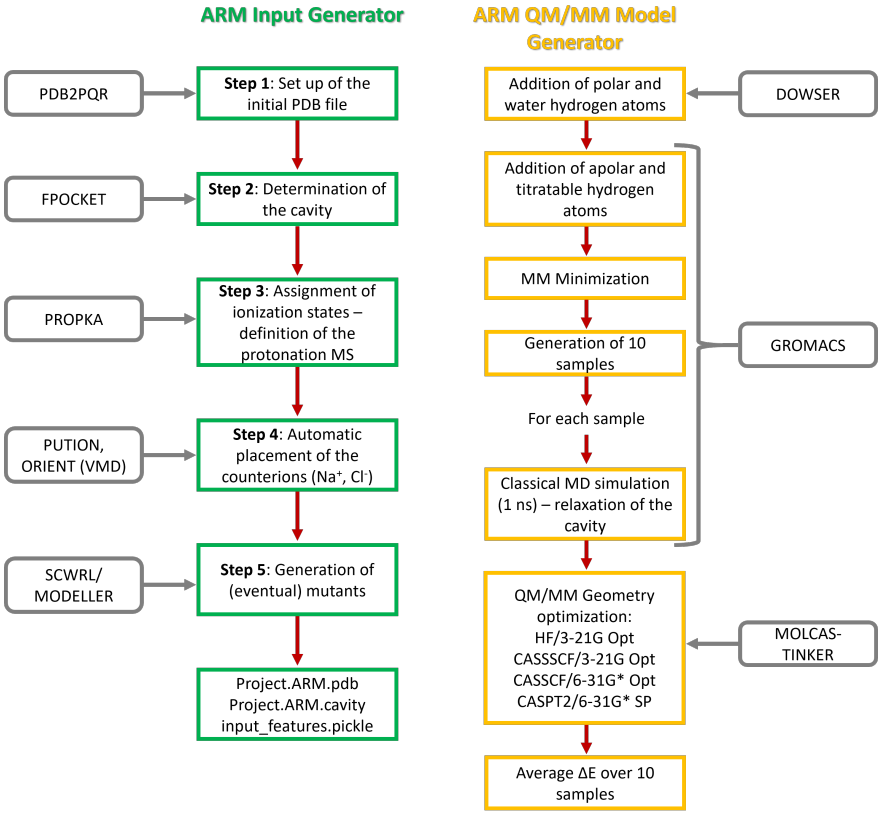

Figure 1: Schematic representation of the ARM protocol subdivided into two parts (each handled by a different driver): the automatic generation of the (PDB and cavity) input files (green, left) and the QM/MM geometry optimization of the 10 replicas that will constitute the final model (orange, right). The `Project.ARM.pdb` and the `Project.ARM.cavity` files produced by the input generator (left) are the files required by the QM/MM model generator (right).

While very expensive yet accurate methods, such as Constant pH Molecular Dynamics (CpHMD),[59] can achieve an efficient sampling of both the motion of the particles and their protonation microstates simultaneously, some of us proposed a cheap and straightforward minimal electrostatic method (MEM)[39] able to quickly screen the protonation microstate space to determine which amino acid (de)protonations may significantly modify the $\lambda_{\max}$ of photoactive proteins, and in what manner. MEM can be used to fit experimental titration curves such as $\lambda_{\max} = f(\text{pH})$, ultimately determining the pK_a values for the most important amino acids. When experimental data are not available, MEM can be used with predicted pK_a values to identify the important amino acid whose titration would modify $\lambda_{\max}$.

Based on the above-mentioned capabilities of the ARM and the MEM methodologies, in this article, we introduce an improved ARM protocol, quoted as ARM+MEM in the following, in which the MEM methodology has been incorporated to account for the contributions to the overall $\lambda_{\max}$ due to different protonation microstates at a given pH. After having recalled the most important features of both the ARM protocol and the MEM methodology – for which slight modifications with respect to the original presentation have been incorporated, we hereafter present how MEM can be easily integrated into the ARM workflow. Moreover, we propose a simple numerical approach for reconstructing the pH-dependent absorption spectrum. Then, we will illustrate how the integrated ARM+MEM scheme works in the case of a toy model made of the retinal chromophore and 3 titrated amino acids. Finally, we apply ARM+MEM to the *Anabaena* Sensory Rhodopsin (ASR) which is known to feature both $\lambda_{\max}$ redshifting and blueshifting pH effects.[38]

2 Methods

2.1 Overview of the ARM protocol

The ARM protocol aims at automatizing the generation of QM/MM models for rhodopsin proteins designed to qualitatively reproduce $\lambda_{\max}$ changes. The protocol has been thoroughly

described in the literature;[55,56] In what follows, we will only provide a brief overview of the methodology. Figure 1 shows a scheme summarizing the ARM protocol, which is subdivided into two parts: the ARM input generator (Figure 1, green), which generates a formatted PDB file containing the structure of the rhodopsin of interest with the appropriate protonation states for all of its titratable residues and the necessary counterions, and the QM/MM model generator, which partially relaxes the rhodopsin structure before calculating the desired $\lambda_{\max}$, as well as more complex properties.

The input generator is handled by the `a_arm_input_generator.py` driver, which is executed in the command line by the user. It consists of a sequence of four steps (five in the case of mutants) which are executed sequentially to generate the PDB input file necessary to execute the QM/MM geometry optimizations. The execution occurs interactively along this sequence of steps (Figure 1, left), in each of which the user intervention is required to either confirm the selection of a specific parameter by the protocol (*e.g.*, the protonation state) or to modify and/or customize such selection. Thus, a fully "automated" execution of the protocol would correspond to confirming all the parameter values selected by default by the protocol, but customization may (and in some cases should) be performed at this stage.[55]

In Step 1, the program is fed with the crystallographic structure (in PDB format) of the rhodopsin of interest. This step aims at cleaning the structure from potential lipids, detecting the chromophore residue, and detecting the potential main and secondary counterions of the rPSB based on some geometric parameters,[55] and adding missing heavy atoms of chain residues employing the PDB2PQR[60] software.

Step 2 consists of the determination of the chromophore cavity, that is, the selection of the residues present in the protein pocket hosting the chromophore which will be allowed to relax during the classical MD and the QM/MM computations. This selection is performed automatically by using the FPOCKET[61] software.

Step 3 focuses on the determination of the charges of the residues considered to be titrated along the protein chain (ASP, GLU, HIS, ARG, LYS), and is of crucial importance

since it characterizes the **single** protonation microstate that will be employed to generate the QM/MM model. In the current version,[55] it is assumed that the protonation state (and thus, the charge) of each titratable residue depends on its pK_a, which in turn, depends on the hydrogen bonds, desolvation effect, and Coulomb interactions that the residue of interest undergoes, as empirically determined using the PROPKA package.[57,58] Afterward, the state of the titratable residue is determined from the side-chain (de)protonation equilibrium using the Henderson-Hasselbach[62] equation:

$$\mathrm{pH} = \mathrm{p}K_a^{Calc} + \log \frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \tag{1}$$

where $\mathrm{p}K_a^{Calc}$ is the pK_a calculated by PROPKA, and [A^-] and [HA] are the concentrations of deprotonated and protonated species, respectively, at a given pH. In particular, the protonation state of each titrated residue is deduced from the Equation (1) using a set of approximated rules:[63,64]

$$\lceil Q^- \rceil = \frac{-1}{1 + 10^{-(\mathrm{pH} - \mathrm{p}K_a^{Calc})}} \tag{2}$$

for ASP and GLU, and

$$\lceil Q^+ \rceil = \frac{+1}{1 + 10^{-(\mathrm{pH} - \mathrm{p}K_a^{Calc})}} \tag{3}$$

for HIS, LYS, and ARG. $\lceil Q^- \rceil$ and $\lceil Q^+ \rceil$ are "rounded half to even" integers. Finally, the following criterion is used to assign the protonation state of the residue of interest:

$$\text{protonation state} = \begin{cases} \text{ASP, GLU,} & \text{if } \lceil Q^- \rceil = -1 \\ \text{ASH, GLH,} & \text{if } \lceil Q^- \rceil \neq -1 \\ \text{ARG, LYS, HIS,} & \text{if } \lceil Q^+ \rceil = +1 \\ \text{ARN, LYD, HID/HIE,} & \text{if } \lceil Q^+ \rceil \neq +1 \end{cases} \tag{4}$$

In the above notations, ASH (ASP) stands for protonated aspartic acid (deprotonated aspartate); GLH (GLU) stands for protonated glutamic acid (deprotonated glutamate); LYS

(LYD) stands for protonated lysine (deprotonated lysine); ARG (ARN) stands for protonated arginine (deprotonated arginine); HIS (HID, HIE) stands for protonated histidine (deprotonated histidine with the remaining proton bonded to nitrogen, respectively N_δ and N_ϵ).

In step 4, the ARM protocol determines the total charges present at the intracellular (IS) and the extracellular (OS) surfaces of the rhodopsin protein and proceeds to neutralize them by positioning suitable counterions (Na^+ or Cl^-) on both surfaces. The IS and OS are determined by first centering the XYZ coordinates at the protein center of mass, and then rotating the protein to align it with the z-axis by using the ORIENT package of VMD.[65] It should be emphasized that more accurate methods are present in the literature to orient a protein inside a lipid membrane, for example those presented by the Orientation of Proteins in Membranes (OPM) database.[66–68] However, the simple method adopted in the ARM protocol serves the main purpose of determining which are the two sides of the protein exposed to the aqueous medium, so as to neutralize them separately, rather than accurately determine the orientation of the protein inside a lipid membrane. The Na^+ or Cl^- positions are then automatically determined by the PUTION module, described in a previous release.[55,69] In the end, the driver generates a PDB file containing the rhodopsin structure with the selected residue protonation states and necessary counterions, a "cavity" file containing a list of residues (to be relaxed) present in the cavity hosting the chromophore, and a python `pickle` file that keeps track of all the parameters set in the previous steps for the current project (see Figure 1 left for reference). Notice that here a project designates a specific rhodopsin for which to generate a model, hence the "Project" prefix in the generated files in Figure 1. After step 4, the user has the necessary PDB and cavity files for the subsequent MD and QM/MM computations. In case the user is interested in studying not only wild-type rhodopsins but also some point mutations, the package has an extra step in which it ultimately generates one project folder for each mutant of interest (step 5 on Figure 1 left).[55]

The driver `a_arm_qmmm_generator_driver.py` handles the generation of the QM/MM

rhodopsin models (Figure 1 right), a script executed in the command line, which exclusively requires (in its default execution mode) as input files the PDB and the cavity files mentioned above. This part of the protocol (termed ARM QM/MM) is described in detail elsewhere.[55,69] In what follows, we summarize the most important steps of the protocol (see Figure 1, right and section 3). At first, the PDB input structure needs to be protonated since, in most cases, the initial PDB files stem from X-ray crystallography experiments. This is performed by the DOWSER[70] and the GROMACS[71] software. Afterwards, an MM minimization and 10 MD simulations of 1 ns are performed (on 10 samples stemming from using different initial conditions), in which both the cavity and the retinal chromophore are relaxed. Subsequently, a sequence of QM/MM (SA-CASSCF[72–74]/AMBER[75]) geometry optimizations are performed on these 10 structures, in which both the chromophore - described quantum mechanically - and the cavity - described by the AMBER force field - are optimized. The final step consists of a single point (SP) complete active space second-order perturbation theory (CASPT2)[76–78] calculation on each geometry to obtain the excitation energies and oscillator strengths.

It should be emphasized that although the standard/default ARM protocol[55,69] automatically considers 10 samples to obtain the $\lambda_{\max}$, the software does allow for the possibility to employ any user-defined number of geometries. However, in the present work, we have adopted the default ARM protocol to assess the effect of incorporating the minimal electrostatic model on a fully automated application of the ARM protocol. As a final remark, in the ARM protocol only the rPSB and the residues composing the cavity hosting the chromophore are relaxed so that for the most part the structure of the protein backbone corresponds to the crystallographic structure. This is an approximation that nonetheless has provided useful models for most of the rhodopsins modeled in the past[24,25,55,69] at the corresponding crystallographic pH.

2.2 Extension of the Minimal Electrostatic Model

Using a single rigid rhodopsin (or any other protein incorporating a chromophore) structure, the simple MEM approach assumes that $\lambda_{\max}$, or more precisely $\Delta E_{I\to J}$, the energy difference between electronic states I and J, is tuned by a pH-dependent weighted ensemble of N amino acid (*i.e.*, residue) (de)protonation probabilities (x_i with $i = 1, \ldots, N$), each of them affecting the electrostatic interaction with the chromophore charge distribution in its initial and final electronic states ($\Delta E^i_{I\to J}$),

$$\Delta E_{I\to J}(\mathrm{pH}) = \Delta E^0_{I\to J} + \sum_i^N x_i(\mathrm{pH})\Delta E^i_{I\to J} \tag{5}$$

In the original MEM derivation, $\Delta E^0_{I\to J}$ was arbitrarily chosen as the vertical transition energy of the fully protonated microstate (*i.e.* without any deprotonated amino acid). As a consequence, x_i could only be connected to a deprotonation event, and eventually, $\Delta E^0_{I\to J}$ was fitted to experimental data.

Hereafter, we present a modified version of the MEM approach in which $\Delta E^0_{I\to J}$ can refer to any reference microstate, for instance, the one coming out of the ARM protocol. Let's start by denoting U_I the total energy of an isolated chromophore in its electronic state I. When the chromophore (composed of N_a atom centers, each carrying an atomic charge Q^a_I with coordinates $\mathbf{R}_a$) electrostatically interacts with N point charges q_i with coordinates $\mathbf{r}_i$, which are located sufficiently far from it, its energy can be expressed as:

$$E^p_I = U_I + \sum_{i=1}^{N}\sum_{j=a}^{N_a} k^i_{\mathrm{eff}} \frac{q^p_i Q^a_I}{|\mathbf{r}_i - \mathbf{R}_a|} \tag{6}$$

in which k^i_{eff} is an effective screening factor already presented in the original MEM article.[39] The superscript p refers to a given protonation microstate. Indeed, the chromophore environment is composed of titratable residues, each of them carrying a single point charge, equal to -1 or 0 or +1 atomic unit, depending on the considered microstate p. U_I being the

energy of the bare chromophore, chromophore-environment mutual polarization effects are ignored in Equation (6).

The vertical transition energy between electronic states I and J, when the environment is in its protonation microstate p, is expressed as:

$$\Delta E^p_{I\to J} = \Delta U_{I\to J} + \sum_{i=1}^{N} q_i^p k^i_{\text{eff}} \sum_{a=1}^{N_a} \frac{Q^a_J - Q^a_I}{|\mathbf{r}_i - \mathbf{R}_a|} \tag{7}$$

where $\Delta U_{I\to J} = U_J - U_I$ is independent of the environment protonation microstate p.

We suppose that $\Delta E^p_{I\to J}$ is already known for a particular protonation microstate (for instance, the one selected by ARM), the reference one which we label with $p = 0$. Using the obvious equality $\Delta E^p_{I\to J} = \Delta E^0_{I\to J} + (\Delta E^p_{I\to J} - \Delta E^0_{I\to J})$, the transition energy can now be rewritten:

$$\begin{aligned}\Delta E^p_{I\to J} &= \Delta E^0_{I\to J} + \sum_{i=1}^{N} (q_i^p - q_i^0) k^i_{\text{eff}} \sum_{a=1}^{N_a} \frac{Q^a_J - Q^a_I}{|\mathbf{r}_i - \mathbf{R}_a|} \\ &= \Delta E^0_{I\to J} + \sum_{i=1}^{N} \alpha_{ip} \Delta\phi^i_{I\to J}\end{aligned} \tag{8}$$

α_{ip} indicates the variation of the charge for residue i when the protonation microstate changes from the reference one to the considered one p. Accordingly, $\alpha_{ip} = q_i^p - q_i^0$ can be equal to +1 (protonation case), -1 (deprotonation case), or 0 (no change). The quantity $\Delta\phi^i_{I\to J}$ is the difference of the electrostatic potential generated by the chromophore in electronic states I and J, calculated at the position of the residue i as

$$\Delta\phi^i_{I\to J} = k^i_{\text{eff}} \sum_{a=1}^{N_a} \frac{Q^a_J - Q^a_I}{|\mathbf{r}_i - \mathbf{R}_a|} \tag{9}$$

This quantity can also be understood as the absolute change in the vertical transition energy due to a +1 charge change at residue i, that is, without any reference to the probability of this change occurring. In other words, $\Delta\phi^i_{I\to J}$ is a pH-independent quantity. According to

Equation (8), the transition energy of the chromophore interacting with its environment in a given protonation microstate p is obtained by adding a well-defined correction term to the transition energy for a reference microstate. This term expresses how the electrostatic interaction between the chromophore and a given residue changes upon its protonation or deprotonation.

Taking into account the weights w_p of the N_m possible protonation microstates at a given pH, the transition energy writes:

$$\Delta E_{I\to J}(\mathrm{pH}) = \sum_{p=1}^{N_m} w_p(\mathrm{pH})\Delta E^p_{I\to J} \tag{10}$$

This expression includes a sum over all the possible microstates. One way to reduce the number of terms in the sum is based on the weights: we can exclude all the microstates whose weight is below a given threshold at the considered pH. However, a microstate with a relatively small weight could significantly contribute to the transition energy because the pH-independent $\Delta E^p_{I\to J}$ term is large. Alternatively, inserting equation (8) in equation (10) leads to:

$$\Delta E_{I\to J}(\mathrm{pH}) = \Delta E^0_{I\to J} + \sum_{i=1}^{N} \Delta\phi^i_{I\to J} \sum_{p=1}^{N_m} w_p(\mathrm{pH})\alpha_{ip} \tag{11}$$

since the sum of all weights is equal to 1. At this point, we need to distinguish between protonation and deprotonation processes. When the residue i is protonated in the reference microstate, then α_{ip} is either 0 (no change) or -1 (deprotonation). In that case, $\sum_{p=1}^{N_m} w_p(\mathrm{pH})\alpha_{ip} = -x_i(\mathrm{pH})$, where $x_i(\mathrm{pH})$ is the deprotonation probability. Conversely, when the residue i is deprotonated in the reference microstate, then α_{ip} is either 0 (no change) or +1 (protonation), resulting in $\sum_{p=1}^{N_m} w_p(\mathrm{pH})\alpha_{ip} = 1 - x_i(\mathrm{pH})$. Hence, equation (11) can be recast by splitting the N residues in N_p protonated residues and N_d deprotonated residues in the reference microstate:

$$\Delta E_{I\to J}(\mathrm{pH}) = \Delta E^0_{I\to J} - \sum_{i=1}^{N_p} \Delta\phi^i_{I\to J} x_i(\mathrm{pH}) + \sum_{i=1}^{N_d} \Delta\phi^i_{I\to J}(1 - x_i(\mathrm{pH})) \tag{12}$$

The application of this final MEM expression requires a reference transition energy $\Delta E^0_{I\to J}$, a N-long list of pH-dependent deprotonation probabilities x_i(pH) (deriving from their corresponding pK_a values) and a N-long list of differential electrostatic potentials $\Delta\phi^i_{I\to J}$. The latter information can typically be obtained from ARM, eventually justifying our proposition to merge MEM into ARM.

2.3 The MEM+ARM protocol

The scope of the present work is to extend the ARM protocol to include the MEM methodology to simulate the absorption spectrum of rhodopsin at a given pH. The spectrum is resolved in terms of the most relevant protonation microstates that contribute to the overall $\lambda_{\max}$ of absorption. Therefore, the role of the MEM methodology is to unravel these microstates based on the ARM calculation on an initial protonation microstate. Thus, the $\Delta E_{I\to J}$ provided by the MEM methodology (and the corresponding contributions due to different amino acids) are to be considered qualitative, whereas the actual spectrum is constructed by performing actual ARM calculations on top of all the protonation microstates suggested by the MEM analysis. In what follows we describe the MEM outcome when applied following a single initial ARM calculation, and subsequently the way it has been incorporated into the protocol, to automatically obtain the excitation energies of all the relevant microstates.

As mentioned in subsection 2.1, the ARM protocol provides all the information needed for running MEM: (i) a structure in a given protonation microstate, (ii) a set of pK_a values for all titratable residues, (iii) a set of chromophore atomic charges for each considered electronic state and (iv) the corresponding transition energies. While (i) and (iii) are used to compute the differential electrostatic potentials $\Delta\phi_{I\to J}$, (ii) is necessary for estimating the deprotonation probabilities x at the considered pH, and (iv) is the reference transition energy $\Delta E^0_{I\to J}$.

The MEM analysis provides the most relevant titratable residues in terms of their contributions to the overall $\Delta E_{I\to J}$ (Equation (12), once a suitable threshold has been chosen.

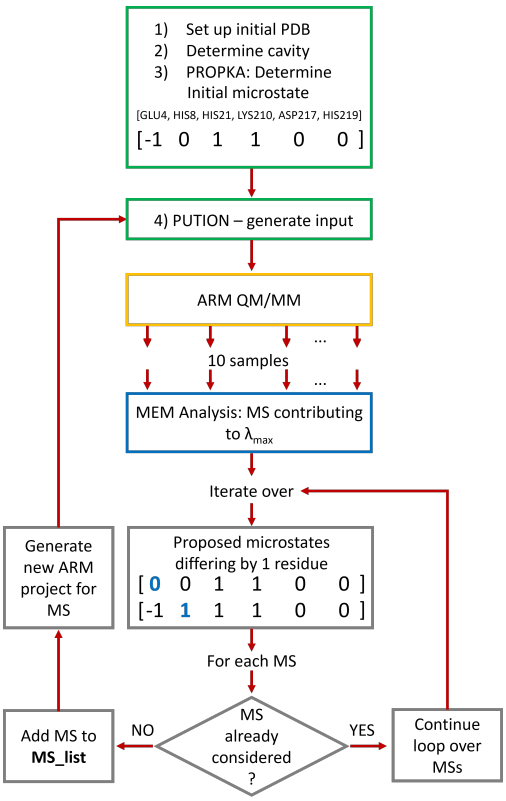

Figure 2: Schematic representation of the general ARM+MEM protocol. Each protonation microstate (MS) is coded in terms of a vector containing the charges of all titratable residues. In this example, the initial MS is a vector of 6 titratable residues of the ASR (in a real case, the length of the vector is 44). The protocol starts with an MS suggested by the ARM input generator (*via* PROPKA, green boxes) after which the standard ARM QMMM protocol is executed (see Figure 1 for reference). A MEM analysis is then performed for each resulting sample (blue). Subsequently, the protocol iterates over the MSs suggested by the MEM analysis differing in the protonation of one residue (gray boxes), and if the new MS has not already been accounted for, it is saved in a **MS_list** variable and the protocol generates a new ARM project for the new MS and executes again the ARM+MEM analysis.

This effectively reduces the number of potentially relevant protonation microstates by reducing the dimensionality of the titratable residue space. Besides the relevant MS, the MEM analysis also provides their corresponding weights (see Equation 13 below). In this work, each protonation microstate is characterized by a vector containing the total charges of all the relevant amino acids (ordered by index) at a specific protonation state. For example, suppose that the ARM protocol applied on the *Anabaena* Sensory Rhodopsin suggests as the (initial) protonation microstate one in which GLU4, HIS8, HIS21, LYS210, ASP217 and HIS219 have charges -1, 0, 1, 1, 0, 0, respectively (represented as the vector [-1 0 1 1 0 0] in Figure 2). Then, once the standard ARM protocol has been applied and has generated 10 samples, a MEM analysis applied on each of these samples will provide the energy contribution to $\Delta E_{I\rightarrow J}$ due to each amino acid (and hence will provide the most relevant amino acids), and a list of all of the possible protonation microstates (in this case, 2^6) with their corresponding weights (or probabilities). The weight of each microstate is computed as the product of the molar fractions of all of the relevant amino acids that define the microstate, each molar fraction being x_i(pH) (*i.e.* the deprotonation probability) if the i^{th} residue appears deprotonated, or $1 - x_i$(pH), if it instead appears protonated. Thus, for example, the weight of [0 0 1 1 0 0] will be given by

$$w_{[001100]} = x_{\mathrm{GLU4}} x_{\mathrm{HIS8}} (1 - x_{\mathrm{LYS210}})(1 - x_{\mathrm{ASP217}}) x_{\mathrm{HIS219}} \tag{13}$$

where the pH dependency of the molar fractions has been omitted for clarity.

Since MEM provides the energy shifts associated with the (de)protonation of one amino acid at a time, only microstates that differ by one protonation state of a specific amino acid with respect to the reference (ARM) microstate - singly (de)protonated as shorthand - can be reliably considered. As a result, to account for all of the relevant microstates at a specific pH, an iterative procedure needs to be adopted, whereby after each ARM+MEM analysis, each of the singly (de)protonated microstates suggested undergoes a new ARM+MEM

protocol, which provides access to "doubly" deprotonated microstates. The procedure is repeated until 1) all of the possible microstates within the reduced space of amino acids have been considered, or 2) all of the microstates above a previously defined threshold have been accounted for, see Figure 2. Notice that when a subsequent ARM+MEM iteration is performed on a microstate other than the initial reference, the input generator is executed, but the default PROPKA execution is ignored, as the protonation MS it suggests is superseded by the selection of the protonation MS of interest of the current iteration. This is exemplified in Figure 2 by indicating that the input generation starts at step 4 – all of the parameters of the input of the initial reference are kept, apart from the protonation microstate.

The number of iterations of the loop shown in Figure 2 is not infinite, since only a finite amount of singly (de)protonated microstates is suggested after the MEM analysis. Each new MS is stored in a **MS_list** variable, so that microstates that appear in subsequent ARM+MEM iterations but have already been stored in **MS_list** are ignored. In this way, an ARM+MEM calculation on the same microstate is only performed once. A chosen test case (subsection 4.1) exemplifies such a behavior.

Finally, note that each of these extra ARM calculations results in a new transition energy, provided by the average of 10 replicas per microstate. Since the weights of the corresponding protonation microstates are already known – Equation (13), the construction of the total spectrum is straightforward, assuming a Gaussian shape for each electronic transition. At a given pH, the absorbance A at a given energy $\omega = \Delta E_{I \to J}$ is computed as a weighted sum of contributions due to n_m protonation microstates ($n_m < N_m$):

$$A(\omega; \mathrm{pH}) = \sum_{p=1}^{n_m} \bar{f}_p(\mathrm{pH}) \exp\left(-\frac{(\omega - \bar{\omega}_p)^2}{2a} \right) \tag{14}$$

in which average oscillator strengths $\bar{f}_p$ and transition energies $\bar{\omega}_p$ are introduced for each

protonation microstate p:

$$\bar{f}_p(\mathrm{pH}) = w_p(\mathrm{pH})\frac{1}{N_s^p}\sum_{s=1}^{N_s^p} f_s^p \tag{15}$$

and

$$\bar{\omega}_p = \frac{1}{N_s^p}\sum_{s=1}^{N_s^p} \omega_s^p \tag{16}$$

N_s^p denotes the number of ARM structures (usually 10) obtained in each microstate, each of them being characterized by a CASPT2 transition energy ω_s^p and its SA-CASSCF oscillator strength f_p^s. Note that $w_p(\mathrm{pH})$ represents the normalized weight of the microstate p (*i.e.* $\sum_p w_p = 1$), as only a reduced set of protonation microstates is considered.

3 Computational Details

All of the computations involving the ARM protocol were performed using a development version of the ARM[55] package. For the simulation of the pH-dependent spectrum of *Anabaena* Sensory Rhodopsin, both the 1XIO[79] and the 2M3G[80] were used as the initial structures for the ARM protocol. As described in subsection 2.1, each ARM calculation involves a single MM energy minimization, followed by 10 classical MD relaxations - in both cases using the GROMACS[71] software - and QM/MM geometry optimizations, in which the cavity residues and the rPSB moiety are relaxed. The system is described by the AMBER94 force field, except for the rPSB, whose force field parameters correspond to those developed and employed in previous works.[51,69,81] The MD relaxation consists of a 50 ps heating following a 150 ps equilibration and an 800 ps production simulations, in all cases in the NVT ensemble using a Berendsen thermostat.[82] The final structures of the MD simulations are used as the initial guesses for the subsequent QM/MM geometry optimizations, in which rPSB (53 atoms) is described quantum mechanically and the rest of the system is described by the AMBER94

force field. The QM/MM boundary is treated using a link atom approach,[83,84] in which a hydrogen atom is placed between the C_δ and C_ϵ atoms of the lysine moiety, where C_ϵ is a QM atom. The interaction between the QM and the MM subsystems was accounted for by employing the electrostatic potential fitting (ESPF) metholodogy.[50]

The QM/MM protocol itself consists of a sequence of geometry optimizations at the HF/3-21G/MM, then at the complete active space self-consistent field (CASSCF) level of theory - CASSCF(12,12)/3-21G/MM first, then CASSCF (12,12)/6-31G*/MM. The CASSCF active space includes the retinal full π system. The final step in the standard ARM protocol[55,69] consists of a three-root State Average (SA) CASSCF(12,12)/6-31G*/MM and a subsequent CASPT2 single point (SP) calculation; for the latter, an imaginary shift of 0.2 Ha was employed to avoid the presence of intruder states, and the IPEA was set to 0.0 Ha, as recommended in the literature.[85]

As the MEM methodology assumes that the point charges of the rPSB are computed in a vacuum, an extra SP three-root SA-CASSCF(12,12)/6-31G*/MM calculation is performed on top of the optimized geometry to retrieve such charges. Thus, the input for the MEM step in the ARM+MEM protocol (blue boxes in Figure 2, right) consists of the CASPT2 vertical excitation energies, the rPSB point charges in a vacuum, and the charges of the titratable residues that define a given protonation microstate. In the MEM analysis, we have applied an Energy threshold of 0.01 kcal/mol. For all the geometry optimizations and the SP calculations involved in the ARM protocol, the OpenMolcas[86] software was used in conjunction with its interface with the Tinker molecular dynamics program.[87]

The rPSB geometry was extracted from one of the ARM CASSCF-optimized geometries for the toy model, and the OpenMolcas inputs were generated using the MoBioTools toolkit.[88]

4 Results and Discussion

4.1 Toy Model

This first application is meant to illustrate how the ARM+MEM protocol works. In order to reduce as much as possible the number of parameters, a toy model consisting of the retinal chromophore in the QM subsystem and three titratable residues in the MM subsystems (denoted as R_1, R_2, and R_3) was chosen. In this model, the three titratable amino acids are represented as three single-point charges whose values are either 0 (protonated case) or -1 (deprotonated case), in atomic units, so that the protonation space contains 8 microstates. In what follows, we will use the notation [n_1 n_2 n_3] ($n_i = 0$ or -1), indicating the protonation states of R_1, R_2 and R_3, respectively, to refer to each microstate. Thus, for example, the microstate [0 0 0] refers to the fully protonated situation, whereas [-1 0 0] corresponds to the situation whereby only R_1 is deprotonated. We have also chosen to bypass the geometry relaxation/optimization steps in ARM, *i.e.*, we are keeping frozen the geometry of the toy model, as shown in Figure 3. The results obtained in this section do not aim at simulating the pH-dependent spectrum of a real system, but instead exemplify the simplest case scenario in which only the variation of the protonation microstates is accounted for.

R_1 is located on the protonated Schiff base side of the rPSB chromophore: its deprotonation is expected to cause a λ_{max} blueshift.[89–91] R_2 is located on the β-ionone side of the retinal chromophore. Its deprotonation is expected to cause a λ_{max} redshift.[89,90,92] R_3 is located close to the middle of the retinal chromophore: the effect of its deprotonation cannot be foreseen easily. To assess the efficiency of the MEM analysis in unraveling the protonation microstates that contribute to the absorption spectrum of the rPSB moiety in the presence of the aforementioned titratable residues, we have arbitrarily fixed their pK_a values to 4.5, 5.0, and 6.0, and considered a range of pH values between 3 and 7. Within this pH range, we ensure that all the possible microstates occurring in our 8-sized protonation space are accounted for while considering the population of each microstate at each pH value in that

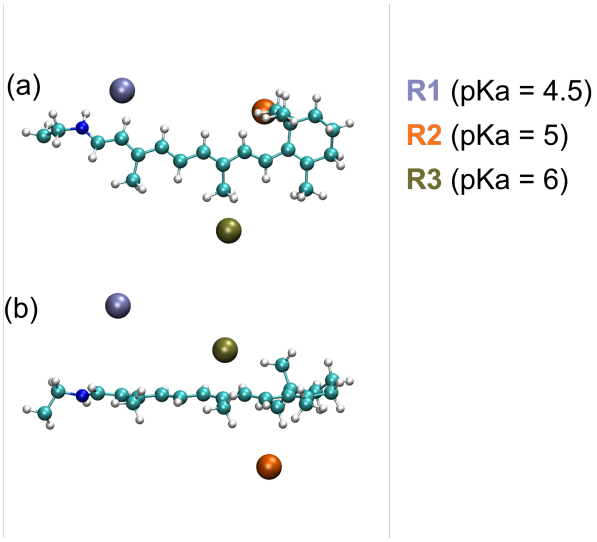

Figure 3: (a) Top and (b) side views of the toy model, featuring an all-*trans* rPSB chromophore and three titrated residues: the $\lambda_{\max}$ blueshifting R_1, the $\lambda_{\max}$ redshifting R_2 and R_3. The distances from the center of mass of the retinal are 8.06 Å, 6.01 Å, and 6.07 Å respectively

range, as can be evidenced from Table 1.

The original ARM protocol considers a single protonation microstate to generate the model of a rhodopsin protein at a given pH. Thus, for each pH value between 3 and 7, we will refer to that protonation state as the "reference" or "initial" microstate, in the context of subsequential applications of ARM and MEM. Regarding the study of our toy model, what should be an ARM+MEM iteration, it will be in practice a CASPT2 + MEM calculation since, as stated above, we are avoiding any geometry relaxation effects. By default, the reference microstate at a given pH value will correspond to the most populated one, so that according to Table 1, at pH=3 and pH=4 it will be [0 0 0], at pH=5 either [-1 0 0] or [-1 -1 0] – in this case, the selection is entirely arbitrary – at pH=6 either [-1 -1 0] or [-1 -1 -1], and at pH=7 it will be [-1 -1- 1]. We will start by considering the outcome of the first ARM+MEM iteration.

Figure 4 (left) shows this situation for the range of pH values under consideration: for each reference microstate, the MEM analysis provides a shift of the reference $S_0 \rightarrow S_1$ transition energy due to single (de)protonation – in other words, each of the weighted $\Delta\phi^i_{S_0 \rightarrow S_1}$ factors

Table 1: Protonation microstate populations at pH values between 3 and 7. In the first 3 columns, the label 0 indicates the corresponding residue is protonated (*i.e.* electrically neutral), and the label -1 means it is deprotonated (*i.e.* electrically negatively charged). The deprotonation probability of a single residue R_i is given by $x_i = (1 + 10^{pK_{a,i} - pH})^{-1}$ and its protonation probability is simply $1 - x_i$.

R_1	R_2	R_3	pH=3	pH=4	pH=5	pH=6	pH=7
0	0	0	0.96	0.68	0.11	0.00	0.00
0	0	-1	0.00	0.00	0.01	0.00	0.00
0	-1	0	0.00	0.07	0.11	0.01	0.00
0	-1	-1	0.00	0.00	0.01	0.01	0.00
-1	0	0	0.03	0.22	0.35	0.04	0.00
-1	0	-1	0.00	0.00	0.03	0.04	0.00
-1	-1	0	0.00	0.02	0.35	0.44	0.09
-1	-1	-1	0.00	0.00	0.03	0.44	0.90

present in either sum (depending on whether it is a protonation or deprotonation event) in Equation (12). With this information, one can determine which protonation microstates need to be accounted for to simulate the absorption spectrum of the chromophore at a given pH. Given the energy threshold of 0.01 kcal/mol, it can be evidenced that at pH=4 (ref. [0 0 0]), both [-1 0 0] and [0 -1 0] induce some relevant energy shifts. Therefore, both microstates need to be considered for a subsequent ARM+MEM iteration.

From the results in Figure 4 (left), it can be observed that the MEM analysis qualitatively predicts the CASPT2 transition energy shifts following the (de)protonation of specific residues in our toy model; for example, the deprotonation of R_1 in [0 0 0] induces a blueshift (and its protonation in [-1 0 0] a redshift, see pH=5), whereas the deprotonation of R_2 induces a redshift. However, these energy shifts are not quantitative. In order to obtain accurate excitation energies due to each of these microstates, a further ARM+MEM iteration needs to be performed using each of them as a reference. These subsequent iterations are shown in Figure 4 (center, right). For example, the results obtained using the microstates suggested by MEM as "secondary" references during the first iteration appear in Figure 4 (center). Noticeably, in the case of pH values 4, 5, and 6, it can be evidenced that new relevant microstates emerge following the second ARM+MEM iteration; for example, with [-1 0 0] as

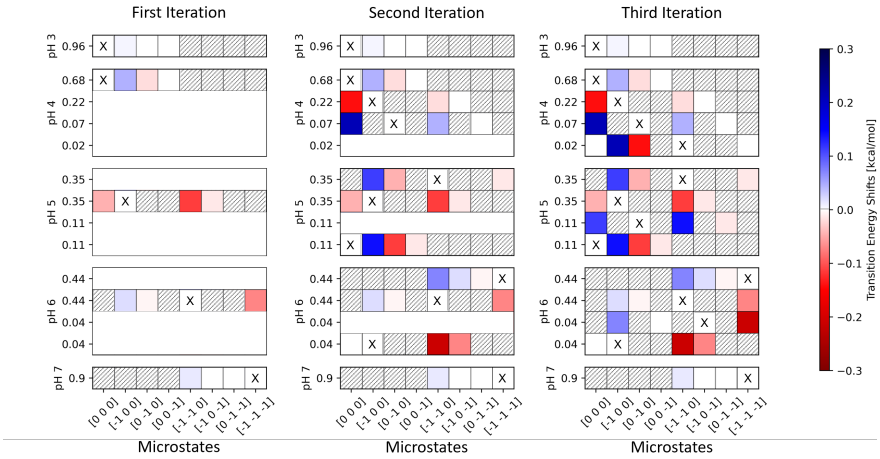

Figure 4: Schematic representation of the outcome of an ARM+MEM calculation applied to the toy model. Each table represents an iteration of the ARM+MEM protocol, using as reference the microstate marked with an X. The 8 possible microstates are represented on the x axis of each table. Given a reference microstate, the filled dashed squares represent microstates that are not accessible by the MEM protocol at a given iteration. For each plot, the y axis gives the reference microstate probabilities (weights) at a given pH. The heatmap colors represent redshift or blueshift of the $S_0 \rightarrow S_1$ transition with respect to the reference microstate in kcal/mol).

the reference microstate at pH=4, the deprotonation of R2 induces a redshift, so that [-1 -1 0] would need to be accounted for in a subsequent ARM+MEM iteration. Similar situations can be evinced at pH=5 and pH=6 – Figure 4 (center). At the third iteration – Figure 4 (right) – at pH values 4, 5, and 6, only microstates that have already been considered in the previous iterations induce relevant energy shifts. Thus, no further iterations are required, *i.e.*, the ARM+MEM protocol has attained convergence.

In the case in which a microstate that stems from a multiple (de)protonation of a given reference induces a relevant energy shift, a single ARM+MEM iteration would not suffice to determine all the relevant microstates at a given pH. That was the case, for example, at pH=5, whereby the initial reference microstate was [-1 0 0], but the energy shift of [0 -1 0] was unraveled only after the second iteration (Figure 4, left and center, respectively). The reason can be evinced from the fact that all of the weighted energy terms present in both sums of Equation 12 are ascribed to single (de)protonation events (*e.g.*, [0 0 0] to [-1

0 0]), whereas the energy shifts due to multiple simultaneous (de)protonation events (*e.g.*, [-1 0 0] to [0 -1 0], as in the case of pH=5) are not directly accessible (unless one assumes that the energy shifts due to individual protonation changes are additive, which need not be the case). This feature, indeed, is the main advantage of the ARM+MEM approach. The microstate subspace is progressively expanded by adding only the most important titratable amino acids, *i.e.*, only the ones that will contribute significantly to the transition energy at a given pH value. This is particularly important when one remembers that the total protonation space is as large as 2^N (N is the total number of titratable residues).

The final step to describe the absorption spectrum of the retinal chromophore in the presence of different titrated residues at a given pH is to consider the excitation energies and oscillator strengths due to all relevant protonation microstates. In order to account for the populations of the different microstates at each pH value, we weigh each oscillator strength by the corresponding population probability. Since we have considered a range of pH values, it is possible to obtain a pH-dependent absorption spectrum for the toy model, as can be evidenced from Figure 5.

It can be seen that at pH=3, the only relevant microstate is [0 0 0], so the spectrum consists of a single absorption band (or stick, in the case the band is not computed via a broadening or sampling technique). As the pH increases, the contributions of different microstates become relevant. For example, at pH=4 one can evidence the contribution of [-1 0 0], whose excitation energy is blueshifted with respect to that of [0 0 0], as expected. This result had also been qualitatively anticipated by the first MEM analysis at pH=4 (Figure 4 left). At pH=5, it can be observed the contributions due to all four microstates suggested by the MEM analysis. In particular, one can evidence the contribution of [0 -1 0], which is strongly redshifted with respect to [0 0 0]. At pH=6 the contributions due to [-1 0 0] and [0 -1 0] are no longer present, whereas a band due to [-1 -1 -1] can now be observed. At pH=7 the only contribution to the spectrum is due to the fully deprotonated microstate [-1 -1 -1].

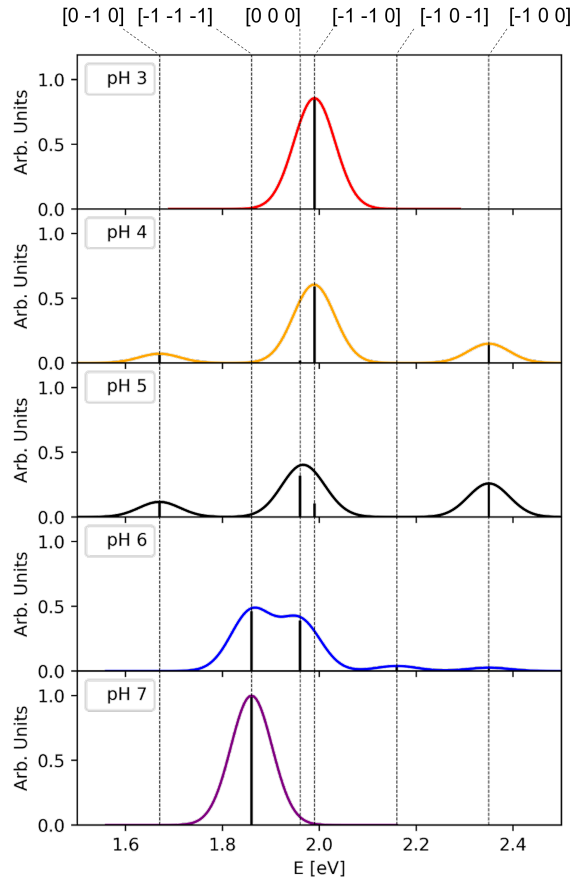

Figure 5: pH-dependent $S_0 \rightarrow S_1$ absorption spectrum of the toy model at the CASPT2(12,12)/6-31G* level of theory, between pH=3 and pH=7. The spectrum is resolved to the contributions of all the relevant microstates (black solid bars) at each pH value. The spectra have been broadened using Gaussian functions of full width at half maximum of 0.1 eV. The spectra have been normalized by the oscillator strength at pH=7 (1.11). All the relevant microstates are shown at the top of the picture, and their excitation energies are evidenced by vertical dashed lines for comparison purposes.

4.2 *Anabaena* Sensory Rhodopsin

ASR is a photoactive transmembrane protein in which the retinal chromophore is found in two configurations, all-*trans* (AT) and 13-*cis* (13C) along the protein photocycle. ASR absorption maximum energy features a tiny but sizable 2 nm redshift between pH=3 and pH=5, then a larger 6 (AT) to 10 nm (13C) blueshift between pH=5 and basic pH.[38] Some of us succeeded in reproducing such a trend using the computationally expensive CpHMD-then-QMMM protocol.[93] Moreover, we were able to highlight which titratable amino acids are responsible for these pH-induced variations.

In Ref. 93, the $\lambda_{\max}$ value was exclusively determined in terms of contributions to the excitation energy due to (de)protonation events, without attempting to simulate the line-shape of the absorption spectrum. In the present work, we simulate the transition spectrum by considering the contribution of all the relevant microstates not only in terms of their excitation energies but also of their oscillator strengths (section 2.3) In particular, notice that the latter depend on the pK_a values *via* Equations 15 and 13. Although not as high in quality as those obtained using CpHMD, in the present work we use the values obtained by PROPKA, which is executed on the fly when running the ARM+MEM protocol. Thus, in what follows we test whether this protocol is capable of reproducing the pH-dependent behavior of ASR for both the AT and the 13C retinal chromophore configurations.

The ASR structures used as the initial input for the ARM+MEM protocol consist of 1XIO,[79] obtained using X-ray crystallography experiments at pH=5.6 (Figure 6(A)) and 2M3G,[80] obtained employing NMR experiments at pH=9 (Figure 6(B)). The cavity, which consists of 20 residues surrounding the rPSB moiety, is the portion of the system (along with the rPSB itself) that will be optimized during the initial ARM protocol and subsequent ARM+MEM iterations. Therefore, most of the protein residues will keep their initial crystallographic structures. In what follows the rPSB moiety will always be assumed to be in the protonated state along the range of pH values considered, since it has been observed that in ASR the pK_a of the deprotonation transition is above 12.5.[38]

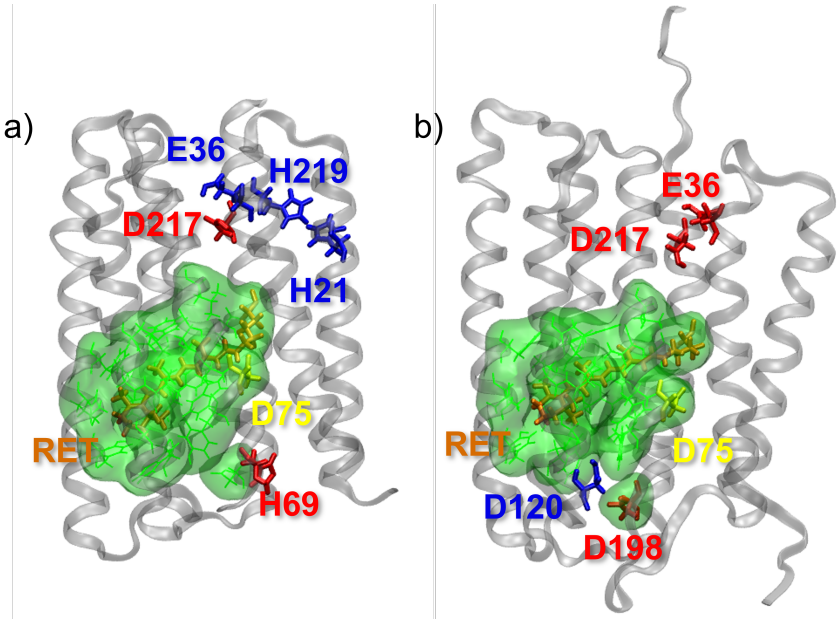

Figure 6: Schematic representation of the structure of the ASR model, obtained from a) the 1XIO[79] crystallographic structure and b) the 2M3G[80] NMR structure. The displayed ribbon representation of the backbone suggests a substantial structural difference. In both cases, the rPSB moiety, the ASP75 counterion and the cavity are represented in orange, yellow, and green, respectively. The cavity consists of the following residues: TYR73, ASP75, TRP76, THR80, LEU83, GLN109, VAL112, LEU113, GLY116, TYR132, GLY135, VAL136, PHE139, TRP176, TYR179, PRO180, TRP183, ASP198, SER209 and LYS210. The most relevant residues that induce significant excitation energy shifts according to the present work are represented in blue (blueshifting) and red (redshifting).

Figure 7 shows a comparison between experimental and the 1XIO-based ARM+MEM calculated absorption spectra of ASR with the retinal chromophore in its AT and 13C configurations, at different pH values.

Both the AT and the 13C calculated spectra are in good agreement with the experimental lineshapes at pH=5 (orange), presenting $\lambda_{\max}$ differences of 5.8 and 14.3 nm, respectively, from the experimental $\lambda_{\max}$. At pH=7 (blue) and pH=9 (purple), the agreement with the experimental $\lambda_{\max}$ worsens. The fact that the discrepancies with the experimental spectra increase as one moves further away from the crystallographic pH (of 5.6 for 1XIO) suggests two possible explanations: (i) structural changes are important when moving towards higher pH values, (ii) PROPKA pK_a predictions at pH=7 and pH=9 are not of the same quality as they are at pH=5. In spite of this, the ARM+MEM protocol correctly predicts a blueshift

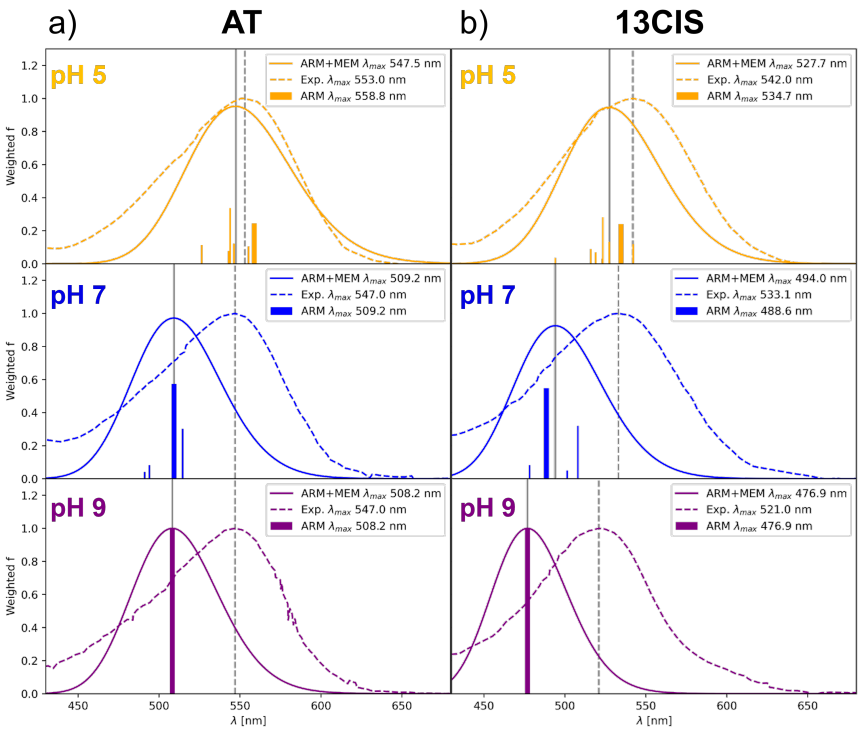

Figure 7: Experimental[38] and simulated spectra (*via* ARM+MEM) of a) all-trans (AT) and b) 13-cis (13C) ASR at pH=5 (orange), 7 (blue) and pH=9 (purple), starting from the 1XIO[79] structure. The vertical bars represent each of the average excitation energy and oscillator strength (Equations 15 and 16) of a given protonation MS. The gray vertical full and dashed lines correspond to the ARM+MEM and experimental $\lambda_{\max}$. A full width at half maximum of 0.3 eV (before converting to nm) was used to construct the Gaussian broadening. The weighted oscillator strengths have been normalized for comparison purposes.

for AT between pH=5 and pH=7, followed by a negligible variation of $\lambda_{\max}$ between pH=7 and pH=9 - Figure 7(a), and a consistent blueshift between pH=5 and pH=9 in the case of 13C - Figure 7(b). Noteworthy, in the case of the stand-alone ARM protocol,[55] the calculated $\lambda_{\max}$ was associated with a single protonation MS, which is the same that in the ARM+MEM protocol is denoted as the initial reference MS - see the thicker vertical bars in Figure 7, labeled as ARM $\lambda_{\max}$. The application of the ARM+MEM protocol however clearly evidences that a single protonation MS does not suffice to properly describe the lineshape of the absorption spectrum of a given rhodopsin. This is particularly evident by the simulated spectra of ASR (1XIO) at pH=5, whereby up to 6 (8) protonation MS have been shown to be relevant to describe the overall spectrum of the AT (13C) configuration of

the rPSB (Figure 7, orange and Table 2). It can also be evidenced from Figure 7 and Table 2 that the initial reference MS suggested by ARM (by applying PROPKA on top of the initial PDB structure) need not be the most populated one. This is due to the fact that the pK_a values used to determine the population of each MS are obtained by applying PROPKA during the MEM analysis on top of the CASSCF optimized geometries, and thus need not coincide with the pK_a values computed at the beginning of the protocol (Figure 2). That further emphasizes the importance of accounting for different protonation microstates and their relative populations.

Another source of information provided by the ARM+MEM protocol is obtained, as in the case of the toy model (subsection 4.1), by analyzing the character of the protonation MS whose excitation energies are blue or redshifted with respect to the initial reference MS. This is particularly important when considering singly (de)protonated microstates, as in those cases one can directly associate the (de)protonation event of a specific amino acid with a blueshifting or redshifting character. Table 2 contains the average energy, the weighted oscillator strength, and the population of all the relevant MS for AT (top) and 13C (bottom), at pH values 5, 7, and 9. In that table, each MS is described as a vector of the amino acids HIS8, HIS21, GLU36, HIS69, ASP198, ASP217, and HIS219 ([H8 H21 E36 H69 D198 D217 H219] for shorthand), whereby at each position it is reported the charge of each amino acid (-1, 0 or 1). Thus, for example, at pH=5 the initial reference MS corresponds to [0 1 0 0 0 0 1], and it can be evinced from Table 2 that the deprotonation of H21, H219, and the protonation of H69 ([0 0 0 0 0 0 1], [0 1 0 0 0 0 0] and [0 1 0 1 0 0 1], respectively) induce a blueshift with respect to the reference. Likewise, at pH=7 it can be evidenced that the deprotonation of E36 induces a blueshift, whereas the protonation of D217 induces a redshift. These blueshifting and redshifting amino acids are evidenced in the schematic representation of the 1XIO structure of ASR in Figure 6(a).

The discrepancy between the 1XIO-based computed spectra and the experimental ones at pH=7 and pH=9 could be either attributed to the lack of pH adaptation of the ASR

Table 2: Average excitation energies (in nm), oscillation strengths, and weights of the microstates that contribute to the overall absorption spectra of AT (top) and 13C (bottom) ASR at pH values 5, 7, and 9, using the 1XIO structure. The microstates are represented following the vector notation introduced in section 2.3, whereby the residues considered are either neutral or have relevant contributions to the overall λ_{max} following a (de)protonation event. For each considered pH the initial (ARM) reference MS is indicated by an asterisk.

1XIO (exp. ASR structure at pH=5.6)

pH	Microstate [H8 H21 E36 H69 D198 D217 H219]	$\bar{\omega}_p$ [nm]	$\bar{f}_p$(pH)	w_p
	AT			
5	[0 0 0 0 0 0 0]	526.3	0.15	0.12
	[0 0 0 0 0 0 1]	542.9	0.10	0.08
	[0 1 0 0 0 0 0]	543.9	0.44	0.34
	[0 1 0 1 0 0 0]	546.3	0.16	0.12
	[0 1 0 1 0 0 1]	555.3	0.14	0.10
	[0 1 0 0 0 0 1]*	558.8	0.32	0.23
7	[0 0 -1 0 -1 0 0]	491.1	0.05	0.04
	[0 0 -1 0 -1 -1 0]	494.0	0.09	0.09
	[0 0 0 0 -1 -1 0]*	509.2	0.64	0.57
	[0 0 0 0 -1 0 0]	514.5	0.34	0.30
9	[0 0 -1 0 -1 -1 0]*	512.5	1.15	1.00
	13C			
5	[0 0 0 1 0 0 0]	511.5	0.04	0.04
	[0 0 0 0 0 0 0]	516.1	0.10	0.09
	[0 0 0 0 0 0 1]	519.1	0.07	0.07
	[0 0 0 1 0 0 1]	523.2	0.03	0.03
	[0 1 0 0 0 0 1]	523.5	0.30	0.29
	[0 1 0 1 0 0 0]	527.4	0.14	0.13
	[0 1 0 0 0 0 1]*	534.7	0.26	0.23
	[0 1 0 1 0 0 0]	542.3	0.13	0.11
7	[0 0 -1 0 -1 -1 0]	478.2	0.08	0.09
	[0 0 0 0 -1 -1 0]*	488.6	0.51	0.56
	[0 0 -1 0 -1 0 0]	501.5	0.05	0.05
	[0 0 0 0 -1 0 0]	508.0	0.30	0.30
9	[0 0 -1 0 -1 -1 0]*	476.9	0.86	1.00

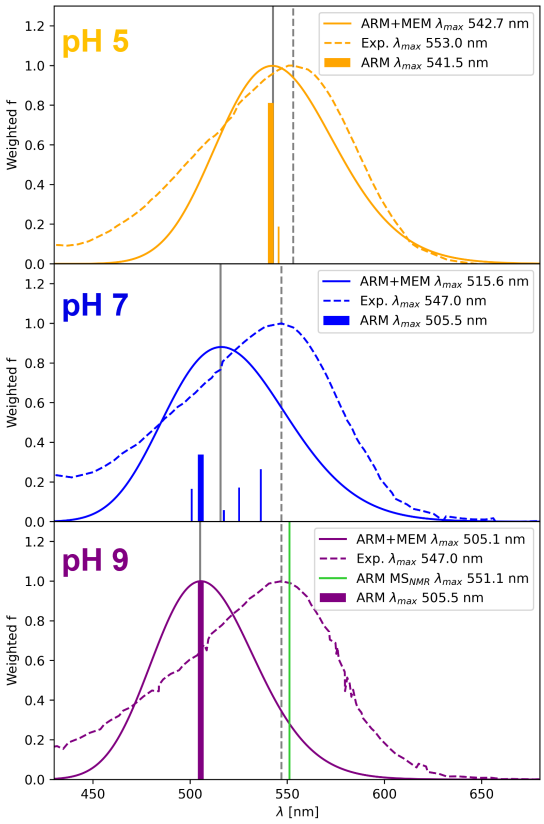

Figure 8: Experimental[38] and simulated spectra (*via* ARM+MEM) of all-trans (AT) ASR at pH=5 (orange), pH=7 (blue) and pH=9 (purple), starting from the 2M3G[80] structure. The vertical bars represent each of the average excitation energy and oscillator strength (Equations 15 and 16) of a given protonation MS. The gray vertical full and dashed lines correspond to the ARM+MEM and experimental $\lambda_{\max}$. The green bar represents an ARM calculation using an MS in accordance with the NMR structure. A full width at half maximum of 0.3 eV (before converting to nm) was used to construct the Gaussian broadening. The weighted oscillator strengths have been normalized for comparison purposes.

structure and/or the bad quality of some predicted pK_a based on the structures at high pH values. In order to disentangle these two effects, we consider another experimental structure, 2M3G, as an alternative input for ARM+MEM. The latter structure has been obtained at pH=9 and provides positions of hydrogen atoms. Accordingly, it should result in a more accurate absorption spectrum at this pH value. Figure 8 shows the experimental[38] and the ARM+MEM calculated absorption spectra of the AT configuration of the retinal protonated Schiff base in ASR using 2M3G (model 6a, the latter showing the smallest protein backbone root mean square deviation with respect to 1XIO, 3.2 Å). The same cavity was used as in the case of the 1XIO-based results so as to have the initial structure as the sole difference between the two sets of results. While the experimentally reported pH-induced blueshift is reproduced, the $\lambda_{\max}$ position is clearly off by more than 40 nm. In the default ARM protocol,[55] the initial structure is considered without hydrogen atoms (even if they are present in the PDB file), and these are added only at a subsequent step, once the pK_a values - and thus, the protonation state of all the titratable amino acids - have been determined. This is important because in the 2M3G structure, hydrogen atoms are present, and in particular it shows histidines H8, H21, H69, and H219 protonated; the corresponding MS would be [1 1 -1 1 -1 -1 -1 1], see Table 3 for reference. However, according to the MS populations derived from PROPKA during the ARM+MEM protocol, no MS having these four histidines protonated is relevant and thus, does not appear among the microstates that contribute to the overall spectrum at pH=9. Considering all the alternative conformations also present in 2M3G, PROPKA pK_a values for these histidines are always lower than 8, *i.e.* they are deprotonated at pH=9 according to PROPKA. Nonetheless, an application of the ARM protocol using this MS provides a $\lambda_{\max}$ of 551.1 nm, much closer to the experimental value (Figure 8). Further confirmation is provided by the computation of the ASR spectrum at pH=9 based on the 1XIO structure and on the 2M3G protonation microstate: $\lambda_{\max}$ =551 nm. Accordingly, changing the input ASR structure results in a 0.1 nm $\lambda_{\max}$ difference only. These results evidence an important point at which the ARM+MEM protocol could

be subject to improvement, as in some cases - and for some initial structures - more accurate and/or reliable pK_a values than those obtained with PROPKA would need to be accounted for.

Table 3: Average excitation energies (in nm), oscillation strengths, and weights of the microstates that contribute to the overall absorption spectra of AT ASR at pH values 5, 7, and 9, using the 2M3G structure. The microstates are represented following the vector notation introduced in section 2.3, whereby the residues considered are either neutral or have relevant contributions to the overall $\lambda_{\max}$ following a (de)protonation event. For each considered pH the initial (ARM) reference MS is indicated by an asterisk.

2M3G (exp. ASR structure at pH=9)

pH	Microstate [H8 H21 E36 H69 D120 D198 D217 H219]	$\bar{\omega}_p$ [nm]	$\bar{f}_p$(pH)	w_p
	AT			
5	[0 1 0 0 0 0 0 1]*	541.5	1.18	0.81
	[0 1 0 0 0 -1 0 1]	545.6	0.27	0.19
7	[0 0 -1 0 0 -1 -1 0]	500.8	0.22	0.17
	[0 0 -1 0 -1 -1 -1 0]*	505.4	0.44	0.35
	[0 0 0 0 -1 -1 -1 0]	517.3	0.08	0.06
	[0 0 -1 0 0 -1 0 0]	525.1	0.22	0.16
	[0 0 -1 0 -1 -1 0 0]	536.4	0.34	0.25
9	[0 0 -1 0 -1 -1 -1 0]*	505.4	1.23	1.00

The interdependence between computed pK_a values and the ASR structure is further evidenced at pH=5, whereby the initial reference MS in the case of 2M3G has the residue ASP120 deprotonated (Table 3) at pH=9, whereas in the case of 1XIO, this amino acid is protonated at all pH values. This difference impacts the number of protonation MS ARM+MEM has to consider. Using 2M3G at pH=5, there are two relevant MS to account for the absorption spectrum, whereas, in the case of 1XIO (Table 2), there were 6 for the AT and 8 for the 13C configurations of the rPSB. In particular, one of the 2M3G microstates at pH=5 involves the deprotonation of D198, a situation not met in the case of 1XIO. This may be associated with the fact that on 2M3G, D198 is present in the cavity (Figure 6), so that its structure is being modified during the CASSCF geometry optimizations, unlike in the case of 1XIO.

Finally, and regardless of the above-mentioned limitations, it is noteworthy that at pH=5

and especially at pH=7, the current ARM+MEM protocol provides a slight improvement to the $\lambda_{\max}$ to those obtained solely with the default ARM protocol, as evidenced in Figure 8.

5 Conclusions and Perspectives

In this work, we have incorporated the minimal electrostatic model (MEM) analysis within the ARM[55] protocol to determine in a cost-efficient way the most relevant protonation microstates describing the absorption spectrum of rhodopsins at a given pH. This has been done by generalizing MEM – developed by some of us in the past[39] – to any reference MS besides the one associated with a fully protonated situation. The new ARM+MEM protocol allows for modeling the absorption spectrum of rhodopsins in those situations beyond the frequently adopted single microstate picture to properly describe the spectral lineshape.

The ARM+MEM protocol starts with an initial ARM execution for a reference microstate, followed by a MEM analysis to determine the microstates that induce an excitation energy shift above a predetermined threshold and have a non-negligible population. These microstates are singly (de)protonated with respect to the reference, so new ARM+MEM calculations are performed self-consistently until all relevant microstates are found. The final spectrum is reconstructed using the excitation energies and the (weighted) oscillator strengths of these microstates.

We applied the ARM+MEM protocol on a toy model with three fixed pKa titratable sites that can tune the excitation energy of the chromophore, showing that at certain pH values, the single protonation microstate approximation breaks down. Subsequently, we have simulated the pH absorption spectrum of ASR protein at pH values 5, 7, and 9, using the 1XIO (pH=5.6, retinal AT and 13C configurations) and the 2M3G (pH=9, retinal AT only) as the initial structures. A good agreement between the experiment and the ARM+MEM simulations was observed at pH=5, although the agreement worsened at higher pH values. Nonetheless, the expected blueshift between pH=5 and pH=7, and (in the case of 13C)

between pH=7 and pH=9 was correctly reproduced, without relying on computationally expensive pK_a determinations. Noteworthy, in some situations such as at pH=7 the $\lambda_{\max}$ obtained with the standard ARM protocol was improved by the ARM+MEM methodology. In the case of 2M3G, there was not a good agreement with the experiment at high pH values, in particular at pH=9. An ARM calculation using the MS with all the histidines protonated – as observed from the NMR experiments[80] – provides a closer agreement with the experimental value. This suggests that there is still room for improvement in the ARM+MEM protocol, in particular concerning the determination of the reference pK_a values.

The above-documented results suggest interesting perspectives. The first (i) point would be the study of the pH-dependent photoisomerization dynamics with initial conditions reflecting the population of the microstates. In other words, these initial conditions would be consistent with the best-computed absorption band. (ii) Given the observed structural deformation possibly induced by the change in pH (see figure 6), the ARM protocol could be further developed to produce a model containing a fully flexible protein and its environment (membrane), especially when the pH of interest is not close to the one at which the crystallographic structure has been experimentally obtained. As a final prospective, although the ARM+MEM protocol in its current state is dedicated to rhodopsin proteins, some extensions to arbitrary photoresponsive proteins are possible with small modifications to the source code of the ARM protocol itself, as the MEM precedure is essentially applicable to arbitrary proteins. In this case, the most demanding part would be the validation of such a general protocol using a sizable set of proteins, as was done for the default ARM protocol itself in the past.[55,69]

Nonetheless, for the most part, the ARM+MEM protocol in its current stage is capable of providing further insights into the underlying structure of the absorption spectra of rhodopsins in particular the influence of the (de)protonation of one or more titratable sites on the transition energy at a specific pH. This knowledge can provide useful information to be accounted for once certain properties that result in the (de)protonation of these groups

– *e.g.*, the pH, the application of an external electric field, etc. – are tuned, or whether mutations on these groups are to be performed.

Supporting Information Available

Further details on the ARM+MEM protocol, specifying its asynchronous execution, the choice of one 2M3G structure, the geometries, OpenMolcas input files and MEM input files for the toy system, and the ARM+MEM results of all the samples for all the calculations involving ASR are available.

Acknowledgements

The authors thank Dr. Laura Pedraza Gonzalez, Dr. Leonardo Barneschi, and Dr. Darío Barreiro Lage for useful discussions. This work has been supported by Agence Nationale de la Recherche, Grant number ANR-2021-CE11-0029-03 (Project ULTRArchae). Centre de Calcul Intensif d'Aix-Marseille is acknowledged for granting access to its high-performance computing resources. MO and NF are grateful to Prof. Roland Lindh for having supported their proposition (back in 2003!) to include a QM/MM model into the Molcas and OpenMolcas packages. A few years and mice later, the `ESPF` module was released. It is still used.

References

(1) Nilsson, D.-E. Photoreceptor Evolution: Ancient Siblings Serve Different Tasks. *Curr. Biol.* **2005**, *15*, R94–R96.

(2) Sharma, A. K.; Zhaxybayeva, O.; Papke, R. T.; Doolittle, W. F. Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. *Environ. Microbiol.* **2008**, *10*, 1039–1056.

(3) Miranda, M. R. M.; Choi, A. R.; Shi, L.; Bezerra, A. G., Jr; Jung, K.-H.; Brown, L. S. The Photocycle and Proton Translocation Pathway in a Cyanobacterial Ion-Pumping Rhodopsin. *Biophys. J.* **2009**, *96*, 1471–1481.

(4) Ernst, O. P.; Lodowski, D. T.; Elstner, M.; Hegemann, P.; Brown, L. S.; Kandori, H. Microbial and Animal Rhodopsins: Structures, Functions, and Molecular Mechanisms. *Chem. Rev.* **2014**, *114*, 126–163.

(5) Gushchin, I.; Shevchenko, V.; Polovinkin, V.; Borshchevskiy, V.; Buslaev, P.; Bamberg, E.; Gordeliy, V. Structure of the light-driven sodium pump KR2 and its implications for optogenetics. *FEBS J.* **2016**, *283*, 1232–1238.

(6) Soppa, J. Two hypotheses - one answer. *FEBS Lett.* **1994**, *342*, 7–11.

(7) Spudich, J. L.; Yang, C.-S.; Jung, K.-H.; Spudich, E. N. Retinylidene Proteins: Structures and Functions from Archaea to Humans. *Annu. Rev. Cell Dev. Biol.* **2000**, *16*, 365–392.

(8) Wald, G. The Molecular Basis of Visual Excitation. *Nature* **1968**, *219*, 800–807.

(9) Stoeckenius, W. Purple membrane of halobacteria: a new light-energy converter. *Acc. Chem. Res.* **1980**, *13*, 337–344.

(10) Motto, M. G.; Sheves, M.; Tsujimoto, K.; Balogh-Nair, V.; Nakanishi, K. Opsin shifts in bovine rhodopsin and bacteriorhodopsin. Comparison of two external point-charge models. *J. Am. Chem. Soc.* **1980**, *102*, 7947–7949.

(11) Ottolenghi, M.; Sheves, M. Synthetic retinals as probes for the binding site and photoreactions in rhodopsins. *J. Membr. Biol.* **1989**, *112*, 193–212.

(12) Liu, R. S. H.; Krogh, E.; Li, X.-Y.; Mead, D.; Colmenares, L. U.; Thiel, J. R.; Ellis, J.; Wong, D.; Asato, A. E. Analyzing the Red-Shift Characteristics of Azulenic, Naphthyl,

other Ring-Fused and Retinyl Pigment Analogs of Bacteriorhodopsin. *Photochem. Photobiol.* **1993**, *58*, 701–705.

(13) Yan, B.; Spudich, J. L.; Mazur, P.; Vunnam, S.; Derguini, F.; Nakanishi, K. Spectral Tuning in Bacteriorhodopsin in the Absence of Counterion and Coplanarization Effects (∗). *J. Biol. Chem.* **1995**, *270*, 29668–29670.

(14) Sakmar, T. P.; Menon, S. T.; Marin, E. P.; Awad, E. S. Rhodopsin: Insights from Recent Structural Studies. *Annu. Rev. Bioph. Biom.* **2002**, *31*, 443–484.

(15) 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.* **2013**, *16*, 1499–1508.

(16) Inagaki, H. K.; Jung, Y.; Hoopfer, E. D.; Wong, A. M.; Mishra, N.; Lin, J. Y.; Tsien, R. Y.; Anderson, D. J. Optogenetic control of Drosophila using a red-shifted channelrhodopsin reveals experience-dependent influences on courtship. *Nat. Methods* **2014**, *11*, 325–332.

(17) Engqvist, M. K.; McIsaac, R. S.; Dollinger, P.; Flytzanis, N. C.; Abrams, M.; Schor, S.; Arnold, F. H. Directed Evolution of Gloeobacter violaceus Rhodopsin Spectral Properties. *J. Mol. Biol.* **2015**, *427*, 205–220.

(18) Cronin, T. W.; Caldwell, R. L.; Marshall, J. Tunable colour vision in a mantis shrimp. *Nature* **2001**, *411*, 547–548.

(19) Wang, W.; Nossoni, Z.; Berbasova, T.; Watson, C. T.; Yapici, I.; Lee, K. S. S.; Vasileiou, C.; Geiger, J. H.; Borhan, B. Tuning the Electronic Absorption of Protein-Embedded All-*trans*-Retinal. *Science* **2012**, *338*, 1340–1343.

(20) Herwig, L.; Rice, A. J.; Bedbrook, C. N.; Zhang, R. K.; Lignell, A.; Cahn, J. K.; Renata, H.; Dodani, S. C.; Cho, I.; Cai, L. et al. Directed Evolution of a Bright

Near-Infrared Fluorescent Rhodopsin Using a Synthetic Chromophore. *Cell Chem. Biol.* **2017**, *24*, 415–425.

(21) Bedbrook, C. N.; Rice, A. J.; Yang, K. K.; Ding, X.; Chen, S.; LeProust, E. M.; Gradinaru, V.; Arnold, F. H. Structure-guided SCHEMA recombination generates diverse chimeric channelrhodopsins. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E2624–E2633.

(22) Morrow, J. M.; Castiglione, G. M.; Dungan, S. Z.; Tang, P. L.; Bhattacharyya, N.; Hauser, F. E.; Chang, B. S. An experimental comparison of human and bovine rhodopsin provides insight into the molecular basis of retinal disease. *FEBS Lett.* **2017**, *591*, 1720–1731.

(23) Marín, M. d. C.; Agathangelou, D.; Orozco-Gonzalez, Y.; Valentini, A.; Kato, Y.; Abe-Yoshizumi, R.; Kandori, H.; Choi, A.; Jung, K.-H.; Haacke, S. et al. Fluorescence Enhancement of a Microbial Rhodopsin via Electronic Reprogramming. *J. Am. Chem. Soc.* **2019**, *141*, 262–271.

(24) Barneschi, L.; Marsili, E.; Pedraza-González, L.; Padula, D.; De Vico, L.; Kaliakin, D.; Blanco-González, A.; Ferré, N.; Huix-Rotllant, M.; Filatov, M. et al. On the fluorescence enhancement of arch neuronal optogenetic reporters. *Nat. Commun.* **2022**, *13*, 6432.

(25) Palombo, R.; Barneschi, L.; Pedraza-González, L.; Padula, D.; Schapiro, I.; Olivucci, M. Retinal chromophore charge delocalization and confinement explain the extreme photophysics of Neorhodopsin. *Nat. Commun.* **2022**, *13*, 6652.

(26) Sugiura, M.; Singh, M.; Tsunoda, S. P.; Kandori, H. A Novel Color Switch of Microbial Rhodopsin. *Biochemistry* **2023**, *62*, 2013–2020.

(27) Kay, E. R.; Leigh, D. A.; Zerbetto, F. Synthetic Molecular Motors and Mechanical Machines. *Angew. Chem., Int. Ed.* **2007**, *46*, 72–191.

(28) Sinicropi, A.; Martin, E.; Ryazantsev, M.; Helbing, J.; Briand, J.; Sharma, D.; Léonard, J.; Haacke, S.; Cannizzo, A.; Chergui, M. et al. An artificial molecular switch that mimics the visual pigment and completes its photocycle in picoseconds. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 17642–17647.

(29) Gueye, M.; Manathunga, M.; Agathangelou, D.; Orozco, Y.; Paolino, M.; Fusi, S.; Haacke, S.; Olivucci, M.; Léonard, J. Engineering the vibrational coherence of vision into a synthetic molecular device. *Nat. Commun.* **2018**, *9*, 313.

(30) Govorunova, E. G.; Sineshchekov, O. A.; Li, H.; Spudich, J. L. Microbial Rhodopsins: Diversity, Mechanisms, and Optogenetic Applications. *Annu. Rev. Biochem.* **2017**, *86*, 845–872.

(31) Kandori, H.; Inoue, K.; Tsunoda, S. P. Light-Driven Sodium-Pumping Rhodopsin: A New Concept of Active Transport. *Chem. Rev.* **2018**, *118*, 10646–10658.

(32) Berndt, A.; Lee, S. Y.; Ramakrishnan, C.; Deisseroth, K. Structure-Guided Transformation of Channelrhodopsin into a Light-Activated Chloride Channel. *Science* **2014**, *344*, 420–424.

(33) Kandori, H. Retinal Proteins: Photochemistry and Optogenetics. *Bull. Chem. Soc. Jpn.* **2020**, *93*, 76–85.

(34) Kojima, K.; Shibukawa, A.; Sudo, Y. The Unlimited Potential of Microbial Rhodopsins as Optical Tools. *Biochemistry* **2020**, *59*, 218–229.

(35) de Grip, W. J.; Ganapathy, S. Rhodopsins: An Excitingly Versatile Protein Species for Research, Development and Creative Engineering. *Front Chem* **2022**, *10*, 879609.

(36) Deisseroth, K. Optogenetics. *Nat. Methods* **2011**, *8*, 26–29.

(37) Hoffmann, M.; Wanko, M.; Strodel, P.; König, P. H.; Frauenheim, T.; Schulten, K.; Thiel, W.; Tajkhorshid, E.; Elstner, M. Color Tuning in Rhodopsins: The Mechanism

for the Spectral Shift between Bacteriorhodopsin and Sensory Rhodopsin II. *J. Am. Chem. Soc.* **2006**, *128*, 10808–10818.

(38) Rozin, R.; Wand, A.; Jung, K.-H.; Ruhman, S.; Sheves, M. pH Dependence of Anabaena Sensory Rhodopsin: Retinal Isomer Composition, Rate of Dark Adaptation, and Photochemistry. *J. Phys. Chem. B* **2014**, *118*, 8995–9006.

(39) Stenrup, M.; Pieri, E.; Ledentu, V.; Ferré, N. pH-Dependent absorption spectrum of a protein: a minimal electrostatic model of Anabaena sensory rhodopsin. *Phys. Chem. Chem. Phys.* **2017**, *19*, 14073–14084.

(40) Fujimoto, K.; Hayashi, S.; Hasegawa, J.-y.; Nakatsuji, H. Theoretical Studies on the Color-Tuning Mechanism in Retinal Proteins. *J. Chem. Theory Comput.* **2007**, *3*, 605–618, PMID: 26637039.

(41) Ryazantsev, M. N.; Nikolaev, D. M.; Struts, A. V.; Brown, M. F. Quantum Mechanical and Molecular Mechanics Modeling of Membrane-Embedded Rhodopsins. *J. Membr. Biol.* **2019**, *252*, 425–449.

(42) Mroginski, M.-A.; Adam, S.; Amoyal, G. S.; Barnoy, A.; Bondar, A.-N.; Borin, V. A.; Church, J. R.; Domratcheva, T.; Ensing, B.; Fanelli, F. et al. Frontiers in Multiscale Modeling of Photoreceptor Proteins. *Photochem. Photobiol.* **2021**, *97*, 243–269.

(43) Warshel, A.; Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **1976**, *103*, 227–249.

(44) Senn, H. M.; Thiel, W. QM/MM Methods for Biomolecular Systems. *Angew. Chem., Int. Ed.* **2009**, *48*, 1198–1229.

(45) van der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular

Mechanics (QM/MM) Methods in Computational Enzymology. *Biochemistry* **2013**, *52*, 2708–2728.

(46) Warshel, A. Multiscale Modeling of Biological Functions: From Enzymes to Molecular Machines (Nobel Lecture). *Angew. Chem., Int. Ed.* **2014**, *53*, 10020–10031.

(47) Sousa, S. F.; Ribeiro, A. J. M.; Neves, R. P. P.; Brás, N. F.; Cerqueira, N. M. F. S. A.; Fernandes, P. A.; Ramos, M. J. Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2017**, *7*, e1281.

(48) Clemente, C. M.; Capece, L.; Martí, M. A. Best Practices on QM/MM Simulations of Biological Systems. *J. Chem. Inf. Model.* **2023**, *63*, 2609–2627.

(49) Field, M. J.; Bash, P. A.; Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry* **1990**, *11*, 700–733.

(50) Ferré, N.; Ángyán, J. G. Approximate electrostatic interaction operator for QM/MM calculations. *Chem. Phys. Lett.* **2002**, *356*, 331–339.

(51) Ferré, N.; Olivucci, M. Probing the Rhodopsin Cavity with Reduced Retinal Models at the CASPT2//CASSCF/AMBER Level of Theory. *J. Am. Chem. Soc.* **2003**, *125*, 6868–6869.

(52) Söderhjelm, P.; Husberg, C.; Strambi, A.; Olivucci, M.; Ryde, U. Protein Influence on Electronic Spectra Modeled by Multipoles and Polarizabilities. *Journal of Chemical Theory and Computation* **2009**, *5*, 649–658, PMID: 26610229.

(53) Olsen, J. M.; Aidas, K.; Kongsted, J. Excited States in Solution through Polarizable Embedding. *Journal of Chemical Theory and Computation* **2010**, *6*, 3721–3734.

(54) Bondanza, M.; Nottoli, M.; Cupellini, L.; Lipparini, F.; Mennucci, B. Polarizable embedding QM/MM: the future gold standard for complex (bio)systems? *Phys. Chem. Chem. Phys.* **2020**, *22*, 14433–14448.

(55) Pedraza-González, L.; De Vico, L.; Marıín, M. d. C.; Fanelli, F.; Olivucci, M. a-ARM: Automatic Rhodopsin Modeling with Chromophore Cavity Generation, Ionization State Selection, and External Counterion Placement. *J. Chem. Theory Comput.* **2019**, *15*, 3134–3152.

(56) Pedraza-González, L.; Barneschi, L.; Marszałek, M.; Padula, D.; De Vico, L.; Olivucci, M. Automated QM/MM Screening of Rhodopsin Variants with Enhanced Fluorescence. *J. Chem. Theory Comput.* **2023**, *19*, 293–310.

(57) Olsson, M. H. M.; Søndergaard, C. R.; Rostkowski, M.; Jensen, J. H. PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions. *J. Chem. Theory Comput.* **2011**, *7*, 525–537.

(58) Søndergaard, C. R.; Olsson, M. H. M.; Rostkowski, M.; Jensen, J. H. Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values. *J. Chem. Theory Comput.* **2011**, *7*, 2284–2295.

(59) Swails, J. M.; York, D. M.; Roitberg, A. E. Constant pH Replica Exchange Molecular Dynamics in Explicit Solvent Using Discrete Protonation States: Implementation, Testing, and Validation. *J. Chem. Theory Comput.* **2014**, *10*, 1341–1352.

(60) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667.

(61) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinf.* **2009**, *10*, 168.

(62) Po, H. N.; Senozan, N. M. The Henderson-Hasselbalch Equation: Its History and Limitations. *J. Chem. Educ.* **2001**, *78*, 1499.

(63) Moore, D. S. Amino acid and peptide net charges: A simple calculational procedure. *Biochem. Educ.* **1985**, *13*, 10–11.

(64) Reijenga, J.; van Hoof, A.; van Loon, A.; Teunissen, B. Development of Methods for the Determination of pKa Values. *Anal. Chem. Insights* **2013**, *8*, ACI.S12304.

(65) Humphrey, W.; Dalke, A.; Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.

(66) Lomize, A. L.; Pogozheva, I. D.; Lomize, M. A.; Mosberg, H. I. Positioning of proteins in membranes: A computational approach. *Protein Science* **2006**, *15*, 1318–1333.

(67) Lomize, M. A.; Pogozheva, I. D.; Joo, H.; Mosberg, H. I.; Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Research* **2011**, *40*, D370–D376.

(68) Lomize, A. L.; Todd, S. C.; Pogozheva, I. D. Spatial arrangement of proteins in planar and curved membranes by PPM 3.0. *Protein Science* **2022**, *31*, 209–220.

(69) Melaccio, F.; del Carmen Marín, M.; Valentini, A.; Montisci, F.; Rinaldi, S.; Cherubini, M.; Yang, X.; Kato, Y.; Stenrup, M.; Orozco-Gonzalez, Y. et al. Toward Automatic Rhodopsin Modeling as a Tool for High-Throughput Computational Photobiology. *J. Chem. Theory Comput.* **2016**, *12*, 6020–6034.

(70) Zhang, L.; Hermans, J. Hydrophilicity of cavities in proteins. *Proteins* **1996**, *24*, 433–438.

(71) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D. et al. GROMACS 4.5: a high-throughput

and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(72) Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. A complete active space SCF method (CASSCF) using a density matrix formulated super-CI approach. *Chem. Phys.* **1980**, *48*, 157 – 173.

(73) Shepard, R. *Adv. Chem. Phys.*; John Wiley & Sons, Ltd, 1987; Chapter 2, pp 63–200.

(74) Roos, B. O. *Adv. Chem. Phys.*; John Wiley & Sons, Ltd, 1987; Chapter 7, pp 399–445.

(75) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(76) Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-order perturbation theory with a CASSCF reference function. *J. Phys. Chem.* **1990**, *94*, 5483–5488.

(77) Andersson, K.; Malmqvist, P.; Roos, B. O. Second-order perturbation theory with a complete active space self-consistent field reference function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

(78) Andersson, K.; Roos, B. O. *Modern Electronic Structure Theory*; Chapter 2, pp 55–109.

(79) Vogeley, L.; Sineshchekov, O. A.; Trivedi, V. D.; Sasaki, J.; Spudich, J. L.; Luecke, H. Anabaena Sensory Rhodopsin: A Photochromic Color Sensor at 2.0 Å. *Science* **2004**, *306*, 1390–1393.

(80) Wang, S.; Munro, R. A.; Shi, L.; Kawamura, I.; Okitsu, T.; Wada, A.; Kim, S.-Y.; Jung, K.-H.; Brown, L. S.; Ladizhansky, V. Solid-state NMR spectroscopy structure

determination of a lipid-embedded heptahelical membrane protein. *Nat. Methods* **2013**, *10*, 1007–1012.

(81) Ferré, N.; Cembran, A.; Garavelli, M.; Olivucci, M. Complete-active-space self-consistent-field/Amber parameterization of the Lys296–retinal–Glu113 rhodopsin chromophore-counterion system. *Theor. Chem. Acc.* **2004**, *112*, 335–341.

(82) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(83) Singh, U. C.; Kollman, P. A. A combined ab initio quantum mechanical and molecular mechanical method for carrying out simulations on complex molecular systems: Applications to the $CH_3Cl + Cl^-$ exchange reaction and gas phase protonation of polyethers. *J. Comput. Chem.* **1986**, *7*, 718–730.

(84) Humbel, S.; Sieber, S.; Morokuma, K. The IMOMO method: Integration of different levels of molecular orbital approximations for geometry optimization of large systems: Test for n-butane conformation and SN2 reaction: $RCl + Cl^-$. *J. Chem. Phys.* **1996**, *105*, 1959–1967.

(85) Zobel, J. P.; Nogueira, J. J.; González, L. The IPEA dilemma in CASPT2. *Chem. Sci.* **2017**, *8*, 1482–1499.

(86) Fdez. Galván, I.; Vacher, M.; Alavi, A.; Angeli, C.; Aquilante, F.; Autschbach, J.; Bao, J. J.; Bokarev, S. I.; Bogdanov, N. A.; Carlson, R. K. et al. OpenMolcas: From Source Code to Insight. *J. Chem. Theory Comput.* **2019**, *15*, 5925–5964.

(87) Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **2018**, *14*, 5273–5289, PMID: 30176213.

(88) Cárdenas, G.; Lucia-Tamudo, J.; Mateo-delaFuente, H.; Palmisano, V. F.; Anguita-Ortiz, N.; Ruano, L.; Pérez-Barcia, A.; Díaz-Tendero, S.; Mandado, M.; Nogueira, J. J. MoBioTools: A toolkit to setup quantum mechanics/molecular mechanics calculations. *J. Comput. Chem.* **2023**, *44*, 516–533.

(89) Ernst, O. P.; Lodowski, D. T.; Elstner, M.; Hegemann, P.; Brown, L. S.; Kandori, H. Microbial and Animal Rhodopsins: Structures, Functions, and Molecular Mechanisms. *Chem. Rev.* **2014**, *114*, 126–163.

(90) Mohseni, M., Omar, Y., Engel, G. S., Plenio, M. B., Eds. *Quantum Effects in Biology*; Cambridge University Press, 2014; p 399.

(91) Inoue, K.; del Carmen Marín, M.; Tomida, S.; Nakamura, R.; Nakajima, Y.; Olivucci, M.; Kandori, H. Red-shifting mutation of light-driven sodium-pump rhodopsin. *Nat. Commun.* **2019**, *10*, 1993.

(92) Kato, H. E.; Kamiya, M.; Sugo, S.; Ito, J.; Taniguchi, R.; Orito, A.; Hirata, K.; Inutsuka, A.; Yamanaka, A.; Maturana, A. D. et al. Atomistic design of microbial opsin-based blue-shifted optogenetics tools. *Nat. Commun.* **2015**, *6*, 7177.

(93) Pieri, E.; Ledentu, V.; Sahlin, M.; Dehez, F.; Olivucci, M.; Ferré, N. CpHMD-Then-QM/MM Identification of the Amino Acids Responsible for the Anabaena Sensory Rhodopsin pH-Dependent Electronic Absorption Spectrum. *J. Chem. Theory Comput.* **2019**, *15*, 4535–4546.

TOC Graphic